

RASP: RELATION-AWARE SEMANTIC PRIOR FOR WEAKLY SUPERVISED INCREMENTAL SEGMENTATION

Subhankar Roy

LTCI, Télécom-Paris

Intitute Polytechnique de Paris

subhankar.roy@telecom-paris.fr

Riccardo Volpi, Gabrela Csurka, Diane Larlus

NAVER LABS Europe

Meylan, France

{name.lastname}@naverlabs.com

ABSTRACT

Class-incremental semantic image segmentation assumes multiple model updates, each enriching the model to segment new categories. This is typically carried out by providing expensive pixel-level annotations to the training algorithm for all new objects, limiting the adoption of such methods in practical applications. Approaches that solely require image-level labels offer an attractive alternative, yet, such coarse annotations lack precise information about the location and boundary of the new objects. In this paper we argue that, since classes represent not just indices but semantic entities, the conceptual relationships between them can provide valuable information that should be leveraged. We propose a weakly supervised approach that exploits such semantic relations to transfer objectness prior from the previously learned classes into the new ones, complementing the supervisory signal from image-level labels. We validate our approach on a number of continual learning tasks, and show how even a simple pairwise interaction between classes can significantly improve the segmentation mask quality of both old and new classes. We show these conclusions still hold for longer and, hence, more realistic sequences of tasks and for a challenging few-shot scenario.

1 INTRODUCTION

When working towards the real-world deployment of artificial intelligence systems, two main challenges arise: such systems should possess the ability to continuously learn, and this learning process should only require limited human intervention. While deep learning models have proved effective in tackling tasks for which large amounts of curated data as well as abundant computational resources are available, they still struggle to learn over continuous and potentially heterogeneous sequences of tasks, especially if supervision is limited.

In this work, we focus on the task of semantic image segmentation (SIS) (Csurka et al., 2022), where the goal is predicting the class label of each pixel in an image. A reliable and versatile SIS model should be able to seamlessly add new categories to its repertoire without forgetting about the old ones. Considering for instance a house robot or a self-driving vehicle with such segmentation capability, we would like it to extend its knowledge to new classes without having to retrain the segmentation model from scratch on the old ones. Such ability is at the core of continual learning research, the main challenge being to mitigate catastrophic forgetting of what has been previously learned (Parisi et al., 2019).

Most learning algorithms for SIS assume training samples with dense pixel-level annotations, an expensive and tedious operation. We argue that this is cumbersome and severely hinders continual learning; adding new classes over time should be an annotation friendly process. This is why, here, we focus on the case where only *image-level* labels are provided (e.g., adding the ‘sheep’ class comes as easily as only providing images guaranteed to contain at least a sheep). However, this task, denoted as Weakly Supervised Class-Incremental (WSCI) SIS, is an

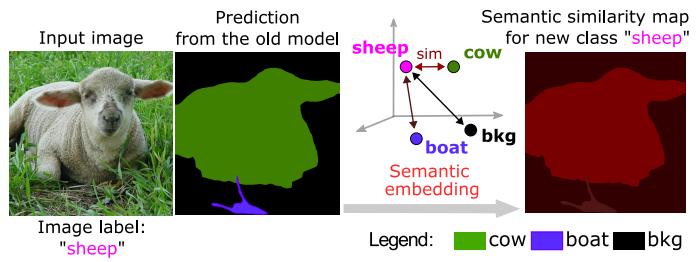


Figure 1: Our proposed Relation-aware Semantic Prior (RaSP) loss is based on the intuition that old class predictions from an existing model provide valuable cues for the segmentation of unseen, semantically related classes. Based on the semantic relatedness of the image label (e.g., sheep) and the model predictions (e.g., cow), our model derives denser maps for the new class and leverages them during training

extremely challenging problem in itself and very few attempts have been made to tackle it in the context of continual learning (Cermelli et al., 2022).

We argue that weakly supervised SIS, despite being a harder problem due to the lack of dense supervision, can efficiently be addressed in an *incremental* scenario, by exploiting the model’s own prior about the objects that have already been learned in the past. This stems from our observation that a SIS model trained to segment, *e.g.*, the class ‘cow’ often misclassifies the pixels of an unseen class ‘sheep’ and assign those to the ‘cow’ class that it knows (see Fig. 1, left). This is caused by the visual similarity between the *bovines* (*e.g.*, cow) and *ovines* (*e.g.*, sheep), both being furry four-legged species. In this work, we take advantage of this behavior by equipping the model with the ability to take into account the *semantic relationship* between the old and new classes, while localizing new objects – as humans do. In other words, the localization cues offered by the model on semantically similar objects (*e.g.*, cow and sheep) is used to approximately convert the coarse image-level supervision of the new class into dense pixel-level supervision (thanks to the semantic similarity maps, as shown in Fig. 1, right). Learning a new class then simply becomes optimizing a pixel-level supervised objective that is erstwhile not available in weakly supervised SIS. Given the similarity maps are derived by leveraging the semantic relationship between the class label names, we term the proposed objective as **Relation-aware Semantic Prior** (RaSP) loss.

The RaSP loss has been designed with the goal of improving *forward transfer* in incremental learning scenarios by converting weaker image-level supervision into pixel-level. It can be seen as a general-purpose plug-and-play module, suitable for any weakly supervised class-incremental SIS framework, as all it needs is a segmentation network that outputs pixel-level predictions and the class label *names* of the previously seen classes. In our experiments we show that RaSP—when integrated with the state-of-the-art method WILSON (Cermelli et al., 2022)—leads to performance improvements, sometimes by large margins, and especially in longer incremental scenarios.

To summarize, our contributions are threefold: **(i)** We propose the RaSP loss to facilitate class-incremental SIS when only image-level labels are available as supervision. It treats class labels as semantic entities and exploits what the model knows about previous classes it has been trained on, to learn new ones at each increment; **(ii)** We broaden the benchmarks previously used for weakly supervised class-incremental SIS and consider longer sequences of tasks (prior work is limited to 2, we extend to up to 11 tasks) and few-shot incremental settings, in both cases with image-level annotations only; **(iii)** We empirically validate that the steady improvement brought by RaSP is also visible in an extended version of our approach that uses an episodic memory, filled with either past samples or web-crawled images for the old classes. We show that, in this context, the memory does not only mitigate catastrophic forgetting, but also and most importantly fosters the learning of new categories.

2 RELATED WORK

This work lies at the intersection of weakly supervised and class-incremental learning of SIS models. Due to the nature of our semantic prior loss, it also relates to text-guided computer vision.

Weakly supervised SIS.

To circumvent the need for expensive pixel-level annotations when learning SIS models, weakly supervised SIS (Borenstein & Ullman, 2004) approaches training SIS models using cheaper and lesser constrained forms of annotations such as image captions (Xu et al., 2022a), bounding boxes (Dai et al., 2015; Ji & Veksler, 2021; Song et al., 2019), scribbles (Lin et al., 2016; Tang et al., 2018), points (Bearman et al., 2016; Qian et al., 2019) and image labels (Kolesnikov & Lampert, 2016; Ahn & Kwak, 2018; Araslanov & Roth, 2020; Xu et al., 2022b). Out of these, learning to segment with only image labels is the most attractive alternative, as the annotation cost is arguably the lowest. Our work falls under the family of methods using image-level supervision, but jointly uses the ground truth image-level labels and the predictions of the old model to provide denser supervision to the new classes. Opposed to several of the previous works, our RaSP loss is simple by design and can be integrated with any SIS model.

Class-incremental SIS. Under the hood of continual learning (Parisi et al., 2019), class-incremental learning consists in exposing a model to sequences of tasks, in which the goal is learning new classes without having access to data from the previous classes. While most class-incremental learning methods have focused on image classification (see Masana et al. (2023) for a survey), some recent works have started focusing on SIS (Cermelli et al., 2020; Michieli & Zanuttigh, 2021a; Douillard et al., 2021; Maracani et al., 2021; Cha et al., 2021). Yet, all aforementioned methods assume pixel-level annotations for all the new classes, which requires a huge, often prohibitively expensive amount of manual work. Therefore, weakly-supervised class-incremental SIS has emerged as a viable alternative in the pioneering work of Cermelli et al. (2022), which formalizes the WSCI task, and proposes the WILSON method to tackle it. In details, the WILSON framework builds on top of standard weakly supervised SIS techniques (Araslanov & Roth, 2020), and explicitly tries to mitigate forgetting using knowledge distillation, akin to the pseudo-labeling approach of

PLOP (Douillard et al., 2021). Orthogonal to the components introduced in WILSON that mostly deal with fortifying backward transfer, our RaSP improves the forward transfer aspect of the WSCI task by providing better supervision to the weakly supervised localizer for segmenting new classes.

Language-guided computer vision. Vision and language have a long history of benefiting from each other, and language, a modality that is inherently more semantic, has often been used as a source of supervision to guide computer vision tasks, such as learning visual representations (Quattoni et al., 2007; Gomez et al., 2017; Sariyildiz et al., 2020; Radford et al., 2021) object detection (Shi et al., 2017), zero-shot segmentation (Zhou et al., 2016; Bucher et al., 2019; Xian et al., 2019; Li et al., 2020; Baek et al., 2021), language-driven segmentation (Zhao et al., 2017; Li et al., 2022; Ghiasi et al., 2022; Xu et al., 2022a) or referring image segmentation (Hu et al., 2016; Liu et al., 2017; Ding et al., 2021; Wang et al., 2022), among others. One of the core ingredients behind the success of language and vision models is the ability to embed the natural language (*e.g.*, captions, class names, etc.) into semantically meaningful spaces using Word2Vec (Mikolov et al., 2013), GloVe (Pennington et al., 2014) or BERT (Devlin et al., 2019)—to name a few. Similarly, our RaSP loss assumes the availability of such similarity metrics to be used between textual pairs consisting of the name of the predicted class label and the name of the ground truth image label, where the strength of the similarity map is determined by the semantic closeness. Contrary to the open-vocabulary segmentation methods that use large-scale datasets with descriptive image captions (Xu et al., 2022a), at times alongside ground truth pixel-level annotations (Wang et al., 2022) or class-agnostic segmentation annotations (Ghiassi et al., 2022), our RaSP just requires class label names and predictions of the model itself.

3 METHODS

We develop a method for the task of Weakly Supervised Class-Incremental SIS (WSCI), where the goal is incrementally learning to segment objects from new classes by using image-level labels only, and thus avoiding the need for pixel-level annotations. Before detailing our method, we formalize our setting.

Problem setup and notations. Following the WSCI setting established by Cermelli et al. (2022) to evaluate WILSON, we likewise assume access to pixel-level annotations for an initial set of categories, followed by incrementally learning on a sequence of new classes using image-level labels only. This can be regarded as well-aligned with practical scenarios for which dense annotations are available for entry-level *primitive* classes, whereas the less frequently occurring objects or specialized variants of the generic classes incrementally come with image labels only.

Let $\mathcal{D}^b = \{(\mathbf{x}_k^b, \mathbf{y}_k^b)\}_{k=1}^{N^b}$ be a dataset for SIS, where $\mathbf{x}^b \in \mathbb{R}^{H \times W \times 3}$ represents an input image and \mathbf{y}^b is a tensor containing the $|\mathcal{C}^b|$ -dimensional one-hot label vectors for each pixel, in a $H \times W$ spatial grid, corresponding to a set of \mathcal{C}^b semantic classes. As typical in SIS, objects that do not belong to any of the foreground classes are annotated as a special background class (‘*bkg*’)—included in \mathcal{C}^b . We refer to \mathcal{D}^b as the *base* task and do not make assumptions on its cardinality. \mathcal{D}^b is used to train a base model, generally defined by an encoder E^b and a decoder F^b , $(E^b \circ F^b): \mathbf{x} \rightarrow \mathbb{R}^{|\mathcal{I}| \times |\mathcal{C}^b|}$, where $|\mathcal{I}| = H' \times W'$ is a spatial grid—corresponding to the input image size or some resized version of it—and $\mathbf{p} = (E^b \circ F^b)(\mathbf{x})$ is the set of class prediction maps, where p_i^c is the probability for the spatial location $i \in \mathcal{I}$ in the input image \mathbf{x} to belong to the class c .

After this base training, we assume the model undergoes a sequence of learning steps, as training sets for new tasks become available. Specifically, at each learning step t , the model is exposed to a new set $\mathcal{D}^t = \{(\mathbf{x}_k^t, \mathbf{l}_k^t)\}_{k=1}^{N^t}$ containing N^t instances labeled for previously unseen \mathcal{C}^t classes, where $\mathbf{l}^t \in \mathbb{R}^{|\mathcal{C}^t|}$ is the vectorized *image-level* label corresponding to an image \mathbf{x}^t . Note that in each incremental step, only weak annotations (image-level labels) are provided for the new classes. This is in sharp contrast with the base task, in which the model is trained with pixel-level annotations.

The goal of WSCI is to update the segmentation model at each incremental step t in a weakly supervised way, without *forgetting* any of the previously learned classes. We learn the function $(E^t \circ F^t): \mathbf{x} \rightarrow \mathbb{R}^{|\mathcal{I}| \times |\mathcal{Y}^t|}$, where $\mathcal{Y}^t = \bigcup_{k=1}^t \{\mathcal{C}^k\} \cup \mathcal{C}^b$ is the set of labels at step t (old and new ones). Note that, in general, we assume that data from previous tasks cannot be stored—that is, there is no episodic memory. We relax this assumption for some of our experiments: see Sec. 4.2 for results related to the setting that includes an episodic memory.

3.1 THE RELATION-AWARE SEMANTIC PRIOR LOSS

In this paper, we propose to leverage the semantic relationship between the new and old classes to improve the segmentation results when only image labels are available. We argue that semantic object categories are not independent, *i.e.*, the new classes \mathcal{C}^t that are being learned at step t may bear semantic resemblance with the old classes from \mathcal{Y}^{t-1} , seen

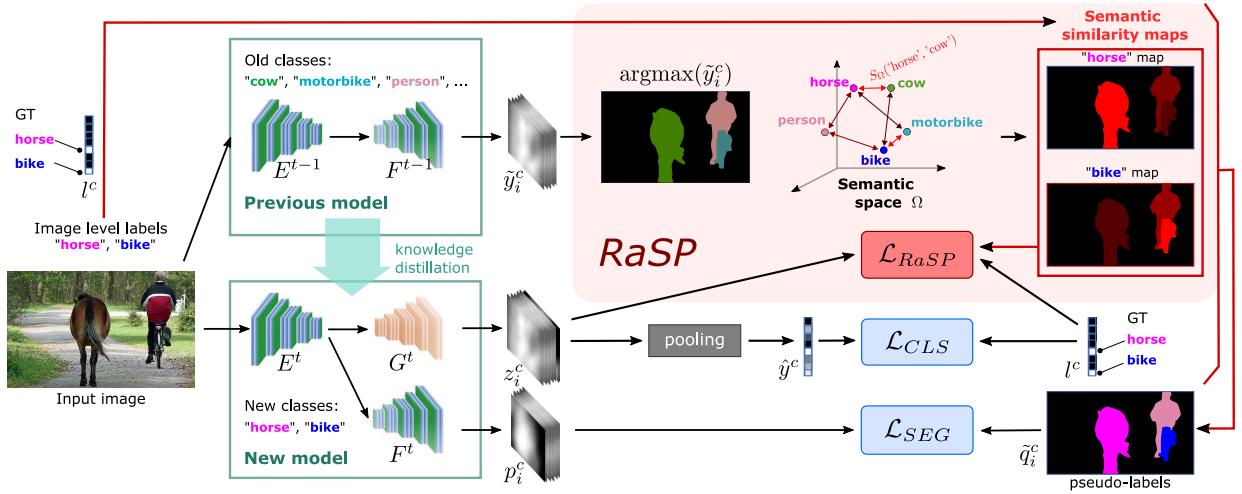


Figure 2: Overview of the **RaSP** loss integrated in a generic WSCI framework. Given the snapshot of a segmentation model $(E \circ F)^{t-1}$, trained to segment ‘cow’, ‘motorbike’, and ‘person’, the training step t is tasked with learning new classes ‘horse’ and ‘bike’ using image-level labels only. The RaSP loss uses the old model predictions \tilde{y}_i^c and image labels l^c to generate semantic maps, where the intensity of the semantic map at pixel location i is proportional to the semantic distance between the embeddings of the two class names. These dense semantic maps are then used as pseudo supervision to train the localizer $(E \circ G)^t$. \mathcal{L}_{RaSP} can be seamlessly combined with any generic weakly supervised \mathcal{L}_{CLS} and not-forgetting losses

by the model during previous training steps. Going back to our initial example, the network may have been trained to segment instances of the ‘cow’ class with dense supervision during the base training, and at any arbitrary incremental step t the segmentation network can be tasked with learning to segment the ‘sheep’ class from weak-supervision. Since cow and sheep are closely related species sharing similar attributes (such as being four-legged, furry mammals), the old snapshot of the model $E^{t-1} \circ F^{t-1}$ (or, for brevity, $(E \circ F)^{t-1}$) can provide valuable cues to localize the ‘sheep’ regions in an image labeled as sheep, despite having never seen this animal before (see Fig. 1). Guided by this insight, instead of using the old model predictions to solely obtain cues about the old classes (if present), as done *e.g.*, in WILSON, we propose a semantically-guided prior that uses the old model predictions to discover more precise object boundaries for the new classes, in the form of semantic similarity maps. Note that the class-incremental SIS methods are often based on the popular background-shift (Cermelli et al., 2020) assumption that unseen objects are *always* classified as background by an old model. Our prior loss challenges this assumption, and is based on our observation that the old model tends to misclassify foreground regions from unseen objects as closely related old classes. We believe that both phenomena are prevalent in incremental learning, and we root our method on the latter. We qualitatively validate our motivation through extensive visualizations in Fig. 3.

Concretely, at step t and using the old model $(E \circ F)^{t-1}$, for each pixel x_i^t we assign the most probable class label $y_i^* = \arg \max_{c \in \mathcal{Y}^{t-1}} \tilde{y}_i^c$ from old classes, yielding the label map y^* . Note that our method expects y_i^* to be a class label *name* instead of a class *index* (*e.g.*, say ‘cow’ instead of an index 5 for the class cow). Then, given the set of ground truth image-level label names $\mathcal{L}(x^t) = \{c | l_c^t = 1\}$ associated with image x^t , we estimate a similarity map s^c between each class l^c in $\mathcal{L}(x^t)$ and the predicted label map y^* :

$$s^c = \{\mathbf{S}_\Omega(\omega(y_i^*), \omega(l^c))\}_{i \in \mathcal{I}}, \quad (1)$$

where $\omega(c)$ is a vectorial embedding of the semantic class c in a semantic embedding space Ω and \mathbf{S}_Ω is a semantic similarity measure defined between the classes in Ω (see Appendix A.1 for details). Different semantic embeddings can be considered, such as Word2Vec (Mikolov et al., 2013), GloVe (Pennington et al., 2014) or BERT (Devlin et al., 2019). These language models were trained such that the dot product between a pair of embedding vectors, \mathbf{S}_Ω , reflects the semantic similarity between their corresponding text. For example in Fig. 1, $\mathbf{S}_\Omega(\omega(\text{'sheep'}), \omega(\text{'cow'})) \gg \mathbf{S}_\Omega(\omega(\text{'sheep'}), \omega(\text{'bkg'}))$, as ‘sheep’ lies closer to ‘cow’ in the semantic space than the ‘background’ class. Intuitively, stronger the similarity between the predicted class label y_i^* at pixel location i and the ground truth image label l^c , higher the likelihood of the pixel i belonging to the new class l^c . In this work, we use BERT (Devlin et al., 2019) for all the experiments (see comparisons with other embeddings in the Tab. A1).

Note that the similarity maps in Eq. (1) are computed exhaustively for every pixel location in a given image with respect to all the previous classes in \mathcal{Y}^{t-1} , which also includes the *bkg* class. As the background can not reliably

provide objectness cues for new object classes, we ensure not to alter the original predictions made on the background class by normalizing the similarity map such that the score for the ‘*bkg*’ class is equal to 1:

$$s_i^c = \frac{\exp(S_\Omega(\omega(y_i^*), \omega(l^c))/\tau)}{\exp(S_\Omega(\omega('bkg'), \omega(l^c))/\tau)}, \quad (2)$$

where τ is a scaling hyperparameter. By exploiting the similarity maps we convert the image labels l^c into pixel-level label maps s^c , one per new class c (see Fig. 1 and Fig. 2). We provide more details in the Appendix (Sec. A.3).

The generality of our proposed RaSP loss is evident from the fact that the similarity maps s^c are derived using the same segmentation model from the previous step $(E \circ F)^{t-1}$ and the image-level labels of the image x^t from the current step. While in many cases the dense similarity maps offered by RaSP might be sufficient (when all the new classes in \mathcal{Y}^t have strong resemblance to the old classes \mathcal{Y}^{t-1}), they can fall short in situations where completely unrelated classes appear in \mathcal{Y}^t or the new class region is predicted as background.

To enable learning in all possible scenarios, including the *edge-cases* where *all* the new classes are dissimilar to the previous ones, we couple our method with approaches from the weakly supervised SIS literature. The most common weakly supervised SIS approach is to exploit the classification activation maps (CAM) (Zhou et al., 2016) from a standard classifier (dubbed as *localizer* head G^t), trained for predicting image-level class labels, to obtain the most discriminative regions as pixel-level pseudo-labels (Kolesnikov & Lampert, 2016; Araslanov & Roth, 2020). In our formulation, where we learn with weaker image labels, we bootstrap the training of the aforementioned localizer (Araslanov & Roth, 2020) using the proposed semantic similarity maps in the form of the following binary cross-entropy (BCE) loss:

$$\mathcal{L}_{\text{RaSP}}(\mathbf{z}, \mathbf{s}) = -\frac{1}{|\mathcal{C}^t||\mathcal{I}|} \sum_{i \in \mathcal{I}} \sum_{c \in \mathcal{C}^t} \sigma(s_i^c) \log(\sigma(z_i^c)) + (1 - \sigma(s_i^c)) \log(1 - \sigma(z_i^c)), \quad (3)$$

where z_i^c is the logit corresponding to the class c assigned by the localizer at location i , $\sigma(\cdot)$ is the sigmoid function. Given a generic loss for a localizer \mathcal{L}_{CLS} (an instance of this loss will be detailed in the next section), we can combine the two terms as $\mathcal{L} = \mathcal{L}_{\text{CLS}} + \lambda \mathcal{L}_{\text{RaSP}}$. Intuitively, our proposed loss serves as a regularizer that encourages forward transfer from the old classes to the new ones.

3.2 FULL INTEGRATION OF RASP

Without loss of generality, we implement our RaSP loss on top of the WILSON framework (Cermelli et al., 2022). We chose WILSON since it is the state of the art and, since it relies on the localizer module introduced by Araslanov & Roth (2020) to tackle WSCI, represents a good fit to test our loss.

Background. WILSON is an end-to-end method for WSCI that incrementally learns to segment new classes with the supervision of pseudo-labels generated by the localizer trained with image-level supervision. More specifically, at step t , WILSON is composed of a shared encoder E^t , a main segmentation head F^t —which is incrementally extended to accommodate new classes—and a localizer head G^t , trained from scratch for every task. It also stores a copy of the model from the previous task, $(E \circ F)^{t-1}$.

Given an image x from the current task, $\tilde{\mathbf{y}} = \sigma((F \circ E)^{t-1}(\mathbf{x})) \in \mathbb{R}^{|\mathcal{I}| \times |\mathcal{Y}^{t-1}|}$ is the output produced by the old model. The scores obtained by the localizer, $\mathbf{z} = (G \circ E)^t(\mathbf{x}) \in \mathbb{R}^{|\mathcal{I}| \times |\mathcal{Y}^t|}$, are aggregated into a one-dimensional vector $\hat{\mathbf{y}} \in \mathbb{R}^{|\mathcal{Y}^t|}$ by using normalized Global Weighted Pooling (see Appendix C.2 for details). The score \hat{y}_c , for each class c , can be seen as the likelihood for image x to contain semantic class c . This allows training the model with image-level labels using the multi-label soft-margin loss:

$$\mathcal{L}_{\text{CLS}}(\hat{\mathbf{y}}, \mathbf{l}) = -\frac{1}{|\mathcal{C}^t|} \sum_{c \in \mathcal{C}^t} l^c \log(\sigma(\hat{y}_c)) + \sum_{c \in \mathcal{C}^t} (1 - l^c) \log(1 - \sigma(\hat{y}_c)). \quad (4)$$

Note that, although the localizer outputs a $|\mathcal{Y}^t|$ -dimensional vector, at task t we are only provided with images and their image-level annotations for the new classes. Therefore, the sum in Eq. (4) is computed only over the new classes. In order to train the localizer for the old classes and prevent the encoder from shifting towards the new classes and forgetting the old ones, WILSON distills knowledge from the old model, by adding two knowledge distillation losses—at intermediate feature and output space. The first one, \mathcal{L}_{KDE} , computes the mean-squared error between the features extracted by the current encoder E^t and those extracted by the previous one E^{t-1} . The second distillation loss \mathcal{L}_{KDL} encourages consistency between the pixel-wise scores for old classes predicted by the localizer $(E \circ G)^t$ and those predicted by the old model $(E \circ F)^{t-1}$ (see details in Appendix C.1).

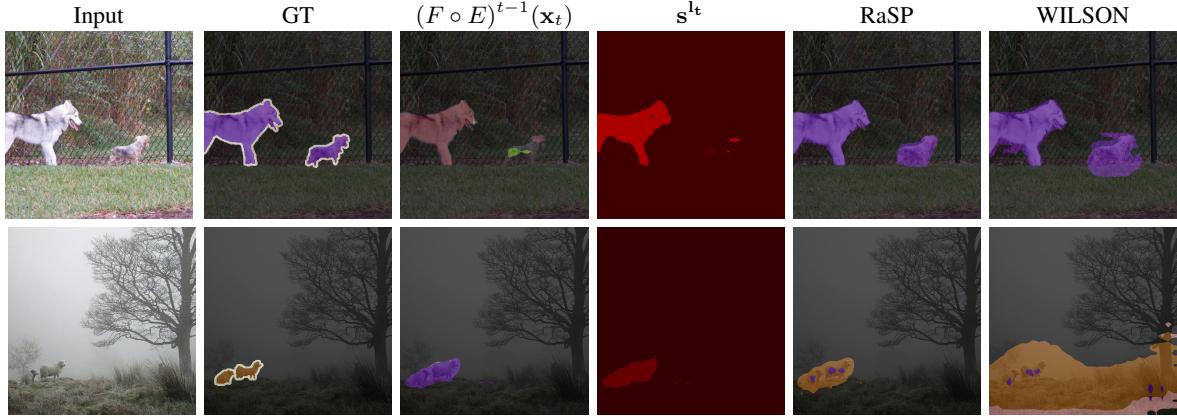


Figure 3: **Visualizations.** Qualitative figures from the *multi-step overlap* incremental protocol on 10-2 VOC. From left to right: input image, GT segmentation overlayed, predicted segmentation from old model, semantic similarity map corresponding to the image label (dog / sheep) computed between this label and old classes, predicted segmentation obtained with RaSP and with WILSON. Semantic similarity maps displayed in OpenCV colormap HOT (low high similarity)

Finally, WILSON combines the localizer output with the old model to generate the pseudo-supervision scores \tilde{q}^c that are used to update the main segmentation module $(E \circ F)^t$, following

$$\mathcal{L}_{\text{SEG}}(\hat{\mathbf{p}}, \tilde{\mathbf{q}}) = -\frac{1}{|\mathcal{Y}^t||\mathcal{I}|} \sum_{i \in \mathcal{I}} \sum_{c \in \mathcal{Y}^t} \tilde{q}_i^c \log(\sigma(\hat{p}_i^c)) + (1 - \tilde{q}_i^c) \log(1 - \sigma(\hat{p}_i^c)), \quad (5)$$

where $\hat{\mathbf{p}} = (E \circ F)^t(\mathbf{x})$ are the predictions from the main segmentation head and $\tilde{\mathbf{q}}$ is the supervisory signal containing: i) the old model’s predictions for the old classes, ii) the localizer’s refined scores for the new classes and iii) the minimum between the old model and the localizer scores for the background. The final objective optimized by WILSON is the non-weighted sum of the different loss terms defined above, $\mathcal{L}_W = \mathcal{L}_{\text{CLS}} + \mathcal{L}_{\text{KDL}} + \mathcal{L}_{\text{KDE}} + \mathcal{L}_{\text{SEG}}$. See Appendix C for more details.

Extending WILSON with RaSP. Since WILSON exploits a localizer-based approach designed for weakly supervised SIS, it constitutes a good starting point to integrate and test our proposed semantic prior—without the need for any ad hoc architectural changes. Therefore, we complement WILSON’s losses with our loss introduced in Eq (3), which simply requires as input: (i) the output from the localizer $\mathbf{z} = (E \circ G)^t(\mathbf{x})$, and (ii) the semantic similarity maps between new and old classes, obtained via Eqs. (1) and (2). Endowed with these, our prior loss can be applied together with WILSON losses by simply optimizing the joint loss $\mathcal{L}_J = \mathcal{L}_W + \lambda \mathcal{L}_{\text{RaSP}}$. The hyperparameter λ controls the strength of our prior loss, which acts as a regularizer fostering forward transfer from the old to the new classes.

4 EXPERIMENTS

Datasets. Following Cermelli et al. (2022), we run experiments on two standard weakly supervised SIS benchmarks: Pascal VOC (Everingham et al., 2010) and MS-COCO (Lin et al., 2014). Note that being the WSCI task more challenging than both supervised SIS (Xie et al., 2021) and class-incremental SIS (Cermelli et al., 2020), state-of-the-art weakly supervised methods (Xu et al., 2022b) do not experiment on ADE20K (Zhou et al., 2017) or Cityscapes (Cordts et al., 2016), which contain ‘stuff’ classes, but restrict themselves to VOC and MS-COCO, focusing on ‘thing’ classes. Similarly, we follow suit with the relevant literature. The VOC benchmark consists of 10,582 training and 1,449 validation images covering 20 semantic categories. MS-COCO is much larger scale and contains 164k training and 5k validation images from 80 ‘thing’ categories. We adopt the same train and testing splits as used in WILSON.

Incremental settings. We provide empirical evaluation under several incremental learning scenarios, which differ in their splitting between the base and new classes. We name the settings following the notation N_b - N_t to indicate that we first learn with *pixel-level supervision* from N_b base classes, and then learn sets of N_t new classes at a time, with *image-level supervision* only. Given a total number of N classes, the number of tasks is $(N - N_b)/N_t + 1$. All the new classes can either be added in a single step, the only scenario explored so far by the WSCI literature, or can be added in multiple learning steps, a more challenging yet more realistic scenario.

As the name suggests, the *Single-step* settings comprise only one incremental learning phase. For instance, in the **15-5 VOC** setting (see Tab. 1), we first train the model on 15 base classes from VOC and then learn the remaining 5 (new) classes in a single incremental step (bringing the total number of classes to 20). The newly introduced *Multi-step* settings add new classes to the model in multiple sequential steps. The **10-2 VOC** setting, for instance, considers 10 base classes and 5 incremental steps which each learn 2 new classes at a time. In each table, we indicate results for base classes as $1-N_b$ and for the new ones as $(N_b + 1)-N$. Differently, the **COCO-to-VOC** setting involves using the 60 classes exclusive to COCO in the base training, and performing the incremental learning step(s) on the 20 classes of VOC (e.g., **60-5** is a 5-step protocol where the 20 VOC classes are learned in 4 increments).

Each incremental setting can be designed following one of these two protocols: i) *Overlap*, if all the training images for a given step contain at least one instance of a new class, but they can also contain previous or even future classes; ii) *Disjoint*, if each step consists of images containing only new or previously seen classes, but never future classes. In both protocols, image-level annotations are available for the *new classes only*. We argue that the multi-step setting with the overlap protocol is the most realistic one. That said, we also consider the original settings of WILSON, since it facilitates fair comparison with the previous work.

Implementation details. Following WILSON (Cermelli et al., 2022), we use DeeplabV3 (Chen et al., 2017) with ResNet101 (He et al., 2016) and Wide-ResNet-38 (Wu et al., 2019) as backbone for the VOC and MS-COCO datasets, respectively. The localizer is composed of 3 convolutional layers, interleaved with BatchNorm (Ioffe, 2021) and Leaky ReLU layers. For each step, we train the model with SGD for 40 epochs using a batch size of 24. Since the localizer can produce noisy outputs early in training, we do not use \mathcal{L}_{SEG} for the first 5 epochs. We set $\tau = 5$ and $\lambda = 1$ and follow the values suggested in WILSON for all other hyperparameters. See Appendix B.1 for sensitivity to τ and λ .

Evaluation metrics. We evaluate all models using the standard mean Intersection over Union (mIoU) (Everingham et al., 2010) metric. We report the mIoU scores evaluated after the last incremental step. We report 3 values: for the base task (considering results on the base classes excluding the background), for the subsequent ones (new classes added during the incremental steps) and finally considering all the classes including the background (All).

4.1 MAIN RESULTS

Comparison with the state of the art. We compare our proposed RaSP with several state-of-the-art class-incremental learning methods that use either pixel-level or image-level annotations in the incremental steps. We mainly focus on WSCI methods, *i.e.*, WILSON, since it is the current state-of-the-art method and it allows fair comparisons (for instance, methods like EPS (Lee et al., 2021) use saliency maps as extra supervision). Pixel-supervised methods are interesting but not comparable as they use a prodigious amount of extra-supervision. The best performing method with image-level and pixel-level supervision are respectively bolded and underlined in tables. Since Cermelli et al. (2022) tested WILSON only for single-step incremental settings, we ran experiments in the other settings using the official implementation provided by the authors (<https://github.com/fcdl194/WILSON>). For comparability, we also re-ran experiments on single-task settings. “WILSON \dagger ” indicates our reproduced results while “WILSON” corresponds to the original numbers from the paper. We further report in tables the relative gain/drop in performance (in %) of our RaSP w.r.t. WILSON \dagger , within brackets.

Results. We summarize the results of the VOC and COCO-to-VOC experiments in Fig. 4. In particular we report RaSP’s relative percentage gain (in %) over WILSON and observe that, as the number of incremental steps increases, the overall gain of RaSP over WILSON for the new classes becomes more and more noteworthy. Next we elaborate the results for each benchmark.

In Tab. 1, we show results for both single-step and multi-step incremental settings—on VOC, using the *overlap* protocol. We observe that our RaSP outperforms WILSON in almost all the considered settings. In particular, the relative gain (in %) w.r.t. WILSON grows wider as the number of incremental steps increases, with RaSP achieving +26.8% relative improvement over WILSON in new class performance, in the 10-2 setting. Not only our semantic-prior loss improves new class performance but also it leads to 15% lesser forgetting w.r.t. WILSON.

We provide qualitative examples in Fig. 3 and in the Fig. D, showing how the semantic maps aid the final segmentation.

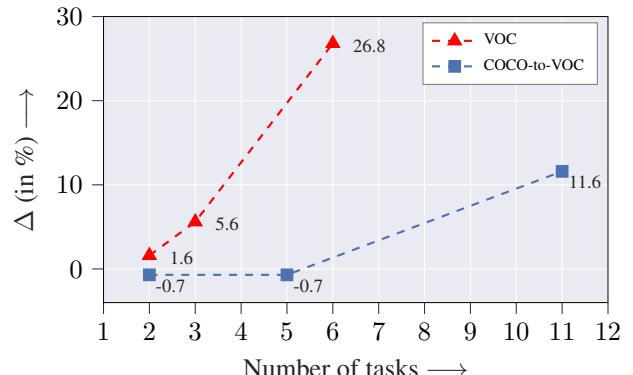


Figure 4: Our RaSP’s relative percentage gain (Δ in %) over WILSON on new class performance for the VOC and COCO-to-VOC tasks

Method	Supervision	15-5 (2 tasks)			10-10 (2 tasks)			
		1-15	16-20	All	1-10	11-20	All	
Fine-Tuning	Pixel	12.5	36.9	18.3	7.8	58.9	32.1	
LWF (<i>Li & Hoiem, 2016</i>)	Pixel	67.0	41.8	61.0	70.7	63.4	67.2	
PLOP (<i>Douillard et al., 2021</i>)	Pixel	75.7	51.7	70.1	69.6	62.2	67.1	
SDR (<i>Michieli & Zanuttigh, 2021b</i>)	Pixel	75.4	52.6	69.9	70.5	63.9	67.4	
RECALL (<i>Maracani et al., 2021</i>)	Pixel	67.7	54.3	65.6	66.0	58.8	63.7	
single-step	CAM (<i>Zhou et al., 2016</i>)	Image	69.9	25.6	59.7	70.8	44.2	58.5
	SEAM (<i>Wang et al., 2020</i>)	Image	68.3	31.8	60.4	67.5	55.4	62.7
	SS (<i>Araslanov & Roth, 2020</i>)	Image	72.2	27.5	62.1	69.6	32.8	52.5
	EPS (<i>Lee et al., 2021</i>)	Image	69.4	34.5	62.1	69.0	57.0	64.3
	WILSON (<i>Cermelli et al., 2022</i>)	Image	74.2	41.7	67.2	70.4	57.1	65.0
	WILSON† (<i>Cermelli et al., 2022</i>)	Image	76.3	44.1	69.3	71.4	56.1	64.9
multi-step	RaSP (Ours)	Image	76.2	47.0	70.0	72.3	57.2	65.9
			(↓0.1%)	(↑6.6%)	(↑1.0%)	(↑1.3%)	(↑1.6%)	(↑1.5%)
single-step								
multi-step								
single-step								
multi-step								
single-step								
multi-step								
single-step								
multi-step								
single-step								
multi-step								
single-step								
multi-step								
single-step								
multi-step								
single-step								
multi-step								
single-step								
multi-step								
single-step								
multi-step								
single-step								
multi-step								
single-step								
multi-step								
single-step								
multi-step								
single-step								
multi-step								
single-step								
multi-step								
single-step								
multi-step								
single-step								
multi-step								
single-step								

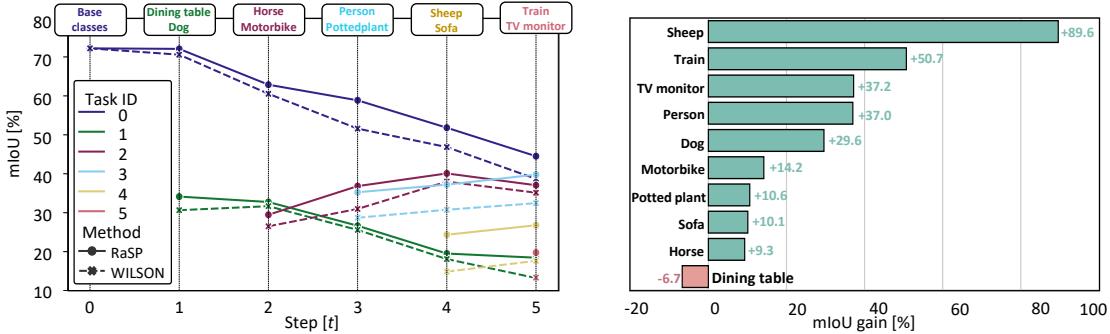


Figure 5: **Left:** Per-task and per-step mIoU for the 10-2 VOC *multi-step overlap* incremental setting. **Right:** Per class gain/drop of RaSP w.r.t. WILSON, evaluated for each class in the step it was learned. We observe that RaSP encounters lesser forgetting and improved learning of the new classes

scenarios, we observe not only improvements for the new classes, but also more limited forgetting of the old ones. This highlights that our RaSP loss is most effective when the sequence of tasks get longer, *i.e.*, in more plausible settings.

4.2 LEARNING ONE NEW CLASS AT A TIME

A major limitation of the state-of-the-art approaches for WSCI is that their performance degrades when learning one new class at a time: the model fails to learn the new class and undergoes drastic forgetting. This is due to the fact that Eq. (4) is optimized for a single positive class: the lack of negative classes leads to gross overestimation of the foreground, causing the localizer to provide poor pseudo labels to the segmentation head, with a negative effect on old classes as well. We show in Tab. 3 (top-half) results of WILSON and RaSP for two single-class incremental settings (15-1 and 10-1), using VOC. Both methods struggle with learning new classes, yielding poor performance compared to pixel-supervised methods. These fully-supervised methods can learn the new classes better since their annotations are composed of both positive foreground-object pixels and negative background pixels.

A solution: episodic memory. To circumvent this issue we store a small number of images from the previous classes in a fixed-size memory \mathcal{M} . Intuitively, the samples in the memory help the localizer by providing negative samples (via pseudo labels on the old objects). We show in Tab. 3 (bottom-half) that storing as little as 100 past samples from the previous classes dramatically improves the learning on new classes for both WILSON and RaSP, with RaSP + \mathcal{M} outperforming WILSON + \mathcal{M} (28.3% vs 20.8% in the 15-1 setting). Unsurprisingly, it also helps retaining performance on the base classes. Similar observations hold for the 10-1 setting.

External data as an alternative. Inspired by the class-incremental SIS method RECALL (Maracani et al., 2021), we consider the option of retrieving samples of the previous classes from external sources. We define this memory as \mathcal{M}_{ext} . Concretely, we retrieve 100 samples per old class from ImageNet (by creating a mapping with the VOC classes). As shown in Tab. 3, this further improves both WILSON + \mathcal{M}_{ext} and RaSP + \mathcal{M}_{ext} compared to the previous episodic memory \mathcal{M} . RECALL performs better on new classes, but i) relies on pixel-level supervision and ii) uses significantly more web-crawled images—therefore, it is not directly comparable.

4.3 CLASS-INCREMENTAL FEW-SHOT SEGMENTATION

We compare RaSP with WILSON on the task of Incremental Few-Shot Segmentation (iFSS) (Cermelli et al., 2021), where the model learns incrementally from only few images per new class (*e.g.*, 2 or 5 images). This is a challenging setting, only tested so far with pixel-level supervision for the new classes. Here, we add the challenging constraints that the training images for the new classes are only weakly annotated, *i.e.*, with image-level labels. Following Cermelli et al. (2021), we consider 4 folds of 5 new classes for PASCAL-5ⁱ and the 4 folds of 20 new classes for COCO-20ⁱ, where each fold is used, in turn, as incremental setting with the other classes defining the base task.

Tab. 4 reports the results averaged over the 4 folds (per-fold results are reported in the Appendix B.6). The bottom rows of Tab. 4 reports the results obtained by the WSCI methods WILSON and RaSP. As expected, in the case of COCO-20ⁱ both methods perform poorly when compared to the pixel-supervised methods, which is not surprising as even the strongly supervised methods (top rows) have difficulties to learn the new classes. On the other hand, in PASCAL-5ⁱ, not only RaSP consistently outperforms WILSON, but in the 5-shot case it also outperforms or performs on par with some of the strongly supervised methods. Finally, we can observe that the performance of RaSP on the

Method	Supervision	15-1 (6 tasks)			10-1 (11 tasks)			
		1-15	16-20	All	1-10	11-20	All	
w/o memory	ILT (Michieli & Zanuttigh, 2019)	Pixel	4.9	7.8	5.7	16.5	1.0	9.1
	MiB (Cermelli et al., 2020)	Pixel	35.1	13.5	29.7	15.1	14.8	15.0
w/ memory	WILSON† (Cermelli et al., 2022)	Image	0.0	2.3	0.6	0.0	0.2	0.1
	RaSP (Ours)	Image	17.7	0.9	13.2	2.0	0.7	1.3
w/ memory	WILSON† + \mathcal{M}	Image	61.5	20.8	52.5	33.4	24.6	30.0
	RaSP (Ours) + \mathcal{M}	Image	63.3	28.3	56.0	38.9	30.9	36.9
w/ memory	WILSON† + \mathcal{M}_{ext}	Image	75.7	32.9	65.9	66.8	34.9	52.3
	RaSP (Ours) + \mathcal{M}_{ext}	Image	75.7	35.2	66.6	66.8	39.1	54.4
RECALL (Web) (Maracani et al., 2021)	Pixel	<u>67.8</u>	<u>50.9</u>	<u>64.8</u>	<u>65.0</u>	<u>53.7</u>	<u>60.7</u>	

Table 3: **Effect of memory.** Results on single-class *multi-step overlap* incremental setting on VOC. \mathcal{M} and \mathcal{M}_{ext} indicate memories of previously seen or external samples, respectively

Method	Supervision	VOC (5-shot)			VOC (2-shot)			COCO (5-shot)			COCO (2-shot)		
		1-15	16-20	HM	1-15	16-20	HM	0-60	61-80	HM	0-60	61-80	HM
Fine-Tuning	Pixel	55.8	29.6	38.7	59.1	19.7	29.5	41.6	12.3	19.0	41.5	7.3	12.4
WI (Qi et al., 2018)	Pixel	63.3	21.7	32.3	63.3	19.2	29.5	43.6	8.7	14.6	44.2	7.9	13.5
AMP Siam et al. ((2019))	Pixel	51.9	18.9	27.7	54.4	18.8	27.9	34.6	11.0	16.7	35.7	8.8	14.2
MiB (Cermelli et al., 2020)	Pixel	<u>65.0</u>	28.1	39.3	63.5	12.7	21.1	44.7	11.9	18.8	44.4	6.0	10.6
PIFS (Cermelli et al., 2021)	Pixel	60.0	33.3	<u>42.8</u>	60.5	<u>26.4</u>	<u>36.8</u>	42.8	<u>15.7</u>	<u>23.0</u>	40.9	<u>11.1</u>	<u>17.5</u>
WILSON† (Cermelli et al., 2022)	Image	64.1	20.5	31.1	63.3	10.2	17.6	45.0	5.8	10.3	43.6	1.9	3.6
RaSP (Ours)	Image	64.4	21.3	32.0	63.5	10.7	18.3	45.1	5.6	10.0	43.5	2.0	3.8

Table 4: **Few-shot results.** The mIoU (in %) scores for the *single-step* (2 tasks) incremental few-shot SiS settings on the PASCAL-5ⁱ and COCO-20ⁱ benchmarks, for 5-shot and 2-shot cases. We show the average results over the 4 folds as in ([Cermelli et al., 2021](#)). For each experiment, columns report performance on old classes, new classes, and the Harmonic-Mean (HM) of the two scores. For RaSP (Ours), we also report the relative gain/drop in performance (in %) w.r.t. WILSON†

base classes remains comparable and is sometimes higher than most of the strongly supervised methods, where the higher performance on the new classes tends to come with a more severe forgetting.

Limitations. In edge cases, where the new classes are very dissimilar to the old classes, our proposed RaSP loss will not bring tangible improvements. However, in practical applications one can reasonably assume that a model has already learned an array of primitive classes (often leveraging stronger pixel-level supervision), and that the incremental learner will encounter new objects that have some degree of resemblance to those primitive classes. We posit that such limitation will become less and less relevant as the model learns a large number of new classes, since the likelihood of finding an old class similar to each new one at hand increases more and more over the model lifetime.

5 CONCLUSIONS

We proposed a method for Weakly Supervised Class-Incremental Semantic Image Segmentation, where the model is tasked with incrementally learning to segment new objects using weaker image labels as supervision. Guided by the observation that new classes to be learned by the model often bear resemblance with the old ones that it already knows, we designed a Relation-aware Semantic Prior (RaSP) loss that fosters forward transfer. It transfers objectness prior from the past model by leveraging the semantic similarity between old and new class names and aids the model in learning new categories. We validated our proposed method on a wide variety of incremental scenarios derived from standard benchmarks. In particular, we demonstrated that our method is resilient in unexplored and challenging scenarios, where the number of tasks is high (or number of classes in each task is low) and reduced data availability for each task.

REFERENCES

- Jiwoon Ahn and Suha Kwak. Learning Pixel-level Semantic Affinity with Image-level Supervision for Weakly Supervised Semantic Segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.
- Nikita Araslanov and Stefan Roth. Single-stage semantic segmentation from image labels. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.
- Donghyeon Baek, Youngmin Oh, and Bumsuk Ham. Exploiting a Joint Embedding Space for Generalized Zero-Shot Semantic Segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021.
- Amy Bearman, Olga Russakovsky, Vittorio Ferrari, and Li Fei-Fei. What's the point: Semantic segmentation with point supervision. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2016.
- Eran Borenstein and Shimon Ullman. Learning to Segment. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2004.
- Maxime Bucher, Tuan-Hung Vu, Matthieu Cord, and Patrick Pérez. Zero-Shot Semantic Segmentation. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.
- Fabio Cermelli, Massimiliano Mancini, Samuel Rota Bulò, 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)*, 2020.
- Fabio Cermelli, Massimiliano Mancini, Yongqin Xian, Zeynep Akata, and Barbara Caputo. Prototype-based Incremental Few-Shot Semantic Segmentation. In *British Machine Vision Conference (BMVC)*, 2021.
- Fabio Cermelli, Dario Fontanel, Antonio Tavera, Marco Ciccone, and Barbara Caputo. Incremental Learning in Semantic Segmentation from Image Labels. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.
- Sungmin Cha, Beomyoung Kim, YoungJoon Yoo, and Taesup Moon. SSUL: Semantic Segmentation with Unknown Label for Exemplar-based Class-Incremental Learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.
- Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017.
- Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.
- Gabriela Csurka, Riccardo Volpi, Boris Chidlovskii, et al. Semantic image segmentation: Two decades of research. *Foundations and Trends® in Computer Graphics and Vision*, 2022.
- Jifeng Dai, Kaiming He, and Jian Sun. BoxSup: Exploiting Bounding Boxes to Supervise Convolutional Networks for Semantic Segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2015.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *NAACL HLT*, 2019.
- Hengui Ding, Chang Liu, Suchen Wang, and Xudong Jiang. Vision-Language Transformer and Query Generation for Referring Segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.
- Arthur Douillard, Yifu Chen, Arnaud Dapogny, and Matthieu Cord. PLOP: Learning without Forgetting for Continual Semantic Segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.
- Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88(2):303–338, 2010.

- Golnaz Ghiasi, Xiuye Gu, Yin Cui, and Tsung-Yi Lin. Scaling Open-Vocabulary Image Segmentation with Image-Level Labels. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2022.
- Spyros Gidaris and Nikos Komodakis. Dynamic Few-Shot Visual Learning without Forgetting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.
- Lluis Gomez, Yash Patel, Marçal Rusinol, Dimosthenis Karatzas, and CV Jawahar. Self-supervised learning of visual features through embedding images into text topic spaces. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.
- Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.
- Ronghang Hu, Marcus Rohrbach, and Trevor Darrell. Segmentation from Natural Language Expressions. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2016.
- Sergey Ioffe. Batch Renormalization: Towards Reducing Minibatch Dependence in Batch-Normalized Models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.
- Zongliang Ji and Olga Veksler. Weakly Supervised Semantic Segmentation: From Box to Tag and Back. In *British Machine Vision Conference (BMVC)*, 2021.
- Alexander Kolesnikov and Christoph H. Lampert. Seed, Expand and Constrain: Three Principles for Weakly-Supervised Image Segmentation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2016.
- 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.
- Boyi Li, Kilian Q Weinberger, Serge Belongie, Vladlen Koltun, and Rene Ranftl. Language-driven semantic segmentation. In *International Conference on Learning Representations (ICLR)*, 2022.
- Peike Li, Yunchao Wei, and Yi Yang. Consistent Structural Relation Learning for Zero-Shot Segmentation. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.
- Zhizhong Li and Derek Hoiem. Learning without Forgetting. *Proceedings of the European Conference on Computer Vision (ECCV)*, 2016.
- Di Lin, Jifeng Dai, Jiaya Jia, Kaiming He, and Jian Sun. Scribblesup: Scribble-supervised convolutional networks for semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.
- 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 *Proceedings of the European Conference on Computer Vision (ECCV)*, 2014.
- Chenxi Liu, Zhe Lin, Xiaohui Shen, Jimei Yang, Xin Lu, and Alan Yuille. Recurrent Multimodal Interaction for Referring Image Segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2017.
- Andrea Maracani, Umberto Michieli, Marco Toldo, and Pietro Zanuttigh. RECALL: Replay-Based Continual Learning in Semantic Segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021.
- Marc Masana, Xiaolei Liu, Bartłomiej Twardowski, Mikel Menta, Andrew D Bagdanov, and Joost van de Weijer. Class-incremental learning: Survey and performance evaluation on image classification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023.
- Umberto Michieli and Pietro Zanuttigh. Incremental Learning Techniques for Semantic Segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2019.
- Umberto Michieli and Pietro Zanuttigh. Knowledge Distillation for Incremental Learning in Semantic Segmentation. *Computer Vision an Image Understanding*, 2021a.

- 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)*, 2021b.
- Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv:1301.3781*, 2013.
- German I. Parisi, Ronald Kemker, Jose L. Part, Christopher Kanan, and Stefan Wermter. Continual Lifelong Learning with Neural Networks: A Review. *Neural Networks*, 113:54–71, 2019.
- Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *In EMNLP*, 2014.
- Hang Qi, Matthew Brown, and David G. Lowe. Low-shot Learning with Imprinted Weights. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.
- 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*, 2019.
- Ariadna Quattoni, Michael Collins, and Trevor Darrell. Learning visual representations using images with captions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2007.
- Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, 2021.
- Mert Bulent Sariyildiz, Julien Perez, and Diane Larlus. Learning visual representations with caption annotations. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020.
- Miaojing Shi, Holger Caesar, and Vittorio Ferrari. Weakly supervised object localization using things and stuff transfer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2017.
- Mennatullah Siam, Boris Oreshkin, and Martin Jagersand. AMP: Adaptive Masked Proxies for Few-Shot Segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019.
- Chunfeng Song, Yan Huang, Wanli Ouyang, and Liang Wang. Box-driven Class-wise Region Masking and Filling Rate Guided Loss for Weakly Supervised Semantic Segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.
- Meng Tang, Abdelaziz Djelouah, Federico Perazzi, Yuri Boykov, and Christopher Schroers. Normalized cut loss for weakly-supervised cnn segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.
- Yonglong Tian, Yue Wang, Dilip Krishnan, Joshua B. Tenenbaum, and Phillip Isola. Rethinking Few-shot Image Classification: A Good Embedding is All You Need? In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020.
- Yude Wang, Jie Zhang, Meina Kan, Shiguang Shan, and Xilin Chen. Self-supervised equivariant attention mechanism for weakly supervised semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.
- Zhaqing Wang, Yu Lu, Qiang Li, Xunqiang Tao, Yandong Guo, Mingming Gong, and Tongliang Liu. CRIS: CLIP-Driven Referring Image Segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.
- Zifeng Wu, Chunhua Shen, and Anton Van Den Hengel. Wider or deeper: Revisiting the resnet model for visual recognition. *Pattern Recognition*, 90:119–133, 2019.
- Yongqin Xian, Subhabrata Choudhury, Yang He, Bernt Schiele, and Zeynep Akata. Semantic Projection Network for Zero- and Few-Label Semantic Segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.
- Enze Xie, Wenhui Wang, Zhiding Yu, Anima Anandkumar, Jose M Alvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

Jiarui Xu, Shalini De Mello, Sifei Liu, Wonmin Byeon, Thomas Breuel, Jan Kautz, and Xiaolong Wang. GroupViT: Semantic Segmentation Emerges from Text Supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022a.

Lian Xu, Wanli Ouyang, Mohammed Bennamoun, Farid Boussaid, and Dan Xu. Multi-class token transformer for weakly supervised semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022b.

Hang Zhao, Xavier Puig, Bolei Zhou, Sanja Fidler, and Antonio Torralba. Open Vocabulary Scene Parsing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2017.

Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ade20k dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

APPENDIX

The Appendix is organized as follows: Sec. A provides implementation details of RaSP. Sec. B includes additional experimental results on ablation study, different class orderings, classwise performance, class-incremental few-shot segmentation. Sec. C lists the details about the WILSON framework. Sec. D provides additional qualitative results. Finally, in Sec. E we conclude with a discussion.

A IMPLEMENTATION DETAILS OF RASP

A.1 SEMANTIC SIMILARITY METRIC

The similarity metric \mathbf{S}_Ω used in the Eq. (2) of the main paper is derived from the cosine distance, which is computed between a pair of class label names as:

$$\mathbf{S}_\Omega = -(1 - \frac{\omega(c_i) \cdot \omega(c_j)}{\|\omega(c_i)\|_2 \|\omega(c_j)\|_2}). \quad (\text{A1})$$

where $\omega(c_i)$ and $\omega(c_j)$ represent the vectorial embeddings for the i^{th} and j^{th} classes. The value of \mathbf{S}_Ω is then substituted to the Eq. (2) of the main paper. Note that higher the semantic similarity between a pair of labels c_i and c_j , higher is the s_i^c value.

We obtain the vectorial embedding $\omega(c)$ corresponding to a class label name l_c using the BERT transformer (Devlin et al., 2019). In details, we prompt the transformer with the class label name to obtain a 768-dimensional vector representation $\omega(c) = \text{Transformer}(\text{"An image of a } \{l_c\}\text{"})$. While one could omit the prompt and simply provide the class label name, we do it to give context to the transformer that the class label name is a noun. Please note that our method can work with other semantic mapping functions, e.g., Word2Vec (see Tab. A1).

A.2 SELECTIVE BACKPROPAGATION OF RASP LOSS

To recap, we compute the semantic similarity maps (described in Eq. (2) of the main paper) only for the new foreground classes \mathcal{C}^t present in an incremental step t . In other words, the semantic map s^{bkg} for the *bkg* class is not computed, and not enforced by the optimization in Eq. (3). Moreover, we selectively backpropagate the RaSP loss $\mathcal{L}_{\text{RaSP}}$ only for those new class channels of the localizer G_t for which ground truth *image labels* are available. As an example, in an incremental step t if there are five new classes, $|\mathcal{C}^t| = 5$, and if for a given image only the new class “dog” is present, then we simply backpropagate the gradients of the RaSP loss for the “dog” channel only. All the other channels, including the *bkg* channel, are ignored during the backpropagation. Given the fact that the old model does not perfectly predict the new classes as *bkg* and is spuriously activated as foreground for the new classes (see the $(F \circ E)^{t-1}(\mathbf{x}_t)$ column in Fig. A5 where new class objects are not *bkg*), the RaSP loss in practice does not largely suppress the CAM loss. We hope that our new findings will encourage future WSCI works to tackle overconfident model predictions on unseen classes.

A.3 ROLE OF NORMALIZATION IN SEMANTIC SIMILARITY

Here we expand on the role of normalization introduced in Eq. (2). Our proposed RaSP first computes the semantic similarity between a given new class c and all the old classes (including the “*bkg*” class) in the image following Eq. 1. It means that for the pixel locations predicted as “*bkg*” by the old model $(E \circ F)^{t-1}$, each new class can have a non-zero semantic similarity with different scale, which can be detrimental for learning the new class. To this end, we enforce that the semantic similarity s_i^c is always lower-bounded by 1 (*a.k.a* normalization) for the pixels corresponding to the “*bkg*” class for every new class c (using Eq. (2)). Then, as shown in the RaSP loss of Eq. (3), we apply a squashing function (sigmoid) such that the network activations corresponding to such pixels are suppressed, and only foreground pixels with high semantic similarities are encouraged by the network.

B ADDITIONAL EXPERIMENTS

B.1 HYPERPARAMETER SENSITIVITY ANALYSIS

In Fig. A1 we show how results are affected when we vary the hyperparameters τ and λ in the case of 10-2 VOC *multi-step overlap* (solid) and **disjoint** (dashed) incremental settings, reporting both performance on old and new classes (in

blue and green, respectively). To recap, τ plays a role in computing the similarity maps via Eq. (2); in particular, it is a scaling factor that controls how steep the decay is, as two semantic entities are more or less similar. Instead, γ controls the strength of our RaSP loss: the larger the γ , the higher the impact of our prior over the other terms.

For our experiments, we have selected $\tau = 5$ and $\gamma = 1$, the former by observing that it provides a sufficiently steep decay, the latter following WILSON’s approach of not assigning different weights to the different terms, since they operate at similar scales. We can observe in Fig. A1 that i) RaSP is satisfactorily robust against the choice of these hyperparameters and ii) better results than the ones proposed in the main paper can be obtained. Notice that using $\lambda = 0$ nullifies the effect of RaSP, making the method equivalent to WILSON; for comparison, WILSON’s mIoU performance for the **disjoint** setting is 36.4 and 20.8 (see Tab. A2, bottom-right) and for the **overlap** setting is 38.7 and 22.4 (see Tab. 1, bottom-right) for old and new classes, respectively. In both cases, significantly below performance of RaSP, regardless of the hyperparameters selected.

Please note that annotated test sets of VOC and COCO are not available. Thus, all the performances are reported on the validation set. Because of this reason, it is hard to tune the hyperparameters without overfitting the validation set. For this reason, we did not spend computational resources into hyperparameter validation and based our decisions on the aforementioned heuristics.

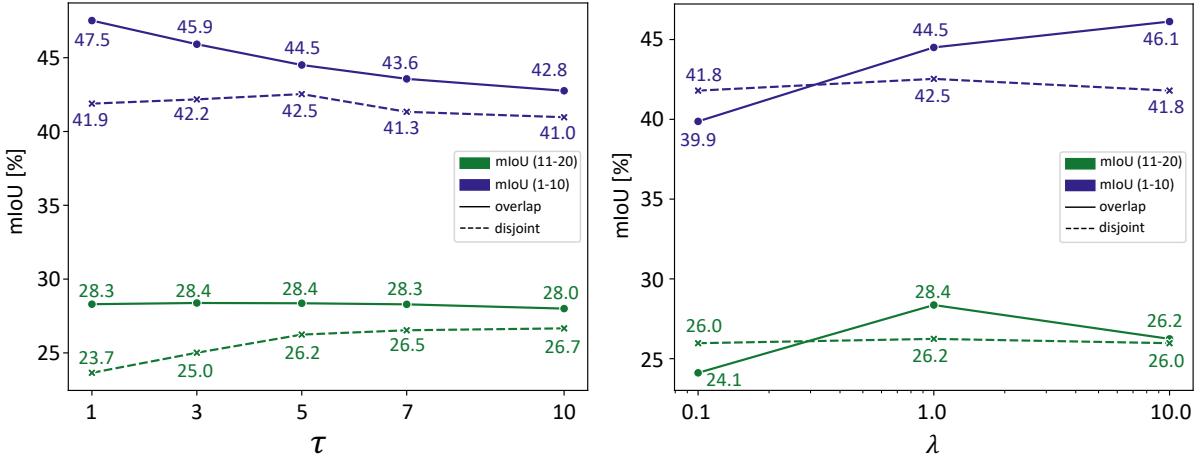


Figure A1: **Ablating τ (left) and γ (right).** RaSP results on VOC, using the 10-2 setting (6 tasks). Solid and dashed lines indicate overlap and disjoint results, respectively. Blue and green lines indicate performance on old and new classes, respectively

Semantic Similarity	10-2 VOC		
	1-10	11-20	All
WordNet	47.6	27.9	39.7
GloVe	43.1	26.8	37.2
BERT	44.5	28.4	38.6
WILSON†	38.7	22.4	32.5

Table A1: **Ablating semantic similarity** on VOC 10-2 *multi-step overlap* incremental setting

Next, we compare different semantic embedding methods for building the similarity between the semantic classes defined in Eq. (3). While by default we used BERT (Devlin et al., 2019) in our experiments, we can also consider other alternatives such as GloVe (Pennington et al., 2014) or a WordNet sub-tree. In the latter case, to compute the similarities between two class, we used 1 over the number of hops (edges) between the two classes (nodes) in the sub-tree. As we can see, using a different semantic embedding yields to relatively similar performance, with a slight drop when we use GLOVE, and significant gain on old classes when we use the WordNet sub-tree. Still, all three methods outperform WILSON: this result further validates the idea that leveraging semantic similarity between old and new classes can improve the localizer and, hence, the final model.

B.2 RESULTS ON VOC USING THE DISJOINT PROTOCOL

We report in Tab. A2 the VOC results in the **disjoint** settings. This table complements the analysis of the Tab. 1, which focused on the **overlap** setting. We can draw similar conclusions: RaSP’s improvements over WILSON \dagger increase as we increase the number of tasks.

Method	Supervision	15-5 (2 tasks)			10-10 (2 tasks)				
		1-15	16-20	All	1-10	11-20	All		
Joint	Pixel	75.5	73.5	75.4	76.6	74.0	75.4		
FT	Pixel	8.4	33.5	14.4	7.7	60.8	33.0		
LWF Li & Hoiem (2016)	Pixel	39.7	33.3	38.2	63.1	61.1	62.2		
ILT Michieli & Zanuttigh (2019)	Pixel	31.5	25.1	30.0	67.7	61.3	64.7		
PLOP Douillard et al. (2021)	Pixel	71.0	42.8	64.3	63.7	60.2	63.4		
SDR Michieli & Zanuttigh (2021b)	Pixel	73.5	47.3	67.2	67.5	57.9	62.9		
RECALL Maracani et al. (2021)	Pixel	69.2	52.9	66.3	64.1	56.9	61.9		
single-step	CAM Zhou et al. (2016)	Image	67.5	25.5	57.8	64.8	41.2	54.2	
	SEAM Wang et al. (2020)	Image	68.9	32.5	61.1	61.5	52.3	58.3	
	SS Araslanov & Roth (2020)	Image	68.9	25.9	60.2	60.3	27.2	45.5	
	EPS Lee et al. (2021)	Image	70.7	36.8	63.6	64.3	53.8	60.5	
	WILSON Cermelli et al. (2022)	Image	72.0	44.1	66.3	64.2	54.5	60.8	
	WILSON \dagger Cermelli et al. (2022)	Image	75.8	45.2	69.3	63.7	51.1	59.0	
RaSP (Ours)	Image	75.9	47.5	69.9	64.5	51.2	59.4		
	Image	(\uparrow 0.1%)	(\uparrow 5.1%)	(\uparrow 0.9%)	(\uparrow 1.3%)	(\uparrow 0.2%)	(\uparrow 0.7%)		
multi-step				10-5 (3 tasks)	10-2 (6 tasks)				
				1-10	11-20	All	1-10	11-20	All
	WILSON \dagger Cermelli et al. (2022)	Image	58.6	45.3	53.6	36.4	20.8	30.6	
	RaSP (Ours)	Image	60.5	46.8	55.3	42.5	26.2	36.6	
				(\uparrow 3.2%)	(\uparrow 3.3%)	(\uparrow 3.2%)	(\uparrow 16.8%)	(\uparrow 26.0%)	(\uparrow 19.6%)

Table A2: The m-IoU (in %) scores for both *single-step* (top half) and *multi-step* (bottom half) **disjoint** incremental settings on the VOC. The best numbers for the pixel supervised and image supervised methods are highlighted in underline and bold, respectively

Furthermore, we report in Tab. A3 the VOC results for the memory-based approaches detailed in Sec. 4.2, for the **disjoint** setting, to complement the analysis we provided in Tab. 3, which focused on the **overlap** setting.

Method	Supervision	15-1 (6 tasks)			10-1 (11 tasks)			
		1-15	16-20	All	1-10	11-20	All	
w/o memory	ILT (Michieli & Zanuttigh, 2019)	Pixel	6.7	1.2	5.4	14.1	0.6	7.5
	MiB (Cermelli et al., 2020)	Pixel	46.2	12.9	37.9	14.9	9.5	12.3
	WILSON \dagger (Cermelli et al., 2022)	Image	0.0	1.4	0.4	0.0	0.2	0.1
	RaSP (Ours)	Image	16.2	1.8	12.4	1.3	1.0	1.1
w/ memory	WILSON \dagger + \mathcal{M}	Image	64.9	24.8	56.0	43.4	21.7	34.1
	RaSP (Ours) + \mathcal{M}	Image	66.7	30.9	59.0	42.7	28.8	37.9
	WILSON \dagger + \mathcal{M}_{ext}	Image	74.2	30.3	64.3	62.0	33.9	49.5
	RaSP (Ours) + \mathcal{M}_{ext}	Image	74.7	35.8	66.1	61.7	37.4	51.2
	RECALL (Web) (Maracani et al., 2021)	Pixel	<u>67.6</u>	<u>49.2</u>	<u>64.3</u>	<u>62.3</u>	<u>50.0</u>	<u>57.8</u>

Table A3: **Effect of memory.** Results on single-class *multi-step disjoint* incremental setting on VOC. \mathcal{M} and \mathcal{M}_{ext} indicate memories of previously seen or external samples, respectively. The best numbers for the pixel supervised and image supervised methods are highlighted in underline and bold, respectively

B.3 RASP PERFORMANCE OVER TASKS

The Fig. A2 extends the plot shown in Fig. 5 (right). We report RaSP’s gains w.r.t. WILSON for different VOC settings. As expected, since RaSP outperforms WILSON more when the number of tasks is larger, the per-class gains are more evident for the 10-2 setting (top) than for the 10-5 one (bottom).

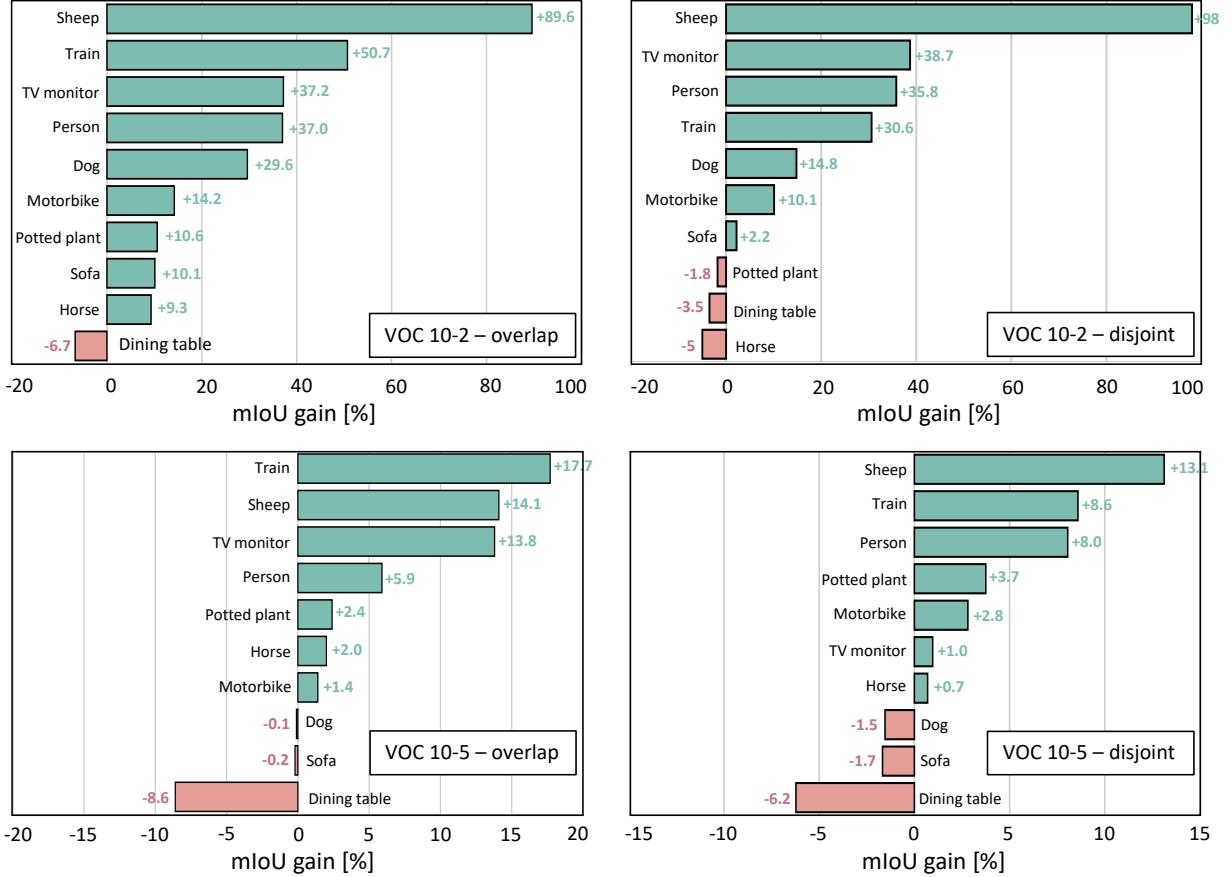


Figure A2: Per class gain/drop of RaSP w.r.t. WILSON, evaluated for each class in the step it was learned. Results computed on VOC. Top plots show **10-2** settings; bottom plots show **10-5** settings; leftmost plots show **overlap** settings; rightmost plots show **disjoint** settings. Note the different scales

The Fig. A3 extends the plot shown in Fig. 5 (left). We report the evolution of the performance across sequence of tasks in the 10-2 VOC setting (6 tasks), for overlap and disjoint protocols (left and right, respectively). For these plots, the conclusions made in the main paper still hold.

B.4 IMPACT OF CLASS ORDERING

To demonstrate that our proposed semantic prior loss $\mathcal{L}_{\text{RaSP}}$ is versatile under different class ordering, we chose the 15-5 VOC disjoint setting, having 15 base classes and 5 novel classes, and randomized the old-novel classes splits. We ran experiments on four of such random splits and report the results in the Tab. A4. From the Tab. A4 it is evident that RaSP outperforms WILSON on the four randomly chosen base-novel classes split, denoted by 15-5a, 15-5b, 15-5c and 15-5d of VOC, indicating that our improvements are consistently better on all of the class orderings. While the improvement by RaSP varies among the base-novel splits, yet most importantly they do not drop below WILSON. Thus we believe that our proposed method is well suited for real world applications where the classes will appear in random (and unknown) order and yet our incremental learner can perform better than its competitors.

B.5 CLASSWISE PERFORMANCE

To get a complete understanding about the performance of each class, we report the classwise mIoU scores in a couple of settings of VOC for RaSP and compare it with WILSON. In details, we report the step-wise performance of both WILSON and our proposed RaSP for the single-step 15-5 VOC and the multi-step 10-2 VOC overlap settings in the Tab. A5 and Tab. A6, respectively. In the Tab. A5 and Tab. A6 the incremental step (*i.e.*, learning with weak labels)

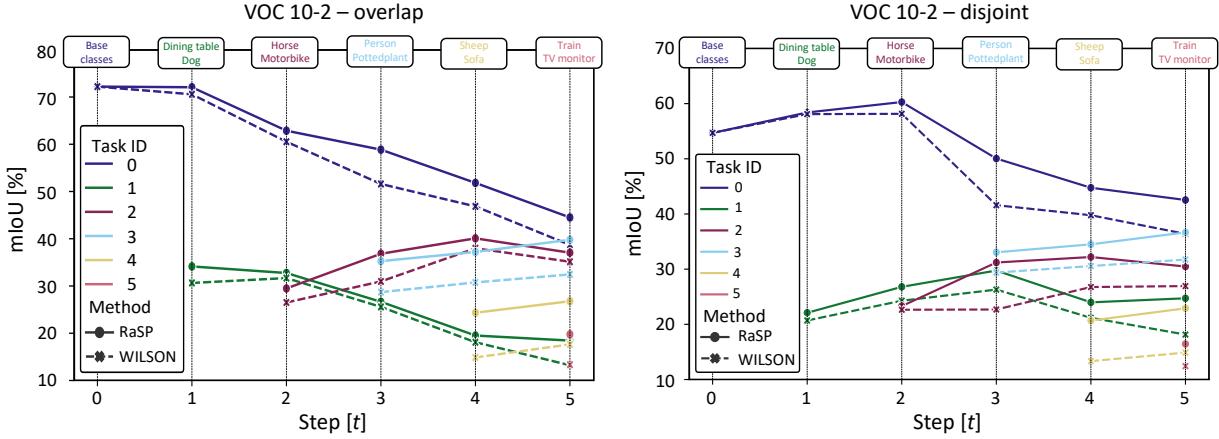


Figure A3: Per-task and per-step mIoU for the 10-2 VOC *multi-step* incremental setting. Leftmost plot shows **overlap** results; rightmost plot shows **disjoint** results. Note the different scales

Method	15-5 (2 tasks)															
	15-5a			15-5b			15-5c			15-5d			Mean			
	1-15	16-20	All	1-15	16-20	All	1-15	16-20	All	1-15	16-20	All	1-15	16-20	All	1-15
WILSON†	75.8	45.2	69.3	71.2	48.5	66.7	68.7	42.7	63.6	66.5	56.2	65.3	70.6	48.2	66.2	
RaSP	75.9	47.5	69.9	71.8	53.3	68.4	70.8	44.5	65.5	66.7	57.8	65.9	71.3	50.8	67.4	

Table A4: Comparison with the state-of-the-art on the 15-5 VOC disjoint incremental setting under different class orderings. The m-IoU (in %) scores have been reported for the methods

starts from step 1, with step 0 being the base training. The summarized versions of the Tab. A5 and Tab. A6 have been reported in the Tab. 1 of the main paper.

Method	Step	Old Classes															New Classes				Aggregate				
		bkg	aeroplane	bicycle	bird	boat	bottle	bus	car	cat	chair	diningtable	dog	horse	motorbike	person	pottedplant	sheep	sofa	train	tv-monitor	1-15	16-20	All	
WILSON†	1	90.3	89.3	42.6	87.0	68.2	79.3	89.0	89.0	92.6	42.0	70.7	58.9	87.9	81.9	80.4	86.3	25.6	52.0	38.4	59.7	44.7	76.3	44.1	69.3
RaSP	1	91.4	89.8	42.6	87.5	65.8	79.3	89.5	89.1	92.0	41.3	70.7	58.7	87.7	81.8	81.7	86.3	26.5	54.6	36.8	70.5	46.5	76.2	47.0	70.0

Table A5: **Classwise results.** The mIoU (in %) scores for the *single-step* 15-5 (2 tasks) **overlap** incremental setting on VOC. The 15 old classes are denoted by green and the 5 new classes are denoted in red. The best numbers are highlighted in bold

From the Tab. A5 we observe that our RaSP improves forward transfer by outperforming WILSON in four out of the five new classes. In-line with our intuition, RaSP’s gain over WILSON is noticeable in the new class “train” (by +10.8 absolute points) since “train” can be considered to have high visual similarity with the old class “bus”. The gain in the other new classes (such as “pottedplant” or “tv-monitor”) is slightly subdued due to the lack of closely resembling old classes. Nevertheless, in terms of new classes (**16-20**) and **All** aggregate performance RaSP outperforms WILSON.

For the multi-step 10-2 VOC setting, reported in the Tab. A6, the improvement of RaSP over WILSON is even more stark compared to the single-step 15-5 VOC setting. In details, RaSP outperforms in 20 out of the 21 classes in the Pascal-VOC benchmark, achieving greatly improved results in both the old (**1-10**) and the new classes (**11-20**). Careful scrutiny of the Tab. A6 reveals that the forward transfer offered by our RaSP has a significant positive impact on the new classes such as “dog”, “horse”, “sheep” and “train”, improving by +10.4, +4.1, +17.5 and +10.9 absolute points, respectively. Interestingly, the old classes suffer from lesser forgetting w.r.t WILSON, with an aggregate improvement of +5.8 absolute points at the end of the final incremental step. We found that in incremental tasks where there are very few new classes (e.g., 2 new classes in the 10-2 VOC) WILSON tends to overestimate the foreground (see Fig. 3 of the main and Fig. A5), thereby forgetting more on the older classes. Contrarily, our RaSP due to the semantic guidance for

Method	Step	Old Classes														New Classes						Aggregate						
		bkg	aeroplane	bicycle	bird	boat	bottle	bus	car	cat	chair	cow	diningtable	dog	horse	motorbike	person	pottedplant	sheep	sofa	train	tv-monitor	1-10	11-20	All			
WILSON†	1	90.6	86.5	41.3	81.4	67.4	82.8	87.8	81.7	85.3	35.1	56.4	30.7	30.6	32.3	20.7	26.6	30.8	20.1	9.7	21.4	5.3	70.6	60.5	51.6			
	2	89.1	84.0	31.8	76.0	66.2	75.5	85.7	56.5	71.5	32.7	25.6	28.4	34.9	30.6	30.6	31.3	29.6	32.0	38.0	21.7	13.6	46.8	42.4	32.5			
	3	79.4	61.0	30.2	68.7	48.1	72.9	52.6	54.5	71.2	30.6	26.2	16.2	35.0	30.6	36.7	39.3	26.7	30.8	32.7	21.4	5.3	38.7	22.4	32.5			
	4	74.6	49.6	27.6	56.9	57.5	62.8	65.2	57.4	59.2	6.2	26.0	0.0	36.2	30.0	36.7	39.3	29.6	32.0	20.1	9.7	21.4	5.3	70.6	60.5	51.6		
	5	72.6	37.9	25.8	59.5	48.9	58.7	48.0	30.1	57.8	5.1	15.1	0.0	26.4	35.5	34.8	32.2	30.8	32.7	21.7	13.6	21.4	5.3	38.7	22.4	32.5		
RaSP	1	92.2	86.3	40.7	83.1	69.5	83.3	88.6	82.1	87.0	35.0	65.2	28.7	39.6	44.8	24.9	40.6	35.3	23.7	38.0	10.7	32.3	7.3	72.1	62.9	58.9		
	2	91.3	83.9	34.2	77.2	68.3	77.8	86.0	58.0	68.0	30.5	44.8	24.9	40.6	35.3	27.9	38.2	38.4	35.3	36.4	34.1	40.4	39.2	36.7	37.7	38.0	10.7	51.8
	3	86.5	64.5	32.5	73.6	59.6	74.4	79.0	62.9	69.1	27.9	45.3	15.1	38.2	35.3	36.4	34.1	36.4	34.1	36.7	37.7	39.2	14.4	32.3	7.3	44.5	28.4	38.6
	4	84.5	52.4	29.3	66.2	62.5	62.2	80.7	60.8	53.7	7.2	43.3	0.0	39.0	41.0	39.2	36.7	37.7	38.0	10.7	32.3	7.3	44.5	28.4	38.6			
	5	82.7	44.2	27.4	67.1	53.2	58.8	65.3	34.5	57.9	6.5	30.1	0.1	36.8	39.6	34.5	40.4	39.2	39.2	14.4	32.3	7.3	44.5	28.4	38.6			

Table A6: **Classwise results.** The mIoU (in %) scores for the multi-step 15-5 (6 tasks) **overlap** incremental setting on VOC. The 10 old classes are denoted by green and the remainder new classes at consecutive steps are color coded as {diningtable, dog}; {horse, motorbike}; {person, pottedplant}; {sheep, sofa}; and {train, tv-monitor}. The best numbers at the end of the final incremental step is highlighted in bold

the foreground objects suffers less from the *recency-bias*. This makes RaSP better suited for the real-world incremental settings where the incremental learner will encounter tasks with very few new classes.

B.6 CLASS-INCREMENTAL FEW-SHOT SEGMENTATION

To push the limits of the WSCI task we also experiment on the weakly supervised few-shot class-incremental scenarios. Given the results on the few-shot settings greatly depend on the chosen few-shot image instances, we run the methods on four different folds of the PASCAL-5ⁱ and COCO-20ⁱ benchmarks. The Tab. A7 is an extended version of Tab. 4 with more pixel-supervised methods. The Tab. A8, Tab. A9, Tab. A10 and Tab. A11 show per-fold results for VOC (5-shot), VOC (2-shot), COCO (5-shot) and COCO (2-shot), respectively. In the per fold tables we only show the results for the pixel-level supervised methods that performed best in average either on the base, new or the harmonic mean (HM score) of the base and the new classes (underlined in Tab. A7). We can observe from these tables that RaSP is perfectly capable of operating in harder incremental scenarios when only few image labelled data are available for the new classes. Despite the overall lower performance of the image-label supervised methods, which is understandable, RaSP can provide better and denser supervision on top of WILSON.

Method	Supervision	VOC (5-shot)			VOC (2-shot)			COCO (5-shot)			COCO (2-shot)		
		1-15	16-20	HM	1-15	16-20	HM	0-60	61-80	HM	0-60	61-80	HM
Fine-Tuning	Pixel	55.8	29.6	38.7	59.1	19.7	29.5	41.6	12.3	19.0	41.5	7.3	12.4
WI (Qi et al., 2018)	Pixel	63.3	21.7	32.3	63.3	19.2	29.5	43.6	8.7	14.6	44.2	7.9	13.5
DWI (Gidaris & Komodakis, 2018)	Pixel	64.9	23.5	34.5	64.8	19.8	30.4	44.9	12.1	19.1	45.0	9.4	15.6
RT (Tian et al., 2020)	Pixel	60.4	27.5	37.8	60.9	21.6	31.9	46.9	13.7	21.2	46.7	8.8	14.8
AMP Siam et al. (2019)	Pixel	51.9	18.9	27.7	54.4	18.8	27.9	34.6	11.0	16.7	35.7	8.8	14.2
SPN (Xian et al., 2019)	Pixel	58.4	33.4	42.5	60.8	26.3	36.7	43.7	15.6	22.9	43.7	10.2	16.5
LWF (Li & Hoiem, 2016)	Pixel	59.7	30.9	40.8	63.6	18.9	29.2	44.6	12.9	20.1	44.3	7.1	12.3
ILT (Michieli & Zanuttigh, 2019)	Pixel	61.4	32.0	42.1	64.2	23.1	34.0	47.0	11.0	17.8	46.3	6.5	11.5
MiB (Cermelli et al., 2020)	Pixel	65.0	28.1	39.3	63.5	12.7	21.1	44.7	11.9	18.8	44.4	6.0	10.6
PIFS (Cermelli et al., 2021)	Pixel	60.0	33.3	42.8	60.5	26.4	36.8	42.8	15.7	23.0	40.9	11.1	17.5
WILSON† (Cermelli et al., 2022)	Image	64.1	20.5	31.1	63.3	10.2	17.6	45.0	5.8	10.3	43.6	1.9	3.6
RaSP	Image	<u>64.4</u>	<u>21.3</u>	<u>32.0</u>	<u>63.5</u>	<u>10.7</u>	<u>18.3</u>	<u>45.1</u>	5.6	10.0	43.5	<u>2.0</u>	<u>3.8</u>

Table A7: **Few-shot results.** The mIoU (in %) scores for the *single-step* (2 tasks) incremental few-shot SiS settings on the PASCAL-5ⁱ and COCO-20ⁱ benchmarks, for 5-shot and 2-shot cases. We show the average results over the 4 folds as in (Cermelli et al., 2021). For each experiment, columns report performance on the base classes, new classes, and the Harmonic-Mean (HM) of the two scores. The best numbers for the pixel supervised and image-label supervised methods are highlighted in underline and bold, respectively

Method	Supervision	Fold 5-0			Fold 5-1			Fold 5-2			Fold 5-3		
		1-15	16-20	HM	1-15	16-20	HM	1-15	16-20	HM	1-15	16-20	HM
FT	Pixel	58.4	22.8	32.8	52.3	42.7	47.0	50.6	29.7	37.5	62.0	23.0	33.6
SPN	Pixel	63.3	28.2	39.0	53.4	43.7	48.1	54.5	33.5	41.5	62.3	28.2	38.8
MiB	Pixel	68.0	24.8	36.4	62.1	35.2	44.9	60.6	27.1	37.4	69.1	25.4	37.2
PIFS	Pixel	64.3	26.7	37.7	53.3	41.0	46.3	57.4	33.8	42.5	65.2	31.6	42.6
WILSON†	Image	66.6	18.8	29.3	60.2	22.5	32.8	61.1	21.3	31.6	68.5	19.2	30.0
RaSP	Image	66.9	19.8	30.6	60.2	23.0	33.3	61.4	21.7	32.1	69.0	20.5	31.6
		(↑0.5%)(↑5.3%)(↑4.4%)(0.0%)(↑2.2%)(↑1.5%)	(↑0.5%)(↑1.9%)(↑1.6%)(↑0.7%)(↑6.8%)(↑5.3%)										

Table A8: **5-shot results per fold.** The m-IoU (in %) scores for the *single-step* (2 tasks) incremental few-shot (**5-shot**) SIS setting on the PASCAL-5ⁱ benchmark. HM signifies the harmonic-mean of the base (0-15) and new classes (16-20) mIoU scores. The best numbers for image-label supervised methods are highlighted in bold

Method	Supervision	Fold 5-0			Fold 5-1			Fold 5-2			Fold 5-3		
		0-15	16-20	HM	0-15	16-20	HM	0-15	16-20	HM	0-15	16-20	HM
FT	Pixel	61.7	12.6	20.9	57.5	31.0	40.3	54.8	20.2	29.5	62.5	15.0	24.2
DWI	Pixel	68.2	15.1	24.7	60.4	30.9	40.9	60.4	17.2	26.8	70.1	16.2	26.3
ILT	Pixel	68.4	16.1	26.1	58.3	33.7	42.7	61.1	25.6	36.1	68.9	17.1	27.4
PIFS	Pixel	64.0	18.9	29.1	53.9	36.6	43.6	58.2	26.5	36.4	65.9	23.6	34.7
WILSON†	Image	65.7	7.7	13.8	60.6	14.7	23.7	60.0	9.4	16.3	66.8	9.0	15.9
RaSP	Image	65.7	8.5	15.1	60.6	14.0	22.7	60.5	9.8	16.9	67.1	10.6	18.3
		(0.0%)(↑10.4%)(↑9.4%)(0.0%)(↓4.8%)(↓4.2%)(↑0.8%)(↑4.3%)(↑3.7%)(↑0.4%)(↑17.8%)(↑15.1%)											

Table A9: **2-shot results per fold.** The m-IoU (in %) scores for the *single-step* (2 tasks) incremental few-shot (**2-shot**) SIS setting on the PASCAL-5ⁱ benchmark. HM signifies the harmonic-mean of the base (0-15) and new classes (16-20) mIoU scores. The best numbers for image-label supervised methods are highlighted in bold

Method	Supervision	Fold 20-0			Fold 20-1			Fold 20-2			Fold 20-3		
		0-61	61-80	HM	0-61	61-80	HM	0-61	61-80	HM	0-61	61-80	HM
FT	Pixel	37.3	7.6	12.6	40.9	15.0	22.0	45.3	13.7	21.0	43.0	12.9	19.8
ILT	Pixel	41.9	7.1	12.2	47.0	13.9	21.5	50.4	11.2	18.3	48.6	11.8	19.0
PIFS	Pixel	40.6	10.7	16.9	41.5	17.7	24.8	45.3	16.9	24.7	43.9	17.5	25.0
WILSON†	Image	41.1	5.6	9.9	44.4	4.6	8.3	48.5	5.9	10.5	46.1	7.1	12.3
RaSP	Image	41.2	5.5	9.7	44.4	4.3	7.8	48.3	5.8	10.4	46.3	6.9	12.0
		(↑0.2%)(↓0.2%)(↓2.0%)(0.0%)(↓6.5%)(↓6.0%)(↓0.4%)(↓1.7%)(↓1.0%)(↑0.4%)(↓2.8%)(↓2.4%)											

Table A10: **5-shot results per fold.** The m-IoU (in %) scores for the *single-step* (2 tasks) incremental few-shot (**5-shot**) SIS setting on the COCO-20ⁱ benchmark. HM signifies the harmonic-mean of the base (0-60) and new classes (61-80) mIoU scores. The best numbers for image-label supervised methods are highlighted in bold

Method	Supervision	Fold 20-0			Fold 20-1			Fold 20-2			Fold 20-3		
		0-60	61-80	HM	0-60	61-80	HM	0-60	61-80	HM	0-60	61-80	HM
FT	Pixel	37.4	4.2	7.6	40.3	9.0	14.7	45.4	7.7	13.2	43.1	8.4	14.0
RT	Pixel	40.6	5.5	9.7	46.8	10.5	17.2	50.8	8.1	14.0	48.5	11.1	18.1
PIFS	Pixel	38.6	6.8	11.6	39.4	13.1	19.7	43.5	11.4	18.1	42.2	13.1	20.0
WILSON†	Image	39.8	2.6	4.9	42.9	1.4	2.7	46.8	1.6	3.1	44.7	1.9	3.6
RaSP	Image	39.7	2.8	5.2	42.5	1.4	2.7	46.8	1.7	3.3	44.9	2.1	4.0
		(↓0.3%)(↑7.7%)(↑6.1%)(↓0.9%)(0.0%)(0.0%)(↑6.3%)(↑6.5%)(↑0.5%)(↑10.5%)(↑11.1%)											

Table A11: **2-shot results per fold.** The m-IoU (in %) scores for the *single-step* (2 tasks) incremental few-shot (**2-shot**) SIS setting on the COCO-20ⁱ benchmark. HM signifies the harmonic-mean of the base (0-60) and new classes (61-80) mIoU scores. The best numbers for image-label supervised methods are highlighted in bold

C ADDITIONAL DETAILS ABOUT WILSON

C.1 KNOWLEDGE DISTILLATION LOSSES

Here we detail the two knowledge distillation losses used by WILSON and RaSP. The first one, \mathcal{L}_{KDE} , – denoted by l_{ENC} in Cermelli et al. (2022) – computes the mean-squared error between the features extracted by the current encoder E^t and those extracted by the old one E^{t-1} :

$$\mathcal{L}_{\text{KDE}}(\mathbf{x}) = \frac{1}{|\mathcal{I}|} \sum_{i \in \mathcal{I}} \|e_i^t - e_i^{t-1}\|, \quad (\text{A2})$$

where e_i^{t-1} and e_i^t are the feature vectors of the pixel i in the feature maps $E^t(\mathbf{x})$ and $E^{t-1}(\mathbf{x})$ respectively.

The second distillation loss \mathcal{L}_{KDL} – denoted by l_{LOC} in Cermelli et al. (2022) – encourages consistency between the pixel-wise scores for old classes predicted by the localizer $(E \circ G)^t$ and those predicted by the old model $(E \circ F)^{t-1}$. It is carried out via the following binary cross-entropy loss:

$$\mathcal{L}_{\text{KDL}}(\mathbf{z}, \tilde{\mathbf{y}}) = -\frac{1}{|\mathcal{Y}^{t-1}| |\mathcal{I}|} \sum_{i \in \mathcal{I}} \sum_{c \in \mathcal{Y}^{t-1}} \tilde{y}_i^c \log(\sigma(z_i^c)) + (1 - \tilde{y}_i^c) \log(1 - \sigma(z_i^c)). \quad (\text{A3})$$

C.2 AGGREGATING PIXEL-LEVEL SCORES

In order to train the localizer with image-level labels, normalized Global Weighted Pooling (nGWP) (Araslanov & Roth, 2020) is used where the channel-wise scores \mathbf{z} are aggregated into a one-dimensional output vector $\hat{\mathbf{y}}_{\text{nGWP}} \in \mathbb{R}^{|\mathcal{Y}^t|}$ as follows:

$$y_{\text{nGWP}}^c = \frac{\sum_{i \in \mathcal{I}} m_i^c z_i^c}{\epsilon + \sum_{i \in \mathcal{I}} m_i^c}, \quad (\text{A4})$$

with $\mathbf{m} = \text{softmax}(\mathbf{z})$ and ϵ is a small constant preventing division by zero. Moreover, to penalize the localizer from predicting very small object masks, as in Araslanov & Roth (2020), the following focal penalty term is added:

$$y_{\text{FOC}} = \left(1 - \frac{\sum_{i \in \mathcal{I}} m_i^c}{|\mathcal{I}|}\right)^\gamma \log\left(\lambda + \frac{\sum_{i \in \mathcal{I}} m_i^c}{|\mathcal{I}|}\right), \quad (\text{A5})$$

where γ and λ are the hyperparameters. The final score from the localizer is then obtained by summing the scores from Eq. (A4) and Eq. (A5) namely $\hat{\mathbf{y}} = \hat{\mathbf{y}}_{\text{nGWP}} + \hat{\mathbf{y}}_{\text{FOC}}$.

C.3 THE PSEUDO-SUPERVISION SCORES \tilde{q}^c

The pixel level predictions of the localizer are combined with the old model predictions to generate the pseudo-supervision scores \tilde{q}^c as follows. First, the predicted binary segmentation maps $\hat{\mathbf{q}}^c$ (hard assignments) are smoothed with the softmax scores:

$$\mathbf{q}^c = \alpha \hat{\mathbf{q}}^c + (1 - \alpha) \mathbf{m}^c, \quad (\text{A6})$$

where $\hat{q}_i^c = 1$ if $c = \arg \max_{k \in \mathcal{Y}^t} m_i^k$ and 0 otherwise.

Then to get the final values to supervise the update of the segmentation module, for the new classes ($c \in \mathcal{C}^t$) the smoothed scores \mathbf{q}^c from the localizer are considered, for the old classes the old model is trusted, while concerning the background the two outputs are combined. Concretely:

$$\tilde{\mathbf{q}}^c = \begin{cases} \min(\tilde{\mathbf{y}}^c, \mathbf{q}^c) & \text{if } c = 'bkg', \\ \mathbf{q}^c & \text{if } c \in \mathcal{C}^t, \\ \tilde{\mathbf{y}}^c & \text{otherwise,} \end{cases} \quad (\text{A7})$$

where $\tilde{\mathbf{y}} = \sigma((F \circ E)^{t-1}(\mathbf{x}))$.

D FURTHER QUALITATIVE RESULTS

We conclude by providing additional qualitative results. In Fig. A4, we show further comparison of RaSP with WILSON on various incremental settings that differ by the number of tasks: 15-5 VOC (2 tasks), 10-5 VOC (3 tasks)



Figure A4: Qualitative results from different *single-step* and *multi-step overlap* incremental settings on VOC. The are from the final step of the corresponding settings

and 10-2 VOC (6 tasks). In Fig. A5 we show further examples with the old model prediction and similarity maps between the image label and old classes. Finally, in Fig. A6 we show failure cases for the new class due to lack of semantically similar class, lack of good region detection or low similarity with the predicted class. In Fig. A7 failure cases are depicted for the old classes where the new class model takes over the old class model (severe forgetting).

From the Fig. A4 we can see that in the 15-5 VOC setting, the WILSON overestimates the “train” pixels due to the fact that it uses CAM-like objective under the hood, which suffers from spuriously correlated “tracks” in the background – a general problem among the WSSS methods (Lee et al., 2021). On the other hand, as RaSP derives dense pseudo-supervision from previously encountered base class, e.g., “bus”, which never occurs alongside “train tracks”, it hinders the CAM-like objective to put mass on the “train tracks”. This is indeed an interesting property offered by the semantic similarity loss of RaSP, which leads to improved segmentation. Similarly, for the other settings we can observe that RaSP leads to improved foreground segmentation. Finally, for the harder 10-2 VOC setting, we notice that WILSON predicts much of the “dog” pixels to be belonging to the class “tv-monitor”, since the “tv-monitor” class is learned in the final task. This happens due to the recency-bias issue described in Sec. B.5. While RaSP also partially suffers from the same problem, but with lesser severity than WILSON.

In the Fig. A5 we provide additional visualizations from the 10-2 VOC setting and highlight the overconfident predictions of the old model on unseen classes. As shown by the $(F \circ E)^{t-1}(x_t)$ column in the Fig. A5, the old model at step $t - 1$ predicts the unseen foreground objects to be belonging to the previously learnt classes. This observation is quite contradictory to the conventional knowledge, established in the class-incremental segmentation literature (Cermelli et al., 2020), that the old model will assign all the unseen classes pixels as the *bkg* due to the *background-shift* issue. As an example, in the first and third rows of the Fig. A5 the old model predicts the “dog” and “horse” (both unseen) as “cat” and “dog” (both previously seen), respectively. Our proposed RaSP capitalizes on these predictions to obtain denser supervision for free.

Indeed there are also some instances, (see the 5th row in the Tab. A5) where the old model rightfully predicts previously unseen objects (“person”) as the class *bkg*, in-line with background-shift issue. Even in such scenarios, RaSP is able to correctly segment the “person” object without suppressing the signal from the CAM objective.. In summary, RaSP

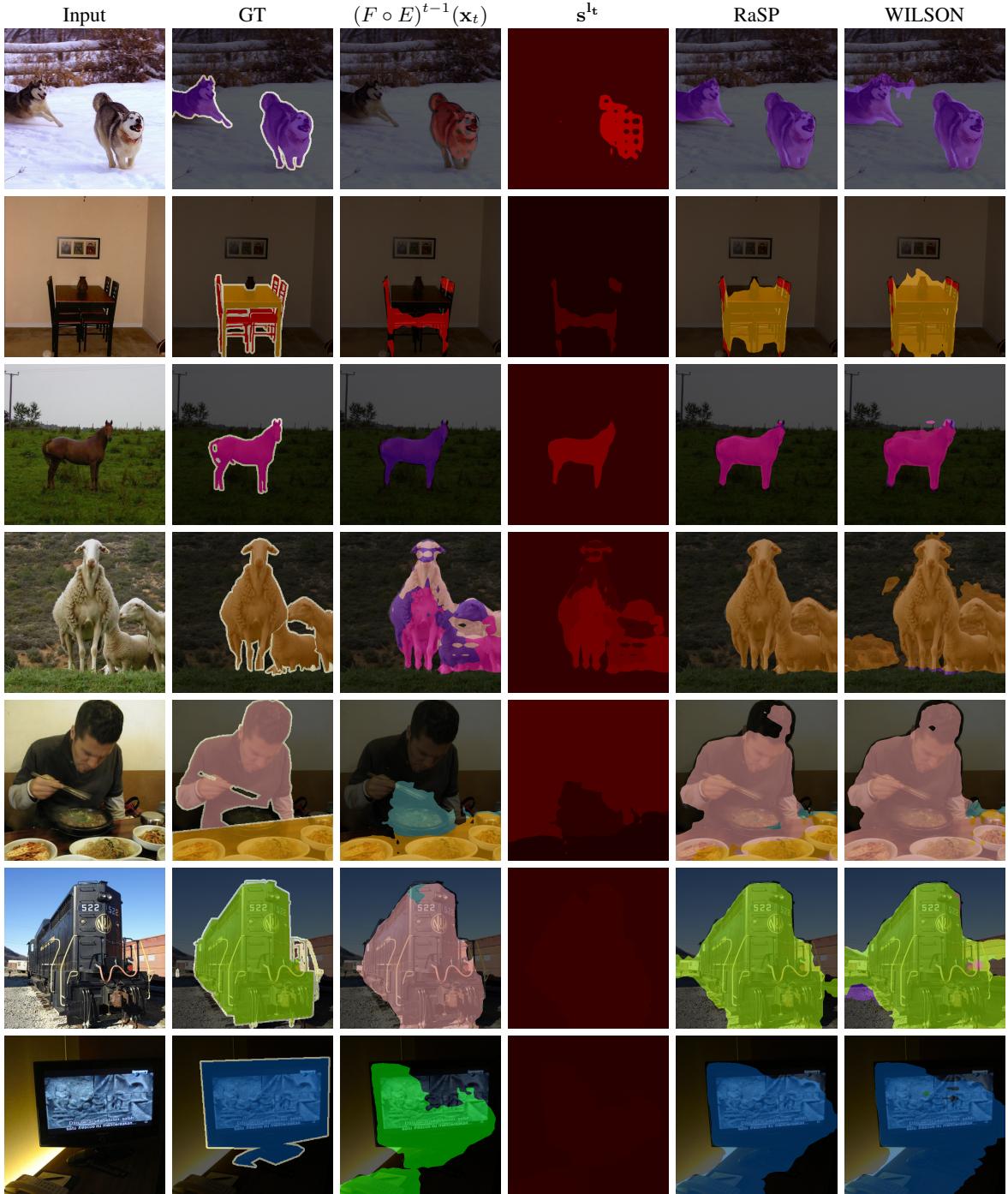


Figure A5: Visualizations. Qualitative figures from the *multi-step overlap* incremental protocol on 10-2 VOC. From left to right: input image, GT segmentation overlayed, predicted segmentation from old model, semantic similarity map computed between the image label and old classes, predicted segmentation obtained with RaSP and with WILSON. Semantic similarity maps displayed in OpenCV colormap HOT (low high similarity)

can inherit all the advantages from the WILSON framework, and even goes further to help refine its predictions when WILSON fails.

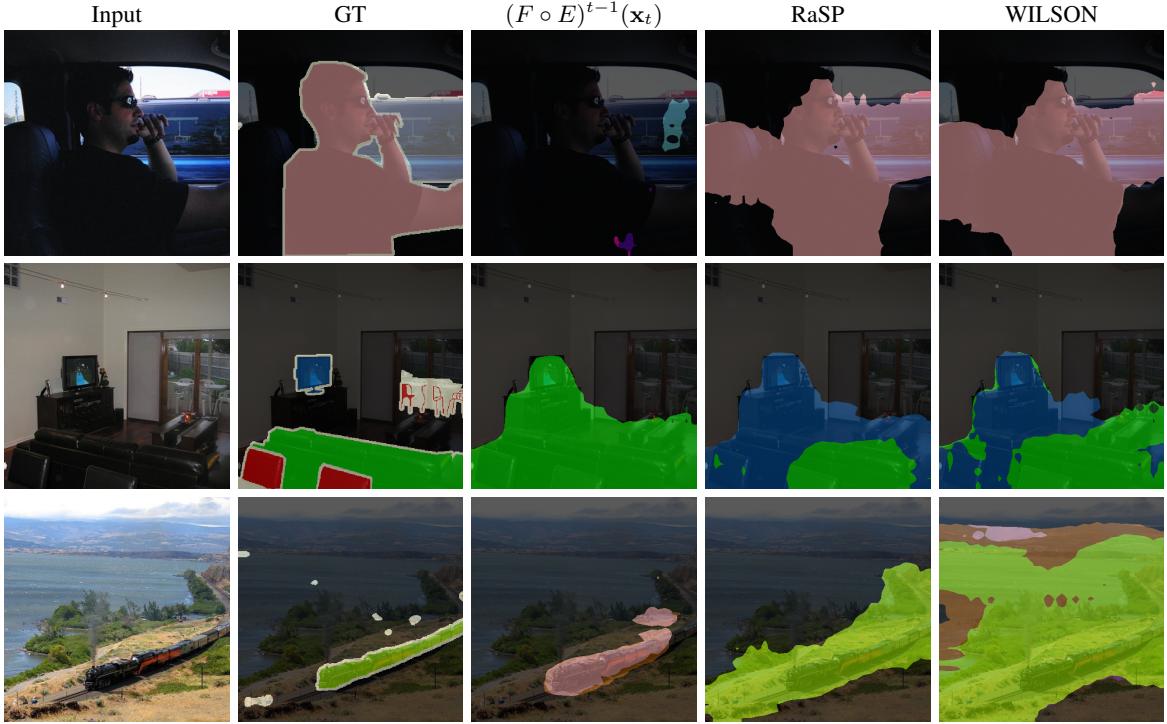


Figure A6: **Failure cases on new classes.** Figures from the *multi-step overlap* incremental protocol on 10-2 VOC. From left to right: input image, GT segmentation overlaid, predicted segmentation from old model, predicted segmentation obtained with RaSP and with WILSON

Despite the successes shown by RaSP, it is far from perfect. We showcase the failure cases on both the new and the old classes in the Tab. A6 and Tab. A7, respectively. In the Tab. A6 we can observe that both WILSON and RaSP fail to satisfactorily segment the weakly-labelled new classes. Given the old model predictions are either not present or insufficient, the proposed RaSP loss can not guide the model to the right regions of the foreground. Simultaneously, we also observe failure on the old classes, which are demonstrated in the Tab. A7. We can see that the base classes “cow”, “bicycle” and “chair”, etc are mostly segmented as the newly learnt classes, both by WILSON and RaSP, despite the old model correctly segmenting them. Given that we use the pseudo-labels supervision from the localizer to re-train the main segmentation head, it wipes away previously learned information about the old classes. Note that this phenomenon is not introduced by the RaSP loss, and is rather caused due to the pseudo-labelling loss of WILSON, as described in Eq. (5) of the main paper.

E DISCUSSION

In this section we discuss some of the edge-cases where our proposed RaSP may fail to provide clean pseudo-labels for supervision. In particular, we discuss about two of such cases where ambiguity in pseudo-supervision may arise.

In the *first* case we can imagine a scenario, where the new class (*e.g.*, “sheep”) co-occurs with an old class (*e.g.*, “cow”) in the current task image. Due to strong visual similarity, the old model will predict “sheep” pixels as “cow”, whereas the localizer will predict “sheep” pixels correctly. Such conflict is introduced by WILSON’s design because it needs to make a decision for the pseudo-label of a pixel given the predictions of both the old model and the localizer (see Eq. (7) in (Cermelli et al., 2022)). The assumption made in WILSON is that the localizer will predict new classes with far higher confidence than the old model. We do not introduce any additional ambiguity for this given use-case because our proposed RaSP loss creates the semantic similarity maps only for the new classes (“sheep” in this case) and not for the old class “cow” (note the subscript in Eq. (3), where $c \in \mathcal{C}_t$, i.e., new classes). In practice, we observe that the old and new classes co-occurring in the new task images do not happen quite often. If such co-occurrences happen only a few times then the model is able to handle this ambiguity.



Figure A7: **Failure cases on old classes.** Figures from the *multi-step overlap* incremental protocol on 10-2 VOC. From left to right: input image, GT segmentation overlayed, predicted segmentation from old model, predicted segmentation obtained with RaSP and with WILSON

In the *second* case, with the introduction of several new classes at once, there is a possibility of confusion in the semantic similarity maps. In detail, the ambiguity will specifically arise when there are more than one new class (*e.g.*, “horse” and “sheep”) in a given new image that has strong visual similarity with an old class (*e.g.*, “cow”). To recap, we have access to image-level labels only for the new classes, and the old model $(E \circ F)^{t-1}$ predicts the “horse” and “sheep” pixels as “cow” owing to strong visual similarity among them. As a result, the estimated semantic similarity maps for the “horse” and “sheep” channels will be simultaneously high for the pixel locations where these two objects are present. This is not ideal because it can drive the model to misclassify “horse” as “sheep” and vice versa. However, such co-occurrences do not happen quite often, which is evident from only a small drop in performance (-0.7%) in the 60-20 COCO-to-VOC. With enough data the model will eventually learn to ignore these noisy pseudo-labels coming from a small fraction of images.

As a future work we plan to reduce the ambiguities introduced by WILSON and RaSP, and as a consequence provide cleaner supervisory signal to the model.