FULL STEP PARAMETER-EFFICIENT CONTINUAL IMAGE SEGMENTATION VIA LORA ADAPTATION

Anonymous authors
Paper under double-blind review

ABSTRACT

In recent years, prevailing class-incremental image segmentation approaches have heavily relied on knowledge distillation—including logit distillation, feature distillation, and pseudo-labeling—which incur substantial training overhead. Meanwhile, vision segmentation models are rapidly increasing in scale; this discrepancy presents a severe dilemma, as heavy distillation paradigms fail to scale sustainably to large foundation models. In this paper, we propose a full step parameter-efficient continual image segmentation framework based on Low-Rank Adaptation (LoRA). Our approach enables lightweight and efficient adaptation of pretrained segmentation models to novel classes from the base step through all subsequent incremental steps. Specifically, we devise a task-specific gradient subspace projection mechanism to minimize mutual interference among LoRA adapters across different steps, thereby mitigating catastrophic forgetting. Furthermore, a background classifier calibration strategy is introduced to alleviate background drift, complemented by a prototype-based loss that widens class-wise separation to suppress category confusion. Extensive experiments on ADE20K and Pascal VOC 2012 demonstrate the effectiveness of our method for semantic segmentation. In addition, we present an exploratory investigation into the applicability of LoRA fine-tuning for panoptic segmentation.

1 INTRODUCTION

Class-incremental image segmentation aims to avoid catastrophic forgetting of previously learned categories when incrementally learning new semantic categories. With the rapid advancement of large vision models, leading segmentation architectures—such as the SAM series Kirillov et al. (2023); Ravi et al. (2024); Carion et al. (2026), InternViT-6B Chen et al. (2024b), BEiT-3 Wang et al. (2023b), ONE-PEACE Wang et al. (2023a), and DINOv3 Siméoni et al. (2025)—typically comprise a large number of parameters. However, parameter-efficient approaches remain largely underexplored in class-incremental image segmentation. Mainstream methods predominantly rely on knowledge distillation Park et al. (2024); Zhu et al. (2025); Cermelli et al. (2023). During each incremental stage, these paradigms typically leverage the previous model for supervision via logit distillation, feature distillation, and pseudo-labeling. Nevertheless, this paradigm inevitably incurs substantial memory overhead, posing a severe bottleneck when scaling continual learning to large-scale segmentation models. Therefore, developing parameter-efficient continual image segmentation frameworks is of paramount importance.

Specifically, distillation-based methods typically resort to conventional full-parameter training in the base step, and thus are not parameter-efficient from the outset. In subsequent incremental steps, knowledge distillation requires additional forward passes through the predecessor model, incurring even higher computational and memory overhead than standard training. Meanwhile, the leading PEFT-based framework, ECLIPSE Kim et al. (2024), employs Visual Prompt Tuning (VPT) Jia et al. (2022) during incremental steps; however, it still requires full-parameter optimization at the base step, and therefore does not ensure efficiency throughout the learning pipeline. As a result, such methods are inherently confined to training small-scale models from scratch, rather than facilitating the incremental adaptation of large segmentation models to novel categories.

As a prominent parameter-efficient fine-tuning (PEFT) paradigm, Low-Rank Adaptation (LoRA) Hu et al. (2021) adapts models to new tasks by updating only a fraction of parameters while keeping

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

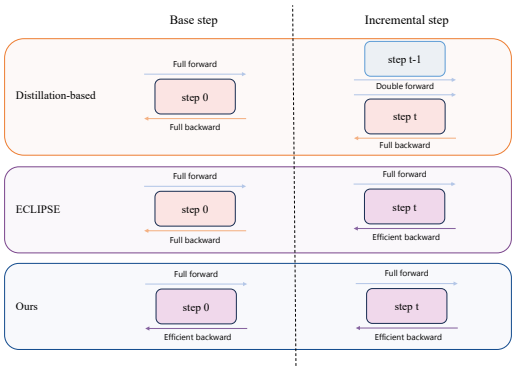


Figure 1: **Computational workflows of different continual segmentation pipelines**. Unlike distillation methods (requiring two forward passes and a full backward pass) and ECLIPSE (confined to efficient updates only at step t), our framework remains parameter-efficient, with lightweight backward updates throughout the entire process.

the pre-trained model frozen. However, sequentially updating LoRA adapters across incremental steps often suffers from representation drift and feature collapse, leading to severe catastrophic forgetting. In practice, continual segmentation fundamentally demands stable and well-separated feature representations. To stabilize continual adaptation, we jointly regularize both parameter updates and the feature space. We propose a full step parameter-efficient continual semantic segmentation framework based on LoRA adaptation. As illustrated in Figure 1, our framework applies parameter-efficient fine-tuning throughout all continual learning steps, from the base step to the final incremental step. To enlarge feature separation among classes across steps, we build on InfLoRA and LoRA Subtraction Liang & Li (2024); Liu & Chang (2025) and introduce a task-specific subspace projection mechanism tailored to class-incremental segmentation. Specifically, prior to training each incremental step, we compute uncentered covariance statistics of representations in the linear layers subject to LoRA adaptation. During training, parameter gradients are then projected onto the corresponding subspace derived from these statistics. This design consolidates intra-step features into dedicated subspaces, substantially mitigating inter-step cross-category interference.

Furthermore, inspired by prototype-based segmentation methods Zhou et al. (2022); Zhang et al. (2023), we devise a pixel-prototype loss to further enlarge intra-step class-wise margins. This loss comprises: (1) a hard contrastive loss that pushes pixel representations away from the top-k most confusing negative class prototypes, and (2) a margin loss that explicitly enforces a sample to be farther from hard negative classes than from its positive prototype. Consequently, this formulation encourages compact intra-class clustering and well-separated decision boundaries while confining updated gradients within the subspace, further reducing cross-step interference.

Finally, given that LoRA is inherently less effective on purely convolutional architectures, we select the widely adopted Mask2Former framework over alternatives like SegFormer Xie et al. (2021) or SAM Kirillov et al. (2023) to ensure a fair comparison with existing baselines.

The contributions of this paper are as follows:

- We propose a class-incremental image segmentation approach that remains fully parameter-efficient from the base step onward, maintaining a consistent training paradigm throughout the continual learning process.
- We adapt a LoRA-based continual learning method to image segmentation and conduct an exploratory study on the feasibility and limitations of LoRA fine-tuning for panoptic segmentation.
- With task-specific gradient subspace projection, a pixel-prototype loss, and background classifier calibration, our method substantially reduces interference among LoRA adapters and mitigates background drift.
- Our method outperforms the strongest fine-tuning-based class-incremental image segmentation baseline and remains competitive with heavy distillation-based methods.

2 Related Work

Image Segmentation. Semantic segmentation assigns each pixel to a semantic class. FCN Long et al. (2015) ushered in the era of deep learning-based semantic segmentation, and subsequent methods such as U-Net Ronneberger et al. (2015), PSPNet Zhao et al. (2017), and DeepLab Chen et al. (2017) further advanced this line of work. Mask R-CNN He et al. (2017) pioneered instance segmentation, which aims to delineate individual object instances. Panoptic segmentation Kirillov et al. (2019) unifies instance and semantic segmentation, handling "stuff" classes in the manner of semantic segmentation and "things" classes in the manner of instance segmentation. Mask2Former Cheng et al. (2022) proposed a transformer-based universal image segmentation head capable of addressing all three tasks. Recently, large segmentation models such as BEiT-3 and DINOv3 have employed Mask2Former as the segmentation head via ViT-Adapter Chen et al. (2023), achieving strong performance. This is one of the reasons we adopt Mask2Former as our base model.

Continual Segmentation. Since LwF Li & Hoiem (2017) introduced knowledge distillation into continual learning, many distillation-based continual image segmentation methods have emerged Fang et al. (2025); Cermelli et al. (2023); Zhu et al. (2025); Park et al. (2024). These methods mainly employ feature or logit distillation from the previous model to retain old knowledge while learning new classes. Following PLOP Douillard et al. (2021), several methods Zhang et al. (2023); Liang et al. (2025); Liu et al. (2026) adopt pseudo-labeling, using predictions of the previous model to supervise the new model, which can be viewed as a form of distillation. Replay-based methods Chen et al. (2024a); Zhu et al. (2025); Maracani et al. (2021) store images or features from previous steps and replay them during subsequent training. These methods incur growing storage costs and raise privacy concerns. Some methods combine distillation and replay: BalConpas Chen et al. (2024a) employs feature distillation and image replay, while SimCLS Zhu et al. (2025) employs feature distillation and feature replay. As the first fine-tuning-based method, ECLIPSE Kim et al. (2024) employs VPT Jia et al. (2022) to fine-tune the model in incremental steps, while still fully training the base step. In addition, ECLIPSE stores new prompts and MLPs at each incremental step and therefore also faces growing storage costs.

Reducing LoRA Interference. For multi-task problems in large language models, LoRI Zhang et al. (2025a) exploits the approximate orthogonality of Gaussian-initialized weights together with sparsity to reduce interference among tasks. FlyLoRA Zou et al. (2025) further employs an implicit router to mitigate interference. For class-incremental image classification, InfLoRA Liang & Li (2024) employs DualGPM Liang & Li (2023) to preserve gradient information of old tasks and uses current-task data to extract the gradient subspace of the new task. It then constrains the learnable low-rank update to the intersection of the new task's gradient space and the orthogonal complement of the old tasks' gradient space, thereby eliminating interference. Related efforts include Zhang et al. (2025b); He et al. (2025); Yipeng Lin & Yang (2026); Qiao & Mahdavi (2026); GUO et al. (2026); Luo et al. (2026).

3 Preliminaries

3.1 Problem Definition

In class-incremental image segmentation, a model learns to segment classes over a sequence of steps $t = 0, 1, 2, \ldots, T$. At each step, the model can access only the current dataset D_t, which consists of image–label pairs $\{x_t, y_t\}$, where x_t denotes an image and y_t denotes the corresponding ground-truth label. In D_t, y_t contains labels only for the current-step classes C_t. During training at step t, labels corresponding to previously seen classes $C_{0:t-1}$ and future classes $C_{t+1:T}$ are inaccessible.

Classes learned at different steps are disjoint, i.e., $C_{t_1} \cap C_{t_2} = \varnothing$ for $t_1 \neq t_2$. After training at step t, the model is required to segment classes in C_t without forgetting previously acquired segmentation capability for $C_{0:t-1}$.

3.2 Low-Rank Adaptation

Low-Rank Adaptation (LoRA) Hu et al. (2021) is a popular parameter-efficient fine-tuning (PEFT) method. For a pre-trained model M_0, let W_0^l denote the weight matrix of layer l, with input dimen-

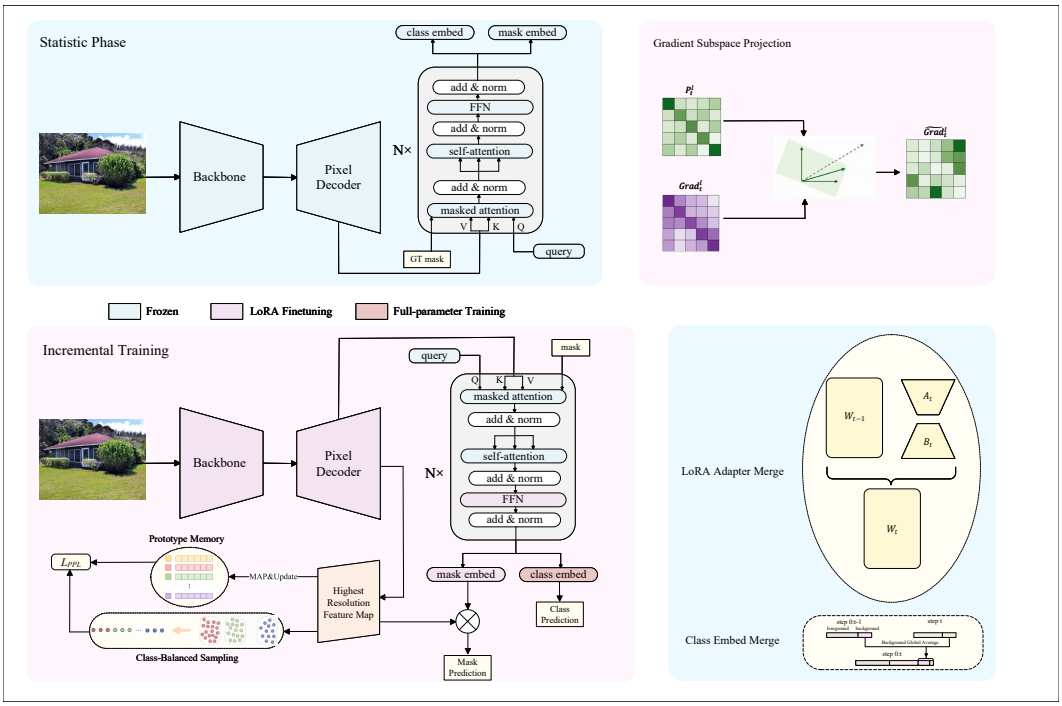


Figure 2: **Overview of our method**. We **extract incremental-step features** in the statistics phase, incrementally train the model with **gradient subspace projection** and a **pixel-prototype loss**, and merge LoRA adapters with **background classifier calibration** after training.

sion d_{in} and output dimension d_{out}. LoRA introduces matrices $A^l \in \mathbb{R}^{d_{\text{in}} \times r}$ and $B^l \in \mathbb{R}^{r \times d_{\text{out}}}$ to parameterize a low-rank update to W_0^l, where typically $r \ll \min(d_{\text{in}}, d_{\text{out}})$. The updated weights are given by:

$$W^l = W_0^l + \triangle W^l = W_0^l + A^l B^l \tag{1}$$

where $\triangle W^l$ denotes the parameter update of layer l and W^l is the updated weight matrix. With the pre-trained weight W_0^l frozen, we update only the low-rank matrices A^l and B^l.

For continual learning, we learn a separate pair of low-rank matrices at each step:

$$W_t^l = W_{t-1}^l + A_t^l B_t^l = W_0^l + \sum_{i=0}^{t} A_i^l B_i^l \tag{2}$$

where W_t^l denotes the weight matrix of layer l after step t, W_{t-1}^l denotes the weight matrix from the previous step, and $A_t^l B_t^l$ is the low-rank update at step t.

3.3 Base Model

LoRA is effective on linear layers such as feed-forward networks (FFNs) and attention projection layers, but it is less effective on convolutional layers. Therefore, we adopt Swin Transformer as the backbone and Mask2Former as the segmentation head, both of which are widely used in class-incremental image segmentation.

The pixel decoder of Mask2Former takes multi-scale feature maps from the backbone and produces finer multi-scale feature maps. The transformer decoder then uses these features together with a set of input queries to classify pixels via masked attention. Following the original Mask2Former formulation Cheng et al. (2022), for the l-th decoder layer, let $X_l \in \mathbb{R}^{N \times d}$ denote the query features and $F \in \mathbb{R}^{H \times W \times d}$ denote the feature maps produced by the pixel decoder. Masked attention is formulated as:

$$X_l = \text{softmax}\left(\frac{Q_l K_l^\top}{\sqrt{d}} + \mathcal{M}_{l-1}\right) V_l + X_{l-1} \tag{3}$$

$$\mathcal{M}_{l-1}(x,y) = \begin{cases} 0, & \text{if } M_{l-1}(x,y) = 1 \\ -\infty, & \text{otherwise} \end{cases} \tag{4}$$

where Q_l is a linear projection of X_{l-1}, and K_l and V_l are linear projections of the feature maps F. The attention mask $\mathcal{M}_{l-1}$ at spatial location (x, y) is derived from the binarized mask prediction of the $(l-1)$-th layer. Specifically, $M_{l-1}(x,y) \in \{0,1\}^{N \times H_l \times W_l}$ comes from the resized mask prediction of the previous transformer decoder layer.

By restricting each query to highly relevant foreground regions, this design yields more focused feature representations and consequently improves segmentation performance under standard training.

4 METHOD

4.1 REDUCING INTERFERENCE OF LORA ADAPTATION

Inspired by InfLoRA Liang & Li (2024) and LoRA Subtraction Liu & Chang (2025), we propose a LoRA adaptation method that reduces cross-step interference. Our approach is parameter-efficient from the first step. An overview is shown in Figure 2.

At the base step, we fine-tune the model on D_0 with standard LoRA. At each incremental step, we initialize a new LoRA adapter while keeping the pre-trained weights and accumulated updates from previous steps fixed. We then estimate the activation subspace of the current step and constrain the optimization of the new adapter accordingly.

4.1.1 EXTRACTING INCREMENTAL-STEP FEATURES

Before training at incremental step t ($t \neq 0$), we pass D_t through the model for one epoch without gradient computation and collect the inputs to each LoRA-adapted linear layer. The backbone and pixel decoder retain their standard forward behavior. In the transformer decoder, we replace the predicted cross-attention mask $M_{l-1}(x,y)$ in Eq. equation 4 with a mask derived from the current ground truth:

$$M_{l-1}(x,y) = \text{Binary}(y_t) \tag{5}$$

$$\text{Binary}(y_t) = \begin{cases} 0, \text{ if the pixel is assigned the ignore label} \\ 1, \text{ otherwise} \end{cases} \tag{6}$$

This mask restricts cross-attention and the corresponding activation statistics to valid annotated regions, yielding representations more relevant to the current step.

Let $a_{t,n}^l \in \mathbb{R}^{d_l}$ be the n-th valid input activation collected at layer l, where activations are flattened over images and spatial tokens, and let M_t^l be their total number. We characterize the current-step input space by the uncentered covariance

$$\mathcal{X}_t^l = \frac{1}{M_t^l} \sum_{n=1}^{M_t^l} a_{t,n}^l \left(a_{t,n}^l\right)^\top. \tag{7}$$

We accumulate the outer products batch by batch, avoiding storage of all activations. Since $\mathcal{X}_t^l$ is symmetric and positive semidefinite, its singular value decomposition is

$$\mathcal{X}_t^l = U_t^l \Sigma_t^l \left(U_t^l\right)^\top. \tag{8}$$

Let $\sigma_{t,j}^l$ denote the singular values in descending order. Following prior work, we retain the smallest k satisfying $\sum_{j=1}^{k} \sigma_{t,j}^l / \sum_{j=1}^{d_l} \sigma_{t,j}^l \geq \rho$, where ρ is a cumulative-energy threshold, and denote the truncated basis by $U_{t,k}^l$. The truncation captures dominant current-step directions while preventing the full-basis transform from reducing to an identity mapping.

4.1.2 CONSTRAINED OPTIMIZATION VIA GRADIENT PROJECTION

The truncated basis defines an orthogonal projector Q_t^l and its Frobenius-normalized transform P_t^l:

$$Q_t^l = U_{t,k}^l \left(U_{t,k}^l\right)^\top, \qquad P_t^l = \frac{Q_t^l}{\left\|Q_t^l\right\|_F}. \tag{9}$$

Method	**100-50** (2 steps)			**100-10** (6 steps)			**100-5** (11 steps)			**50-50** (3 steps)		
	1-100	101-150	all	1-100	101-150	all	1-100	101-150	all	1-50	51-150	all
FT	0.0	26.7	8.9	0.0	2.3	0.8	0.0	1.1	0.3	0.0	1.7	1.1
MiB Cermelli et al. (2020)	37.0	24.1	32.6	31.8	14.1	25.9	21.0	6.1	16.1	45.6	21.0	29.3
PLOP Douillard et al. (2021)	44.2	26.2	38.2	34.8	15.9	28.5	39.5	13.6	30.9	54.9	30.2	38.4
CoMFormer Cermelli et al. (2023)	44.7	26.2	38.4	40.6	15.6	32.3	39.5	13.6	30.9	49.2	26.6	34.1
CoMasTRe Gong et al. (2024)	45.7	26.0	39.2	42.3	18.4	34.4	40.8	15.8	32.6	49.8	26.6	34.5
ECLIPSE Kim et al. (2024)	45.0	21.7	37.1	43.4	17.4	34.6	43.3	16.3	34.2	-	-	-
Ours (Swin-S)	34.2	44.7	37.7	31.8	41.5	35.1	30.4	40.7	33.8	31.8	38.3	36.2
Ours (Swin-B)	38.2	48.5	**41.6**	36.1	46.5	**39.6**	34.5	44.9	**38.0**	35.1	42.6	**40.1**
Joint	44.3	54.3	47.6	44.3	54.3	47.6	44.3	54.3	47.6	41.2	50.9	47.6

Table 1: **Continual semantic segmentation** results on ADE20K in terms of mIoU.

Method	**19-1** (2 steps)			**15-5** (2 steps)			**15-1** (6 steps)		
	1-19	20	all	1-15	16-20	all	1-15	16-20	all
MiB Cermelli et al. (2020)	71.4	23.5	69.1	76.3	49.9	70.0	34.2	13.5	29.2
PLOP Douillard et al. (2021)	75.3	37.3	73.5	75.7	51.7	70.0	65.1	21.1	54.6
CoMFormer Cermelli et al. (2023)	75.3	24.0	72.9	74.6	48.4	68.4	48.9	23.2	42.5
CoMasTRe Gong et al. (2024)	75.1	69.5	74.8	79.7	51.9	73.1	69.7	43.6	63.5
Ours	93.9	62.6	**92.4**	91.2	83.1	**89.2**	81.6	23.9	**67.2**
Joint	94.1	97.0	94.3	94.6	93.3	94.3	94.6	93.3	94.3

Table 2: **Continual semantic segmentation** results on Pascal VOC 2012 in terms of mIoU.

Normalization controls the scale variation caused by different retained dimensions without changing the range of the projector. At optimization iteration i, let $D_{\theta,t}^{l,i}$ be the unconstrained preconditioned direction produced by AdamW for a compatible LoRA factor $\theta_t^l \in \{A_t^l, B_t^l\}$. We update it as

$$\widetilde{D}_{\theta,t}^{l,i} = \Pi_{P_t^l}\left(D_{\theta,t}^{l,i}\right), \qquad \theta_t^{l,i} = \theta_t^{l,i-1} - \alpha \widetilde{D}_{\theta,t}^{l,i}, \tag{10}$$

where α is the learning rate. The operator $\Pi_{P_t^l}$ left- or right-multiplies P_t^l along the matching feature dimension and applies it blockwise to fused projections; factors without a compatible dimension retain their original direction. Updating A_t^l and B_t^l independently preserves the low-rank parameterization. Consequently, the new adapter is biased toward directions supported by current-step representations, suppressing unconstrained drift when it is merged with previous updates and thereby mitigating cross-step interference.

4.2 Pixel-Prototype Loss

4.2.1 Class-Balanced Pixel Sampling

To alleviate severe class imbalance, we adopt a class-balanced pixel sampling strategy so that each category contributes equally to prototype learning. Class prototypes are computed and updated via masked average pooling (MAP) and momentum updates. Further details are provided in Appendix A.

4.2.2 Hard Negative Loss

Although gradient projection reduces cross-step interference, class-wise confusion can still degrade performance. For each sampled pixel n, let s_n^+ denote its cosine similarity to the positive prototype and $\mathcal{H}_n^k$ the set of its top-k negative similarities. We define

$$L_{\text{HardPPC}} = -\frac{1}{N}\sum_{n=1}^{N} \log \frac{\exp(s_n^+/\tau)}{\exp(s_n^+/\tau) + \sum_{s\in\mathcal{H}_n^k}\exp(s/\tau)}, \tag{11}$$

where τ is the temperature. We further enforce a margin m between the positive and hardest negative similarities:

$$L_{\text{Margin}} = \frac{1}{N}\sum_{n=1}^{N} \max\left(0, m + \max \mathcal{H}_n^k - s_n^+\right). \tag{12}$$

Table 3: Performance of LoRA fine-tuning for **panoptic segmentation**.

Method	PQ	SQ	RQ
LoRA Fine-tuning	36.487	72.595	43.726
Full Training	48.293	83.175	57.175

The pixel-prototype loss L_{PPL} and total loss L_{total} are

$$L_{\mathrm{PPL}} = L_{\mathrm{HardPPC}} + \lambda_m L_{\mathrm{Margin}} \tag{13}$$

$$L_{\mathrm{total}} = L_{\mathrm{Seg}} + \lambda_p L_{\mathrm{PPL}} \tag{14}$$

where L_{Seg} denotes the original segmentation loss, and λ_m and λ_p are hyperparameters that balance the loss terms.

4.3 Background Classifier Calibration

In continual image segmentation, the background classifier is optimized at every incremental step, whereas each foreground classifier is learned only when its corresponding class first appears. Consequently, after adapter merging, the background classifier accumulates substantially more optimization updates than the foreground classifiers, inducing a bias toward background predictions.

Let $\mathbf{w}_{bg}^{(t)}$ denote the background classifier learned at step t. We first obtain a merged background classifier by global averaging. To alleviate the accumulated background bias, we further apply a calibration factor:

$$\mathrm{w}_{bg}^{*} = \gamma \bar{\mathrm{w}}_{bg} = \gamma \left(\frac{\mathrm{t}}{\mathrm{t}+1} \mathrm{w}_{bg}^{0:t-1} + \frac{1}{\mathrm{t}+1} \mathrm{w}_{bg}^{\mathrm{t}} \right), (t > 0) \tag{15}$$

where γ down-weights over-confident background responses or up-weights under-confident ones. The calibrated background classifier is then used together with all foreground classifiers at inference time. The background weights of each step inherently encapsulate both cross-step shared background semantics and class-specific information associated with other steps. The proposed "merge-and-average" procedure reinforces the invariant shared background representations while attenuating the lingering class-specific components, thereby effectively suppressing background drift.

5 Experiments

5.1 Experimental Setup

Datasets and evaluation metrics. Following prior work Cermelli et al. (2023); Fang et al. (2025); Kim et al. (2024); Zhu et al. (2025), we evaluate our approach on ADE20K Balazevic et al. (2023) for semantic and panoptic segmentation. We further evaluate semantic segmentation on Pascal VOC 2012 Everingham et al. (2012), following prior studies Gong et al. (2024); Zhang et al. (2023); Park et al. (2024). ADE20K contains 150 classes and provides both semantic and panoptic annotations, whereas VOC 2012 contains 20 classes. Both datasets are widely used benchmarks for continual image segmentation.

Following previous work, we evaluate semantic segmentation with mean Intersection over Union (mIoU) and panoptic segmentation with Panoptic Quality (PQ), where PQ is composed of SQ (Segmentation Quality) and RQ (Recognition Quality). After the complete continual learning process, we report performance on base classes C_1 and incremental classes $C_{2:T}$.

Incremental protocol. Following previous continual image segmentation studies, we denote each setting as "base classes–incremental classes". For example, "100-10" indicates that 100 classes are learned at the base step and 10 classes are added at each incremental step.

Implementation details. We mainly evaluate our Mask2Former-based framework with a Swin-B backbone. For comparison with methods that use ResNet-101, we also report results with Swin-S, which has a comparable number of parameters. To avoid evaluating on data already seen during pre-training, we initialize from Mask2Former weights pre-trained on Cityscapes Cordts et al. (2016)

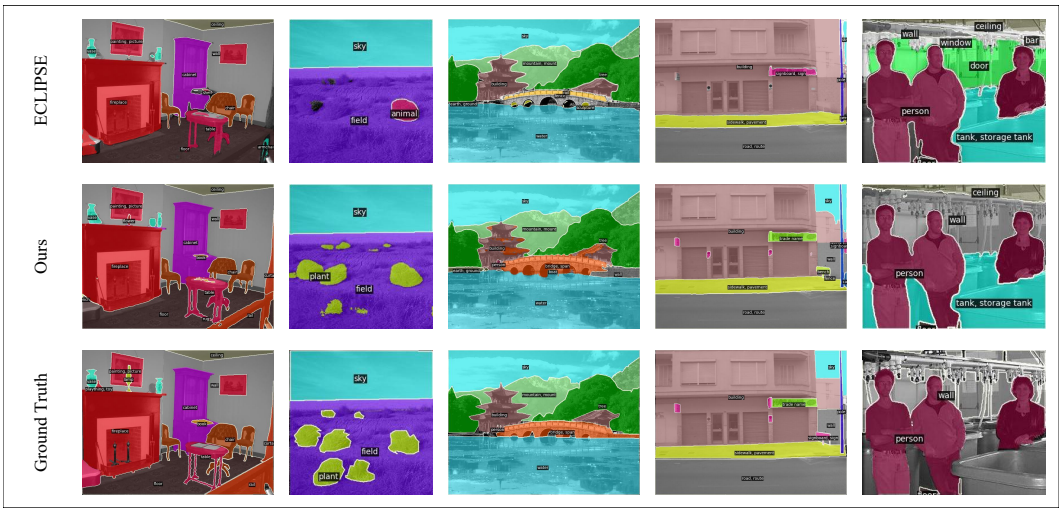


Figure 3: **Qualitative comparisons between our method and ECLIPSE** on the ADE20K 100-10 continual semantic segmentation task.

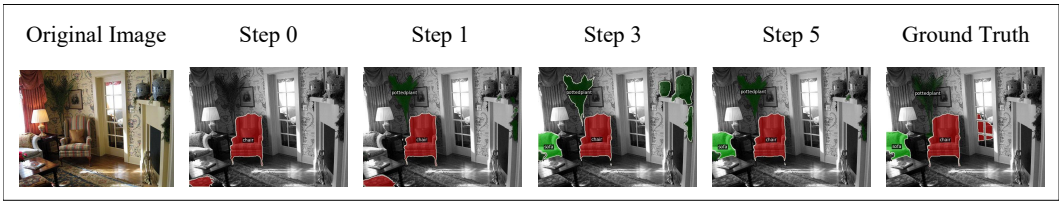


Figure 4: **Qualitative examples of continual learning** on the Pascal VOC 2012 15-1 continual semantic segmentation task.

for ADE20K and weights pre-trained on ADE20K for VOC2012. We set the LoRA rank to 64. On ADE20K, we shuffle the class order with random seed 42. Additional details are provided in Appendix B.

5.2 Main Results

Continual semantic segmentation on ADE20K. Table 1 reports our results on ADE20K. "FT" denotes training a plain Mask2Former model without our method, and "Joint" denotes training all classes at the base step with our method. Across all steps, only 10–20% of the parameters are trained. As shown in Table 1, with Swin-S, our method is competitive with heavy distillation-based approaches, and with Swin-B it outperforms them. Moreover, our Swin-S variant is competitive with or superior to ECLIPSE, which applies efficient updates only at incremental steps while fully training the base step. Qualitative results are shown in Figure 3.

Continual semantic segmentation on Pascal VOC 2012. Table 2 reports results on VOC 2012. Our method outperforms distillation-based approaches such as CoMFormer Cermelli et al. (2023) and CoMasTRe Gong et al. (2024), while remaining parameter-efficient. Qualitative continual learning results are shown in Figure 4.

Preliminary study of LoRA suitability for panoptic segmentation on ADE20K. Table 3 reports the performance of LoRA fine-tuning for panoptic segmentation. Even without a continual learning protocol, LoRA fine-tuning retains only about 75.5% of the PQ obtained by full-parameter training. Specifically, it retains 87.3% of Segmentation Quality (SQ) and 76.4% of Recognition Quality (RQ). These results indicate that LoRA preserves segmentation quality reasonably well but is less effective at maintaining recognition quality, and is therefore less suitable for panoptic segmentation. This finding helps delineate the boundaries of LoRA in dense prediction tasks, offering a preliminary reference for future research on tailored PEFT approaches for panoptic segmentation.

Table 4: Effect of gradient projection.

GP	AM	**100-50** (ADE20K) 1-100	101-150	all
		18.5	48.7	28.6
✓		37.3	48.9	41.2
✓	✓	38.2	48.5	41.6

Table 5: Effect of background classifier calibration.

BA	Scale	**100-50** (ADE20K) 1-100	101-150	all
		34.3	43.9	37.5
✓		37.8	48.3	41.3
✓	✓	38.2	48.5	41.6

Table 6: Effect of pixel-prototype loss.

$L_{HardPPC}$	L_{Margin}	**100-50** (ADE20K) 1-100	101-150	all	**15-5** (VOC2012) 1-15	16-20	all
		36.2	47.7	40.1	90.0	81.4	87.9
✓		37.8	47.9	41.1	91.2	81.1	88.6
✓	✓	38.2	48.5	41.6	91.2	83.1	89.2

6 Ablation Studies

Gradient projection. As shown in Table 4, without gradient projection (GP) the model substantially forgets knowledge from the previous step. Applying gradient projection with attention-mask (AM) filtering further reduces this forgetting, indicating that gradient projection effectively mitigates interference among LoRA adapters.

Background classifier calibration. As shown in Table 5, omitting background-weight averaging leads to a sharp performance drop, while background-weight scaling yields a further slight improvement.

Pixel-prototype loss. As shown in Table 6, L_{HardPPC} slightly improves performance on the current step and reduces interference with the previous step. The auxiliary loss L_{Margin} concentrates more on the performance of the current step.

Additional ablation results are provided in Appendix C.

7 Conclusion

In this paper, we propose a full step parameter-efficient approach for continual image segmentation. By reducing cross-step interference via a task-specific subspace and a prototype-based loss, our method substantially mitigates catastrophic forgetting. Background classifier calibration further rescales the background weights and alleviates background drift. Under a full step parameter-efficient training regime, our approach achieves competitive or even superior semantic segmentation performance relative to knowledge distillation-based methods. Ablation studies confirm the effectiveness of each component. We additionally investigate LoRA fine-tuning for panoptic segmentation.

AI use statement

In this work, we used generative AI tools to assist with language translation and to implement method-related code. We did not use generative AI tools to propose hypotheses, design the research methodology, or interpret experimental results. Developing theoretical frameworks and formulating mathematical proofs are not applicable to this work.

We also used generative AI tools to suggest experimental hyperparameters and to aid revision.

We have reviewed all AI-assisted content. In particular, LLM-assisted translations were checked against the original meaning, generated code and visualization scripts were debugged and verified, and suggested hyperparameters were validated through controlled experiments. We take full responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

REFERENCES

Ivana Balazevic, David Steiner, Nikhil Parthasarathy, Relja Arandjelović, and Olivier Henaff. Towards in-context scene understanding. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 63758–63778. Curran Associates, Inc., 2023. doi: 10.52202/075280-2785. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/c94a632545000531f0b47000e9caa5b6-Paper-Conference.pdf.

Nicolas Carion, Laura Gustafson, Yuan-Ting Hu, Shoubhik Debnath, Ronghang Hu, Didac Suris Coll-Vinent, Chaitanya Ryali, Kalyan Vasudev Alwala, Haitham Khedr, Andrew Huang, Jie Lei, Tengyu Ma, Baishan Guo, Arpit Kalla, Markus Marks, Joseph Greer, Meng Wang, Peize Sun, Roman Rädle, Triantafyllos Afouras, Effrosyni Mavroudi, Katherine Xu, Tsung-Han Wu, Yu Zhou, Liliane Momeni, RISHI HAZRA, Shuangrui Ding, Sagar Vaze, Francois Porcher, Feng Li, Siyuan Li, Aishwarya Kamath, Ho Kei Cheng, Piotr Dollar, Nikhila Ravi, Kate Saenko, Pengchuan Zhang, and Christoph Feichtenhofer. Sam 3: Segment anything with concepts. In C. Vondrick, B. Hariharan, C. Raffel, L. Pinto, D. Yang, and A. Faust (eds.), *International Conference on Learning Representations*, volume 2026, pp. 138846–138923, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/e0982cbc81401df3430ee1ff780dc7a2-Paper-Conference.pdf.

Fabio Cermelli, Massimiliano Mancini, Samuel Rota Bulò, Elisa Ricci, and Barbara Caputo. Modeling the background for incremental learning in semantic segmentation, 2020. URL https://arxiv.org/abs/2002.00718.

Fabio Cermelli, Matthieu Cord, and Arthur Douillard. Comformer: Continual learning in semantic and panoptic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3010–3020, June 2023.

Jinpeng Chen, Runmin Cong, Yuxuan Luo, Horace Ho Shing Ip, and Sam Kwong. Strike a balance in continual panoptic segmentation, 2024a. URL https://arxiv.org/abs/2407.16354.

Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L. Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs, 2017. URL https://arxiv.org/abs/1606.00915.

Zhe Chen, Yuchen Duan, Wenhai Wang, Junjun He, Tong Lu, Jifeng Dai, and Yu Qiao. Vision transformer adapter for dense predictions, 2023. URL https://arxiv.org/abs/2205.08534.

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, Bin Li, Ping Luo, Tong Lu, Yu Qiao, and Jifeng Dai. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 24185–24198, June 2024b.

Bowen Cheng, Ishan Misra, Alexander G. Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1290–1299, June 2022.

Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

Arthur Douillard, Yifu Chen, Arnaud Dapogny, and Matthieu Cord. Plop: Learning without forgetting for continual semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4040–4050, June 2021.

Mark Everingham, Luc Van Gool, Christopher K. I. Williams, John Winn, and Andrew Zisserman. The PASCAL Visual Object Classes Challenge 2012 (VOC2012) Results. http://www.pascal-network.org/challenges/VOC/voc2012/workshop/index.html, 2012.

Kai Fang, Anqi Zhang, Guangyu Gao, Jianbo Jiao, Chi Harold Liu, and Yunchao Wei. Combo: Conflict mitigation via branched optimization for class incremental segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 25667–25676, June 2025.

Yizheng Gong, Siyue Yu, Xiaoyang Wang, and Jimin Xiao. Continual segmentation with disentangled objectness learning and class recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3848–3857, June 2024.

Tao GUO, Junwei Chen, and Laizhong Cui. Fed-duet: Dual expert-orchestrated framework for continual federated vision-language learning. In C. Vondrick, B. Hariharan, C. Raffel, L. Pinto, D. Yang, and A. Faust (eds.), *International Conference on Learning Representations*, volume 2026, pp. 14388–14412, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/17e342b35a001993b539a678020e2a3f-Paper-Conference.pdf.

Jiangpeng He, Zhihao Duan, and Fengqing Zhu. Cl-lora: Continual low-rank adaptation for rehearsal-free class-incremental learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 30534–30544, June 2025.

Kaiming He, Georgia Gkioxari, Piotr Dollar, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models, 2021. URL https://arxiv.org/abs/2106.09685.

Menglin Jia, Luming Tang, Bor-Chun Chen, Claire Cardie, Serge Belongie, Bharath Hariharan, and Ser-Nam Lim. Visual prompt tuning, 2022. URL https://arxiv.org/abs/2203.12119.

Beomyoung Kim, Joonsang Yu, and Sung Ju Hwang. Eclipse: Efficient continual learning in panoptic segmentation with visual prompt tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3346–3356, June 2024.

Alexander Kirillov, Kaiming He, Ross Girshick, Carsten Rother, and Piotr Dollar. Panoptic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollar, and Ross Girshick. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4015–4026, October 2023.

Zhizhong Li and Derek Hoiem. Learning without forgetting, 2017. URL https://arxiv.org/abs/1606.09282.

Yan-Shuo Liang and Wu-Jun Li. Adaptive plasticity improvement for continual learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7816–7825, June 2023.

Yan-Shuo Liang and Wu-Jun Li. Inflora: Interference-free low-rank adaptation for continual learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 23638–23647, June 2024.

Zichen Liang, Yusong Hu, Fei Yang, and Xialei Liu. Enhancing continual semantic segmentation via uncertainty and class balance re-weighting. *IEEE Transactions on Image Processing*, 34: 3689–3702, 2025. doi: 10.1109/TIP.2025.3576477.

Weixiang Liu, Jianbin Huang, Deyu Zeng, Jiawei Pan, Zongze Wu, Xiaopin Zhong, and Yuanlong Deng. Dcgseg: Distance correlation graph for nonlinear inter-class relation distillation in continual semantic segmentation. *Neurocomputing*, 680:133334, 2026. ISSN 0925-2312. doi: https://doi.org/10.1016/j.neucom.2026.133334. URL https://www.sciencedirect.com/science/article/pii/S0925231226007319.

Xuan Liu and Xiaobin Chang. Lora subtraction for drift-resistant space in exemplar-free continual learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15308–15318, June 2025.

Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015.

Mao-Lin Luo, Zi-Hao Zhou, Yi-Lin Zhang, Yuanyu Wan, Min-Ling Zhang, and Tong Wei. Keeplora: Continual learning with residual gradient adaptation. In C. Vondrick, B. Hariharan, C. Raffel, L. Pinto, D. Yang, and A. Faust (eds.), *International Conference on Learning Representations*, volume 2026, pp. 136017–136040, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/dc3ca8bcd613e43ce540352b58d55d6d-Paper-Conference.pdf.

Andrea Maracani, Umberto Michieli, Marco Toldo, and Pietro Zanuttigh. Recall: Replay-based continual learning in semantic segmentation. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 7006–7015. IEEE, 2021.

Gilhan Park, WonJun Moon, SuBeen Lee, Tae-Young Kim, and Jae-Pil Heo. Mitigating background shift in class-incremental semantic segmentation. *arXiv preprint arXiv:2407.11859*, 2024.

Fuli Qiao and Mehrdad Mahdavi. Merge before forget: A single lora continual learning via continual merging. In C. Vondrick, B. Hariharan, C. Raffel, L. Pinto, D. Yang, and A. Faust (eds.), *International Conference on Learning Representations*, volume 2026, pp. 8133–8159, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/0dedca10992d34035c9a9776254cfa8c-Paper-Conference.pdf.

Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, Eric Mintun, Junting Pan, Kalyan Vasudev Alwala, Nicolas Carion, Chao-Yuan Wu, Ross Girshick, Piotr Dollár, and Christoph Feichtenhofer. Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*, 2024. URL https://arxiv.org/abs/2408.00714.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In Nassir Navab, Joachim Hornegger, William M. Wells, and Alejandro F. Frangi (eds.), *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, pp. 234–241, Cham, 2015. Springer International Publishing.

Oriane Siméoni, Huy V. Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, Francisco Massa, Daniel Haziza, Luca Wehrstedt, Jianyuan Wang, Timothée Darcet, Théo Moutakanni, Leonel Sentana, Claire Roberts, Andrea Vedaldi, Jamie Tolan, John Brandt, Camille Couprie, Julien Mairal, Hervé Jégou, Patrick Labatut, and Piotr Bojanowski. Dinov3, 2025. URL https://arxiv.org/abs/2508.10104.

Peng Wang, Shijie Wang, Junyang Lin, Shuai Bai, Xiaohuan Zhou, Jingren Zhou, Xinggang Wang, and Chang Zhou. One-peace: Exploring one general representation model toward unlimited modalities, 2023a. URL https://arxiv.org/abs/2305.11172.

Wenhui Wang, Hangbo Bao, Li Dong, Johan Bjorck, Zhiliang Peng, Qiang Liu, Kriti Aggarwal, Owais Khan Mohammed, Saksham Singhal, Subhojit Som, and Furu Wei. Image as a foreign language: Beit pretraining for vision and vision-language tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19175–19186, June 2023b.

Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M. Alvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 12077–12090. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/64f1f27bf1b4ec22924fd0acb550c235-Paper.pdf.

Fengqiang Wan Yipeng Lin and Yang Yang. Gr-lora: Gradient-recycling low-rank adaptation for class-incremental learning. In *Forty-third International Conference on Machine Learning*, 2026.

Juzheng Zhang, Jiacheng You, Ashwinee Panda, and Tom Goldstein. Lori: Reducing cross-task interference in multi-task low-rank adaptation, 2025a. URL https://arxiv.org/abs/2504.07448.

Xin Zhang, Liang Bai, Xian Yang, and Jiye Liang. C-lora: Continual low-rank adaptation for pre-trained models, 2025b. URL https://arxiv.org/abs/2502.17920.

Zekang Zhang, Guangyu Gao, Jianbo Jiao, Chi Harold Liu, and Yunchao Wei. Coinseg: Contrast inter- and intra- class representations for incremental segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 843–853, October 2023.

Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

Tianfei Zhou, Wenguan Wang, Ender Konukoglu, and Luc Van Gool. Rethinking semantic segmentation: A prototype view. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2582–2593, June 2022.

Yuchen Zhu, Cheng Shi, Dingyou Wang, Jiajin Tang, Zhengxuan Wei, Yu Wu, Guanbin Li, and Sibei Yang. Rethinking query-based transformer for continual image segmentation, 2025. URL https://arxiv.org/abs/2507.07831.

Heming Zou, Yunliang Zang, Wutong Xu, Yao Zhu, and Xiangyang Ji. Flylora: Boosting task decoupling and parameter efficiency via implicit rank-wise mixture-of-experts. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pp. 10386–10419. Curran Associates, Inc., 2025. doi: 10.52202/085713-0349. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/0f234479d4be0102b6a9ec5ebf75e881-Paper-Conference.pdf.

A Details of Class-Balanced Pixel Sampling

Let $P = \{(f_i, y_i)\}_{i=1}^{N}$ denote the pixel set from the highest-resolution feature map produced by the pixel decoder, where $f_i \in \mathbb{R}^d$ is the pixel feature, $y_i \in C_t$ is the pixel label at the current step t, and N is the number of selected pixels. For a single class c, $P_c = \{(f_i, y_i) \in P \mid y_i = c\}$, and the number of pixels is N_c. To alleviate severe class imbalance in semantic segmentation, we adopt a class-balanced pixel sampling strategy before computing the prototype-based loss. Given a sampling budget E and K classes present in the current mini-batch, we allocate an equal quota $e = E/K$ to each class. For class c, we randomly sample at most e pixels:

$$\widetilde{\mathcal{P}}_c = \begin{cases} \text{RandomSample}(\mathcal{P}_c, e), & N_c > e \\ \mathcal{P}_c, & N_c \le e \end{cases} \tag{16}$$

The final sampled pixel set is $\widetilde{\mathcal{P}} = \bigcup_{c \in \mathcal{C}_b} \widetilde{\mathcal{P}}_c$, where $\mathcal{C}_b$ denotes the classes present in mini-batch b. The number of sampled pixels is

$$\widetilde{N} = \sum_{c \in \mathcal{C}_b} \min(N_c, e) \tag{17}$$

This strategy prevents dominant categories from overwhelming the contrastive objective and ensures that rare classes contribute equally to prototype learning.

Prototypes are computed by classical masked average pooling (MAP) and updated with momentum.

MAP is defined as:

$$P_c = \frac{\sum_{i=1}^{H} \sum_{j=1}^{W} F\,(i,j) \odot M_c(i,j)}{\sum_{i=1}^{H} \sum_{j=1}^{W} M_c(i,j)} \tag{18}$$

where $P_c \in \mathbb{R}^d$ is the prototype of a single class $c \in C_t$, and $M_c \in \mathbb{R}^{H \times W}$ is the binarized mask derived from the ground-truth label. Analogous to Eq. equation 6, M_c masks out pixels that do not belong to class c.

Table 7: Effect of modifying attention on ADE20K.

Method	**100-50** (Incremental)			**Panoptic Segmentation** (Full)		
	1-100	101-150	all	RQ	SQ	PQ
Original Attention	38.2	48.5	41.6	43.726	72.595	36.487
Modified Attention	34.8	45.3	38.3	45.457	74.656	37.802

Table 8: Effect of different pre-trained weights on Pascal VOC 2012.

Base Weights	**15-5** (2 steps)			**15-1** (6 steps)		
	1-15	16-20	all	1-15	16-20	all
Pre-trained on Cityscapes	83.8	57.0	77.1	80.9	6.6	62.4
Pre-trained on ADE20K	91.2	83.1	89.2	81.6	23.9	67.2

B Other Implementation Details

For the ADE20K dataset, the learning rate is set to 1×10^{-4} for steps comprising 50 or 100 classes, with the number of training iterations set to $8\times$ the number of images in that step. For steps with 10 classes, we set the learning rate to 5×10^{-5} and the iteration count to $9\times$ the step image count. In steps with 5 classes, the learning rate is adjusted to 2×10^{-5}, and the iterations are set to $10\times$ the corresponding image count. For the Pascal VOC 2012 dataset, the learning rate at the base step is configured to 1×10^{-4} with an iteration budget of $36\times$ the base-step image count. During the incremental steps, the learning rate is set to 2×10^{-4} with $20\times$ the step image count for the 19-1 and 15-5 protocols, whereas for the 15-1 protocol, it is set to 3×10^{-5} with $40\times$ the step image count.

The cumulative singular-value threshold ρ is set to 0.99. For the pixel-prototype loss $\mathcal{L}_{\text{PPL}}$, both loss weights λ_m and λ_p are set to 0.5. Sampling budget E is set to 2048, temperature τ is set to 0.1, and margin m is set to 0.2. Regarding the foundation checkpoint pre-trained on Cityscapes, we resize the output dimension of the classification head (*class_embed*) and reinitialize it using Kaiming uniform initialization. To facilitate continual learning, we zero out the gradients of parameters in *class_embed* that correspond to inactive classes at each step, and set the weight decay of the AdamW optimizer to 0. Furthermore, we do not apply LoRA adaptation to the classifier head; instead, parameters corresponding to active categories in *class_embed* are directly updated via full-parameter training.

C Extended Ablation Study

Modifying attention for LoRA fine-tuning. In the original Mask2Former implementation, both *self-attention* and *cross-attention* blocks are constructed using PyTorch's native *nn.MultiheadAttention*. By default, the PEFT library targets standalone *nn.Linear* layers for LoRA injection; within *nn.MultiheadAttention*, however, only *output_proj* is instantiated as an *nn.Linear* module. Consequently, LoRA cannot access or adapt the underlying projections such as *query_proj*, *key_proj*, and *value_proj*. To explore more comprehensive adaptation, we refactored *nn.MultiheadAttention* into explicit linear layers so that all projection matrices are accessible to PEFT. As reported in Table 7, adapting these additional projections for continual semantic segmentation unexpectedly exacerbates inter-step interference; we therefore retain the default attention configuration in our primary experiments. In contrast, Table 7 also reports the effect of fine-tuning the refactored attention on panoptic segmentation. The observed marginal gain in Panoptic Quality (PQ) verifies the functional correctness of our refactored attention.

Different pre-trained weights. Table 8 compares different pre-trained initializations on Pascal VOC 2012. Initializing from ADE20K consistently outperforms initializing from Cityscapes under both protocols. The improvement is particularly pronounced for the incremental classes: their mIoU increases from 57.0 to 83.1 in the 15-5 setting and from 6.6 to 23.9 in the more challenging 15-1 setting. The corresponding overall gains are 12.1 and 4.8 mIoU, respectively. These results indicate that stronger and more transferable segmentation representations provide a substantially better starting point for parameter-efficient continual adaptation, especially when learning novel classes.

Table 9: Effect of cross-step pixel-prototype loss.

Method	**100-50** (ADE20K)			**15-5** (VOC2012)		
	1-100	101-150	all	1-15	16-20	all
Cross-step PPL	36.0	46.5	39.5	91.2	83.1	89.2
Intra-step PPL	38.2	48.5	41.6	89.1	71.1	84.6

Table 10: Effect of full-parameter training at the base step.

Method	**100-50** (ADE20K)		
	1-100	101-150	all
Full Train	40.1	46.0	42.0
LoRA fine-tuning	38.2	48.5	41.6

Cross-step pixel-prototype loss. Table 9 reports the impact of retaining prototypes from previous steps when computing the pixel-prototype loss at the current step. In our experiments, preserving previous-step prototypes degrades performance on ADE20K, but consistently improves results on Pascal VOC 2012. This disparity can be attributed to differences in scene complexity: an ADE20K image contains 10.5 categories on average, compared with only 1.4 on VOC 2012. Consequently, old categories co-occur frequently with novel classes on ADE20K, but much less often on VOC 2012. These observations suggest that whether cross-step prototype utilization conflicts with task-specific gradient subspace projection depends on how frequently old classes co-occur in images that contain incremental classes.

Full-parameter training at the base step. Table 10 reports the impact of training the base step without LoRA fine-tuning. As expected, this configuration achieves a higher overall mIoU, driven mainly by gains on base classes, at the cost of degradation on subsequent incremental steps. Such a strategy, however, compromises the end-to-end parameter efficiency that is a core tenet of our framework. We therefore retain parameter-efficient tuning across all sessions and do not adopt full-parameter training at the base step.

Behavior and regularity of scaling factor γ. Table 11 reports the impact of varying γ under different initialization and training regimes. Intriguingly, the deviation of the optimal factor from unity, $|\gamma^* - 1.00|$, serves as an empirical indicator of representation fidelity. Under strong pre-trained representations (e.g., ADE20K-initialized weights), the classifier weights across steps naturally align in scale, yielding optimal γ^* values that strictly remain within $|\gamma^* - 1.00| \leq 0.05$ across both 15-5 and 15-1 protocols. Conversely, suboptimal initializations (e.g., Cityscapes) or under-converged schedules induce significant logit discrepancies, requiring γ^* to deviate noticeably from 1.00 (down to 0.63) to compensate for the background bias. This finding confirms that while γ provides a useful compensatory tool for weaker models, our framework inherently converges toward a tuning-free regime ($\gamma \approx 1.00$) when built atop high-quality vision backbones.

Computational Efficiency Analysis. To quantitatively assess the parameter and computational efficiency of our framework, we report the proportion of trainable parameters and the average training FLOPs per image on Pascal VOC 2012 in Table 12. Full-parameter training updates 100 % of the

Table 11: Effect of the scaling factor γ on Pascal VOC 2012.

γ	**Cityscapes-pt**(19-1)			γ	**ADE20K-pt**(19-1)			γ	**ADE20K-pt**(15-5)		
	1-19	20	all		1-19	20	all		1-15	16-20	all
1.00	86.779	3.851	82.632	1.00	93.861	60.891	92.212	1.00	91.247	82.914	89.163
0.80	87.122	11.111	83.321	1.10	93.934	61.766	92.326	1.03	91.259	82.981	89.189
0.70	87.629	25.723	84.533	1.15	93.969	62.251	92.382	1.05	91.237	83.109	**89.205**
0.65	87.774	37.075	85.238	1.18	93.982	62.501	92.407	1.08	91.199	82.993	89.147
0.63	87.874	41.481	**85.554**	1.20	93.988	62.655	**92.420**	-	-	-	-
0.60	87.454	39.381	85.049	1.23	93.911	62.882	92.359	-	-	-	-

Table 12: Quantitative analysis of parameter efficiency on Pascal VOC 2012. Training FLOPs per image are profiled via fvcore over 20 iterations under standard batch configurations, accounting for both forward and backward computational graphs.

Method	**Trainable Params %**	**Training FLOPs / img** (G)
Full Parameter Training	100.0	823.8
Ours (base step)	13.9	734.4
Ours (incremental step)	13.9	742.2

model parameters and incurs 823.8 G FLOPs per image. In contrast, our method trains only 13.9 % of the parameters at both the base step and every incremental step. Correspondingly, the training cost is reduced to 734.4 G FLOPs per image at the base step and 742.2 G FLOPs per image at incremental steps—approximately 10–11 % lower than full-parameter training—while remaining fully parameter-efficient throughout the continual learning process.