1. Additional Implementation Details

This section contains pseudo-code (Alg. 1) to help clarify the overall training procedure as well as further implementation details on the localizer training included in the WILSON framework. We recall that the localizer, denoted as \( g \), takes as input the features coming from the segmentation encoder \( e \) to predict a score for all classes (background, old and new ones) i.e., \( z = g(e(x)) \in \mathbb{R}^{Y^t \times H \times W} \), where \( Y^t \) is the set of classes seen (new and old) at the incremental step \( t \). In the main paper, we described two loss functions (the classification loss \( \ell_{CLS} \) and the localization prior \( \ell_{LOC} \)) that we used to train the localizer. However, for the sake of space, we left to the supplementary material two additional losses.

**Pixel-wise refinement of the localizer output.** Following [1], we introduced a self-supervised segmentation loss on the localizer output. This loss aims to force classification scores to be locally consistent, i.e. near pixels with similar appearance should be assigned the same classification score. First, we apply a refinement process on the classification output based on the pixel-level image similarity. Then, we force the classification scores to be similar to the refined version by using a pixel-level segmentation loss only on the localizer without directly affecting the segmentation output.

To refine the classification scores, similarly to [1], we employ the Pixel Adaptive Mask Refinement (PAMR). PAMR is a parameter-free module that iteratively refines the score of each pixel. Starting from the normalized classification score \( m = \psi(z) \), PAMR refines it by considering the neighbour pixels \( N(i) \). We initialize \( m^{ref,1} = m \) and, at the \( t \)'th iteration the refined mask \( m^{ref,t} \) is computed as:

\[
m^{ref,t}_i = \sum_{n \in N(i)} \alpha_{i,n} \cdot m^{ref,t-1}(n), \tag{A1}
\]

where the pixel-level level affinity \( \alpha_{i,n} \) is a value that measures the similarity among two pixels which is computed using a kernel function \( k \) on the pixel intensities, so that:

\[
\alpha_{i,n} = \frac{e^{k(i,n)}}{\sum_{l \in N(i)} e^{k(l,i)}}, \tag{A2}
\]

where we followed the same definition of the kernel function \( k \) defined in [1] that considers the average similarity in pixel intensity on the RGB channels. As suggested in [1], we used \( 3 \times 3 \) neighborhood with different dilation rates, that we set to \( \{1, 2, 4, 8, 12\} \), and we stopped after 10 iterations. Please refer to [1] for additional details.

The refined classification scores are then converted to a pseudo ground-truth mask to compute the self-supervised segmentation loss. We ignored clashing pixels and we selected only pixels with a confidence higher than 60% of the maximum value (greater than 70% for the background class). The localizer is then trained by optimizing a weighted cross-entropy loss:

\[
\ell_{SSS} = -\sum_{i \in I} \sum_{e \in Y^t} w_e \log m_{c,i} \tag{A3}
\]

with \( w_e = \frac{|I|-M_e}{|I|} \) and \( M_e = \sum_{i \in I} m^{ref}_{c,i} \), indicating with \( c,i \) the score of class \( c \) at pixel \( i \).

**Encoder feature distillation loss.** Since the losses applied on the localizer are backpropagated on the segmentation encoder, it is possible that they will cause a shift in encoder representation, impacting negatively the segmentation performance.

We use an additional feature distillation loss to prevent the encoder’s representation from shifting towards new classes and forgetting old ones. In particular, we used a mean-squared error function between the features extracted by the current encoder \( e^t \) and the ones extracted at previous step \( e^{t-1} \). Formally, given an image \( x \), the loss is computed as:

\[
\ell_{ENC} = \frac{1}{|I|} \sum_{i \in I} (e^t(x)_i - e^{t-1}(x)_i)^2, \tag{A4}
\]

where \( I \) is the set of pixels in the image and the suffix indicates the value at pixel \( i \).

---

*Equal contribution*
Overall training procedure. To sum up, the localizer has been trained, for the first 5 epochs, to minimize the following loss function:

\[ \ell_{TOT_5} = \lambda_1 \ell_{CLS} + \lambda_2 \ell_{LOC} + \lambda_3 \ell_{ENC}, \] (A5)

where \( \lambda_1, \lambda_2, \lambda_3 \) are all set to 1.

After the fifth epoch, we introduce the self-supervised segmentation loss, as in [1]:

\[ \ell_{TOT_{\lambda}} = \lambda_1 \ell_{CLS} + \lambda_2 \ell_{LOC} + \lambda_3 \ell_{ENC} + \lambda_4 \ell_{SSS}, \] (A6)

with \( \lambda_4 = 1 \).

Code. The code to replicate WILSON has been attached to the supplementary material. The code provides the scripts to replicate WILSON and the baselines for all the settings. For the offline weakly-supervised methods, we refer to the official implementations\(^1\) to generate the pseudo-labels.

2. Detailed results

2.1. Dataset class splits

We provide an extensive evaluation of WILSON on the two standard benchmarks Pascal VOC 2012 [2] and COCO [3]. Following previous work, we used two data settings on Pascal VOC: 15-5 and 10-10. To split the dataset, we follow the standard practice, and we divide them according to the alphabetic order. Tab. 1 and Tab. 2 report the classes for the 15-5 and 10-10 settings, respectively. For the COCO dataset, we split the classes according to their presence in the Pascal VOC dataset. In particular, classes in the Pascal VOC dataset are in the incremental step, while the others are in the base one. The split is reported in Tab. 3.

2.2. Class-by-class results

In this section, we report per class results on all the settings considered in the main paper. We considered the offline weakly supervised methods (WSSS) as baselines. At the same time, we could not report values for the incremental learning methods since these are not available in the considered published works.

Single step addition of five classes (15-5). In Tab. 4 and Tab. 5 are reported the results for the disjoint and overlapped settings, respectively. From the tables, we can see that WILSON outperforms all the WSSS baselines on most of the classes. In particular, considering the disjoint setting, it obtains better results on 11 out of 15 old classes and 4 out of 5 novel ones. On the overlapped, WILSON obtains even better results, being best on 12 out of 15 old classes and all the new ones. Moreover, in comparison with the

---

\(^1\)https://github.com/YudeWang/SEAM

\(^2\)https://github.com/halbielle/EPSS

\(^3\)https://github.com/visinf/1-stage-wseg
joint training with pixel-level supervision, it achieves close results and, surprisingly, even superior on some old classes (e.g., bike, boat, bottle, chair, d.table, dog, person). Differently, on the new classes, we still note a considerable performance gap, especially on classes with highly variable shapes, such as plant.

**Single step addition of five classes (10-10).** Tab. 6 and Tab. 7 report the results for the disjoint and overlapped settings. In this more challenging setting, the performance gap between WILSON and the WSSS baselines is reduced, but WILSON still obtains better results on the majority of the classes. In particular, WILSON outperforms the baselines on 2 and 3 out of 10 old classes in the disjoint and overlapped scenario, respectively, and on 4 out of 10 new classes in both scenarios. The difficulty of the setting is also confirmed by the comparison with the Joint training baseline. In particular, in the disjoint setting, WILSON achieves 10.6% less mIoU on old classes and 19.7% on the new ones, while in the overlapped setting, the performance improves on the old classes (-4.6% mIoU) but decreases on the new ones (-16.9%).

**COCO-to-VOC.** COCO-to-VOC is the most challenging scenario proposed in the paper. Not only the classes to learn are more, but they also come from different datasets, i.e., COCO on the base step and Pascal VOC on the incremental one. We report the results on the incremental classes on VOC on Tab. 8. From the results, we see that WILSON obtains the best results, outperforming SS (the second best) by 1.5% mIoU. Moreover, it achieves better results than the WSSS baselines on 11 out of 20 classes. The results on the COCO dataset are reported in Tab. 9. We note that the most challenging classes are small objects that often appear with the person class, such as skii, handbag, baseball bat, skateboard, toaster, hair-drier. The low performances can be explained considering that, at step 0, we removed from

<table>
<thead>
<tr>
<th>step</th>
<th>classes</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>aeroplane, bicycle, bird, boat, bottle, bus, car, cat, chair, cow, table, dog, horse, motorbike, person</td>
</tr>
<tr>
<td>1</td>
<td>plant, sheep, sofa, train, tv-monitor</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>step</th>
<th>classes</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>table, dog, horse, motorbike, person, plant, sheep, sofa, train, tv-monitor</td>
</tr>
<tr>
<td>1</td>
<td>person, bicycle, car, motorcycle, airplane, bus, train, boat, bird, cat, dog, horse, sheep, cow, bottle, chair, couch, potted plant, dining table, tv</td>
</tr>
</tbody>
</table>

**Table 1. Pascal VOC 15-5 class split.**

**Table 2. Pascal VOC 10-10 class split.**

**Table 3. COCO class split.**

COCO all the images containing at least a pixel from a class of VOC, including person, significantly reducing the number of samples occurring often with it (e.g., skateboard, ski). The low performances can be explained considering that, at step 0, we removed from

**3. Additional Qualitative Results**

In the main paper, we reported qualitative results for the VOC 10-10 setting. We introduce additional qualitative results here, showing results for each setting (VOC 15-5 and COCO-to-VOC) and some failure cases.

**Single step addition of five classes (15-5).** Fig. 1 shows evaluations on the Pascal VOC 15-5 setting in which five classes are added in a single step. As we can see from the images, WILSON predictions on new classes sofa, train and tv-monitor are much more accurate than those produced by EPS, even if the latter exploits an off-the-shelf saliency detector to better capture object shapes. The performances are even higher than CAM, SEAM, and SS, which tend to extend the new class predictions over the background pixels inaccurately. On old classes, WILSON is significantly more resilient to catastrophic forgetting than the other competitors, being able to properly segment both boat and cat pixels. It is also worth noting that WILSON is the only approach capable of accurately classifying the majority of the old class boat pixels, avoiding the uncertainty towards the class train that the other competitors manifest.

**COCO-to-VOC.** Fig. 2 and Fig. 3 provide the results on COCO-to-VOC setting in which additional 20 VOC classes are added in a second step, respectively evaluating each method on VOC and COCO validation sets. Fig. 2 confirms the strong performances on newer classes observed in the VOC 15-5 and 10-10 settings, demonstrating how only
WILSON is able to segment the leg of the motorcyclist and the background of the chair with significantly fewer inaccuracies than the others competitors. The results achieved on plant category show the ability of WILSON in successfully segmenting small objects as well, compared to EPS which entirely fails in accurately predicting small object boundaries. Even on the COCO validation dataset, WILSON outperforms the other state-of-the-art weakly-supervised semantic segmentation methods, as shown in Fig. 3. Indeed, it is able to segment the old classes umbrella, sandwich, parking meter and the new ones person, dog, cat, tv-monitor with much less uncertainty. Both CAM and SS perform poorly in this scenario, while SEAM and EPS still show some misclassified pixels. We also note that none of the provided approaches can classify the tennis racket correctly. We attribute this behavior to a context bias, as many training images involve a person holding a skateboard, and the models have learned to associate a person holding something to a skateboard.

**References**


Table 6. Per class results on the Pascal VOC 10-10 Disjoint setting, expressed in mIoU. Best Image-supervised method in bold.

| Method | Sup | bkg | aplane | bike | bird | boat | booth | bus | car | cat | chair | cow | dtable | dog | horse | mbike | person | plant | sheep | sofa | train | tv | Old | New | All |
|--------|-----|-----|--------|------|------|------|-------|-----|-----|-----|-------|-----|--------|-----|-------|------|--------|-------|-------|------|-------|-----|-----|-----|
| CAM    | Image | 78.5 | 79.7 | 35.6 | 81.6 | 62.3 | 77.9 | 76.6 | 78.1 | 73.1 | 24.5 | 56.1 | 38.1 | 41.9 | 28.9 | 50.2 | 47.5 | 37.6 | 40.1 | 34.1 | 49.0 | 45.2 | 65.4 | 41.3 | 54.5 |
| SEAM [5] | Image | 87.4 | 88.8 | 35.7 | 84.4 | 63.5 | 77.2 | 71.4 | 66.6 | 76.4 | 27.6 | 59.3 | 42.1 | 65.2 | 45.6 | 64.3 | 59.3 | 37.9 | 67.9 | 44.5 | 56.1 | 52.5 | 65.1 | 53.5 | 60.7 |
| SS [] | Image | 81.2 | 81.5 | 25.3 | 85.4 | 64.5 | 80.5 | 78.0 | 68.4 | 60.1 | 25.4 | 38.4 | 14.8 | 17.2 | 6.4 | 26.1 | 24.4 | 29.7 | 33.8 | 21.1 | 38.5 | 45.5 | 60.8 | 25.7 | 45.1 |
| EPS [] | Image | 89.1 | 79.2 | 35.3 | 85.2 | 66.8 | 80.2 | 61.1 | 70.6 | 74.2 | 26.0 | 60.3 | 38.1 | 64.8 | 51.1 | 68.7 | 66.8 | 43.4 | 50.8 | 39.4 | 73.8 | 44.5 | 64.2 | 54.1 | 60.6 |
| WILSON (Ours) | Image | 89.7 | 78.8 | 33.1 | 87.7 | 57.8 | 80.9 | 75.5 | 71.3 | 75.7 | 28.7 | 55.0 | 27.7 | 66.8 | 43.3 | 64.3 | 59.5 | 45.9 | 63.2 | 37.6 | 74.3 | 60.4 | 64.5 | 54.3 | 60.8 |

Table 7. Per class results on the Pascal VOC 10-10 Overlapped setting, expressed in mIoU. Best Image-supervised method in bold.

| Method | Sup | bkg | aplane | bike | bird | boat | booth | bus | car | cat | chair | cow | dtable | dog | horse | mbike | person | plant | sheep | sofa | train | tv | Old | New | All |
|--------|-----|-----|--------|------|------|------|-------|-----|-----|-----|-------|-----|--------|-----|-------|------|--------|-------|-------|------|-------|-----|-----|-----|
| CAM    | Image | 85.4 | 85.9 | 35.7 | 82.9 | 68.4 | 78.9 | 82.7 | 85.6 | 83.0 | 34.2 | 67.5 | 39.0 | 51.1 | 40.3 | 53.7 | 45.0 | 39.1 | 39.0 | 31.1 | 50.1 | 44.5 | 76.1 | 44.0 | 58.5 |
| SEAM [5] | Image | 87.1 | 84.2 | 38.5 | 84.4 | 63.3 | 77.2 | 71.7 | 76.9 | 83.1 | 33.7 | 69.8 | 44.5 | 73.8 | 60.5 | 67.3 | 59.2 | 38.0 | 65.1 | 41.3 | 54.9 | 49.3 | 67.5 | 55.4 | 62.7 |
| SS [] | Image | 79.1 | 85.8 | 38.3 | 84.1 | 67.1 | 79.4 | 84.5 | 84.8 | 78.8 | 34.0 | 59.1 | 17.4 | 40.6 | 31.9 | 40.5 | 23.6 | 29.9 | 37.9 | 24.0 | 38.4 | 44.1 | 69.6 | 32.8 | 52.5 |
| EPS [] | Image | 88.9 | 87.3 | 37.7 | 84.4 | 68.7 | 77.6 | 60.5 | 75.5 | 84.8 | 35.8 | 72.8 | 38.9 | 76.2 | 65.5 | 73.7 | 66.7 | 42.4 | 50.8 | 38.9 | 76.0 | 41.9 | 69.0 | 57.0 | 64.3 |
| WILSON (Ours) | Image | 89.1 | 83.0 | 38.1 | 86.5 | 62.8 | 79.6 | 79.3 | 83.7 | 85.4 | 33.0 | 72.5 | 29.0 | 77.8 | 64.5 | 73.9 | 59.8 | 40.9 | 67.9 | 37.5 | 57.0 | 62.3 | 70.4 | 57.1 | 65.0 |

Table 8. Per class results on Pascal VOC for the COCO-to-VOC setting, expressed in mIoU.

| Method | Sup | bkg | aplane | bike | bird | boat | booth | bus | car | cat | chair | cow | dtable | dog | horse | mbike | person | plant | sheep | sofa | train | tv | Old | New | All |
|--------|-----|-----|--------|------|------|------|-------|-----|-----|-----|-------|-----|--------|-----|-------|------|--------|-------|-------|------|-------|-----|-----|-----|
| CAM    | Image | 68.5 | 49.4 | 22.9 | 16.1 | 54.0 | 37.8 | 24.3 | 51.0 | 31.7 | 47.7 | 40.4 | 39.8 | 49.4 | 45.2 | 40.2 | 50.1 | 22.8 | 22.5 | 39.4 | 33.4 | 34.5 | 39.1 |
| SEAM [5] | Image | 76.3 | 65.8 | 26.1 | 42.6 | 60.9 | 54.9 | 9.7 | 51.3 | 30.9 | 76.6 | 63.3 | 63.3 | 60.2 | 52.7 | 47.1 | 59.0 | 22.9 | 28.0 | 40.2 | 37.4 | 37.2 | 48.0 |
| SS [] | Image | 81.8 | 71.2 | 34.8 | 33.7 | 71.3 | 80.7 | 41.7 | 77.5 | 57.2 | 81.1 | 69.2 | 65.6 | 65.0 | 51.3 | 51.4 | 63.7 | 27.6 | 22.6 | 49.5 | 26.4 | 37.8 | 53.5 |
| EPS [] | Image | 79.9 | 70.2 | 29.0 | 43.1 | 65.3 | 63.2 | 44.3 | 60.7 | 42.4 | 78.6 | 70.3 | 64.6 | 64.4 | 32.2 | 49.2 | 68.1 | 28.2 | 30.0 | 49.4 | 24.9 | 42.6 | 52.4 |
| WILSON | Image | 86.6 | 72.9 | 38.1 | 46.0 | 67.8 | 77.9 | 34.4 | 55.6 | 54.2 | 74.2 | 74.9 | 70.6 | 65.2 | 66.6 | 59.0 | 60.1 | 29.4 | 27.1 | 42.4 | 21.2 | 48.7 | 55.7 |

Table 9. Per class results on COCO for the COCO-to-VOC setting, expressed in mIoU. Best method in bold. VOC classes in red.
Figure 1. Qualitative results on the 15-5 VOC setting comparing different weakly supervised semantic segmentation methods. From left to right: image, CAM, SEAM [5], SS [1], EPS [3], WILSON and the ground-truth. Best viewed in color.
Figure 2. Qualitative results on the COCO-to-VOC setting evaluated on VOC validation set. From left to right: image, CAM, SEAM [5], SS [1], EPS [3], WILSON and the ground-truth. Best viewed in color.
Figure 3. Qualitative results on the COCO-to-VOC setting evaluated on COCO validation set. From left to right: image, CAM, SEAM [5], SS [1], EPS [3], WILSON and the ground-truth. Best viewed in color.
Figure 4. Qualitative results of failure on each setting, in order: VOC 15-5, VOC 10-10, COCO-to-VOC (VOC validation set) and COCO-to-VOC (COCO validation set). From left to right: image, CAM, SEAM [5], SS [1], EPS [3], WILSON and the ground-truth. Best viewed in color.