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

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

Abstract

Deep neural networks suffer from the major limitation of catastrophic forgetting old tasks when learning new ones. In this paper we focus on class incremental continual learning in semantic segmentation, where new categories are made available over time while previous training data is not retained. The proposed continual learning scheme shapes the latent space to reduce forgetting whilst improving the recognition of novel classes. Our framework is driven by three novel components which we also combine on top of existing techniques effortlessly. First, prototypes matching enforces latent space consistency on old classes, constraining the encoder to produce similar latent representation for previously seen classes in the subsequent steps. Second, features sparsification allows to make room in the latent space to accommodate novel classes. Finally, contrastive learning is employed to cluster features according to their semantics while tearing apart those of different classes. Extensive evaluation on the Pascal VOC2012 and ADE20K datasets demonstrates the effectiveness of our approach, significantly outperforming state-of-the-art methods.

1. Introduction

Semantic segmentation is a challenging computer vision problem with many real-world applications ranging from robot sensing, to autonomous driving, video surveillance, virtual reality, and many others. For most applications, continuously improving the set of classes to be distinguished is a fundamental requirement. Current state-of-the-art semantic segmentation approaches are typically based on auto-encoder structures and on fully convolutional models [38] that are trained in a single-shot requiring all the dataset to be available at once. Indeed, existing architectures are not designed to incrementally update their inner classification model to accommodate new categories. This issue is well-known for deep neural networks and it is called *catastrophic forgetting* [41, 18, 20], as deep architectures fail to update their parameters for learning new categories while preserving good performance on the old ones.

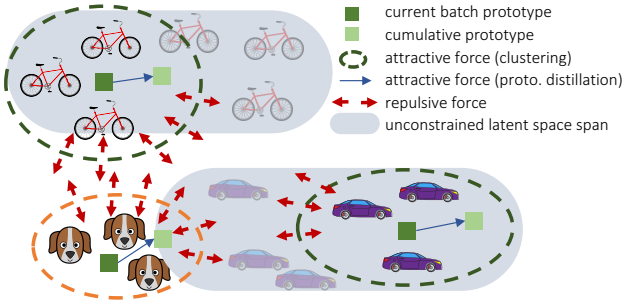

Figure 1. Our continual learning scheme is driven by 3 main components: latent contrastive learning, prototypes matching and features sparsity. Latent representations of old classes are preserved via prototypes matching and clustering, whilst also making room for accommodating new classes via sparsity and repulsive force of contrastive learning. The decoder preserves previous knowledge via output-level distillation. In the figure, bike and cars represent old classes and leave more space to new classes (the dog) thanks to the novel constraints (green dotted ovals versus gray-filled ovals).

Continual learning has been widely studied in image classification [32, 36] and object detection [56, 34], while has been tackled only recently in the semantic segmentation field [42, 58, 4, 33]. In this paper, we investigate class-incremental continual learning in semantic segmentation. Differently from the majority of previous approaches both in image classification [36, 51, 3] and semantic segmentation [58, 42, 4, 33], we do not mainly or solely rely on output-level knowledge distillation. In this work, we focus on latent space organization which has been only marginally investigated in the current literature, and we empirically prove it to be complementary to other existing techniques. The main idea is depicted in Fig. 1, where some of the latent space constraints are introduced. First, a prototype matching is devised to enforce features extraction consistency on old classes between the cumulative prototype computed using all previous samples and the current prototype (*i.e.*, the prototype computed on the current batch only). In other

words, we force the encoder to produce similar latent representations for previously seen classes in the new steps. Second, a features sparsification constraint makes room in the latent space to accommodate novel classes. To further regularize the latent space, we introduce an attraction-repulsion rule similar in spirit to the recent advancements in contrastive learning. Finally, to enforce the decoder to preserve discriminability on previous categories during classification, we employ a targeted output-level distillation.

Although continual semantic segmentation has only been faced recently, it already comes with different experimental protocols depending on how the incremental data are considered (see Section 3.1): namely, *sequential* (new images are labeled with both new and old classes), *disjoint* (new images are labeled with only new classes, old classes are assigned to the background) and *overlapped* (new images are labeled with only new classes, images are repeated across training steps with different semantic maps associated to them). In this paper we devise a common framework which allows to tackle all these scenarios and can be applied in combination with previous techniques, which has never been attempted before. We evaluate on standard semantic segmentation datasets, like Pascal VOC2012 [16] and ADE20K [76], in many scenarios.

Summing up, the main contributions of this work are: 1) We investigate class-incremental learning in semantic segmentation, providing a common framework for different experimental protocols. 2) We explore the latent space organization and we propose complementary techniques with respect to the existing ones. 3) We propose novel knowledge preservation techniques based on prototypes matching, contrastive learning and features sparsity. 4) We benchmark our approach on standard semantic segmentation datasets outperforming state-of-the-art continual learning methods.

2. Related Work

Continual Learning. Deep learning models are prone to *catastrophic forgetting* [20, 30, 48], *i.e*., training a model with new information interferes with previously learned knowledge and typically greatly degrades performance. This phenomenon has been widely studied in image classification task and most of the current techniques fall into the following categories [10, 48]: regularization approaches [5, 32, 73, 13, 36], dynamic architectures [69, 64, 35], parameter isolation [17, 53, 40] and replay-based methods [66, 46, 55, 26]. Regularization-based approaches are by far the most widely employed and mainly come in two flavours, *i.e*., penalty computing and knowledge distillation [25]. Penalty computing approaches [73, 32, 32] protect important weights inside the models to prevent forgetting. Knowledge distillation [52, 66, 36, 13] relies on a teacher (old) model transferring or remembering knowledge related to previous tasks to a student model which is trained to learn also additional tasks. Parameter isolation approaches [40, 39] reserve a subset of weights for a specific task to avoid degradation. Dynamic architectures [64, 35] grow new branches for new tasks. Replay-based models exploit stored [3, 26, 51] or generated [66, 46, 55] examples during the learning process of new tasks.

Continual Semantic Segmentation. Nowadays, deep learning architectures have achieved outstanding results in semantic segmentation [19, 21]. Current approaches are based on fully convolutional models [38, 7, 6, 75, 72] and exploit various techniques to cope with multi-scale and spatial dependency. All these approaches, however, require training data and segmentation maps to be available at once (*i.e*., *joint* setting) and they experience catastrophic forgetting if new tasks (*e.g*., new classes to learn) are made available sequentially [42]. Hence, it emerged the need for continual approaches specifically targeted to solve the semantic segmentation task [47, 58, 42, 43, 33, 4]. Earlier works focus on the continual semantic segmentation problem in specific scenarios, *e.g*., in medical imaging [47] or remote sensing [58], extending standard image-level classification methods. More recently, standard semantic segmentation datasets and targeted methods have been proposed. In [42, 43] an exploration on knowledge distillation techniques is proposed to alleviate forgetting: the authors designed output-level and features-level distillation losses coupled with freezing the encoder's weights. Klingner *et al*. [33] extend previous work not requiring old labels during the incremental steps and proposing class importance weighting to emphasize gradients on difficult classes. Cermelli *et al*. [4] study the distribution shift of the background class when it incorporates previous and/or future classes (*disjoint* and *overlapped* protocols, respectively). Background shift is addressed via unbiased versions of cross entropy and output-level knowledge distillation losses together with an unbiased weight initialization rule for the classifier. Nevertheless, previous works neglect accurate investigation of the latent space in continual learning.

Latent Space Organization. The analysis of the latent space organization is becoming crucial towards understanding and improvement of classification models [68, 49]. Recently, some attention has been devoted to latent regularization in continual image classification [1, 2, 27].
Besides this, one of the emerging paradigms is constrastive learning applied to visual representations. Dating back to [22], these approaches learn representations by contrasting positive against negative pairs and have been recently re-discovered for deep learning. Many works use a memory bank to store the instance class representation vector [67, 77, 59, 23, 44, 9], while some others explore the usage of in-batch negative samples instead [14, 71, 28, 31]. The contrastive learning objective proposed in this work moves from opposition of positive and negative pairs and

also recalls features clustering (if features belong to the same class) and separation (if features belong to different classes), which has been recently applied to adapt semantic segmentation models across domains [29, 37, 61].
Prototypes-based regularizing terms gained a great interest and, in particular, have been largely used in the literature of few-shot learning [15, 63, 60], to learn prototypical representations of each category, and domain adaptation, to enforce orthogonality [50, 65] or centroid matching [70, 12].
Finally, to minimize the interference among features we drive them to be channel-wise sparse. Only limited attention has been given on sparsity for deep learning architectures [2]; however, some prior techniques exist for domain adaptation on linear models exploiting sparse codes on a shared dictionary between the domains [54, 74].

Our work is the first combining together contrastive learning, sparsity and prototypes matching to regularize latent space for segmenting new categories over time.

3. Problem Definition and Setups

Before presenting the proposed strategies, we first introduce the semantic segmentation task, which assigns a class to each pixel in an image. We denote the input image space with $\mathcal{X} \in \mathbb{R}^{H \times W \times 3}$ with spatial dimensions H and W, the set of classes (or categories) with $\mathcal{C} = \{c_i\}_{i=0}^{C-1}$ and the output space with $\mathcal{Y} \in \mathcal{C}^{H \times W}$ (*i.e.*, the segmentation map). Given a training set $\mathcal{T} = \{(\mathbf{x}_n, \mathbf{y}_n)\}_{n=1}^{N}$, where $(\mathbf{x}_n, \mathbf{y}_n) \in \mathcal{X} \times \mathcal{Y}$, we aim at finding a map M from the input space to a pixel-wise class probability vector $M : \mathcal{X} \mapsto \mathbb{R}^{H \times W \times C}$. Then, the output segmentation mask is computed as $\hat{\mathbf{y}}_n = \arg\max_{c \in \mathcal{C}} M(\mathbf{x}_n)[h, w, c]$, where $h = 1, .., H$, $w = 1, ..., W$ and $M(\mathbf{x}_n)[h, w, c]$ is the probability for class c in pixel (h, w). Nowadays, M is typically some auto-encoder model made by an encoder E and a decoder D (*i.e.*, $M = E \circ D$). We call $\mathbf{F}_n = E(\mathbf{x}_n)$ the feature map of $\mathbf{x}_n$, and $\mathbf{y}_n^*$ the downsampled segmentation map matching the spatial dimensions of $\mathbf{F}_n$.

In the standard supervised setting it is assumed that the training set $\mathcal{T}$ is available at once and the model is learned in one shot. In the continual learning scenario, instead, training is achieved over multiple iterations each carrying a novel category to learn and a subset of the training data. More formally, at each learning step k the previous label set $\mathcal{C}_{k-1}$ is expanded with a set of novel classes $\mathcal{S}_k$ forming a new label set $\mathcal{C}_k = \mathcal{C}_{k-1} \cup \mathcal{S}_k$. Additionally, a new training subset $\mathcal{T}_k \subset \mathcal{X} \times \mathcal{C}_k$ is made available and used to update the previous model into a new model M_k. Step $k = 0$ consists of a standard supervised training performed with only a subset of training data and classes. As in the standard incremental class learning scenario, we assume the different sets of new classes to be disjoint with the exception of the peculiar background class c_0, *i.e.*, $\mathcal{S}_i \cap \mathcal{S}_j = \{c_0\}$.

3.1. Experimental Protocols

Despite being quite a recent field, continual learning in semantic segmentation already comes in different flavors.
Sequential: this setup has been proposed in [42, 43]. Each learning step contains a unique set of images, whose pixels belong to classes seen either in the current or in the previous learning steps. At each step, labels for pixels of both old and novel classes are present.
Disjoint: this setup has been proposed in [4]. At each learning step, the unique set of images is identical to the *sequential* setup. The difference with respect to the *sequential* setup lies in the set of labels. At each step, only labels for pixels of novel classes are present, while the old ones are labeled as background in the ground truth.
Overlapped: this setup moves from the work of [56] for object detection and has been adapted to semantic segmentation in [4]. Each training step contains all the images that have at least one pixel of a novel class, with only the novel classes annotated while the rest is set to background. Differently from the other settings, here images may contain pixels of classes that will be learned in future learning steps, but they are labeled as background in the current step.

4. Method

In this section, we provide a detailed description of the core modules of the proposed method. Our approach leverages a contrastive learning objective applied over the feature representations, with novel prototypes matching and sparsity constraints. Specifically, features repulsion and attraction based on the semantic classes are enforced by grouping together features of the same class, while simultaneously pushing away those of different categories. We further regularize the distribution of latent representations by the joint application of prototypes matching and sparsity. While prototypes matching seeks for an invariant representation of the features extracted for the old classes, the sparsity objective encourages a lower volume of active feature channels from latent representations (*i.e.*, it concentrates the energy of features on few dimensions) to free up space for new classes.

An overall scheme of our approach is shown in Fig. 2: the training objective is given by the combination of a cross-entropy loss ($\mathcal{L}_{ce}$) with the proposed modules. $\mathcal{L}_{ce}$ is the usual cross-entropy loss for all the classes except for the background. The ground truth of the background, indeed, is not directly compared with its probabilities, but with the probability of having either an old class or the background in the current model [4]. Formally, at step k the background probabilities $M(\mathbf{x}_n)[h, w, c_0]$ are replaced by $\sum_{c \in \mathcal{C}_{k-1}} M(\mathbf{x}_n)[h, w, c]$. The rationale behind this is that the background class could incorporate statistics of previous classes in both the disjoint and overlapped protocols.

The other components are a prototypes matching target

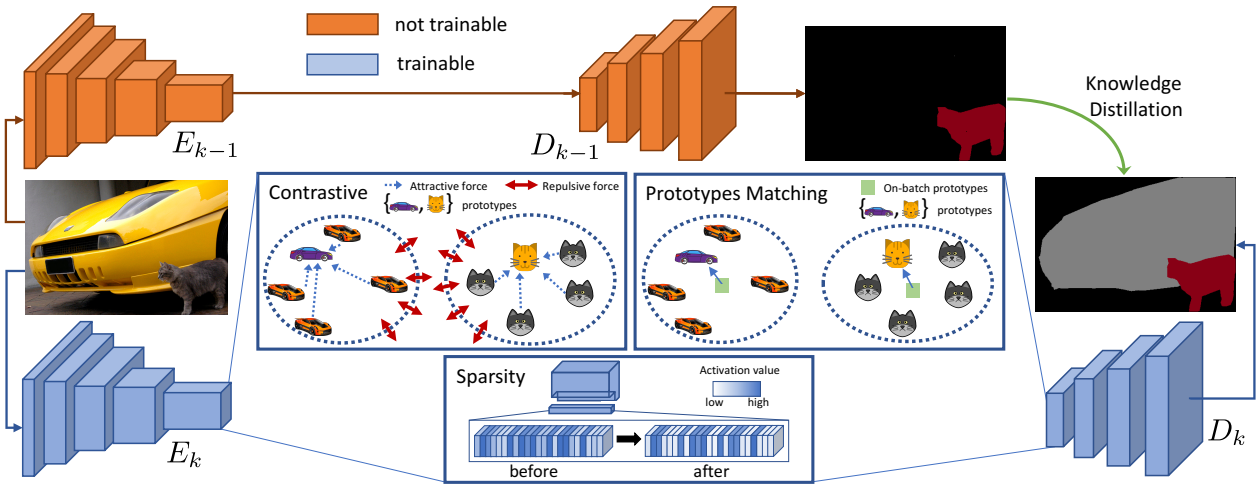

Figure 2. Overview of the proposed approach, with an old class (*cat*) and a new class (*car*). Latent representations of old classes are preserved over time via prototypes matching and clustering, whilst also making room for accommodating new classes via sparsity and repulsive force in contrastive learning. The decoder is constrained to act as in previous steps on previous classes via output-level distillation.

($\mathcal{L}_{pm}$), a contrastive learning objective ($\mathcal{L}_{cl}$) and a sparsity constraint ($\mathcal{L}_{sp}$), which will be detailed in the following sections. The training objective is then computed as:

$$\mathcal{L}'_{tot} = \mathcal{L}_{ce} + \lambda_{pm} \cdot \mathcal{L}_{pm} + \lambda_{cl} \cdot \mathcal{L}_{cl} + \lambda_{sp} \cdot \mathcal{L}_{sp} \quad (1)$$

where the λ parameters balance the multiple losses and have been tuned using a validation set (see Section 5). Our aim is to seek for disentangled latent representations characterized by semantic-driven regularization and to show that this approach can achieve comparable or superior results with respect to standard regularization methods (*e.g.*, output-level knowledge distillation). We further integrate the proposed framework with an output-level knowledge distillation objective [43] and we show that its effect is highly not overlapping, achieving increased accuracy. The training objective comprising an unbiased output-level distillation module is defined as:

$$\mathcal{L}_{tot} = \mathcal{L}'_{tot} + \lambda_{kd} \cdot \mathcal{L}_{kd} \quad (2)$$

4.1. Prototypes Matching

Prototypes (*i.e.*, class-centroids) are vectors that are representative of each category that appears in the dataset. During training, the features extracted by the encoder contribute in forming the latent prototypical representation of each class. To preserve the geometrical structure of the features of old classes we apply prototypes matching. Current prototypes $\hat{\mathbf{p}}_c$ (*i.e.*, computed on the current batch of images) are forced to be placed close to their representation learned from the previous steps $\mathbf{p}_c$. We use the Frobenius norm $||\cdot||_F$ as metric distance [57, 45, 63]. More formally:

$$\mathcal{L}_{pm} = \frac{1}{|\mathcal{C}_{k-1}|} ||\mathbf{p}_c - \hat{\mathbf{p}}_c||_F \quad c \in \mathcal{C}_{k-1} \quad (3)$$

The prototypes are computed in-place with a running average updated at each training step with supervision. At training step t with batch $\mathcal{B}$ of B images, the prototypes are updated for a generic class c as:

$$\mathbf{p}_c[t] = \frac{1}{Bt}\left(B(t-1)\mathbf{p}_c[t-1] + \sum_{\mathbf{x}_n \in \mathcal{B}} \frac{\sum_{\mathbf{f}_i \in \mathbf{F}_n} \mathbf{f}_i \mathbb{1}[y_i^* = c]}{|\mathbb{1}[\mathbf{y}_n^* = c]|}\right) \quad (4)$$

initialized to $\mathbf{p}_c[0] = \mathbf{0}\ \forall c$. $\mathbf{f}_i \in \mathbf{F}_n$ is a generic feature vector and y_i^* the corresponding pixel in $\mathbf{y}_n^*$, $\mathbb{1}[\mathbf{y}_n^* = c]$ indicates the pixels in $\mathbf{y}_n^*$ associated to c and $|\cdot|$ denotes cardinality.

We update the prototypes only when we have ground truth labels for that class to avoid incorporating the mutable statistics of the background class: we exclude the background from the incremental steps in the disjoint protocol (as it could contain old classes) and in the overlapped scenario (as it could contain old and future classes).

For the current batch $\mathcal{B}$ of an incremental training stage, the current (or in-batch) prototypes $\hat{\mathbf{p}}_c[t]$ are computed as:

$$\hat{\mathbf{p}}_c[t] = \frac{1}{B} \sum_{\mathbf{x}_n \in \mathcal{B}} \begin{cases} \frac{\sum_{\mathbf{f}_i \in \mathbf{F}_n} \mathbf{f}_i \mathbb{1}[y_i^* = c]}{|\mathbb{1}[\mathbf{y}_n^* = c]|} & \text{if sequential} \\ \frac{\sum_{\mathbf{f}_i \in \mathbf{F}_n} \mathbf{f}_i \mathbb{1}[\hat{z}_i^* = c]}{|\mathbb{1}[\hat{\mathbf{z}}_n^* = c]|} & \text{otherwise} \end{cases} \quad (5)$$

where $\hat{\mathbf{z}}_n^*$ (with pixels $\hat{z}_i^*$) is a pseudo-labeled segmentation map computed from the ground truth data by replacing the background region with the prediction from the previous model, since in the disjoint and overlapped protocols old classes are labeled as background. The difference between (4) and (5) lies in the usage of pseudo-labels: we use them in (5) to compute prototypes for old classes in the current batch since we may not have any label for them, but we avoid to use them in (4), since there is no need to update prototypes computed using the ground truth at previous steps with data from less reliable pseudo-labels.

4.2. Contrastive Learning

The second component is similar to recent contrastive learning [9, 59] and clustering [29, 37] approaches to constraint the latent space organization. The underlying idea is to structure the latent space in order to have features of the same category clustered near their prototype and at the same time to force prototypes to be far one from the other. We argue that this organization helps also in continual learning to mitigate forgetting and to facilitate the addition of novel classes, as features are clustered and there is more separation between the clusters. In formal terms, the constraint is defined by a loss $\mathcal{L}_{cl}$ made of an attractive term $\mathcal{L}^a_{cl}$ and a repulsive term $\mathcal{L}^r_{cl}$, as follows:

$$\mathcal{L}^a_{cl} = \frac{1}{|c_j \in \mathbf{y}^*_n|} \sum_{c_j \in \mathbf{y}^*_n} \sum_{\mathbf{f}_i \in \mathbf{F_n}} \|(\mathbf{f}_i - \mathbf{p}_{c_j}) \mathbb{1}[y^*_i = c_j]\|_F \quad (6)$$

$$\mathcal{L}^r_{cl} = \frac{1}{|c_j \in \mathbf{y}^*_n|} \sum_{c_j \in \mathbf{y}^*_n} \sum_{\substack{c_k \in \mathbf{y}^*_n \\ c_k \neq c_j}} \frac{1}{\|\hat{\mathbf{p}}_{c_j} - \hat{\mathbf{p}}_{c_k}\|_F} \quad (7)$$

The objective is composed of two terms: $\mathcal{L}^a_{cl}$ measures how close features are from their respective centroids and $\mathcal{L}^r_{cl}$ how spaced out prototypes corresponding to different semantic classes are. Hence, the effect provided by the loss minimization is twofold: firstly, feature vectors from the same class are tightened around class feature centroids; secondly, features from separate classes are subject to a repulsive force applied to feature centroids, moving them apart.

4.3. Features Sparsity

To enforce the regularizing effect brought by contrastive learning, we introduce a further feature-wise objective on the latent space. We propose a sparsity loss to decrease the number of active feature channels of latent vectors. First, to give the same importance to all classes, we normalize each feature vector with respect to the maximum value any of the feature channels for that particular class assumes, *i.e.*:

$$\bar{\mathbf{f}}_i = \frac{\mathbf{f}_i}{\max_{\substack{g_{j,l} \in \mathbf{g}_j \\ y^*_j = y^*_i}} g_{j,l}} \qquad \mathbf{f}_i, \mathbf{g}_j \in \mathbf{F}_n \quad (8)$$

We design the sparsity constraint as the ratio between a stretching function (we used the sum of exponentials) and a linear function (*i.e.*, the sum) applied over each feature vector, which is minimized when the energy is concentrated in a few channels (since the normalized features assume values ≤ 1). The sparsity constraint is thus defined as:

$$\mathcal{L}_{sp} = \frac{1}{|\mathbf{f}_i \in \mathbf{F}_n|} \sum_{\mathbf{f}_i \in \mathbf{F}_n} \frac{\sum_j \exp\left(\bar{f}_{i,j}\right)}{\sum_j \bar{f}_{i,j}} \quad (9)$$

While the contrastive learning objective forces features to lie within tight semantically-consistent well-distanced clusters, the sparsity constraint aims at narrowing the set of active channels with the aim of letting room for the representation of upcoming classes. In other words, by constraining features of the same classes to be tightly clustered and to be spaced apart from features of other classes and sparse, we can preserve geometrical space (few active channels) and expressiveness (division in well-separated clusters) for the latent representation of future classes. Empirically, we found entropy-based minimization methods in the latent space [62] to be less reliable for our task. In the *Supplementary Material* we show how to handle degenerate cases of (9) and an ablation on other sparsifying strategies.

4.4. Output-Level Knowledge Distillation

The last component of our work is an output-level knowledge distillation which we show to be complementary to the previously introduced strategies. Indeed, we add knowledge distillation on top of all the other components to transfer knowledge from the old model's classifier to the current one. While previous constraints regularize the latent space achieving simultaneously an invariant features extraction with respect to previous steps and an easier addition of novel categories, output-level knowledge distillation directly acts on the classifier, to preserve its discriminative ability regarding old classes. In particular, we start from the preliminary considerations of [42, 43] and we employ the unbiased distillation proposed in [4] as natural extension to the case in which the background may contain other categories. In this case we avoid to re-normalize the probabilities from the previous step and, instead, we compare the background probability from the previous step with the probability of having either a new class or the background (this accounts for the fact that the background in the previous steps may include samples of the new classes, see [4]).

5. Training Procedure

To train and benchmark our approach we resort to two publicly available datasets following [56, 42, 43, 4]. The **Pascal VOC 2012** [16] contains 10582 images in the training split and 1449 in the validation split (that we used for testing, as done by all competing works being the test set not publicly available). Each pixel of each image is assigned to one semantic label chosen among 21 different classes (20 plus the background). The **ADE20K** [76] is a large-scale dataset of 22210 images, 2000 of which form the validation split. The typical benchmark defined in [76] includes 150 classes, representing both stuff (*e.g.*, sky, building) and object classes (*e.g.*, bottle, chair), differently from VOC 2012.

The proposed strategy is agnostic to the backbone architecture. For the experimental evaluation of all the compared methods we use a standard Deeplab-v3+ [8] architecture with ResNet-101 [24] as backbone (differently from [4] for wider reproducibility) with output stride of 16. The back-

bone has been initialized using a pre-trained model on ImageNet [11] (see the *Supplementary Material* for a detailed discussion of the impact of different pre-training strategies). We optimize the network weights following [7] with SGD and with same learning rate policy, momentum and weight decay. The first learning step involves an initial learning rate of 10^{-2}, which is decreased to 10^{-3} for the following steps as done in [56, 4]. The learning rate is decreased with a polynomial decay rule with power 0.9. In each learning step we train the models with a batch size of 8 for 30 epochs for Pascal VOC 2012 and a batch size of 4 for 60 epochs for ADE20K. Following [7], we crop the images to 512×512 during both training and validation and we apply the same data augmentation (*i.e.*, random scaling the input images of a factor from 0.5 to 2.0 and random left-right flipping during training). In order to set the hyper-parameters of each method, we follow the same continual learning protocol of [10, 4], i.e, we used 20% of the training set as validation and we report the results on the original validation set of the datasets. We use Pytorch to develop and train all the models on a NVIDIA 2080 Ti GPU. The code is available at: https://lttm.dei.unipd.it/paper_data/SDR/.

6. Experimental Results

We evaluate the performance of our method (denoted in the tables with **SDR**, *i.e.*, Sparse and Disentangled Representations) against some state-of-the-art continual learning frameworks. We report as a lower limit the performance of the naïve fine-tuning approach (FT), which consists in training the model on the newly available training data with no additional provisions, while the upper limit is given by the offline single-shot training (offline) on the whole dataset $\mathcal{T}$ and on all the classes at once. Then, we compare with 3 recent continual semantic segmentation schemes: ILT [42], which combines latent and output level knowledge distillation, CIL [33], which adds class importance weighting to output-level knowledge distillation, and MiB [4], which deals with the background distribution shift and proposes an unbiased weight initialization rule. We also report the results on LwF [36] (together with its single-headed version LwF-MC [51]), that according to [4] is the best performing continual image classification algorithm when adapted to semantic segmentation. For a fair comparison, all the methods have been re-trained with a standard Deeplab-v3+ [8] architecture with ResNet-101 [24] as backbone.

6.1. Pascal VOC2012

Following previous works [56, 42, 43, 4], we design three main experiments adding one class (19-1), five classes at once (15-5) and five classes sequentially (15-1) added in alphabetical order. In Table 1 we report comprehensive results on the three experimental protocols defined in Section 3.1. Results are averaged for mIoU of classes in the base step (*old*), for classes in the incremental steps (*new*) and for *all* classes, and are reported at the end of all the incremental steps. For [4] we also report the original results in their paper (denoted with MiB†), that uses a different backbone (thus different pre-trained model) and batch size.

We can appreciate forgetting of previous classes and intransigence in learning new ones even when adding as little as one class (the *tv/monitor* class is added) in the scenario 19-1. FT always leads to the worst mIoU in terms of *old*, *new* and *all* classes. Incremental methods designed for semantic segmentation allow for a stable improvement across the experimental protocols, in particular MiB, that is specifically targeted to solve the disjoint and the overlapped scenarios, while CIL and ILT encounter difficulties in the overlapped scenario. Also LwF allows for a good improvement while its single-headed version has lower performance in this scenario. Our method (SDR) significantly outperforms all the competitors in the disjoint and overlapped scenarios (with a gap of more than 3% against the best competing approach in the disjoint setup), while in the sequential setup the gap is smaller. Further adding on top of our method the MiB framework (*i.e.*, unbiased cross entropy, knowledge distillation and classifier initialization), which we regard as the current state-of-the-art approach for class incremental semantic segmentation, the results increase on all the scenarios, showing that proposed techniques are complementary with respect to previous schemes.

When moving to the addition of 5 classes at once (*i.e.*, *potted plant*, *sheep*, *sofa*, *train*, *tv/monitor*) we immediately notice an overall increased drop of performance of all compared methods, especially in disjoint and overlapped protocols, due to the increased domain shift occurring when adding more classes at once with very variable content. In this and in the following scenario, indeed, we are adding to the model classes belonging to different macroscopic groups, according to [16], which are responsible for a variegate distribution: three indoor classes (*potted plant*, *sofa* and *tv/monitor*), one animal class (*sheep*) and one vehicle class (*train*). All compared methods obtain a relevant improvement with respect to FT but are always surpassed by SDR, which in particular outrun the best competing method (MiB) by more than 20% in the disjoint scenario.

In the final scenario we add the last 5 classes sequentially in 5 consecutive learning steps. This approach leads to the largest accuracy drop being the model exposed to a reiterated addition of single classes, which are also coming from different semantic contexts. In the sequential scenario LwF and MiB (which is designed for background shift) show poor final accuracy. ILT and CIL, instead, show results comparable to our approach. In the disjoint and in the overlapped scenarios all the methods heavily suffer from the semantic shift undergone by the background class: LwF (both versions) and ILT have poor performance in these scenar-

Table 1. mIoU on multiple incremental scenarios and protocols on VOC2012. Best in **bold**, runner-up underlined. †: results from [4].

	19-1									15-5									15-1								
	sequential			disjoint			overlapped			sequential			disjoint			overlapped			sequential			disjoint			overlapped		
Method	old	new	all	old	new	all	old	new	all	old	new	all	old	new	all	old	new	all	old	new	all	old	new	all	old	new	all
FT	63.4	21.2	61.4	35.2	13.2	34.2	34.7	14.9	33.8	62.0	38.1	56.3	8.4	33.5	14.4	12.5	36.9	18.3	49.0	17.8	41.6	5.8	4.9	5.6	4.9	3.2	4.5
LwF [36]	67.2	26.4	65.3	65.8	28.3	64.0	62.6	23.4	60.8	68.0	43.0	62.1	39.7	33.3	38.2	67.0	41.8	61.0	33.7	13.7	29.0	26.2	15.1	23.6	24.0	15.0	21.9
LwF-MC [51]	49.2	0.9	46.9	38.5	1.0	36.7	37.1	2.3	35.4	70.6	19.5	58.4	41.5	25.4	37.6	59.8	22.6	51.0	12.1	1.9	9.7	6.9	2.1	5.7	6.9	2.3	5.8
ILT [42]	64.3	22.7	62.3	66.9	23.4	64.8	50.2	29.2	49.2	71.3	**47.8**	65.7	31.5	25.1	30.0	69.0	46.4	63.6	49.2	**30.3**	**48.3**	6.7	1.2	5.4	5.7	1.0	4.6
CIL [33]	64.1	22.8	62.1	62.6	18.1	60.5	35.1	13.8	34.0	63.8	39.8	58.1	42.6	35.0	40.8	14.9	37.3	20.2	52.4	22.3	45.2	33,3	**15.9**	29.1	6.3	4.5	5.9
MiB†[4]	-	-	-	69.6	25.6	67.4	70.2	22.1	67.8	-	-	-	71.8	43.3	64.7	75.5	49.4	69.0	-	-	-	46.2	12.9	37.9	35.1	13.5	29.7
MiB [4]	68.2	31.9	66.5	67.0	26.0	65.1	69.6	23.8	67.4	73.0	44.4	66.1	47.5	34.1	44.3	73.1	44.5	66.3	35.7	11.0	29.8	39.0	15.0	33.3	44.5	11.7	36.7
SDR (ours)	68.4	**35.3**	66.8	69.9	**37.3**	68.4	69.1	**32.6**	67.4	73.6	46.7	67.2	73.5	**47.3**	67.2	75.4	**52.6**	69.9	**58.5**	10.1	47.0	59.2	12.9	48.1	44.7	**21.8**	39.2
SDR + MiB	**70.6**	24.8	**68.5**	**70.8**	31.4	**68.9**	**71.3**	23.4	**69.0**	**74.6**	43.8	**67.3**	**74.6**	44.1	**67.3**	**76.3**	50.2	**70.1**	58.1	11.8	47.1	**59.4**	14.3	**48.7**	**47.3**	14.7	**39.5**
offline	75.5	73.5	75.4	75.5	73.5	75.4	75.5	73.5	75.4	77.5	68.5	75.4	77.5	68.5	75.4	77.5	68.5	75.4	77.5	68.5	75.4	77.5	68.5	75.4	77.5	68.5	75.4

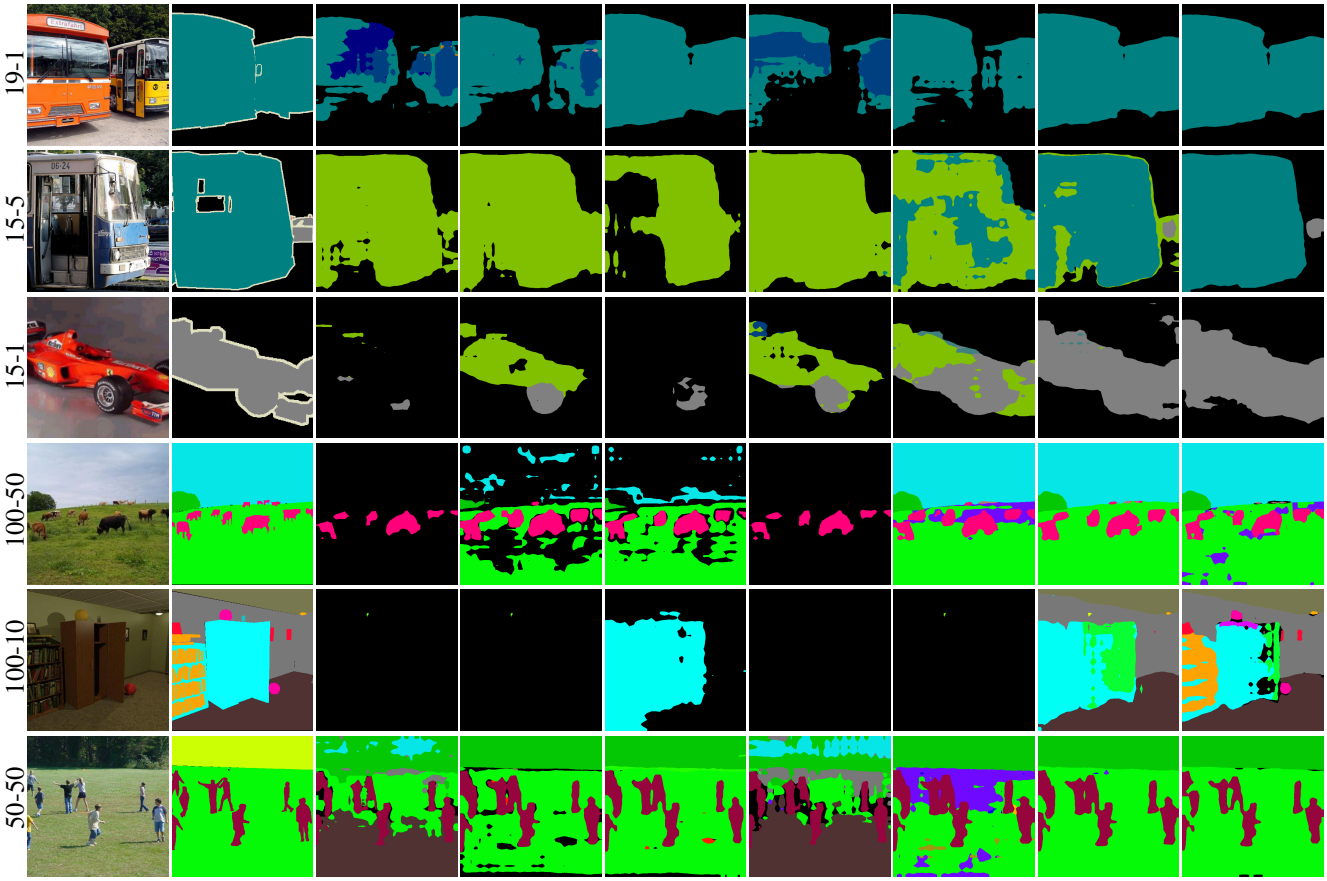

Figure 3. Qualitative results of competing approaches in different scenarios on disjoint VOC 2012 and ADE20K (*best viewed in colors*).

ios, while CIL is able to achieve some improvement only in the disjoint scenario. The best competitor is again MiB that is able to obtain a mIoU of 33% and 36.7% in the disjoint and overlapped scenarios respectively. Our approach (SDR) is able to significantly increase the final mIoU in both scenarios to 48.1% and 39.2%; it achieves a remarkable result especially in the disjoint scenario thanks to the novel features-level constraints which help the model to maintain accuracy on old classes while learning new ones.

Visual results for each scenario in the disjoint protocol are shown in the first three rows of Fig. 3, where our method is compared against all the competitors consistently obtaining better segmentation maps. For example, our method does not mislead the *bus* windows with *tv/monitor* instances in row 1 differently from several competitors (which are more biased toward predicting the novel class), and it is the only one able to distinguish the *bus* in row 2 and the *car* in row 3 from the similar-looking *train* class. Here, *train* is added in the incremental step causing catastrophic forgetting of similar classes in competing approaches.

Table 2. mIoU over multiple incremental scenarios on disjoint setup of ADE20K. Best in **bold**, runner-up underlined.

Method	100-50			100-10			50-50		
	old	new	all	old	new	all	old	new	all
FT	0.0	22.5	7.5	0.0	2.5	0.8	13.9	12.0	12.6
LwF [36]	25.0	23.9	24.6	5.4	5.6	5.5	32.2	22.9	26.0
LwF-MC [51]	8.6	0.0	5.8	0.0	0.9	0.3	2.8	0.5	1.2
ILT [42]	27.2	21.7	25.4	0.0	0.2	0.8	41.9	21.1	28.0
CIL [33]	0.0	22.5	7.5	0.0	2.0	0.6	14.0	11.9	12.6
MiB [4]	**37.6**	24.7	33.3	21.0	5.3	15.8	39.1	22.6	28.1
SDR (ours)	37.4	24.8	33.2	**28.9**	7.4	21.7	40.9	23.8	29.5
SDR+MiB	37.5	**25.5**	**33.5**	**28.9**	**11.7**	**23.2**	**42.9**	**25.4**	**31.3**
offline	43.9	27.2	38.3	43.9	27.2	38.3	50.9	32.1	38.3

Table 3. Ablation on disjoint VOC2012 15-1 in terms of mIoU.

$\mathcal{L}_{ce}$	$\mathcal{L}_{pm}$	$\mathcal{L}_{sp}$	$\mathcal{L}_{cl}$	$\mathcal{L}'_{kd}$	$\mathcal{L}_{kd}$	old	new	all
✓						5.8	4.9	5.6
✓				✓		30.0	11.0	25.4
✓	✓					18.7	9.0	16.4
✓	✓	✓				40.4	12.9	33.9
✓	✓	✓	✓			41.0	13.2	34.4
✓	✓	✓	✓	✓		50.0	**15.9**	41.9
✓	✓	✓	✓		✓	**59.2**	12.9	**48.1**

6.2. ADE20K

Following [4] we split the dataset into disjoint image sets with the only constraint that a minimum number of images (*i.e.*, 50) have labeled pixels on $\mathcal{C}_k$. Classes are ordered according to [76]. In this comparison we report the same competing methods of Section 6.1. The scenarios we consider are the addition of the last 50 classes at once (100-50), of the last 50 classes 10 at a time (100-10) and of the last 100 classes in 2 steps of 50 classes each (50-50). The results are summarized in Table 2, where we can appreciate that the proposed approach outperforms competitors in every scenario, in particular with a larger gain when multiple incremental steps are performed. When adding 50 classes at a time LwF-MC and CIL achieve low results and are outperformed by the other competitors (*i.e.*, LwF, ILT and MiB), which in turn are always consistently surpassed by our framework. In the scenario 100-10, instead, all competing approaches (except for MiB) are unable to provide useful outputs leading to extremely low results, while our method stands out from competitors outperforming also MiB by a good margin. Visual results for each scenario are reported as last rows of Fig. 3, which confirm our considerations showing how SDR produces less noisy predictions and does not overestimate the background as some competitors.

6.3. Ablation Study

To evaluate the effect of each component, we report an ablation analysis in Table 3 on the Pascal dataset in the challenging 15-1 scenario. As already noticed, FT leads to a great degradation of mIoU. Early continual semantic segmentation approaches use a classical output-level knowledge distillation [42, 43, 33] which show discrete benefits boosting the mIoU by almost 20%. Each component of the approach significantly contributes to the final mIoU providing non-overlapping and mutual benefits. Matching prototypes, sparsifying features vectors and constraining them via the contrastive objective regularize the latent space bringing large improvements on both old and new classes. We observe that also the contrastive loss brings a significant contribution if applied alone improving the mIoU of 13.5%. Introducing standard output-level knowledge distillation on top increases the accuracy on old classes mainly, and its unbiased version prevents forgetting even more accounting for the background shift across the incremental learning steps.

Table 4. Standard (non-incremental) semantic segmentation.

$\mathcal{L}_{ce}$	$\mathcal{L}_{sp}$	$\mathcal{L}_{cl}$	$\text{mIoU}_{\text{VOC2012}}$	$\text{mIoU}_{\text{ADE20K}}$
✓			75.4	38.3
✓	✓		75.8	38.7
✓		✓	75.8	38.8
✓	✓	✓	**76.3**	**39.3**

Finally, we show that two of the proposed approaches (namely, sparsity and contrastive learning) may be beneficial also for the more general case of standard (*i.e.*, non incremental) semantic segmentation. Hence, we conduct some additional experiments on Pascal VOC2012 and ADE20K, reported in Table 4, showing the clear benefit of the two components in this setup. On both datasets the outcome is consistent, gaining 0.9% and 1% respectively, even starting from an architecture (*i.e.*, Deeplab-v3+) which is already state of the art.

7. Conclusion

In this paper we presented some latent representation shaping techniques to prevent forgetting in continual semantic segmentation. In particular, the proposed constraints on the latent space regularize the learning process reducing forgetting whilst simultaneously improving the recognition of novel classes. A prototypes matching constraint enforces latent space consistency on old classes, a features sparsification objective reduces the number of active channels limiting cross-talk between features of different classes, and contrastive learning clusters features according to their semantic while tearing apart those of different classes. Our evaluation shows the effectiveness of the proposed techniques, which can also be seamlessly applied in combination of previous methods (*e.g.*, knowledge distillation). Future research will exploit the proposed techniques in different tasks, such as standard semantic segmentation and class-incremental open-set domain adaptation, and explore the combination of our approach with output-level techniques.

References

[1] Alessandro Achille, Tom Eccles, Loic Matthey, Chris Burgess, Nicholas Watters, Alexander Lerchner, and Irina Higgins. Life-long disentangled representation learning with cross-domain latent homologies. In *Neural Information Processing Systems (NeurIPS)*, pages 9873–9883, 2018. 2

[2] Rahaf Aljundi, Marcus Rohrbach, and Tinne Tuytelaars. Selfless sequential learning. *Proceedings of the International Conference on Learning Representations (ICLR)*, 2018. 2, 3

[3] Francisco M Castro, Manuel J Marín-Jiménez, Nicolás Guil, Cordelia Schmid, and Karteek Alahari. End-to-end incremental learning. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 233–248, 2018. 1, 2

[4] 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, 4, 9, 10, 11

[5] Arslan Chaudhry, Puneet K Dokania, Thalaiyasingam Ajanthan, and Philip HS Torr. Riemannian walk for incremental learning: Understanding forgetting and intransigence. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 532–547, 2018. 2

[6] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 40(4):834–848, 2018. 2

[7] Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. Rethinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv:1706.05587*, 2017. 2, 6

[8] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018. 5, 6

[9] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. *Proceedings of the International Conference on Machine Learning (ICML)*, 2020. 2, 5

[10] Matthias De Lange, Rahaf Aljundi, Marc Masana, Sarah Parisot, Xu Jia, Ales Leonardis, Gregory Slabaugh, and Tinne Tuytelaars. Continual learning: A comparative study on how to defy forgetting in classification tasks. *arXiv preprint arXiv:1909.08383*, 2019. 2, 6

[11] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 248–255. Ieee, 2009. 6

[12] Zhijie Deng, Yucen Luo, and Jun Zhu. Cluster alignment with a teacher for unsupervised domain adaptation. In *Proceedings of the International Conference on Computer Vision (ICCV)*, pages 9944–9953, 2019. 3

[13] Prithviraj Dhar, Rajat Vikram Singh, Kuan-Chuan Peng, Ziyan Wu, and Rama Chellappa. Learning without memorizing. *arXiv preprint arXiv:1811.08051*, 2018. 2

[14] Carl Doersch and Andrew Zisserman. Multi-task self-supervised visual learning. In *Proceedings of the International Conference on Computer Vision (ICCV)*, pages 2051–2060, 2017. 2

[15] Nanqing Dong and Eric P Xing. Few-shot semantic segmentation with prototype learning. In *Proceedings of the British Machine Vision Conference (BMVC)*, volume 3, 2018. 3

[16] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge 2012 Results, 2012. 2, 5, 6

[17] Chrisantha Fernando, Dylan Banarse, Charles Blundell, Yori Zwols, David Ha, Andrei A Rusu, Alexander Pritzel, and Daan Wierstra. Pathnet: Evolution channels gradient descent in super neural networks. *arXiv preprint arXiv:1701.08734*, 2017. 2

[18] Robert M French. Catastrophic forgetting in connectionist networks. *Trends in Cognitive Sciences*, 3(4):128–135, 1999. 1

[19] Alberto Garcia-Garcia, Sergio Orts-Escolano, Sergiu Oprea, Victor Villena-Martinez, Pablo Martinez-Gonzalez, and Jose Garcia-Rodriguez. A survey on deep learning techniques for image and video semantic segmentation. *Applied Soft Computing*, 70:41–65, 2018. 2

[20] Ian J Goodfellow, Mehdi Mirza, Da Xiao, Aaron Courville, and Yoshua Bengio. An empirical investigation of catastrophic forgetting in gradient-based neural networks. *Proceedings of the International Conference on Learning Representations (ICLR)*, 2014. 1, 2

[21] Yanming Guo, Yu Liu, Theodoros Georgiou, and Michael S Lew. A review of semantic segmentation using deep neural networks. *International Journal of Multimedia Information Retrieval*, 7(2):87–93, 2018. 2

[22] Raia Hadsell, Sumit Chopra, and Yann LeCun. Dimensionality reduction by learning an invariant mapping. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 2, pages 1735–1742. IEEE, 2006. 2

[23] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9729–9738, 2020. 2

[24] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016. 5, 6

[25] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *Neural Information Processing Systems, Deep Learning and Representation Learning Workshop*, 2015. 2

[26] Saihui Hou, Xinyu Pan, Chen Change Loy, Zilei Wang, and Dahua Lin. Learning a unified classifier incrementally via rebalancing. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 831–839, 2019. 2

[27] Khurram Javed and Martha White. Meta-learning representations for continual learning. In *Neural Information Processing Systems (NeurIPS)*, pages 1820–1830, 2019. 2

[28] Xu Ji, João F Henriques, and Andrea Vedaldi. Invariant information clustering for unsupervised image classification and segmentation. In *Proceedings of the International Conference on Computer Vision (ICCV)*, pages 9865–9874, 2019. 2

[29] Guoliang Kang, Lu Jiang, Yi Yang, and Alexander G Hauptmann. Contrastive adaptation network for unsupervised domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4893–4902, 2019. 3, 5

[30] Ronald Kemker, Marc McClure, Angelina Abitino, Tyler Hayes, and Christopher Kanan. Measuring catastrophic forgetting in neural networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2018. 2

[31] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. *arXiv preprint arXiv:2004.11362*, 2020. 2

[32] 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. 1, 2, 7, 4, 5, 6

[33] 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, 6, 7, 8, 3, 4, 5, 10, 11

[34] Dawei Li, Serafettin Tasci, Shalini Ghosh, Jingwen Zhu, Junting Zhang, and Larry Heck. Rilod: near real-time incremental learning for object detection at the edge. In *Proceedings of the 4th ACM/IEEE Symposium on Edge Computing*, pages 113–126, 2019. 1

[35] Xilai Li, Yingbo Zhou, Tianfu Wu, Richard Socher, and Caiming Xiong. Learn to grow: A continual structure learning framework for overcoming catastrophic forgetting. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2019. 2

[36] Zhizhong Li and Derek Hoiem. Learning without forgetting. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 40(12):2935–2947, 2018. 1, 2, 6, 7, 8, 3, 10, 11

[37] Jian Liang, Ran He, Zhenan Sun, and Tieniu Tan. Distant supervised centroid shift: A simple and efficient approach to visual domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2975–2984, 2019. 3, 5

[38] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3431–3440, 2015. 1, 2

[39] Arun Mallya, Dillon Davis, and Svetlana Lazebnik. Piggyback: Adapting a single network to multiple tasks by learning to mask weights. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 67–82, 2018. 2

[40] Arun Mallya and Svetlana Lazebnik. Packnet: Adding multiple tasks to a single network by iterative pruning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7765–7773, 2018. 2

[41] Michael McCloskey and Neal J Cohen. Catastrophic interference in connectionist networks: The sequential learning problem. In *Psychology of learning and motivation*, volume 24, pages 109–165. Elsevier, 1989. 1

[42] 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, 4, 10, 11

[43] Umberto Michieli and Pietro Zanuttigh. Knowledge distillation for incremental learning in semantic segmentation. *Computer Vision and Image Understanding*, 205:103167, 2021. 2, 3, 4, 5, 6, 8, 1

[44] Ishan Misra and Laurens van der Maaten. Self-supervised learning of pretext-invariant representations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6707–6717, 2020. 2

[45] 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. 4, 1

[46] Oleksiy Ostapenko, Mihai Puscas, Tassilo Klein, Patrick Jahnichen, and Moin Nabi. Learning to remember: A synaptic plasticity driven framework for continual learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11321–11329, 2019. 2

[47] Firat Ozdemir and Orcun Goksel. Extending pretrained segmentation networks with additional anatomical structures. *International journal of computer assisted radiology and surgery*, 14(7):1187–1195, 2019. 2

[48] German I Parisi, Ronald Kemker, Jose L Part, Christopher Kanan, and Stefan Wermter. Continual lifelong learning with neural networks: A review. *Neural Networks*, 2019. 2

[49] Xingchao Peng, Zijun Huang, Ximeng Sun, and Kate Saenko. Domain agnostic learning with disentangled representations. *Proceedings of the International Conference on Machine Learning (ICML)*, 2019. 2

[50] Pedro O Pinheiro. Unsupervised domain adaptation with similarity learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8004–8013, 2018. 3

[51] 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. 1, 2, 6, 7, 8, 10, 11

[52] Jonathan Schwarz, Jelena Luketina, Wojciech M Czarnecki, Agnieszka Grabska-Barwinska, Yee Whye Teh, Razvan Pascanu, and Raia Hadsell. Progress & compress: A scalable

framework for continual learning. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2018. 2

[53] Joan Serra, Didac Suris, Marius Miron, and Alexandros Karatzoglou. Overcoming catastrophic forgetting with hard attention to the task. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2018. 2

[54] Sumit Shekhar, Vishal M Patel, Hien V Nguyen, and Rama Chellappa. Generalized domain-adaptive dictionaries. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 361–368, 2013. 3

[55] Hanul Shin, Jung Kwon Lee, Jaehong Kim, and Jiwon Kim. Continual learning with deep generative replay. In *Neural Information Processing Systems (NeurIPS)*, pages 2990–2999, 2017. 2

[56] Konstantin Shmelkov, Cordelia Schmid, and Karteek Alahari. Incremental learning of object detectors without catastrophic forgetting. In *Proceedings of the International Conference on Computer Vision (ICCV)*, pages 3400–3409, 2017. 1, 3, 5, 6

[57] Jake Snell, Kevin Swersky, and Richard Zemel. Prototypical networks for few-shot learning. In *Neural Information Processing Systems (NeurIPS)*, pages 4077–4087, 2017. 4, 1

[58] Onur Tasar, Yuliya Tarabalka, and Pierre Alliez. Incremental Learning for Semantic Segmentation of Large-Scale Remote Sensing Data. *arXiv preprint arXiv:1810.12448*, 2018. 1, 2

[59] Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive multiview coding. *arXiv preprint arXiv:1906.05849*, 2019. 2, 5

[60] Zhuotao Tian, Xin Lai, Li Jiang, Michelle Shu, Hengshuang Zhao, and Jiaya Jia. Generalized few-shot semantic segmentation. *arXiv preprint arXiv:2010.05210*, 2020. 3

[61] Marco Toldo, Umberto Michieli, and Pietro Zanuttigh. Unsupervised domain adaptation in semantic segmentation via orthogonal and clustered embeddings. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 1358–1368, 2021. 3

[62] 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. 5, 2

[63] 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. 3, 4, 1

[64] Yu-Xiong Wang, Deva Ramanan, and Martial Hebert. Growing a brain: Fine-tuning by increasing model capacity. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2471–2480, 2017. 2

[65] Si Wu, Jian Zhong, Wenming Cao, Rui Li, Zhiwen Yu, and Hau-San Wong. Improving domain-specific classification by collaborative learning with adaptation networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 5450–5457, 2019. 3

[66] Yue Wu, Yinpeng Chen, Lijuan Wang, Yuancheng Ye, Zicheng Liu, Yandong Guo, Zhengyou Zhang, and Yun Fu. Incremental classifier learning with generative adversarial networks. *arXiv preprint arXiv:1802.00853*, 2018. 2

[67] Zhirong Wu, Yuanjun Xiong, Stella X Yu, and Dahua Lin. Unsupervised feature learning via non-parametric instance discrimination. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3733–3742, 2018. 2

[68] Yongqin Xian, Zeynep Akata, Gaurav Sharma, Quynh Nguyen, Matthias Hein, and Bernt Schiele. Latent embeddings for zero-shot classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 69–77, 2016. 2

[69] Tianjun Xiao, Jiaxing Zhang, Kuiyuan Yang, Yuxin Peng, and Zheng Zhang. Error-driven incremental learning in deep convolutional neural network for large-scale image classification. In *Proceedings of the ACM International Conference on Multimedia*, pages 177–186. ACM, 2014. 2

[70] Shaoan Xie, Zibin Zheng, Liang Chen, and Chuan Chen. Learning semantic representations for unsupervised domain adaptation. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 5423–5432, 2018. 3

[71] Mang Ye, Xu Zhang, Pong C Yuen, and Shih-Fu Chang. Unsupervised embedding learning via invariant and spreading instance feature. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6210–6219, 2019. 2

[72] Fisher Yu, Vladlen Koltun, and Thomas Funkhouser. Dilated residual networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 472–480, 2017. 2

[73] Friedemann Zenke, Ben Poole, and Surya Ganguli. Continual learning through synaptic intelligence. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 3987–3995, 2017. 2

[74] Heng Zhang, Vishal M Patel, Sumit Shekhar, and Rama Chellappa. Domain adaptive sparse representation-based classification. In *IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, volume 1, pages 1–8. IEEE, 2015. 3

[75] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2881–2890, 2017. 2

[76] Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ade20k dataset. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 633–641, 2017. 2, 5, 8

[77] Chengxu Zhuang, Alex Lin Zhai, and Daniel Yamins. Local aggregation for unsupervised learning of visual embeddings. In *Proceedings of the International Conference on Computer Vision (ICCV)*, pages 6002–6012, 2019. 2

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

Umberto Michieli and Pietro Zanuttigh

University of Padova,
Department of Information Engineering

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 8. Then, Section 9 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 10, 11 and 12.

8. 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 [63, 57, 45]) 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 [42, 43, 4, 33] (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:

$$\bar{\mathbf{f}}_i = \frac{\mathbf{f}_i}{\max_{f_{i,j} \in \mathbf{f}_i} f_{i,j}} \qquad \mathbf{f}_i \in \mathbf{F}_n \tag{S1}$$

$$\bar{\mathbf{f}}_i = \frac{\mathbf{f}_i}{\max_{g_{j,l} \in \mathbf{g}_j} g_{j,l}} \qquad \mathbf{f}_i, \mathbf{g}_j \in \mathbf{F}_n \tag{S2}$$

$$\bar{\mathbf{f}}_i = \frac{\mathbf{f}_i}{||\mathbf{f}_i||_2} \qquad \mathbf{f}_i \in \mathbf{F}_n \tag{S3}$$

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:

$$\mathcal{L}_{sp} = \frac{1}{|\mathbf{f}_i \in \mathbf{F}_n|} \sum_{\mathbf{f}_i \in \mathbf{F}_n} \frac{\sum_j \bar{f}_{i,j}^2}{\sum_j \bar{f}_{i,j}} \tag{S4}$$

$$\mathcal{L}_{sp} = \frac{1}{|\mathbf{f}_i \in \mathbf{F}_n|} \sum_{\mathbf{f}_i \in \mathbf{F}_n} \frac{\sum_j \bar{f}_{i,j}^3}{\sum_j \bar{f}_{i,j}}. \tag{S5}$$

Finally, following the success of recent works exploiting entropy minimization [62] 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:

$$\bar{\mathbf{f}}_i = \frac{\mathbf{f}_i}{||\mathbf{f}_i||_1} \qquad \mathbf{f}_i \in \mathbf{F}_n \tag{S6}$$

$$\bar{\mathbf{f}}_i = \frac{\exp(\mathbf{f}_i)}{\sum_j \exp(f_{i,j})} \qquad \mathbf{f}_i \in \mathbf{F}_n \tag{S7}$$

$$\mathcal{L}_{sp} = \frac{1}{|\mathbf{f}_i \in \mathbf{F}_n|} \sum_{\mathbf{f}_i \in \mathbf{F}_n} \sum_j -\bar{f}_{i,j} \cdot \log\left(\bar{f}_{i,j}\right) \tag{S8}$$

Table S1 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 S1. 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**

9. 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 [42, 4, 43, 33]. Additional considerations have been already addressed in [43], 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 latent 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%).

10. 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 S1, S2 and S3 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 com-

petitors, *i.e.*, LwF [36], ILT [42], CIL [33] and MiB [4]. The images show how our approach alleviates forgetting and at the same time accommodates 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 S4 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.

11. 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 [4]).

Pascal VOC2012. The results on two sample scenes from this dataset are reported in Figure S5 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). Latent representations got even more damaged across subsequent steps, while our approach (SDR) can reduce the interference on latent representations of old classes. Similar considerations also holds for the second set of images, although forgetting is less evident in this scenario: SDR achieves 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 S6, and 2 incremental steps each adding 50 classes to the model (50-50) in Figure S7.

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 S6, 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 S7 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.

12. Quantitative Results: per-Class Accuracy

We also analyze per-class accuracy for all 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 S2 and S3), 15-5 (Tables S4 and S5) and 15-1 (Tables S6 and S7) scenarios on the VOC2012 dataset.

Even when adding as little as 1 class (scenario 19-1 in Tables S2 and S3) we 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 can 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 S4 and S5 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 S6 and S7. 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.

Figure S1. 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*).

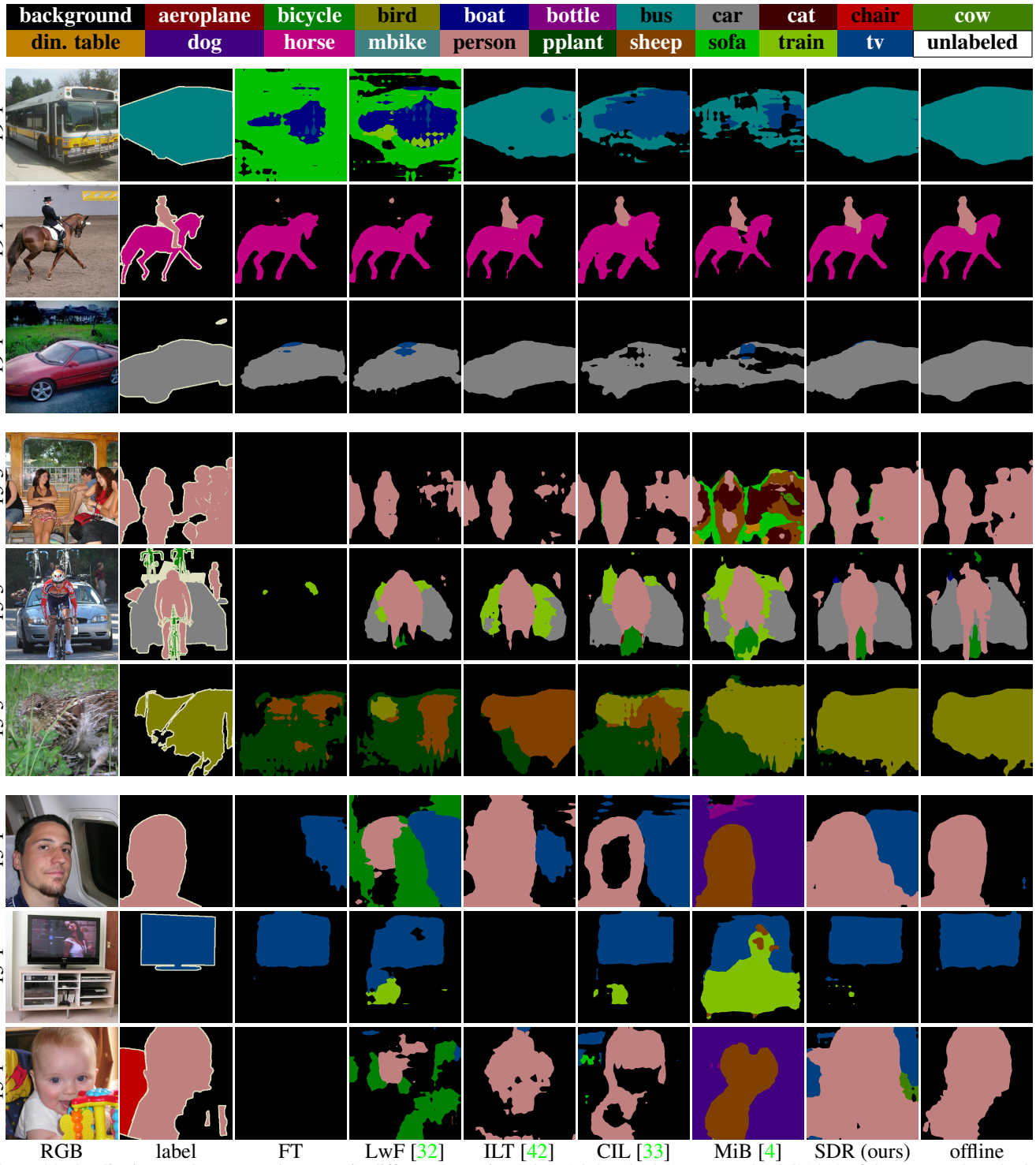

Figure S2. Qualitative results on sample scenes in different scenarios (19-1, 15-5 and 15-1) on Pascal VOC 2012 of the proposed method and of competing approaches in the disjoint setup (*best viewed in colors*).

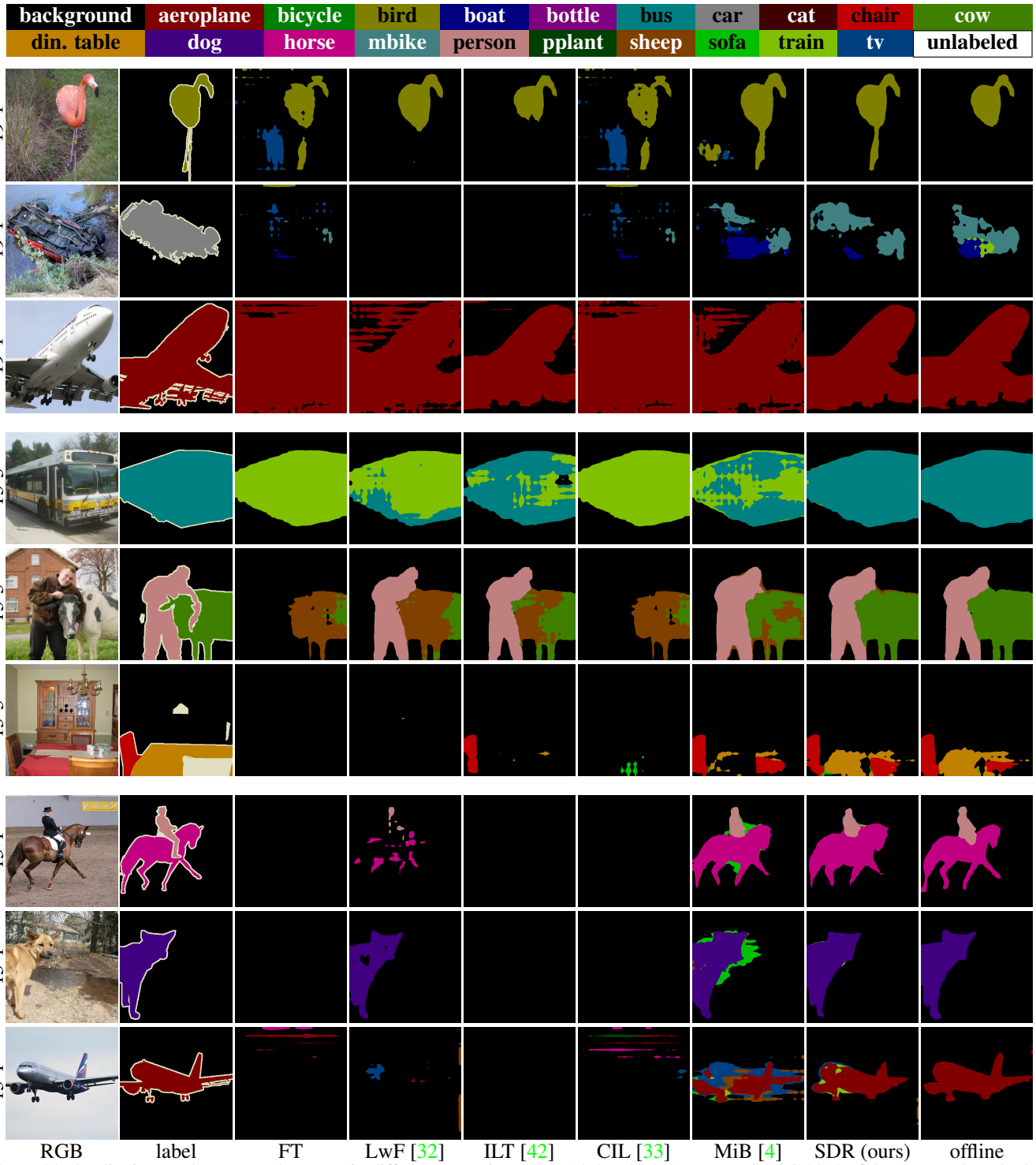

Figure S3. Qualitative results on sample scenes in different scenarios (19-1, 15-5 and 15-1) on Pascal VOC 2012 of the proposed method and of competing approaches in the overlapped setup (*best viewed in colors*).

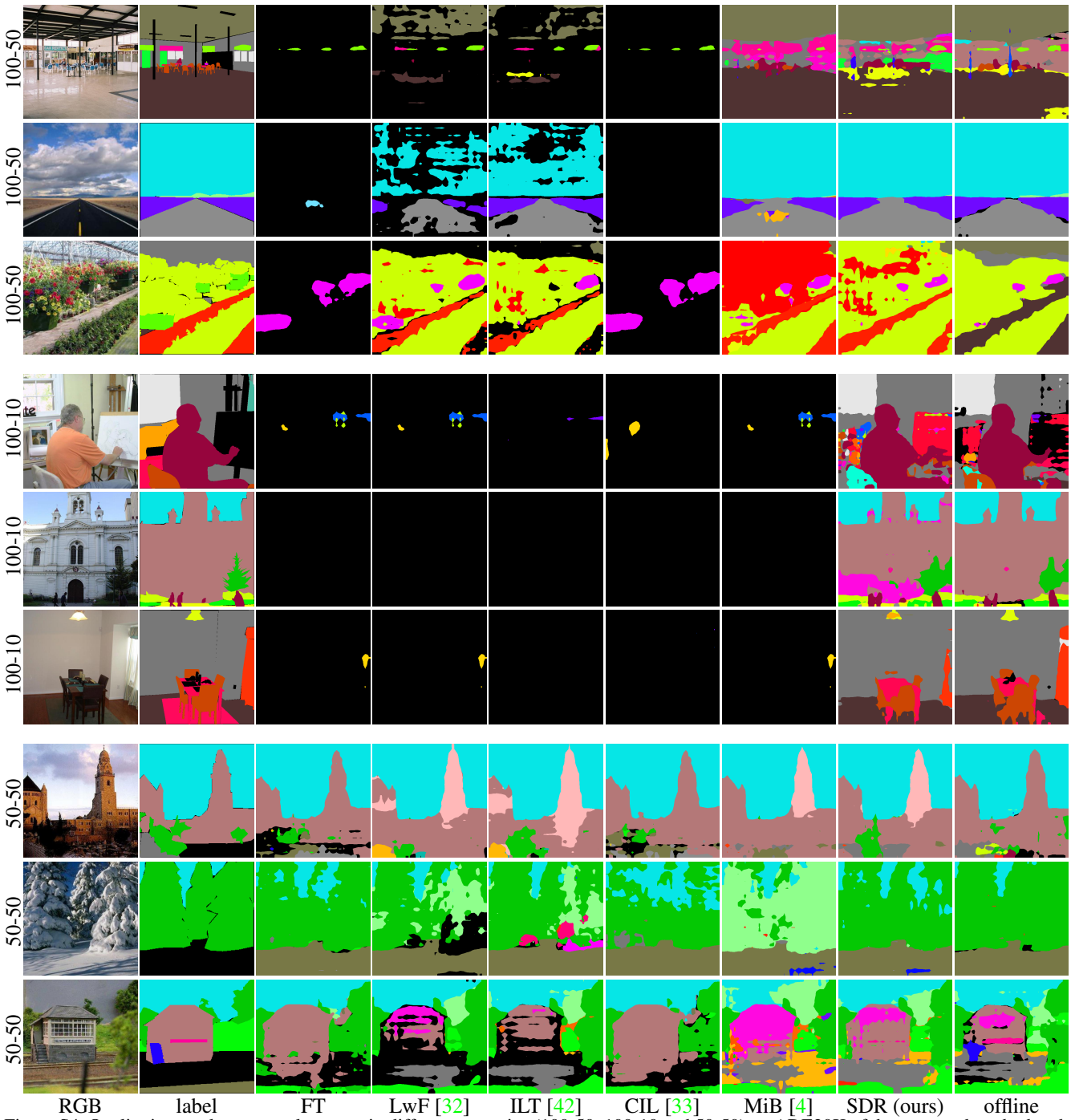

Figure S4. Qualitative results on sample scenes in different scenarios (100-50, 100-10 and 50-50) on ADE20K of the proposed method and of competing approaches (*best viewed in colors*).

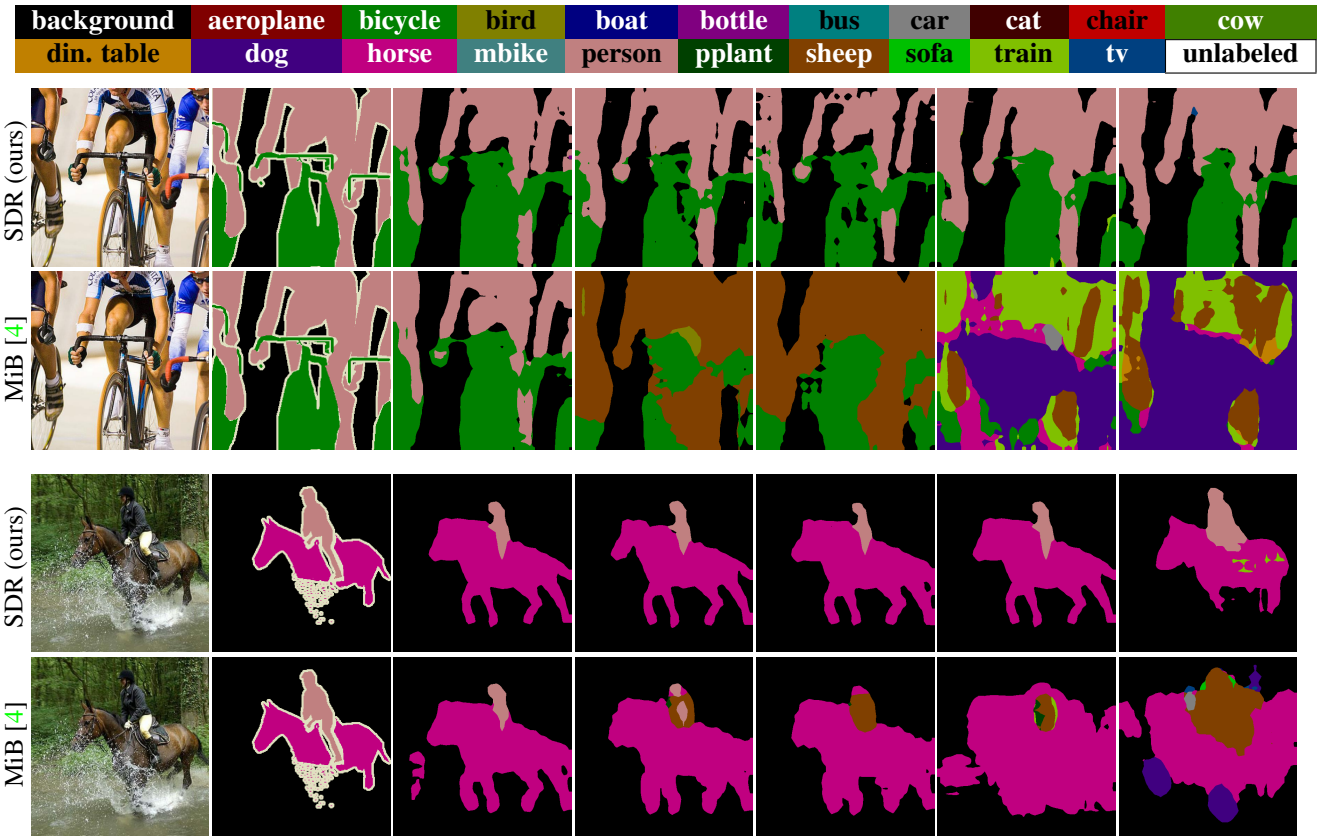

Figure S5. Qualitative results on sample scenes in the disjoint experimental protocol 15-1 on Pascal VOC 2012 during the various incremental steps (*best viewed in colors*).

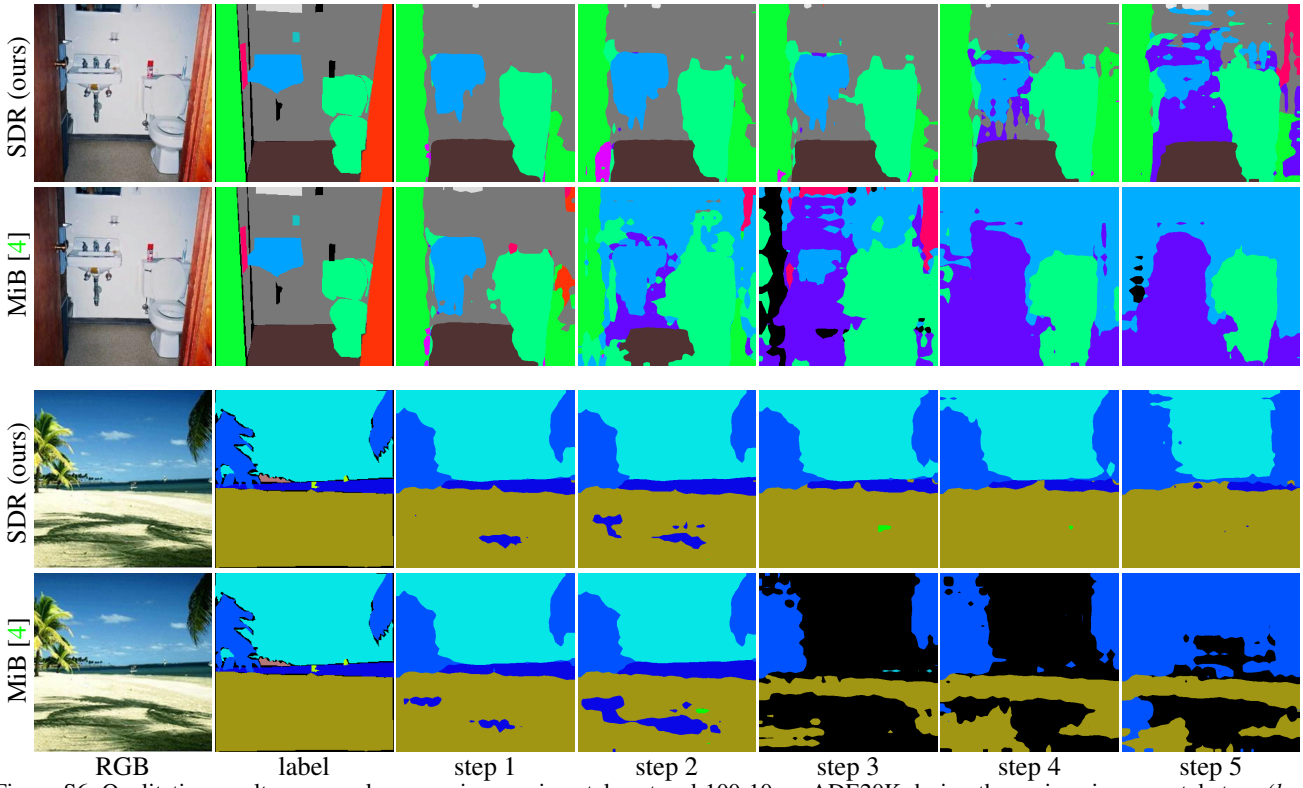

Figure S6. Qualitative results on sample scenes in experimental protocol 100-10 on ADE20K during the various incremental steps (*best viewed in colors*).

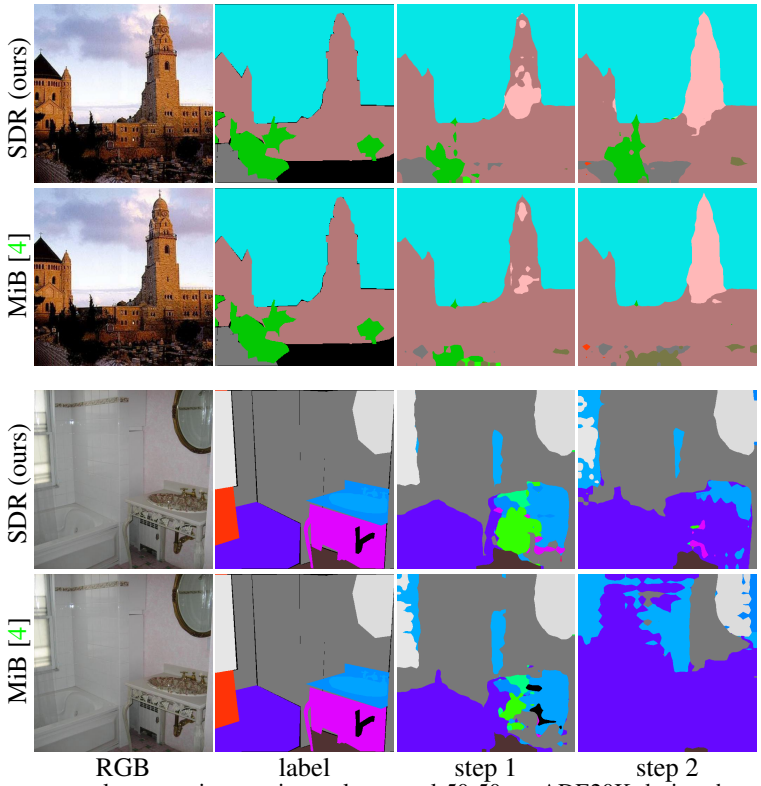

Figure S7. Qualitative results on sample scenes in experimental protocol 50-50 on ADE20K during the various incremental steps (*best viewed in colors*).

Table S2. 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 [36]	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 [51]	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 [42]	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 [33]	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 [4]	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 S3. 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 [36]	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 [51]	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 [42]	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 [33]	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 [4]	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 S4. 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 [36]	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 [51]	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 [42]	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 [33]	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 [4]	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 S5. 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 [36]	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 [51]	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 [42]	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 [33]	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 [4]	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 S6. 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 [36]	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 [51]	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 [42]	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 [33]	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 [4]	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 S7. 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 [36]	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 [51]	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 [42]	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 [33]	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 [4]	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