Continual Attentive Fusion for Incremental Learning in Semantic Segmentation

Guanglei Yang, Enrico Fini, Dan Xu, Paolo Rota, Mingli Ding*, Hao Tang,
Xavier Alameda-Pineda, *Senior Member, IEEE,* Elisa Ricci*, *Member, IEEE*

***Abstract*—Over the past years, semantic segmentation, similar to many other tasks in computer vision, has benefited from the progress in deep neural networks, resulting in significantly improved performance. However, deep architectures trained with gradient-based techniques suffer from catastrophic forgetting, which is the tendency to forget previously learned knowledge while learning new tasks. Aiming at devising strategies to counteract this effect, incremental learning approaches have gained popularity over the past years. However, the first incremental learning methods for semantic segmentation appeared only recently. While effective, these approaches do not account for a crucial aspect in pixel-level dense prediction problems, *i.e.*, the role of attention mechanisms. To fill this gap, in this paper, we introduce a novel attentive feature distillation approach to mitigate catastrophic forgetting while accounting for semantic spatial- and channel-level dependencies. Furthermore, we propose a continual attentive fusion structure, which takes advantage of the attention learned from the new and the old tasks while learning features for the new task. Finally, we also introduce a novel strategy to account for the background class in the distillation loss, thus preventing biased predictions. We demonstrate the effectiveness of our approach with an extensive evaluation on Pascal-VOC 2012 and ADE20K, setting a new state of the art.**

***Index Terms*—Knowledge Distillation, Incremental Learning, Semantic Segmentation.**

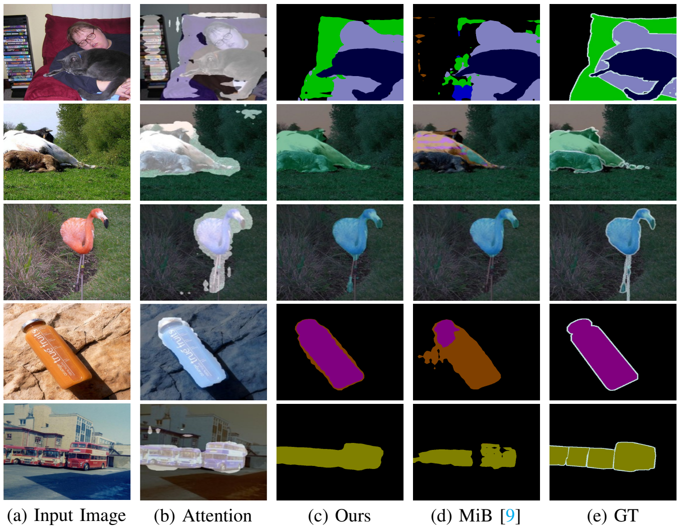

(a) Input Image (b) Attention (c) Ours (d) MiB [9] (e) GT

Fig. 1. Given an input image, by leveraging from attention maps (b) computed with the proposed continual attentive fusion (CAF) module, our method produces segmentation maps (c) more similar to ground truths (e) than those computed with previous methods such as MiB [9] in (d).

I. INTRODUCTION

During the last decade, the emergence of deep learning has led to several breakthroughs in many computer vision and multimedia tasks. Semantic segmentation, the problem of assigning a semantic label to each pixel in an image, is no exception to this trend [1]. Sophisticated deep neural networks such as fully convolutional networks (FCNs) [2] or dilated convolution models [3], together with the availability of large human-annotated datasets and powerful hardware, have led to exceptional results on challenging semantic segmentation benchmarks [2]–[7].

Although deep learning models are approaching and even exceeding human-level performance on numerous tasks, including semantic segmentation, artificial neural networks encounter serious difficulties in regard to incremental learning (IL). In other words, they struggle to preserve past knowledge when attempting to learn multiple tasks sequentially. This is due to the well-known catastrophic forgetting issue [8], which is the tendency of a deep network to forget previously learned tasks. Unfortunately, this problem is intrinsic to the optimization techniques (*e.g.*, gradient descent) used to train neural networks. Nonetheless, the ability to learn a sequence of tasks is unarguably a highly desirable property of artificial intelligent systems.

For this reason, IL has recently received considerable attention, particularly incremental class learning (ICL), where the ability of a model to discriminate current and past classes simultaneously without knowing the task at test time is evaluated. Over the past years, several works have proposed ICL approaches, mostly focusing on image classification [10]–[17], object detection [18], [19], image retrieval [20], [21] and emotion recognition [22]–[24]. To date, much less attention has been devoted to ICL for semantic segmentation. Recently, Cermelli *et al.* [9] showed promising results by introducing a distillation-based framework that accounts for background

Guanglei Yang and Mingli Ding are with the School of Instrument Science and Engineering, Harbin Institute of Technology (HIT), Harbin, China. E-mail: {yangguanglei,dingml}@hit.edu.cn.

Enrico Fini, Paolo Rota, and Elisa Ricci are with the Department of Information Engineering and Computer Science, University of Trento, Italy. E-mail: {enrico.fini, paolo.rota, elisa.ricci}@unitn.it.

Dan Xu is with the Department of Computer Science and Engineering, Hong Kong University of Science and Technology. E-mail: danxu@cse.ust.hk.

Hao Tang is with the Department of Information Technology and Electrical Engineering, ETH Zurich. E-mail: hao.tang@vision.ee.ethz.ch

Xavier Alameda-Pineda is with the RobotLearn Group, INRIA. E-mail: xavier.alameda-pineda@inria.fr.

Elisa Ricci is with the Deep Visual Learning group at Fondazione Bruno Kessler, Trento, Italy.

*Corresponding authors.

distribution shifts between tasks to prevent biased predictions. Concurrently, recent studies have demonstrated the effectiveness of attention mechanisms on semantic segmentation [25]–[28]. Although attempts were made to exploit attention for learning tasks sequentially in the context of object classification [29] and detection [19], they cannot be effectively adopted for incremental semantic segmentation. In this paper, we bridge this gap and propose the first attention-based ICL method for semantic segmentation.

Our contributions can be summarized as follows:

- We propose a novel deep architecture for ICL that embeds a new continual attentive fusion (CAF) module. Given the features of the current and previous models, CAF produces structured self-attention weights to update the current features. We demonstrate that this module by itself significantly reduces catastrophic forgetting and leads to improved segmentation maps with respect to previous methods such as MiB [9] (see Figure 1).
- We introduce a novel attentive distillation loss that leverages both channelwise and spatial attention to transfer more relevant knowledge into the current model. By correctly weighting the importance of each channel, the attentive distillation loss enables the model to maintain its performance on old tasks without using any explicit information from the past.
- We devise a simple and effective method for improving the balance between old classes and background probabilities that only depends on the inferred ratio of old and new classes in a given image, thus avoiding introducing additional hyperparameters.
- The proposed approach sets the new state-of-the-art on two common datasets for ICL for semantic segmentation, namely, Pascal-VOC 2012 [30] and ADE20K [31].

II. RELATED WORKS

Incremental Learning. Modern artificial neural networks are haunted by the well-known catastrophic forgetting problem: the tendency of neural models to severely degrade performance on previous tasks when trained on new tasks. This issue has been largely studied in the literature in the last few decades [8], leading to a wide variety of IL approaches. According to [32], prior art in this field can be organized into three categories: replay-based methods [13], [14], [33]–[36], regularization-based techniques [10]–[12], [29], [37] and parameter isolation-based approaches [38], [39]. Replay-based methods consist of storing [13], [14], [33], [40] or generating [34]–[36] examples of the first task, which are then reused in subsequent learning stages. Regularization-based methods either penalize changes in a subset of parameters while learning on new tasks [11], [12], [37], [41] or employ distillation to force the network not to forget past knowledge [10], [17], [29]. Parameter isolation-based approaches are built on the idea of having a task-specific set of parameters. Despite the interest in the problem, the large majority of the literature focuses on classification. A pioneering work by Shmelkov et al. [18] exploited distillation [10] for class discovery in detection. In this paper, we address ICL in semantic segmentation.

Incremental Learning in Semantic Segmentation. Deep learning has brought great progress in semantic segmentation [2]–[5], [42]. Despite the abundant literature on this task, very few works are addressing ICL for this task [9], [43]–[46]. Moreover, these studies address the problem from different perspectives and utilize contrasting experimental settings. For instance, in [43], the authors were the first to study this task, proposing an approach that operates both on the output and on the intermediate representations of the segmentation model. [45] presented a method to sample prototypical examples of the old classes to be used as a rehearsal in the new task. However, both [43], [45] assume that some information from the previous task will be available during the training of the second task. Other approaches [44], [46] described ICL methods that are specialized for certain subfields (*i.e.*, remote sensing and computer assisted radiology and surgery), lacking generality. More recently, [9] attempted to fix the semantic distribution shift in the background class, showing a significant performance boost. However, their approach operates at the loss level, while network architectural changes to improve segmentation maps are not considered. In contrast, in this paper, we overcome several limitations of the previous literature and propose a new architectural solution for rehearsal-free ICL in semantic segmentation.

Attention Mechanisms. Several works considered attention models within deep architectures to improve performance [47]–[54]. Focusing only on pixelwise dense prediction, Chen *et al.* [25] first described an attention model to combine multiscale features learned by an FCN for semantic segmentation. Zhang *et al.* [6] designed EncNet, a network equipped with a channel attention mechanism to model global context. Similarly, Zhao *et al.* [55] proposed accounting for pixelwise dependencies by introducing relative position information across the spatial dimension within the convolutional layers. Other works [27] introduced attention to model contextual and semantic dependencies. Zhong *et al.* [56] considered spatial and channel interdependencies in their squeeze-and-attention network. Xu *et al.* [26] described attention gates introduced to control the message passing among variables, thus integrating attention into a probabilistic model formulation. Our work differs significantly from previous works since (i) we use spatial and channelwise attention to help the network discover new classes and (ii) attention also counteracts the semantic distribution shift of the background class between the two tasks.

III. PROPOSED METHOD

The problem of image segmentation is that of inferring the correct label s_p for each pixel p in the input image $\mathbf{x}$ among the available class labels $\mathcal{S}$. For the sake of simplicity, all images are assumed to be of the same size $|\mathbf{x}| = P$. Tools for semantic segmentation can vary significantly in nature, but we will consider the general case of function $\phi_\omega : \mathcal{X} \rightarrow \mathbb{R}^{|\mathcal{S}| \times P}$ conceived to predict a probability distribution over the class labels for each pixel of the input image, where $\mathcal{X}$ is the input image space and ω are the parameters of this function. More precisely, $\phi_\omega(\mathbf{x})[w, h, s]$ is supposed to represent the probability of the pixel at position (w, h) in image $\mathbf{x}$ belonging

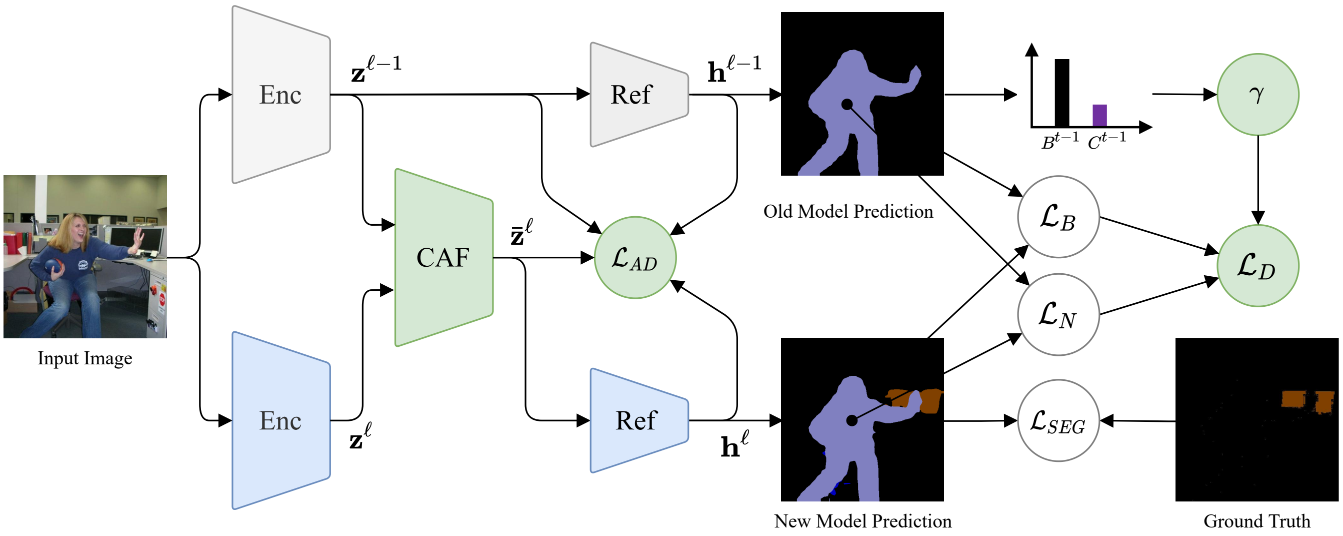


Fig. 2. The overall scheme of our approach at incremental learning step ℓ. The gray modules correspond to the network learned at the previous incremental learning step, which is frozen at step ℓ, while the blue modules correspond to the network trained at step ℓ. Our three contributions, namely, continual attentive fusion (CAF), attentive distillation loss ($\mathcal{L}_{\mathrm{AD}}$) and balanced knowledge distillation ($\mathcal{L}_{\mathrm{D}}$), are highlighted in green.

to class $s \in \mathcal{S}$. Specifically, the function approximator ϕ_ω can be further divided into three subnetworks (see Figure 2): (i) an encoder realized by a residual neural network that extracts a feature map $\mathbf{z}$, (ii) a refinement model that projects the features into a refined version $\mathbf{h}$, and (iii) a classifier that produces a probability distribution p over the class labels. The set of parameters can be learned with the help of training set $\mathcal{D} \subset \mathcal{X} \times \mathcal{S}^P$.

Traditionally, semantic segmentation considers a fixed set of classes $\mathcal{S}$. More recently, the community started looking at learning these classes in a *continuous* setup, *i.e.*, considering a series of *incremental learning steps*, indexed by ℓ. At each learning step, an extra set of categories is added to the semantic segmentation task. In the following, we denote by $\mathcal{S}_{\ell-1}$ the set of classes learned –seen– before learning step ℓ, and by $\mathcal{U}_\ell$ the set of new –unseen– classes added at learning step ℓ. The following holds: $\mathcal{S}_\ell = \mathcal{U}_\ell \cup \mathcal{S}_{\ell-1}$ and $\mathcal{U}_\ell \cap \mathcal{S}_{\ell-1} = \emptyset$. This means that at every learning step, the size of the output classifier increases with respect to the previous learning step. Consequently, the training set at step ℓ is denoted by $\mathcal{D}_\ell \subset \mathcal{X} \times \mathcal{U}_\ell^P$ and the inference function by $\phi_\omega^\ell : \mathcal{X} \to \mathbb{R}^{|\mathcal{S}_\ell| \times P}$. The index ℓ also applies to the various features extracted by our architecture in Figure 2, namely, the convolutional features $\mathbf{z}^\ell$ and the refined features $\mathbf{h}^\ell$.

There are three methodological contributions of the proposed method. First, a continual attentive fusion module exploits the convolutional features of both tasks, $\mathbf{z}^{\ell-1}$ and $\mathbf{z}^\ell$, to compute a structured attention tensor used to transform $\mathbf{z}^\ell$, considering $\mathbf{z}^{\ell-1}$ (see Section III-A). At test time, where $\mathbf{z}^{\ell-1}$ are unavailable, several strategies are evaluated in the experiments. Second, a self-attention feature distillation loss that leverages both channelwise and spatial attention to transfer more relevant knowledge into the current step. The loss acts on both the convolutional $\mathbf{z}$ and the refined $\mathbf{h}$ features (see Section III-B). Third, a balanced knowledge distillation loss accounts for the overpresence of the background class. Indeed, when comparing old and new segmentation maps, the new classes must be merged with the background, thus overestimating the number of background pixels. We propose a simple yet very effective method to rebalance the learning in Section III-C. In the following, we describe in detail our three contributions.

A. Continual Attentive Fusion

The continual attentive fusion (CAF) module is specifically conceived to compute self-attention from the features of the current IL step ℓ as well as from the previous step $\ell - 1$. In practice, it is composed of two main blocks, with the purpose of (i) computing features independently for $\mathbf{z}^\ell$ and $\mathbf{z}^{\ell-1}$ and (ii) fusing these features into a structured attention tensor. The diagram of CAF is shown in Figure 3.

To compute the features to be fused, denoted by $\mathbf{v}^\ell$ and $\mathbf{v}^{\ell-1}$, we draw inspiration from nonlocal neural networks [57]. Each input feature map $\mathbf{z}$ is fed to three different 1×1 convolutional layers. The output of the first two convolutions is used to obtain a nonlocal self-attention tensor, which is then used to weight the output of the third convolution, thus obtaining $\mathbf{v}$. The nonlocal features corresponding to $\mathbf{z}^{\ell-1}$ and $\mathbf{z}^\ell$ are denoted by $\mathbf{v}^{\ell-1}$ and $\mathbf{v}^\ell$, respectively, and correspond to the upper and bottom nonlocal blocks of Figure 3. The projected features corresponding to $\mathbf{z}^{\ell-1}$ and $\mathbf{z}^\ell$ are denoted by $\mathbf{v}^{\ell-1}$ and $\mathbf{v}^\ell$, respectively, and correspond to the upper and bottom feature refinement blocks in Figure 3.

As stated above, the projected features are aggregated to compute a structured attention tensor. To do so, we employ a fusion module (see Figure 3) that first concatenates the projected features along the channel dimension and then feeds them to a convolutional layer to reduce the number of channels by two, thus returning to the original amount. The output of the fusion module is a tensor $\mathbf{v}^{(\ell-1,\ell)}$ containing the information

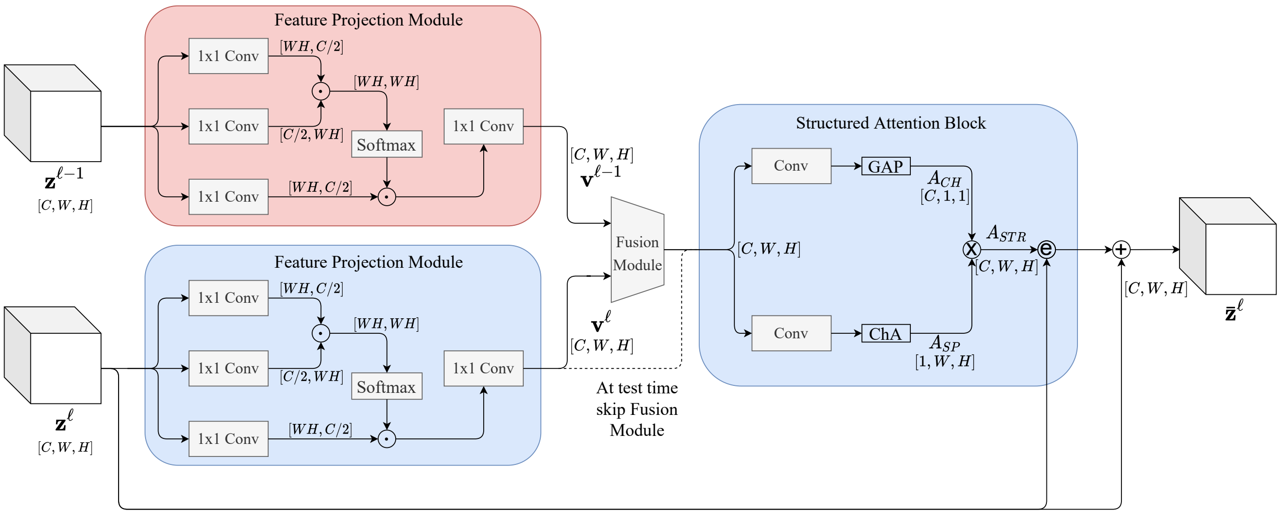


Fig. 3. A detailed view of the CAF module. Nonlocal features $\mathbf{v}^{\ell-1}$ and $\mathbf{v}^{\ell}$ are computed independently from the features of the two learning steps $\mathbf{z}^{\ell-1}$ and $\mathbf{z}^{\ell}$ via a series of 1×1 convolutions and matrix multiplications $\odot$. After a fusion module, a channel attention vector and a spatial attention map are computed and then aggregated via a tensor product $\otimes$ into a structured attention tensor. The elementwise product (circled e) is used to apply the attention to the original features. A residual connection provides the updated feature map $\bar{\mathbf{z}}^{\ell}$. The whole structure represents the architecture at training time, while red background parts and the fusion module are discarded at test time.

coming from both models. Note that this module is used at training time but can be skipped at test time if $\mathbf{z}^{\ell-1}$ is not computed (*e.g.,* to reduce resource usage).

The fusion tensor, $\mathbf{v}^{(\ell-1,\ell)}$, is fed to two convolutional layers to extract a channelwise attention vector and a spatial attention map via global average pooling (GAP) and channelwise average (ChA) operations, respectively:

$$\boldsymbol{A}^{\ell}_{\text{SP}}[w,h] = \frac{1}{C}\sum_{c=1}^{C}(\omega_{\text{SP}} * \mathbf{v}^{(\ell-1,\ell)})[c,w,h], \tag{1}$$

$$\boldsymbol{A}^{\ell}_{\text{CH}}[c] = \frac{1}{WH}\sum_{h,w=1}^{H,W}(\omega_{\text{CH}} * \mathbf{v}^{(\ell-1,\ell)})[c,w,h], \tag{2}$$

where ω_{SP} and ω_{CH} are the weights of the convolutions. The attention vector and spatial map are used to construct a structured attention tensor via a tensor product:

$$\boldsymbol{A}_{\text{STR}} = \boldsymbol{A}_{\text{CH}} \otimes \boldsymbol{A}_{\text{SP}}, \tag{3}$$

which is finally used, together with a residual connection, to obtain the updated feature map $\bar{\mathbf{z}}^{\ell}$:

$$\bar{\mathbf{z}}^{\ell} = (\boldsymbol{I} + \boldsymbol{A}_{\text{STR}})\,\mathbf{v}^{(\ell-1,\ell)} = (\boldsymbol{I} + \boldsymbol{A}_{\text{CH}} \otimes \boldsymbol{A}_{\text{SP}})\,\mathbf{v}^{(\ell-1,\ell)}. \tag{4}$$

In this way, the new features $\bar{\mathbf{z}}^{\ell}$ are computed via a structured self-attention tensor, which is computed from an aggregation of projected features from both the current and the previous learning steps $\mathbf{z}^{\ell}$ and $\mathbf{z}^{\ell-1}$. Alternatives to this fusion step that avoid using $\mathbf{z}^{\ell-1}$ at test time are discussed in the experimental section.

B. Attentive Feature Distillation

Many existing methods apply feature distillation to preserve previous knowledge [10], [19], [29], [44]. Most of them treat the channels in the feature map equally, *i.e.*, channels are weighted uniformly in the distillation loss. However, the features tend to drift to a new configuration to discriminate the classes of the task at hand, regardless of the type of precautions employed to avoid forgetting. Hopefully, some portion of the features will change substantially to adapt to the new task, while most of them will remain reasonably close to their previous configuration. This undermines the assumption that all the channels should be treated equally. We conjecture that this is a primary limitation to previous works in ICL for semantic segmentation. To overcome this issue, we employ the squeeze-and-excitation (SE) module [58] to generate channelwise attention as follows:

$$\text{AD}_{\text{CH}}\left(\mathbf{m}\right) = \psi\left(\omega_2^{\mathbf{m}} * \sigma\left(\omega_1^{\mathbf{m}} * \text{AvgPool}(\mathbf{m})\right)\right), \tag{5}$$

where $\psi(\cdot)$ and $\sigma(\cdot)$ represent the sigmoid and ReLU activation functions, respectively, and $\mathbf{m}$ is a generic feature map we use here as a placeholder. Note the superscript in the weight matrices $\omega_1^{\mathbf{m}}$ and $\omega_2^{\mathbf{m}}$, meaning that those weights are specific to the feature map input to $\text{AD}_{\text{CH}}\left(\cdot\right)$. For a generic feature map $\mathbf{m}$ of dimensions $[C, W, H]$, $\text{AD}_{\text{CH}}\left(\cdot\right)$ will be of size $[C, 1, 1]$.

Similarly, since the background is very complex and can excite features of other classes, context information is crucial for semantic segmentation. We leverage the inferred probability distributions of the classes in the background to improve the distillation process. However, for the channels, it might be appropriate to let the network decide which parts of the background are more important to distill. Hence, we utilize self-spatial attention:

$$\text{AD}_{\text{SP}}\left(\mathbf{m}\right) = \frac{\sum_{j=1}^{C}\mathbf{m}_j^2}{\left\|\sum_{j=1}^{C}\mathbf{m}_j^2\right\|_{\text{F}}}. \tag{6}$$

The size of AD_{SP} will be $[1, W, H]$. Spatial and channelwise attention are combined through a tensor product:

$$\text{AD}\left(\mathbf{m}\right) = \left(\text{AD}_{\text{CH}}\left(\mathbf{m}\right) \otimes \text{AD}_{\text{SP}}\left(\mathbf{m}\right) + 1\right)\,\mathbf{m}, \tag{7}$$

where the second product is elementwise between two tensors of the same size. Overall, $\mathrm{AD}(\mathbf{m})$ weights the original tensor with a structured and learned self-attention tensor. In our framework, we use this formulation to distill knowledge from the previous task into the new task, therefore defining a structured self-attention distillation loss:

$$\mathcal{L}_{\mathrm{AD}} = \|\mathrm{AD}(\bar{\mathbf{z}}^{\ell}) - \mathrm{AD}(\mathbf{z}^{\ell-1})\|_{\mathrm{F}}^{2} + \|\,\mathrm{AD}(\mathbf{h}^{\ell}) - \mathrm{AD}(\mathbf{h}^{\ell-1})\|_{\mathrm{F}}^{2}. \tag{8}$$

Very importantly, $\mathcal{L}_{\mathrm{AD}}$ is applied to both features $\mathbf{z}$ and $\mathbf{h}$ of both the old and new incremental learning steps $\ell-1$ and ℓ.

C. Balanced Knowledge Distillation

As mentioned, in the context of incremental learning in semantic segmentation, distillation [59] plays a core role in transferring knowledge from the old model into the new model, mitigating catastrophic forgetting. A typical definition for the distillation loss is:

$$\mathcal{L}_{\mathrm{UD}} = \beta \sum_{w,h=0}^{W,H} \sum_{s=0}^{|\mathcal{S}_{\ell-1}|} \phi_{\omega}^{\ell-1}(\mathbf{x})[w,h,s] \, \log \phi_{\omega}^{\ell}(\mathbf{x})[w,h,s], \tag{9}$$

where $\phi_{\omega}^{\ell}(\mathbf{x})[w,h,s]$ and $\phi_{\omega}^{\ell-1}(\mathbf{x})[w,h,s]$ are the probabilities of a pixel at position (h,w) belonging to class s as inferred by the new and old models, respectively, while $\beta = -\frac{1}{HW}$ is a normalization factor.

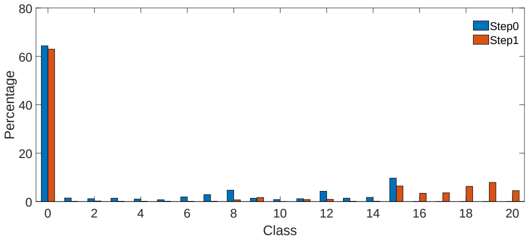
Assuming that the background class was part of the previous learning step, $\mathrm{B}_{\ell-1} \in \mathcal{S}_{\ell-1}$, we can decompose the previous loss into two contributions: from the background $\mathcal{L}_{\mathrm{B}}$ and from the other classes $\mathcal{L}_{\mathrm{N}}$. During the learning of the previous step $\ell-1$, the pixels belonging to the classes unknown at step $\ell-1$ but known at step ℓ, that is, $\mathcal{U}_{\ell}$, are assigned to the background class. This leads to an imbalance in $\mathcal{L}_{\mathrm{B}}$ because $\phi_{\omega}^{\ell-1}$ is not aware of the new classes. To address this issue, [9] rewrites ϕ_{ω}^{ℓ} as:

$$\hat{\phi}_{\omega}^{\ell}(\mathbf{x})[h,w,s] = \begin{cases} \phi_{\omega}^{\ell}(\mathbf{x})[h,w,s] & s \neq \mathrm{B}_{\ell-1}, \\ \sum_{s' \in \mathcal{U}_{\ell} \cup \{\mathrm{B}_{\ell}\}} \phi_{\omega}^{\ell}(\mathbf{x})[h,w,s'] & s = \mathrm{B}_{\ell-1}, \end{cases} \tag{10}$$

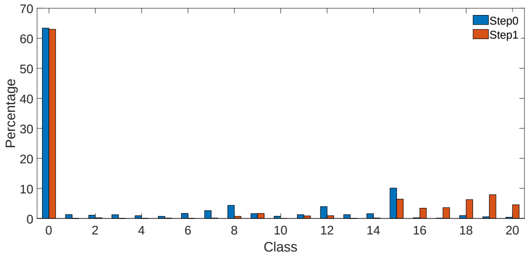
thus aggregating the new background class B_{ℓ} together with the unseen classes $\mathcal{U}_{\ell}$ to emulate the background class at the previous learning step $\mathrm{B}_{\ell-1}$. The opposite also holds, since we assume that annotations for classes of previous time steps are not available and therefore belong to the background class at step ℓ. This has no impact on the distillation loss but on the supervised segmentation loss in (13).

According to Figure 4, even if the new $\hat{\phi}_{\omega}^{\ell}$ accounts for the imbalance between the old and new probability distributions for distillation, the background is usually the most represented class by far. Thus, $\mathcal{L}_{\mathrm{B}}$ has a much stronger contribution to the overall loss compared to $\mathcal{L}_{\mathrm{N}}$. As a consequence, the network basically ignores the information of old classes, preventing effective knowledge distillation. To overcome this issue, we propose introducing a balancing parameter γ to reweight the influence between $\mathcal{L}_{\mathrm{B}}$ and $\mathcal{L}_{\mathrm{N}}$ in the distillation loss. Formally, the expression of γ is written as:

$$\gamma = \frac{\sum_{s \in \mathcal{S}_{t-1}/\{\mathrm{B}_{\ell-1}\}} \mathrm{Softmax}\left(\mathrm{AvgPool}\left(\phi_{\omega}^{\ell-1}(\mathbf{x})\,[s]\right)\right)}{\mathrm{Softmax}\left(\mathrm{AvgPool}\left(\phi_{\omega}^{\ell-1}\,(\mathbf{x})\,[\mathbf{B}_{\ell-1}]\right)\right)}, \tag{11}$$



(a) Disjoint



(b) Overlapped

Fig. 4. The statistical distribution of different classes under different steps in the VOC 2012 15-5 disjoint setting and overlapped setting.

and allows us to step further than what was proposed in [9] and weight the distillation as follows:

$$\mathcal{L}_{D} = \gamma \mathcal{L}_{B} + \mathcal{L}_{N}, \tag{12}$$

where $\mathcal{L}_{B}$ and $\mathcal{L}_{N}$ and the background and nonbackground contributions to the unweighted distillation loss in (9).

D. Overall Loss

The two losses described in the previous sections are used, together with the standard supervised loss for semantic segmentation, to train the overall architecture. In more detail, the final loss is expressed as:

$$\mathcal{L} = \mathcal{L}_{\mathrm{SEG}} + \lambda_{\mathrm{AD}} \mathcal{L}_{\mathrm{AD}} + \lambda_{\mathrm{D}} \mathcal{L}_{\mathrm{D}}, \tag{13}$$

where λ_{AD} and λ_{D} are the weights of the attention distillation and knowledge distillation losses defined in Sections III-B and III-C, respectively, and $\mathcal{L}_{\mathrm{SEG}}$ is the supervised segmentation loss (pixelwise cross-entropy) previously defined as follows:

$$\mathcal{L}_{\mathrm{SEG}} = -\frac{1}{HW} \sum_{w,h=0}^{W,H} \sum_{s=0}^{|\mathcal{S}_{\ell}|} \log \tilde{\phi}_{\omega}^{\ell}(\mathbf{x})[w,h,s], \tag{14}$$

where:

$$\tilde{\phi}_{\omega}^{\ell}(\mathbf{x})[h,w,s] = \begin{cases} \phi_{\omega}^{\ell}(\mathbf{x})[h,w,s] & s \neq \mathrm{B}_{\ell}, \\ \sum_{s' \in \mathcal{S}_{\ell-1} \cup \{\mathrm{B}_{\ell}\}} \phi_{\omega}^{\ell}(\mathbf{x})[h,w,s'] & s = \mathrm{B}_{\ell}. \end{cases} \tag{15}$$

Note that the output probabilities $\tilde{\phi}$ defined here for the segmentation loss are different from the output probabilities $\hat{\phi}$ defined in the main paper for the background distillation loss. While $\tilde{\phi}$ aggregates the previous classes to the current background class (so that all previous classes become the background for the segmentation loss), the output probabilities defined in the paper $\hat{\phi}$ aggregate the new classes to the current background because the background at the previous incremental step includes the new classes.

TABLE I
MEAN IOU ON THE PASCAL-VOC 2012 DATASET FOR DIFFERENT INCREMENTAL CLASS LEARNING SCENARIOS. $*$ INDICATES THAT RESULTS COME FROM REIMPLEMENTATION.

Method	19-1						15-5						15-1					
	Disjoint			Overlapped			Disjoint			Overlapped			Disjoint			Overlapped		
	1-19	20	all	1-19	20	all	1-15	16-20	all	1-15	16-20	all	1-15	16-20	all	1-15	16-20	all
FT	5.8	12.3	6.2	6.8	12.9	7.1	1.1	33.6	9.2	2.1	33.1	9.8	0.2	1.8	0.6	0.2	1.8	0.6
PI [37]	5.4	14.1	5.9	7.5	14.0	7.8	1.3	34.1	9.5	1.6	33.3	9.5	0.0	1.8	0.4	0.0	1.8	0.4
EWC [11]	23.2	16.0	22.9	26.9	14.0	26.3	26.7	37.7	29.4	24.3	35.5	27.1	0.3	4.3	1.3	0.3	4.3	1.3
RW [12]	19.4	15.7	19.2	23.3	14.2	22.9	17.9	36.9	22.7	16.6	34.9	21.2	0.2	5.4	1.5	0.0	5.2	1.3
LwF [10]	53.0	9.1	50.8	51.2	8.5	49.1	58.4	37.4	53.1	58.9	36.6	53.3	0.8	3.6	1.5	1.0	3.9	1.8
LwF-MC [13]	63.0	13.2	60.5	64.4	13.3	61.9	67.2	41.2	60.7	58.1	35.0	52.3	4.5	7.0	5.2	6.4	8.4	6.9
ILT [43]	69.1	16.4	66.4	67.1	12.3	64.4	63.2	39.5	57.3	66.3	40.6	59.9	3.7	5.7	4.2	4.9	7.8	5.7
MiB [9]	69.6	25.6	67.4	70.2	22.1	67.8	71.8	43.3	64.7	75.5	49.4	69.0	46.2	12.9	37.9	35.1	13.5	29.7
SDR [60]	70.8	31.4	68.9	71.3	23.4	69.0	74.6	**44.1**	**67.3**	76.3	**50.2**	70.1	**59.4**	14.3	**48.7**	47.3	14.7	39.5
PLOP* [61]	75.1	**38.2**	73.2	75.0	39.1	73.2	66.5	39.6	59.8	74.7	49.8	68.5	49.0	13.8	40.2	**65.2**	**22.4**	**54.5**
Ours	**75.5**	30.8	**73.3**	**75.5**	**34.8**	**73.4**	**72.9**	42.1	65.2	**77.2**	49.9	**70.4**	57.2	**15.5**	46.7	55.7	14.1	45.3
Joint	77.4	78.0	77.4	77.4	78.0	77.4	79.1	72.6	77.4	79.1	72.6	77.4	79.1	72.6	77.4	79.1	72.6	77.4

TABLE II
MEAN IOU ON THE ADE20K DATASET FOR DIFFERENT INCREMENTAL CLASS LEARNING SCENARIOS. $*$ INDICATES THAT RESULTS COME FROM REIMPLEMENTATION.

Method	100-50			100-10							50-50			
	1-100	101-150	all	1-100	100-110	110-120	120-130	130-140	140-150	all	1-50	51-100	101-150	all
FT	0.0	24.9	8.3	0.0	0.0	0.0	0.0	0.0	16.6	1.1	0.0	0.0	22.0	7.3
LwF [10]	21.1	25.6	22.6	0.1	0.0	0.4	2.6	4.6	16.9	1.7	5.7	12.9	22.8	13.9
LwF-MC [13]	34.2	10.5	26.3	18.7	2.5	8.7	4.1	6.5	5.1	14.3	27.8	7.0	10.4	15.1
ILT [43]	22.9	18.9	21.6	0.3	0.0	1.0	2.1	4.6	10.7	1.4	8.4	9.7	14.3	10.8
Inc. Seg [62]	36.6	0.4	24.6	32.4	0.0	0.2	0.0	0.0	0.0	21.7	40.2	1.3	0.3	14.1
MiB [9]	**37.9**	27.9	34.6	31.8	10.4	14.8	12.8	**13.6**	**18.7**	25.9	35.5	22.2	**23.6**	27.0
SDR [60]	37.5	25.5	33.5	28.9	-	-	-	-	-	23.2	42.9	-	**-**	31.3
PLOP* [61]	29.8	4.2	22.2	32.1	1.9	10.0	0.8	1.2	0.1	22.3	19.2	0.4	0.4	6.6
Ours	37.3	**31.9**	**35.5**	**39.0**	**14.6**	**22.0**	**25.4**	12.1	13.1	**31.8**	**47.5**	**30.6**	23.0	**33.7**
Joint	44.3	28.2	38.9	44.3	26.1	42.8	26.7	28.1	17.3	38.9	51.1	38.3	28.2	38.9

IV. EXPERIMENTS

In this section, we demonstrate the effectiveness of our approach through extensive experiments on the two publicly available benchmarks for ICL in semantic segmentation.

A. *Experimental Setups*

Datasets. We consider two datasets in our experiments. The PASCAL-VOC 2012 dataset [30] contains 10,582 and 1,449 in the training and validation sets, respectively. Pixels can be associated with 21 different classes (20 plus the background). Following [9], [18], [43], [44], we define two experimental settings: disjoint and overlapped. Following [43], the disjoint setup assumes that the new set is disjointed from previously used samples, ı.e., $(\cup_{j=0,\ldots,k-1}\mathcal{D}_j \cap \mathcal{D}_k = \emptyset)$. Following [18], in the overlapped setting each training step contains all the images that have at least one pixel of a novel class, regardless of what other classes are also included. It is important to know that in this case, training images may contain pixels of unseen classes (thus labeled background). This setup is more realistic since it does not make any restriction on the objects present in the images. Following previous work [9], [18], [43], we perform three different experiments concerning the addition of one class (19-1), five classes all at once (15-5), and five classes added one-by-one in alphabetical order (15-1), and report mean IoU.

ADE20K [31] is a large-scale dataset with 150 classes. Different from Pascal-VOC 2012, this dataset contains nonobject classes (*e.g.*, sky, building, wall). We create the incremental datasets $\mathcal{D}_\ell$ by splitting the whole dataset into disjoint image sets without any constraint except ensuring a minimum number of images (*i.e.*, 50) containing new classes. Obviously, each $\mathcal{D}_\ell$ provides annotations only for current classes, while old and future classes are annotated as the background. We report the mean IoU obtained by averaging the results as in [31], and we perform three different experiments: single-step addition of 50 classes (100-50), multistep addition of 50 classes (100-10) and three steps of 50 classes (50-50).

Baselines. We compare our method against previous ICL methods originally designed for image classification, following [9], namely, path integral (PI) [37], elastic weight consolidation (EWC) [11], and Riemannian walks (RW) [12]. We also compare our method with learning without forgetting (LwF) [10] and its multiclass version (LwF-MC) [13]. Finally, we consider previous ICL approaches for segmentation, *i.e.* MiB [9], ILT [43], SDR [60], and PLOP [61]. In addition to the state-of-the-art methods, we report results for fine tuning (FT) and joint training (joint). These results serve as a lower and upper bound. In FT, we train on the new task via simple fine tuning, while in joint a unique training of both tasks is performed.

Implementation Details. We choose ResNet-101 [63] as our

TABLE III

Ablation study of networks' components on the Pascal-VOC 2012 disjoint 15-5 setup. Per-class IoU of the evaluated methods when the last five classes are added are reported. CAF denotes our continual attentive fusion module, AD attentive feature distillation, BKD the balanced knowledge distillation loss and KD knowledge distillation as in (9) and (10).

Method	aero	bike	bird	boat	bottle	bus	car	cat	chair	cow	din. table	dog	horse	mbike	person	mIoU old	plant	sheep	sofa	train	tv	mIoU
Baseline	70.6	33.5	73.5	59.2	66.6	49.1	72.6	74.7	28.6	34.8	43.3	72.6	70.7	69.9	70.2	59.3	27.4	31.0	25.9	41.5	44.5	53.0
+ KD	77.9	36.9	81.1	65.3	73.4	54.2	80.1	82.4	31.6	38.4	47.7	80.1	78.0	77.1	77.4	65.4	30.3	34.2	28.6	45.8	49.1	58.5
+ BKD	79.1	37.5	82.3	66.3	74.5	55.0	81.4	83.7	32.1	39.0	48.5	81.4	79.2	78.3	78.6	66.5	30.7	34.7	29.1	46.5	49.8	59.4
+ AD	79.9	37.9	83.2	67.0	75.3	55.6	82.2	84.5	32.4	39.4	48.9	82.2	80.0	79.1	79.4	67.1	31.0	35.0	29.3	46.9	50.3	60.0
+ AD, BKD	83.3	39.5	86.7	69.8	78.5	58.0	85.7	88.2	33.8	41.1	51.1	85.7	83.4	82.5	82.8	70.0	32.4	36.5	30.6	49.0	52.5	62.6
+ CAF	85.4	40.5	88.9	71.6	80.5	59.4	87.8	90.3	34.6	42.1	52.3	87.8	85.5	84.5	84.9	71.7	33.2	37.4	31.4	50.2	53.8	64.1
+ CAF, BKD	86.1	40.8	89.6	72.1	81.1	59.9	88.6	91.1	34.9	42.4	52.7	88.5	86.2	85.2	85.6	72.3	33.5	37.8	31.6	50.6	54.2	64.6
+ CAF, AD	86.3	40.9	89.8	72.3	81.3	60.0	88.8	91.3	35.0	42.5	52.9	88.7	86.4	85.4	85.8	72.5	33.5	37.8	31.7	50.7	54.4	64.8
+ CAF, AD, BKD	**86.8**	**41.1**	**90.4**	**72.8**	**81.8**	**60.4**	**89.3**	**91.9**	**35.2**	**42.8**	**53.2**	**89.3**	**86.9**	**85.9**	**86.3**	**72.9**	**33.7**	**38.1**	**31.9**	**51.0**	**54.7**	**65.2**

TABLE IV

mIoU for different incremental learning scenarios.

Method	VOC 15-5			ADE 100-50		
	1-15	16-20	all	1-100	101-150	all
Only Step 1	**79.6**	-	-	**42.7**	-	-
Fine Tuning	1.1	33.6	9.2	0.0	24.9	8.3
Ours	72.9	**42.1**	**65.2**	37.3	**31.9**	**35.5**

TABLE V

Ablation study of our continual attentive fusion on the Pascal-VOC 2012 disjoint 15-5 setup and ADE disjoint 100-50 setup. "Baseline + KD" corresponds to the second row in Table III, "Projection" denotes the feature projection module, and "SAB" indicate the structured attention block. "Without continual attentive fusion" indicates that the attention weights only depend on the current module, while in "with continual attentive fusion" new and old model features are fused with the fusion module to generate the attention weights.

Method	VOC 15-5			ADE 100-50		
	1-15	16-20	mIoU	1-100	101-150	mIoU
Without continual attentive fusion						
Baseline + KD	65.4	37.6	58.5	**37.9**	27.9	34.6
+ Projection	65.4	38.0	58.6	36.5	31.3	34.8
+ Projection, SAB	66.5	38.2	59.4	36.9	31.5	35.1
With continual attentive fusion						
+ Fusion	70.7	38.7	62.7	36.8	31.5	35.0
+ Fusion, Projection	71.1	40.6	63.5	37.1	31.7	35.3
+ Fusion, Projection, SAB (CAF)	**71.7**	**41.2**	**64.1**	37.3	**31.9**	**35.5**

backbone, DeepLab-v3 architecture [64] as the refinement model and nonlocal block [57] as the feature projection module. We initialize our backbone with ImageNet pretraining [65] and train the full network as in [64] for learning rate policy, momentum, and weight decay. We use an initial learning rate of 10^{-2} for the first learning step and 10^{-3} and 10^{-2} for the following steps in the Pascal-VOC 2012 and ADE20K datasets, respectively. We train the model with a batch size of 24 for 30 epochs for Pascal-VOC 2012 and 60 epochs for ADE20K in every learning step. For our loss, λ_D and λ_{AD} are set to 10 and 1,000, respectively. We apply the same data augmentation of [64] and crop the images to 512 × 512 during both training and testing. To set the hyperparameters of each method, we use the protocol of IL defined in [9], [32], using 20% of the training set as validation. The final results are reported on the standard validation sets.

B. Experimental Results

We report some quantitative and qualitative results associated with our method, as well as the results of an ablation study, to demonstrate the merit of our technical contributions.

Comparison with State-of-the-Art Methods. Table I and Table II compare our approach with state-of-the-art ICL methods. Looking at the results in Table I, it is clear that in the case of the Pascal-VOC 2012 dataset, our method outperforms all the competitors in almost all the overlapped and disjoint settings, often by a large margin. Comparative results on ADE20K are shown in Table II. Our model exhibits competitive performance in all tasks and often better performance than the current art by several points. Only the last step of the 100-10 task is associated with lower performance, but it is compensated by far if we consider the overall task score.

Demonstration of Catastrophic Forgetting. The catastrophic forgetting phenomenon is clearly shown in Table IV. Fine-tuning suffers from catastrophic forgetting (-78.5% on VOC and -42.7% on ADE20K) on the first task after training the second task. In contrast, the performance of our method only decreases by 6.7% on VOC and 5.4% on ADE20K, which demonstrates the effectiveness of our method for addressing the issue of catastrophic forgetting.

Ablation Study. We perform an ablation study on the VOC 2012 dataset to demonstrate the impact of each component of our model. Table III shows the variants of our method, obtained by gradually adding one component at a time. We decided to use a very simple baseline, which takes no precautions against catastrophic forgetting, apart from using the revisited cross-entropy loss $\mathcal{L}_{\text{SEG}}$ as defined in [9]. In the top part of the table, we test different combinations of loss functions without considering the CAF module. Undoubtedly, according to the results, balancing knowledge distillation (BKD) (see Section III-C and (12)) increases the performance over the standard distillation loss formulation (KD). The results show that attentive feature distillation (AD) mitigates catastrophic forgetting significantly better than just distilling the output probability distribution, improving the mIoU by more than 4%. The model achieves the best performance when AD and BKD are combined. In the bottom part of Table III, we report the same experiments but at this time activate our CAF module. Interestingly, when no IL technique is used to alleviate catastrophic forgetting, adding the CAF module dramatically

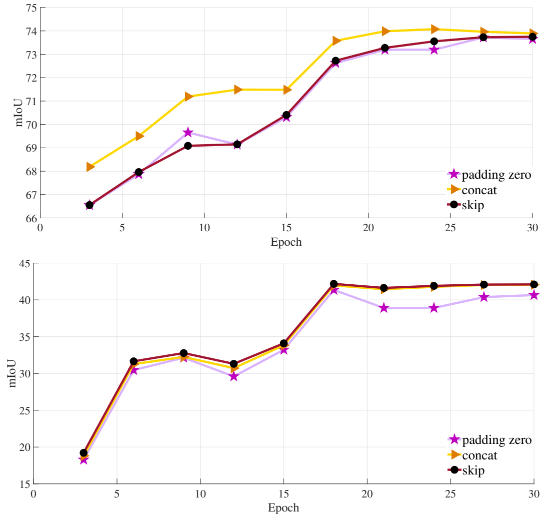


Fig. 5. Ablation study: comparison of different strategies for using the fusion module (setting 15-5, top: old categories, bottom: new categories). The curves refer to the concatenation of $\mathbf{v}^{\ell}$ and zero padding (purple), to the concatenation of $\mathbf{v}^{\ell}$ and $\mathbf{v}^{\ell-1}$ (yellow) and to skipping the fusion module and using $\mathbf{v}^{\ell}$ (red).

TABLE VI

ABLATION STUDY ON ATTENTION METHODS ON PASCAL-VOC 2012. FD DENOTES NORMAL FEATURE DISTILLATION (WITHOUT ATTENTION).

Method	15-5 disjoint			19-1 disjoint		
	1-15	16-20	all	1-19	20	all
Baseline	59.3	34.1	53.0	69.7	24.7	67.4
+ FD	63.2	39.5	57.3	60.3	16.3	58.1
+ AD_{SP}	71.5	42.7	64.3	74.2	29.2	71.9
+ AD_{CH}	10.5	11.2	10.6	1.7	16.2	2.4
+ AD_{SP} & AD_{CH}	**72.5**	**41.6**	**64.8**	**75.1**	**30.0**	**72.8**

increases the accuracy (+11%). As before, both BKD and AD further improve performance, and the whole model outperforms all the other variants.

In addition, we also show the results of an ablation study on the structure of the CAF module in Table V. First, in the top block, we show that our architectural modifications (feature projection module, structured attention block) have minimal impact if not used in composition with continual attentive fusion, showing less than 1% improvement. Instead, a significant improvement (approximately 4%) is achieved using a continual attentive fusion scheme (fusion module is active at training time). Moreover, feature projection and SAT seem to be more helpful when used in combination with CAF. This evidence suggests that the improvement does not come from the additional parameters introduced in the feature projection and structured attention block. If that were the case, the performance would be boosted even without using the old model. Rather, it is clear that the information transfer between the two models in the CAF module is the real catalyst that mitigates catastrophic forgetting.

Finally, we also analyze strategies for minimizing the computational complexity and the memory usage at test time. In particular, we believe that a truly continual learner should retain only one model, i.e., the old model used for distillation and continual attentive fusion should be discarded.

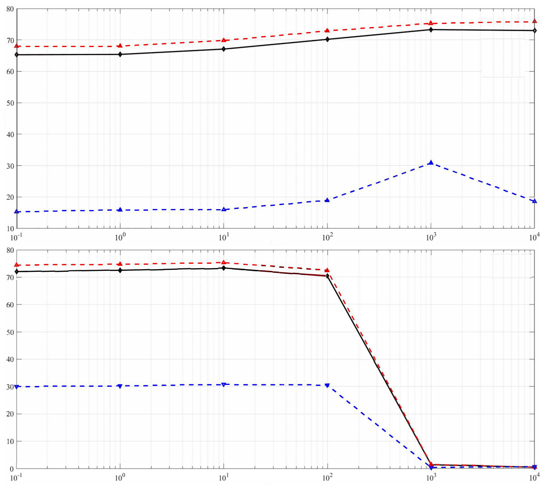


Fig. 6. The mIoU of the method for different values of the loss weights on the VOC2012 dataset (19-1 disjoint setup). (a) λ_{AD} is fixed at 1,000. (b) λ_{D} is fixed to 10. The black line denotes the overall mIoU, the red dashed line indicates the old class mIoU, and the blue dashed line indicates the new class mIoU.

TABLE VII

ABLATION STUDY ABOUT APPLIED POSITION ON PASCAL- VOC 2012 DATASET.

Combination	15-1 disjoint			19-1 disjoint		
	1-15	16-20	all	1-19	20	all
$z_{\ell} + h_{\ell}$	**57.2**	**15.5**	**46.7**	**75.5**	**30.8**	**73.3**
h_{ℓ}	56.2	15.2	45.9	74.3	29.6	72.1
z_{ℓ}	56.4	15.3	46.1	74.8	30.1	72.5

Hence, in Figure 5, we show different solutions for using the fusion module at test time. Possible strategies include (i) **skip**: completely skipping the fusion module at test time, *i.e.* passing only $\mathbf{v}^{\ell}$ to subsequent layers; (ii) **padding zero**: concatenating $\mathbf{v}^{\ell}$ with a zero-padding vector of the same size and evaluating the fusion module; (iii) **concat**: concatenating $\mathbf{v}^{\ell}$ and $\mathbf{v}^{\ell-1}$ at training time. To compare the performance of these strategies, we evaluate the accuracy of the network on both new and old classes every three epochs during the training trajectory of the second task (15-5 setting). Interestingly, we obtain comparable overall performance when skipping the fusion module (strategy (i)), while zero-padding seems slightly suboptimal. Additionally, it is important to note that at the beginning of the training, concatenating old and new features brings great performance improvements on the old classes, while the difference is negligible at convergence. It validates the idea that the old model is helping the new model through continual attentive fusion, transferring valuable information that the new model then uses to counteract catastrophic forgetting. However, as expected, the accuracy is unchanged when using concatenation on the new classes. This is reasonable since the old model does not possess any knowledge of the new classes and therefore cannot help the new model.

Sensitivity analysis of $\lambda_{\mathbf{AD}}$ and $\lambda_{\mathbf{D}}$. We provide a more thorough analysis of the loss weighting parameters λ_{AD} and λ_{D}; see Figure 6. We evaluate the average IoU of our method for different values of the loss weights around the working

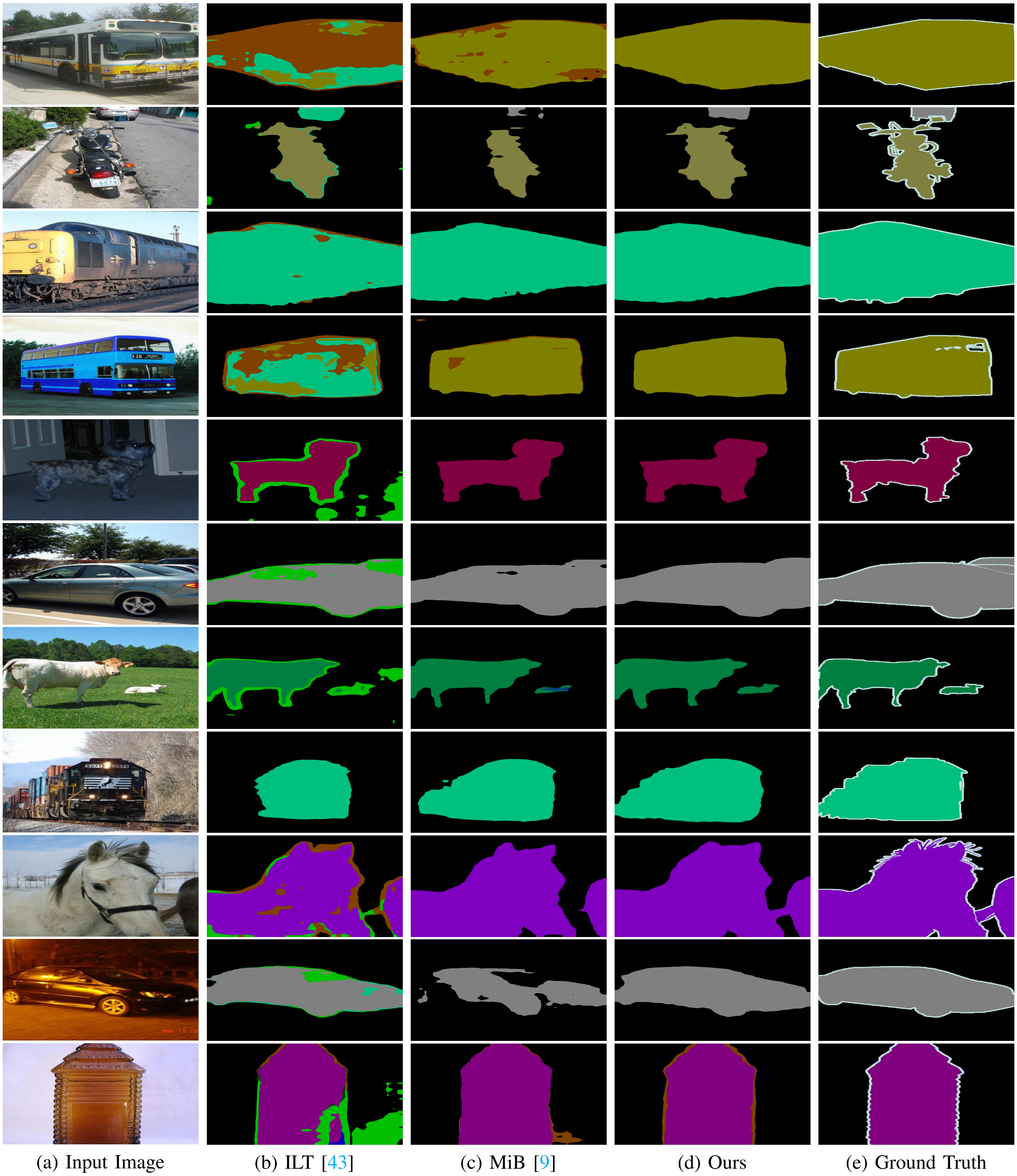

(a) Input Image (b) ILT [43] (c) MiB [9] (d) Ours (e) Ground Truth

Fig. 7. Qualitative results on the VOC 2012 dataset 19-1 (the first five rows) and 15-5 (the last six rows). They show the superiority of our approach on both new (*e.g.* train) and old (*e.g.* car, cow, bus) classes.

point. We can see that λ_{AD} is more of a critical choice than λ_{D}. Indeed, the operating region of λ_{AD} is approximately 10^3, while the operating region of λ_{D} is much larger, and its choice does not have much impact as long as it is kept below 10^2.

Attentive Feature Distillation Impact. In this section, we deeply analyze the effect of our attentive feature distillation. The results are shown in Table VI. In the table, FD, AD_{SP} and AD_{CH} denote normal feature distillation without attention,

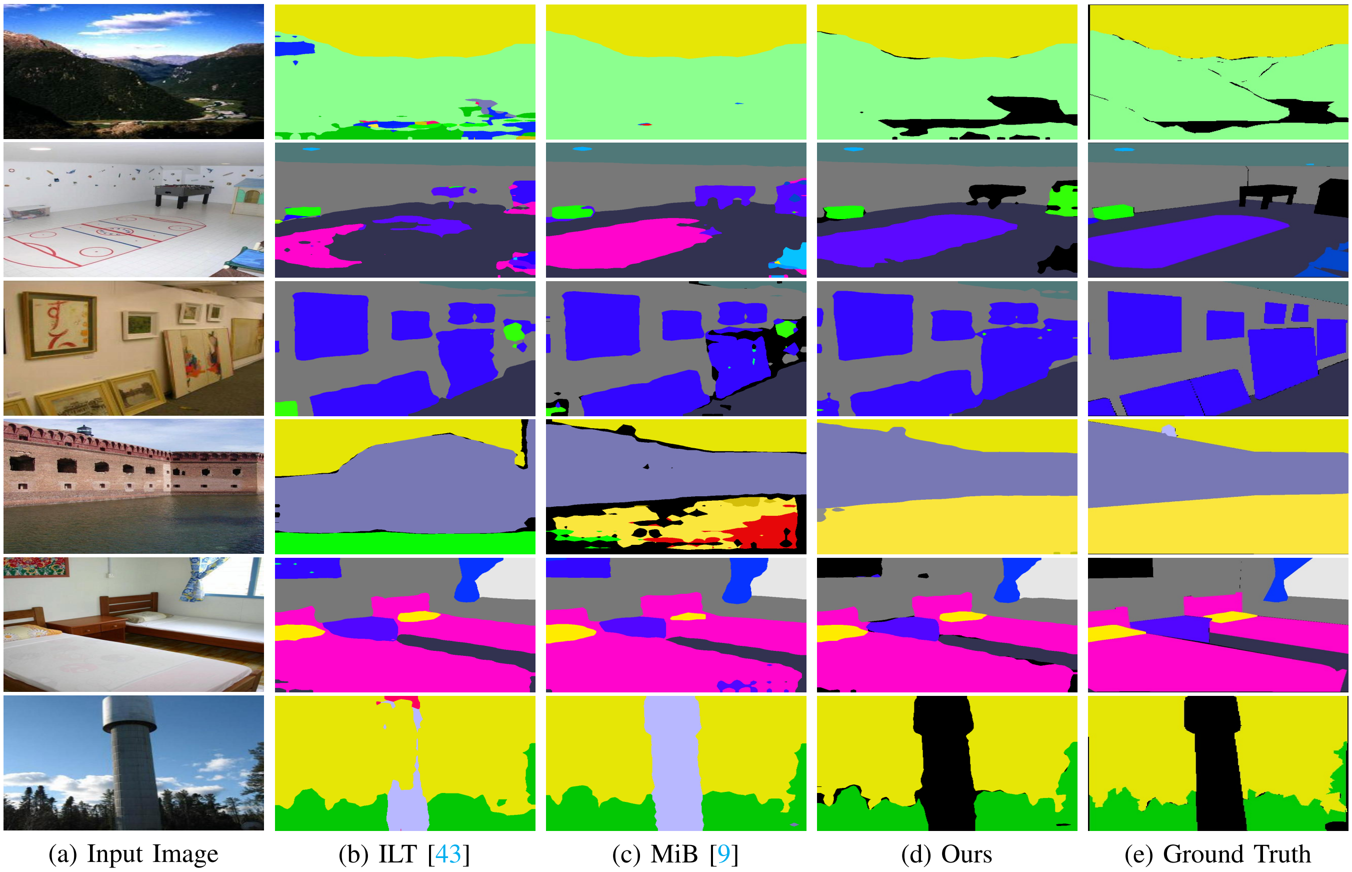


Fig. 8. Qualitative results on the ADE20K dataset using the 100-50 setup. The image demonstrates the superiority of our approach on both new (*e.g.* sky, grass, wall) and old (*e.g.* paint, pool, building) classes.

spatialwise attentive feature distillation, and channelwise attentive feature distillation, respectively. According to Table VI, adding only channelwise attentive feature distillation leads to training failure, while spatialwise attentive feature distillation significantly improves the performance. The training failure phenomenon with channelwise attention is caused by collapsing features in the same attention layers. We believe that the structured attention block prevents this collapse because the interaction between the channelwise and spatial attention makes it easier for the model to leverage channel information. Additionally, a combination of both attentive feature distillation further boosts the result. This proves that the performance improvement is not due to complex attention modules such as the squeeze-and-attention module but the choice of an appropriate attention module and a suitable combination. Moreover, we also run a position sensitivity analysis of attentive feature distillation in Table VII. According to the results, we choose $z_\ell + h_\ell$ as the input of attentive feature distillation for all experiments. The results also confirm that using the old model $\phi^{\ell-1}$ during training can significantly improve the performance compared to only using ϕ^{ℓ}. Furthermore, using the CAF module without the old model does not improve performance, suggesting that the improvement does not come from the additional parameters introduced in the CAF module.

Qualitative Results. The qualitative results associated with our method on the ADE20K and VOC 2012 datasets are shown in Figures 7 and 8. We found that CAF not only preserves more knowledge on the old classes with respect to MiB and ILT but also produces accurate segmentations for the objects of the new categories.

Figure 9 shows the predictions for both MiB and our method on VOC 15-1 across time. It seems clear that MiB quickly forgets the previous classes and becomes biased toward new classes. In contrast, our method's predictions are much more stable, owing to the CAF module for alleviating catastrophic forgetting by spatially constraining representations and to attentive feature distillation for dealing with the background shift. To analyze the relationship between continual attentive fusion and incremental learning, we show the qualitative results of our attention on the VOC 15-1 disjoint task in Figure 10. In detail, the first and third rows show the networks' attention on each step, while the second and fourth rows show the known (seen) classes on each step. According to Figure 10, due to the continual attentive fusion, our model can keep focusing on the old classes. When a new class is introduced (*e.g.* plant, television), the model can focus on the new object without losing attention for the old objects.

Figure 11 shows examples of the attention maps that our model learns on ADE20K (50-50 setting). It is quite clear that the network can focus on regions containing objects of the old classes (person, chair, building) as well as new classes (sideboard, animal, shower).

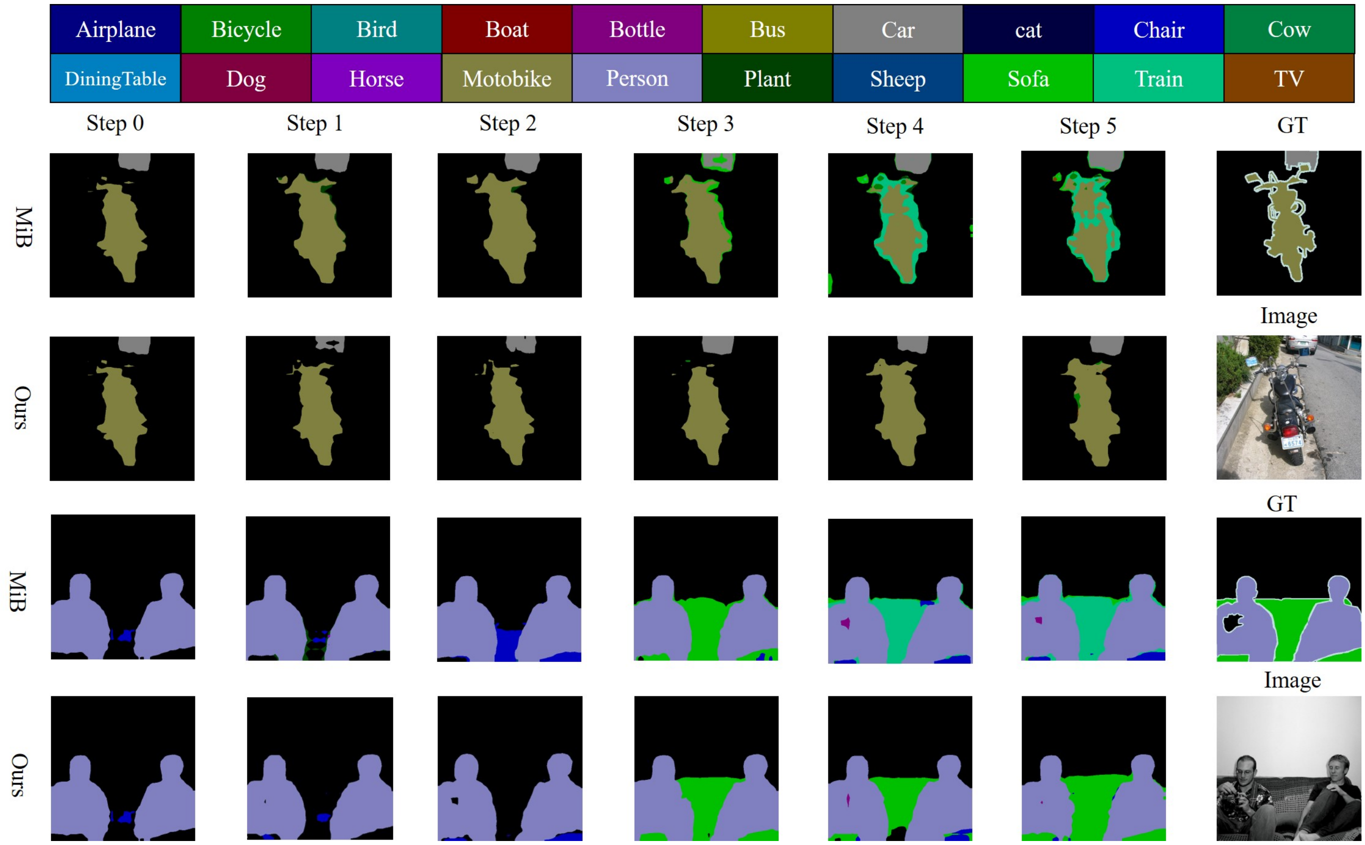


Fig. 9. Visualization of MiB and our method predictions across time in VOC 15-1 for two test images.

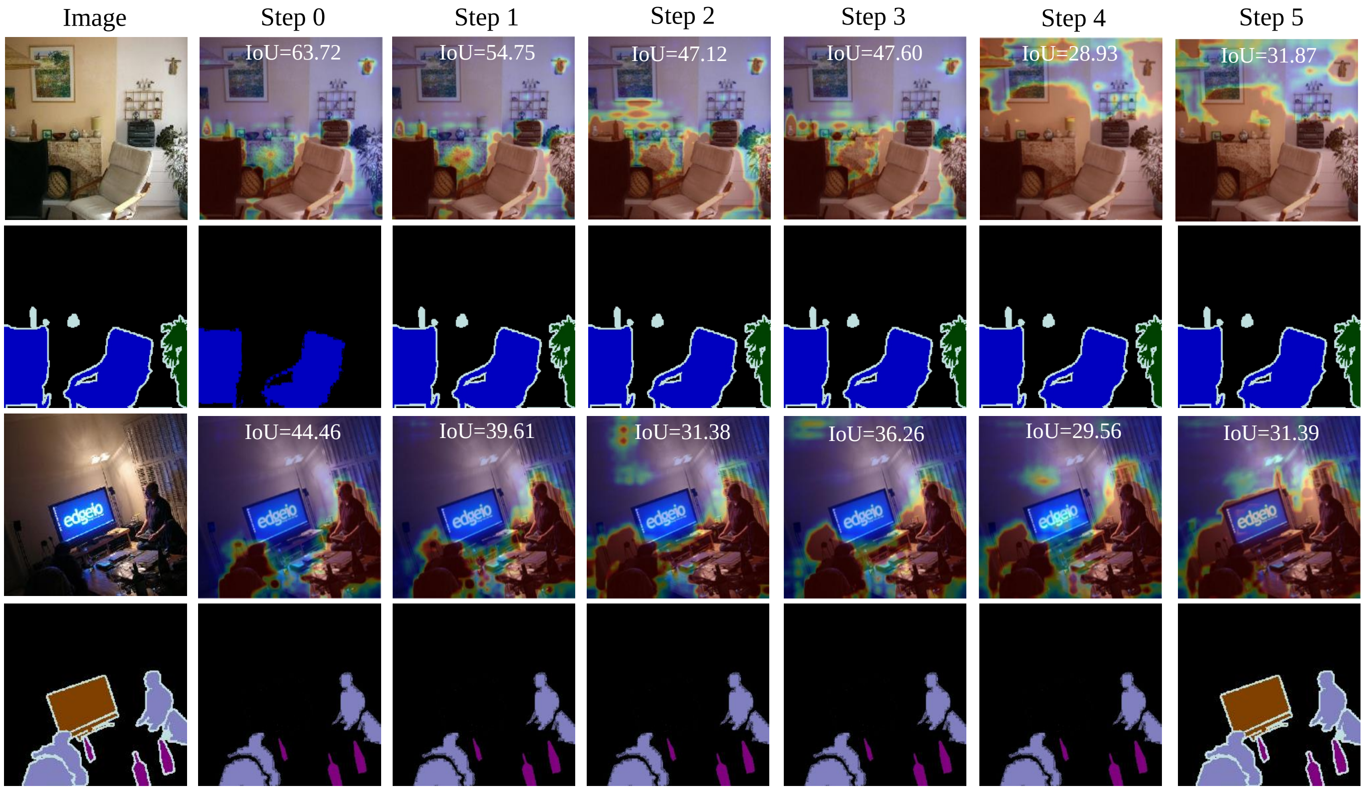


Fig. 10. Visualization of our method's attention across time in VOC 15-1 for two test images. The IoU is calculated between the old class ground truth and the attention map (Threshold=0.5).

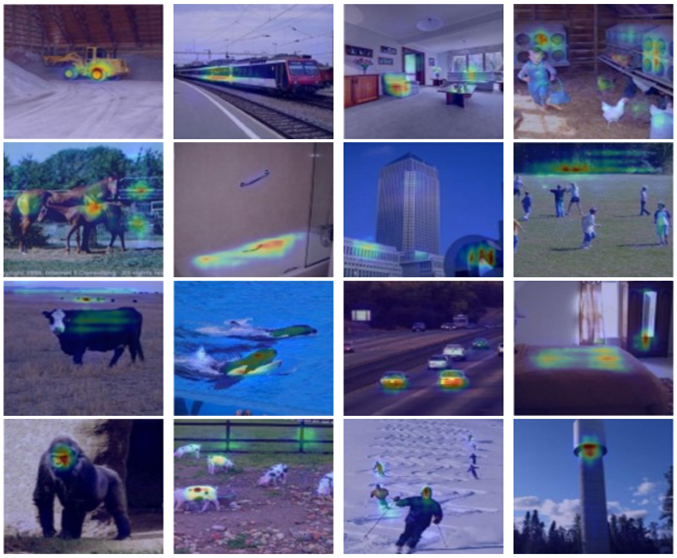


Fig. 11. Attention map on the ADE20K dataset (50-50).

V. CONCLUSIONS

We proposed the first attention-based ICL method for semantic segmentation. Our methodological contribution is threefold. First, a new continual attentive fusion module that updates the current features by using the information of the previous model was proposed. While the information from the previous model is used at training time through a fusion module, it is discarded at test time to save resources. We also proposed a new attentive distillation loss that leverages both channelwise and spatial attention to transfer compelling information. Finally, we introduced a new method for balancing old and new background probabilities in the distillation loss. Our extensive experimental evaluation demonstrates the outstanding performance of our method in several datasets (VOC 2012 and ADE20K) and settings (14 in total).

ACKNOWLEDGMENT

The authors would like to thank Nvidia Corporation for GPU donations to support our research. This work was also supported by the Caritro Foundation and the Deep Learning lab of the ProM Facility, as well as by the ANR-JCJC ML3RI (ANR-19-CE33-0008-01) and H2020 SPRING (funded by EC under GA #871245) projects.

REFERENCES

[1] S. Minaee, Y. Boykov, F. Porikli, A. Plaza, N. Kehtarnavaz, and D. Terzopoulos, "Image segmentation using deep learning: A survey," *arXiv preprint arXiv:2001.05566*, 2020. 1

[2] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *CVPR*, 2015. 1, 2

[3] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *ECCV*, 2018. 1, 2

[4] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *CVPR*, 2017. 1, 2

[5] Z. Zhang, X. Zhang, C. Peng, X. Xue, and J. Sun, "Exfuse: Enhancing feature fusion for semantic segmentation," in *ECCV*, 2018. 1, 2

[6] H. Zhang, K. Dana, J. Shi, Z. Zhang, X. Wang, A. Tyagi, and A. Agrawal, "Context encoding for semantic segmentation," in *CVPR*, 2018. 1, 2

[7] Y. Yuan, X. Chen, and J. Wang, "Object-contextual representations for semantic segmentation," *ECCV*, 2020. 1

[8] M. McCloskey and N. J. Cohen, "Catastrophic interference in connectionist networks: The sequential learning problem," in *Psychology of learning and motivation*. Elsevier, 1989, vol. 24, pp. 109–165. 1, 2

[9] F. Cermelli, M. Mancini, S. R. Bulo, E. Ricci, and B. Caputo, "Modeling the background for incremental learning in semantic segmentation," in *CVPR*, 2020. 1, 2, 5, 6, 7, 9, 10

[10] Z. Li and D. Hoiem, "Learning without forgetting," *TPAMI*, vol. 40, no. 12, pp. 2935–2947, 2017. 1, 2, 4, 6

[11] J. Kirkpatrick, R. Pascanu, N. Rabinowitz, J. Veness, G. Desjardins, A. A. Rusu, K. Milan, J. Quan, T. Ramalho, A. Grabska-Barwinska, *et al.*, "Overcoming catastrophic forgetting in neural networks," *Proceedings of the national academy of sciences*, vol. 114, no. 13, pp. 3521–3526, 2017. 1, 2, 6

[12] A. Chaudhry, P. K. Dokania, T. Ajanthan, and P. H. Torr, "Riemannian walk for incremental learning: Understanding forgetting and intransigence," in *ECCV*, 2018. 1, 2, 6

[13] S.-A. Rebuffi, A. Kolesnikov, G. Sperl, and C. H. Lampert, "icarl: Incremental classifier and representation learning," in *CVPR*, 2017. 1, 2, 6

[14] S. Hou, X. Pan, C. C. Loy, Z. Wang, and D. Lin, "Learning a unified classifier incrementally via rebalancing," in *CVPR*, 2019. 1, 2

[15] A. Douillard, M. Cord, C. Ollion, T. Robert, and E. Valle, "Podnet: Pooled outputs distillation for small-tasks incremental learning: Supplementary material," *ECCV*, 2020. 1

[16] A. Iscen, J. Zhang, S. Lazebnik, and C. Schmid, "Memory-efficient incremental learning through feature adaptation," *arXiv preprint arXiv:2004.00713*, 2020. 1

[17] E. Fini, S. Lathuilière, E. Sangineto, M. Nabi, and E. Ricci, "Online continual learning under extreme memory constraints," in *ECCV*, 2020. 1, 2

[18] K. Shmelkov, C. Schmid, and K. Alahari, "Incremental learning of object detectors without catastrophic forgetting," in *ICCV*, 2017. 1, 2, 6

[19] X. Liu, H. Yang, A. Ravichandran, R. Bhotika, and S. Soatto, "Multi-task incremental learning for object detection," in *CVPR*, 2020. 1, 2, 4

[20] W. Chen, Y. Liu, N. Pu, W. Wang, L. Liu, and M. S. Lew, "Feature estimations based correlation distillation for incremental image retrieval," *IEEE Transactions on Multimedia*, 2021. 1

[21] X. Tian, W. Ng, H. Wang, and S. Kwong, "Complementary incremental hashing with query-adaptive re-ranking for image retrieval," *IEEE Transactions on Multimedia*, 2020. 1

[22] S. Thuseethan, S. Rajasegarar, and J. Yearwood, "Deep continual learning for emerging emotion recognition," *IEEE Transactions on Multimedia*, 2021. 1

[23] H. Zhang and M. Xu, "Weakly supervised emotion intensity prediction for recognition of emotions in images," *IEEE Transactions on Multimedia*, 2020. 1

[24] K. Fujii, D. Sugimura, and T. Hamamoto, "Hierarchical group-level emotion recognition," *IEEE Transactions on Multimedia*, 2020. 1

[25] L.-C. Chen, Y. Yang, J. Wang, W. Xu, and A. L. Yuille, "Attention to scale: Scale-aware semantic image segmentation," in *CVPR*, 2016. 2

[26] D. Xu, W. Ouyang, X. Alameda-Pineda, E. Ricci, X. Wang, and N. Sebe, "Learning deep structured multi-scale features using attention-gated crfs for contour prediction," in *NeurIPS*, 2017. 2

[27] J. Fu, J. Liu, H. Tian, Y. Li, Y. Bao, Z. Fang, and H. Lu, "Dual attention network for scene segmentation," in *CVPR*, 2019. 2

[28] H. Zhan, R. Garg, C. Saroj Weerasekera, K. Li, H. Agarwal, and I. Reid, "Unsupervised learning of monocular depth estimation and visual odometry with deep feature reconstruction," in *CVPR*, 2018, pp. 340–349. 2

[29] P. Dhar, R. V. Singh, K.-C. Peng, Z. Wu, and R. Chellappa, "Learning without memorizing," in *CVPR*, 2019. 2, 4

[30] M. Everingham, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes challenge 2007 results," 2007. 2, 6

[31] B. Zhou, H. Zhao, X. Puig, S. Fidler, A. Barriuso, and A. Torralba, "Scene parsing through ade20k dataset," in *CVPR*, 2017. 2, 6

[32] M. De Lange, R. Aljundi, M. Masana, S. Parisot, X. Jia, A. Leonardis, G. Slabaugh, and T. Tuytelaars, "Continual learning: A comparative study on how to defy forgetting in classification tasks," *arXiv preprint arXiv:1909.08383*, vol. 2, no. 6, 2019. 2, 7

[33] F. M. Castro, M. J. Marín-Jiménez, N. Guil, C. Schmid, and K. Alahari, "End-to-end incremental learning," in *ECCV*, 2018. 2

[34] H. Shin, J. K. Lee, J. Kim, and J. Kim, "Continual learning with deep generative replay," in *NeurIPS*, 2017. 2

[35] O. Ostapenko, M. Puscas, T. Klein, P. Jahnichen, and M. Nabi, "Learning to remember: A synaptic plasticity driven framework for continual learning," in *CVPR*, 2019. 2
[36] C. Wu, L. Herranz, X. Liu, J. van de Weijer, B. Raducanu, *et al.*, "Memory replay gans: Learning to generate new categories without forgetting," in *NeurIPS*, 2018. 2
[37] F. Zenke, B. Poole, and S. Ganguli, "Continual learning through synaptic intelligence," *Proceedings of machine learning research*, vol. 70, p. 3987, 2017. 2, 6
[38] A. Mallya and S. Lazebnik, "Packnet: Adding multiple tasks to a single network by iterative pruning," in *CVPR*, 2018. 2
[39] A. Mallya, D. Davis, and S. Lazebnik, "Piggyback: Adapting a single network to multiple tasks by learning to mask weights," in *ECCV*, 2018. 2
[40] Y. Wu, Y. Chen, L. Wang, Y. Ye, Z. Liu, Y. Guo, and Y. Fu, "Large scale incremental learning," in *CVPR*, 2019. 2
[41] R. Aljundi, F. Babiloni, M. Elhoseiny, M. Rohrbach, and T. Tuytelaars, "Memory aware synapses: Learning what (not) to forget," in *ECCV*, 2018. 2
[42] G. Lin, A. Milan, C. Shen, and I. Reid, "Refinenet: Multi-path refinement networks for high-resolution semantic segmentation," in *CVPR*, 2017. 2
[43] U. Michieli and P. Zanuttigh, "Incremental learning techniques for semantic segmentation," in *ICCV Workshops*, 2019. 2, 6, 9, 10
[44] O. Tasar, Y. Tarabalka, and P. Alliez, "Incremental learning for semantic segmentation of large-scale remote sensing data," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 12, no. 9, pp. 3524–3537, 2019. 2, 4, 6
[45] F. Ozdemir, P. Fuernstahl, and O. Goksel, "Learn the new, keep the old: Extending pretrained models with new anatomy and images," in *MICCAI*, 2018. 2
[46] F. Ozdemir and O. Goksel, "Extending pretrained segmentation networks with additional anatomical structures," *International journal of computer assisted radiology and surgery*, vol. 14, no. 7, pp. 1187–1195, 2019. 2
[47] T. Xiao, Y. Xu, K. Yang, J. Zhang, Y. Peng, and Z. Zhang, "The application of two-level attention models in deep convolutional neural network for fine-grained image classification," in *CVPR*, 2015. 2
[48] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, and Y. Bengio, "Show, attend and tell: Neural image caption generation with visual attention," in *ICML*, 2015. 2
[49] B. Duan, H. Tang, W. Wang, Z. Zong, G. Yang, and Y. Yan, "Audio-visual event localization via recursive fusion by joint co-attention," in *WACV*, 2021, pp. 4013–4022. 2
[50] J. K. Chorowski, D. Bahdanau, D. Serdyuk, K. Cho, and Y. Bengio, "Attention-based models for speech recognition," in *NeurIPS*, 2015. 2
[51] L. Ding, H. Tang, and L. Bruzzone, "Lanet: Local attention embedding to improve the semantic segmentation of remote sensing images," *IEEE TGRS*, vol. 59, no. 1, pp. 426–435, 2020. 2
[52] B. Duan, W. Wang, H. Tang, H. Latapie, and Y. Yan, "Cascade attention guided residue learning gan for cross-modal translation," in *ICPR*, 2021, pp. 1336–1343. 2
[53] Y. Li, H. Liu, and H. Tang, "Multi-modal perception attention network with self-supervised learning for audio-visual speaker tracking," in *AAAI*, 2022. 2
[54] D. Xu, W. Wang, H. Tang, H. Liu, N. Sebe, and E. Ricci, "Structured attention guided convolutional neural fields for monocular depth estimation," in *CVPR*, 2018, pp. 3917–3925. 2
[55] H. Zhao, Y. Zhang, S. Liu, J. Shi, C. Change Loy, D. Lin, and J. Jia, "Psanet: Point-wise spatial attention network for scene parsing," in *ECCV*, 2018. 2
[56] Z. Zhong, Z. Q. Lin, R. Bidart, X. Hu, I. B. Daya, Z. Li, W.-S. Zheng, J. Li, and A. Wong, "Squeeze-and-attention networks for semantic segmentation," in *CVPR*, 2020. 2
[57] X. Wang, R. Girshick, A. Gupta, and K. He, "Non-local neural networks," in *CVPR*, 2018. 3, 7
[58] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *CVPR*, 2018, pp. 7132–7141. 4
[59] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," *arXiv preprint arXiv:1503.02531*, 2015. 5
[60] U. Michieli and P. Zanuttigh, "Continual semantic segmentation via repulsion-attraction of sparse and disentangled latent representations," *CVPR*, 2021. 6
[61] A. Douillard, Y. Chen, A. Dapogny, and M. Cord, "Plop: Learning without forgetting for continual semantic segmentation," *CVPR*, 2021. 6
[62] S. Yan, J. Zhou, J. Xie, S. Zhang, and X. He, "An em framework for online incremental learning of semantic segmentation," *arXiv*, 2021. 6
[63] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *CVPR*, 2016. 6
[64] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam, "Rethinking atrous convolution for semantic image segmentation," *arXiv preprint arXiv:1706.05587*, 2017. 7
[65] S. Rota Bulò, L. Porzi, and P. Kontschieder, "In-place activated batchnorm for memory-optimized training of dnns," in *CVPR*, 2018. 7

Guanglei Yang received a B.S. degree in instrument science and technology from Harbin Institute of Technology (HIT), Harbin, China, in 2016. He is currently pursuing a Ph.D. degree at the School of Instrumentation Science and Engineering, Harbin Institute of Technology (HIT), Harbin, China. He has been working at the University of Trento as a visiting student since 2020. His research interests include domain adaptation and pixel-level prediction .

Enrico Fini is a Ph.D. student at the University of Trento. His research focuses on continual learning and self-supervised learning. He received a B.S. degree in computer engineering from the University of Parma, Italy, in 2015 and an M.S. degree in computer science and engineering from Politecnico di Milano, Italy, in 2019. In 2018, he spent one year at the European Space Astronomy Centre of the European Space Agency in Madrid, Spain, working on machine learning for automatic sunspot detection.

Dan Xu is an assistant professor in the Department of Computer Science and Engineering at HKUST. He was a postdoctoral research fellow in VGG at the University of Oxford. He was a Ph.D. student in the Department of Computer Science at the University of Trento. He was also an MM Lab research assistant at the Chinese University of Hong Kong. He received the best scientific paper award at ICPR 2016 and a Best Paper Nominee at ACM MM 2018. He served as an area chairs of ACM MM 2020, WACV 2021 and ICPR 2020.

Paolo Rota is an assistant professor (RTDa) at the University of Trento (in the MHUG group) working on computer vision and machine learning. He received his Ph.D. in Information and Communication Technologies from the University of Trento in 2015. Prior to joining UniTN, he worked as post-doc at the TU Wien and the Italian Institute of Technology (IIT) of Genova. He is also collaborating with the ProM facility of Rovereto on assisting companies in inserting machine learning into their production chain.

Mingli Ding received the B.S., M.S. and Ph.D. degrees in instrument science and technology from Harbin Institute of Technology (HIT), Harbin, China, in 1996, 1997 and 2001, respectively. He worked as a visiting scholar in France from 2009 to 2010. Currently, he is a professor in the School of Instrumentation Science and Engineering at Harbin Institute of Technology. Prof. Ding's research interests are intelligence tests and information processing, automation test technology, computer vision, and machine learning. He has published over 40 papers in peer-reviewed journals and conferences.

Hao Tang is currently a postdoctoral with Computer Vision Lab, ETH Zurich, Switzerland. He received a master's degree from the School of Electronics and Computer Engineering, Peking University, China and a Ph.D. degree from Multimedia and Human Understanding Group, University of Trento, Italy. He was a visiting scholar in the Department of Engineering Science at the University of Oxford. His research interests are deep learning, machine learning, and their applications to computer vision.

Xavier Alameda-Pineda is a (tenured) Research Scientist at Inria and the Leader of the RobotLearn Team. He obtained an M.Sc. (equivalent) in Mathematics in 2008, in Telecommunications in 2009 from BarcelonaTech, and in Computer Science in 2010 from Université Grenoble-Alpes (UGA). He then worked toward his Ph.D. in mathematics and computer science, which he obtained in 2013 from UGA. After a two-year postdoc period at the Multimodal Human Understanding Group, at the University of Trento, he was appointed to his current position. Xavier is an active member of SIGMM, a senior member of IEEE, and a member of ELLIS. He is the coordinator of the H2020 Project SPRING: Socially Pertinent Robots in Gerontological Health care and is coleading the "Audio-visual machine perception and interaction for companion robots" chair of the Multidisciplinary Institute of Artificial Intelligence. Xavier's research interests are in combining machine learning, computer vision, and audio processing for scene and behavior analysis and human-robot interaction.

Elisa Ricci received a Ph.D. degree from the University of Perugia in 2008. She is an associate professor at the University of Trento and a researcher at Fondazione Bruno Kessler. She has since been a postdoctoral researcher at Idiap, Martigny, and Fondazione Bruno Kessler, Trento. She was also a visiting researcher at the University of Bristol. Her research interests are in the areas of computer vision and machine learning. She is a member of the IEEE.