# RECALL: Replay-based Continual Learning in Semantic Segmentation
## *Supplementary Material*

Andrea Maracani*, Umberto Michieli*, Marco Toldo*†, Pietro Zanuttigh
Department of Information Engineering, University of Padova
andreamaracani@gmail.com, {umberto.michieli,toldomarco,zanuttigh}@dei.unipd.it

In this document, we present some additional material to better motivate the design choices behind our method, RECALL, along with some additional experiments. More in detail, we start by discussing the impact of the pre-training dataset used for the initialization of the ResNet101 backbone on the performance of continual semantic segmentation algorithms. Then, we further comment on the Class Mapping Module needed to perform the conversion from the class space of the GAN to the class space of the considered segmentation dataset. Then, we report some additional insights on the experimental results on the Pascal VOC2012 benchmark and include some preliminary analyses on the ADE20K dataset. Finally, some preliminary results combining RECALL with competitors are reported.

## 1. Analyses on Pre-Training

The weights of semantic segmentation deep learning architectures are typically initialized with pre-trained values computed on a large dataset for a related (but usually different) task. The most common pre-training strategy consists in using weights computed on image classification large-scale datasets, such as ImageNet [3]. On the other hand, it has been shown [8, 2] that pre-training weights on a related segmentation dataset, such as MS COCO [5], could further boost results on semantic segmentation benchmarks. However, using another semantic segmentation dataset raises some concerns about the fact that the pre-training data could contain information about the tasks to be learned in the incremental steps, thus following previous works [6, 7, 1] we decided for a more conservative approach in the main paper using ImageNet pre-training.

To further investigate this aspect we show extensive results in Table 1 comparing ImageNet and MS COCO pre-training strategies for all the considered incremental setups. Here we can see that, as expected, pre-training on MS COCO always helps incremental semantic segmentation. We argue that the motivation is at least two-fold: first, a better initialized model on the same target task could converge to a better solution, and second, a model pre-trained on MS COCO could have already learned some spatial and semantic information of classes added to the model in incremental steps (this second point is the reason why we decided to avoid using this strategy even if it leads to better results). More in detail, we can observe that incremental approaches show significant improvements of up to $15\%$. This quite large gap could be due also to the encoder freezing procedure we employ in our our work (similarly to [6, 8]) that reduces catastrophic forgetting, but at the same time does not allow the network to update the feature extraction module according to the information in the samples of the new classes at each incremental step, which can be used only to update the decoder. Indeed, a model initialized with pre-trained weights on MS COCO (which has enough variability and semantic content information) needs less training steps to adapt to a new (related) semantic segmentation dataset (*i.e.*, different domain but same task). On the other side, pre-training on a different task and different domain (*e.g.*, on ImageNet) requires more training steps to adapt to the new scenario. We can verify this claim looking at Table 1, where we can observe how the mIoU gap between the two pre-training strategies is larger when the initial step has fewer classes (*e.g.*, 10).

## 2. Class Mapping Module

Here we provide some further analysis and insights on the Class Mapping Module (introduced in Section 5 of the main paper), which is used to translate each class of the semantic segmentation incremental dataset (*e.g.*, Pascal VOC2012 [4]) to the most similar class of the GAN's training dataset (*e.g.*, ImageNet [3]). Notice that properly mapping the labels between the different domains is an important step, since incorrect pairings may easily harm the accuracy of the final model.

To solve the task, we took an Image Classifier $I$ pre-trained to address an image classification task on the GAN's

*These authors share the first authorship.

†Our work was in part supported by the Italian Minister for Education (MIUR) under the "Departments of Excellence" initiative (Law 232/2016).

dataset. Then, for each class $c$ in the current label set we select the corresponding training subset (*i.e.*, all the samples of the current training set associated to class $c$), and we sum the resulting class probability vectors from the classification output (according to $I$). An $\arg\max$ operation is then performed, to identify the GAN's class $c_G$ with the highest probability score. To show the effectiveness of the proposed classification, we report in Table 3 the 3 classes from the GAN's dataset with the highest score for each class of the Pascal VOC2012 dataset. We can see that for all the classes the top selected pairings appear reasonable at first (notice that only the best matching class is selected in the proposed approach). At a closer look, we find that the classifier selects an unexpected label only in a single case, that is the *person* class being translated into *cowboy hat*; however, we remark that the ImageNet dataset does not contain the *person* class, thus inherently lacking a close match for that category. In light of this, we believe that the chosen class (*i.e.*, *cowboy hat*) is a reasonable choice and may still help in retaining high accuracy on the *person* class, being the *cowboy hat* always shown on top of people's heads. This situation is interesting as it shows the robustness of our approach not only to different domains with different statistical distributions (ImageNet domain versus Pascal VOC2012 one), but also to different labeling domains (the label set of VOC2012 is not a subset of the ImageNet one).

The effectiveness of the mapping can also be visually appreciated from the sample generated images shown in Figure 1 of Section 4. In particular, notice how even for the *person* class mapped to the *cowboy hat* the images look reasonable, even if the variability in this case is much smaller if compared to the original VOC class data.

## 3. Per-Class Quantitative Results

For a more detailed evaluation, we present the per-class IoU values for some of the proposed approaches and scenarios. We considered the following methods in the disjoint scenario on all the experimental protocols: fine-tuning (FT), background inpainting, RECALL (GAN), RECALL (Web) and joint training. The results are summarized in Table 5. From here, we can appreciate how fine-tuning always catastrophically forgets previous classes when learning new ones. The simple background inpainting strategy allows to largely alleviate such phenomenon bringing a similar effect to recent knowledge distillation approaches [8, 1]. On top of this, we apply GAN or Web-based replay strategies to regularize training and background content inpainting scheme to reduce bias toward the background. While these strategies are specifically designed to preserve old knowledge, they also allow to achieve large mIoU gains on new classes reducing the false positive rate (*i.e.*, the detection of new classes in locations containing the old ones).

In order to better understand the effect of our proposed modules, we report in Table 4 the Pixel Accuracy (PA) and the IoU for the class being added at each step of the disjoint 10-1 scenario. The results demonstrate that, on the newly introduced class, FT generally achieves a very high PA (top-left) and a per-class IoU (top-right) comparable to the other approaches. Yet, FT concurrently shows very low mIoU over all classes learned up to the current step (bottom-right), as well as over only previously seen categories (bottom-left). All combined, this is indicative of an overestimation of the new class. In other words, FT progressively forgets foregoing semantic information, while predicting more often the newly seen class (which experiences high PA but low IoU, due to many false positive predictions). Our approach, instead, can effectively improve knowledge preservation thanks to replay data and background inpainting, providing steady mIoU results throughout the incremental steps.

## 4. Additional Qualitative Results

We report some additional qualitative results: Figure 1 show some examples of images produced by the GAN replay strategy. Figure 2 displays the segmentation output in the disjoint scenario for all the experimental incremental training protocols (*i.e.*, FT, background inpainting, RECALL with GAN or Web and joint). In particular, we show the results for a couple of samples in each of the 6 considered setups (*i.e.*, 19-1, 15-5, 15-1, 10-10, 10-5 and 10-1). Finally, Figure 3, instead, shows the evolution of the output maps across the incremental steps in the 15-1 scenario for a couple of sample images.

The sample generated images in Figure 1 allow to verify the effectiveness of the conditioned image generation strategy. Notice how in most cases the images are very similar to the Pascal VOC ones, even for the *person* class that does not have a direct mapping as discussed in Section 2. For the sake of comparison, the figure also reports 2 randomly sampled images for each class taken either from the Pascal VOC2012 dataset (first 2 columns of Figure 1) or from Flickr for the Web approach (last 2 columns of Figure 1).

In Figure 2 it is possible to see that the background inpainting strategy constitutes a clear improvement with respect to the simple fine-tuning approach, allowing to reduce catastrophic forgetting, which is very critical in FT. However, forgetting is still fairly noticeable with the sole inpainting strategy, where the output maps are quite noisy and relevant parts of the objects get lost in many scenes, typically overestimating the background class. However, the addition of replay data in both the GAN and the Web-based solutions proves to be very effective in further reducing the forgetting phenomenon, thus providing a final segmentation performance very close to the joint-training reference except for some details, which are typically close to the boundaries of the objects. Furthermore, our approaches do not mislead

Table 1: Mean IoU achieved by the proposed approach on the Pascal-VOC 2012 dataset for different incremental setups and pre-training strategies.

| | | **19-1** | | | | | | **15-5** | | | | | | **15-1** | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | **Disjoint** | | | **Overlapped** | | | **Disjoint** | | | **Overlapped** | | | **Disjoint** | | | **Overlapped** | | |
| **Method** | **Init** | 1-19 | 20 | all | 1-19 | 20 | all | 1-15 | 16-20 | all | 1-15 | 16-20 | all | 1-15 | 16-20 | all | 1-15 | 16-20 | all |
| GAN | ImageNet | 65.2 | 50.1 | 65.8 | 67.9 | 53.5 | 68.4 | 66.3 | 49.8 | 63.5 | 66.6 | 50.9 | 64.0 | 66.0 | 44.9 | 62.1 | 65.7 | 47.8 | 62.7 |
| | MSCOCO | 68.7 | 58.4 | 69.3 | 68.6 | 59.6 | 69.3 | 70.3 | 58.4 | 68.5 | 70.3 | 59.5 | 68.7 | 70.4 | 55.5 | 67.8 | 70.8 | 57.5 | 68.6 |
| Web | ImageNet | 65.0 | 47.1 | 65.4 | 68.1 | 55.3 | 68.6 | 69.2 | 52.9 | 66.3 | 67.7 | 54.3 | 65.6 | 67.6 | 49.2 | 64.3 | 67.8 | 50.9 | 64.9 |
| | MSCOCO | 69.7 | 55.3 | 70.1 | 68.8 | 60.7 | 69.5 | 70.7 | 59.2 | 69.0 | 70.7 | 59.9 | 69.1 | 70.9 | 57.4 | 68.7 | 71.2 | 55.7 | 68.5 |

| | | **10-10** | | | | | | **10-5** | | | | | | **10-1** | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | **Disjoint** | | | **Overlapped** | | | **Disjoint** | | | **Overlapped** | | | **Disjoint** | | | **Overlapped** | | |
| **Method** | **Init** | 1-10 | 11-20 | all | 1-10 | 11-20 | all | 1-10 | 11-20 | all | 1-10 | 11-20 | all | 1-10 | 11-20 | all | 1-10 | 11-20 | all |
| GAN | ImageNet | 62.6 | 56.1 | 60.8 | 65.0 | 58.4 | 63.1 | 60.0 | 52.5 | 57.8 | 60.8 | 52.9 | 58.4 | 58.3 | 46.0 | 53.9 | 59.5 | 46.7 | 54.8 |
| | MSCOCO | 68.2 | 64.5 | 67.6 | 68.3 | 66.1 | 68.4 | 68.2 | 62.1 | 66.4 | 67.3 | 61.9 | 65.8 | 68.2 | 58.3 | 64.6 | 67.8 | 60.5 | 65.4 |
| Web | ImageNet | 64.1 | 56.9 | 61.9 | 66.0 | 58.8 | 63.7 | 63.2 | 55.1 | 60.6 | 64.8 | 57.0 | 62.3 | 62.3 | 50.0 | 57.8 | 65.0 | 53.7 | 60.7 |
| | MSCOCO | 68.0 | 64.9 | 67.7 | 68.2 | 66.4 | 68.4 | 68.6 | 63.9 | 67.4 | 67.6 | 64.6 | 67.3 | 68.0 | 58.0 | 64.3 | 68.5 | 62.5 | 66.7 |

Table 2: mIoU on VOC2012 disjoint 15-1 with replay data. G: GAN, F: Flickr. Naïve: only decoder of last step is used for pseudo-labeling, ours: our complete approach (RECALL) is used.

| | none | + naïve (G) | + naïve (F) | + ours (G) | + ours (F) |
|---|---|---|---|---|---|
| ILT | 5.4 | 37.8 (+32.4) | 39.9 (+34.5) | 49.6 (+44.2) | 51.5 (+46.1) |
| MiB | 37.9 | 49.3 (+11.4) | 50.5 (+12.6) | 63.5 (+25.6) | 65.7 (+27.8) |
| SDR | 48.1 | 53.1 (+05.0) | 55.8 (+07.7) | 65.5 (+17.4) | 66.5 (+18.4) |

previous classes with similar ones introduced in the incremental steps (*e.g.*, FT and inpainting mislead the *cow* with *sheep* in row 3 and the *bus* with *train* in row 4).

The accuracy boost introduced by the proposed replay strategies can be further appreciated in Figure 3, where we report the segmentation output computed after each incremental step of the 15-1 disjoint setup for a couple of image samples. The improvement can be noted by looking, for example, at the images on the second and fourth incremental steps, where the new classes *sheep* and *train* are introduced respectively. When FT or background inpainting are adopted, the segmentation network tends to experience a severe forgetting of the old classes *cow* and *bus* (which are mistaken for visually similar novel ones). This behavior is corrected by providing replay training data to the network: both GAN and Web-based strategies are able to preserve an accurate recognition of old classes, even when semantically similar ones are incrementally added.

## 5. Combining RECALL with Competitors

To the best of our knowledge, no works on continual semantic segmentation using GAN-generated or web-crawled data exist. The aim of our work is to provide a general framework to retrieve and employ unlabeled replay data. In this section, we demonstrate that our framework can be applied on top of competing approaches to improve their performance: some experimental results are shown in Table 2. Adding replay data with naïve pseudo-labeling (*i.e.*, using the decoder of the previous step) already leads to a performance improvement, but combining our method with previous approaches leads to much higher results with improvements ranging from 17% to 46%, proving the effectiveness and general applicability of the modules introduced in RECALL.

## 6. Preliminary Analyses on ADE20K

In the main paper we reported the experiments on Pascal VOC2012, which contains object-level classes. However, when addressing real-world tasks, the complete understanding of the surroundings is usually required: for instance, to distinguish a mixture of *stuff* and object-level classes, as in the ADE20K dataset [9]. Indeed, the ADE20k dataset poses a great challenge due to the vast class set, comprising *stuff* categories not present in Pascal VOC2012. We remark that exact correspondence between GAN's conditioning class space and semantic segmentation category set is not required by our replay strategy. For example, as we have already observed, the *person* class is missing in the ImageNet dataset (used for pre-training the generative model), but images of people can still be retrieved from generated images of semantically related categories, such as *hat* (see Table 3). Thus, even when the generative model cannot be

explicitly conditioned to reproduce some specific segmentation classes (*e.g.*, *stuff* categories), it is possible to retrieve instances of them just relying on semantically correlated categories. This retrieval (*i.e.*, mapping) operation is performed automatically by our approach. This is even more true for the web-replay strategy, where we have complete control over the keywords used for the search.

Hence, we argue that our approach is suitable for continual semantic segmentation even when non-object categories are present. To make our point, we run preliminary experiments on ADE20k. We consider the $100$-$10$ setting, where $100$ classes are learned in the first step and the others are added in batches of $10$ at each incremental step. The FT baseline reaches a very low mIoU of $0.8\%$, while our RECALL with GAN-based replay samples improves the score up to $11.4\%$, showing that our methods mitigate catastrophic forgetting even in this challenging setup. To achieve these results, we did not perform any parameter tuning (*i.e.*, we kept the same pre-processing, learning parameters, batch and image sizes used for VOC2012). Further experiments and tuning will allow us to provide a proper comparison with other works on this benchmark.

## References


[1] Fabio Cermelli, Massimiliano Mancini, Samuel Rota Bulò, Elisa Ricci, and Barbara Caputo. Modeling the background for incremental learning in semantic segmentation. In *Int. Conf. on Computer Vision and Pattern Recognition*, 2020.

[2] 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 Trans. on Pattern Analysis and Machine Intelligence*, 40:834–848, 2018.

[3] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Fei-Fei Li. Imagenet: A large-scale hierarchical image database. In *Int. Conf. on Computer Vision and Pattern Recognition*, pages 248–255, 2009.

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

[5] 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 *European Conference on Computer Vision*, pages 740–755. Springer, 2014.

[6] Umberto Michieli and Pietro Zanuttigh. Incremental Learning Techniques for Semantic Segmentation. In *International Conference on Computer Vision Workshops*, 2019.

[7] Umberto Michieli and Pietro Zanuttigh. Continual semantic segmentation via repulsion-attraction of sparse and disentangled latent representations. In *Int. Conf. on Computer Vision and Pattern Recognition*, 2021.

[8] Umberto Michieli and Pietro Zanuttigh. Knowledge distillation for incremental learning in semantic segmentation. *Computer Vision and Image Understanding*, 2021.

[9] Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ade20k dataset. In *Int. Conf. on Computer Vision and Pattern Recognition*, 2017.

Table 3: Class mapping between Pascal VOC and ImageNet datasets. The table shows the 3 best matching ImageNet classes for each Pascal VOC 2012 class. (*): matching classes for *tv/monitor* are not computed since replay data is not needed.

| **Pascal** | | **ImageNet** | | |
|---|---|---|---|---|
| index | class | 1st class | 2nd class | 3rd class |
| 1 | airplane | airliner | warplane | wing |
| 2 | bicycle | mountain bike | tandem | tricycle |
| 3 | bird | kite (bird) | dipper | quail |
| 4 | boat | catamaran | lakeside | fireboat |
| 5 | bottle | beer bottle | soda bottle | water bottle |
| 6 | bus | trolleybus | carriage | minibus |
| 7 | car | racing car | station wagon | minivan |
| 8 | cat | tabby cat | Egyptian cat | tiger cat |
| 9 | chair | rocking chair | dining table | folding chair |
| 10 | cow | ox | oxcart | water ox |
| 11 | dining table | dining table | china closet | restaurant |
| 12 | dog | Labrador retriever | pit bull terrier | beagle |
| 13 | horse | sorrel | ox | fox squirrel |
| 14 | motorbike | moped | scooter | disc brake |
| 15 | person | cowboy hat | crash helmet | crutch |
| 16 | potted plant | pot | pencil case | greenhouse |
| 17 | sheep | ram | llama | bighorn sheep |
| 18 | sofa | studio couch | quilt | rocking chair |
| 19 | train | carriage | electric locomotive | freight car |
| 20 | tv/monitor* | - | - | - |

Table 4: Per-round accuracy measures in the 10-1 disjoint scenario. In the top part we report the PA (left) and IoU (right) of the last class currently introduced. The bottom part, instead, shows the mean IoU over the old classes up to the ongoing step (left), as well as the overall mean IoU including the new classes (right). The classes added at each incremental step are: 1:*dining table*, 2:*dog*, 3:*horse*, 4:*motorbike*, 5:*person*, 6:*potted plant*, 7:*sheep*, 8:*sofa*, 9:*train* and 10:*tv/monitor*. Best in **bold**.

| PA (new) | step 1 | step 2 | step 3 | step 4 | step 5 | step 6 | step 7 | step 8 | step 9 | step 10 |
|---|---|---|---|---|---|---|---|---|---|---|
| FT | 69.2 | **90.9** | **87.1** | **79.2** | 88.3 | 3.8 | 26.9 | **49.9** | 55.0 | 57.5 |
| ILT [6] | 70.2 | 88.3 | 44.3 | 52.8 | 86.9 | 58.7 | 22.5 | 43.6 | 48.8 | 56.3 |
| MiB [1] | **75.5** | 87.2 | 38.5 | 51.2 | **89.6** | **61.0** | 6.1 | 38.2 | 47.4 | **60.2** |
| SDR [7] | 68.7 | 73.2 | 34.0 | 31.0 | 84.5 | 55.6 | 15.7 | 39.0 | 45.1 | 55.8 |
| RECALL (GAN) | 24.0 | 56.7 | 38.1 | 58.5 | 76.2 | 27.1 | 36.8 | 25.4 | 53.0 | 45.0 |
| RECALL (Web) | 23.8 | 57.3 | 46.1 | 62.3 | 79.3 | 36.9 | **42.8** | 26.9 | **64.2** | 57.4 |

| IoU (new) | step 1 | step 2 | step 3 | step 4 | step 5 | step 6 | step 7 | step 8 | step 9 | step 10 |
|---|---|---|---|---|---|---|---|---|---|---|
| FT | 16.2 | 37.0 | 15.8 | 12.6 | **78.4** | 3.2 | 9.2 | 16.3 | 9.7 | 19.9 |
| ILT [6] | 14.6 | 37.2 | 14.5 | 11.5 | 68.6 | 5.5 | 4.9 | 10.3 | 11.1 | 13.8 |
| MiB [1] | 14.6 | 38.0 | 12.2 | 9.5 | 66.0 | 9.6 | 1.7 | 5.5 | 10.9 | 7.0 |
| SDR [7] | **25.0** | 54.0 | 13.3 | 10.9 | 69.2 | 10.9 | 5.4 | 8.9 | 15.4 | 17.8 |
| RECALL (GAN) | 22.4 | 53.9 | 36.1 | 54.6 | 64.4 | 23.6 | 34.4 | 23.6 | 49.3 | 38.8 |
| RECALL (Web) | 22.1 | **54.5** | **43.7** | **58.4** | 67.9 | **30.5** | **39.6** | **24.7** | **58.3** | **45.0** |

| mIoU (old) | step 1 | step 2 | step 3 | step 4 | step 5 | step 6 | step 7 | step 8 | step 9 | step 10 |
|---|---|---|---|---|---|---|---|---|---|---|
| FT | 50.4 | 23.9 | 26.1 | 20.5 | 4.6 | 4.7 | 4.2 | 3.7 | 3.8 | 3.5 |
| ILT [6] | 59.4 | 33.5 | 35.2 | 20.9 | 11.7 | 11.2 | 10.9 | 10.2 | 7.1 | 7.2 |
| MiB [1] | 64.3 | 53.6 | 61.1 | 26.0 | 25.3 | 27.6 | 19.0 | 16.8 | 9.7 | 12.6 |
| SDR [7] | 64.9 | 57.5 | 61.6 | 35.2 | 30.2 | 32.4 | 28.7 | 27.5 | 19.2 | 21.0 |
| RECALL (GAN) | **74.7** | **64.7** | 63.4 | 63.2 | 60.7 | 62.9 | 57.9 | 56.5 | 53.5 | 54.6 |
| RECALL (Web) | 74.3 | 63.9 | **64.6** | **65.0** | **63.5** | **65.5** | **60.6** | **60.1** | **58.3** | **58.4** |

| mIoU (all) | step 1 | step 2 | step 3 | step 4 | step 5 | step 6 | step 7 | step 8 | step 9 | step 10 |
|---|---|---|---|---|---|---|---|---|---|---|
| FT | 47.3 | 25.0 | 25.3 | 19.9 | 9.5 | 4.6 | 4.5 | 4.4 | 4.1 | 4.3 |
| ILT [6] | 55.3 | 33.8 | 33.6 | 20.2 | 15.5 | 10.8 | 10.5 | 10.2 | 7.3 | 7.5 |
| MiB [1] | 59.8 | 52.3 | 57.3 | 24.8 | 28.0 | 26.5 | 18.0 | 16.2 | 9.8 | 12.3 |
| SDR [7] | 61.3 | 57.2 | 57.9 | 33.5 | 32.8 | 31.1 | 27.3 | 26.5 | 19.0 | 20.8 |
| RECALL (GAN) | **70.4** | **63.9** | 61.5 | 62.7 | 60.9 | 60.6 | 56.6 | 54.8 | 53.3 | 53.9 |
| RECALL (Web) | 69.9 | 63.2 | **63.1** | **64.6** | **63.7** | **63.5** | **59.4** | **58.3** | **58.3** | **57.8** |

Table 5: Per-class IoU of compared methods in disjoint experimental protocol on multiple scenarios of Pascal VOC 2012.

| | Method | backgr. | aero | bike | bird | boat | bottle | bus | car | cat | chair | cow | din. table | dog | horse | mbike | person | plant | sheep | sofa | train | tv | old | new | all |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 19-1 | FT | 72.4 | 62.4 | 6.7 | 45.0 | 47.1 | 39.5 | 33.7 | 40.9 | 25.7 | 4.3 | 54.0 | 8.0 | 25.0 | 50.4 | 50.6 | 0.0 | 35.3 | 43.0 | 0.8 | 59.5 | 13.2 | 35.2 | 13.2 | 34.2 |
| | Inpainting | 91.0 | 83.9 | 35.1 | 77.3 | 62.3 | 70.7 | 77.9 | 73.4 | 85.7 | 31.5 | 73.1 | 48.0 | 81.3 | 74.4 | 64.6 | 81.0 | 44.1 | 75.7 | 41.3 | 74.5 | 30.4 | 66.1 | 30.4 | 65.6 |
| | GAN | 91.7 | 82.8 | 32.3 | 82.6 | 62.8 | 74.1 | 86.2 | 79.6 | 86.0 | 30.0 | 58.9 | 45.9 | 80.5 | 67.9 | 73.4 | 80.6 | 35.3 | 62.9 | 39.6 | 77.9 | 50.1 | 65.2 | 50.1 | 65.8 |
| | Web | 91.4 | 82.8 | 35.9 | 83.4 | 59.9 | 73.5 | 85.3 | 73.7 | 85.7 | 31.3 | 59.4 | 40.9 | 81.1 | 67.1 | 73.4 | 80.5 | 43.1 | 61.5 | 42.6 | 74.4 | 47.1 | 65.0 | 47.1 | 65.4 |
| | Joint | 92.5 | 89.9 | 39.2 | 87.6 | 65.2 | 77.3 | 91.1 | 88.5 | 92.9 | 34.8 | 84.0 | 53.7 | 88.9 | 85.0 | 85.1 | 84.9 | 60.0 | 79.7 | 47.0 | 82.2 | 73.5 | 75.5 | 73.5 | 75.4 |
| 15-5 | FT | 72.4 | 62.4 | 6.7 | 45.0 | 47.1 | 39.5 | 33.7 | 40.9 | 25.7 | 4.3 | 54.0 | 8.0 | 25.0 | 50.4 | 50.6 | 0.0 | 35.3 | 43.0 | 0.8 | 59.5 | 13.2 | 35.2 | 13.2 | 34.2 |
| | Inpainting | 89.0 | 68.7 | 36.0 | 68.2 | 48.4 | 71.4 | 12.8 | 77.3 | 85.6 | 26.7 | 8.1 | 48.8 | 80.3 | 61.6 | 68.8 | 78.7 | 20.1 | 29.0 | 26.3 | 38.4 | 51.8 | 56.1 | 33.1 | 52.2 |
| | GAN | 90.4 | 78.8 | 35.0 | 79.5 | 60.3 | 75.7 | 79.3 | 78.7 | 85.9 | 22.8 | 55.0 | 46.6 | 80.0 | 67.4 | 72.1 | 77.8 | 37.3 | 60.2 | 32.2 | 64.4 | 55.1 | 66.3 | 49.8 | 63.5 |
| | Web | 90.8 | 82.2 | 35.5 | 81.7 | 63.9 | 75.3 | 85.0 | 77.8 | 86.3 | 28.0 | 67.5 | 48.7 | 81.0 | 72.7 | 73.8 | 78.0 | 40.4 | 65.7 | 31.9 | 69.1 | 57.6 | 69.2 | 52.9 | 66.3 |
| | Joint | 92.5 | 89.9 | 39.2 | 87.6 | 65.2 | 77.3 | 91.1 | 88.5 | 92.9 | 34.8 | 84.0 | 53.7 | 88.9 | 85.0 | 85.1 | 84.9 | 60.0 | 79.7 | 47.0 | 82.2 | 73.5 | 77.5 | 68.5 | 75.4 |
| 15-1 | FT | 74.2 | 27.2 | 0.0 | 1.6 | 15.1 | 11.3 | 0.0 | 4.1 | 0.5 | 0.0 | 0.0 | 0.0 | 0.0 | 0.2 | 0.2 | 0.0 | 27.0 | 25.6 | 28.9 | 33.5 | 52.2 | 8.4 | 33.5 | 14.4 |
| | Inpainting | 85.9 | 38.9 | 31.4 | 79.4 | 41.5 | 71.3 | 28.9 | 62.6 | 85.6 | 32.2 | 29.6 | 50.2 | 76.6 | 69.2 | 55.3 | 80.2 | 18.5 | 37.4 | 36.3 | 19.8 | 17.9 | 55.5 | 26.0 | 49.9 |
| | GAN | 90.5 | 80.7 | 34.5 | 79.5 | 59.1 | 75.5 | 72.7 | 78.2 | 85.3 | 25.3 | 59.0 | 39.9 | 79.9 | 68.8 | 72.5 | 78.6 | 23.2 | 58.0 | 39.2 | 60.1 | 43.8 | 66.0 | 44.9 | 62.1 |
| | Web | 90.5 | 82.1 | 34.4 | 81.5 | 62.6 | 76.0 | 82.3 | 77.0 | 85.1 | 27.4 | 63.6 | 39.4 | 80.3 | 71.9 | 72.2 | 78.4 | 35.4 | 64.4 | 35.7 | 61.9 | 48.7 | 67.6 | 49.2 | 64.3 |
| | Joint | 92.5 | 89.9 | 39.2 | 87.6 | 65.2 | 77.3 | 91.1 | 88.5 | 92.9 | 34.8 | 84.0 | 53.7 | 88.9 | 85.0 | 85.1 | 84.9 | 60.0 | 79.7 | 47.0 | 82.2 | 73.5 | 77.5 | 68.5 | 75.4 |
| 10-10 | FT | 82.1 | 0.2 | 0.0 | 1.2 | 0.0 | 1.4 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 52.3 | 73.2 | 49.8 | 73.1 | 81.8 | 41.4 | 49.7 | 49.1 | 76.1 | 62.2 | 7.7 | 60.9 | 33.0 |
| | Inpainting | 90.9 | 81.8 | 34.1 | 73.1 | 58.6 | 73.3 | 85.6 | 78.8 | 78.2 | 29.0 | 29.1 | 43.7 | 66.6 | 47.7 | 73.0 | 74.2 | 29.6 | 57.3 | 38.8 | 70.9 | 61.4 | 62.2 | 56.3 | 60.7 |
| | GAN | 90.8 | 83.3 | 30.4 | 75.8 | 61.4 | 73.5 | 80.8 | 77.2 | 72.8 | 23.6 | 46.8 | 48.0 | 65.4 | 55.3 | 66.1 | 72.5 | 36.8 | 58.3 | 36.1 | 67.1 | 55.6 | 62.6 | 56.1 | 60.8 |
| | Web | 90.9 | 82.3 | 32.7 | 75.4 | 63.2 | 72.8 | 81.7 | 73.5 | 76.2 | 24.2 | 58.5 | 46.5 | 68.8 | 60.2 | 64.7 | 73.3 | 38.3 | 58.3 | 34.2 | 68.5 | 56.2 | 64.1 | 56.9 | 61.9 |
| | Joint | 92.5 | 89.9 | 39.2 | 87.6 | 65.2 | 77.3 | 91.1 | 88.5 | 92.9 | 34.8 | 84.0 | 53.7 | 88.9 | 85.0 | 85.1 | 84.9 | 60.0 | 79.7 | 47.0 | 82.2 | 73.5 | 76.6 | 74.0 | 75.4 |
| 10-5 | FT | 78.2 | 0.0 | 0.0 | 0.0 | 0.0 | 1.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 16.4 | 13.7 | 23.4 | 55.7 | 46.7 | 37.4 | 39.8 | 47.9 | 75.6 | 62.3 | 7.2 | 41.9 | 23.7 |
| | Inpainting | 88.5 | 58.8 | 31.9 | 55.4 | 58.2 | 69.2 | 0.2 | 78.3 | 83.2 | 28.2 | 5.0 | 36.4 | 71.6 | 34.7 | 61.4 | 74.1 | 20.4 | 26.2 | 25.5 | 34.0 | 47.9 | 46.8 | 43.2 | 47.1 |
| | GAN | 89.3 | 77.9 | 28.9 | 72.1 | 59.3 | 73.6 | 75.1 | 75.8 | 79.5 | 20.5 | 37.2 | 44.2 | 67.8 | 50.9 | 59.5 | 71.3 | 31.4 | 51.9 | 32.3 | 63.1 | 53.0 | 60.0 | 52.5 | 57.8 |
| | Web | 89.5 | 80.8 | 31.2 | 74.6 | 61.6 | 72.0 | 81.6 | 74.3 | 80.6 | 19.8 | 55.1 | 44.3 | 69.2 | 56.9 | 56.5 | 71.7 | 39.8 | 59.0 | 30.2 | 69.7 | 54.2 | 63.2 | 55.1 | 60.6 |
| | Joint | 92.5 | 89.9 | 39.2 | 87.6 | 65.2 | 77.3 | 91.1 | 88.5 | 92.9 | 34.8 | 84.0 | 53.7 | 88.9 | 85.0 | 85.1 | 84.9 | 60.0 | 79.7 | 47.0 | 82.2 | 73.5 | 76.6 | 74.0 | 75.4 |
| 10-1 | FT | 69.4 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 19.9 | 6.3 | 2.0 | 4.3 |
| | Inpainting | 85.0 | 35.3 | 28.6 | 72.6 | 37.2 | 67.9 | 16.9 | 65.3 | 83.0 | 32.1 | 13.3 | 35.1 | 41.4 | 5.0 | 22.5 | 71.7 | 21.5 | 16.9 | 34.5 | 19.2 | 13.0 | 45.2 | 28.1 | 39.0 |
| | GAN | 88.8 | 77.9 | 26.6 | 71.8 | 58.6 | 73.2 | 63.5 | 74.0 | 75.7 | 20.5 | 41.3 | 34.2 | 60.5 | 43.6 | 58.6 | 66.2 | 15.8 | 51.7 | 37.4 | 53.0 | 38.8 | 58.3 | 46.0 | 53.9 |
| | Web | 89.1 | 79.1 | 31.0 | 74.4 | 62.2 | 66.5 | 81.7 | 74.1 | 78.7 | 19.4 | 56.2 | 41.8 | 62.7 | 58.5 | 62.1 | 66.6 | 8.5 | 59.3 | 36.8 | 59.2 | 45.0 | 62.3 | 50.0 | 57.8 |
| | Joint | 92.5 | 89.9 | 39.2 | 87.6 | 65.2 | 77.3 | 91.1 | 88.5 | 92.9 | 34.8 | 84.0 | 53.7 | 88.9 | 85.0 | 85.1 | 84.9 | 60.0 | 79.7 | 47.0 | 82.2 | 73.5 | 76.6 | 74.0 | 75.4 |


Figure 1: Original images from the incremental Pascal VOC dataset, together with replay data generated by GAN or retrieved by Flickr's web crawler.

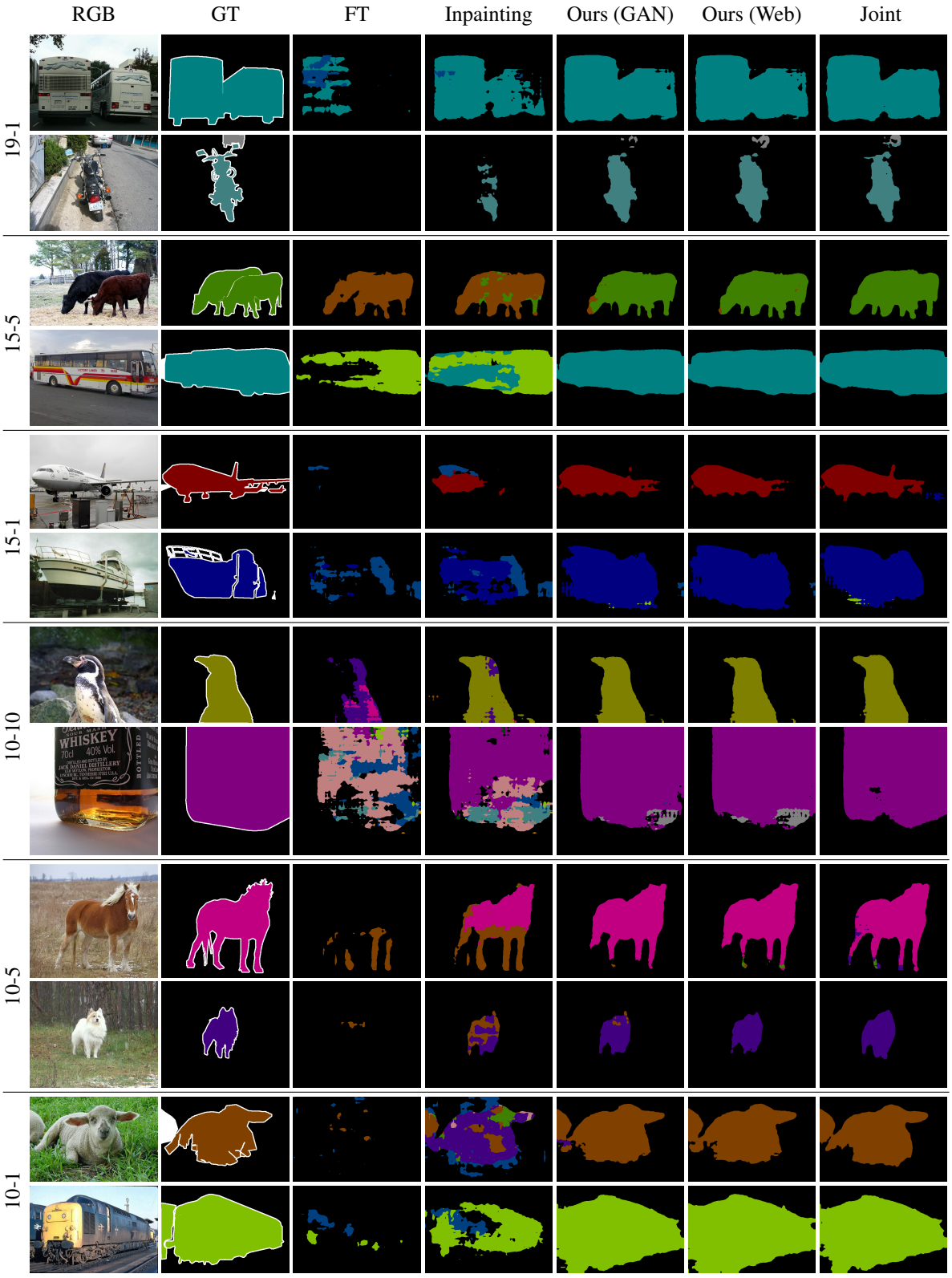


Figure 2: Qualitative results on disjoint incremental setups (*best viewed in colors*).

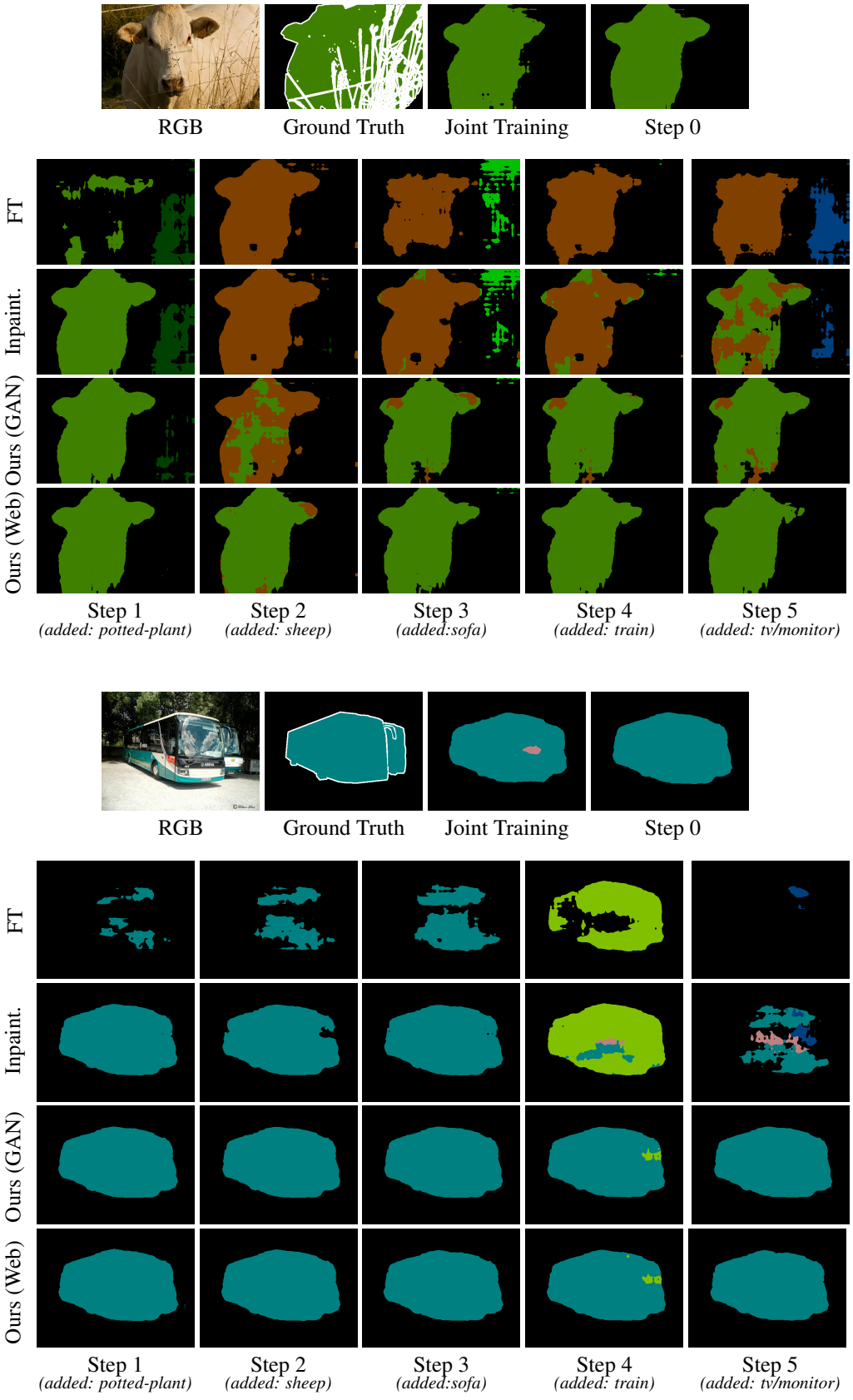


Figure 3: Per-step prediction maps on the 15-1 disjoint incremental setup for different training strategies.