ECLIPSE: Efficient Continual Learning in Panoptic Segmentation with Visual Prompt Tuning

Beomyoung Kim[1,2] Joonsang Yu[1] Sung Ju Hwang[2]

NAVER Cloud, ImageVision[1] KAIST[2]

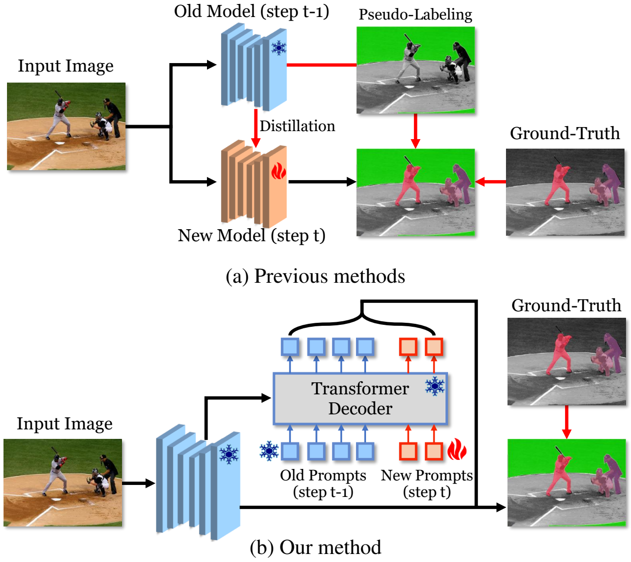
{beomyoung.kim, joonsang.yu}@navercorp.com, sjhwang82@kaist.ac.kr

Abstract

Panoptic segmentation, combining semantic and instance segmentation, stands as a cutting-edge computer vision task. Despite recent progress with deep learning models, the dynamic nature of real-world applications necessitates continual learning, where models adapt to new classes (plasticity) over time without forgetting old ones (catastrophic forgetting). Current continual segmentation methods often rely on distillation strategies like knowledge distillation and pseudo-labeling, which are effective but result in increased training complexity and computational overhead. In this paper, we introduce a novel and efficient method for continual panoptic segmentation based on Visual Prompt Tuning, dubbed ECLIPSE. Our approach involves freezing the base model parameters and fine-tuning only a small set of prompt embeddings, addressing both catastrophic forgetting and plasticity and significantly reducing the trainable parameters. To mitigate inherent challenges such as error propagation and semantic drift in continual segmentation, we propose logit manipulation to effectively leverage common knowledge across the classes. Experiments on ADE20K continual panoptic segmentation benchmark demonstrate the superiority of ECLIPSE, notably its robustness against catastrophic forgetting and its reasonable plasticity, achieving a new state-of-the-art. The code is available at https://github.com/clovaai/ECLIPSE.

1. Introduction

Image segmentation is a fundamental computer vision task, and it involves dividing an image into meaningful segments to facilitate easier analysis. One of the most advanced forms of image segmentation is *panoptic segmentation*, which merges semantic segmentation (categorizing pixels into set categories) with instance segmentation (identifying individual objects within categories). Recent panoptic segmentation studies have made significant progress by leveraging convolutional neural networks [7, 22, 41, 46] and transformer-based architectures [1, 8, 9, 17, 25].



(a) Previous methods

(b) Our method

Figure 1. **Comparison of the overview of (a) previous methods and (b) our method.** Previous methods rely on distillation strategies such as knowledge distillation and pseudo-labeling, demanding more training complexity and computational overhead. In contrast, our method freezes all trained parameters and fine-tunes only a small set of prompt embeddings, robustly keeping the previous knowledge and extending the scalability of the model.

Despite these advances, the dynamic nature of the real world demands that models not only understand the present but also evolve over time. Continual image segmentation addresses this need by enabling models to learn new classes incrementally over time without forgetting the old classes. It is critical in real-world applications where new classes emerge unpredictably, such as in robotics [37] and surveillance [36]. However, it is greatly challenging to preserve the previous class knowledge (avoiding **catastrophic forgetting** [13]) and integrate new class information efficiently (**plasticity**) simultaneously.

Recently, various continual segmentation methods [2, 5, 10, 34, 35, 38, 44, 45, 47] have emerged, addressing the key challenges and showing notable improvements. Most continual segmentation approaches often employ distillation strategies like knowledge distillation [2, 10, 45] and pseudo-labeling [5, 10, 35], as shown in Figure 1a. Knowledge distillation can alleviate catastrophic forgetting by transferring knowledge from an old model to a new model, and pseudo-labeling allows the new model to train with labels of the previously learned classes. Though groundbreaking, these approaches involve trade-offs such as the need for doubled network forwarding and careful tuning of the hyperparameters (*e.g.*, distillation loss weights and threshold for pseudo-labels), which increases the training complexity and computational overhead. As the number of classes increases incrementally, maintaining a scalable and efficient distillation process can become challenging. Moreover, while much of the research has focused on continual semantic segmentation, continual panoptic segmentation, which is more challenging in incorporating both semantic- and instance-level segmentation tasks, has been relatively underexplored.

In this paper, we propose a novel method, dubbed **ECLIPSE**, for **E**fficient **C**ontinual **L**earning **I**n **P**anoptic **SE**gmentation that leverages the potential of Visual Prompt Tuning (VPT) [18] and obviates the need for conventional distillation strategies. Our approach begins with freezing all parameters of the base model and repeatedly fine-tunes a set of new prompt embeddings as new classes emerge, as shown in Figure 1b. Our method inherently addresses catastrophic forgetting through model freezing and enhances plasticity via prompt tuning. To the best of our knowledge, our approach is the first distillation-free continual panoptic segmentation, significantly reducing the trainable parameters and simplifying the continual segmentation process.

Despite these strengths, we confront inherent challenges in continual panoptic segmentation that necessitate further improvement. Although model freezing preserves the prior knowledge, it can simultaneously propagate prior errors forward. Moreover, the definition of `no-obj` class, which is required in inference to distinguish whether an output mask is no object or not, changes at each continual learning step, which is known as the semantic drift problem. To circumvent the challenges, we propose a simple and effective strategy, called logit manipulation. It allows the model to leverage the inter-class knowledge of all learned classes to more meaningfully manipulate the `no-obj` logit. The dynamically updated `no-obj` logit helps suppress prior error predictions and mitigate the semantic drift issue at once.

Our comprehensive experiments on ADE20K [48] demonstrate that ECLIPSE achieves a new state-of-the-art in continual panoptic segmentation with a mere 1.3% of the trainable parameters. Notably, ECLIPSE shows outstanding robustness against catastrophic forgetting, especially as the number of continual steps increases, making a substantial improvement over existing methods.

In summary, the contributions of our paper are:

- We successfully integrate VPT into continual panoptic segmentation, effectively mitigating catastrophic forgetting and efficiently extending the scalability of the model.
- We propose an effective logit manipulation strategy that circumvents the inherent challenges in continual panoptic segmentation: error propagation and semantic drift.
- We achieve state-of-the-art results on ADE20K with a significantly few number of trainable parameters.

2. Related Work

Panoptic Segmentation is a cutting-edge task in computer vision, blending the concepts of semantic and instance segmentation to provide a comprehensive understanding of both 'stuff' (amorphous regions like sky or grass) and 'things' (countable objects like cars or people). Pioneering works [7, 22, 23] integrated semantic and instance segmentation tasks into a unified framework. Following this, some methods [26, 41, 46] introduced significant improvements by employing dynamic convolutions in a fully convolutional paradigm. More recently, transformer-based architectures [8, 9, 25] have further advanced the field by leveraging the power of attention mechanisms. However, the challenge of continual panoptic segmentation, particularly in dynamically adapting to new classes without forgetting the old ones, remains a frontier for ongoing research.

Continual Segmentation. To address the dynamic nature of real-world applications, continual segmentation emerges as an advanced task. Pioneering work [2] has discovered the distinct challenge in continual segmentation tasks, semantic drift, caused by *background* class. Most approaches [2, 5, 10, 34, 35, 45, 47] utilize distillation strategies such as knowledge distillation and pseudo-labeling to mitigate the semantic drift issue. More recently, Incrementer [38] leverages the architectural advantages of a transformer-based model using incremented class embeddings and multiple distillation strategies. However, most of the research has focused on continual semantic segmentation, and continual panoptic segmentation, which is a more challenging task, is less explored. CoMFormer [4] is a pioneer work in continual panoptic segmentation using the universal segmentation model (*i.e.*, Mask2Former [9]) to perform both panoptic and semantic segmentation tasks with query-based distillation strategy. However, such distillation-based approaches increase the training complexity and computational overhead and demand careful tuning of hyperparameters such as loss weights and temperatures of distillation and threshold for pseudo-labeling. Unlike them, our method is the first distillation-free approach for both panoptic and semantic segmentation tasks, simplifying the continual learning process and reducing training computations.

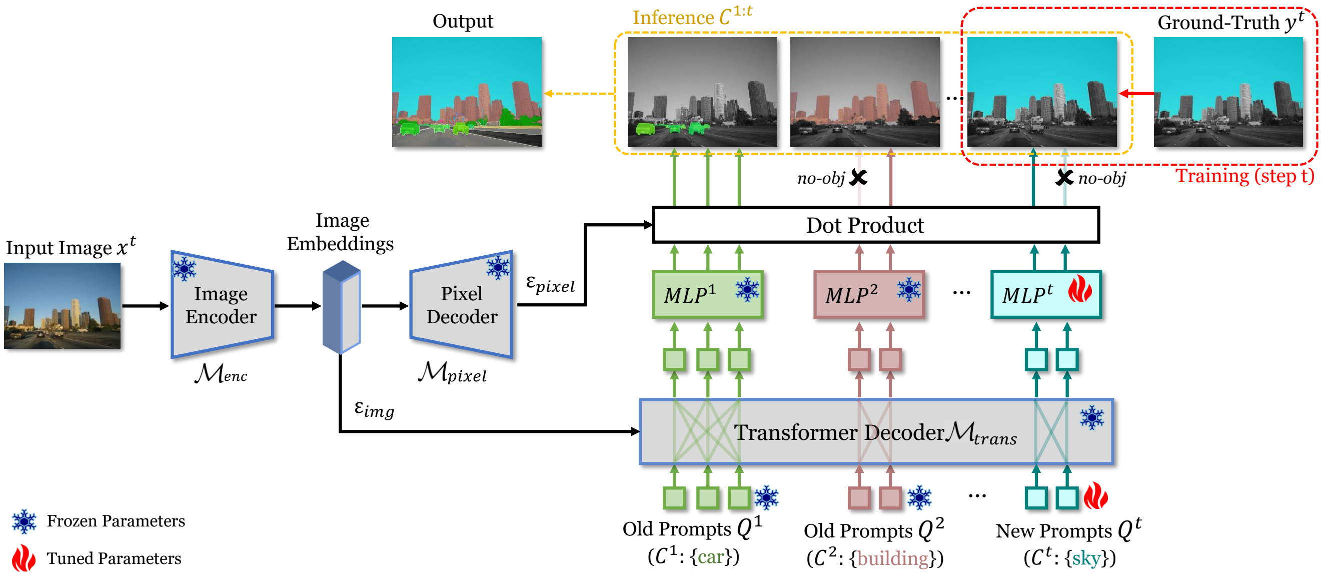


Figure 2. **Overview of ECLIPSE**. We freeze all trained parameters and fine-tune only a set of prompt embeddings $\mathbf{Q}^t$ alongside MLP layers to recognize a set of classes $\mathcal{C}^t$. In inference, we aggregate outputs from all prompt sets $\mathbf{Q}^{1:t}$ to segment all learned classes $\mathcal{C}^{1:t}$.

Visual Prompt Tuning (VPT) in Continual Learning. VPT [18] introduced an efficient and effective fine-tuning method for vision transformer models. By freezing the pre-trained parameters, they fine-tuned only a set of learnable prompts and achieved remarkable performance. In the continual image classification field, there are several attempts to utilize the VPT. Namely, L2P [43] and DualPrompt [42] freeze the pre-trained model and select the most relevant prompts from a prompts pool in a key-value mechanism. They achieved noticeable performance and demonstrated the potential of leveraging VPT in continual image classification. We are the first VPT-based continual segmentation method tailored to address several distinct challenges in continual panoptic segmentation.

3. Preliminary

3.1. Problem Setting

Panoptic segmentation is an advanced task that unifies semantic- and instance-level segmentation tasks. Compared to semantic segmentation without discrimination of individual instances, panoptic segmentation necessitates a more comprehensive unified framework to identify both 'things' (*e.g.*, individual car) and 'stuff' (*e.g.*, sky or road).

This paper mainly focuses on continual learning in panoptic segmentation, following the same setting in [4]. Over continually arriving tasks at timesteps $t = 1, \dots, T$, at each step t, a training dataset $\mathcal{D}^t$ is introduced which contains image-label pairs $(\boldsymbol{x}^t, \boldsymbol{y}^t)$, where $\boldsymbol{x}^t$ represents an image and $\boldsymbol{y}^t$ its corresponding segmentation label. Here, $\boldsymbol{y}^t$ is labeled only for the set of current classes $\mathcal{C}^t$, and other classes (previous $\mathcal{C}^{1:t-1}$ and future $\mathcal{C}^{t+1:T}$ classes) are not accessible. Once a task is completed, the model is expected to segment for all classes in $\mathcal{C}^{1:t}$ while preventing catastrophic forgetting for $\mathcal{C}^{1:t-1}$ and incrementally learning for $\mathcal{C}^t$ (plasticity).

3.2. Network Architecture: Mask2Former

We adopt Mask2Former [9] as our baseline architecture, which is a transformer-based model for universal segmentation tasks including panoptic, instance, and semantic segmentation. Unlike the previous paradigm of per-pixel classification, Mask2Former directly predicts a set of masks including their classes, termed mask classification. Mask2Former consists of three kinds of modules: image encoder $\mathcal{M}_{enc}$, pixel decoder $\mathcal{M}_{pixel}$ and transformer decoder $\mathcal{M}_{trans}$. The image encoder extracts image embedding $\mathcal{E}_{img}$ from the input image, and the pixel decoder converts image embedding into per-pixel embedding $\mathcal{E}_{pixel}$. The transformer decoder takes N learnable queries and generates N mask embeddings by self-attention and cross-attention with image embedding. Here, each query takes charge of representing an object (or no object). Finally, N mask proposals are produced via a dot product between N mask embedding and per-pixel embedding. Moreover, the category of each mask is assigned using the MLP-based classifier. When a mask proposal is classified as *no-obj* class, it is dropped during inference.

For simplicity, we express the model $\mathcal{M}$ as a function which takes an image $\mathbf{x} \in \mathbb{R}^{3\times H\times W}$ and queries $\mathbf{Q} \in \mathbb{R}^{N\times D}$ as inputs and output masks $\mathbf{m} \in \mathbb{R}^{N\times H\times W}$ and class logits $\mathbf{s} \in \mathbb{R}^{N \times C}$:

$$(\mathbf{s}, \mathbf{m}) = \mathcal{M}(\mathbf{x}, \mathbf{Q}), \tag{1}$$

$$\mathcal{M}(\mathbf{x}, \mathbf{Q}) = \text{MLP}(\mathcal{M}_{trans}(\mathcal{E}_{img}, \mathbf{Q})) \otimes \mathcal{E}_{pixel}, \tag{2}$$

$$\mathcal{E}_{img} = \mathcal{M}_{enc}(\mathbf{x}),\ \mathcal{E}_{pixel} = \mathcal{M}_{pixel}(\mathcal{E}_{img}), \tag{3}$$

where N is the number of queries, D is the dimension for the query embedding, and C is the number of classes.

4. Method

4.1. Prompt Tuning for Continual Segmentation

We present a novel approach, named ECLIPSE, for efficient continual panoptic segmentation which leverages Visual Prompt Tuning (VPT) [18] for the integration of new classes without the conventional distillation strategies. Our approach begins with the initial training (t=1) of all parameters of the model $\mathcal{M}$ on the set of base classes $\mathcal{C}^1$:

$$(\mathbf{s}^1, \mathbf{m}^1) = \mathcal{M}(\mathbf{x}, \mathbf{Q}^1). \tag{4}$$

After this initial training, we apply a *freeze-and-tune* strategy repeatedly. Namely, when new classes are introduced (t>1), we freeze all trained parameters to preserve the acquired knowledge and fine-tune a set of learnable prompt embeddings $\mathbf{Q}^t \in \mathbb{R}^{N^t \times D}$ alongside unshared MLP layers to recognize new classes $\mathcal{C}^t$:

$$(\mathbf{s}^t, \mathbf{m}^t) = \mathcal{M}(\mathbf{x}, \mathbf{Q}^t)\ (t > 1), \tag{5}$$

$$\mathcal{M}(\mathbf{x}, \mathbf{Q}^t) = \text{MLP}^t(\mathcal{M}_{trans}(\mathcal{E}_{img}, \mathbf{Q}^t)) \otimes \mathcal{E}_{pixel}, \tag{6}$$

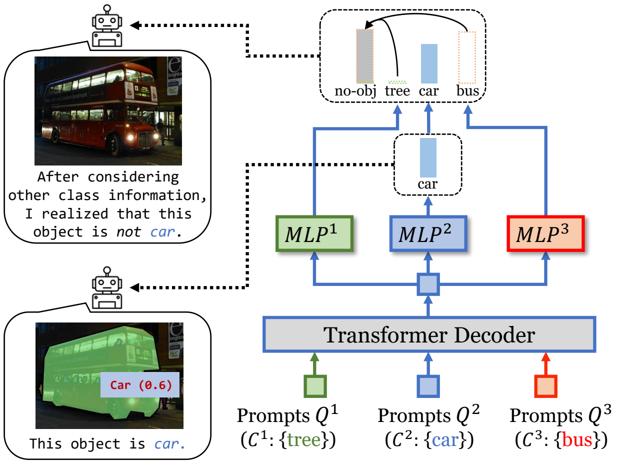
$$\mathcal{E}_{img} = \mathcal{M}_{enc}(\mathbf{x}),\ \mathcal{E}_{pixel} = \mathcal{M}_{pixel}(\mathcal{E}_{img}), \tag{7}$$

where ● and ● denote trainable and frozen parameters, respectively. In short, we treat each set of prompts $\mathbf{Q}^t$ as a discrete task-specific module, solely devoted to the recognition of $\mathcal{C}^t$ classes. As the continual steps progress, we stably extend the scalability of the model through the lightweight task-specific prompt sets. During the inference phase, we aggregate outputs from all the prompt sets across steps. $\mathbf{Q}^{1:t}$, allowing the model to segment all learned classes $\mathcal{C}^{1:t}$:

$$(\mathbf{s}^{1:t}, \mathbf{m}^{1:t}) = \mathcal{M}(\mathbf{x}, \mathbf{Q}^{1:t}). \tag{8}$$

Our approach ensures that previous knowledge is preserved with model freezing, which prevents catastrophic forgetting, while also enabling the efficient integration of new knowledge through prompt tuning, enhancing the model's plasticity.

We introduce two prompt tuning strategies for continual panoptic segmentation, termed *shallow* and *deep*, inspired by [18]. The *shallow* means tuning the prompt embeddings at the first transformer layer only, $\mathbf{Q}^t_{shallow} \in \mathbb{R}^{N^t \times D}$. Whereas, the *deep* denotes tuning the embeddings across all transformer layers, $\mathbf{Q}^t_{deep}$=$\{\mathbf{Q}^t_1, \mathbf{Q}^t_2, \cdots, \mathbf{Q}^t_L\}$ where L is the number of transformer layers. By default, we adopt the *deep* prompt tuning and will present a detailed analysis in the experimental section.



After considering other class information, I realized that this object is *not* car.
no-obj
tree
car
bus
car
MLP^1
MLP^2
MLP^3
Transformer Decoder
Car (0.6)
This object is car.
Prompts Q^1 (C^1: {tree})
Prompts Q^2 (C^2: {car})
Prompts Q^3 (C^3: {bus})

Figure 3. **Illustration of logit manipulation.** To alleviate semantic drift of `no-obj` class, we make a new `no-obj` logit leveraging the inter-class knowledge of all learned classes. Moreover, an erroneous prediction caused by semantic confusion of prior frozen parameters can be fixed through logit manipulation.

4.2. Resolving Semantic Confusion and Drift

Our method effectively addresses both catastrophic forgetting and plasticity in continual panoptic segmentation. However, we confront an inherent issue, called error propagation. This issue arises because freezing the model, while helpful in preventing catastrophic forgetting, also means carrying prior errors forward. These errors often originate from **semantic confusion**, where the model misclassifies objects due to their visual similarities with other classes. In a continual learning setting, semantic confusion becomes significant due to unawareness of future classes. For example, a model that has learned to identify `car` might mistakenly recognize `bus` as `car` because it hasn't learned to distinguish between them yet.

Moreover, there is a distinct issue in continual segmentation, called **semantic drift**, as discussed in prior works [2, 5, 10]. Our baseline has a unique `no-obj` class with a corresponding MLP-classifier layer to identify whether an output mask is no object or not. However, the definition of `no-obj` shifts with each continual step, including future classes not learned yet, past classes already learned, and the background. The reliability of the *no-obj* class is essential during inference, so semantic drift significantly impacts the performance of the model.

To simultaneously tackle the semantic confusion and drift issues, we propose a simple yet effective method, called **logit manipulation**. First, we eliminate the MLP classifier for `no-obj` because the `no-obj` logit is no longer reliable due to semantic drift. Instead, we generate a new `no-obj` logit by leveraging the mutual information of old and new classes to make `no-obj` information

Method	Backbone	Trainable Params	KD	**100-5** (11 tasks)			**100-10** (6 tasks)			**100-50** (2 tasks)		
				1-100	*101-150*	*all*	*1-100*	*101-150*	*all*	*1-100*	*101-150*	*all*
FT	R50	44.9M		0.0	25.8	8.6	0.0	2.9	1.0	0.0	1.3	0.4
MiB [2]	R50	44.9M	✓	24.0	6.5	18.1	27.1	10.0	21.4	35.1	19.3	29.8
PLOP [10]	R50	44.9M	✓	28.1	15.7	24.0	30.5	17.5	26.1	40.2	22.4	34.3
CoMFormer [4]	R50	44.9M	✓	34.4	15.9	28.2	36.0	17.1	29.7	40.2	23.5	34.6
ECLIPSE	R50	**0.60M**		**41.1**	**16.6**	**32.9**	**41.4**	**18.8**	**33.9**	**41.7**	**23.5**	**35.6**
	Swin-L	0.60M		48.0	20.6	38.9	48.6	25.5	40.9	48.2	29.8	42.0
joint	R50			43.2	32.1	39.5	43.2	32.1	39.5	43.2	32.1	39.5

(a)

Method	Backbone	Trainable Params	KD	**50-10** (11 tasks)			**50-20** (6 tasks)			**50-50** (3 tasks)		
				1-50	*51-150*	*all*	*1-50*	*51-150*	*all*	*1-50*	*51-150*	*all*
FT	R50	44.9M		0.0	1.7	1.1	0.0	4.4	2.9	0.0	12.0	8.1
MiB [2]	R50	44.9M	✓	34.9	7.7	16.8	38.8	10.9	20.2	42.4	15.5	24.4
PLOP [10]	R50	44.9M	✓	39.9	15.0	23.3	43.9	16.2	25.4	45.8	18.7	27.7
CoMFormer [4]	R50	44.9M	✓	38.5	15.6	23.2	42.7	17.2	25.7	45.0	19.3	27.9
ECLIPSE	R50	**0.60M**		**45.9**	**17.3**	**26.8**	**46.4**	**19.6**	**28.6**	**46.0**	**20.7**	**29.2**
	Swin-L	0.60M		52.8	22.9	32.9	53.2	25.7	34.8	53.0	25.3	34.5
joint	R50			50.2	34.1	39.5	50.2	34.1	39.5	50.2	34.1	39.5

(b)

Table 1. **Continual Panoptic Segmentation** results on ADE20K dataset in PQ when the number of base classes $|\mathcal{C}^1|$ is (a) 100 and (b) 50. *KD* denotes using distillation strategies, which demands more trainable parameters and computational overhead. All methods use the same network of Mask2Former [9] with ResNet-50 [15] backbone. *joint* means an oracle setting training all classes offline at once.

more meaningful. For example, as shown in Figure 3, the decoder output of prompt $\mathbf{Q}^2$ is fed into other *MLP*^{1} and *MLP*^{3} layers and then the `no-obj` logit is manipulated by aggregating logits from the *MLP* layers. In our approach, as the outputs from $\mathbf{Q}^t$ are responsible for prediction only for $\mathbf{C}^t$ classes, logits of other classes that do not belong to $\mathbf{C}^t$ can be treated as `no-obj` class. Given the set of prompts $\mathbf{Q}_t$ at step t, `no-obj` logits s_t^{no-obj} are manipulated as:

$$s_t^{\mathbf{C}^{1:T}} = \text{MLP}^{1:T}(\mathbf{Q}_t), \tag{9}$$

$$s_t^{no-obj} = \delta \times (\sum_{k=1}^{t-1} s_t^{\mathbf{C}^k} + \sum_{k=t+1}^{T} s_t^{\mathbf{C}^k}), \tag{10}$$

$$c_t = \text{argmax}(s_t^{no-obj}, s_t^{\mathbf{C}^t}), \tag{11}$$

where c_t is the class indexes of output masks m_t and δ is a scalar hyperparameter for logit modulation. The dynamically manipulated `no-obj` logit helps in suppressing the propagated erroneous predictions and inherently resolves semantic drift because the `no-obj` logit is meaningfully updated. We note that the logit manipulation is applied only at the inference stage and δ is the post-processing hyperparameter. Unlike our baseline employs *softmax* activation (relative logits) in the classification, we use *sigmoid* activation (independent logits) to leverage the distinct information associated with each class, inspired by [5].

5. Experiments

5.1. Experimental Setting.

Dataset and Evaluation Metrics. We conduct experiments on ADE20K [48] dataset that consists of 150 classes with 100 things and 50 stuff categories and provides both panoptic and semantic segmentation benchmarks. Compared to COCO [29] containing an average of 7.7 instances and 3.5 classes per image and VOC [12] containing an average of 2.3 instances and 1.4 classes per image, ADE20K contains an average of 19.5 instances and 9.9 classes. We adopt Panoptic Quality (PQ) for evaluating continual panoptic segmentation performance and mean Inter-over-Union (mIoU) for continual semantic segmentation. In detail, the PQ is defined by recognition quality (RQ) and segmentation quality (SQ):

$$\text{PQ} = \underbrace{\frac{\sum_{(p,g)\in TP} \text{IoU}(p,g)}{|TP|}}_{\text{segmentation quality(SQ)}} \times \underbrace{\frac{|TP|}{|TP| + \frac{1}{2}|FP| + \frac{1}{2}|FN|}}_{\text{recognition quality(RQ)}}, \tag{12}$$

where $\text{IoU}(p, g)$ is the intersection-over-union between the predicted mask p and the ground truth g, and TP, FP, and FN denote true-positive, false-positive, and false-negative, respectively. After the last continual step T, we report the performances for the base classes ($\mathcal{C}^1$), new classes ($\mathcal{C}^{2:T}$), and all classes ($\mathcal{C}^{1:T}$).

Method	Backbone	Trainable Params	KD	**100-5** (11 tasks)			**100-10** (6 tasks)			**100-50** (2 tasks)		
				1-100	*101-150*	*all*	*1-100*	*101-150*	*all*	*1-100*	*101-150*	*all*
SDR† [34]	R101	60.4M	✓	-	-	-	28.9	7.4	21.7	37.4	24.8	33.2
UCD† [44]	R101	60.4M	✓	-	-	-	40.8	15.2	32.3	42.1	15.8	33.3
SPPA† [30]	R101	60.4M	✓	-	-	-	41.0	12.5	31.5	42.9	19.9	35.2
RCIL† [45]	R101	58.0M	✓	38.5	11.5	29.6	39.3	17.6	32.1	42.3	18.8	34.5
SSUL† [5]	R101	1.78M	✓	39.9	17.4	32.5	40.2	18.8	33.1	41.3	18.0	33.6
REMINDER† [35]	R101	60.4M	✓	-	-	-	39.0	21.3	33.1	41.6	19.2	34.1
MiB [2]	R101	63.4M	✓	21.0	6.1	16.1	23.5	10.6	26.6	37.0	24.1	32.6
PLOP [10]	R101	63.4M	✓	33.6	14.1	27.1	34.8	15.9	28.5	43.4	25.7	37.4
CoMFormer [4]	R101	63.4M	✓	39.5	13.6	30.9	40.6	15.6	32.3	43.6	26.1	37.6
ECLIPSE	R101	**0.60M**		**43.3**	**16.3**	**34.2**	**43.4**	**17.4**	**34.6**	**45.0**	**21.7**	**37.1**
joint	R101			46.9	35.6	43.1	46.9	35.6	43.1	46.9	35.6	43.1

Table 2. **Continual Semantic Segmentation** results on ADE20K dataset in mIoU. All methods use the same backbone network of ResNet-101 [15]. † denotes that using DeepLab-V3 [6] network, otherwise using Mask2Former [9] network architecture.

Incremental Protocol. Following previous continual segmentation methods [2, 4], we construct numerous challenging incremental protocols, termed as (BASE CLASSES)-(NEW CLASSES). For instance, 100-10 scenario means firstly learning 100 base classes and incrementally learning 10 new classes 5 times (T=6). The larger number of continual steps implies a more challenging scenario. The seminal work [2] introduced two different settings, *disjoint* and *overlap*. Here, we mainly follow the *overlap* setting that is more challenging and realistic and provide results with the *disjoint* setting in our supplementary material.

Implementation Details. Our implementation is based on the previous continual panoptic segmentation method, CoMFormer [4]. Specifically, we implement our method on Mask2Former [9] codebase using the backbone ResNet-50 [15] for continual panoptic segmentation and ResNet-101 for continual semantic segmentation. Moreover, we set the dimension of prompt embeddings D to 256, the number of transformer layers L to 9, and MLP^t consists of 2 hidden layers of 256 channels. We set the number of incremented prompts to the number of incremented classes, N^t=$|\mathcal{C}^t|$, and the minimum value of N^t to 10 to handle images containing more than 10 objects in $\mathcal{C}^t$. For objective functions, we employ dice and binary cross-entropy loss functions as mask loss and binary cross-entropy loss function as classification loss, after *bipartite matching* between predictions and ground-truths. We train the model for 1600 iterations per class with a learning rate of 0.0001 for the first step and 0.0005 for the following steps.

5.2. Experimental Results

Continual Panoptic Segmentation. We evaluate our approach against three previous methods (*i.e.*, MiB [2], PLOP [10], and CoMFormer [4]) and the basic fine-tuning (FT) approach, all using the Mask2Former [9] network with ResNet-50 backbone. We conduct experiments on six incremental scenarios and report their reproduced performances using the official implementation of CoMFormer. The three previous methods depend on knowledge distillation and pseudo-labeling that involve more trainable parameters and doubled network forwarding. Unlike them, our method stands out as the first without distillation, streamlining the continual learning process significantly. As shown in Table 1, our method achieved a new state-of-the-art performance across all tested scenarios, requiring only 1.3% of total trainable parameters. Notably, our method demonstrates superior retention of previously learned knowledge, even with a large number of continual steps, as evidenced in scenarios such as 100-5 and 50-10. In addition, our visual prompt tuning approach with the proposed logit manipulation strategy effectively develops the plasticity of the model for new classes, even when the base knowledge of the model is diminished from 100 to 50 classes, as shown in Table 1 (b). Our superiority can also be found in the qualitative results in Figure 5.

Continual Semantic Segmentation. We extend our evaluation to the semantic segmentation benchmark and compare our method against six previous methods [5, 30, 34, 35, 44, 45] using DeepLab-V3 [6] network and three methods (MiB, PLOP, and CoMFormer) using Mask2Former network, all using the same ResNet-101 backbone. Here again, our method is the only one without distillation. As shown in Table 2, our method achieves the best trade-off between catastrophic forgetting and plasticity, especially in difficult scenarios like 100-5. Although Mask2Former-based methods show slightly higher mIoU scores of new classes than ours in the 100-50 scenario, the reason is that our method focuses more on catastrophic forgetting than

Method	Num Prompts	Trainable Params	GPU Memory	FLOPs	**100-10** (6 tasks)		
					1-100	*101-150*	*all*
CoM-Former [4]	100	44.38M	3.28G	97.46G	36.0	17.1	29.7
	150	44.40M	3.66G	99.35G	36.5	15.9	29.6
	200	44.43M	4.03G	101.27G	36.7	12.1	28.5
ECLIPSE	100 (=50+10×5)	0.55M	0.59G	97.43G	41.2	18.7	33.7
	150 (=100+10×5)	0.55M	0.59G	99.27G	**41.4**	**18.8**	**33.9**
	200 (=100+20×5)	0.57M	0.66G	101.14G	41.0	18.7	33.4
	300 (=100+40×5)	0.62M	0.79G	104.83G	39.6	18.4	32.5

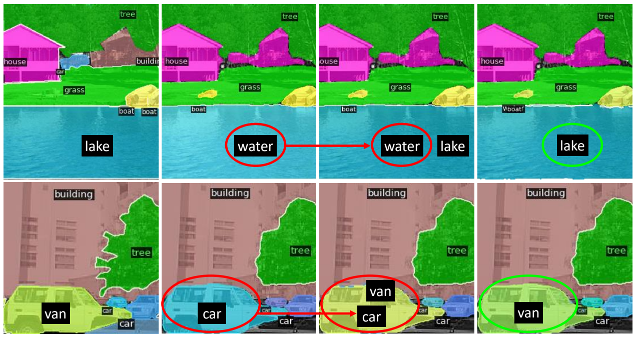
Table 3. **Effect of the number of prompts** and resulting **computational complexity**. We measure the GPU memory per a single training image. The number of prompts in ECLIPSE is denoted as `(NUM BASE PROMPTS)+(NUM NEW PROMPTS)×(NUM STEPS)`.

δ	**100-10** (6 tasks)		
	1-100	*101-150*	*all*
0.3	39.5	14.2	31.0
0.4	40.9	17.0	32.9
0.5	**41.4**	**18.8**	**33.9**
0.6	40.1	18.3	32.8
0.7	38.4	15.3	30.7

Table 4. **Effect of δ** that is the post-processing hyperparameter for modulation of logit manipulation.

Prompt Tuning	Deep Prompt	Logit Mani	**100-10** (6 tasks)		
			1-100	*101-150*	*all*
			0.2	3.9	1.3
✓			11.1	3.9	8.7
✓	✓		8.2	15.0	12.8
✓		✓	40.5	14.0	31.7
✓	✓	✓	41.4	18.8	33.9

Table 5. **Effect of the proposed components**.



Ground-Truth Step 1 Step 6 wo/ mani Step 6 w/ mani

Figure 4. **Qualitative samples for logit manipulation.** At step 1, the model, which learned classes $\mathcal{C}^1$ containing `water` and `car`, can produce incorrect predictions due to semantic confusion with unexplored classes; these errors propagate forward continuously, resulting in overlapping predictions for one object (3rd column). After the model learns new classes containing `lake` and `van` at step 6, the logit manipulation can suppress the prior errors.

plasticity and their distillation strategy is more effective in easier scenarios. Furthermore, compared to SSUL [5] that employs model freezing and fine-tunes new classifier layers and enhances plasticity using off-the-shelf saliency maps, our approach outperforms without using saliency maps.

6. Analysis

In this section, we delve into the details of our method using the ADE20K panoptic segmentation `100-10` scenario.

Effect of the number of queries. In Table 3, we conduct experiments to analyze the effect of the number of prompts in continual panoptic segmentation. The total number of prompts N is defined as $N{=}B{+}C{\times}T$, where B is the number of base prompts (step 1), C is the number of incremented prompts (step $t{>}1$), and T is the total continual steps. By default, we set $B{=}|\mathcal{C}^t|$ and $C{=}|\mathcal{C}^t|$, that is, $N{=}150{=}100{+}10{\times}5$ for `100-10` scenario. When we use the same number of prompts as our baseline work, CoMFormer [4], our ECLIPSE still significantly outperforms them. We also found that increasing the number of prompts in ECLIPSE does not help in improving the performance because the more incremented prompts lead to more false-positive predictions due to over-weighting to new classes; this tendency also appears in CoMFormer. Conversely, decreasing the number of prompts from 150 to 100 can save the FLOPs but marginally drop the performance by 0.2% due to the reduced capacity of the model.

Computational complexity. In Table 3, we measure the computational complexity of ECLIPSE and CoMFormer according to the number of prompts. Compared to CoMFormer, our ECLIPSE, which obviates the need for distillation strategies, shows 80 times less trainable parameters and 5.6 times less GPU memory usage for training. In addition, CoMFormer processes all prompts together, leading to complexity that grows quadratically with the number of queries, $O(N^2)$. In contrast, our ECLIPSE processes each set of queries $\mathbf{Q}^t$ separately, multiplying the complexity by the number of steps, $O(B^2){+}O(C^2){\times}T$, where $N{=}B{+}C{\times}T$. Even if the total number of queries is the same ($e.g.$, $N{=}100$), our method shows slightly advanced computational complexity (97.43G $v.s.$ 97.46G). However, we argue that increasing the scalability of the model as the number of classes gets larger is the natural step even in fully supervised models and our increased computation due to new prompts is marginal compared to the total FLOPs of the model (3.8% FLOPs increasing by 100 additional prompts).

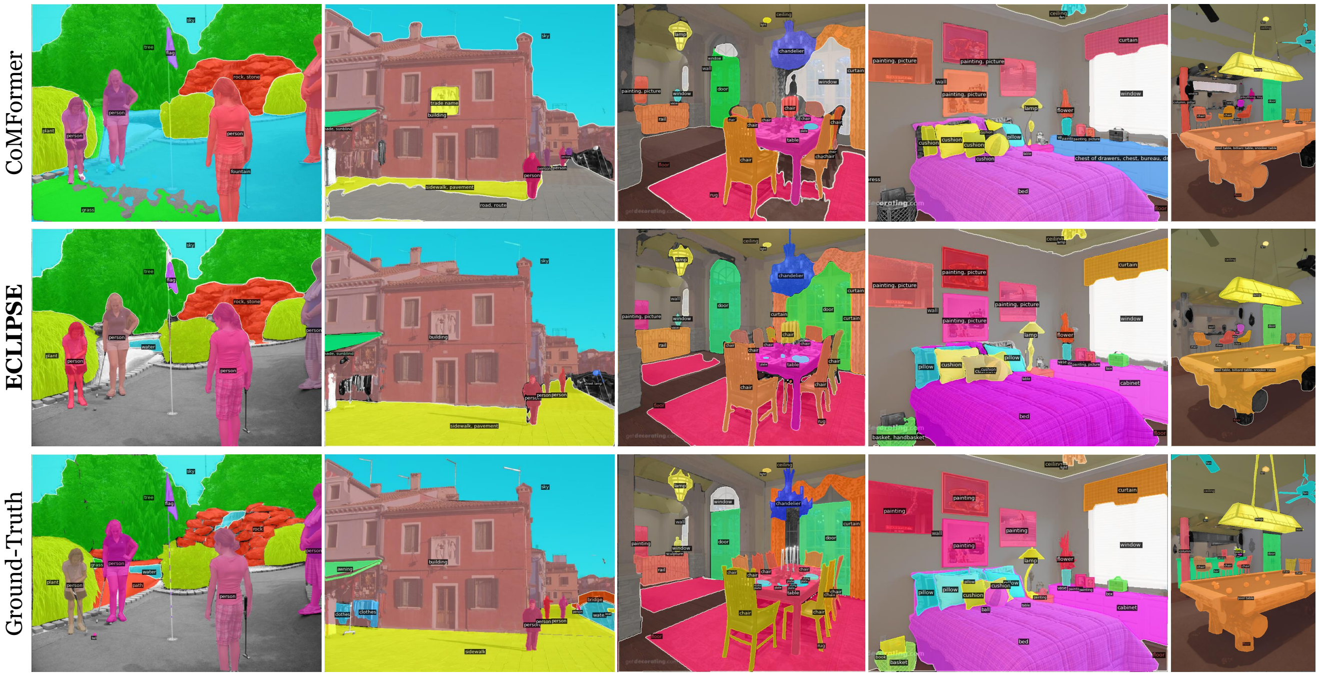


Figure 5. **Qualitative comparisons between ECLIPSE and CoMFormer [4]** on the ADE20K `100-10` continual panoptic segmentation scenario. Our ECLIPSE shows more robust results against catastrophic forgetting without reliance on distillation strategies.

Effect of visual prompt tuning. To validate the impact of the prompt tuning, we skip the model freezing and fine-tune all parameters of the model including new prompt sets without distillation strategies. As shown in the last row of Table 5, the performance is extremely dropped because the model substantially suffers from catastrophic forgetting due to the absence of model freezing or distillation strategies.

Moreover, as mentioned in Section 4.1, we have two prompt tuning strategies, termed *shallow* and *deep*. We adopt the *deep* strategy by default and the *deep* requires 100K more trainable parameters than the *shallow*. The result in the third row of Table 5 shows that adopting the *shallow* strategy noticeably drops the PQ for new classes (18.8%→14.0%) due to the reduced plasticity of the model. Considering the total number of model parameters is 63.4M, the 100K additional parameters in the *deep* strategy are efficient in developing the plasticity of the model.

Effect of the logit manipulation. To analyze the effect of the logit manipulation, we conduct an ablation study, as shown in the second row of Table 5. Without the logit manipulation, the prior errors caused by semantic confusion propagate forward (Figure 4), and the definition of `no-obj` continuously shifts as continual learning progresses, extremely diminishing the performance of old classes.

In addition, we have a post-processing hyperparameter δ for the logit manipulation. Table 4 shows that δ of 0.5 is a suitable value. We note that since the δ is a post-processing hyperparameter, it requires much less endeavor for tuning.

Exploring advanced frozen parameters. To demonstrate the potential for further improving ECLIPSE, we explore the impact of using more advanced frozen parameters of the base model. When employing the more powerful backbone network, Swin-L [31], we observe significant improvements in all tested scenarios, as shown in Table 1. Moreover, leveraging pre-trained weights from other datasets (*e.g.*, COCO [29]) substantially boosts the performance of ours, as shown in our supplementary material.

7. Conclusion and Future Direction

We presented a groundbreaking method in the field of continual panoptic segmentation. By integrating VPT with our innovative logit manipulation technique, we effectively addressed key challenges in the field. Our method not only ensures the preservation of previously learned information but also adapts to new class information. The experimental results on ADE20K dataset highlight the superiority of our approach, achieving state-of-the-art performance with a notable reduction in the number of trainable parameters. Our future direction may involve optimizing increased computational complexity resulting from expanding prompt sets, particularly when dealing with a massive number of classes.

Acknowledgment. This work was supported by Institute for Information & communications Technology Promotion(IITP) grant funded by the Korea government(MSIT) (No.2019-0-00075 Artificial Intelligence Graduate School Program(KAIST).

References

[1] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European conference on computer vision*, pages 213–229. Springer, 2020. 1

[2] Fabio Cermelli, Massimiliano Mancini, Samuel Rota Bulo, Elisa Ricci, and Barbara Caputo. Modeling the background for incremental learning in semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9233–9242, 2020. 2, 4, 5, 6

[3] Fabio Cermelli, Antonino Geraci, Dario Fontanel, and Barbara Caputo. Modeling missing annotations for incremental learning in object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3700–3710, 2022.

[4] 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*, pages 3010–3020, 2023. 2, 3, 5, 6, 7, 8

[5] Sungmin Cha, YoungJoon Yoo, Taesup Moon, et al. Ssul: Semantic segmentation with unknown label for exemplar-based class-incremental learning. *Advances in neural information processing systems*, 34:10919–10930, 2021. 2, 4, 5, 6, 7

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

[7] Bowen Cheng, Maxwell D Collins, Yukun Zhu, Ting Liu, Thomas S Huang, Hartwig Adam, and Liang-Chieh Chen. Panoptic-deeplab: A simple, strong, and fast baseline for bottom-up panoptic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12475–12485, 2020. 1, 2

[8] Bowen Cheng, Alex Schwing, and Alexander Kirillov. Per-pixel classification is not all you need for semantic segmentation. *Advances in Neural Information Processing Systems*, 34:17864–17875, 2021. 1, 2

[9] 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*, pages 1290–1299, 2022. 1, 2, 3, 5, 6

[10] 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*, pages 4040–4050, 2021. 2, 4, 5, 6

[11] Arthur Douillard, Alexandre Ramé, Guillaume Couairon, and Matthieu Cord. Dytox: Transformers for continual learning with dynamic token expansion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9285–9295, 2022.

[12] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88:303–338, 2010. 5

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

[14] Yanan Gu, Cheng Deng, and Kun Wei. Class-incremental instance segmentation via multi-teacher networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 1478–1486, 2021.

[15] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 5, 6

[16] Yu-Hsing Hsieh, Guan-Sheng Chen, Shun-Xian Cai, Ting-Yun Wei, Huei-Fang Yang, and Chu-Song Chen. Class-incremental continual learning for instance segmentation with image-level weak supervision. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1250–1261, 2023.

[17] Jitesh Jain, Jiachen Li, Mang Tik Chiu, Ali Hassani, Nikita Orlov, and Humphrey Shi. Oneformer: One transformer to rule universal image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2989–2998, 2023. 1

[18] Menglin Jia, Luming Tang, Bor-Chun Chen, Claire Cardie, Serge Belongie, Bharath Hariharan, and Ser-Nam Lim. Visual prompt tuning. In *European Conference on Computer Vision*, pages 709–727. Springer, 2022. 2, 3, 4

[19] Beomyoung Kim, Sangeun Han, and Junmo Kim. Discriminative region suppression for weakly-supervised semantic segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 1754–1761, 2021.

[20] Beomyoung Kim, Youngjoon Yoo, Chae Eun Rhee, and Junmo Kim. Beyond semantic to instance segmentation: Weakly-supervised instance segmentation via semantic knowledge transfer and self-refinement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4278–4287, 2022.

[21] Beomyoung Kim, Joonhyun Jeong, Dongyoon Han, and Sung Ju Hwang. The devil is in the points: Weakly semi-supervised instance segmentation via point-guided mask representation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11360–11370, 2023.

[22] Alexander Kirillov, Ross Girshick, Kaiming He, and Piotr Dollár. Panoptic feature pyramid networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6399–6408, 2019. 1, 2

[23] Alexander Kirillov, Kaiming He, Ross Girshick, Carsten Rother, and Piotr Dollár. Panoptic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9404–9413, 2019. 2

[24] James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, et al. Overcoming catastrophic forgetting in neural networks. *Proceedings of the national academy of sciences*, 114(13):3521–3526, 2017.

[25] Feng Li, Hao Zhang, Huaizhe Xu, Shilong Liu, Lei Zhang, Lionel M Ni, and Heung-Yeung Shum. Mask dino: Towards a unified transformer-based framework for object detection and segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3041–3050, 2023. 1, 2

[26] Yanwei Li, Hengshuang Zhao, Xiaojuan Qi, Liwei Wang, Zeming Li, Jian Sun, and Jiaya Jia. Fully convolutional networks for panoptic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 214–223, 2021. 2

[27] Zhizhong Li and Derek Hoiem. Learning without forgetting. *IEEE transactions on pattern analysis and machine intelligence*, 40(12):2935–2947, 2017.

[28] Zhiqi Li, Wenhai Wang, Enze Xie, Zhiding Yu, Anima Anandkumar, Jose M Alvarez, Ping Luo, and Tong Lu. Panoptic segformer: Delving deeper into panoptic segmentation with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1280–1289, 2022.

[29] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014. 5, 8

[30] Zihan Lin, Zilei Wang, and Yixin Zhang. Continual semantic segmentation via structure preserving and projected feature alignment. In *European Conference on Computer Vision*, pages 345–361. Springer, 2022. 6

[31] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021. 8

[32] Andrea Maracani, Umberto Michieli, Marco Toldo, and Pietro Zanuttigh. Recall: Replay-based continual learning in semantic segmentation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 7026–7035, 2021.

[33] Umberto Michieli and Pietro Zanuttigh. Incremental learning techniques for semantic segmentation. In *Proceedings of the IEEE/CVF international conference on computer vision workshops*, pages 0–0, 2019.

[34] Umberto Michieli and Pietro Zanuttigh. Continual semantic segmentation via repulsion-attraction of sparse and disentangled latent representations. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1114–1124, 2021. 2, 6

[35] Minh Hieu Phan, Son Lam Phung, Long Tran-Thanh, Abdesselam Bouzerdoum, et al. Class similarity weighted knowledge distillation for continual semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16866–16875, 2022. 2, 6

[36] David A Ross, Jongwoo Lim, Ruei-Sung Lin, and Ming-Hsuan Yang. Incremental learning for robust visual tracking. *International journal of computer vision*, 77:125–141, 2008. 1

[37] Andrei A Rusu, Neil C Rabinowitz, Guillaume Desjardins, Hubert Soyer, James Kirkpatrick, Koray Kavukcuoglu, Razvan Pascanu, and Raia Hadsell. Progressive neural networks. *arXiv preprint arXiv:1606.04671*, 2016. 1

[38] Chao Shang, Hongliang Li, Fanman Meng, Qingbo Wu, Heqian Qiu, and Lanxiao Wang. Incrementer: Transformer for class-incremental semantic segmentation with knowledge distillation focusing on old class. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7214–7224, 2023. 2

[39] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[40] Huiyu Wang, Yukun Zhu, Hartwig Adam, Alan Yuille, and Liang-Chieh Chen. Max-deeplab: End-to-end panoptic segmentation with mask transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5463–5474, 2021.

[41] Xinlong Wang, Rufeng Zhang, Tao Kong, Lei Li, and Chunhua Shen. Solov2: Dynamic and fast instance segmentation. *Advances in Neural information processing systems*, 33:17721–17732, 2020. 1, 2

[42] Zifeng Wang, Zizhao Zhang, Sayna Ebrahimi, Ruoxi Sun, Han Zhang, Chen-Yu Lee, Xiaoqi Ren, Guolong Su, Vincent Perot, Jennifer Dy, et al. Dualprompt: Complementary prompting for rehearsal-free continual learning. In *European Conference on Computer Vision*, pages 631–648. Springer, 2022. 3

[43] Zifeng Wang, Zizhao Zhang, Chen-Yu Lee, Han Zhang, Ruoxi Sun, Xiaoqi Ren, Guolong Su, Vincent Perot, Jennifer Dy, and Tomas Pfister. Learning to prompt for continual learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 139–149, 2022. 3

[44] Guanglei Yang, Enrico Fini, Dan Xu, Paolo Rota, Mingli Ding, Moin Nabi, Xavier Alameda-Pineda, and Elisa Ricci. Uncertainty-aware contrastive distillation for incremental semantic segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(2):2567–2581, 2022. 2, 6

[45] Chang-Bin Zhang, Jia-Wen Xiao, Xialei Liu, Ying-Cong Chen, and Ming-Ming Cheng. Representation compensation networks for continual semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7053–7064, 2022. 2, 6

[46] Wenwei Zhang, Jiangmiao Pang, Kai Chen, and Chen Change Loy. K-net: Towards unified image segmentation. *Advances in Neural Information Processing Systems*, 34:10326–10338, 2021. 1, 2

[47] Hanbin Zhao, Fengyu Yang, Xinghe Fu, and Xi Li. Rbc: Rectifying the biased context in continual semantic segmentation. In *European Conference on Computer Vision*, pages 55–72. Springer, 2022. 2

[48] Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through

ade20k dataset. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 633–641, 2017. 2, 5