Continual Semantic Segmentation via Repulsion-Attraction of Sparse and Disentangled Latent Representations
Supplementary Material

Umberto Michieli Pietro Zanuttigh
Department of Information Engineering
University of Padova
{umberto.michieli, zanuttigh}@dei.unipd.it

In this document we present some additional material to better motivate our method and we conduct some supplementary experiments. More in detail, we start by discussing some of the design choices that led to the models of the losses and constraints presented in the main paper in Section 1. Then, Section 2 shows some additional ablation studies. Finally, many additional qualitative and quantitative results for both the Pascal VOC2012 and the ADE20K datasets are presented in Sections 3, 4 and 5.

1. Design Choices

In this section we present some additional discussion and results motivating the design choices behind the various modules exploited in our work.

Prototypes Matching enforces latent space consistency on old classes, forcing the encoder to produce similar latent representation for previously seen classes in the subsequent steps. The target is achieved by considering the Euclidean distance in the latent space (see Section 4.1 of the paper). Although different distance metrics could have been used in principle (*e.g.*, cosine distance [11, 9, 7]) we found that a simple Euclidean distance was easier to understand and very computationally efficient results similar to more complex schemes.

Contrastive Learning aims at clustering features according to their semantics while tearing apart those of different classes (see Section 4.2 of the paper): we implement it as an attractive force between latent representations with their prototypical representation, against a repulsive one between prototypes of different semantic categories. This attraction-repulsion rule is enforced again using an Euclidean distance metric.

Knowledge Distillation is employed to constraint the decoder to preserve previous knowledge at the output-level and it is implemented as a standard cross-entropy on the output softmax probabilities between old model and current model predictions [5, 6, 1, 3] (see Section 4.4 of the paper).

Sparsity: We think that the most peculiar constraint is represented by the sparsity objective. However, the underlying concept is simple: applying some latent-level sparsification we allow the model to retain enough discriminative power to accommodate the upcoming representations of novel classes without cross-talk with previous ones (see Section 4.3). Here, a wide range of possibilities could be considered to address the aforementioned task and one may wonder why the sparsity constraint was designed as it is presented in the main paper. First, common sparsity losses are the L0 or L1 norms of feature vectors; however, we show that they achieve lower accuracy. In this work, we define the sparsity objective as the ratio between a stretching function (*i.e.*, the sum of exponentials) and a linear function (*i.e.*, the sum) applied to the feature vectors which were previously normalized with respect to the maximum value that is assumed by any of the feature channels for that particular class. The idea is that by keeping fixed the sum of features, then the proposed loss in Eq. (9) of the main paper is directly proportional to the degree of distribution across the channels: the value is low when the energy is concentrated in a single or in a few channels, while it increases when distributed (with a gradual change). In some extreme cases, the model of Eq. (9) could lead to degenerate solutions, however we argue that these do not happen in practice on a model learning compact representations. We checked to avoid the zero division in the practical implementation, while the *all-ones* case is degenerate in the sense that energy cannot be re-distributed in any way since all channels are already onset to the maximum value and, furthermore, this configuration would not be informative for the decoder.

Although we believe that normalizing the features with a class-conditioned guidance is reasonable (sometimes, features of few particular classes may just be on average more active than features of other classes), we can think of getting rid of it and normalizing with other strategies, *e.g.*, with respect to:

- the maximum value for each feature (*norm max*);
- the overall maximum value (*norm max overall*);
- the *L2* norm of each feature (*norm L2*).

In such cases, Eq. (8) would become respectively:

$$\bar{\mathbf{f}}_i = \frac{\mathbf{f}_i}{\max_{f_{i,j} \in \mathbf{f}_i} f_{i,j}} \qquad \mathbf{f}_i \in \mathbf{F}_n \tag{1}$$

$$\bar{\mathbf{f}}_i = \frac{\mathbf{f}_i}{\max_{g_{j,l} \in \mathbf{g}_j} g_{j,l}} \qquad \mathbf{f}_i, \mathbf{g}_j \in \mathbf{F}_n \tag{2}$$

$$\bar{\mathbf{f}}_i = \frac{\mathbf{f}_i}{||\mathbf{f}_i||_2} \qquad \mathbf{f}_i \in \mathbf{F}_n \tag{3}$$

Furthermore, in principle any stretching function could be applied in spite of the sum of exponentials over the linear sum. For instance, the sum of squares (*power 2*) or sum of the cubic powers (*power 3*) could be used as stretching functions: *i.e.*, formulating Eq. (9) respectively as:

$$\mathcal{L}_{sp} = \frac{1}{|\mathbf{f}_i \in \mathbf{F}_n|} \sum_{\mathbf{f}_i \in \mathbf{F}_n} \frac{\sum_j \bar{f}_{i,j}^2}{\sum_j \bar{f}_{i,j}} \tag{4}$$

$$\mathcal{L}_{sp} = \frac{1}{|\mathbf{f}_i \in \mathbf{F}_n|} \sum_{\mathbf{f}_i \in \mathbf{F}_n} \frac{\sum_j \bar{f}_{i,j}^3}{\sum_j \bar{f}_{i,j}}. \tag{5}$$

Finally, following the success of recent works exploiting entropy minimization [10] techniques, an alternative strategy could be to minimize the entropy of the latent representations opportunely preceded by *L1* or *softmax* normalization of each feature vector in order to obtain a probability distribution over the channels. More formally:

$$\bar{\mathbf{f}}_i = \frac{\mathbf{f}_i}{||\mathbf{f}_i||_1} \qquad \mathbf{f}_i \in \mathbf{F}_n \tag{6}$$

$$\bar{\mathbf{f}}_i = \frac{\exp(\mathbf{f}_i)}{\sum_j \exp(f_{i,j})} \qquad \mathbf{f}_i \in \mathbf{F}_n \tag{7}$$

$$\mathcal{L}_{sp} = \frac{1}{|\mathbf{f}_i \in \mathbf{F}_n|} \sum_{\mathbf{f}_i \in \mathbf{F}_n} \sum_j -\bar{f}_{i,j} \cdot \log\left(\bar{f}_{i,j}\right) \tag{8}$$

Table 1 shows the performance of the aforementioned approaches in the 19-1 and 15-1 disjoint scenarios on Pascal VOC2012. Different normalization rules bring to consistently lower results, proving the efficacy of using class guidance during normalization. Also, different stretching functions are found to be less adequate for our purpose reducing the final mIoU of about 2% to 4%. Finally, entropy minimization techniques obtain competitive and comparable results in the $15-1$ scenario, while they experience a drop of about $2-3\%$ of mIoU when only one class is added.

Table 1. Comparison of different $\mathcal{L}_{sp}$ in terms of mIoU in the disjoint scenarios 19-1 and 15-1 on Pascal VOC2012 dataset.

Method	mIoU_{19-1}	mIoU_{15-1}
L0	66.7	46.3
L1	65.9	45.4
norm max	67.4	47.8
norm max overall	67.5	45.6
norm L2	64.8	44.3
power 2	66.3	44.2
power 3	66.6	45.3
entropy (L1)	65.3	48.0
entropy (softmax)	66.0	48.0
ours	**68.4**	**48.1**

2. Additional Ablation Studies

In this section, we report a couple of additional ablation studies concerning the dataset size and the pre-training.

Random Split. Looking at Table 1 of the main paper, we see that in some cases, especially on the 15-1 setup, the proposed method is still far from the offline reference. An interesting question is whether this is due to the difficulty of handling new classes or if, more fundamentally, it is due to an inherent difficulty to train a network using only a small subset of the data at each step. To answer this, we split the dataset equally in 5 parts (each part containing all classes, thus removing the complexity of learning new classes) and then we trained the network sequentially on each of this parts. We obtained 69.9% of mIoU against 75.4% of the joint training, 5.6% of the FT (disjoint) and 48.1% of SDR (disjoint). The difference with respect to joint training is relatively small, and it could be due to sub-optimal network weights estimation (samples are taken from the 5 parts accessed subsequently, instead of the full dataset); on the other side, the difference with respect to FT is very large proving that handling unseen classes is the key issue and the proposed latent constraints aim at addressing it.

Considerations on Pre-Training. The results reported in the main paper have been obtained initializing the weights of the backbone ResNet-101 approach on the ImageNet dataset. This is the standard setup in continual semantic segmentation approaches [5, 1, 6, 3]. Additional considerations have been already addressed in [6], where it has been shown that pre-training on a segmentation benchmark could boost the accuracy; nonetheless, the ranking of the proposed strategies is mainly maintained.

On the other hand, even ImageNet contains visual samples for many of the elements present in the Pascal dataset (for classification task instead of segmentation), potentially limiting the magnitude of decay on *old* tasks, and likely raising accuracies for *new* concepts that are not necessarily completely new to the encoder. Here, we show how the network performs without such a strong prior on the la-

tent representations. The results are strongly affected by the fact that datasets for in-the-wild segmentation are often too small to reliably train complex deep networks from scratch. We trained on VOC2012 without pre-training and we achieved a low mIoU of 24.4% when training for 30 epochs, as we do in the main paper, and 40.9%, when training for 120 epochs (about 30 hours of computation). In continual learning, the final mIoU are also lower (as the starting point is much lower), but the improvements achieved by our approach and the ranking of the various methods are preserved, for instance in VOC2012 15-1 disjoint the accuracy of SDR (13.5%) is still significantly above FT (4.1%) and MiB (10.9%).

3. Additional Qualitative Results

Many qualitative experimental results are reported for all the different scenarios, experimental protocols (*i.e.*, sequential, disjoint and overlapped) and datasets.

Pascal VOC2012. The results for this dataset are reported in Figures 1, 2 and 3 respectively for sequential, disjoint and overlapped protocols. In each figure, 3 images for each scenario (*i.e.*, 19-1, 15-5 and 15-1) are depicted. We compare our method with naïve fine tuning and the competitors, *i.e.*, LwF [4], ILT [5], CIL [3] and MiB [1]. The images show how our approach is able to alleviate forgetting and at the same time accommodate for new classes to learn. On the other side, the fine-tuning and the compared approaches often deviate (*i.e.*, are biased) in predicting novel classes being added or the special background class.

ADE20K. We report several visual results in Figure 4 also for this dataset. In particular, we show 3 images for each scenario (*i.e.*, 100-50, 100-10, 50-50). Again, we can appreciate how our method largely outperforms compared approaches in all scenarios better capturing the details of the shapes of the objects (e.g, in rows 1-4) and not degenerating into an overestimation of the background (*e.g.*, in the 100-10 scenario). In particular, we notice how compared approaches have big difficulties in handling multiple additions of multiple classes (they struggle in tackling catastrophic forgetting in the 100-10 scenario), while our method can achieve reasonably good output segmentation maps also in the most challenging scenarios.

4. Qualitative Results Across Incremental Steps

In this section we analyze the performance across the various incremental steps, comparing our method with the top performing competitor (*i.e.*, MiB [1]).

Pascal VOC2012. The results on two sample scenes from this dataset are reported in Figure 5 for the disjoint 15-1 experimental protocol, where an initial training stage over 15 classes is followed by 5 incremental learning steps each carrying one class to be learned. In the first row our method shows quite robust results across the different learning steps, being able to preserve content semantics. MiB, instead, is able to avoid catastrophic forgetting for one incremental step but it degenerates after introducing the *sheep* class (step 2), which is predicted in spite of *person* probably due to the confusion of the arms and legs (caused also by their similar color). The latent representations got even more damaged across subsequent steps, while our approach (SDR) is able to reduce the interference on latent representations of old classes. Similar considerations also holds for the second set of images, although in this scenario forgetting is less evident: our approach is able to achieve superior performance thanks to correct spatial localization and latent disentanglement.

ADE20K. For this dataset we consider two distinct scenarios: *i.e.*, 5 incremental steps each adding 10 categories to the model (100-10) in Figure 6, and 2 incremental steps each adding 50 classes to the model (50-50) in Figure 7.

The first scenario is definitely the most challenging one as the model need to adapt 5 times to discover new (and possibly unrelated) classes. Nevertheless, we can appreciate that our model obtain quite robust results across the various steps in the 2 sample scenes shown in Figure 6, while MiB suffers more from catastrophic forgetting previous knowledge. In the first sample scene our approach shows a small gradual degradation across the multiple steps, while MiB firstly completely looses the wall on the background in step 2, then the curtain in step 3 and finally also the hand basin in step 4. Similarly, in the second scene our approach maintains very good results across all the steps, while MiB at the third step misleads the sky on the background.

In Figure 7 we consider the case in which only two incremental steps with 50 classes each are performed. It can be appreciated how in the first step the predicted segmentation maps are quite precise according to both our approach and MiB, but, in both examples, MiB produces a less precise map after the second incremental step. More in detail, we remark some differences: our model can identify the tree (green) in the first image, that MiB only partially captures in the first step and completely misses it in the second. Similarly, SDR preserves the walls (gray) in the second image that MiB misleads in the second step. Again, the latent space regularization helps in preserving previous classes representations and in accommodating new classes.

5. Quantitative Results: per-Class Accuracy

We also analyze the per-class accuracy for all the compared methods in some scenarios. We report the results of per-class IoU and per-class pixel accuracy (PA) on the disjoint 19-1 (Tables 2 and 3), 15-5 (Tables 4 and 5) and 15-1 (Tables 6 and 7) scenarios on the Pascal VOC2012 dataset.

Even when adding as little as 1 class (scenario 19-1 in

Tables 2 and 3) we can appreciate how FT and LwF-MC are generally able to learn the new class to some extent but they catastrophically forget previous classes resulting in a poor final mIoU. This performance drop is typically due to a biased prediction toward the new class (high per-class PA for that class but low IoU). The other competing approaches and our proposal, instead, are more balanced across the various classes and are able to greatly alleviate forgetting (with performance gains distributed across the classes) when learning the new class, thus resulting in higher mIoUs.

Analyzing the per-class IoUs on the 15-5 case in Tables 4 and 5 we can appreciate how FT is completely unable to preserve knowledge about previous classes which are heavily forgotten. The competitors can better preserve knowledge related to previous classes while learning new classes but our approach shows superior results in both retaining old classes knowledge and in learning new ones.

The last 15-1 scenario is shown in Tables 6 and 7. Here we can confirm most of the previous considerations; our method outperforms all the competitors proving its scalability when multiple incremental steps are made. From the per-class pixel accuracy we can observe that most of competing approaches are biased toward the prediction of the very few last classes added to the model, thus reducing the IoU for the other classes.

References

[1] 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 Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 1, 2, 3, 5, 6, 7, 8, 9, 10, 11, 12

[2] James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, Demis Hassabis, Claudia Clopath, Dharshan Kumaran, and Raia Hadsell. Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences (PNAS)*, 114(13):3521–3526, 2017. 5, 6, 7, 8

[3] Marvin Klingner, Andreas Bär, Philipp Donn, and Tim Fingscheidt. Class-incremental learning for semantic segmentation re-using neither old data nor old labels. *International Conference on Intelligent Transportation Systems*, 2020. 1, 2, 3, 5, 6, 7, 8, 11, 12

[4] Zhizhong Li and Derek Hoiem. Learning without forgetting. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 40(12):2935–2947, 2018. 3, 11, 12

[5] Umberto Michieli and Pietro Zanuttigh. Incremental Learning Techniques for Semantic Segmentation. In *Proceedings of the International Conference on Computer Vision Workshops (ICCVW)*, 2019. 1, 2, 3, 5, 6, 7, 8, 11, 12

[6] Umberto Michieli and Pietro Zanuttigh. Knowledge distillation for incremental learning in semantic segmentation. *Computer Vision and Image Understanding (CVIU)*, 2021. 1, 2

[7] Boris Oreshkin, Pau Rodríguez López, and Alexandre Lacoste. Tadam: Task dependent adaptive metric for improved few-shot learning. *Neural Information Processing Systems (NeurIPS)*, 31:721–731, 2018. 1

[8] Sylvestre-Alvise Rebuffi, Alexander Kolesnikov, Georg Sperl, and Christoph H Lampert. icarl: Incremental classifier and representation learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2001–2010, 2017. 11, 12

[9] Jake Snell, Kevin Swersky, and Richard Zemel. Prototypical networks for few-shot learning. In *Neural Information Processing Systems (NeurIPS)*, pages 4077–4087, 2017. 1

[10] Tuan-Hung Vu, Himalaya Jain, Maxime Bucher, Matthieu Cord, and Patrick Pérez. Advent: Adversarial entropy minimization for domain adaptation in semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2517–2526, 2019. 2

[11] Kaixin Wang, Jun Hao Liew, Yingtian Zou, Daquan Zhou, and Jiashi Feng. Panet: Few-shot image semantic segmentation with prototype alignment. In *Proceedings of the International Conference on Computer Vision (ICCV)*, pages 9197–9206, 2019. 1

Figure 1. Qualitative results on sample scenes in different scenarios (19-1, 15-5 and 15-1) on Pascal VOC 2012 of the proposed method and of competing approaches in the sequential setup (*best viewed in colors*).

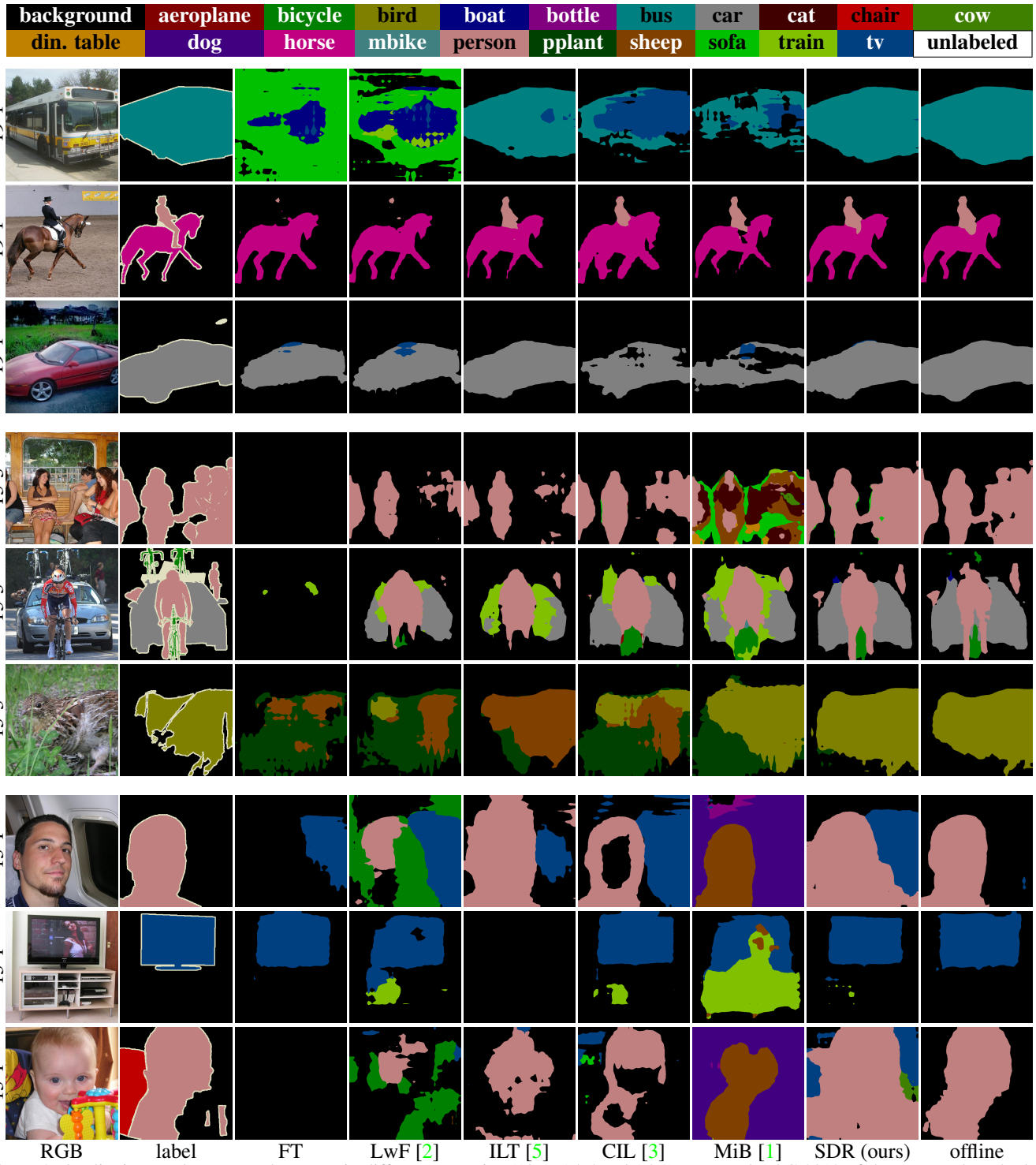

Figure 2. Qualitative results on sample scenes in different scenarios (19-1, 15-5 and 15-1) on Pascal VOC 2012 of the proposed method and of competing approaches in the disjoint setup (*best viewed in colors*).

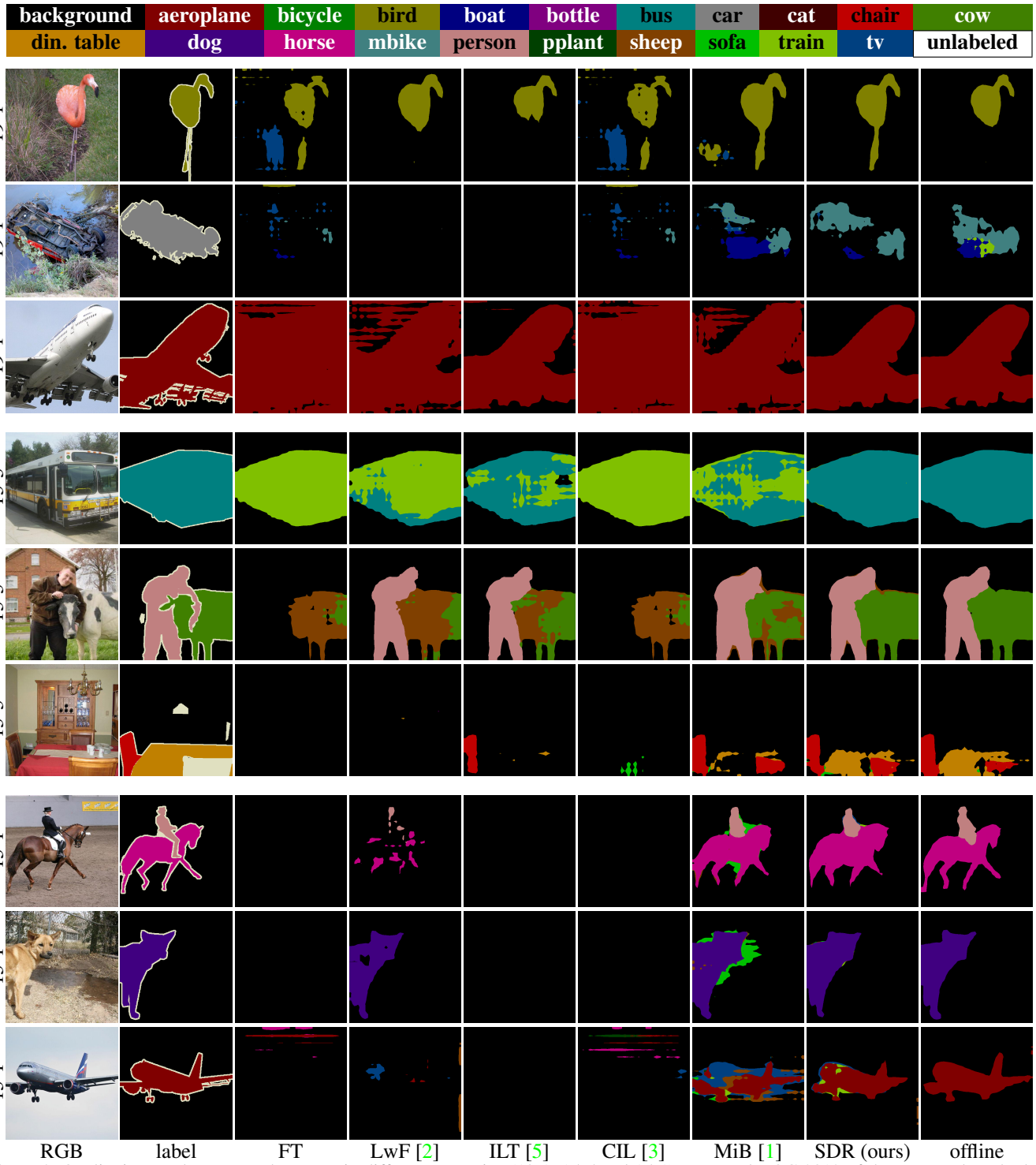

Figure 3. Qualitative results on sample scenes in different scenarios (19-1, 15-5 and 15-1) on Pascal VOC 2012 of the proposed method and of competing approaches in the overlapped setup (*best viewed in colors*).

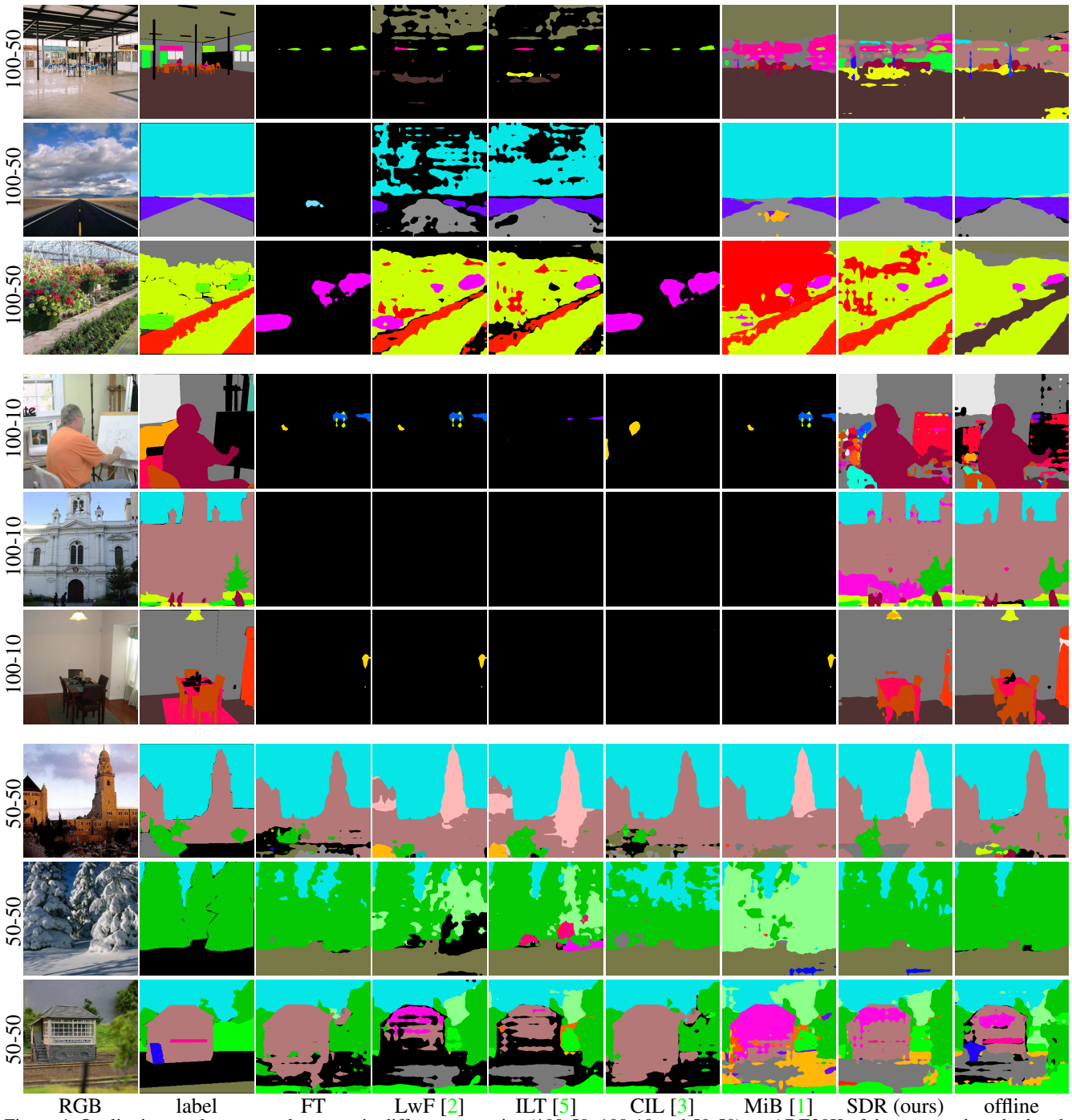

Figure 4. Qualitative results on sample scenes in different scenarios (100-50, 100-10 and 50-50) on ADE20K of the proposed method and of competing approaches (*best viewed in colors*).

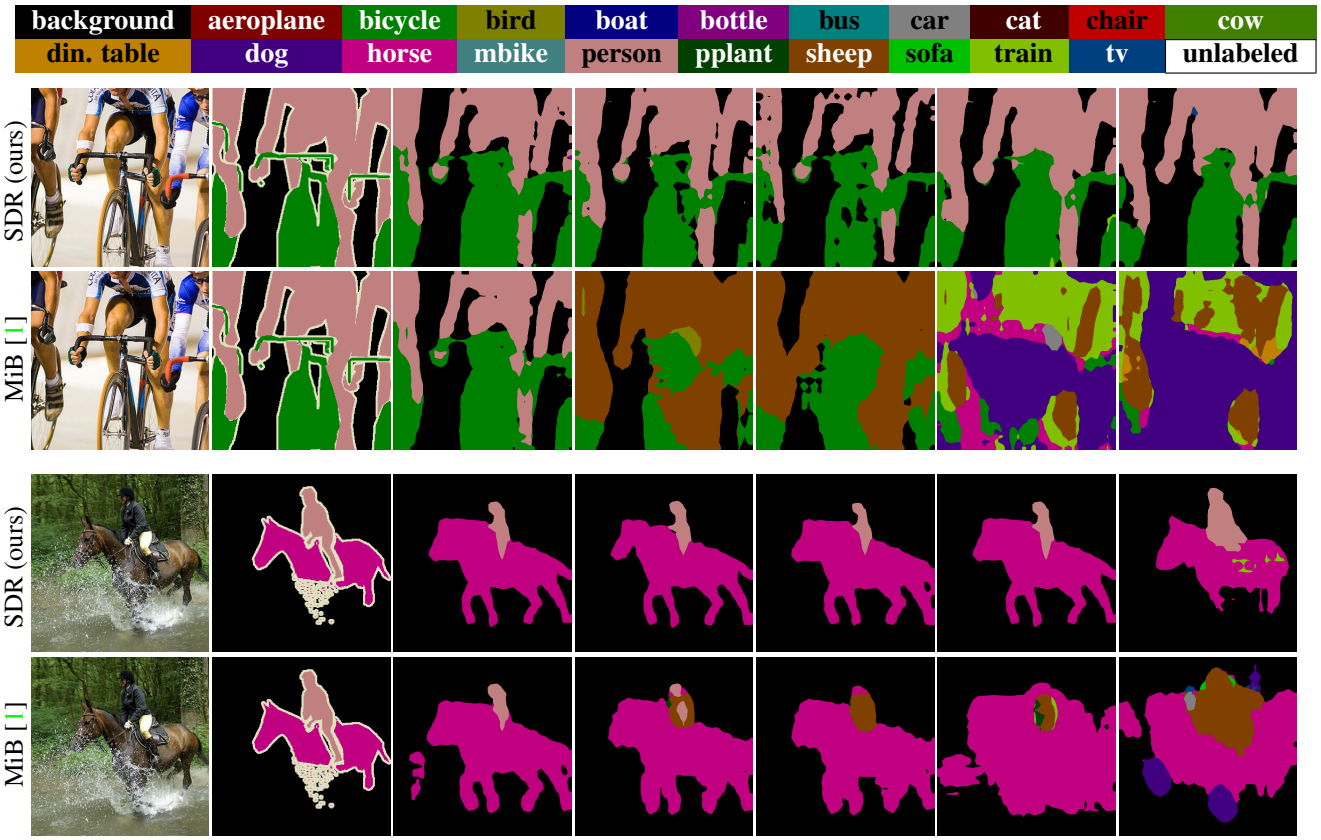

Figure 5. Qualitative results on sample scenes in the disjoint experimental protocol 15-1 on Pascal VOC 2012 during the various incremental steps (*best viewed in colors*).

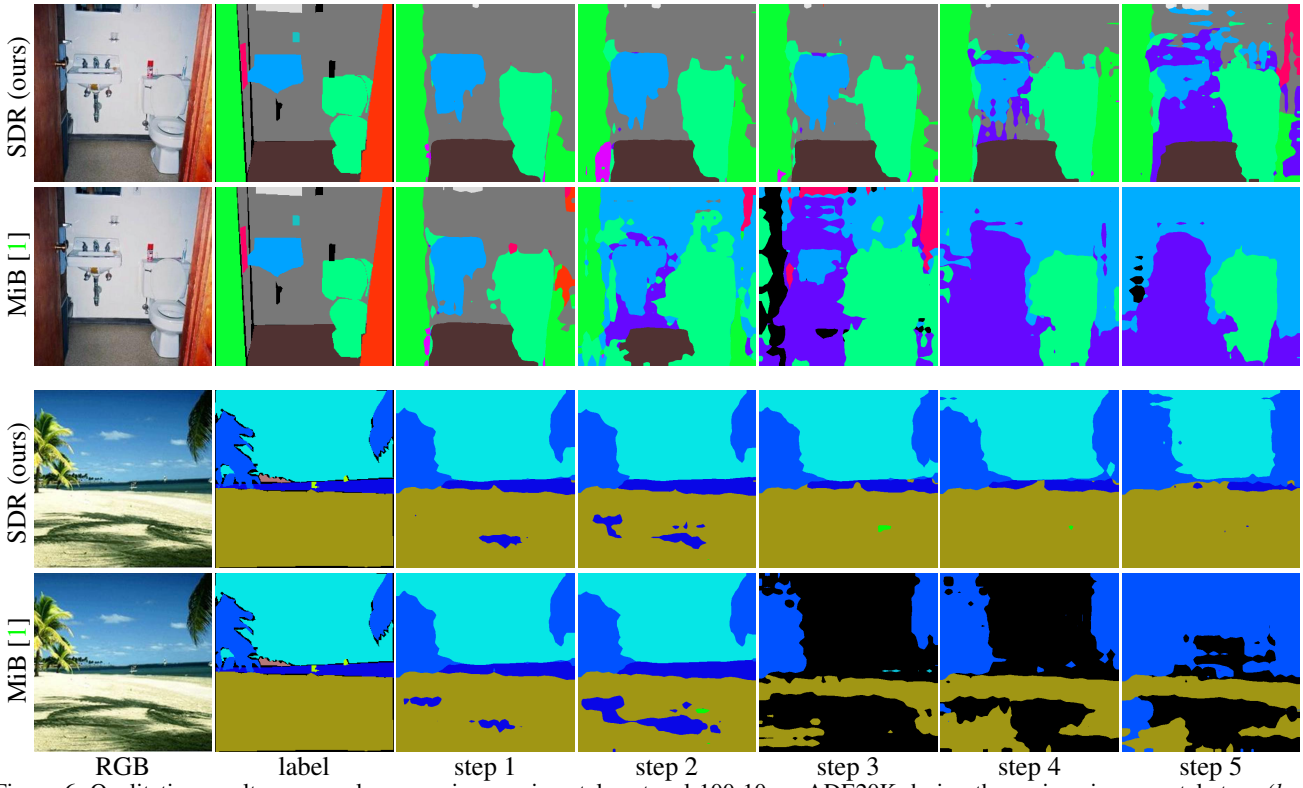

Figure 6. Qualitative results on sample scenes in experimental protocol 100-10 on ADE20K during the various incremental steps (*best viewed in colors*).

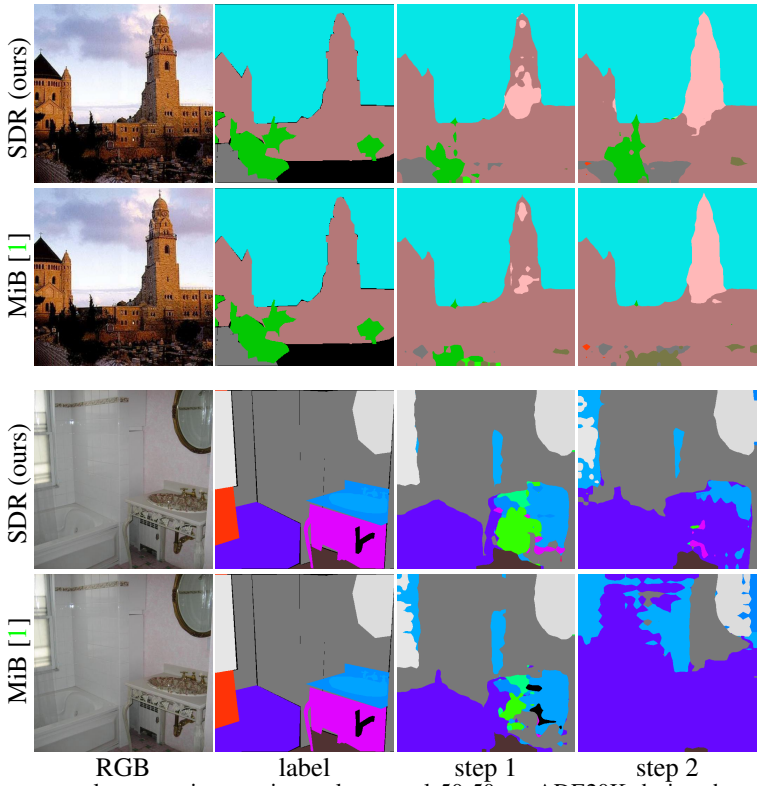

Figure 7. Qualitative results on sample scenes in experimental protocol 50-50 on ADE20K during the various incremental steps (*best viewed in colors*).

Table 2. Per-class IoU of compared methods in disjoint experimental protocol on scenario 19-1 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
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
LwF [4]	87.6	75.4	31.1	71.7	50.8	66.0	81.6	79.0	87.9	32.1	66.9	49.9	84.1	66.2	77.3	79.4	51.8	68.5	42.1	65.8	28.3	65.8	28.3	64.0
LwF-MC [8]	78.6	63.6	0.4	61.2	10.6	35.2	52.8	35.1	75.5	0.4	63.9	1.5	75.5	67.8	32.6	13.1	13.0	63.4	0.7	25.9	1.0	38.5	1.0	36.7
ILT [5]	87.7	79.5	31.6	77.4	54.5	66.5	70.9	79.0	90.4	31.4	66.5	52.9	85.1	67.7	78.1	82.0	56.0	67.3	41.4	72.3	23.4	66.9	23.4	64.8
CIL [3]	85.3	71.4	33.6	75.2	56.5	59.3	45.8	67.2	85.9	27.6	62.7	46.9	85.2	67.9	75.2	83.7	47.4	67.0	42.3	66.0	18.1	62.6	18.1	60.5
MiB [1]	86.9	73.5	35.7	64.0	50.5	71.0	89.5	87.0	84.8	33.7	62.9	56.9	82.1	61.8	79.5	82.4	56.2	62.0	46.0	75.9	26.0	67.0	26.0	65.1
SDR (ours)	89.6	85.3	35.9	78.6	55.2	73.6	86.2	81.9	89.1	34.2	71.4	56.6	86.5	72.7	78.0	83.0	54.1	71.0	45.5	70.4	37.3	69.9	37.3	68.4
SDR+MiB	89.5	84.4	39.0	76.5	53.6	75.1	89.1	87.6	89.0	33.7	67.8	55.4	85.2	72.8	80.8	83.4	57.8	71.3	46.3	78.4	31.4	70.8	31.4	68.9
offline	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

Table 3. Per-class pixel accuracy of compared methods in disjoint experimental protocol on scenario 19-1 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
FT	91.5	79.9	7.2	74.9	71.1	44.0	34.3	46.4	26.1	4.5	72.6	8.1	25.4	78.0	53.9	0.0	40.6	58.5	0.8	64.3	82.0	35.2	13.2	34.2
LwF [4]	94.1	85.6	58.7	91.2	59.1	76.3	84.4	80.3	94.1	39.3	93.5	52.3	91.7	95.3	84.0	82.3	76.5	84.1	48.2	68.4	69.6	65.8	28.3	64.0
LwF-MC [8]	99.8	65.5	0.4	63.1	10.7	39.6	53.1	35.3	78.4	0.5	66.5	1.5	77.8	72.0	34.0	13.1	14.4	65.9	0.7	25.9	1.0	38.5	1.0	36.7
ILT [5]	93.5	88.2	59.5	94.3	77.1	83.2	72.0	81.5	96.2	38.7	93.5	55.9	93.8	94.2	84.9	85.7	79.0	91.3	47.1	77.0	63.4	66.9	23.4	64.8
CIL [3]	91.9	77.6	68.6	90.8	66.0	67.6	46.0	67.9	97.3	31.3	95.8	48.6	95.4	94.6	78.9	87.7	82.1	86.4	48.2	68.2	82.1	62.6	18.1	60.5
MiB [1]	89.8	95.0	91.6	97.7	83.9	93.0	93.7	91.2	96.9	52.3	94.2	60.8	96.8	96.2	95.5	88.0	81.9	88.5	56.7	83.6	73.8	67.1	26.1	65.1
SDR (ours)	95.0	90.1	66.5	95.1	67.9	87.7	88.0	83.0	96.4	44.9	93.0	61.3	95.9	95.3	82.7	86.8	81.8	92.9	53.3	72.9	57.9	69.9	37.3	68.4
SDR+MiB	93.1	96.0	86.9	97.3	85.5	91.5	92.1	90.5	96.7	48.8	92.4	58.6	95.7	94.8	91.3	88.9	78.9	90.3	56.1	84.4	69.5	70.8	31.4	68.9
offline	96.1	96.6	85.4	94.4	87.2	92.2	94.7	93.5	96.9	50.2	95.4	56.5	95.8	91.8	94.7	90.8	80.8	92.1	54.8	89.5	83.5	75.5	73.5	75.4

Table 4. Per-class IoU of compared methods in disjoint experimental protocol on scenario 15-5 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
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
LwF [4]	83.4	59.1	21.7	16.7	36.8	47.0	18.7	62.5	52.3	6.6	4.8	37.7	35.9	44.9	55.5	51.6	22.6	27.8	25.3	39.6	51.1	39.7	33.3	38.2
LwF-MC [8]	85.4	54.2	16.9	59.7	29.7	46.0	34.4	65.9	38.1	5.2	35.9	7.5	62.4	44.3	48.7	29.1	11.4	37.3	8.9	42.1	27.1	41.5	25.4	37.6
ILT [5]	81.7	47.6	18.4	1.6	29.7	19.4	3.8	52.5	56.7	0.5	4.6	20.7	43.1	35.4	33.6	54.8	22.7	22.4	15.9	30.1	34.3	31.5	25.1	30.0
CIL [3]	81.0	45.4	28.8	30.4	31.1	54.5	9.4	67.8	52.1	10.5	9.2	47.9	53.0	35.3	66.3	58.4	23.9	33.3	25.2	39.1	53.9	42.6	35.1	40.8
MiB [1]	78.4	58.3	30.8	52.5	35.5	60.5	60.2	74.8	38.2	14.0	21.6	41.8	42.9	34.8	67.4	48.8	23.2	31.0	24.4	46.3	45.8	47.5	34.1	44.3
SDR (ours)	88.7	82.9	40.5	82.4	62.8	69.2	83.8	88.2	91.6	28.9	71.1	54.2	86.8	80.3	79.7	84.4	39.4	51.4	23.7	63.3	58.7	73.5	47.3	67.2
SDR + MiB	89.4	87.1	39.9	84.8	67.3	75.2	85.1	88.2	91.3	29.9	67.8	54.4	86.1	81.8	80.5	85.0	33.8	43.6	24.7	61.7	56.6	74.6	44.1	67.3
offline	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

Table 5. Per-class pixel accuracy of compared methods in disjoint experimental protocol on scenario 15-5 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
FT	95.3	27.5	0.0	1.6	15.4	11.5	0.0	4.1	0.5	0.0	0.0	0.0	0.0	0.2	0.2	0.0	72.0	90.0	77.2	89.7	80.7	8.4	33.5	14.4
LwF [4]	91.9	79.4	35.4	16.9	50.9	49.0	19.4	71.0	78.8	8.0	5.2	39.7	36.3	78.5	59.2	53.3	67.1	91.2	74.2	81.6	76.5	39.7	33.3	38.2
LwF-MC [8]	96.6	80.7	30.3	68.5	62.0	60.4	37.7	79.7	62.5	10.8	46.2	9.2	73.2	84.4	64.8	31.7	11.4	39.7	9.1	60.1	27.1	41.5	25.4	37.6
ILT [5]	94.6	61.4	26.4	1.6	30.8	19.5	4.0	57.7	71.7	0.5	5.1	20.9	45.9	43.7	34.6	56.7	42.5	86.0	38.8	71.0	44.9	31.5	25.1	30.0
CIL [3]	85.0	80.5	56.3	31.6	57.2	59.5	10.0	81.9	87.6	16.6	12.3	53.9	58.1	86.1	74.1	61.5	84.4	95.7	88.8	93.5	87.1	42.6	35.1	40.8
MiB [1]	80.7	92.6	64.8	64.5	74.0	68.3	65.0	84.3	93.7	23.6	36.2	50.9	49.8	91.2	85.7	52.0	73.9	86.6	87.6	89.9	83.7	47.5	34.1	44.3
SDR (ours)	91.2	95.1	82.1	96.5	80.1	86.3	93.3	92.2	97.0	51.8	93.0	64.3	96.0	91.0	92.0	91.1	68.9	64.1	69.6	74.0	82.9	73.5	47.3	67.2
SDR + MiB	91.7	94.7	80.1	93.4	79.1	88.7	90.6	91.4	96.3	51.0	82.4	64.6	94.9	90.2	91.7	91.8	68.6	67.8	70.3	79.7	81.3	74.6	44.1	67.3
offline	96.1	96.6	85.4	94.4	87.2	92.2	94.7	93.5	96.9	50.2	95.4	56.5	95.8	91.8	94.7	90.8	80.8	92.1	54.8	89.5	83.5	77.5	68.5	75.4

Table 6. Per-class IoU of compared methods in disjoint experimental protocol on scenario 15-1 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
FT	70.4	5.5	0.0	5.9	5.2	0.5	0.2	1.6	0.4	0.0	3.3	0.0	0.0	0.2	0.0	0.0	0.1	0.0	0.0	9.4	14.8	5.8	4.9	5.6
LwF [4]	77.1	12.0	6.9	52.6	14.3	23.1	18.4	27.3	56.3	20.5	48.9	8.3	17.8	12.6	15.6	8.3	0.0	17.0	21.0	18.6	19.1	26.2	15.1	23.6
LwF-MC [8]	69.5	0.1	0.0	8.0	0.1	7.2	0.0	0.1	8.1	0.0	6.6	0.0	8.0	1.7	0.3	0.1	0.0	0.0	0.0	2.4	8.1	6.9	2.1	5.7
ILT [5]	69.4	0.0	2.1	0.0	0.0	0.1	0.0	4.0	0.0	0.0	0.0	0.0	1.4	0.0	0.0	19.2	0.0	0.0	0.0	1.4	4.6	6.7	1.2	5.4
CIL [3]	78.4	2.4	23.6	47.9	4.6	32.9	0.3	29.9	45.4	15.4	30.3	2.4	54.5	13.0	8.7	59.7	15.2	17.5	12.1	20.9	19.2	33.3	15.9	29.1
MiB [1]	70.6	56.2	24.8	41.7	45.8	34.9	44.9	52.8	64.1	17.8	40.4	28.2	16.1	30.3	55.3	0.1	5.9	8.2	16.5	27.2	17.3	39.0	15.0	33.3
SDR (ours)	86.2	47.1	34.2	69.1	37.9	61.3	67.2	72.5	81.1	17.9	51.3	40.8	72.9	67.6	68.5	70.8	8.3	4.8	2.7	24.5	24.2	59.2	12.9	48.1
SDR+MiB	86.9	32.0	29.8	76.0	42.8	60.7	67.4	64.7	85.8	19.2	50.3	39.4	75.1	73.0	69.3	78.2	3.4	2.7	11.5	34.0	20.1	59.4	14.3	48.7
offline	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

Table 7. Per-class pixel accuracy of compared methods in disjoint experimental protocol on scenario 15-1 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
FT	98.5	5.6	0.0	5.9	5.4	0.5	0.2	1.6	0.4	0.0	3.4	0.0	0.0	0.2	0.0	0.0	0.1	0.0	0.0	9.8	80.1	5.8	4.9	5.6
LwF [4]	95.1	12.0	47.8	54.0	15.0	23.2	18.4	27.4	57.3	35.1	62.8	8.3	18.0	12.7	15.7	8.3	0.0	22.0	35.8	47.9	70.7	26.2	15.1	23.6
LwF-MC [8]	99.9	0.1	0.0	8.0	0.1	7.4	0.0	0.1	8.1	0.0	6.6	0.0	8.0	1.7	0.3	0.1	0.0	0.0	0.0	2.9	8.2	6.9	2.1	5.7
ILT [5]	20.9	0.0	73.2	0.0	0.0	0.0	2.3	89.3	19.0	16.3	14.8	1.4	48.3	0.0	23.2	0.4	4.6	0.0	0.0	1.8	4.9	6.7	1.2	5.4
CIL [3]	90.1	16.8	40.0	48.4	15.3	32.7	9.0	28.2	60.1	17.1	75.0	20.4	53.8	28.7	13.5	60.0	31.0	11.8	49.7	50.1	87.0	33.3	15.9	29.1
MiB [1]	72.7	61.7	58.6	60.7	52.3	69.4	45.8	59.2	88.3	30.2	62.3	53.9	68.6	60.7	70.9	0.1	7.0	84.3	28.8	84.9	65.6	39.0	15.0	33.3
SDR (ours)	92.7	47.6	72.3	91.9	44.5	69.2	76.5	74.7	89.3	60.9	92.8	53.1	94.9	75.5	88.3	73.8	11.5	5.1	3.0	35.7	76.6	59.2	12.9	48.1
SDR+MiB	92.7	33.2	45.0	84.7	47.0	67.6	72.1	65.2	96.6	59.1	95.7	45.1	85.3	80.5	83.5	84.2	4.4	2.8	17.2	57.1	76.6	59.4	14.3	48.7
offline	96.1	96.6	85.4	94.4	87.2	92.2	94.7	93.5	96.9	50.2	95.4	56.5	95.8	91.8	94.7	90.8	80.8	92.1	54.8	89.5	83.5	77.5	68.5	75.4