Rethinking Query-based Transformer for Continual Image Segmentation

Yuchen Zhu* Cheng Shi* Dingyou Wang Jiajin Tang Zhengxuan Wei
Yu Wu Guanbin Li Sibei Yang†
School of Information Science and Technology, ShanghaiTech University
Sun Yat-sen University Wuhan University

Abstract

Class-incremental/Continual image segmentation (CIS) aims to train an image segmenter in stages, where the set of available categories differs at each stage. To leverage the built-in objectness of query-based transformers, which mitigates catastrophic forgetting of mask proposals, current methods often decouple mask generation from the continual learning process. This study, however, identifies two key issues with decoupled frameworks: loss of plasticity and heavy reliance on input data order. To address these, we conduct an in-depth investigation of the built-in objectness and find that highly aggregated image features provide a shortcut for queries to generate masks through simple feature alignment. Based on this, we propose SimCIS, a simple yet powerful baseline for CIS. Its core idea is to directly select image features for query assignment, ensuring "perfect alignment" to preserve objectness, while simultaneously allowing queries to select new classes to promote plasticity. To further combat catastrophic forgetting of categories, we introduce cross-stage consistency in selection and an innovative "visual query"-based replay mechanism. Experiments demonstrate that SimCIS consistently outperforms state-of-the-art methods across various segmentation tasks, settings, splits, and input data orders. All models and codes will be made publicly available at https://github.com/SooLab/SimCIS.

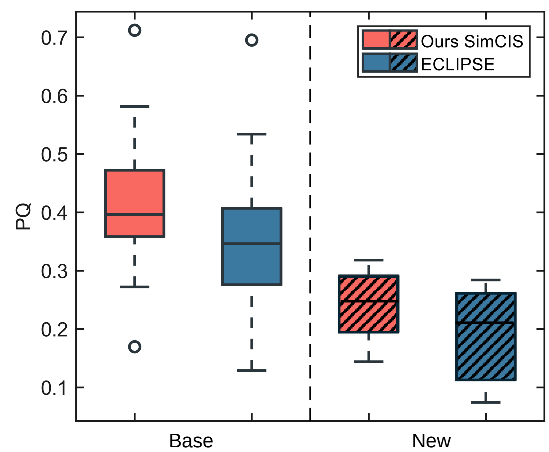

Figure 1. **Boxplots** of PQ metric for our SimCIS and previous SOTA [43] on ADE20K. We train each model on randomly shuffled continual data input orders and report average PQ for base and novel classes. We observe that recent query-based transformers suffer from a loss of plasticity (low average PQ) and heavy reliance on the input data order (high variance).

1. Introduction

Continual learning empowers models to progressively acquire, learn, and assimilate new knowledge from an ever-evolving environment. It serves as a fundamental task in image classification [5, 10, 20, 22, 28, 35, 46, 49, 55, 56, 64, 65, 67, 71, 80, 81, 84] where models are required to recognize new classes (**plasticity**) and preserve old class knowledge (avoid **catastrophic forgetting**). Extending beyond classification, continual image segmentation adapts this to the image segmentation, unlocking a myriad of practical applications [57, 60]. However, it also confronts more challenges: 1) **Additional catastrophic forgetting** of mask prediction, beyond that of class prediction; 2) **Background semantic shift** occurs when the current foreground becomes background in subsequent stages, driven by the need for image segmentation to predict the background class and the constraint of only having class annotations from current stage. Recently, query-based transformers [12, 19, 38, 39, 62, 63, 66, 70, 88] are introduced into continual image segmentation, as their **built-in objectness** has been shown to mitigate catastrophic forgetting in mask generation. Leveraging this built-in objectness, many studies [8, 29, 43, 83] decouple mask segmentation from the continual learning process by freezing the parameters associated with mask proposal generation. However, we observe two notable yet suboptimal behaviors in the aforementioned methods.

*Equal contribution
†Corresponding author

Figure 2. **Clustering results** from feature map. Pixel feature provides sufficient semantic priors (Person) even after finetuning.

- The advantage of objectness diminishes and even has a detrimental effect on plasticity as the task sequence shortens. In the shortest two-task setting, they typically achieve performance comparable to or even slightly lower than the baseline.
- The built-in objectness is fragile and lacks robustness, showing heavy dependence on the split and order of input data. As shown in Fig 1, in ten random trials, the worst trial shows a significant performance drop on new classes compared to the default setting.

Therefore, in this work, we aim to understand the built-in objectness and achieve consistent improvements (especially on plasticity) across different task lengths and varying data input orders. This is crucial, as it is impractical to assume fixed task lengths and data sequences in real-world scenarios. The conclusion from a series of investigations is:

- ❶ ***The built-in objectness emerges from the alignment between the query and the semantic priors within the image feature, mediated by the decoder.*** As shown in Fig 2, the clustering results indicate that image features contain sufficient semantic priors where pixels belonging to the same semantic are grouped together) even after finetune. Meanwhile, the query continuously aligns with specific regions of the feature map at each layer of the decoder as shown in Fig 3 (right). In summary, the highly aggregated image feature provides a shortcut for queries to generate masks by simply aligning themselves to semantic priors in the image feature through the decoder.
- ❷ ***The built-in objectness diminishes over training stages due to the query's failure to align with the semantic priors of the feature map.*** As shown in Fig 3 (left), since semantic priors vary at different stages due to background semantic shift, causing the updated learnable query to gradually misalign with the pixel feature from old classes in previous stages, even after the decoder's post-alignment (observed in ❶).

Inspired by ❶ and ❷, to ensure objectness is preserved throughout the continual learning stages, we propose a **lazy Query Pre-Alignment (QPA)** method, where query features are selected from specific locations in the image feature map, rather than being learned from scratch, to "*perfectly*" pre-align query feature with semantic priors. Specif-

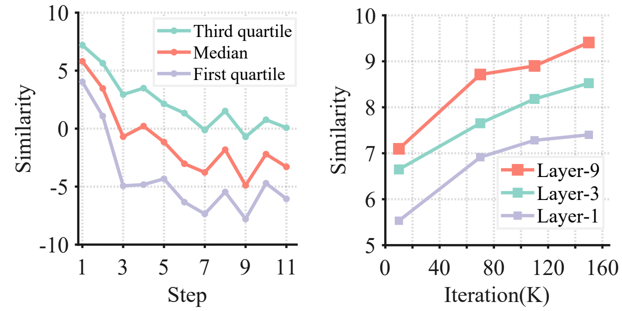

Figure 3. **Similarity** between queries and feature map changes across decoder layers and training stages (right). The query gradually misaligns with the pixel feature (left).

ically, based on the current stage's semantic classes, we select the most semantically significant locations in the image feature, preserving objectness at each stage. However, objectness is still lost across stages due to varying semantic classes in different stages.

To overcome cross-stage selection issues, a naive solution involves distillation on the feature map or query features between stages. However, in turn, while it preserves old priors from previous stages, it re-introduces incorrect priors for current stages (where old priors label current semantics as background), leading to a loss of plasticity. Fortunately, thanks to our query pre-alignment method, we can easily maintain old classes by keeping queries corresponding to old class positions, while enabling the selection of remaining queries for new classes in the current stage. Thus, we propose a **Consistent Selection Loss (CSL)** to ensure that, for the same image, the most semantically significant locations selected in the previous stage are revisited in the current stage.

With QPA and CSL, objectness in the query-based transformer is fully utilized to generate mask proposals. However, for class prediction, catastrophic forgetting may still occur. Previous methods typically rely on image replay to mitigate catastrophic forgetting. In contrast, thanks to our query pre-alignment, our query inherently contains category semantics. By storing the query feature, we can simulate specific semantics without requiring the actual image to contain the corresponding category. Therefore, we propose a novel **Virtual Query (VQ)** strategy to replay the virtual queries corresponding to previous classes in the decoder layer to avoid catastrophic forgetting. Compared to conventional image replay methods, our approach reduces storage requirements by 10x, is independent of input data order, and preserves dataset privacy.

In summary, our contributions are multi-fold:

- We provide a thorough analysis of the built-in objectness, revealing the reasons behind its emergence and demise.
- By addressing the root cause, we can successfully leverage built-in objectness to mitigate catastrophic forgetting

and background semantic shift through the introduction of three simple yet novel modules—QPA, CSL, and VQ.

- Our model, SimCIS, consistently and significantly outperforms state-of-the-art results on ADE20K in both continual panoptic and semantic segmentation.
- We introduce new dataset splits to evaluate the model's robustness to input order in continual learning. SimCIS shows superior robustness over state-of-the-art methods, thanks to the effective utilization of built-in objectness.

2. Related Work

Continual Learning is a longstanding field which possesses significant importance in addressing dynamic environments, enhancing model adaptability, and improving resource efficiency. The objective of continuous learning is to enable the model to efficiently acquire and adapt to new tasks and data, while retaining previously learned knowledge as it encounters additional information. The greatest challenge of continual learning is catastrophic forgetting [27, 56, 72]. The early research are categorized into three primary types: those that rely on regularization constraints [10, 11, 21, 22, 45, 48], those employing replay techniques [53, 56, 67], and those based on dynamic structures [24, 49, 50, 68, 80, 84] . Regularization-based methods aim to reduce the interference of new tasks on old knowledge by constraining the learning process of the model, ensuring that the model parameters remain closely aligned with previously learned representations when updated due to task changes. Replay-based methods employ strategies to store, replay [5, 37, 55, 75], or generate [53, 67, 74] samples from old tasks to mitigate catastrophic forgetting. Those methods based on dynamic structure [49, 50, 59] allocate distinct subsets of parameters to various subtasks by facilitating the expansion of their network architecture.

Universal Image Segmentation. Before MaskFormer proposed, traditional segmentation methods developed specialized architectures and models for each task to achieve top performance [3, 14–17, 31, 34, 40, 69, 77, 82, 87]. MaskFormer [18] is the first unified segmentation architecture to achieve state-of-the-art performance across three image segmentation tasks. Mask2Former [19] improves MaskFormer by adapting multi-scale features and introducing mask attention mechanism and achieve better performance. Follow its success in segmentation, we use Mask2Former as our baseline aims to extend its capability into the field of continual learning.

Continual Segmentation is the application of continual learning within the field of image segmentation. The challenge of continual segmentation tasks lies in the ability to identify new categories while generating high-quality masks for each category. This dual requirement underscores the complexity of maintaining accurate segmentation performance while adapting to an evolving set of class labels. Methods for continual segmentation are also categorized into three types as previously mentioned: regularization-based [6, 7, 23, 51, 52, 54, 61, 76, 83, 86], replay-based [8, 13, 25, 85, 90], and dynamic structure-based [1, 29, 30, 43, 78]. Among these methods, those query-based architectures demonstrate notable performance. CoMFormer [7] is the first query-based method in the field of continuous panoptic segmentation, employing distillation and pseudo label to combat catastrophic forgetting. CoMasTRe [29] is inspired by the methods of CoMFormer and, while maintaining the use of distillation loss, decouples mask and class predictions in continuous segmentation tasks. ECLIPSE [43] adapts the strategy of VPT [42], freezing the majority of model parameters and providing a set of trainable queries for fine-tuning across different tasks. BalConpas [13] attempts to combat catastrophic forgetting by employing a method that combines feature-based distillation and a replay sample set, aiming to learn new classes without negatively impacting previously acquired knowledge.

3. Preliminary

3.1. Problem Setting

Following the same continual learning setting in [7] , we train our model over T steps. At each step t, the model $\mathcal{M}^t$ has access only to a subset $\mathcal{D}^t = \{\boldsymbol{x}^t, \boldsymbol{y}^t\}$ of the entire dataset $\mathcal{D}^{1:T}$, where $\boldsymbol{x}^t \in \mathbb{R}^{C\times H\times W}$ denotes the image at the current step and $\boldsymbol{y}^t$ represents the corresponding annotations (where it can only contain annotations for classes $\mathcal{C}^t$). This setup, where each stage involves learning different classes, makes the model highly susceptible to catastrophic forgetting as it tends to lose previously acquired knowledge at each training step. Meanwhile, as the same image may appear across different learning steps with entirely different annotations, we also face the issue of so-called background shift [6]. Given these challenges, our objective is to design a model $\mathcal{M}$ such that, at any stage t, the model $\mathcal{M}^t$ not only effectively learns from $\mathcal{D}^t$ but also preserve the previous class knowledge from $\mathcal{D}^{1:t-1}$.

3.2. Mask2Former

We leverage Mask2former [19] as our meta-architecture for image segmentation. Mask2Former is a transformer-based model, which predicts a set of binary masks instead of per-pixel classification, for universal segmentation tasks. It primarily consists of three components: 1) An image encoder as backbone f_{backbone} to extract image embeddings. 2) A pixel decoder f_{pixel} to embed image embeddings to multi-scale pixel features, which we denote as F:

$$F = \{\mathcal{F}_{(l,h,w)} \,|\, \forall(l,h,w) \in \Omega\},\ \mathcal{F} \in \mathbb{R}^{D\times H_l\times W_l}, \quad (1)$$

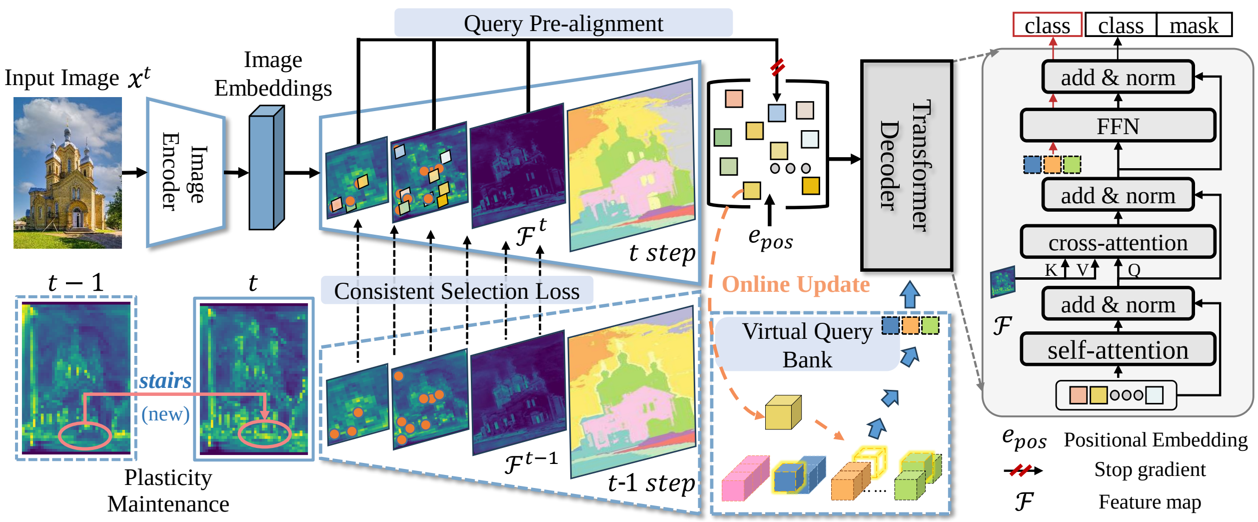

Figure 4. **The Overall Architecture of our SimCIS**: a lazy Query Pre-Alignment (Sec 4.1) with a Consistent Selection loss (Sec 4.2) to ensure built-in objectness inner and across stages, and Virtual Query (Sec 4.3) to avoid catastrophic forgetting in class prediction.

where l denotes the multi-scale layer, D represents the hidden dimension, $\mathcal{F}_{(l,h,w)}$ refers to the feature point at position (h, w) on the l-th layer and Ω represents the spatial set of multi-scale features. 3) A transformer decoder f_{decoder} takes N learnable queries $Q_N = \{q_1, q_2, \dots, q_N\} \in \mathbb{R}^{N\times D}$ with positional encodings $e_{pos} \in \mathbb{R}^{N\times D}$ to first conduct cross-attention and then self-attention with $\mathcal{F}$ as follows:

$$Q'_N = \text{FFN}(\text{SA}(\text{CA}(Q_N + e_{pos}, F))), \tag{2}$$

where CA(,) denotes the cross-attention, SA$(\cdot)$ represents self-attention, and Q'_N denotes the updated query feature. The final prediction for each query is $Z_N = \{(c_i, m_i)\}_{i=1}^{N}$, where $c_i \in \mathbb{R}^C$ and $m_i \in \mathbb{R}^{H\times W}$ represent the predicted class and mask for q_i, respectively.

4. Method

In this section, we introduce the overall architecture of our proposed SimCIS model for continual image segmentation. As shown in Fig 4, SimCIS contains three modules: 1) Lazy Query Pre-alignment (Sec 4.1), 2) Consistent Selection Loss (Sec 4.2) and Virtual Query (Sec 4.3).

4.1. Lazy Query Pre-alignment

To preserve the objectness across continual learning stages, we propose to pre-align the object query Q_N with semantic priors in the pixel feature $\mathcal{F}_{(l,h,w)}$ by directly initializing query feature with the most semantically significant pixel feature. To determine the semantic score of each pixel feature, we learn a prototype for each category and select pixel features as initial features by calculating the similarity between the pixel feature and each prototype.

Specifically, for each training step t, we maintain a set of trainable prototypes $\{p^i \,|\, i \in C^t\}$, $p^i \in \mathbb{R}^D$ for each class in C^t. By concatenating the prototypes of the past step, $\mathcal{P}^{t-1}$, with those of the current classes, we obtain the current prototype set $\mathcal{P}^t$ as follows,

$$\mathcal{P}^t = \text{concat}(\mathcal{P}^{t-1}, \{p^i \,|\, i \in C^t\}). \tag{3}$$

Then for each feature point on F, we compute its similarity with $\mathcal{P}^t$ to select the best feature points. The selection process is as follows:

$$\mathcal{I}^t = \text{topK}\left(\left\{\max\, S(\mathcal{F}^t_{(l,h,w)}, \mathcal{P}^t) \mid \forall (l,h,w) \in \Omega\right\}, N\right), \tag{4}$$

$$Q_N = \mathcal{E}^{n=t}_{m=t} = \left\{\mathcal{F}^{m=t}_i \mid i \in \mathcal{I}^{n=t}\right\}, \tag{5}$$

where $\mathcal{I} = \{(l_i, h_i, w_i)\}_{i=0}^{N} \in \Omega$ represents the spatial positions of the selected feature points, $\mathcal{E}^n_m$ represents the feature points from $\mathcal{F}^m$ selected by $\mathcal{I}^n$ and $S(,)$ denotes the similarity calculation by dot product. The topK(X, Y) function returns the indices of the Y largest values in X and N is the number of object query Q_N . We select N feature points with the highest similarity with the prototype to initialize Q_N. To supervise our selection process, we use a classification loss during training and update $\mathcal{P}^t$ through backpropagation [58]. Additionally, we apply stop gradient on Q_N to ensure that the information in F is not disrupted during training, keeping the objectness information stable across different stages.

4.2. Consistent Selection Loss

To ensure selection $\mathcal{I}$ is stable for the same image across stages, we propose a consistent selection loss. Specifically, when training our model $\mathcal{M}^t$ at current stage, we can easily obtain feature points $\mathcal{E}_t^{t-1} = \{\mathcal{F}_i^t \mid i \in \mathcal{I}^{t-1}\}$. Then, to maintain consistency in object selection across different steps, we calculate the similarity between selected feature points with $\mathcal{P}^{t-1}$, after that, we use the Kullback-Leibler (KL) divergence loss [36] to compute the loss:

$$L_{csl} = \frac{1}{|\mathcal{I}^{t-1}|} \sum_{i=1}^{|\mathcal{I}^{t-1}|} S(\mathcal{E}_{t-1}^{t-1}, \mathcal{P}^{t-1}) \log \frac{S(\mathcal{E}_{t-1}^{t-1}, \mathcal{P}^{t-1})}{S(\mathcal{E}_{t}^{t-1}, \mathcal{P}^{t-1})}. \tag{6}$$

In this way, we successfully maintain the most semantically significant locations from the previous stage, ensuring that the selection of Q_N remains stable across stages.

4.3. Virtual Query

To overcome catastrophic forgetting in class prediction, we propose the virtual query to bypass the limitations of previous methods that rely on data order. Virtual Query replays the previous query feature in the decoder layer to simulate semantics. Specifically, our innovative virtual query strategy can be divided into three steps: Firstly, we use the results of bipartite matching to select object queries and build our VQ bank. Then we analyze the pseudo-distribution to focus on rare categories in the current stage. Finally, we sample VQs in the new stage according to the pseudo-distribution and concatenate them into the object query Q_N for input into the decoder.

(1) Query Storage. During training, we maintain a queue of length h for each class, forming our virtual query bank

$$\mathcal{B}_{\text{vq}} = \{b_1^h, b_2^h, \dots, b_{|c^{1:T}|}^h\}, \tag{7}$$

where b_i^h represents a queue of length h for class i where b_i^h is the queue for class i. Queries matched through bipartite matching [4] from the decoder's final layer output, Z_N (defined in Sec 3.2), are stored in the appropriate class queues based on their bipartite matching results with ground truth $\boldsymbol{y}$.

$$\begin{cases} \mathcal{I}_b = \text{Bipartite}(Z_N, \boldsymbol{y}), \\ \mathcal{B}_{\text{vq}} \leftarrow \underset{\forall i=(i_q,i_y)\in\mathcal{I}_b}{\text{Enqueue}} (Q_N(i_q), b_{\hat{y}^{(i_y)}}), \end{cases} \tag{8}$$

where N denotes the number of queries. The set $\mathcal{I}_b$ consists of tuples, where each tuple $i = (i_q, i_y)$ represents the correspondence between query and ground truth. Here, i_q denotes the query index, and i_y denotes the ground truth index. $\hat{y}^i$ represents the class label of the i^{th} ground truth.

(2) Pseudo-Distribution Statistics. In each continual learning step, the category distribution of images changes at each stage. To ensure the decoder retains the category information for all old classes, we use the pre-trained last-stage model $\mathcal{M}^{t-1}$'s outputs on current stage's dataset D^t to simulate the distribution of real classes which helps mitigate the forgetting of rare classes in the current stage. We use this pseudo-distribution statistics by calculating

$$\omega = \left\{ \left((\sum_{i=1}^{m} \sigma_i)/\sigma_j \right)^{\frac{1}{2}} \right\}_{j=1}^{m}, \tag{9}$$

where σ_i is the pseudo number of class i in the current stage and $m = |c^{1:t-1}|$ represents the number of categories from the previous stages.

(3) VQ Utilization. Based on the pseudo-distribution statistics, in each iteration, we sample j virtual queries $Q_j = \{vq_1, \dots, vq_j\}$ for each batch based on ω. These queries are then concatenated with Q_N as

$$Q_{N+j} = \{q_1, \cdots, q_N, vq_1, \cdots, vq_j\}, \tag{10}$$

and fed into the decoder. As shown in Fig 4, within the decoder, we design a skip attention strategy for the VQs. Specifically, since the objects represented by the VQs do not appear in the image, to prevent the VQs from influencing Q_N during the self-attention and cross-attention processes, we allow the VQs to bypass the attention layers and directly affect the FFN layers as follows:

$$Q'_{N+j} = \text{FFN}(\text{concat}[\text{CA}(\text{SA}(Q_N + e_{pos}, F)), Q_j]). \tag{11}$$

Finally, the virtual query only computes L_{class} to address the model's category forgetting.

5. Experiments

5.1. Experimental Setup

Dataset and Evaluation Metric. Following previous works [7, 13, 43], we compare our SimCIS with other approaches using the ADE20K dataset [89] to evaluate its effectiveness. The images in the dataset include annotations for 150 classes, which are ranked by their total pixel ratios in the whole dataset. Among these 150 classes, 50 amorphous background classes are labeled as "stuff" classes, while 100 discrete object classes are labeled as "thing" classes. Following[7], we use Panoptic Quality (PQ) as the performance metric for continual panoptic segmentation and mean Inter-over-Union (mIoU) for continual semantic segmentation. After incremental learning steps, we report results for base classes ($\mathcal{C}^1$), new classes ($\mathcal{C}^{2:T}$), all classes ($\mathcal{C}^{1:T}$), and an average of all visible classes at each step (avg), respectively.

Continual Learning Protocol. Following existing continual segmentation methods [6–8, 13, 23, 29, 43], we evaluate our method on different continual learning settings. In particular, our incremental learning tasks are represented in the form of A-B, where A denotes the number of base classes partitioned from the dataset, and B denotes the number of

Method	**100-5** (11 tasks)				**100-10** (6 tasks)				**100-50** (2 tasks)			
	1-100	*101-150*	*all*	*avg*	*1-100*	*101-150*	*all*	*avg*	*1-100*	*101-150*	*all*	*avg*
FT	0.0	2.2	0.7	4.7	0.0	4.8	1.6	8.9	0.0	32.4	10.8	26.8
MiB [6]	2.3	0.0	1.5	13.4	6.8	0.2	4.6	19.1	23.3	14.9	20.5	31.7
PLOP [23]	31.1	11.9	24.7	31.3	37.7	23.3	32.9	37.8	42.4	23.7	36.2	39.5
SSUL [8]	30.2	7.9	22.8	27.9	31.6	11.9	25.0	30.3	35.9	18.1	30.0	33.8
CoMFormer [7]	34.4	15.9	28.2	34.0	36.0	17.1	29.7	35.3	41.1	27.7	36.7	38.8
BalConpas [13]	36.1	20.3	30.8	35.8	40.7	22.8	34.7	38.8	42.8	25.7	37.1	40.0
ECLIPSE [43]	41.1	16.6	32.9	-	41.4	18.8	33.9	-	41.7	23.5	35.6	-
Our SimCIS	**42.1**	**21.9**	**35.4**	**38.7**	**42.2**	**30.1**	**38.1**	**40.5**	**44.7**	**30.8**	**40.0**	**42.7**
joint	43.6	34.2	40.4	-	43.6	34.2	40.4	-	43.6	34.2	40.4	-

Table 1. **Continual Panoptic Segmentation** results on ADE20K dataset in PQ. All methods use the same network of Mask2Former [19] with ResNet-50 [33] backbone. *joint* means an oracle setting training all classes offline at once.

Method	**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	0.0	1.7	1.1	0.0	4.4	2.9	0.0	12.0	8.1
MiB [6]	34.9	7.7	16.8	38.8	10.9	20.2	42.4	15.5	24.4
PLOP [23]	39.9	15.0	23.3	43.9	16.2	25.4	45.8	18.7	27.7
CoMFormer [7]	38.5	15.6	23.2	42.7	17.2	25.7	45.0	19.3	27.9
ECLIPSE [43]	45.9	17.3	26.8	46.4	19.6	28.6	46.0	20.7	29.2
BalConpas [13]	44.6	24.8	31.4	49.2	28.2	35.2	51.2	26.5	34.7
Our SimCIS	**48.8**	**30.0**	**36.3**	**51.6**	**31.9**	**38.5**	**52.1**	**30.7**	**37.9**
joint	51.1	35.1	40.4	51.1	35.1	40.4	51.1	35.1	40.4

Table 2. **Continual Panoptic Segmentation** results on ADE20K dataset in PQ. All methods use Mask2Former [19] with ResNet-50 [33].

new classes. For both continual panoptic (CPS) and semantic segmentation (CSS), we conduct tasks of 100 - 5, 100 - 10, and 100 - 50. Additionally, we conduct tasks of 50 - 10, 50 - 20, and 50 - 50 for panoptic segmentation.

Implementation Details. We adapt an pre-trained ResNet-50 [33] backbone for CPS and an pre-trained ResNet-101 for CSS. Following previous work [13], the input image resolution for the CPS tasks is set to 640×640, while for the CSS tasks, it is set to 512×512. For the number of virtual queries N, it be set up to 80. For more detailes, please refer to the Appendix.

5.2. Quantitative Results

Tab 1, Tab 2 and Tab 3 present the performance of SimCIS and other approaches on the continual panoptic segmentation and semantic segmentation benchmark. In these tables, "FT" refers to fine-tuning the base model without employing continual learning methods, while "joint" indicates training the base model using all available data. They represent the lower and upper-performance bounds for continual learning methods, respectively.

Continual Panoptic Segmentation. Tab 1 and Tab 2 present the performance of SimCIS and other approaches under different continual panoptic segmentation settings. **(1)** Compared to regularization-based methods MiB [6], PLOP [23], and CoMFormer [7], SimCIS achieves superior results on both new and base classes. Notably, compared to CoMFormer, the best-performing among them, SimCIS improves PQ by $+6.0\%$ on new classes and $+7.7\%$ on base classes in the 100 - 5 task, maintaining a consistent lead in the 100 - 10 and 100 - 50 tasks. Especially in the 100 - 10 task, it surpasses CoMFormer by $+6.2\%$ PQ on base and $+13.0\%$ PQ on new classes. When using 50 base classes, SimCIS significantly outperforms these methods, demonstrating its superiority. **(2)** Compared with the method also using built-in objectness, SimCIS achieves better performance on new classes without freezing the model parameters. In the 100 - 5 , 100 - 10 , and 100 - 50 tasks, SimCIS outperforms ECLIPSE [43] by $+5.3\%$ PQ, $+11.3\%$ PQ, and $+7.6\%$ PQ, respectively. In the tasks with 50 classes as base classes, SimCIS outperforms ECLIPSE [43] by over $+10\%$ PQ on new classes, demonstrating the stability of our approach. **(3)** BalConpas [13] is a continual learning method based on the Mask2Former [19] architecture. In the 100 - 10 and 100 - 50 tasks, SimCIS outperforms BalConpas [13] by more than $+5.0\%$ PQ on new classes. In the longer step sequence of the 100 - 5 task, SimCIS surpasses BalConpas [13] by $+6.0\%$ PQ on base classes. In the 50 -

Model	100-5 (11 tasks)				100-10 (6 tasks)				100-50 (2 tasks)			
	1-100	*101-150*	*all*	*avg*	*1-100*	*101-150*	*all*	*avg*	*1-100*	*101-150*	*all*	*avg*
FT	0.0	0.3	0.1	5.6	0.0	0.1	0.0	9.1	0.0	3.2	1.1	26.3
MiB [6]	36.0	5.7	26.0	-	31.8	14.1	25.9	-	37.9	27.9	34.6	-
PLOP [23]	39.1	7.8	28.8	35.3	40.5	14.1	31.6	36.6	41.9	14.9	32.9	37.4
SSUL [8]	42.9	17.8	34.6	-	42.9	17.7	34.5	-	42.8	17.5	34.4	-
EWF [76]	41.4	13.4	32.1	-	41.5	16.3	33.2	-	41.2	21.3	34.6	-
CoMFormer [7]	39.5	13.6	30.9	36.5	40.6	15.6	32.3	37.4	39.5	26.2	38.4	41.2
ECLIPSE [43]	43.3	16.3	34.2	-	43.4	17.4	34.6	-	45.0	21.7	37.1	-
BalConpas [13]	42.1	17.2	33.8	41.3	47.3	24.2	38.6	43.6	49.9	30.1	43.3	47.4
CoMasTRe [29]	40.8	15.8	32.6	38.6	42.3	18.4	34.4	38.4	45.7	26.0	39.2	41.6
Our SimCIS	**46.7**	**22.8**	**38.7**	**47.4**	**49.7**	**27.4**	**42.3**	**49.2**	**54.9**	**36.0**	**48.6**	**52.0**
Joint	57.1	39.1	51.2	-	57.1	39.1	51.2	-	57.1	39.1	51.2	-

Table 3. **Continual Semantic Segmentation** results on the ADE20K dataset, measured by mIoU.

Psd	QPA	CSL	VQ	Panoptic 100-5 (11 tasks)			Semantic 100-5 (11 tasks)		
				1-100	*101-150*	*all*	*1-100*	*101-150*	*all*
✓				31.6	21.3	28.2	15.6	8.5	13.2
✓	✓			30.7	22.3	27.9	37.4	16.7	30.5
✓	✓	✓		35.7	24.0	31.8	43.2	17.0	34.5
✓	✓		✓	35.1	23.3	31.2	42.5	19.5	34.8
✓	✓	✓	✓	42.1	21.9	35.4	46.7	22.8	38.7

Table 4. **Ablation Study on Proposed Components.** Psd: pseudo label, QPA: lazy query pre-alignment, CSL: consistent selection loss, and VQ: virtual query.

20 and 50 - 50 tasks, SimCIS maintains strong performance, averaging +4% PQ higher than BalConpas [13] on new classes. In the longer step sequence of the 50 - 10 task, SimCIS exceeds BalConpas [13] by +4.2% PQ on base classes. It is noteworthy that in the 100 - 50 task, SimCIS almost matches the performance of the "joint", with base classes performance even exceeding that of the "joint".

Continual Semantic Segmentation. As shown in Tab 3, we further compare SimCIS with state-of-the-art works in continual semantic segmentation. **(1)** Across three tasks, SimCIS surpasses prior approaches by at least +4% mIoU on base classes. For new classes, it outperforms SSUL [8] by +5.0% and +9.7% mIoU in the 100 - 5 and 100 - 10 tasks, respectively. In the 100 - 50 task, SimCIS surpasses MiB [6], which achieves 27.9% mIoU, by +8.1% mIoU. **(2)** Among Mask2Former [19]-based methods, SimCIS also achieves the best results. In the 100 - 5 task, it outperforms ECLIPSE [43] on base classes by +3.4% mIoU and BalConpas [13] on new classes by +5.6% mIoU. In the 100 - 10 task, SimCIS achieves the performance of new classes exceeding all other architectures by at least +3.0% mIoU while maintaining high performance on base classes.

5.3. Qualitative Comparison.

Comparison with Previous SOTAs. We compare SimCIS with BalConpas [13] in the 100 - 5 continual panoptic segmentation task of the ADE20K dataset, and the visual results are illustrated in Fig 5. In the first, second, and fifth examples, BalConpas [13] encounters forgetting on base classes such as path, bus, and building. Additionally, in the third example, BalConpas incorrectly classifies the microwave and bag as cabinet and box, respectively. Benefiting from the VQ, our SimCIS has a significant advantage in preserving class information, allowing it to perform well in these examples. Furthermore, BalConpas [13] fails to provide segmentation masks for the bus and refrigerator instances in the second and third examples. In contrast, our proposed the keep built-in objectness strategy effectively preserves object information within the encoder, enabling SimCIS to accurately segment object instances.

Comparison in Different Steps. To further illustrate the effectiveness of our method, we select certain visual examples from the continual learning steps of the 100 - 5 task. In the two examples shown in Fig 6, our method is able to correct errors during the continual learning steps, such as the microwave and bag in the first image, as well as the sink, vase, and stair in the second image. SimCIS refines itself during the continual learning process, ultimately achieving accurate classification and segmentation of object instances based on our proposed flexible VQ.

5.4. Ablation Study

In this section, we report the results of the ablation experiments to validate the effectiveness of each component and configuration in our SimCIS. We select the 100 - 5 task in CPS and CSS to report the performance of SimCIS.

Main Components. As shown in Tab 4, each component contributes to the overall performance. We take Mask2Former [19] with pseudo label as our baseline performance. The second row of the table shows the performance of QPA with an increase of +18.2% mIoU on base classes and an increase of +8.2% mIoU on new classes. With the help of CSL (the third row), the CSL strategy achieves increases of +8.2% PQ and +5.8% mIoU for base classes,

Figure 5. **Qualitative comparisons** between SimCIS and BalConpas [13] on the ADE20K 100-5 continual panoptic segmentation scenario. Our SimCIS demonstrates significant results, highlighting the effectiveness of our strategies.

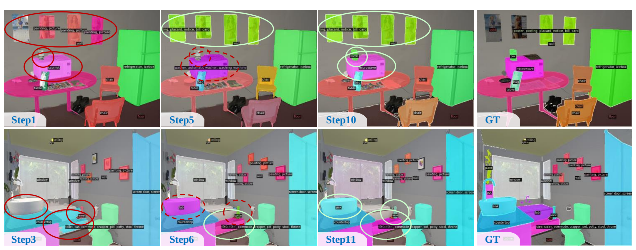

Figure 6. **Qualitative examples in continual learning.**

Reply Type	Num Samples	Disk Memory	**100-5** (11 tasks) *base*	*all*
Image	0 (*20)	0.0MB	35.7	31.8
	75 (*20)	3.4MB	38.9	33.4
	150 (*20)	6.1MB	38.9	34.0
	300 (*20)	11.8MB	38.5	33.7
	600 (*20)	21.9MB	39.2	34.3
Virtual Query	0 (*150)	0.0MB	35.7	31.8
	20 (*150)	1.5MB	40.6	34.6
	40 (*150)	3.0MB	40.4	34.1
	80 (*150)	5.9MB	**42.1**	**35.4**
	160 (*150)	12.0MB	40.9	34.2

Table 5. **Effect of Replay Type** and **Storage Requirements**.

respectively.

Effectiveness of VQ. As shown in Tab 5, compared to the conventional image replay method, our VQ strategy demonstrates significant improvements in both storage efficiency and performance. Firstly, when using 300 samples for the image replay and 80 samples for VQ, we achieve a $+1.4\%$ increase in PQ across all classes while using almost the same disk memory. When comparing the optimal cases for both storage methods, our VQ strategy outperforms the conventional image replay method by $+1.1\%$ PQ, while utilizing only 27% of the storage space.

Method	**100-10** (6 tasks) *1-100*	*101-150*	*all*
BalConpas [13]	38.9(39.4)	**27.8(26.8)**	35.2
ECLIPSE [43]	32.7(32.1)	22.3(23.8)	29.3
Ours	**40.3(40.2)**	25.4(25.7)	**35.3**
Joint	(43.6)	(34.2)	(40.4)

Table 6. **Continual Panoptic Segmentation** with random order. We also report the performance evaluated in the original class order in $(\cdot)$. For detailed experiments, please refer to the Appendix.

Robust to Input Data Order. As shown in Tab 6, our model has great robustness in random data order. We have a $+0.1\%$ PQ increase compared to BalConpas and a $+6.0\%$ PQ increase against ECLIPSE across all classes.

6. Conclusion

In this work, we present a novel class-incremental image segmentation (CIS) method called SimCIS, which addresses the challenges of catastrophic forgetting and background shift. We first explore the emergence and diminishing of built-in objectness in query-based transformers and then propose two novel modules: lazy query pre-alignment and consistent selection loss, to ensure both intra-stage and cross-stage built-in objectness. Additionally, we introduce virtual queries to mitigate catastrophic forgetting in class prediction. Comparisons with previous state-of-the-art CIS methods and our ablation study demonstrate the superiority of each individual component in our model, highlighting its effectiveness in overcoming the challenges of incremental learning. **Acknowledgment:** This work was supported by the National Natural Science Foundation of China (No.62206174).

References

[1] Donghyeon Baek, Youngmin Oh, Sanghoon Lee, Junghyup Lee, and Bumsub Ham. Decomposed knowledge distillation for class-incremental semantic segmentation. *Advances in Neural Information Processing Systems*, 35:10380–10392, 2022. 3

[2] Donghyeon Baek, Youngmin Oh, Sanghoon Lee, Junghyup Lee, and Bumsub Ham. Decomposed knowledge distillation for class-incremental semantic segmentation. *Advances in Neural Information Processing Systems*, 35:10380–10392, 2022. 16

[3] Zhaowei Cai and Nuno Vasconcelos. Cascade r-cnn: Delving into high quality object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6154–6162, 2018. 3

[4] 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. 5

[5] Francisco M Castro, Manuel J Marín-Jiménez, Nicolás Guil, Cordelia Schmid, and Karteek Alahari. End-to-end incremental learning. In *Proceedings of the European conference on computer vision (ECCV)*, pages 233–248, 2018. 1, 3

[6] 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. 3, 5, 6, 7, 14

[7] 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. 3, 5, 6, 7, 14, 15

[8] 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. 1, 3, 5, 6, 7

[9] 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. 16

[10] Arslan Chaudhry, Puneet K Dokania, Thalaiyasingam Ajanthan, and Philip HS Torr. Riemannian walk for incremental learning: Understanding forgetting and intransigence. In *Proceedings of the European conference on computer vision (ECCV)*, pages 532–547, 2018. 1, 3

[11] Arslan Chaudhry, Marc'Aurelio Ranzato, Marcus Rohrbach, and Mohamed Elhoseiny. Efficient lifelong learning with a-gem. *arXiv preprint arXiv:1812.00420*, 2018. 3

[12] Chaoqi Chen, Yushuang Wu, Qiyuan Dai, Hong-Yu Zhou, Mutian Xu, Sibei Yang, Xiaoguang Han, and Yizhou Yu. A survey on graph neural networks and graph transformers in computer vision: A task-oriented perspective. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024. 1

[13] Jinpeng Chen, Runmin Cong, Yuxuan Luo, Horace Ho Shing Ip, and Sam Kwong. Strike a balance in continual panoptic segmentation, 2024. 3, 5, 6, 7, 8, 16, 18

[14] Liang-Chieh Chen. Semantic image segmentation with deep convolutional nets and fully connected crfs. *arXiv preprint arXiv:1412.7062*, 2014. 3

[15] Liang-Chieh Chen. Rethinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv:1706.05587*, 2017.

[16] 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. *IEEE transactions on pattern analysis and machine intelligence*, 40(4):834–848, 2017.

[17] Bowen Cheng, Liang-Chieh Chen, Yunchao Wei, Yukun Zhu, Zilong Huang, Jinjun Xiong, Thomas S Huang, Wen-Mei Hwu, and Honghui Shi. Spgnet: Semantic prediction guidance for scene parsing. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 5218–5228, 2019. 3

[18] 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. 3

[19] 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, 3, 6, 7, 15, 16

[20] Qiyuan Dai and Sibei Yang. Curriculum point prompting for weakly-supervised referring image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13711–13722, 2024. 1

[21] Prithviraj Dhar, Rajat Vikram Singh, Kuan-Chuan Peng, Ziyan Wu, and Rama Chellappa. Learning without memorizing. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5138–5146, 2019. 3

[22] Arthur Douillard, Matthieu Cord, Charles Ollion, Thomas Robert, and Eduardo Valle. Podnet: Pooled outputs distillation for small-tasks incremental learning. In *Computer vision–ECCV 2020: 16th European conference, Glasgow, UK, August 23–28, 2020, proceedings, part XX 16*, pages 86–102. Springer, 2020. 1, 3

[23] 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. 3, 5, 6, 7

[24] 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. 3

[25] Mostafa ElAraby, Ali Harakeh, and Liam Paull. Bacs: Background aware continual semantic segmentation. *arXiv preprint arXiv:2404.13148*, 2024. 3

[26] 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. 14

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

[28] Weifeng Ge, Sibei Yang, and Yizhou Yu. Multi-evidence filtering and fusion for multi-label classification, object detection and semantic segmentation based on weakly supervised learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1277–1286, 2018. 1

[29] 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*, pages 3848–3857, 2024. 1, 3, 5, 7, 14, 16

[30] Dipam Goswami, René Schuster, Joost van de Weijer, and Didier Stricker. Attribution-aware weight transfer: A warm-start initialization for class-incremental semantic segmentation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 3195–3204, 2023. 3

[31] Bharath Hariharan, Pablo Arbeláez, Ross Girshick, and Jitendra Malik. Simultaneous detection and segmentation. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part VII 13*, pages 297–312. Springer, 2014. 3

[32] JA Hartigan. Clustering algorithms. *John Wiley google schola*, 2:25–47, 1975. 16

[33] 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. 6

[34] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017. 3

[35] Xiang He, Sibei Yang, Guanbin Li, Haofeng Li, Huiyou Chang, and Yizhou Yu. Non-local context encoder: Robust biomedical image segmentation against adversarial attacks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 8417–8424, 2019. 1

[36] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *stat*, 1050:9, 2015. 5

[37] Saihui Hou, Xinyu Pan, Chen Change Loy, Zilei Wang, and Dahua Lin. Learning a unified classifier incrementally via rebalancing. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 831–839, 2019. 3

[38] Hanzhuo Huang, Yufan Feng, Cheng Shi, Lan Xu, Jingyi Yu, and Sibei Yang. Free-bloom: Zero-shot text-to-video generator with llm director and ldm animator. *Advances in Neural Information Processing Systems*, 36:26135–26158, 2023. 1

[39] Hanzhuo Huang, Yuan Liu, Ge Zheng, Jiepeng Wang, Zhiyang Dou, and Sibei Yang. Mvtokenflow: High-quality 4d content generation using multiview token flow. *arXiv preprint arXiv:2502.11697*, 2025. 1

[40] Zilong Huang, Xinggang Wang, Lichao Huang, Chang Huang, Yunchao Wei, and Wenyu Liu. Ccnet: Criss-cross attention for semantic segmentation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 603–612, 2019. 3

[41] 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. 16

[42] 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. 3, 15

[43] 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)*, pages 3346–3356, 2024. 1, 3, 5, 6, 7, 8, 14, 15, 16

[44] 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. 16

[45] Zhizhong Li and Derek Hoiem. Learning without forgetting. *IEEE transactions on pattern analysis and machine intelligence*, 40(12):2935–2947, 2017. 3

[46] Liang Lin, Pengxiang Yan, Xiaoqian Xu, Sibei Yang, Kun Zeng, and Guanbin Li. Structured attention network for referring image segmentation. *IEEE Transactions on Multimedia*, 24:1922–1932, 2021. 1

[47] 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. 15

[48] David Lopez-Paz and Marc'Aurelio Ranzato. Gradient episodic memory for continual learning. *Advances in neural information processing systems*, 30, 2017. 3

[49] Arun Mallya and Svetlana Lazebnik. Packnet: Adding multiple tasks to a single network by iterative pruning. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 7765–7773, 2018. 1, 3

[50] Arun Mallya, Dillon Davis, and Svetlana Lazebnik. Piggyback: Adapting a single network to multiple tasks by learning to mask weights. In *Proceedings of the European conference on computer vision (ECCV)*, pages 67–82, 2018. 3

[51] 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. 3

[52] 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. 3

[53] Oleksiy Ostapenko, Mihai Puscas, Tassilo Klein, Patrick Jahnichen, and Moin Nabi. Learning to remember: A synaptic plasticity driven framework for continual learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11321–11329, 2019. 3

[54] 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. 3

[55] Sylvestre-Alvise Rebuffi, Alexander Kolesnikov, Georg Sperl, and Christoph H Lampert. icarl: Incremental classifier and representation learning. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 2001–2010, 2017. 1, 3

[56] Anthony Robins. Catastrophic forgetting, rehearsal and pseudorehearsal. *Connection Science*, 7(2):123–146, 1995. 1, 3

[57] 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

[58] David E Rumelhart, Geoffrey E Hinton, and Ronald J Williams. Learning representations by back-propagating errors. *nature*, 323(6088):533–536, 1986. 4

[59] 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. 3

[60] 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

[61] 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. 3

[62] Cheng Shi and Sibei Yang. Edadet: Open-vocabulary object detection using early dense alignment. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 15724–15734, 2023. 1

[63] Cheng Shi and Sibei Yang. Logoprompt: Synthetic text images can be good visual prompts for vision-language models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2932–2941, 2023. 1

[64] Cheng Shi and Sibei Yang. The devil is in the object boundary: Towards annotation-free instance segmentation using foundation models. *arXiv preprint arXiv:2404.11957*, 2024. 1

[65] Cheng Shi, Yulin Zhang, Bin Yang, Jiajin Tang, Yuexin Ma, and Sibei Yang. Part2object: Hierarchical unsupervised 3d instance segmentation. In *European Conference on Computer Vision*, pages 1–18. Springer, 2024. 1

[66] Cheng Shi, Yuchen Zhu, and Sibei Yang. Plain-det: A plain multi-dataset object detector. In *European Conference on Computer Vision*, pages 210–226. Springer, 2024. 1

[67] Hanul Shin, Jung Kwon Lee, Jaehong Kim, and Jiwon Kim. Continual learning with deep generative replay. *Advances in neural information processing systems*, 30, 2017. 1, 3

[68] Pravendra Singh, Vinay Kumar Verma, Pratik Mazumder, Lawrence Carin, and Piyush Rai. Calibrating cnns for lifelong learning. *Advances in Neural Information Processing Systems*, 33:15579–15590, 2020. 3

[69] Robin Strudel, Ricardo Garcia, Ivan Laptev, and Cordelia Schmid. Segmenter: Transformer for semantic segmentation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 7262–7272, 2021. 3

[70] Jiajin Tang, Ge Zheng, Cheng Shi, and Sibei Yang. Contrastive grouping with transformer for referring image segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 23570–23580, 2023. 1

[71] Jiajin Tang, Ge Zheng, and Sibei Yang. Temporal collection and distribution for referring video object segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15466–15476, 2023. 1

[72] Sebastian Thrun. Lifelong learning algorithms. In *Learning to learn*, pages 181–209. Springer, 1998. 3

[73] 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. 14

[74] Chenshen Wu, Luis Herranz, Xialei Liu, Joost Van De Weijer, Bogdan Raducanu, et al. Memory replay gans: Learning to generate new categories without forgetting. *Advances in neural information processing systems*, 31, 2018. 3

[75] Yue Wu, Yinpeng Chen, Lijuan Wang, Yuancheng Ye, Zicheng Liu, Yandong Guo, and Yun Fu. Large scale incremental learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 374–382, 2019. 3

[76] Jia-Wen Xiao, Chang-Bin Zhang, Jiekang Feng, Xialei Liu, Joost van de Weijer, and Ming-Ming Cheng. Endpoints weight fusion for class incremental semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7204–7213, 2023. 3, 7, 16

[77] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M Alvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. *Advances in neural information processing systems*, 34: 12077–12090, 2021. 3

[78] Zhengyuan Xie, Haiquan Lu, Jia-wen Xiao, Enguang Wang, Le Zhang, and Xialei Liu. Early preparation pays off: New classifier pre-tuning for class incremental semantic segmentation. *arXiv preprint arXiv:2407.14142*, 2024. 3

[79] Zhengyuan Xie, Haiquan Lu, Jia-wen Xiao, Enguang Wang, Le Zhang, and Xialei Liu. Early preparation pays off: New

classifier pre-tuning for class incremental semantic segmentation. In *European Conference on Computer Vision*, pages 183–201. Springer, 2025. 16

[80] Shipeng Yan, Jiangwei Xie, and Xuming He. Der: Dynamically expandable representation for class incremental learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3014–3023, 2021. 1, 3

[81] Sibei Yang, Meng Xia, Guanbin Li, Hong-Yu Zhou, and Yizhou Yu. Bottom-up shift and reasoning for referring image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11266–11275, 2021. 1

[82] Yuhui Yuan, Lang Huang, Jianyuan Guo, Chao Zhang, Xilin Chen, and Jingdong Wang. Ocnet: Object context network for scene parsing. *arXiv preprint arXiv:1809.00916*, 2018. 3

[83] 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. 1, 3

[84] Gengwei Zhang, Liyuan Wang, Guoliang Kang, Ling Chen, and Yunchao Wei. Slca: Slow learner with classifier alignment for continual learning on a pre-trained model. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 19148–19158, 2023. 1, 3

[85] Zekang Zhang, Guangyu Gao, Zhiyuan Fang, Jianbo Jiao, and Yunchao Wei. Mining unseen classes via regional objectness: A simple baseline for incremental segmentation. *Advances in neural information processing systems*, 35:24340–24353, 2022. 3

[86] 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*, pages 843–853, 2023. 3

[87] 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*, pages 2881–2890, 2017. 3

[88] Ge Zheng, Bin Yang, Jiajin Tang, Hong-Yu Zhou, and Sibei Yang. Ddcot: Duty-distinct chain-of-thought prompting for multimodal reasoning in language models. *Advances in Neural Information Processing Systems*, 36:5168–5191, 2023. 1

[89] 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. 5, 14

[90] Lanyun Zhu, Tianrun Chen, Jianxiong Yin, Simon See, and Jun Liu. Continual semantic segmentation with automatic memory sample selection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3082–3092, 2023. 3

Rethinking Query-based Transformer for Continual Image Segmentation

Supplementary Material

In this supplementary material, we provide additional information regarding:

- Overall Workflow of our SimCIS with Pseudocode (In Sec. 7).
- More Dataset and Implementation Details (In Sec. 8).
- Comprehensive Experiments of Random Class Rrder (In Sec. 9).
- More Ablation Studies on the Stop-Gradient Strategy. (In Sec. 10).
- More Visualization Results of the Continual Semantic Segmentation task (In Sec. 12).
- More Visualization Results of Objectness Information (In Sec. 13).
- Discussion, Limitation and Future Work (In Sec. 15).

7. Pseudocode for our SimCIS

In this section, we present the overall workflow of our method in the pseudo-code Algo. 1. At the beginning, we define some modules, functions, and variables. For the current stage t and the previous stage $t-1$, we define the backbone modules f^t_{backbone} and $f^{t-1}_{\text{backbone}}$, the pixel decoder modules f^t_{pixel} and f^{t-1}_{pixel}, the prototypes $\mathcal{P}^t$ and $\mathcal{P}^{t-1}$ respectively. For clarity and readability of the pseudocode, some formulas introduced in the main text are encapsulated as functions. These include the select feature points function Φ (Eq. 4), the consistent selection loss function l_{csl} (Eq. 6), the calculate sample weights function g (Eq. 9), the virtual query bank $\mathcal{B}_{vq}$ update function $\mathcal{U}$ (Eq. 8), and the decoder layer with skip attention Θ(Eq. 11). We also define the input image for the current stage as x^t, the Virtual Query Bank $\mathcal{B}_{vq}$, and the total training iteration M. Specifically, our lazy **Query Pre-alignment** strategy is described in the line-4 and line-8-9, our **Consistent Selection Loss** strategy is described in the line-5-7, and our **Virtual Query** strategy is described in line-11-12, line-10-16. All model and code will be made publicly available.

Algorithm 1 Pseudocode for SimCIS

Input: Backbone f_{backbone}, pixel decoder f_{pixel} and prototype $\mathcal{P}$ at stage t and $t-1$;
Select feature points function Φ (Eq. 4);
Consistent selection loss function l_{csl}(Eq. 6);
Calculate sample weights function g (Eq. 9);
$\mathcal{B}_{vq}$ update function $\mathcal{U}$(Eq. 8);
Decoder layer with skip attention Θ (Eq. 11);
Image of current stage x^t;
Virtual Query Bank $\mathcal{B}_{vq}$;
Training iteration M.
Output: $\mathcal{M}^t$: model of current stage.

1: $\sigma \leftarrow$ Collect pseudo-distribution statistics
2: $\omega \leftarrow g(\sigma)$
3: **for** $i \leftarrow 1, \ldots, M$ **do**
4: $\quad F^t \leftarrow f_{\text{pixel}}(f_{\text{encoder}}(x^t))$
5: $\quad F^{t-1} \leftarrow f^{t-1}_{\text{pixel}}(f^{t-1}_{\text{encoder}}(x^t))$
6: $\quad \mathcal{I}^{t-1} \leftarrow \Phi(F^{t-1}, \mathcal{P}^{t-1})$
7: $\quad \mathcal{L}_{csl} \leftarrow l_{\text{cls}}(F^t, F^{t-1}, \mathcal{I}^{t-1}, \mathcal{P}^{t-1})$ ▷ Sec. 4.2 end.
8: $\quad \mathcal{I}^t \leftarrow \Phi(F^t, P^t)$
9: $\quad Q_N \leftarrow$ Object query on F^t by $\mathcal{I}^t$. ▷ Sec. 4.1 end.
10: $\quad Q_j \leftarrow$ Sample j virtual query from $\mathcal{B}_{vq}$ using ω.
11: $\quad Q_{N+j} \leftarrow \{Q_N, Q_j\}$
12: $\quad$ **for** $l \leftarrow 1, \ldots, L$ **do**
13: $\quad\quad Q_{N+j} \leftarrow \Theta(Q_{N+j})$
14: $\quad$ **end for**
15: $\quad Z_N \leftarrow$ Get Q_N's prediction results.
16: $\quad \mathcal{B}_{vq} \leftarrow \mathcal{U}(Z_N, Q_N, y)$ ▷ Sec. 4.3 end.
17: $\quad$ Calculate L_{class} using Q_{N+j}.
18: $\quad$ Calculate L_{mask} using Q_N.
19: $\quad L_{\text{total}} \leftarrow L_{\text{class}} + L_{\text{mask}} + L_{\text{csl}}$
20: $\quad$ Update parameters via backpropagation.
21: **end for**

8. More Dataset and Implementation Details

Dataset Information. Following previous works [7, 29, 43], we use ADE20k [89] to train and evaluate our model for both continual panoptic segmentation and continual semantic segmentation tasks. The ADE20K dataset contains $20,210$ training images and $2,000$ validation images, with each image averaging 19.5 instances and 10.5 classes. Compared with other datasets, such as VOC [26], which contains an average of 2.3 instances and 1.4 classes per image. ADE20K is a particularly challenging dataset that highlights our robustness during continual training stages.

Implementation Details. To ensure a fair comparison, we strictly follow previous works [6, 7, 43, 73]. In the initial training step, the learning rate is set up to 1e-4, and during the incremental learning phase, it is reduced to $5e$-5. The total training iteration is set to $160,000$ in the first step and $1,000$ iterations for each class in incremental steps. We utilize a multi-step strategy to dynamically adjust our learning rate for optimizing our model, with a decay factor set to 0.1. Following [6], there are two different experimental protocols: disjoint and overlap. In the disjoint setting, each task has its own exclusive image data, while the overlap setting allows different images to appear across tasks. We choose the more challenging overlap setting as our experimental

Random ID	Our SimCIS			ECLIPSE		
	1-100	*101-150*	*all*	*1-100*	*101-150*	*all*
1	41.2	28.9	37.1	33.4	20.4	29.1
2	42.2	30.2	38.2	32.1	23.0	29.1
3	41.1	29.8	37.3	32.2	23.3	29.3
4	42.2	29.6	38.0	30.4	18.0	26.3
5	41.2	30.5	37.6	32.2	22.8	29.1
6	41.7	27.5	37.0	28.5	24.3	27.1
7	41.9	28.8	37.6	34.3	18.8	29.2
8	40.0	29.9	36.6	30.4	22.7	27.9
9	42.0	28.7	37.6	32.7	22.2	29.2
10†	39.1	33.8	37.4	11.3	0.0	7.6
Origin	42.2	30.1	38.1	41.4	18.8	33.9

Table 7. **Continual Panoptic Segmentation with 10 random order** on the ADE20K 100-5 continual panoptic segmentation scenario. † means descending order. Origin means original ascending order.

protocol. Except for setting consistent select loss weight to 2.0, we follow Mask2former [19] to set other loss weights.

9. Continual Learning with Random Order

Experiment Details. As shown in Tab. 7, we conduct extensive experiments on our model and ECLIPSE [43] under the ten random orders (detailed orders shown in Tab. 10), where nine of them were completely randomly generated using the `random` module in `Numpy` without any manual selection. As ADE20k's classes are ranked by their total pixel ratios in the entire dataset, we deliberately set the last order to descending to evaluate the model's dependency on base categories. Specifically, the descending order forces the model first to learn rare categories, enabling us to assess its continual learning ability under such challenging conditions.

Comparison with ECLIPSE. The results are shown in Tab. 7. Our model achieves SOTAs across all 10 random orders. Overall, our model achieves an increase of 41.9% across all classes compared to ECLIPSE. Specifically, the average performance of old classes improves by $+11.5\%$ PQ, and new classes see an average improvement of $+10.2\%$ PQ. In the final experiment, where we set the categories in descending order, the performance of ECLIPSE is relatively dropped by 73.9%. This demonstrates that ECLIPSE's approach, which freezes other parameters and employs the VPT [42] strategy for model updates, strongly depends on the base class during continual learning. In contrast, our model remains stable even under this highly challenging setup.

10. More Ablation Study for Stop Gradient

As we mention in the main text, we apply stop gradient on selected object query Q_N after the QPA strategy, to ensure that the information in feature map F is not disrupted during training, keeping the objectness information stable across different stages. As shown in the Tab. 8. After using the stop gradient strategy, we achieve an increase of +2.1% PQ across all classes. All the experiments in the main text use this strategy unless otherwise specified.

Psd	QPA	CSL	VQ	SG	**Panoptic 100-5** (11 tasks)		
					1-100	*101-150*	*all*
✓	✓	✓	✓		39.5	20.7	33.3
✓	✓	✓	✓	✓	42.1	21.9	35.4

Table 8. **Ablation Study on Stop Gradient.** Psd: pseudo label, QPA: lazy query pre-alignment, CSL: consistent selection loss, and SG: stop gradient.

11. Performance on COCO Ponaptic

To demonstrate the robustness of our approach across diverse datasets, we present its performance on the COCO panoptic segmentation dataset. As illustrated in Tabel 9, our method demonstrates strong adaptability across diverse datasets. Baseline means Mask2Former [19] with only pseudo label strategy [7].

Method	**Panoptic 83-5** (11 tasks)		
	1-83	84-133	all
Baseline	34.3	20.9	29.3
Our SimCIS	39.5	23.7	33.6

Table 9. **Continual Panoptic Segmentation.** Results on COCO [47] panoptic segmentation dataset where the total number of classes is 133 in PQ under the overlap setting.

12. More Visualization Results for CSS

As shown in Fig. 7, we additionally compare our SimCIS with BalConpas [13] in the 100-5 continual semantic segmentation task. In the first, second, and fourth row from Fig. 7, BalConpas encounters misclassification of the TV and lamps. In the fourth image, Balconpas fails to predict the building's accurate mask. While benefiting from the proper utilization of semantic priors in pixel feature and VQ strategy's ability to preserve class information, our SimCIS performs well in these cases.

13. Built-in Objectness Maintenance

Detailed clustering implementation. In the multi-scale feature generated by the pixel decoder, we choose the feature with the highest resolution for clustering. To evaluate the quality of objectness information contained in the features, we applied the K-means [32] algorithm for clustering. Regarding the hyperparameter settings, for the images shown in Fig. 8, we set the number of clustering centers from top to bottom as $[15, 10, 15, 15, 15, 15, 15]$.

SimCIS provides stable built-in objectness. Although pixel features can generally provide semantic priors across various methods, our observations indicate that they are still influenced by the continual learning process. In this section, we visually demonstrate that our SimCIS has the ability to maintain object information. As shown in Fig. 8, in the first image, the clustering results of Balconpas around the jeep exhibit significantly more noise. In the last image, Balconpas fails to capture the entire helicopter, while our feature successfully preserves the complete object information.

14. The Order of Attention Layers

In Mask2Former [19], the authors employ a cross then self-attention mechanism, as they argue that query features to the first self-attention layer are image-independent and do not have signals from the image, thus applying self-attention is unlikely to enrich information. However, in our proposed Lazy Query Pre-alignment strategy, the query features have rich information. Therefore, we revert to the conventional sequence of cross then self-attention. This modification, however, does not exhibit any significant impact on the experimental outcomes.

15. Discussion, Limitation and Future Work

Discussion of the choice of meta-architecture for image segmentation. To ensure a fair comparison, we adopt the same Mask2Former [19] as our meta-architecture for image segmentation. However, recent years have witnessed rapid advancements in transformer-based universal image segmentor [41, 44], which achieves a much stronger performance on the segmentation benchmark. We leave the investigation of other meta-architectures as future work.

Discussion of other common techniques/tricks in CIS. To maintain the simplicity and elegance of our SimCIS, we have discarded certain continual learning techniques/tricks commonly used in previous methods, such as model weight fusion across stages [76], specific initialization methods [2, 9, 79] for the classifier head, and freezing model parameters [29, 43]. Whether these techniques/tricks can further improve SimCIS's performance remains an open question for future work.

ID	Category Order
1	[71, 135, 3, 60, 74, 1, 10, 40, 118, 91, 52, 50, 59, 146, 33, 42, 66, 148, 41, 78, 46, 14, 26, 57, 73, 96, 89, 55, 149, 84, 13, 2, 77, 54, 32, 138, 64, 81, 129, 104, 93, 86, 62, 130, 21, 125, 128, 136, 12, 65, 79, 43, 4, 134, 68, 145, 99, 15, 58, 29, 111, 51, 56, 11, 117, 102, 140, 105, 116, 131, 18, 120, 22, 19, 85, 28, 0, 123, 38, 95, 115, 17, 70, 61, 20, 112, 109, 67, 98, 133, 30, 76, 49, 8, 101, 47, 25, 48, 147, 132, 100, 44, 69, 6, 53, 126, 7, 75, 90, 83, 107, 106, 9, 113, 37, 122, 121, 143, 103, 137, 80, 144, 94, 142, 110, 63, 124, 87, 35, 24, 88, 39, 139, 27, 92, 23, 114, 119, 141, 108, 5, 45, 72, 31, 36, 127, 82, 16, 97, 34]
2	[11, 114, 103, 122, 48, 41, 85, 92, 113, 64, 3, 80, 110, 10, 112, 30, 96, 101, 102, 9, 7, 21, 17, 37, 93, 77, 73, 94, 59, 135, 2, 123, 98, 130, 49, 129, 25, 66, 50, 145, 76, 147, 83, 90, 63, 111, 27, 126, 1, 65, 75, 119, 12, 78, 5, 143, 15, 29, 71, 22, 89, 115, 84, 16, 120, 139, 38, 68, 146, 116, 35, 124, 97, 23, 39, 117, 13, 18, 108, 138, 33, 134, 141, 62, 105, 142, 40, 26, 8, 46, 144, 95, 131, 99, 104, 19, 60, 132, 6, 42, 4, 140, 128, 55, 32, 70, 118, 100, 125, 127, 87, 52, 45, 31, 81, 88, 44, 24, 20, 56, 82, 61, 28, 34, 148, 14, 53, 121, 47, 133, 57, 137, 67, 136, 106, 36, 58, 109, 107, 72, 91, 86, 43, 74, 69, 0, 149, 51, 79, 54]
3	[74, 149, 75, 46, 113, 67, 118, 89, 130, 7, 119, 33, 77, 39, 96, 81, 112, 37, 124, 1, 34, 105, 35, 80, 135, 13, 143, 53, 9, 101, 22, 57, 139, 138, 12, 123, 48, 63, 60, 69, 117, 71, 4, 65, 127, 84, 97, 59, 70, 91, 128, 142, 41, 99, 136, 32, 108, 120, 42, 145, 148, 104, 87, 132, 52, 5, 85, 61, 10, 121, 49, 44, 17, 115, 93, 134, 68, 3, 110, 36, 133, 102, 0, 16, 55, 90, 83, 54, 62, 94, 126, 6, 19, 18, 26, 51, 114, 31, 43, 45, 76, 131, 25, 66, 92, 29, 50, 40, 100, 58, 109, 20, 30, 98, 86, 14, 28, 107, 122, 11, 111, 64, 21, 72, 103, 137, 23, 88, 125, 140, 47, 146, 27, 116, 141, 78, 79, 24, 95, 2, 144, 38, 82, 56, 106, 129, 147, 73, 8, 15]
4	[60, 110, 89, 119, 147, 123, 116, 35, 22, 1, 36, 99, 58, 17, 43, 11, 109, 130, 113, 138, 65, 94, 74, 8, 106, 12, 29, 118, 24, 136, 140, 21, 6, 93, 142, 9, 71, 135, 54, 114, 121, 77, 16, 105, 117, 5, 67, 86, 61, 97, 20, 76, 18, 84, 103, 46, 96, 0, 141, 100, 63, 131, 31, 45, 81, 73, 13, 124, 79, 48, 40, 132, 102, 112, 107, 44, 27, 49, 134, 85, 144, 66, 83, 104, 75, 88, 101, 82, 19, 47, 87, 122, 125, 115, 72, 137, 7, 128, 78, 15, 90, 51, 145, 39, 2, 126, 64, 139, 41, 55, 34, 26, 3, 129, 69, 68, 120, 98, 92, 57, 59, 70, 23, 80, 148, 10, 149, 52, 38, 42, 53, 108, 127, 91, 50, 95, 146, 56, 33, 30, 111, 25, 62, 32, 4, 37, 14, 143, 133, 28]
5	[77, 20, 111, 65, 117, 53, 43, 90, 28, 79, 134, 45, 116, 98, 92, 105, 137, 10, 6, 59, 67, 34, 44, 99, 55, 147, 1, 80, 122, 54, 56, 12, 31, 49, 37, 61, 108, 133, 143, 130, 70, 95, 132, 2, 115, 118, 81, 47, 51, 121, 14, 3, 8, 21, 22, 62, 78, 72, 39, 25, 23, 142, 149, 50, 83, 11, 52, 141, 129, 113, 4, 148, 144, 136, 91, 146, 35, 114, 46, 138, 97, 16, 69, 84, 131, 64, 66, 5, 24, 13, 68, 9, 102, 104, 139, 106, 74, 126, 19, 0, 58, 60, 96, 32, 41, 94, 7, 48, 93, 30, 119, 75, 42, 15, 57, 38, 127, 120, 124, 100, 135, 123, 63, 33, 103, 71, 128, 17, 145, 26, 86, 29, 107, 82, 88, 73, 110, 112, 85, 89, 27, 125, 109, 40, 76, 87, 36, 101, 18, 140]
6	[54, 27, 42, 13, 38, 94, 134, 97, 95, 109, 130, 26, 117, 67, 107, 96, 69, 78, 141, 113, 4, 147, 129, 108, 144, 145, 49, 44, 128, 115, 148, 104, 19, 58, 114, 89, 98, 21, 106, 39, 138, 63, 43, 7, 12, 17, 81, 84, 103, 45, 120, 5, 23, 142, 143, 14, 102, 56, 116, 112, 136, 60, 50, 92, 65, 82, 127, 139, 8, 91, 10, 93, 131, 83, 73, 74, 85, 75, 121, 105, 40, 25, 123, 149, 118, 52, 29, 88, 126, 51, 110, 1, 122, 133, 47, 99, 137, 80, 55, 57, 62, 71, 125, 140, 32, 20, 2, 61, 132, 30, 111, 37, 76, 64, 15, 77, 79, 28, 33, 100, 31, 124, 72, 119, 9, 6, 90, 36, 16, 68, 22, 59, 86, 18, 0, 70, 53, 3, 34, 41, 46, 35, 24, 135, 146, 101, 66, 87, 11, 48]
7	[87, 70, 74, 1, 60, 111, 0, 26, 59, 35, 57, 128, 55, 24, 20, 53, 108, 49, 140, 29, 54, 6, 84, 10, 101, 5, 94, 32, 79, 63, 15, 9, 31, 107, 110, 104, 38, 33, 77, 132, 43, 149, 72, 119, 37, 56, 112, 114, 124, 13, 51, 58, 47, 83, 69, 45, 11, 145, 127, 123, 52, 97, 98, 8, 73, 95, 117, 86, 46, 89, 65, 93, 62, 61, 129, 28, 39, 125, 78, 67, 133, 120, 14, 99, 21, 141, 121, 7, 136, 42, 88, 17, 146, 19, 131, 96, 102, 4, 34, 44, 30, 22, 50, 90, 142, 137, 81, 82, 16, 118, 130, 100, 103, 64, 18, 113, 135, 41, 12, 85, 2, 115, 147, 134, 80, 76, 66, 68, 36, 109, 3, 105, 106, 92, 75, 138, 148, 27, 126, 71, 40, 48, 25, 139, 91, 122, 116, 23, 143, 144]
8	[22, 119, 103, 67, 40, 38, 95, 43, 72, 34, 54, 88, 132, 94, 0, 107, 91, 104, 71, 21, 133, 16, 1, 27, 48, 125, 139, 144, 35, 75, 129, 25, 53, 82, 117, 7, 140, 124, 128, 147, 120, 23, 70, 122, 108, 106, 93, 12, 90, 73, 149, 99, 52, 47, 146, 28, 61, 55, 37, 87, 76, 136, 112, 148, 29, 57, 49, 45, 65, 100, 13, 32, 68, 78, 58, 69, 56, 2, 9, 130, 110, 51, 116, 123, 111, 118, 101, 19, 138, 59, 109, 4, 85, 98, 17, 141, 131, 50, 92, 8, 81, 30, 6, 41, 79, 97, 46, 74, 126, 115, 31, 11, 15, 3, 33, 5, 63, 105, 83, 62, 64, 134, 39, 137, 113, 36, 42, 10, 18, 114, 145, 80, 84, 66, 60, 77, 86, 89, 14, 127, 24, 96, 121, 142, 20, 143, 26, 44, 135, 102]
9	[83, 53, 93, 75, 14, 89, 54, 2, 115, 80, 110, 24, 56, 124, 62, 113, 1, 30, 100, 107, 86, 82, 87, 95, 129, 149, 0, 130, 143, 103, 43, 122, 29, 106, 19, 34, 5, 17, 74, 90, 6, 97, 44, 139, 51, 31, 35, 135, 96, 9, 72, 18, 66, 33, 40, 126, 125, 91, 23, 145, 94, 77, 3, 78, 49, 27, 7, 50, 63, 28, 41, 55, 84, 73, 123, 42, 38, 8, 102, 109, 112, 119, 65, 121, 144, 88, 133, 132, 25, 114, 134, 105, 92, 10, 11, 120, 79, 26, 47, 16, 46, 137, 71, 141, 117, 48, 20, 101, 142, 15, 104, 21, 127, 136, 147, 140, 128, 32, 108, 70, 57, 98, 69, 45, 22, 111, 12, 99, 59, 60, 36, 52, 116, 58, 13, 68, 76, 4, 131, 146, 67, 39, 148, 37, 138, 64, 118, 85, 61, 81]
10*	[149, 148, 147, 146, 145, 144, 143, 142, 141, 140, 139, 138, 137, 136, 135, 134, 133, 132, 131, 130, 129, 128, 127, 126, 125, 124, 123, 122, 121, 120, 119, 118, 117, 116, 115, 114, 113, 112, 111, 110, 109, 108, 107, 106, 105, 104, 103, 102, 101, 100, 99, 98, 97, 96, 95, 94, 93, 92, 91, 90, 89, 88, 87, 86, 85, 84, 83, 82, 81, 80, 79, 78, 77, 76, 75, 74, 73, 72, 71, 70, 69, 68, 67, 66, 65, 64, 63, 62, 61, 60, 59, 58, 57, 56, 55, 54, 53, 52, 51, 50, 49, 48, 47, 46, 45, 44, 43, 42, 41, 40, 39, 38, 37, 36, 35, 34, 33, 32, 31, 30, 29, 28, 27, 26, 25, 24, 23, 22, 21, 20, 19, 18, 17, 16, 15, 14, 13, 12, 11, 10, 9, 8, 7, 6, 5, 4, 3, 2, 1, 0]

Table 10. **Random orders.**

Figure 7. **Qualitative comparisons** between SimCIS and BalConpas [13] on the ADE20K 100-5 continual semantic segmentation.

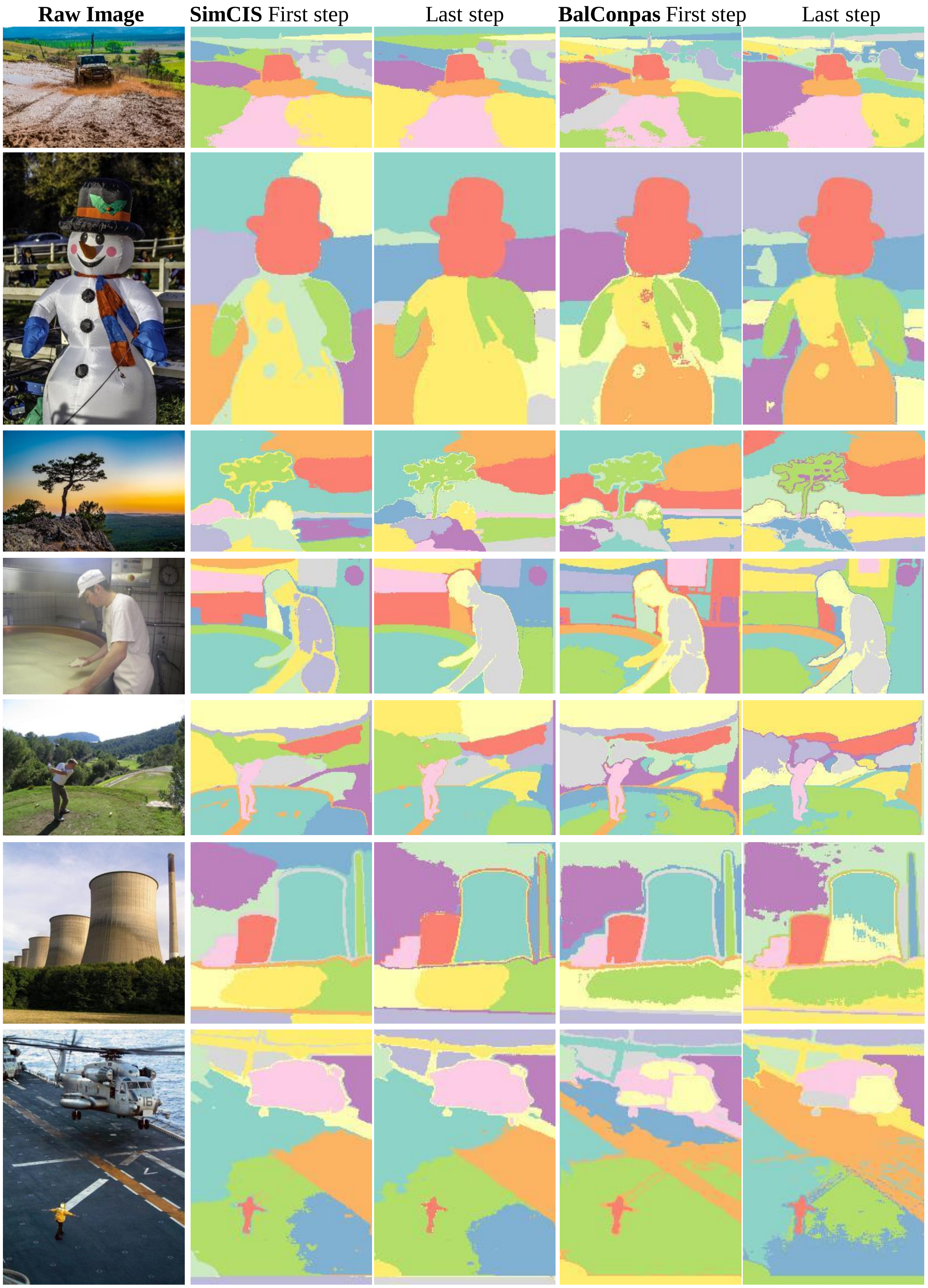

Figure 8. **Clustering results comparison between SimCIS and BalConpas.** Our SimCIS maintains the semantic priors in the pixel feature.