Supplementary Material:
Semantic Prior for Weakly Supervised Class-Incremental Segmentation

The Supplementary material 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.

A Implementation details of RaSP

A.1 Semantic Similarity Metric

The similarity metric $\mathbf{S}_\Omega$ 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:

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

where $\omega(c_i)$ and $\omega(c_j)$ represent the vectorial embeddings for the i^{th} and j^{th} classes. The value of $\mathbf{S}_\Omega$ 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 $\omega(c) = \texttt{Transformer}(\text{“}\texttt{An image of a } \{l_c\}\text{”})$. 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 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.

B Additional experiments

B.1 Ablations

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, respctively). 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 VOC and COCO do not provide train/test/validation sets, hence, it is hard to tune the hyperparameters without overfitting the test set. For this reason, we did not spend computational resources into hyperparameter validation and based our decisions on the aforementioned heuristics.

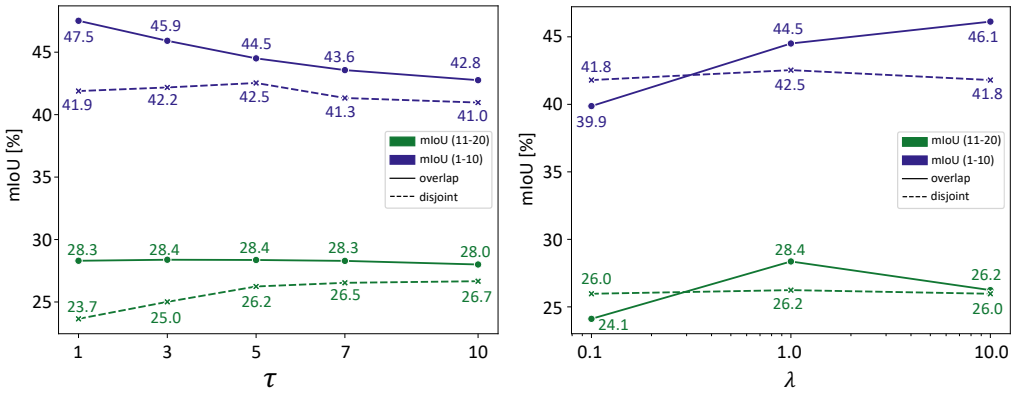

Figure A1: **Ablating τ (left) and γ (right).** RaSP results on VOC, using the 10-2 setting (6 tasks). Solid and dashed lines indicate overlap and disjoint results, respectively. Blue and green lines indicate performance on old and new classes, respectively.

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 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† increase as we increase the number of tasks.

Method	Supervision	**15-5** (2 tasks) 1-15	16-20	All	**10-10** (2 tasks) 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 (2021)	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
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† Cermelli et al. (2022)	Image	75.8	45.2	69.3	63.7	51.1	59.0
RaSP (Ours)	Image	**75.9** (↑0.1%)	**47.5** (↑5.1%)	**69.9** (↑0.9%)	**64.5** (↑1.3%)	51.2 (↑0.2%)	59.4 (↑0.7%)
		10-5 (3 tasks) 1-10	11-20	All	**10-2** (6 tasks) 1-10	11-20	All
WILSON† Cermelli et al. (2022)	Image	58.6	45.3	53.6	36.4	20.8	30.6
RaSP (Ours)	Image	**60.5** (↑3.2%)	**46.8** (↑3.3%)	**55.3** (↑3.2%)	**42.5** (↑16.8%)	**26.2** (↑26.0%)	**36.6** (↑19.6%)

Table A2: The m-IoU (in %) scores for both *single-step* (top) and *multi-step* (bottom) **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 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) 1-15	16-20	All	**10-1** (11 tasks) 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† (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† + $\mathcal{M}$	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† + $\mathcal{M}_{\text{ext}}$	Image	74.2	30.3	64.3	**62.0**	33.9	49.5
	RaSP (Ours) + $\mathcal{M}_{\text{ext}}$	Image	**74.7**	**35.8**	**66.1**	61.7	**37.4**	**51.2**
	RECALL (Web) (Maracani et al., 2021)	Pixel	67.6	49.2	64.3	62.3	50.0	57.8

Table A3: **Effect of memory.** Results on single-class *multi-step* **disjoint** incremental setting on VOC. $\mathcal{M}$ and $\mathcal{M}_{\text{ext}}$ 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 plots shown in Fig. 4 (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).

The Fig. A3 extends the plots shown in Fig. 4 (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.

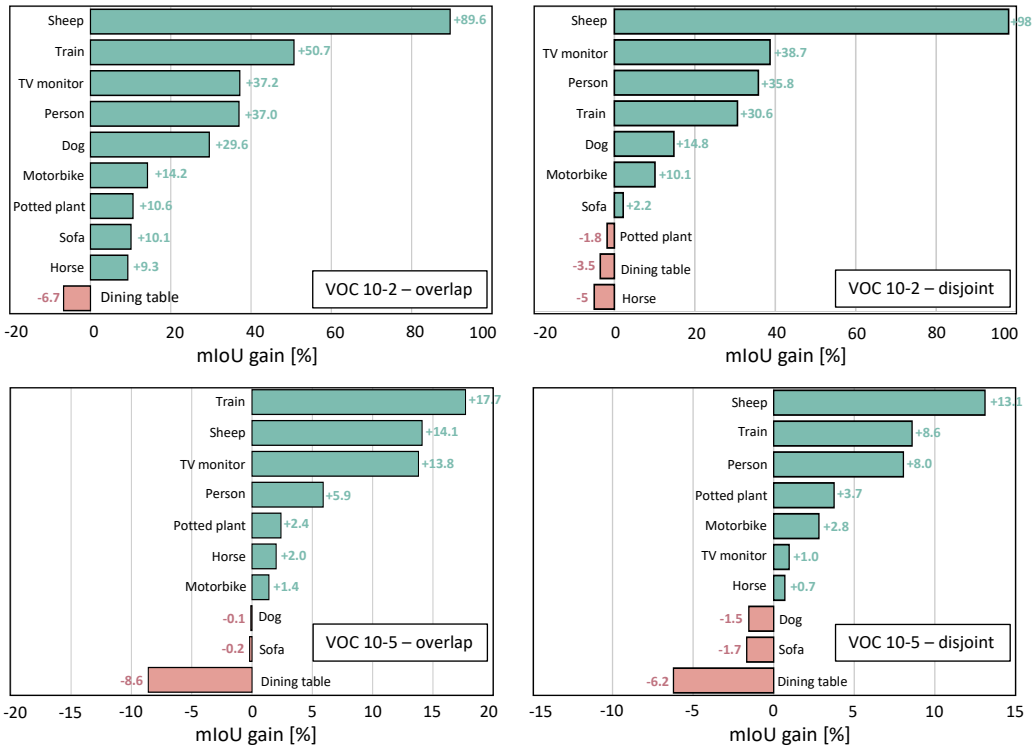

Figure A2: Per class gain/drop of RaSP w.r.t. WILSON, evaluated for each class in the step it was learned. Results computed on VOC. Top plots show **10-2** settings; bottom plots show **10-5** settings; leftmost plots show **overlap** settings; rightmost plots show **disjoint** settings. Note the different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

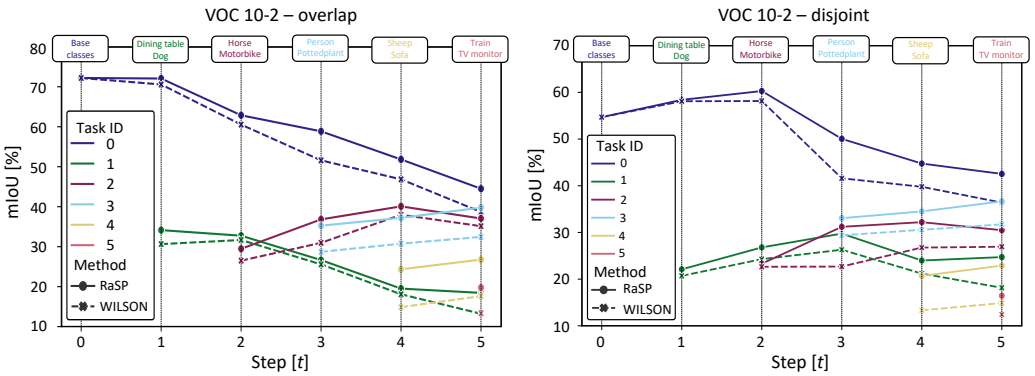

Figure A3: Per-task and per-step mIoU for the 10-2 VOC *multi-step* incremental setting. Leftmost plot shows **overlap** results; rightmost plot shows **disjoint** results. Note the different scales.

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.

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
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.

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. The summarized versions of the Tab. A5 and Tab. A6 have been reported in the Tab. 1 of the main paper.

		Old Classes																New Classes					Aggregate		
Method	Step	bkg	aeroplane	bicycle	bird	boat	bottle	bus	car	cat	chair	cow	dinningtable	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.

		Old Classes											New Classes									Aggregate		
Method	Step	bkg	aeroplane	bicycle	bird	boat	bottle	bus	car	cat	chair	cow	dinningtable	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									70.6		
	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	32.3	20.7							60.5		
	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	31.3	26.6	30.8					51.6		
	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	36.7	39.3	29.6	32.0	20.1	9.7			46.8		
	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	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									72.1		
	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	23.7							62.9		
	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	38.4	35.3	36.4	34.1					58.9		
	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			51.8		
	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 {dinningtable, 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.

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 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-5^i and COCO-20^i benchmarks. The Tab. A7 is an extended version of Tab. 4 with more pixel-supervised methods. The Tabs. A8 to 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	**64.4** (↑0.5%)	**21.3** (↑3.9%)	**32.0** (↑2.9%)	**63.5** (↑0.3%)	**10.7** (↑4.9%)	**18.3** (↑4.0%)	**45.1** (↑0.2%)	5.6 (↓3.4%)	10.0 (↓2.9%)	43.5 (↓0.2%)	**2.0** (↑5.3%)	**3.8** (↑5.6%)

Table A7: **Few-shot results.** The mIoU (in %) scores for the *single-step* (2 tasks) incremental few-shot SiS settings on the PASCAL-5^i and COCO-20^i 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** (↑0.5%)	**19.8** (↑5.3%)	**30.6** (↑4.4%)	**60.2** (0.0%)	**23.0** (↑2.2%)	**33.3** (↑1.5%)	**61.4** (↑0.5%)	**21.7** (↑1.9%)	**32.1** (↑1.6%)	**69.0** (↑0.7%)	**20.5** (↑6.8%)	**31.6** (↑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-5^i 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-5^i 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-20^i 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%)	(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-20^i 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, $\mathcal{L}_{\text{KDE}}$, – 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}:

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

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 $\mathcal{L}_{\text{KDL}}$ – 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:

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

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:

$$y^c_{\text{nGWP}} = \frac{\sum_{i \in \mathcal{I}} m_i^c z_i^c}{\epsilon + \sum_{i \in \mathcal{I}} m_i^c} \tag{A4}$$

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:

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

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:

$$\mathbf{q}^c = \alpha \hat{\mathbf{q}}^{c^*} + (1 - \alpha)\mathbf{m}^c \tag{A6}$$

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:

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

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) 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 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

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.

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 $(F \circ E)^{t-1}(\mathbf{x}_t)$ 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 5^{th} row in the Fig. 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 can inherit all the advantages from the WILSON framework, and even goes further to help refine its predictions when WILSON fails.

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 Fig. A6 and Fig. A7, respectively. In the Fig. 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 Fig. A7. We can see that the base classes "cow", "bicycle" and

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).

"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.

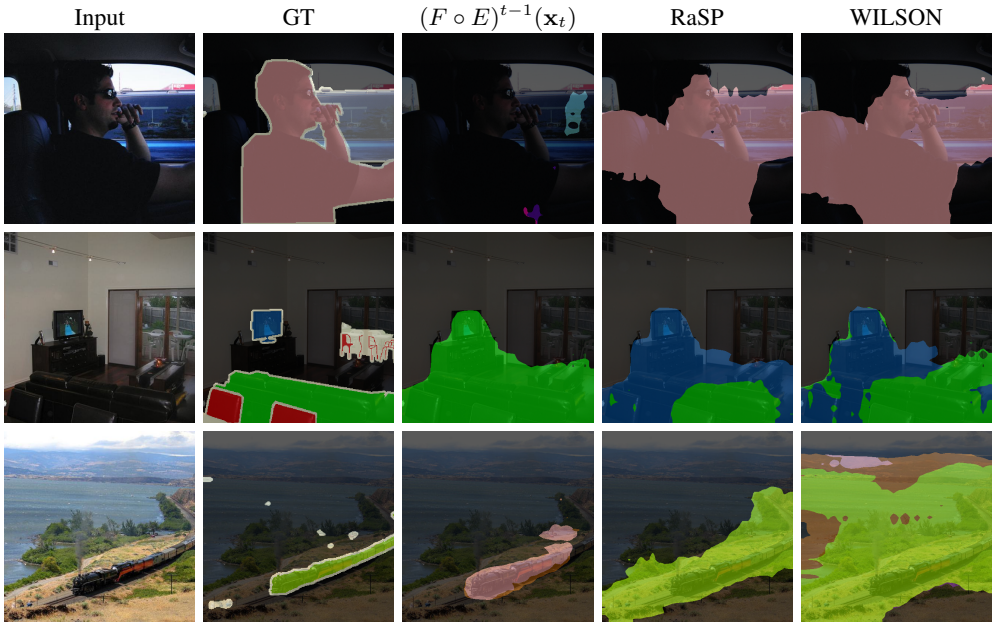

Figure A6: **Failure cases on new classes.** Figures from the *multi-step* **overlap** incremental protocol on 10-2 VOC. From left to right: input image, GT segmentation overlayed, predicted segmentation from old model, predicted segmentation obtained with RaSP and with WILSON.

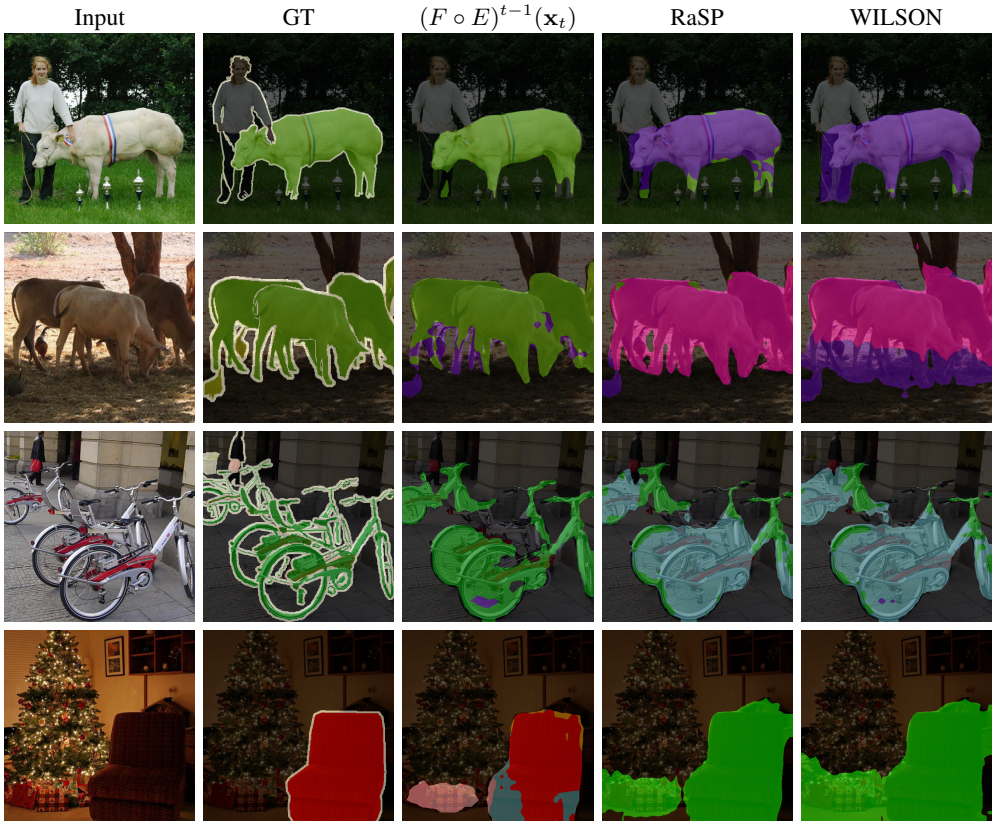

Figure A7: **Failure cases on old classes.** Figures from the *multi-step* **overlap** incremental protocol on 10-2 VOC. From left to right: input image, GT segmentation overlayed, predicted segmentation from old model, predicted segmentation obtained with RaSP and with WILSON.

REFERENCES

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*, 2020.

Fabio Cermelli, Massimiliano Mancini, Samuel Rota Bulò, Elisa Ricci, and Barbara Caputo. Modeling the Background for Incremental Learning in Semantic Segmentation. In *CVPR*, 2020.

Fabio Cermelli, Massimiliano Mancini, Yongqin Xian, Zeynep Akata, and Barbara Caputo. Prototype-based Incremental Few-Shot Semantic Segmentation. In *BMVC*, 2021.

Fabio Cermelli, Dario Fontanel, Antonio Tavera, Marco Ciccone, and Barbara Caputo. Incremental Learning in Semantic Segmentation from Image Labels. In *CVPR*, 2022.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *NAACL HLT*, 2019.

Arthur Douillard, Yifu Chen, Arnaud Dapogny, and Matthieu Cord. PLOP: Learning without Forgetting for Continual Semantic Segmentation. In *CVPR*, 2021.

Spyros Gidaris and Nikos Komodakis. Dynamic Few-Shot Visual Learning without Forgetting. In *CVPR*, 2018.

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*, pp. 5495–5505, 2021.

Zhizhong Li and Derek Hoiem. Learning without Forgetting. *ECCV*, 2016.

Andrea Maracani, Umberto Michieli, Marco Toldo, and Pietro Zanuttigh. RECALL: Replay-Based Continual Learning in Semantic Segmentation. In *ICCV*, 2021.

Umberto Michieli and Pietro Zanuttigh. Incremental Learning Techniques for Semantic Segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pp. 0–0, 2019.

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*, pp. 1114–1124, 2021.

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 *CVPR*, 2018.

Mennatullah Siam, Boris Oreshkin, and Martin Jagersand. AMP: Adaptive Masked Proxies for Few-Shot Segmentation. In *ICCV*, 2019.

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 *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*, pp. 12275–12284, 2020.

Yongqin Xian, Subhabrata Choudhury, Yang He, Bernt Schiele, and Zeynep Akata. Semantic Projection Network for Zero- and Few-Label Semantic Segmentation. In *CVPR*, 2019.

Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *CVPR*, 2016.