

Background Adaptation with Residual Modeling for Exemplar-Free Class-Incremental Semantic Segmentation

Anqi Zhang[✉] and Guangyu Gao^{*✉}

School of Computer Science and Technology,
Beijing Institute of Technology, Beijing, China

andy_zaq@outlook.com

guangyugao@bit.edu.cn

Abstract. Class Incremental Semantic Segmentation (CISS), within Incremental Learning for semantic segmentation, targets segmenting new categories while reducing the *catastrophic forgetting* on the old categories. Besides, *background shifting*, where the background category changes constantly in each step, is a special challenge for CISS. Current methods with a shared background classifier struggle to keep up with these changes, leading to decreased stability in background predictions and reduced accuracy of segmentation. For this special challenge, we designed a novel background adaptation mechanism, which explicitly models the background residual rather than the background itself in each step, and aggregates these residuals to represent the evolving background. Therefore, the background adaptation mechanism ensures the stability of previous background classifiers, while enabling the model to concentrate on the easy-learned residuals from the additional channel, which enhances background discernment for better prediction of novel categories. To precisely optimize the background adaptation mechanism, we propose Pseudo Background Binary Cross-Entropy loss and Background Adaptation losses, which amplify the adaptation effect. Group Knowledge Distillation and Background Feature Distillation strategies are designed to prevent forgetting old categories. Our approach, evaluated across various incremental scenarios on Pascal VOC 2012 and ADE20K datasets, outperforms prior exemplar-free state-of-the-art methods with mIoU of 3.0% in VOC 10-1 and 2.0% in ADE 100-5, notably enhancing the accuracy of new classes while mitigating catastrophic forgetting. Code is available in <https://andyzaq.github.io/barmsite/>.

1 Introduction

With the development of deep learning and the support of convolutional neural networks (CNNs) [27, 30, 31, 51] and Transformers [20, 33, 53, 54], current semantic segmentation networks [10–13, 15, 16, 24, 34, 46] have achieved remarkable results on standard supervised learning tasks. However, real-world data, arriving as an

^{*} Corresponding Author.

unstable continuous stream, poses the challenge of retaining prior knowledge while accommodating new concepts, termed as *catastrophic forgetting* [38, 39].

To tackle this problem, *incremental learning* is proposed to adapt to changing data streams for new concepts, but also avoid forgetting old knowledge, especially for the classification task, i.e., Class-Incremental Learning (CIL) [14, 28, 32, 59, 65]. Cermelli et al. [6] introduced CIL into Semantic Segmentation as Class Incremental Semantic Segmentation (CISS). Since the semantic segmentation tasks involve dense predictions, the issue of catastrophic forgetting typically becomes more challenging. Most recent works [6, 8, 21] have struggled to mitigate *catastrophic forgetting* through knowledge distillation [3, 21, 40] or parameter freezing [8, 61]. Although these strategies effectively alleviate *catastrophic forgetting*, it comes at the cost of the model’s plasticity, making it challenging to learn novel classes. In current state-of-the-art approaches [3, 61], the performance of established old classes approaches the upper limit of *joint* training, yet there still remains a significant gap in mIoU for novel classes. However, in the context of *lifelong learning* scenarios featuring an infinite influx of novel classes, the incremental learner’s plasticity becomes increasingly crucial. Therefore, the optimal solution should be fine-tuning more parameters for novel classes, while *catastrophic forgetting* needs to be handled properly.

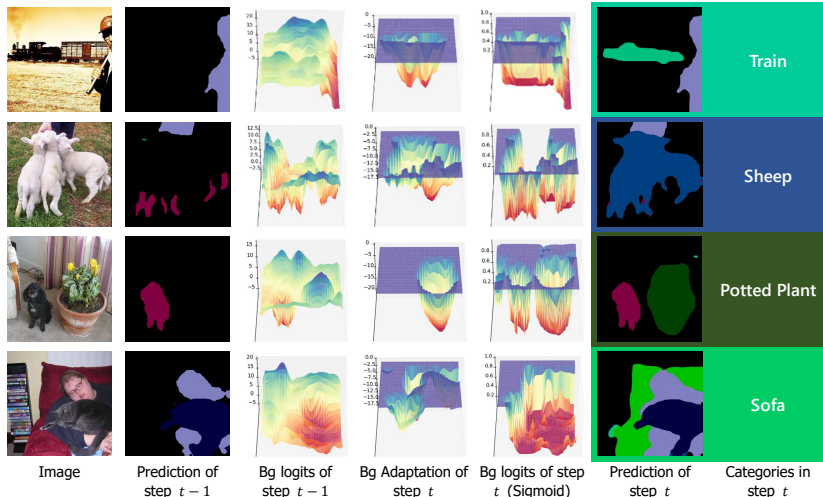


Fig. 1: 3D Visualization of Background Adaptation results. The background logits for step t combine those from step $t - 1$ with the adaptation logits learned in step t , preventing disorderly adjustments and focusing on residuals. The current step background logits are processed using the Sigmoid function. Note that colors range from red (small values) to blue (large values).

Nevertheless, *background shifting*, where the background category changes constantly in each step, is another challenge for CISS. Upon further investiga-

tion, it was observed that previous methods within CISS often utilize a shared background classifier across each step. However, novel classes are typically considered as the background in previous learning steps, indicating that the background class undergoes continuous changes throughout incremental learning. This way results in an incessantly shifting background in each step, leading to disruption in background prediction and inaccurate learning of novel classes. Some approaches [8, 9, 41] store old exemplars in extra memory space to prevent *background shifting*, while long-term storage of such data may not be feasible in real-world scenarios. To enhance the *incremental* nature of the CISS method and reduce background instability, we revise the representation of the background class and propose an exemplar-free background adaptation mechanism. In particular, the background adaptation mechanism explicitly models the background residual rather than the background itself in each incremental step. The learned residuals of the additional channel are aggregated with the previously learned background logits to represent the evolving background, as illustrated in Fig. 1. By employing the background adaptation mechanism, we only need to optimize the additional channel of residuals for constructing the background during the incremental steps. This approach eliminates the need for repeatedly learning background classifiers associated with old classes, thereby avoiding the instability of optimizing a shared background classifier.

Moreover, we customize our training approaches in order to fit the training process of the background adaptation structure and promote its performance. We first introduce the widely used pseudo-labeling [8, 21] to generate a pseudo-background, and then we design a PB-BCE loss to focus on optimizing the classifier of novel classes and related background adaptation residual. Besides, for the Background Adaptation channel, the objective is to minimize the values in the regions of novel categories, while ensuring that the values in other regions remain above 0. Thus, two auxiliary Background Adaptation losses are specifically applied to the background adaptation channel for precise adjustment. Given our pursuit of plasticity on novel categories without external memory, we have devised more comprehensive exemplar-free knowledge distillation strategies to combat *catastrophic forgetting*. We further adopt knowledge distillation strategies on the logit maps of old categories and the intermediate features, namely the Group Knowledge Distillation and Background Feature Distillation, respectively. Finally, our approach outperforms previous methods in multiple benchmarks, especially in realistic and challenging long-term scenarios such as 10-1 of VOC2012 and 100-5 of ADE20K. Our approach achieves significant performance gains of **4.5%** and **5.2%** on novel classes, respectively, compared to the previous state-of-the-art.

Overall, our contributions are summarized as follows:

- We propose a Background Adaptation strategy in incremental steps by optimizing an additional channel for residual modeling instead of the shared background classifier, which avoids unstable optimization and prevents chaos in background logits.

- We design PB-BCE loss and Background Adaptation losses to enhance the effectiveness of Background Adaptation, while two distillation strategies ensure maintaining knowledge of old categories.
- Experiments validate the effectiveness of our proposed method, showcasing a new state-of-the-art performance on CISS benchmarks.

2 Related Works

2.1 Class Incremental Learning

Class Incremental Learning (CIL) breaks the constraints of the standard training process in an incremental form on new classes. CIL task aims to continually learn the new classes without retraining on the data of old classes, which significantly reduces the training cost. As a compromise, the CIL task has to effectively solve the problem of *catastrophic forgetting* [38, 39]. Current CIL methods are divided into several main categories. Replay-based methods recall the data of old classes during incremental training. Some of the approaches [2, 4, 35] reserve memory buffers and temporarily store representative past data. Others [42, 49, 55] introduce powerful generative models to generate pseudo-data for rehearsal. Regularization-based methods introduce regularization terms to prevent forgetting the old classes. These methods utilize loss functions as penalties [1, 29, 45, 47] or knowledge distillation from the previous model [19, 32, 44]. Representation-based methods attempt to transfer representations to certain categories, which apply adaptation on a fixed pre-trained backbone [5, 22, 50, 60] or self-supervised learning [7, 36, 43].

2.2 Class Incremental Semantic Segmentation

The achievements in CIL tasks encourage the Incremental Learning pipeline to develop in other dense prediction tasks, including Semantic Segmentation. Class Incremental Semantic Segmentation (CISS) task is first introduced in MiB [6], which focused on reconstructing the background region under the restriction from knowledge distillation. Since then, various methods [57, 63] have been introduced in the CISS task. PLOP [21] defined the pseudo-labeling and multi-scale local distillation to remember the old categories. SDR [40] introduced category-wise prototype matching and contrastive learning methods to enlarge the difference between categories. RCIL [58] and EWF [56] explored methods for integrating parameters of new and old models. DKD [3] decomposed the class logits into positive and negative logits for better knowledge distillation and applied threshold to generate background regions. Incrementer [48] borrows prompt learning from ViT [20] and applies the structure of Segmenter [52] instead of DeeplabV3 [12], which is the basic structure of other methods. Besides, several approaches utilize external memory [66] or models [37, 62] for better performance. SSUL [8] introduced a Saliency-map detector to identify the saliency region and store some of the previous instances in memory. MicroSeg [61] utilized proposals from pretrained Mask2Former [15] as a guidance. ALIFE [41] memorized category features and adapted for remembering the old classes.

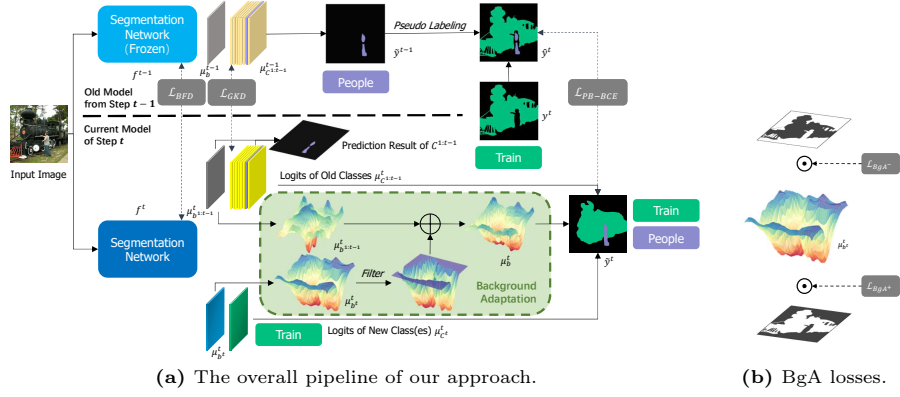


Fig. 2: Overview of our framework that consists of a Background Adaptation mechanism and proposed losses. (a) The Background Adaptation mechanism uses the negative part of the adaptation channel to rectify the background logits. (b) Two Background Adaptation losses separately optimize the regions *w/* and *w/o* novel categories to enhance the performance of the Background Adaptation.

3 Methods

3.1 Preliminaries

The Class-Incremental Semantic Segmentation (CISS) task performs training in T steps. During step t , the training data D^t containing the labels of the novel classes C^t are applied for training, while the labels of old classes $C^{1:t-1}$ are expunged. In particular, each sample of the training data D^t consists of an RGB image $x^t \in \mathbb{R}^{3 \times H \times W}$ and a corresponding ground truth label y^t on novel classes C^t . Besides, the ground truth label y^t replaces the labels of the $C^{1:t-1}$ with the background label c_b as the old classes will not appear in C^t to ensure $C^{1:t-1} \cap C^t = \emptyset$. Therefore, the model f^t being trained at step t not only needs to learn to segment novel categories but also has to retain the capability to segment old categories. However, while learning the novel categories, the model f^t tends to overfit the novel categories C^t as f^t cannot contact with samples of $C^{1:t-1}$, causing severe *catastrophic forgetting*. Moreover, the novel categories can only appear in the background region during the previous steps. We are more likely to classify them from the previous background prediction region, which motivates us to focus on background adaptation.

3.2 Background Adaptation

Previous studies often apply only one background classifier. They have to readjust the parameters of the background classifier whenever the model learns novel categories, which causes *catastrophic forgotten* even if the classifier of the old categories remains the same probability output. The unstable background classifier

could be a thorny constraint factor. To tackle the problems above, we design a Background Adaptation mechanism. Different from the previous distinct background classifier, we introduce an additional background adaptation channel for each classifier of the incremental process.

During the training process of incremental step t , the current model f^t inherits the parameters from the previous model f^{t-1} and adds a new classifier $f_{C^t}^t$ for parallel classification of categories in C^t alongside the previous classifiers. The encoder f_e^t and the decoder f_d^t transform the input image x^t into features map $F_d^t \in \mathbb{R}^{h \times w \times c}$:

$$F_d^t = f_d^t(f_e^t(x^t)). \quad (1)$$

Compared to the standard classifier of the previous methods [8, 21, 41], our classifier of each incremental step has an additional channel for background adaptation. Therefore, for each group of categories C^i from step i , the classifier $f_{C^i}^t$ of the model f^t generates the corresponding classification logit maps $\mu_{C^i \cup \{b_i\}}^t \in \mathbb{R}^{h \times w \times (|C^i|+1)}$:

$$\mu_{C^i \cup \{b_i\}}^t = f_{C^i}^t(F_d^t), \quad (2)$$

where $i \in \{1, 2, \dots, t\}$. We define b^1 to represent the background class of the initial step, and b^i denotes the background adaptation for subsequent incremental steps when $i \geq 2$. The symbol b refers to the adapted background class. Thus, $\mu_{C^i \cup \{b_i\}}^t$ contains the probability maps μ_{C^i} of $|C^i|$ categories from step i and an additional channel $\mu_{b^i}^t$ for constructing the background logits μ_b^t . Considering that the region of the novel categories C^t is more likely inherited from the background region of the previous step, the classifier $f_{C^t}^t$ adjusts the background logits through $\mu_{b^t}^t$, which decreases the probability values in the region $\mathcal{R}_{C^t}$ of C^t . The parameters of previous classifiers are not required to continuously optimize for new background regions, thereby avoiding instability in the classification of old categories.

The background adaptation channel $\mu_{b^t}^t$ can correct the background logits $\mu_b^t \in \mathbb{R}^{h \times w \times 1}$ in $\mathcal{R}_{C^t}$, yet the noise in other regions of $\mu_{b^t}^t$ may interfere with the background prediction. Particularly, when old categories are present in the image, there is a risk that these regions could be incorrectly covered by the adapted background. Hence, we design a *Filter* operation that replaces positive values of $\mu_{b^i}^t$ with 0. As shown in Fig. 2a, with the *Filter* operation applied, the residual modeling of background logits during the inference process is as follows:

$$\mu_b^t = \mu_{b^1}^t + \sum_{i=2}^t \text{Filter}(\mu_{b^i}^t), \quad (3)$$

where the background logits from the initial step $\mu_{b^1}^t$ stay the same and the background adaptation logits from the incremental steps $\mu_{b^i}^t$ only remain the negative values via the *Filter*. In particular, to make μ_b^t more friendly for gradient backpropagation, we abandoned the *Filter* on the background adaptation logits

of novel classes $\mu_{b^t}^t$ during the training process:

$$\mu_b^t = \mu_{b^1}^t + \sum_{i=2}^{t-1} Filter(\mu_{b^i}^t) + \mu_{b^t}^t. \quad (4)$$

3.3 Incremental Training Strategy

Pseudo Labeling. During the incremental steps, even if objects of old categories $C^{1:t-1}$ exist in the image, we cannot obtain corresponding labels from the ground truth. In order to simulate a ground truth label that contains the categories we have learned, we follow [8, 21, 61] to introduce the prediction probability of the previous step $\varphi^{t-1} \in \mathbb{R}^{h \times w \times |C^{1:t-1}|}$ from $f^{t-1}(x^t)$ with Sigmoid function and combine it with the current ground truth label y^t :

$$\hat{y}^t = \begin{cases} y^t, & (y^t \in C^t) \vee \{(y^t \in c_b) \wedge (\max(\varphi^{t-1}) < \tau)\} \\ \tilde{y}^{t-1}, & (y^t \in c_b) \wedge (\max(\varphi^{t-1}) \geq \tau) \end{cases} \quad (5)$$

where $\hat{y}^t \in \mathbb{R}^{h \times w}$ denotes the generated pseudo label and $\tilde{y}^{t-1} = \operatorname{argmax}(\varphi^{t-1})$ represents the final prediction from $f^{t-1}(x^t)$.

PB-BCE Loss. Binary Cross-Entropy (BCE) loss is extensively employed for learning to segment objects of novel categories. We follow the usage of the Sigmoid function and activate μ^t as φ^t . Some of the CISS methods [8, 21, 61] utilize the pseudo mask as the ground truth for supervision, which effectively replaces the background label in the regions of the old categories. However, the ambiguous supervision of the old categories could improperly optimize the classifiers of $C^{1:t-1}$ and destroy their identification capability. Therefore, we design a Pseudo Background BCE (PB-BCE) loss which cuts off the backpropagation of old categories, meaning that we only take the labels of current categories $\hat{y}_{C^t}^t$ and pseudo-background $\hat{y}_{c_b}^t$ for optimization:

$$\mathcal{L}_{PB-BCE} = -\frac{1}{hw} \sum_{k=1}^{hw} \sum_{c \in \{c_b\} \cap C^t} CE(\varphi_{c,k}^t, \hat{\psi}_{c,k}^t), \quad (6)$$

where k represents the index of pixel and $\hat{\psi}_c^t$ denotes the binary map of class c in pseudo map $\hat{y}^t$. CE is the cross-entropy calculation:

$$CE(\varphi_*^t, \hat{\psi}_*^t) = \hat{\psi}_*^t \cdot \log \varphi_*^t + (1 - \hat{\psi}_*^t) \cdot \log(1 - \varphi_*^t). \quad (7)$$

Although the whole background prediction logits μ_b^t are involved in the loss calculation, only the background adaptation channel for current categories $\mu_{b^t}^t$ serves for optimization. Previous background logits are detached to prevent disorderly evolving on parameters of the old classifiers.

Background Adaptation Losses. The Background Adaptation (BgA) losses, as shown in Fig. 2b, are designed for further refined adjustments on the background adaptation channel $\mu_{b^t}^t$. We propose two losses for two different regions of $\mu_{b^t}^t$ according to the target effect. For regions $\mathcal{R}_{C^t}$ belonging to the current classes C^t , the ideal target is to have negative values as small as possible. To achieve this, we introduce cross-entropy loss $\mathcal{L}_{BgA^+}$ for supervision:

$$\mathcal{L}_{BgA^+} = -\frac{1}{|\mathcal{R}_{C^t}|} \sum_{k \in \mathcal{R}_{C^t}} \log \frac{\exp(-\mu_{b^t,k}^t)}{1 + \exp(-\mu_{b^t,k}^t)}, \quad (8)$$

where $|\mathcal{R}_{C^t}|$ represents the number of pixels in $\mathcal{R}_{C^t}$. Besides, as for the regions of the background that might contain old classes, we hope the value of $\mu_{b^t}^t$ could be 0 for maintaining the historical background prediction result. However, achieving precise control over the whole region to be a precise value is challenging and may be impractical. Therefore, we set the positive logits in the adaptation channel of the old categories to 0 in Eq. 4 to avoid the noise and apply a Triplet loss $\mathcal{L}_{BgA^-}$ to supervise the background region:

$$\mathcal{L}_{BgA^-} = \frac{1}{hw - |\mathcal{R}_{C^t}|} \sum_{k \notin \mathcal{R}_{C^t}} \max[0, (1 - \varphi_{b^t,k}^t)^2 - (\varphi_{b^t,k}^t - 0)^2]. \quad (9)$$

We first activate the background adaptation channel as $\varphi_{b^t}^t$ via Sigmoid function, then the logits are mapped to the range from 0 to 1. We aim for the logits $\mu_{b^t}^t$ of this region to be greater than 0 while not overfitting to the objects in the background, meaning that values of $\varphi_{b^t}^t$ lie within the range of 0.5 to 1 and not close to 1. In particular, The Triplet loss $\mathcal{L}_{BgA^-}$ respectively set 0 and 1 as negative and positive anchors. We simultaneously constrain the loss not to go below 0 by setting a lower bound. Therefore, once the value of the pixel exceeds 0.5, *i.e.*, the middle of the positive and negative anchors, there will be no more gradient on the pixel for backpropagation, which mitigates overfitting to the objects in the background.

Knowledge Distillation Strategies. In the training process of incremental steps for CISS tasks, substantial alterations to the existing parameters of f^t frequently occur, giving rise to situations of *catastrophic forgetting*. Guided by the previous methods, we have devised a series of knowledge distillation strategies tailored to our background adaptation mechanism.

- **Group Knowledge Distillation.** In Sec. 3.2, we have introduced the classification logit map $\mu_{C^i \cup \{b_i\}}^t$ for each group of categories C^i , which contains an additional channel for background adaptation. We introduce a Group Knowledge Distillation loss $\mathcal{L}_{GKD}$ to ensure the stability of the old category logits, as well as the previous background adaptation channel:

$$\mathcal{L}_{GKD} = -\frac{1}{hw} \sum_{k=1}^{hw} \sum_{i=1}^{t-1} \sum_{c \in \{c_b^i\} \cup C^i} CE(\varphi_{c,k}^t, \varphi_{c,k}^{t-1}) \quad (10)$$

- where φ_*^{t-1} and φ_*^t denote the activated logits of μ_*^{t-1} and μ_*^t , respectively.
- **Background Feature Distillation.** When the network learning novel categories, not only the parameters of the novel classifier $f_{C^t}^t$ are optimized, but the encoder f_e^t and the decoder f_d^t are affected as well. Merely utilizing Group Knowledge Distillation loss $\mathcal{L}_{GKD}$ is insufficient in effectively mitigating *catastrophic forgetting* as it primarily focuses on constraining classifiers of the old categories. However, excessively constraining parameter optimization can impede learning for novel categories. Taking all these factors into account, we designed a Background Feature Distillation (BFD) method that only acts on the features out of the regions of current categories $\mathcal{R}_{C^t}$:

$$\mathcal{L}_{BFD} = \sum_{i=1}^{t-1} \|\psi_b^t \cdot F_{m^i}^t - \psi_b^t \cdot F_{m^i}^{t-1}\|_2^2, \quad (11)$$

where $F_{m^i}^t$ denotes the intermediate features in $f_{C^i}^t$, ψ_b^t represents the binary mask out of $\mathcal{R}_{C^t}$, and $\|\cdot\|_2$ calculates the Euclidean distance between the two matrix. The Background Feature Distillation allows other losses to supervise the region of novel categories $\mathcal{R}_{C^t}$ while restricting the negative effect outside $\mathcal{R}_{C^t}$, where old categories $C^{1:t-1}$ could appear.

3.4 Objective Function

With our proposed losses above, we construct our objective function as follows:

$$\mathcal{L} = \mathcal{L}_{PB-BCE} + \lambda_1 \cdot \mathcal{L}_{BgA^+} + \lambda_2 \cdot \mathcal{L}_{BgA^-} + \lambda_3 \cdot \mathcal{L}_{GKD} + \lambda_4 \cdot \mathcal{L}_{BFD}, \quad (12)$$

where $\mathcal{L}_{PB-BCE}$ supervises the learning of novel classes, $\mathcal{L}_{PB-BCE}$ and $\mathcal{L}_{BgA^-}$ enhance the effect of Background Adaptation, $\mathcal{L}_{GKD}$ and $\mathcal{L}_{BFD}$ mitigate forgetting old classes. Hyperparameters $\lambda_1, \lambda_2, \lambda_3, \lambda_4$ balance these constraints.

4 Experiments

4.1 Experimental Setup

Datasets. We follow the settings of most previous work and evaluate our proposed methods on Pascal VOC 2012 [23, 26] and ADE20K [64] datasets. Pascal VOC 2012 is composed of 20 different classes with 10,582 images for training and 1,449 images for validation. ADE20K has 150 classes that contain 100 thing classes and 50 stuff classes. There are 20,210 images for training and 2,000 images for validation in ADE20K.

Protocols. Earlier work [6, 21] has defined two types of training settings, i.e. *overlapping* and *disjoint*. The *overlapping* setting contains future classes in the previous steps, whereas the *disjoint* setting forbids future classes until they belong to the training classes. We follow recent work [3, 41, 56, 63] and adopt the *overlapping* setting for evaluating as it is more realistic. Besides, following the previous training protocols [6, 21, 56], we set each sub-task from the dataset as

$N_{ini} - N_{inc}$, where N_{ini} and N_{inc} denotes the quantity of the categories in the initial step and a single incremental step, respectively. For example, the 15-1 sub-task of Pascal VOC 2012 has 15 classes for the initial training step. During each incremental step, one novel class is applied for incremental training without previous classes. Specifically, we define 4 sub-tasks on Pascal VOC 2012, including 19-1 (2 steps), 15-1 (6 steps), 10-1 (11 steps), and 5-3 (6 steps). On ADE20K, we define 4 sub-tasks containing 100-50 (2 steps), 100-10 (6 steps), 50-50 (3 steps), and 100-5 (11 steps).

Evaluation metrics. We evaluate our proposed method with widely-used mean Intersection-over-Union (mIoU). The Intersection-over-Union (IoU) is defined as $IoU = \frac{TP}{TP+FP+FN}$, where TP , FP , and FN represent the quantity of true-positive, false-positive, and false-negative prediction pixels. The mIoU calculates the mean value of IoU corresponding to each category. To effectively measure the performance of CISS tasks, we compute the mIoU values of the initial classes C^0 , the incremental classes $C^{1:t}$, and all classes $C^{0:t}$, respectively.

Implementation details. Following the previous studies [6, 8, 40], we apply Deeplab-v3 [12] as the segmentation network with a ResNet-101 [27] for feature extraction. The ResNet-101 is initially pretrained on ImageNet-1K [18]. We use SGD optimizer to optimize the network. The initial learning rate of the initial step is set to 10^{-2} , while in the incremental steps, the initial learning rate is set to 10^{-3} for Pascal VOC 2012 and 10^{-2} for ADE20K. The momentum value is 0.9 and the weight decay is 10^{-4} in all steps. We apply the poly learning rate schedule following [3, 8, 12, 21]. During the initial step, the training epoch is set to 50 for Pascal VOC 2012 and 60 for ADE20K, while for each incremental step, we train the network for 20 epochs on Pascal VOC and 100 epochs on ADE20K. The batch size is 16 for Pascal VOC 2012 and 8 for ADE20K in all steps. The data augmentation follows the standard of previous work, which is composed of random scaling, random flipping, and random crop. The coefficients λ_1 , λ_2 , λ_3 , λ_4 , of the objective function are set to 1, 5, 1, 4. The threshold of pseudo labeling $\tau = 0.7$ is set to all settings. The implementation is built on Pytorch. The experiments are conducted on NVIDIA RTX2080Ti GPUs.

4.2 Quantitative Results

Comparison on Pascal VOC 2012. We evaluate our proposed method with current exemplar-free DeeplabV3-based methods on the Pascal VOC 2012 dataset. The sub-tasks of Pascal VOC 2012 have distinct characteristics. For example, the sub-task 19-1 has more categories in the initial step, while 5-3 has fewer classes for the initial training and more incremental steps, which is much more challenging. The results in Tab. 1 show that our proposed method outperforms other methods on novel classes and overall mIoU results of all sub-tasks. Particularly, our approach has 1.7% and 3.0% promotion of mIoU in more challenging sub-tasks of 10-1 and 5-3, respectively. Except for DKD [3] which refine the background class, our approach achieves at least 3.0% and 7.0% of promotion on mIoU. Compared to the methods with a standard shared background classifier, demonstrating the effectiveness of the Background Adaptation mechanism.

Table 1: Quantitative comparison with state-of-the-art exemplar-free methods on Pascal VOC 2012 in mIoU. Scores of novel classes and all classes in **bold** are the best while underlined are the second best.

Method	15-1 (6 steps)			10-1 (11 steps)			5-3 (6 steps)			19-1 (1 step)		
	0-15	16-20	all	0-10	11-20	all	0-5	6-20	all	0-19	20	all
MiB [CVPR20] [6]	38.0	13.5	32.2	12.2	13.1	12.6	57.1	42.5	46.7	71.2	22.1	68.9
SDR [CVPR21] [40]	47.3	14.7	39.5	32.4	17.1	25.1	-	-	-	69.1	32.6	67.4
PLOP [CVPR21] [21]	65.1	21.1	54.6	44.0	15.5	30.5	25.7	30.0	28.7	75.4	37.3	73.5
RCIL [CVPR22] [58]	70.6	23.7	59.4	55.4	15.1	34.3	63.1	34.6	42.8	77.0	31.5	74.7
EFW [CVPR23] [56]	77.7	32.7	67.0	71.5	30.3	51.9	61.7	42.2	47.7	77.9	6.7	74.5
AWT [WACV23] [25]	59.1	17.2	49.1	33.2	18.0	26.0	61.8	45.9	50.4	-	-	-
IDEC [TPAMI23] [63]	77.0	36.5	67.3	70.7	46.3	59.1	67.1	49.0	54.1	-	-	-
DKD [NeurIPS22] [3]	78.1	<u>42.7</u>	<u>69.7</u>	73.1	<u>46.5</u>	<u>60.4</u>	69.6	<u>53.5</u>	<u>58.1</u>	77.8	<u>41.5</u>	<u>76.0</u>
Ours	77.6	45.9	70.0	72.2	51.0	62.1	71.3	57.0	61.1	78.2	42.2	76.4
<i>Joint</i>	79.8	72.4	77.4	78.4	76.4	77.4	76.9	77.6	77.4	77.5	77.0	77.4

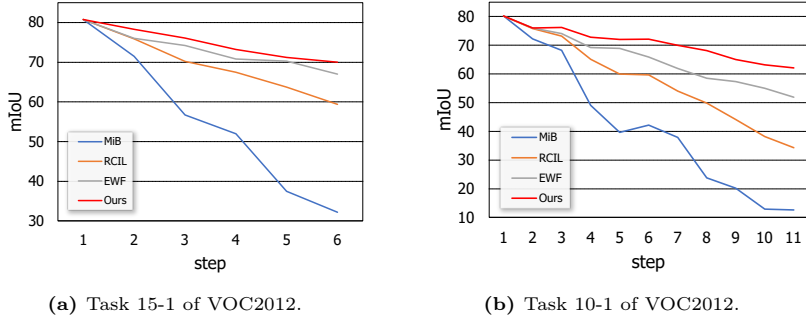


Fig. 3: The step-wise mIoU comparison of our approach with previous methods under the sub-tasks 10-1 and 15-1 of Pascal VOC 2012.

Besides, our method achieve an impressive advantage on the novel classes, with significant improvements of 4.5% and 3.5% compared to the previous state-of-the-art method. Many previous methods emphasize restricting catastrophic forgetting by boosting the mIoU of larger-proportion old classes in specific tasks (e.g. 15-1) to make the overall mIoU appear to be higher. However, our approach not only retains the ability to segment old classes but also significantly improves the accuracy of novel classes, aligning more closely with the *incremental* objective emphasized in CISS.

Comparison on ADE20K. We evaluate our method on the sub-tasks of more challenging ADE20K, including 100-50, 100-10, 100-5, and 50-50, as shown in Tab. 2. Although our method is slightly lower than the state-of-the-art method in short-term sub-tasks, we are remarkably ahead of other methods in more challenging long-term 100-10 and 100-5 tasks. The performance in the novel classes is still outstanding, which demonstrates the universality of our method.

Table 2: Quantitative comparison with state-of-the-art exemplar-free methods on ADE20K in mIoU. Scores of novel classes and all classes in **bold** are the best while underlined are the second best.

Method	100-50 (2 steps)			100-10 (6 steps)			50-50 (3 steps)			100-5 (11 steps)		
	0-100	101-150	all	0-100	101-150	all	0-50	51-150	all	0-100	101-150	all
MiB [6]	40.5	17.2	32.8	38.2	11.1	29.2	45.6	21.0	29.3	36.0	5.7	26.0
PLOP [21]	41.9	14.9	32.9	40.5	13.6	31.6	48.8	21.0	30.4	39.1	7.8	28.8
RCIL [58]	42.3	18.8	34.5	39.3	17.6	32.1	48.3	25.0	32.5	38.5	11.5	29.6
EWf [56]	41.2	21.3	34.6	41.5	16.6	33.2	-	-	-	41.4	13.4	<u>32.1</u>
AWT [25]	40.9	24.7	35.6	39.1	<u>21.4</u>	33.2	46.6	27.0	33.5	38.6	<u>16.0</u>	31.1
GSC [17]	42.4	19.2	34.8	40.8	16.2	32.6	46.2	26.4	33.0	-	-	-
IDEC [63]	42.0	18.2	34.1	40.3	17.6	32.7	47.4	26.0	33.1	39.2	14.6	31.0
DKD [3]	42.4	<u>22.9</u>	36.0	41.6	19.5	<u>34.3</u>	48.8	26.3	33.9	-	-	-
Ours	42.0	23.0	<u>35.7</u>	41.1	23.1	35.2	47.9	<u>26.5</u>	<u>33.7</u>	40.5	21.2	34.1
<i>Joint</i>	<i>44.3</i>	<i>28.2</i>	<i>38.9</i>	<i>44.3</i>	<i>28.2</i>	<i>38.9</i>	<i>51.1</i>	<i>33.3</i>	<i>38.9</i>	<i>44.3</i>	<i>28.2</i>	<i>38.9</i>

4.3 Ablation Studies

Component Ablations. In this section, we analyze our proposed components, including the Background Adaptation mechanism, Background Feature Distillation loss $\mathcal{L}_{BFD}$, Group Knowledge Distillation loss $\mathcal{L}_{GKD}$, Background Adaptation losses $\mathcal{L}_{BgA-}$ and $\mathcal{L}_{BgA+}$. The experiments are carried out on the representative sub-task 15-1 of Pascal VOC 2012. We separately evaluate different combinations of components and present the results in Tab. 3. The baseline method applies pseudo-labeling, $\mathcal{L}_{PB-BCE}$, and freezes the whole network except the classifiers of background and novel classes. Under the same experimental settings, the Background Adaptation mechanism significantly promotes 5.2% of overall mIoU, demonstrating its effectiveness in retaining the knowledge of old classes, and more notably, improving the performance of novel classes. The usage of distillation strategies with $\mathcal{L}_{BFD}$ and $\mathcal{L}_{GKD}$ preserves the performance of old classes while further boosting mIoU by 6.0% on novel classes. Moreover, the precise supervision of the Background Adaptation using $\mathcal{L}_{BgA-}$ and $\mathcal{L}_{BgA+}$ once again enhances the segmentation of novel classes with an additional 9.4% of mIoU. Subsequently, there is a growth of 2.4% on the overall mIoU, which proves that additional precise supervision is necessary. The results above validate the effectiveness of the Background Adaptation mechanism and other related losses. There is a notable improvement, particularly in the performance of novel classes, exceeding twice that of the baseline.

Background Adaptation Schemes. Sec. 4.3 has presented the importance of the Background Adaptation losses. We have devised multiple schemes to achieve more optimal results of Background Adaptation, as shown in Tab. 4. Initially, we aim to have the Background Adaptation value for the current ground truth background region equal to 0, using Mean Square Error (MSE) loss to minimize

Table 3: Ablation Study on the components of our proposed method on 15-1. BgA denotes Background Adaptation. In particular, the experiments *w/o* distillation losses (i.e. $\mathcal{L}_{BFD}$ and $\mathcal{L}_{GKD}$) freeze the whole network except the classifier f^t following the settings in [8]. All experiments contain $\mathcal{L}_{PB-BCE}$.

BgA	$\mathcal{L}_{BFD}$	$\mathcal{L}_{GKD}$	$\mathcal{L}_{BgA-}$	$\mathcal{L}_{BgA+}$	mIoU _{ini}	mIoU _{inc}	mIoU
					73.4	21.8	61.1
✓					77.4	30.5	66.3
✓	✓				76.7	37.1	67.3
✓	✓	✓			77.3	36.5	67.6
✓	✓	✓	✓		77.1	43.4	69.0
✓	✓	✓		✓	76.6	34.1	66.5
✓	✓	✓	✓	✓	77.6	45.9	70.0

Table 4: Ablation study on schemes of Background Adaptation. **Table 5:** Ablation Study on the feature distillation (FD) strategies.

	Filter	mIoU _{ini}	mIoU _{inc}	mIoU
MSE-0	✗	74.2	41.8	66.5
	✓	74.6	42.2	66.9
BCE-1	✓	77.4	44.7	69.6
	✗	65.7	45.4	60.8
Ours	✓	77.6	45.9	70.0

	mIoU _{ini}	mIoU _{inc}	mIoU
<i>w/o</i> FD	56.9	33.2	51.2
$\mathcal{L}_{KD}$	70.1	38.3	62.6
$\mathcal{L}_{MSE}$	77.5	44.3	69.6
$\mathcal{L}_{BFD}$	77.6	45.9	70.0

the deviation from 0 (MSE-0). However, MSE-0 couldn’t restrict the background value to absolute 0, leading to noise accumulation and a lower mIoU compared to Background Adaptation *w/* knowledge distillation. Later, we incorporate considerations for the current class region and introduce a Binary Cross-Entropy (BCE) loss (BCE-1). The expectation is to have the value in the current class region as small as possible, while the value in the background class region is as large as possible. Besides, we apply the *Filter* to remove the noise above 0. The BCE-1 scheme with Filter improved mIoU by 2%, but supervision in the background region tends to overfit to background objects. Therefore, we retain the BCE loss in the novel class region and redesign a triplet loss as in Sec. 3.3. The final Background Adaptation losses further enhance the segmentation capability for novel classes, resulting in a 1.2% improvement.

Background Feature Distillation. To validate the effectiveness of Background Feature Distillation loss $\mathcal{L}_{BFD}$ in Sec. 3.3, we compare $\mathcal{L}_{BFD}$ with the method *w/o* feature distillation (*w/o* FD), *w/* standard knowledge distillation ($\mathcal{L}_{KD}$), and *w/* mean square error ($\mathcal{L}_{MSE}$). As shown in Tab. 5, our $\mathcal{L}_{BFD}$ outperforms other representative distillation methods. Particularly, the Background distillation with MSE, i.e., $\mathcal{L}_{BFD}$, shows a 1.6% promotion of mIoU on novel classes with no decline in old classes compared to $\mathcal{L}_{MSE}$. This demonstrates the necessity of preserving a tunnel for learning novel classes.

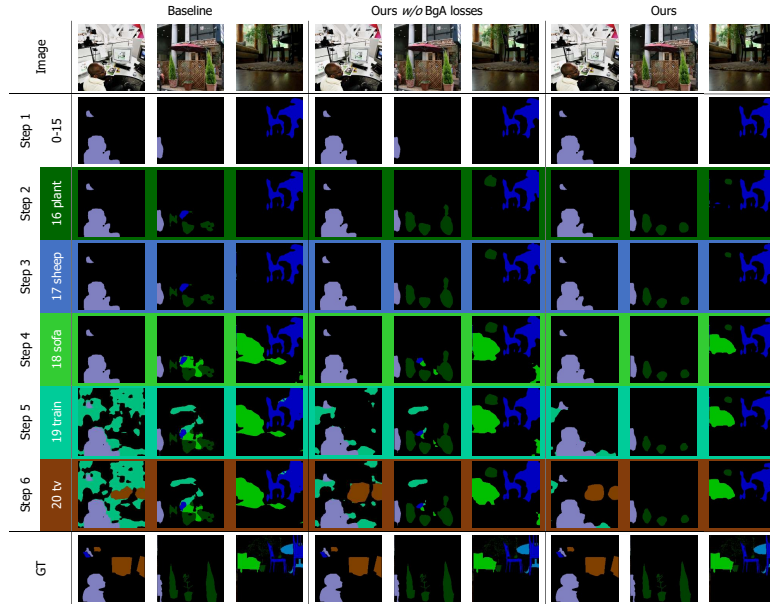


Fig. 4: Qualitative analysis on 15-1 of Pascal VOC 2012. The *Baseline* and *Ours w/o BgA losses* are in the 1st row and the 4th row of Tab. 3. *Ours* applies additional Background Adaptation losses, compared to the *Ours w/o BgA losses*.

4.4 Qualitative Analysis

We’ve conducted a qualitative analysis of our approach with *Baseline* and *Ours w/o BgA losses* in the 1st row and the 4th row of Tab. 3. The results shown in Fig. 4 are visualized from the 15-1 sub-task of VOC 2012. Our approach in the last three columns reduces the misclassification (e.g., *train* in the first two columns of each method) and enhances the precision (e.g., *sofa* in the last column of each method) of the novel classes, while maintaining segmentation ability on the old classes. The results showcase the stability and plasticity of our approach.

5 Conclusions

We have introduced a novel CISS method featuring a Background Adaptation mechanism, designed to prevent the continual shifting of a shared background class and promote the learning of novel classes. The mechanism explicitly models the background residual rather than the background at each step and aggregates these residuals to represent the evolving background. To further enhance the performance of the Background Adaptation, we customized corresponding losses on the adaptation channel and improved the distillation scheme. Experiments on our proposed components verify their effectiveness under the Background Adaptation mechanism. Moreover, our method surpasses previous state-of-the-art methods in most sub-tasks, notably excelling in performance of novel classes.

Acknowledgement

This work was supported by the National Natural Science Foundation of China (No. 61972036) and the Industry-University-Institute Cooperation Foundation of the Eighth Research Institute of China Aerospace Science and Technology Corporation (No. SAST2022-049).

References

1. Ahn, H., Cha, S., Lee, D., Moon, T.: Uncertainty-based continual learning with adaptive regularization. *NeurIPS* **32** (2019) 4
2. Aljundi, R., Lin, M., Goujaud, B., Bengio, Y.: Gradient based sample selection for online continual learning. *NeurIPS* **32** (2019) 4
3. Baek, D., Oh, Y., Lee, S., Lee, J., Ham, B.: Decomposed knowledge distillation for class-incremental semantic segmentation. *NeurIPS* **35**, 10380–10392 (2022) 2, 4, 9, 10, 11, 12
4. Borsos, Z., Mutny, M., Krause, A.: Coresets via bilevel optimization for continual learning and streaming. *NeurIPS* **33**, 14879–14890 (2020) 4
5. Boschini, M., Bonicelli, L., Porrello, A., Bellitto, G., Pennisi, M., Palazzo, S., Spampinato, C., Calderara, S.: Transfer without forgetting. In: *ECCV*. pp. 692–709 (2022) 4
6. Cermelli, F., Mancini, M., Bulò, S.R., Ricci, E., Caputo, B.: Modeling the background for incremental learning in semantic segmentation. In: *CVPR*. pp. 9233–9242 (2020) 2, 4, 9, 10, 11, 12
7. Cha, H., Lee, J., Shin, J.: Co2l: Contrastive continual learning. In: *ICCV*. pp. 9516–9525 (2021) 4
8. Cha, S., Yoo, Y., Moon, T., et al.: Ssul: Semantic segmentation with unknown label for exemplar-based class-incremental learning. *NeurIPS* **34**, 10919–10930 (2021) 2, 3, 4, 6, 7, 10, 13
9. Chen, J., Cong, R., Luo, Y., Ip, H., Kwong, S.: Saving 100x storage: Prototype replay for reconstructing training sample distribution in class-incremental semantic segmentation. *Advances in Neural Information Processing Systems* **36** (2024) 3
10. Chen, L.C., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A.L.: Semantic image segmentation with deep convolutional nets and fully connected crfs. *arXiv preprint arXiv:1412.7062* (2014) 1
11. Chen, L.C., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A.L.: Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE TPAMI* **40**(4), 834–848 (2017) 1
12. Chen, L.C., Papandreou, G., Schroff, F., Adam, H.: Rethinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv:1706.05587* (2017) 1, 4, 10
13. Chen, L.C., Zhu, Y., Papandreou, G., Schroff, F., Adam, H.: Encoder-decoder with atrous separable convolution for semantic image segmentation. In: *ECCV*. pp. 801–818 (2018) 1
14. Chen, Z., Liu, B.: *Lifelong machine learning*, vol. 1. Springer (2018) 2
15. Cheng, B., Misra, I., Schwing, A.G., Kirillov, A., Girdhar, R.: Masked-attention mask transformer for universal image segmentation. In: *CVPR*. pp. 1290–1299 (2022) 1, 4
16. Cheng, B., Schwing, A., Kirillov, A.: Per-pixel classification is not all you need for semantic segmentation. *NeurIPS* **34**, 17864–17875 (2021) 1

17. Cong, W., Cong, Y., Dong, J., Sun, G., Ding, H.: Gradient-semantic compensation for incremental semantic segmentation. *arXiv preprint arXiv:2307.10822* (2023) [12](#)
18. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: *CVPR*. pp. 248–255 (2009) [10](#)
19. Dhar, P., Singh, R.V., Peng, K.C., Wu, Z., Chellappa, R.: Learning without memorizing. In: *CVPR*. pp. 5138–5146 (2019) [4](#)
20. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. In: *ICLR* (2021) [1](#), [4](#)
21. Douillard, A., Chen, Y., Dapogny, A., Cord, M.: Plop: Learning without forgetting for continual semantic segmentation. In: *CVPR*. pp. 4040–4050 (2021) [2](#), [3](#), [4](#), [6](#), [7](#), [9](#), [10](#), [11](#), [12](#)
22. Ermis, B., Zappella, G., Wistuba, M., Rawal, A., Archambeau, C.: Memory efficient continual learning with transformers. *NeurIPS* **35**, 10629–10642 (2022) [4](#)
23. Everingham, M., Van Gool, L., Williams, C.K., Winn, J., Zisserman, A.: The pascal visual object classes (voc) challenge. *IJCV* **88**, 303–338 (2010) [9](#)
24. Fang, Y., Zhu, F., Cheng, B., Liu, L., Zhao, Y., Wei, Y.: Locating noise is halfway denoising for semi-supervised segmentation. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. pp. 16612–16622 (2023) [1](#)
25. Goswami, D., Schuster, R., van de Weijer, J., Stricker, D.: Attribution-aware weight transfer: A warm-start initialization for class-incremental semantic segmentation. In: *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. pp. 3195–3204 (2023) [11](#), [12](#)
26. Hariharan, B., Arbeláez, P., Bourdev, L., Maji, S., Malik, J.: Semantic contours from inverse detectors. In: *ICCV*. pp. 991–998 (2011) [9](#)
27. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: *CVPR*. pp. 770–778 (2016) [1](#), [10](#)
28. Jung, H., Ju, J., Jung, M., Kim, J.: Less-forgetting learning in deep neural networks. *arXiv preprint arXiv:1607.00122* (2016) [2](#)
29. Kirkpatrick, J., Pascanu, R., Rabinowitz, N., Veness, J., et al.: Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences* **114**(13), 3521–3526 (2017) [4](#)
30. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. *NeurIPS* **25** (2012) [1](#)
31. LeCun, Y., Bottou, L., Bengio, Y., Haffner, P.: Gradient-based learning applied to document recognition. *Proceedings of the IEEE* **86**(11), 2278–2324 (1998) [1](#)
32. Li, Z., Hoiem, D.: Learning without forgetting. *IEEE TPAMI* **40**(12), 2935–2947 (2017) [2](#), [4](#)
33. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows. In: *ICCV*. pp. 10012–10022 (2021) [1](#)
34. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. In: *CVPR*. pp. 3431–3440 (2015) [1](#)
35. Lopez-Paz, D., Ranzato, M.: Gradient episodic memory for continual learning. *NeurIPS* **30** (2017) [4](#)
36. Madaan, D., Yoon, J., Li, Y., Liu, Y., Hwang, S.J.: Representational continuity for unsupervised continual learning. In: *ICLR* (2022) [4](#)
37. Maracani, A., Michieli, U., Toldo, M., Zanuttigh, P.: Recall: Replay-based continual learning in semantic segmentation. In: *ICCV*. pp. 7026–7035 (2021) [4](#)

38. McClelland, J.L., McNaughton, B.L., O'Reilly, R.C.: Why there are complementary learning systems in the hippocampus and neocortex: insights from the successes and failures of connectionist models of learning and memory. *Psychological review* **102**(3), 419 (1995) 2, 4
39. McCloskey, M., Cohen, N.J.: Catastrophic interference in connectionist networks: The sequential learning problem. In: *Psychology of learning and motivation*, vol. 24, pp. 109–165. Elsevier (1989) 2, 4
40. Michieli, U., Zanuttigh, P.: Continual semantic segmentation via repulsion-attraction of sparse and disentangled latent representations. In: *CVPR*. pp. 1114–1124 (2021) 2, 4, 10, 11
41. Oh, Y., Baek, D., Ham, B.: Alife: Adaptive logit regularizer and feature replay for incremental semantic segmentation. *NeurIPS* **35**, 14516–14528 (2022) 3, 4, 6, 9
42. Ostapenko, O., Puscas, M., Klein, T., Jahnichen, P., Nabi, M.: Learning to remember: A synaptic plasticity driven framework for continual learning. In: *CVPR*. pp. 11321–11329 (2019) 4
43. Pham, Q., Liu, C., Hoi, S.: Dualnet: Continual learning, fast and slow. *NeurIPS* **34**, 16131–16144 (2021) 4
44. Rebuffi, S.A., Kolesnikov, A., Sperl, G., Lampert, C.H.: icarl: Incremental classifier and representation learning. In: *CVPR*. pp. 2001–2010 (2017) 4
45. Ritter, H., Botev, A., Barber, D.: Online structured laplace approximations for overcoming catastrophic forgetting. *NeurIPS* **31** (2018) 4
46. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: *Proc. International Conference on Medical Image Computing and Computer-Assisted Intervention*. pp. 234–241 (2015) 1
47. Schwarz, J., Czarnecki, W., Luketina, J., Grabska-Barwinska, A., Teh, Y.W., Pascanu, R., Hadsell, R.: Progress & compress: A scalable framework for continual learning. In: *Proc. International Conference on Machine Learning*. pp. 4528–4537. PMLR (2018) 4
48. Shang, C., Li, H., Meng, F., Wu, Q., Qiu, H., Wang, L.: Incrementer: Transformer for class-incremental semantic segmentation with knowledge distillation focusing on old class. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 7214–7224 (2023) 4
49. Shin, H., Lee, J.K., Kim, J., Kim, J.: Continual learning with deep generative replay. *NeurIPS* **30** (2017) 4
50. Shon, H., Lee, J., Kim, S.H., Kim, J.: Dlcft: Deep linear continual fine-tuning for general incremental learning. In: *ECCV*. pp. 513–529 (2022) 4
51. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. In: *ICLR* (2015) 1
52. Strudel, R., Garcia, R., Laptev, I., Schmid, C.: Segmenter: Transformer for semantic segmentation. In: *Proceedings of the IEEE/CVF international conference on computer vision*. pp. 7262–7272 (2021) 4
53. Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A., Jégou, H.: Training data-efficient image transformers & distillation through attention. In: *Proc. International Conference on Machine Learning*. pp. 10347–10357 (2021) 1
54. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. *NeurIPS* **30** (2017) 1
55. Wu, C., Herranz, L., Liu, X., Van De Weijer, J., Raducanu, B., et al.: Memory replay gans: Learning to generate new categories without forgetting. *NeurIPS* **31** (2018) 4

56. Xiao, J.W., Zhang, C.B., Feng, J., Liu, X., van de Weijer, J., Cheng, M.M.: End-points weight fusion for class incremental semantic segmentation. In: CVPR. pp. 7204–7213 (2023) [4](#), [9](#), [11](#), [12](#)
57. Yang, G., Fini, E., Xu, D., Rota, P., Ding, M., Nabi, M., Alameda-Pineda, X., Ricci, E.: Uncertainty-aware contrastive distillation for incremental semantic segmentation. IEEE TPAMI **45**(2), 2567–2581 (2022) [4](#)
58. Zhang, C.B., Xiao, J.W., Liu, X., Chen, Y.C., Cheng, M.M.: Representation compensation networks for continual semantic segmentation. In: CVPR. pp. 7053–7064 (2022) [4](#), [11](#), [12](#)
59. Zhang, G., Wang, L., Kang, G., Chen, L., Wei, Y.: Slca: Slow learner with classifier alignment for continual learning on a pre-trained model. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 19148–19158 (2023) [2](#)
60. Zhang, J.O., Sax, A., Zamir, A., Guibas, L., Malik, J.: Side-tuning: a baseline for network adaptation via additive side networks. In: ECCV. pp. 698–714 (2020) [4](#)
61. Zhang, Z., Gao, G., Fang, Z., Jiao, J., Wei, Y.: Mining unseen classes via regional objectness: A simple baseline for incremental segmentation. NeurIPS **35**, 24340–24353 (2022) [2](#), [4](#), [7](#)
62. Zhang, Z., Gao, G., Jiao, J., Liu, C.H., Wei, Y.: Coinseg: Contrast inter-and intra-class representations for incremental segmentation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 843–853 (2023) [4](#)
63. Zhao, D., Yuan, B., Shi, Z.: Inherit with distillation and evolve with contrast: Exploring class incremental semantic segmentation without exemplar memory. IEEE TPAMI (2023) [4](#), [9](#), [11](#), [12](#)
64. Zhou, B., Zhao, H., Puig, X., Fidler, S., Barriuso, A., Torralba, A.: Scene parsing through ade20k dataset. In: CVPR. pp. 633–641 (2017) [9](#)
65. Zhu, H., Wei, Y., Liang, X., Zhang, C., Zhao, Y.: Ctp: Towards vision-language continual pretraining via compatible momentum contrast and topology preservation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 22257–22267 (2023) [2](#)
66. Zhu, L., Chen, T., Yin, J., See, S., Liu, J.: Continual semantic segmentation with automatic memory sample selection. In: CVPR. pp. 3082–3092 (2023) [4](#)