RBC: Rectifying the Biased Context in Continual Semantic Segmentation

Hanbin Zhao[1]*, Fengyu Yang[4]*, Xinghe Fu[1], and Xi Li[1,2,3]**

[1] College of Computer Science & Technology, Zhejiang University, CHINA
{zhaohanbin, xinghefu, xilizju}@zju.edu.cn
[2] Shanghai Institute for Advanced Study, Zhejiang University, CHINA
[3] Shanghai AI Laboratory, CHINA
[4] University of Michigan, USA
fredyang@umich.edu

Abstract. Recent years have witnessed a great development of Convolutional Neural Networks in semantic segmentation, where all classes of training images are simultaneously available. In practice, new images are usually made available in a consecutive manner, leading to a problem called Continual Semantic Segmentation (CSS). Typically, CSS faces the forgetting problem since previous training images are unavailable, and the semantic shift problem of the background class. Considering the semantic segmentation as a context-dependent pixel-level classification task, we explore CSS from a new perspective of context analysis in this paper. We observe that the context of old-class pixels in the new images is much more biased on new classes than that in the old images, which can sharply aggravate the old-class forgetting and new-class overfitting. To tackle the obstacle, we propose a biased-context-rectified CSS framework with a context-rectified image-duplet learning scheme and a biased-context-insensitive consistency loss. Furthermore, we propose an adaptive re-weighting class-balanced learning strategy for the biased class distribution. Our approach outperforms state-of-the-art methods by a large margin in existing CSS scenarios. Code is available in https://github.com/sntc129/RBC.

Keywords: Continual Semantic Segmentation, Class-incremental Learning, Continual Learning, Biased Context

1 Introduction

Semantic segmentation is a classic pixel-level classification problem in the computer vision area, where deep learning approaches have led to marvelous effect when a large-scale pixel-wise labeled dataset is given [32,51,5,48]. However, in a more practical scenario, deep neural networks are required to learn a sequence of tasks with incremental classes and data which is known as the continual learning

[*] The first two authors contributed equally to this paper.
[**] Corresponding Author

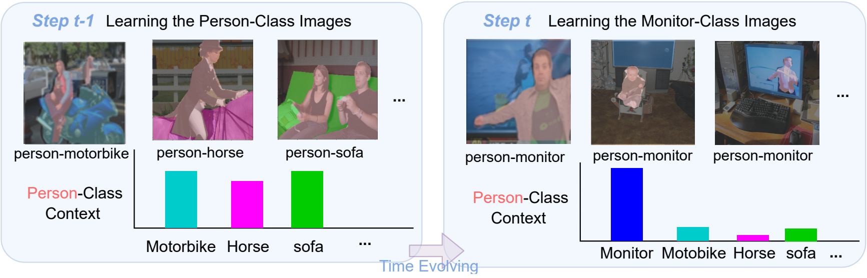


Fig. 1. Illustration of biased context correlation between the old-class and new-class pixels in the added new images. At step $t-1$, the context of *person*-class contains different types (e.g. person-horse, person-sofa) while the model learning person-class images firstly. However, the person context mainly contains person-monitor while the model learning the added *monitor*-class images at step t. Thus, a new-class-biased context for the old-class (person) pixels exists in the added new-class (monitor) images, which aggravates the old-class forgetting and new-class overfitting problem of the model.

setup. Semantic segmentation under the setting of continual learning is referred as Continual Semantic Segmentation (CSS) [3,14,39]. The study of CSS aims at alleviating the forgetting of the network on past tasks and the overfitting on the current task without past data available.

Currently, there are two main challenges in the study of CSS problem. The first challenge is the catastrophic forgetting phenomenon in continual learning [36]. In CSS, the images for past tasks are usually unavailable while the model learning the current task, and only the pixels belonging to new semantic classes are labeled. The model tends to forget the ability to distinguish pixels belonging to old classes due to the shortage of labeled old-class data in the training stage. The second challenge is CSS-specific and called semantic shift of background class [3]. In the current task of CSS, only new-class pixels are labeled as a semantic class and other pixels including old-class pixels are labeled as background class. This semantic shift of pixel-wise labels causes the ambiguous meaning of old-class pixels during the continual learning process and brings an obstacle to the correct model prediction. Since the semantic segmentation is usually considered as a context-dependent pixel-level classification task [10], we explore CSS from the perspective of context. As shown in Figure 1, we find out there is another CSS-specific challenge that has not drawn attention. The context of old-class pixels in the new images is much more biased on new classes than that in the old images, which can cause the sharp aggravation of old-class forgetting and new-class overfitting. We call this challenge “biased context” in CSS.

In the literature, a number of pseudo-labeling-based CSS methods [14,47] attempt to solve the first two main challenges by labeling the mislabeled pixels of old classes with the model obtained from the last learning step (as shown in Figure 2). However, the incrementally updated segmentation model still suffers

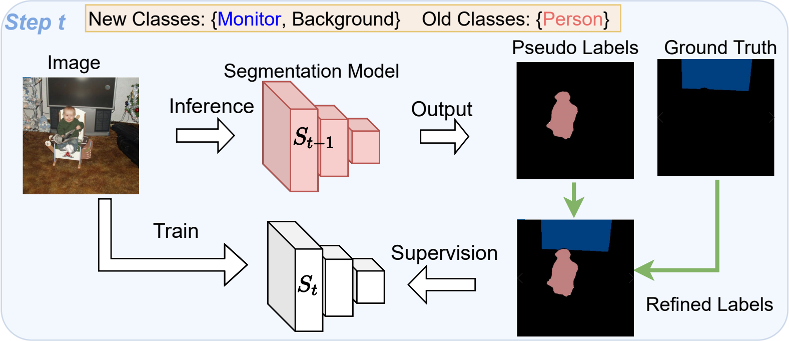


Fig. 2. Pseudo-labeling-based CSS methods. At step t, only the new-class (*monitor*) pixels in the added images are labeled and other pixels are all "background class" pixels. With the old model S_{t-1} from the last step, the mislabeled old-class pixels (*person*) can be pseudo-labeled, and then the model S_t is updated with the old-class pseudo labels and new-class ground truth labels.

from the biased prediction towards new classes on account of the following two observations: 1) the new images contain the new-class-biased context for the old-class pixels, and 2) the number of new-class pixels included in the new task is much larger than that of old-class pixels, which is commonly termed as an imbalanced class distribution problem.

Motivated by the observations above, we try to address the CSS from the following two aspects: 1) building a biased-context-rectified CSS learning scheme that is less sensitive to the biased context information of old-class pixels in the incremental images, and 2) developing a class-balance CSS learning strategy for the imbalanced class distribution at different learning steps. We propose a biased-context-insensitive consistency loss, which resorts to a consistency constraint on the context of old classes in an image pair. The duplet of images, consisting of the original image (containing the new-class pixels) and the corresponding erased image (erasing the new-class pixels in the original image), rectify the context of old classes with respect to new classes. Furthermore, we propose an adaptive class-balance CSS learning strategy to cope with the biased class distribution, which adaptively assigns higher weights to the old-class pixels.

Overall, the main contributions of this paper are three-fold: (1) We first consider the biased context in the CSS scenario and propose a biased-context-rectified CSS framework, which aims to avoid overfitting on new classes while not forgetting old classes. (2) We design a novel context-rectified image-duplet learning scheme and a biased-context-insensitive consistency loss that ingeniously rectifies the context of old classes with respect to new classes. To cope with the imbalanced class distribution, we propose an adaptive re-weighting class-balanced learning strategy for CSS. (3) Extensive experiments demonstrate the effectiveness of our method. Our method outperforms several previous CSS approaches by a large margin and obtains state-of-the-art performance.

2 Related Work

Continual Learning. The last years have seen great interest in continual learning (i.e. also called incremental learning or lifelong learning) [7]. Continual learning is first explored on the image classification task with the catastrophic forgetting problem. These are three major families of works: 1) architectural methods, 2) rehearsal methods, and 3) regularization methods. Architectural methods [29,54,34,33,2,1] adjust the network architecture to maintain the learned knowledge from old tasks and acquire new information from the current task. Rehearsal methods [41,19,53,46,31] replay the knowledge of old tasks when learning the new task, and the old knowledge is memorized by storing previous tasks' exemplars or the distribution of old tasks data via generative models. Regularization methods [30,9,13] alleviate forgetting by regularization loss terms enabling the updated parameters of networks to retain past knowledge. Continual learning is usually conducted under two scenarios (task-incremental or the class-incremental learning). The latter is more challenging because the task identity is unavailable at inference time. Recently, continual learning has been also explored on several other computer vision tasks, *e.g.*, incremental object detection [25], incremental video classification [52], incremental instance segmentation [17], continual semantic segmentation [3,14,47,37,16,49,35,20,56,43,27], incremental domain adaptation [28,11,12]. Our work focuses on the CSS problem which can be considered as the class-incremental learning scenario on semantic segmentation. Exploring the imbalanced class distribution problem is important for continual learning (e.g., the methods [19,46] are proposed to address the problem in classic class-incremental image classification scenario) and our work utilizes an adaptive re-weighting class-balanced learning strategy to alleviate this problem in CSS scenario.

Continual Semantic Segmentation. The forgetting problem in CSS is first considered in ILT [37] and the more challenging CSS-tailored problem (background shift) is proposed in MiB [3]. To cope with the problems, some regularization based CSS methods [14,38] utilizes a confidence-based pseudo-labeling method and a feature-based multi-scale pooling distillation scheme or employs a prototype consistency constraint at the latent space, and some replay-based CSS methods [35] utilize an extra memory to replay the data for old classes by an extra generative adversarial network or web crawling process. Semantic segmentation is a pixel-wise classification problem [6,42,22,32,48,44,40,24] and classifying a local pixel with context information is helpful for reducing the local ambiguities [55,51,45,21,23]. Our work first analyzes the effect of biased context in CSS, and we design several biased-context-rectified continual learning strategies tailored for CSS problem.

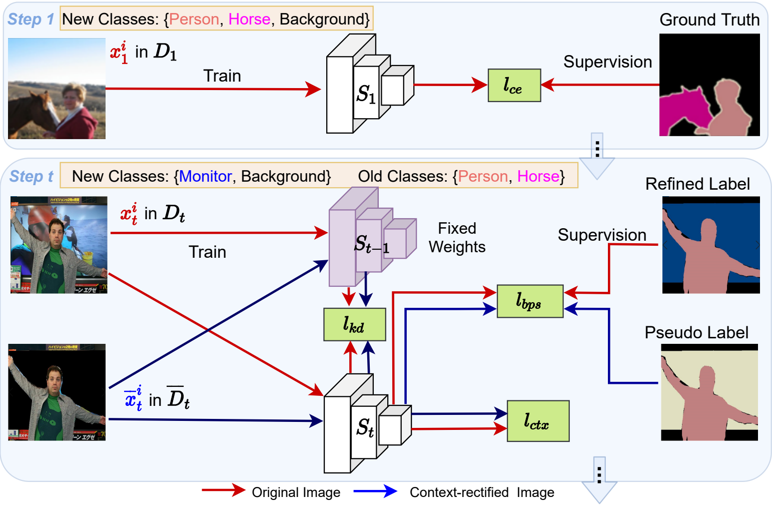


Fig. 3. Illustration of our biased-context-rectified CSS framework. At step 1, the semantic segmentation model is trained from scratch via the classic cross-entropy loss l_{ce} on D_1. At the latter steps (e.g. step t), we first obtain the context-rectified image-duplet $(D_t, \overline{D}_t)$ and update the model by our context-rectified image-duplet learning scheme with the balanced pseudo-labeling loss l_{bps} and the distillation loss l_{kd} and our biased-context-insensitive consistency loss l_{ctx}.

3 Method

3.1 CSS Problem Formulation

In a continual semantic segmentation scenario, a segmentation model learns several image segmentation tasks continually, and the image subset in each learning step contains pixels from one or several new classes [3,14,39]. We suppose the training image set for the t-th learning step is D_t that consists of a set of pairs $(\mathbf{x}_t^i, \mathbf{y}_t^i)$, where $\mathbf{x}_t^i \in \mathbb{R}^{H\times W\times 3}$ and $\mathbf{y}_t^i \in \tilde{\mathcal{Y}}_t^{H\times W}$ denote the i-th input image of size $W\times H$ and the corresponding ground truth segmentation mask, respectively. New categories C_t are introduced and required to be learnt at the t-th step. $\mathbf{y}_t^i$ only contains the labels of C_t and all other labels (e.g., old classes $C_{1:t-1}$) are collapsed into the background class C_0.

We assume a typical semantic segmentation model S with parameters Θ, which consists of an encoder-decoder backbone network F extracting a dense feature map and a convolution head G producing the segmentation score map. Classically, we utilize $S(\mathbf{x}) = G \cdot F(\mathbf{x})$ to represent the output predicted segmentation mask of $\mathbf{x}$, $S^{w,h,c}(\mathbf{x})$ denotes the prediction score (about the class c) of the pixel at the location (w, h) of $\mathbf{x}$, and $\hat{S}(\mathbf{x}) = \text{Softmax}\left(S(\mathbf{x})\right)$ denotes the output of the network. Then, S_t with parameters Θ_t is updated on D_t at the t-th step. Our goal is to obtain S_t which performs well on both previously seen classes $C_{1:t-1}$ and the current classes C_t. CSS task is faced with three dilemmas: 1) S_t is only updated on D_t without the previously seen data $D_{1:t-1}$ and suffers from a significant performance drop on pixels of old classes (i.e., the catastrophic forgetting problem); 2) some of pixels in $\mathbf{x}_t^i$ of D_t are mislabeled as C_0 but actually belong to $C_{1:t-1}$ (i.e., the background shift problem); 3) the context for the old-class pixels in D_t is biased to new classes, since the new-class pixels are usually dominant in the added images D_t.

To address the first two issues, pseudo-labeling CSS methods [14,47] are proposed by labeling the mislabeled pixels with the model obtained from the last step, which is described in Section 3.2. These methods alleviate the forgetting problem since a few pixels of old classes are introduced during learning the new images (similar to the replay-based continual learning strategy [41,46,19]), and reduce the background shift due to correcting the mislabeled "background class" pixels. However, the updated segmentation model by these methods still suffers from the biased prediction towards new classes because of the following two observations: the biased context (shown in Figure 1) and the common imbalanced class distribution in the new images. To alleviate the above issues, we propose a biased-context-rectified CSS framework including a context-rectified image-duplet learning scheme and a biased-context-insensitive consistency loss in Section 3.3 and the illustration of our framework is shown in Figure 3 and propose an adaptive class-balance strategy for tackle the biased class distribution in Section 3.4.

3.2 Pseudo-Labeling-Based CSS

To alleviate the forgetting and background shift problems, pseudo-labeling-based methods [14,47] are utilized in CSS. Specifically, at the t-th learning step, we can access to S_{t-1} from the last step and correct the mislabeled "background class" pixels with S_{t-1} (as shown in Figure 2). For each $(\mathbf{x}_t^i, \mathbf{y}_t^i)$ in D_t, the pixels belonging to the new classes C_t have ground-truth labels and some of the other pixels belonging to the old classes $C_{1:t-1}$ are mislabeled as C_0. The predictions of the old model for these mislabeled pixels $\hat{S}_{t-1}(\mathbf{x}_t^i)$ are utilized as clues if they belong to any of the old classes. After that, each $\mathbf{x}_t^i$ in D_t can have a refined segmentation label $\tilde{S}_t(\mathbf{x}_t^i)$ by combining the pseudo label $\hat{S}_{t-1}(\mathbf{x}_t^i)$ and the ground truth $\mathbf{y}_t^i$ (as shown in Figure 2). Then the model S_t is updated by optimizing the following objective function:

$$\mathcal{L}_{total}(\Theta_t) = \frac{1}{|D_t|} \sum_{(\mathbf{x},\mathbf{y}) \in D_t} l(\mathbf{x}; \Theta_t), \tag{1}$$

where $l(\mathbf{x}; \Theta_t)$ is usually composed of a cross-entropy loss term with pseudo-labeling and a knowledge distillation term:

$$l(\mathbf{x}; \Theta_t) = l_{ps}(\mathbf{x}; \Theta_t) + \alpha l_{kd}(\mathbf{x}; \Theta_t), \tag{2}$$

where α is a hyper-parameter balancing the importance of the loss terms. $l_{ps}(\mathbf{x}; \Theta_t)$ is utilized to maintain the performance on old classes and reduce the ambiguity of old-class pixels labeled as background class at step t:

$$l_{ps}(\mathbf{x};\Theta_t) = -\frac{\beta}{WH}\sum_{w,h}^{W,H}\sum_{c\in C_{0:t}} \tilde{S}_t^{w,h,c}(\mathbf{x}) \log \hat{S}_t^{w,h,c}(\mathbf{x}), \tag{3}$$

where β is the ratio of accepted old classes pixels over the total number of such pixels. $l_{kd}(\mathbf{x};\Theta_t)$ is added to the backbone network $F(\cdot)$ to retain information of the old classes:

$$l_{kd}(\mathbf{x};\Theta_t) = \left\|\Phi(F_t(\mathbf{x})) - \Phi(F_{t-1}(\mathbf{x}))\right\|^2, \tag{4}$$

where $\|\cdot\|$ and $\Phi(F(\mathbf{x})) \in \mathbb{R}^{(H+W)\times C}$ denotes the Euclidean distance and concatenation operation, respectively. The concatenation operation function $\Phi(F(\mathbf{x}))$ is formulated as follows:

$$\Phi(F(\mathbf{x})) = \left[\frac{1}{W}\sum_{w=1}^{W} F^{:,w,:}(\mathbf{x}) \big|\big| \frac{1}{H}\sum_{h=1}^{H} F^{h,:,:}(\mathbf{x})\right], \tag{5}$$

where $[\cdot||\cdot]$ denotes concatenation over the channel axis.

3.3 Biased-context-rectified framework

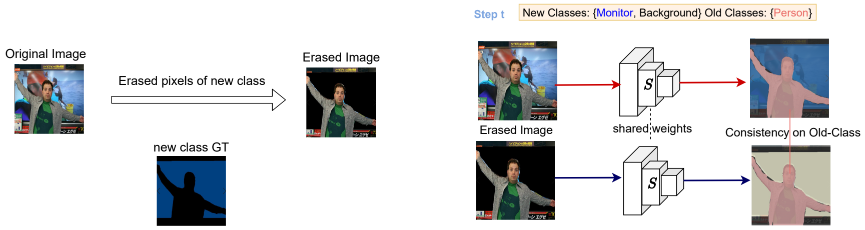
To alleviate the biased context correlation between the old-class and new-class pixels in CSS, we propose a biased-context-rectified framework with a context-rectified image-duplet learning scheme and a biased-context-insensitive consistency loss. Taking the t-th step as an example (shown in Figure 1), the incrementally added images D_t mainly contain the new-class-related context for the old-class pixels, which leads to the aggravation of the old-class forgetting and new-class overfitting problems.

Context-rectified Image-Duplet Learning. As for the new-class-related context, we observe that the contextual information of old-class pixels included in the incremental images is biased to the pixels of new classes (shown in Figure 1). In order to continually learn a semantic segmentation model that is less sensitive to the entangled new-class-context, we firstly rectify the biased context between new classes and old classes in these new images by erasing the new-class pixels of the original image (shown in Figure 4(a)). At the t-th step, we obtain the corresponding erased image $\overline{\mathbf{x}}_t^i$ for each new image $\mathbf{x}_t^i$ in D_t. Then an image-duplet $(\mathbf{x}_t^i, \mathbf{y}_t^i, \overline{\mathbf{x}}_t^i, \overline{\mathbf{y}}_t^i)$ is constructed from the erased image and the corresponding original image. The set of image-duplets with D_t and $\overline{D}_t$ are denoted as:

$$\begin{aligned} &(D_t, \overline{D}_t) = \left\{(\mathbf{x}_t^i, \mathbf{y}_t^i, \overline{\mathbf{x}}_t^i, \overline{\mathbf{y}}_t^i)\right\}_{i=1}^{|D_t|} \\ &s.t.\ (\mathbf{x}_t^i, \mathbf{y}_t^i) \in D_t, (\overline{\mathbf{x}}_t^i, \overline{\mathbf{y}}_t^i) \in \overline{D}_t, \end{aligned} \tag{6}$$

With the image-duplets $(D_t, \overline{D}_t)$ at the t-th step, our method updates the model S_t by optimizing the following loss function:

$$\mathcal{L}_{total}(\Theta_t) = \frac{1}{|D_t|+|\overline{D}_t|}\sum_{(\mathbf{x},\overline{\mathbf{x}})}[l_{dup}(\mathbf{x},\overline{\mathbf{x}};\Theta_t) + \gamma l_{ctx}(\mathbf{x},\overline{\mathbf{x}};\Theta_t)], \tag{7}$$



(a) original (b) original

Fig. 4. The illustration of (a): generating the context-rectified image-duplet, (b): a biased-context-insensitive consistency loss.

where γ is a hyper-parameter balancing the importance of the loss terms. The first loss term $l_{dup}(\mathbf{x}, \overline{\mathbf{x}}; \Theta_t)$ takes the similar form of Equation (2) on the original image $\mathbf{x}$ and the corresponding erased image $\overline{\mathbf{x}}$:

$$l_{dup}(\mathbf{x}, \overline{\mathbf{x}}; \Theta_t) = l(\mathbf{x}; \Theta_t) + l(\overline{\mathbf{x}}; \Theta_t), \tag{8}$$

Biased-context-insensitive Consistency Loss. To further address the biased context, the second loss term $l_{ctx}(\mathbf{x}, \overline{\mathbf{x}}; \Theta_t)$ is introduced and utilized to keep a biased-context-insensitive consistency between the original image $\mathbf{x}$ and the corresponding erased image $\overline{\mathbf{x}}$ (as shown in Figure 4(b)). For the old-class pixels, the new-class-related context are included in the original image $\mathbf{x}$ and erased in the corresponding erased image $\overline{\mathbf{x}}$. For simplicity, we utilize $O(\mathbf{x})$ to represent the locations $\{(w_o^j, h_o^j)\}_{j=1}^{O(\mathbf{x})}$ of old-class pixels included in the image $\mathbf{x}$, . To reduce the effect of biased context between the old-class and new-class, the prediction of the updated model S_t on the old-class pixels with the new-class-related context should be consistent with that without the new-class-related context. Then $l_{ctx}(\mathbf{x}, \overline{\mathbf{x}}; \Theta_t)$ is formulated as follows:

$$l_{ctx}(\mathbf{x}, \overline{\mathbf{x}}; \Theta_t) = \sum_{(w,h) \in O(\mathbf{x})} \sum_{c \in C_{1:t-1}} \left\| S_t^{w,h,c}(\overline{\mathbf{x}}) - S_t^{w,h,c}(\mathbf{x}) \right\|^2, \tag{9}$$

3.4 Adaptive Class-Balance CSS

As for the imbalanced class distribution problem, we observe that the number of new-class pixels included in the new images is much larger than that of pseudo-labeled old-class pixels. This class-imbalance problem usually results in the updated classifier being biased towards the new classes. To cope with the problem, we propose to adaptively assign different weights to the pixels of different classes based on the number of pixels. We optimize the biased classifier by a balanced pseudo-labeling cross-entropy loss $l_{bps}(\mathbf{x}; \Theta_t)$ with different weights.

To address the class-imbalance problem in CSS, the balanced pseudo-labeling cross-entropy loss is formulated as follows:

$$l_{bps}(\mathbf{x};\Theta_t) = -\frac{\beta}{WH}\sum_{w,h}^{W,H}\sum_{c\in C_{0:t}} \eta^{w,h}(\mathbf{x})\tilde{S}_t^{w,h,c}(\mathbf{x})\log \hat{S}_t^{w,h,c}(\mathbf{x}), \tag{10}$$

where $\eta^{w,h}(\mathbf{x})$ denotes the weight of the pixel at the location (w,h) in the image $\mathbf{x}$. $\eta^{w,h}(\mathbf{x})$ depends on the category of the pixel and the number of pixels from different classes in the image:

$$\eta^{w,h}(\mathbf{x}) = \begin{cases} 0.5 + \sigma(\frac{N^{old}(\mathbf{x})}{N^{new}(\mathbf{x})}) & {\scriptstyle (w,h)\in O(\mathbf{x})} \\ 1 & \text{otherwise,} \end{cases} \tag{11}$$

where $N^{old}(\mathbf{x})$, $N^{new}(\mathbf{x})$ and $\sigma(\cdot)$ are the number of pixels belonging to old classes $C_{1:t-1}$, the total number of pixels belonging to the new classes C_t and the sigmoid function respectively. Then $l_{dup}(\mathbf{x},\overline{\mathbf{x}};\Theta_t)$ in Equation (8) with the balanced pseudo-labeling is formulated as follows:

$$l_{dup}(\mathbf{x},\overline{\mathbf{x}};\Theta_t) = l^{'}(\mathbf{x};\Theta_t) + l(\overline{\mathbf{x}};\Theta_t), \tag{12}$$

where $l^{'}(\mathbf{x};\Theta_t)$ is denoted as follows:

$$l^{'}(\mathbf{x};\Theta_t) = l_{bps}(\mathbf{x};\Theta_t) + \alpha l_{kd}(\mathbf{x};\Theta_t), \tag{13}$$

4 Experiments

4.1 Datasets

We follow previous CSS works [35,3,14] and utilize the commonly used semantic segmentation datasets PASCAL VOC 2012 [15] and ADE20k [57] for experiments: VOC contains $10,582$ fully-annotated images for training and $1,449$ for testing, over 20 foreground object classes. ADE20k is a large-scale dataset that has $20,210$ training images and $2,000$ testing images in 150 classes. For all datasets, we resize the images to 512×512, with a center crop and employ the random horizontal flip augmentation strategy as the practice in PLOP [14] at training time.

4.2 Experimental Setup

Continual Semantic Segmentation Setting: MiB [3] introduces two different CSS settings (*Disjoint* and *Overlapped*). In the *Disjoint* setting, the incremental new images D_t at t-th step contain pixels belonging to old and current new classes $(C_{1:t-1} \cup C_t)$, each training step contains a unique set of images, whose pixels belong to classes seen either in the current or in the previous learning steps. In the *Overlapped* setting, the new images contains the pixels belonging

Table 1. CSS results under the *Disjoint* setting on VOC-19-1, VOC-15-5 and VOC-15-1 benchmarks. † means the results from [14,38]. Best in **bold**.

Method	**19-1** (2 steps)			**15-5** (2 steps)			**15-1** (6 steps)		
	0-19	20	*all*	0-15	16-20	*all*	0-15	16-20	*all*
FT	5.80	12.30	6.20	1.10	33.60	9.20	0.20	1.80	0.60
PI† [50]	5.40	14.10	5.90	1.30	34.10	9.50	0.00	1.80	0.40
EWC† [26]	23.20	16.00	22.90	26.70	37.70	29.40	0.30	4.30	1.30
RW† [4]	19.40	15.70	19.20	17.90	36.90	22.70	0.80	3.60	1.30
LwF† [30]	53.00	9.10	50.80	58.40	37.40	53.10	0.80	3.60	1.50
LwF-MC† [41]	63.00	13.20	60.50	67.20	41.20	60.70	4.50	7.00	5.20
ILT† [37]	69.10	16.40	66.40	63.20	39.50	57.30	3.70	5.70	4.20
MiB† [3]	69.60	25.60	67.40	71.80	43.30	64.70	46.20	12.90	37.90
SDR† [38]	69.90	37.30	68.40	73.50	47.30	67.20	59.20	12.90	48.10
PLOP† [14]	75.37	38.89	73.64	71.00	42.82	64.29	57.86	13.67	46.48
Ours	**76.43**	**45.79**	**75.01**	**75.12**	**49.71**	**69.89**	**61.68**	**19.52**	**51.60**
Joint	77.40	78.00	77.40	79.10	72.56	77.39	79.10	72.56	77.39

to old, current new and future classes ($C_{1:t-1} \cup C_t \cup C_{t+1:T}$), each step contains all the images that have at least one pixel of a novel class, with only the latter annotated. The *Overlapped* setting is usually more challenging than the *Disjoint* setting.

Evaluation Protocol: We evaluate our method under these two CSS settings on the commonly used CSS benchmarks (VOC-19-1, VOC-15-5, VOC-15-1, ADE-100-50, ADE-50-50 and ADE-100-10), where 19-1 means learning 19 then 1 class (2 learning steps), 15-5 learning 15 then 5 classes (2 steps) and 15-1 learning 15 classes followed by five times 1 class (6 steps). The benchmarks on ADE20k are 100-50 (2 steps), 50-50 (3 steps) and 100-10 (6 steps). The benchmark with higher number of steps is usually more challenging. Each method is trained on the CSS benchmark in several steps. At the last step, we follow [14,38] and report the traditional mean Intersection over Union (mIoU) for the initial classes C_1, for the incremented classes $C_{2:T}$, for all classes $C_{1:T}$ (*all*).

Training Details: We implement our models with Pytorch and use SGD for optimization. Following [3,14], we use the Deeplab-V3 [5] architecture with a ResNet-101 [18] pre-trained on ImageNet [8] as the backbone network. As for our proposed context-rectified image-duplet learning scheme, we train our model with a batch size of 24 on both Pascal VOC and ADE20k datasets. At the first CSS step, all the images are the original images since no old model is kept for pseudo-labeling and the learning rate is set to 0.01 on both VOC and ADE20k CSS benchmarks. At other CSS steps, the sample duplets are generated by the old model from the last step (half of the images in each batch are the original images and half of them are the corresponding new-class-erased images). The learning rate on VOC/ADE20k is set to 0.001/0.005. The loss weight γ of the biased-context-insensitive consistency loss term in Equation (7) is set to 0.01 for all datasets. More experimental results are included in the supplementary materials.

Table 2. CSS results under the *Overlapped* setting on VOC-19-1, VOC-15-5 and VOC-15-1 benchmarks. † means the results from [14,38]. Best in **bold**.

Method	**19-1** (2 steps)			**15-5** (2 steps)			**15-1** (6 steps)		
	0-19	20	*all*	0-15	16-20	*all*	0-15	16-20	*all*
FT	6.80	12.90	7.10	2.10	33.10	9.80	0.20	1.80	0.60
PI† [50]	7.50	14.00	7.80	1.60	33.30	9.50	0.00	1.80	0.50
EWC† [26]	26.90	14.00	26.30	24.30	35.50	27.10	0.30	4.30	1.30
RW† [4]	23.30	14.20	22.90	16.60	34.90	21.20	0.00	5.20	1.30
LwF† [30]	51.20	8.50	49.10	58.90	36.60	53.30	1.00	3.90	1.80
LwF-MC† [41]	64.40	13.30	61.90	58.10	35.00	52.30	6.4	8.40	6.90
ILT† [37]	67.75	10.88	65.05	67.08	39.23	60.45	8.75	7.99	8.56
MiB† [3]	71.43	23.59	69.15	76.37	49.97	70.08	34.22	13.50	29.29
SDR† [38]	69.10	32.60	67.40	75.40	52.60	69.90	44.70	21.80	39.20
PLOP† [14]	75.35	37.35	73.54	75.73	51.71	70.09	65.12	21.11	54.64
Ours	**77.26**	**55.60**	**76.23**	**76.59**	**52.78**	**70.92**	**69.54**	**38.44**	**62.14**
Joint	77.40	78.00	77.40	79.10	72.56	77.39	79.10	72.56	77.39

4.3 Comparison to State-of-the-Art Methods

In this section, we evaluate the CSS performance of our proposed method on Pascal VOC and ADE20k datasets, against existing state-of-the-art methods, including PI [50], EWC [26], RW [4], LwF [30], LwF-MC [41], ILT [37], MiB [3] , SDR [38] and PLOP [14]. In the tables, we also provide the results of the other two methods: the simple fine-tuning approach which trains the model on the new images with no additional constraints (denoted by "FT"), and training the model on all classes off-line (denoted by "Joint"). The former can be regarded as a lower limit and the latter as an upper limit.

Results on Pascal VOC. Table 1 and Table 2 summarizes the experimental results for the *Disjoint* and *Overlapped* settings of three VOC benchmarks respectively. Under the *Disjoint* setting, it is observed that the performance of our method consistently surpasses the other methods at the last learning step on each evaluated benchmark. On VOC-19-1, we can see that the mIOU of our method on new classes (20) is 6.90% higher than that of PLOP. On the VOC-15-1 with a large number of learning steps, our method consistently performs better than other methods. All of these results indicate the effectiveness of our method to catastrophic forgetting of past classes and overfitting on the current classes. Under the *Overlapped* setting, we can see that the performance of our method consistently outperforms that of other methods by a sizable margin on all evaluated VOC benchmarks (i.e., 19-1, 15-5 and 15-1). On VOC-19-1, the forgetting of old classes (1-19) is reduced by 1.91% while performance on new classes is greatly improved by 18.25%. On the most challenging benchmark VOC-15-1, it is worth noting that the performance of our method on all the seen classes outperforms its closest contender PLOP [14] by around 7.50%. All of these results indicate the effectiveness of our method to catastrophic forgetting of past classes and overfitting on the current classes.

Table 3. CSS results under the *Overlapped* setting on ADE-100-50, ADE50-50 and ADE-100-10 benchmarks. $*$ means the results from re-production.

Method	**100-50** (2 steps)			**50-50** (3 steps)			**100-10** (6 steps)		
	0-100	101-150	*all*	0-50	51-150	*all*	0-100	101-150	*all*
FT	0.00	22.50	7.50	13.90	12.00	12.60	0.00	2.50	9.20
ILT† [37]	18.29	14.40	17.00	3.53	12.85	9.70	0.11	3.06	1.09
MiB† [3]	40.52	17.17	32.79	45.57	21.01	29.31	38.21	11.12	29.24
PLOP* [14]	41.66	15.42	32.97	47.75	21.60	30.43	**39.42**	13.63	30.88
Ours	**42.90**	**21.49**	**35.81**	**49.59**	**26.32**	**34.18**	39.01	**21.67**	**33.27**
Joint	43.90	27.20	38.30	50.90	32.10	38.30	43.90	27.20	38.30

Table 4. Ablation experimental results on VOC-*Overlapped*-15-1.

Ablation	**Method**	**15-1** (6 steps)		
		0-15	16-20	*all*
Ablation I	Baseline	65.12	21.11	54.64
	Baseline+double	60.23	11.95	48.73
	Baseline+duplet	**70.54**	**31.06**	**61.14**
Ablation II	Baseline+duplet	**70.54**	31.06	61.14
	Baseline+duplet+ctx	69.54	**38.44**	**62.14**
Ablation III	Baseline	65.12	21.11	54.64
	Baseline+balance	**65.35**	**24.89**	**55.72**

Results on ADE20k. We have also evaluated our method under the *Overlapped* setting of ADE-100-50, ADE-50-50 and ADE-100-10 benchmarks and the results are shown in Table 3. This dataset is very hard because the mIoU of the joint model is only 38.30%. On these ADE CSS benchmarks, our method improves the mIoU on new classes by a sizable margin (more than 4.5%) and shows comparable performance on previous classes with its closest contender PLOP. The overview on the performance of new classes reveals that our approach is greatly helpful to avoid the overfitting on new classes while maintaining the performance on previous classes.

4.4 Ablation Study

In this section, we first carry out ablation experiments to validate the effectiveness of the context-rectified image-duplet. Then we conduct experiments to validate our biased-context-insensitive consistency loss and adaptive class-balance strategy. All of the ablation experiments are conducted on the challenging *Overlapped* setting of the benchmark VOC-15-1.

Baseline. Our main baseline is given based on a classical pseudo-labeling-based CSS method PLOP [14], which utilizes a multi-scale pooling distillation scheme to preserve the performance on previously seen classes and an entropy-based pseudo-labeling strategy on the mislabeled background class pixels to reduce the background shift.

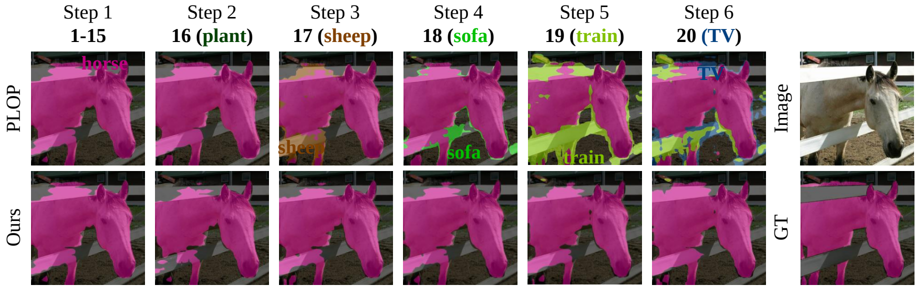


Fig. 5. Visualization of PLOP and our method at different steps under the *Overlapped* setting of VOC-15-1.

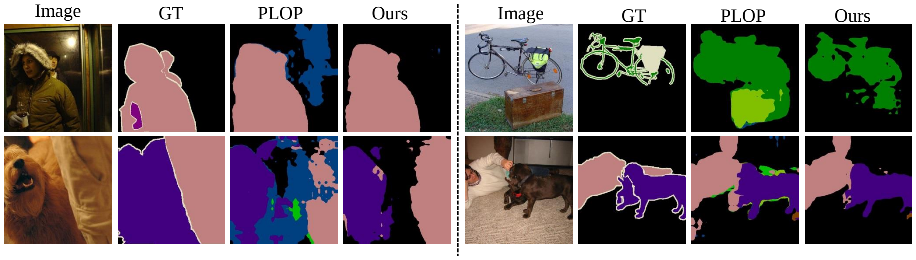


Fig. 6. The predictions of PLOP and our method for different images at the last step under the *Overlapped* setting of VOC-15-1.

Effect of the context-rectified image-duplet. In order to demonstrate the effectiveness of the context-rectified image-duplet, we compare the performance of "Baseline" with our image-duplet including the original image and the corresponding new-class-erased image (denoted by "Baseline+duplet"). The results are shown in Table 4. We can see that the performance of Baseline+duplet surpasses that of Baseline. In particular, the mIoU on new classes of Baseline+duplet outperforms that of Baseline by a large margin (9.95%). Moreover, we also compare Baseline+duplet with Baseline+double to reduce the influence of increasing the number of samples. In a minibatch, Baseline+duplet utilizes the original images and the corresponding erased images. Baseline+double utilizes the original images and the corresponding copied original images. As shown in Table 4, the performance of Baseline+duplet is higher than Baseline+double, which demonstrates that directly increasing the number of images can not lead to performance improvement.

Effect of biased-context-insensitive consistency constraint. We evaluate the performance of "Baseline+duplet" with our biased-context-insensitive consistency loss (denoted by "Baseline+duplet+ctx") and Table 4 summarizes the experimental results on the *Overlapped* setting of the benchmark VOC-15-1.

The mIoU on new classes of Baseline+duplet+ctx is around 7.38% higher than that of Baseline+duplet, which demonstrates that the biased-context-insensitive consistency constraint can greatly improve the performance on new classes and is essential to avoid the overfitting on new classes. To further validate the effectiveness of our biased-context-rectified CSS learning framework, we also compare our method with Baseline and show the average mIoU curves in Figure 7. It is observed that Ours achieves better performance than Baseline at every step.

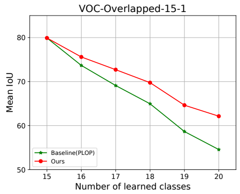


Fig. 7. The mIoU evolution of ours and Baseline(PLOP) on VOC-15-1.

Effect of adaptive class-balance strategy. To demonstrate the effectiveness of our adaptive class-balance strategy, we evaluate the performance of Baseline with our adaptive class-balance strategy (Baseline+balance). In Table 4, we report the experimental results after the last learning step. As for the old classes (0-15), Baseline+balance achieves better performance than Baseline. Regarding new classes (16-20), Baseline+balance exceeds Baseline by around 4%.

Effect of biased context. To demonstrate the effect of the biased context, we visualize the predictions for both PLOP (Baseline) and our method on 15-1 protocol of the benchmark VOC-*Overlapped*. As shown in Figure 5, PLOP is more prone to overfitting on new classes (sheep, sofa, train, TV) than ours at the latter steps. Besides, we visualize the predictions of ours and PLOP for different samples at the last step in Figure 6. Ours achieve less forgetting on old classes (person, dog, bicycle) than PLOP, illustrating that the biased context aggravates the old-class forgetting and new-class overfitting.

5 Conclusion

In this paper, we first consider the biased context problem in CSS and design a novel biased-context-rectified CSS framework for it. Firstly, our method utilizes a context-rectified image-duplet learning scheme and a biased-context-insensitive consistency loss to rectify the biased context correlation between the old-class pixels and new-class pixels, which effectively alleviates the old-class forgetting and new-class overfitting. Secondly, we propose an adaptive re-weighting class-balanced learning strategy to cope with the dynamiclly changing imbalanced class distribution in CSS. Lastly, we perform intensive evaluations of our method and other CSS methods, showing the effectiveness of our method.

6 Acknowledgments

This work is supported in part by National Key Research and Development Program of China under Grant 2020AAA0107400, Zhejiang Provincial Natural Science Foundation of China under Grant LR19F020004, National Natural Science Foundation of China under Grant U20A20222.

References

1. Abati, D., Tomczak, J., Blankevoort, T., Calderara, S., Cucchiara, R., Bejnordi, B.E.: Conditional channel gated networks for task-aware continual learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3931–3940 (2020)
2. Aljundi, R., Chakravarty, P., Tuytelaars, T.: Expert gate: Lifelong learning with a network of experts. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 3366–3375 (2017)
3. Cermelli, F., Mancini, M., Bulo, S.R., Ricci, E., Caputo, B.: Modeling the background for incremental learning in semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9233–9242 (2020)
4. Chaudhry, A., Dokania, P.K., Ajanthan, T., Torr, P.H.: Riemannian walk for incremental learning: Understanding forgetting and intransigence. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 532–547 (2018)
5. Chen, L.C., Papandreou, G., Schroff, F., Adam, H.: Rethinking atrous convolution for semantic image segmentation. arXiv preprint arXiv:1706.05587 (2017)
6. Chen, L.C., Zhu, Y., Papandreou, G., Schroff, F., Adam, H.: Encoder-decoder with atrous separable convolution for semantic image segmentation. In: Proceedings of the European conference on computer vision (ECCV). pp. 801–818 (2018)
7. Delange, M., Aljundi, R., Masana, M., Parisot, S., Jia, X., Leonardis, A., Slabaugh, G., Tuytelaars, T.: A continual learning survey: Defying forgetting in classification tasks. IEEE Transactions on Pattern Analysis and Machine Intelligence (2021)
8. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: 2009 IEEE conference on computer vision and pattern recognition. pp. 248–255. IEEE (2009)
9. Dhar, P., Singh, R.V., Peng, K.C., Wu, Z., Chellappa, R.: Learning without memorizing. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5138–5146 (2019)
10. Ding, H., Jiang, X., Shuai, B., Liu, A.Q., Wang, G.: Context contrasted feature and gated multi-scale aggregation for scene segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 2393–2402 (2018)
11. Dong, J., Cong, Y., Sun, G., Fang, Z., Ding, Z.: Where and how to transfer: Knowledge aggregation-induced transferability perception for unsupervised domain adaptation. IEEE Transactions on Pattern Analysis and Machine Intelligence pp. 1–1 (2021). https://doi.org/10.1109/TPAMI.2021.3128560
12. Dong, J., Cong, Y., Sun, G., Zhong, B., Xu, X.: What can be transferred: Unsupervised domain adaptation for endoscopic lesions segmentation. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 4022–4031 (June 2020)
13. Dong, J., Wang, L., Fang, Z., Sun, G., Xu, S., Wang, X., Zhu, Q.: Federated class-incremental learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 10164–10173 (June 2022)
14. Douillard, A., Chen, Y., Dapogny, A., Cord, M.: Plop: Learning without forgetting for continual semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4040–4050 (2021)
15. Everingham, M., Van Gool, L., Williams, C.K., Winn, J., Zisserman, A.: The pascal visual object classes (voc) challenge. International journal of computer vision **88**(2), 303–338 (2010)

16. Fontanel, D., Cermelli, F., Mancini, M., Caputo, B.: Detecting anomalies in semantic segmentation with prototypes. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 113–121 (2021)
17. Ganea, D.A., Boom, B., Poppe, R.: Incremental few-shot instance segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1185–1194 (2021)
18. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
19. Hou, S., Pan, X., Loy, C.C., Wang, Z., Lin, D.: Learning a unified classifier incrementally via rebalancing. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 831–839 (2019)
20. Huang, Z., Hao, W., Wang, X., Tao, M., Huang, J., Liu, W., Hua, X.S.: Half-real half-fake distillation for class-incremental semantic segmentation. arXiv preprint arXiv:2104.00875 (2021)
21. Huang, Z., Wang, X., Huang, L., Huang, C., Wei, Y., Liu, W.: Ccnet: Criss-cross attention for semantic segmentation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 603–612 (2019)
22. Huang, Z., Wei, Y., Wang, X., Shi, H., Liu, W., Huang, T.S.: Alignseg: Feature-aligned segmentation networks. IEEE Transactions on Pattern Analysis and Machine Intelligence (2021)
23. Ji, W., Li, X., Wei, L., Wu, F., Zhuang, Y.: Context-aware graph label propagation network for saliency detection. IEEE Transactions on Image Processing **29**, 8177–8186 (2020)
24. Ji, W., Li, X., Wu, F., Pan, Z., Zhuang, Y.: Human-centric clothing segmentation via deformable semantic locality-preserving network. IEEE Transactions on Circuits and Systems for Video Technology **30**(12), 4837–4848 (2019)
25. Joseph, K., Khan, S., Khan, F.S., Balasubramanian, V.N.: Towards open world object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5830–5840 (2021)
26. Kirkpatrick, J., Pascanu, R., Rabinowitz, N., Veness, J., Desjardins, G., Rusu, A.A., Milan, K., Quan, J., Ramalho, T., Grabska-Barwinska, A., et al.: Overcoming catastrophic forgetting in neural networks. Proceedings of the national academy of sciences **114**(13), 3521–3526 (2017)
27. Klingner, M., Bär, A., Donn, P., Fingscheidt, T.: Class-incremental learning for semantic segmentation re-using neither old data nor old labels. In: 2020 IEEE 23rd International Conference on Intelligent Transportation Systems (ITSC). pp. 1–8. IEEE (2020)
28. Kundu, J.N., Venkatesh, R.M., Venkat, N., Revanur, A., Babu, R.V.: Class-incremental domain adaptation. In: European Conference on Computer Vision. pp. 53–69. Springer (2020)
29. Li, X., Zhou, Y., Wu, T., Socher, R., Xiong, C.: Learn to grow: A continual structure learning framework for overcoming catastrophic forgetting. In: International Conference on Machine Learning. pp. 3925–3934. PMLR (2019)
30. Li, Z., Hoiem, D.: Learning without forgetting. IEEE transactions on pattern analysis and machine intelligence **40**(12), 2935–2947 (2017)
31. Liu, Y., Su, Y., Liu, A.A., Schiele, B., Sun, Q.: Mnemonics training: Multi-class incremental learning without forgetting. In: Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition. pp. 12245–12254 (2020)

32. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3431–3440 (2015)
33. Mallya, A., Davis, D., Lazebnik, S.: Piggyback: Adapting a single network to multiple tasks by learning to mask weights. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 67–82 (2018)
34. Mallya, A., Lazebnik, S.: Packnet: Adding multiple tasks to a single network by iterative pruning. In: Proceedings of the IEEE conference on Computer Vision and Pattern Recognition. pp. 7765–7773 (2018)
35. Maracani, A., Michieli, U., Toldo, M., Zanuttigh, P.: Recall: Replay-based continual learning in semantic segmentation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 7026–7035 (2021)
36. 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)
37. Michieli, U., Zanuttigh, P.: Incremental learning techniques for semantic segmentation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops. pp. 0–0 (2019)
38. Michieli, U., Zanuttigh, P.: Continual semantic segmentation via repulsion-attraction of sparse and disentangled latent representations. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1114–1124 (2021)
39. Michieli, U., Zanuttigh, P.: Knowledge distillation for incremental learning in semantic segmentation. Computer Vision and Image Understanding **205**, 103167 (2021)
40. Pohlen, T., Hermans, A., Mathias, M., Leibe, B.: Full-resolution residual networks for semantic segmentation in street scenes. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 4151–4160 (2017)
41. Rebuffi, S.A., Kolesnikov, A., Sperl, G., Lampert, C.H.: icarl: Incremental classifier and representation learning. In: Proceedings of the IEEE conference on Computer Vision and Pattern Recognition. pp. 2001–2010 (2017)
42. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical image computing and computer-assisted intervention. pp. 234–241. Springer (2015)
43. Stan, S., Rostami, M.: Unsupervised model adaptation for continual semantic segmentation. arXiv preprint arXiv:2009.12518 (2020)
44. Wang, J., Sun, K., Cheng, T., Jiang, B., Deng, C., Zhao, Y., Liu, D., Mu, Y., Tan, M., Wang, X., et al.: Deep high-resolution representation learning for visual recognition. IEEE transactions on pattern analysis and machine intelligence (2020)
45. Wang, X., Girshick, R., Gupta, A., He, K.: Non-local neural networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 7794–7803 (2018)
46. Wu, Y., Chen, Y., Wang, L., Ye, Y., Liu, Z., Guo, Y., Fu, Y.: Large scale incremental learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 374–382 (2019)
47. Yan, S., Zhou, J., Xie, J., Zhang, S., He, X.: An em framework for online incremental learning of semantic segmentation. In: Proceedings of the 29th ACM International Conference on Multimedia. pp. 3052–3060 (2021)
48. Yu, F., Wang, D., Shelhamer, E., Darrell, T.: Deep layer aggregation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2403–2412 (2018)

49. Yu, L., Liu, X., van de Weijer, J.: Self-training for class-incremental semantic segmentation. arXiv preprint arXiv:2012.03362 (2020)
50. Zenke, F., Poole, B., Ganguli, S.: Continual learning through synaptic intelligence. In: International Conference on Machine Learning. pp. 3987–3995. PMLR (2017)
51. Zhang, H., Dana, K., Shi, J., Zhang, Z., Wang, X., Tyagi, A., Agrawal, A.: Context encoding for semantic segmentation. In: Proceedings of the IEEE conference on Computer Vision and Pattern Recognition. pp. 7151–7160 (2018)
52. Zhao, H., Qin, X., Su, S., Fu, Y., Lin, Z., Li, X.: When video classification meets incremental classes. In: Proceedings of the 29th ACM International Conference on Multimedia. pp. 880–889 (2021)
53. Zhao, H., Wang, H., Fu, Y., Wu, F., Li, X.: Memory efficient class-incremental learning for image classification. IEEE Transactions on Neural Networks and Learning Systems (2021)
54. Zhao, H., Zeng, H., Qin, X., Fu, Y., Wang, H., Omar, B., Li, X.: What and where: Learn to plug adapters via nas for multidomain learning. IEEE Transactions on Neural Networks and Learning Systems (2021)
55. Zhao, H., Shi, J., Qi, X., Wang, X., Jia, J.: Pyramid scene parsing network. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2881–2890 (2017)
56. Zheng, E., Yu, Q., Li, R., Shi, P., Haake, A.: A continual learning framework for uncertainty-aware interactive image segmentation. In: Proceedings of the AAAI Conference on Artificial Intelligence. pp. 6030–6038 (2021)
57. Zhou, B., Zhao, H., Puig, X., Fidler, S., Barriuso, A., Torralba, A.: Scene parsing through ade20k dataset. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 633–641 (2017)