LandCIS: Hierarchical Semantic Anchoring for Concept-Centric Continual Segmentation

Yuyin Ma[1] Yijian Wu[1] Xinyu Wang[2] Yijun Lu[3*]
Zhen Tian[2] Ming Yan[1] Yunni Xia[4]
[1]Xinjiang University, China
[2]University of Glasgow, United Kingdom
[3]Waseda University, Japan
[4]Chongqing University, China

Abstract

In continual segmentation, decoder queries in query-based models serve as object-centric latent tokens, yet their semantic roles often drift across incremental stages, hindering concept reuse and interpretability. We propose ***LandCIS****, a framework that treats decoder queries as explicit* concept carriers *by anchoring them to a hierarchical semantic space spanning low-level visual patterns, mid-level perceptual structure, and high-level scene context. To mitigate concept drift, we introduce cross-stage concept consistency to regularize query–anchor affinity distributions over time, together with augmented semantic memory replay to preserve concept-relevant representations without raw-image rehearsal. On ADE20K, LandCIS improves over SimCIS by +1.4 PQ (all) for continual panoptic segmentation under 100-5 and by +1.4 mIoU for continual semantic segmentation under 100-10, while approaching the joint-training reference. Further analyses of concept purity and embedding structure show that hierarchical anchoring yields more specialized and stable query semantics throughout continual learning.*

1. Introduction

Visual concept discovery aims to learn compact, structured representations that capture reusable semantic units of the visual world [1, 8, 14]. A central question is whether such representations can remain stable and interpretable under non-stationary learning, where new categories are introduced incrementally.

Continual scene understanding [6] provides a natural testbed for this question. It examines whether learned concepts can be retained across stages, whether new classes can be incorporated without disrupting existing representations, and whether concept quality is reflected in dense prediction performance rather than only classification accuracy. Recent work has extended this setting toward continual universal segmentation, suggesting that continual adaptation should preserve not only category-level knowledge but also broader scene representation capacity across segmentation tasks [13].

Query-based segmentation models [4, 5, 10] are particularly well suited to concept-centric modeling because their decoder queries function as object-centric latent tokens. However, existing continual segmentation methods mainly address forgetting at the output level. MiB [2] and PLOP [7] distill logits or features without explicitly modeling query semantics. CoMFormer [3] adapts query initialization but imposes no structural constraints. CoMasTRe [9] reformulates continual segmentation through disentangled objectness learning and class recognition, highlighting the benefits of query-based segmenters with built-in objectness. ECLIPSE [11] improves continual panoptic segmentation through efficient visual prompt tuning. SimCIS [17] regularizes consistency, yet still treats queries as task-specific variables that remain prone to semantic drift. These limitations raise a natural question: *can decoder queries be explicitly structured as visual concepts, improving not only segmentation accuracy but also concept stability and interpretability under continual learning?*

This question is related to structured representation learning. Object-centric methods such as slot attention [14] learn compositional latent tokens, but they assume a fixed set of generic slots and do not address incremental class expansion. Concept bottleneck models [12] improve interpretability through explicit concept layers, yet they are mostly studied in classification rather than dense prediction. In contrast, our method uses hierarchical anchoring to encourage decoder queries to develop stable semantic roles across continual learning stages without requiring concept supervision.

*Corresponding author.

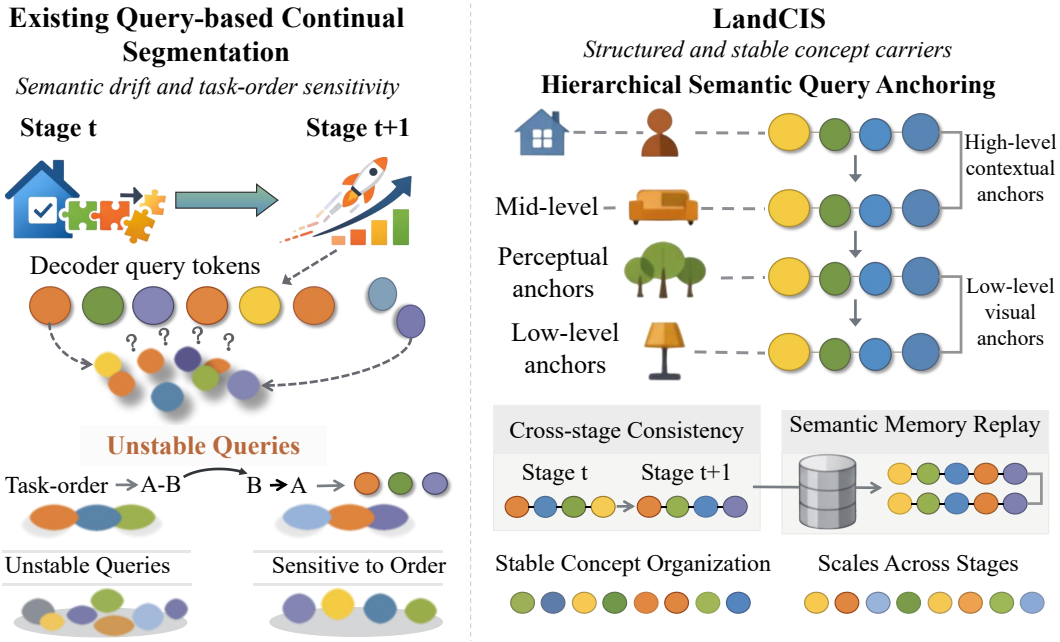

Figure 1. Overview of LandCIS. Visual features initialize hierarchically anchored concept queries. The decoder is regularized by cross-stage concept consistency, while augmented semantic memory replay preserves concept-relevant representations from previous stages.

To this end, we propose **LandCIS**, which organizes decoder queries through *hierarchical semantic anchoring*, stabilizes them via *cross-stage concept consistency*, and preserves concept-relevant information through *augmented semantic memory replay*. Our main contributions are:

- We propose **LandCIS**, a concept-centric continual segmentation framework that reinterprets decoder queries as explicit *concept carriers* and organizes them through hierarchical semantic anchoring across low-level, mid-level, and contextual abstraction levels.
- We introduce two complementary mechanisms to mitigate concept drift during incremental learning: cross-stage concept consistency, which regularizes query semantics in anchor space, and augmented semantic memory replay, which preserves concept-level representations without raw-image rehearsal.
- We broaden evaluation beyond task accuracy by analyzing concept purity, task-order robustness, and embedding structure, showing that LandCIS learns more stable, specialized, and interpretable query representations throughout continual learning.

2. Method

LandCIS reinterprets decoder queries as *concept carriers* that encode reusable visual semantics across incremental stages. The framework consists of three components: hierarchical semantic query anchoring (HSQA), cross-stage concept consistency (CSC), and augmented semantic memory replay (ASMR). An overview is shown in Fig. 1.

2.1. Hierarchical Semantic Query Anchoring

Visual concepts naturally span multiple levels of abstraction. Instead of initializing queries from unconstrained latent tokens, we anchor them to a structured semantic space. At stage t, we maintain hierarchical anchors $\mathcal{H}^t = \{\mathcal{H}_l^t\}_{l=1}^L$ across L levels, corresponding to low-level visual, mid-level perceptual, and high-level contextual semantics. Given backbone features $F \in \mathbb{R}^{H\times W\times D}$, we compute a semantic saliency score at each spatial location:

$$\Phi_{(h,w)} = \sum_{l=1}^{L} \lambda_l \cdot \max_{a_l \in \mathcal{H}_l^t} S_l(F_{(h,w)}, a_l), \tag{1}$$

where $S_l(\cdot)$ is scaled cosine similarity and λ_l is a per-level weighting coefficient. The top-N spatial locations, where N denotes the number of concept queries, are used to initialize the decoder queries:

$$Q = \{F_i \mid i \in \mathrm{TopK}(\{\Phi_{(h,w)}\}, N)\}. \tag{2}$$

The three levels correspond to feature maps at different backbone depths: early layers capture low-level patterns, intermediate layers encode mid-level structure, and deeper layers provide high-level context. At each level l, we apply K-means clustering to the stage-t training features to obtain $\mathcal{H}_l^t$. At the next stage, we freeze existing anchors and append centroids computed from newly introduced classes, yielding $\mathcal{H}_l^{t+1} = \mathcal{H}_l^t \cup \Delta\mathcal{H}_l^{t+1}$. This keeps existing anchors stable while accommodating new semantics.

Thus, queries are grounded in a hierarchical semantic space rather than learned as unconstrained latent variables.

2.2. Cross-Stage Concept Consistency

To promote concept stability across incremental stages, we regularize how anchored query representations evolve over time. Using anchoring indices from stage $t{-}1$, we extract decoder query embeddings $e_i^{(\tau)}$ from both the previous ($\tau{=}t{-}1$) and current ($\tau{=}t$) models for each query i and compare their affinity distributions over the previous anchor set $\mathcal{H}_l^{t-1}$. For each level l:

$$p_l^{(\tau)}(i) = \mathrm{Softmax}\left(\{S_l(e_i^{(\tau)}, a_l)\}\right),\ \tau \in \{t{-}1, t\}. \quad (3)$$

We minimize a symmetrized divergence:

$$\mathcal{L}_{\mathrm{CSC}} = \frac{1}{NL}\sum_{i=1}^{N}\sum_{l=1}^{L}\Bigg(D_{\mathrm{KL}}\left(p_l^{(t-1)}(i)\,\|\,p_l^{(t)}(i)\right) + D_{\mathrm{KL}}\left(p_l^{(t)}(i)\,\|\,p_l^{(t-1)}(i)\right)\Bigg). \quad (4)$$

This loss regularizes concept-level semantic assignments directly, rather than only outputs, thereby reducing drift in previously established query roles.

2.3. Augmented Semantic Memory Replay

Instead of replaying raw images, we build a compact semantic memory that stores concept-relevant representations. For each matched query, we store $m = (z, v_{ctx}, v_{att})$, where z denotes the decoder output, v_{ctx} a contextual descriptor, and v_{att} an attention descriptor. At stage t, sampled memory items are concatenated with current queries and processed jointly. The replay loss reconstructs the stored descriptors, where $\hat{z}_j$ is the current model's decoder output for the j-th replayed item, ϕ_{ctx} and ϕ_{att} are lightweight projection heads, and β balances the two reconstruction terms:

$$\mathcal{L}_{\mathrm{ASMR}} = \frac{1}{J}\sum_{j=1}^{J}\Bigg(\left\|\phi_{\mathrm{ctx}}(\hat{z}_j) - v_{\mathrm{ctx}}^{(j)}\right\|_2^2 + \beta\left\|\phi_{\mathrm{att}}(\hat{z}_j) - v_{\mathrm{att}}^{(j)}\right\|_2^2\Bigg). \quad (5)$$

This strategy preserves concept-relevant representations directly, making replay better aligned with concept retention than raw-image rehearsal.

2.4. Overall Objective

The total loss is $\mathcal{L} = \mathcal{L}_{task} + \alpha \cdot \mathcal{L}_{CSC} + \eta \cdot \mathcal{L}_{ASMR}$, where $\mathcal{L}_{task}$ is the standard Mask2Former objective combining mask and classification losses. The full objective balances task performance with concept structure preservation.

Table 1. Continual Panoptic Segmentation on ADE20K, reported in PQ (%). Best in **bold**, second best underlined.

Method	100-5 (11 tasks)				100-10 (6 tasks)				100-50 (2 tasks)			
	0-100	new	all	avg	0-100	new	all	avg	0-100	new	all	avg
MiB	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	31.1	11.9	24.7	31.3	37.7	23.3	32.9	37.8	42.4	23.7	36.2	39.5
CoMFormer	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	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	41.1	16.6	32.9	-	41.4	18.8	33.9	-	41.7	23.5	35.6	-
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
LandCIS	**43.2**	**23.7**	**36.8**	**40.1**	**43.0**	**32.4**	**39.4**	**41.9**	**45.1**	**33.3**	**41.8**	**44.0**
joint	*40.4*				*40.4*				*40.4*			

Table 2. Continual Semantic Segmentation on ADE20K, 100-10 setting, reported in mIoU (%). Best in **bold**, second best underlined.

Method	1-100	101-150	all	avg
MiB	31.8	14.1	25.9	-
PLOP	40.5	14.1	31.6	36.6
CoMFormer	40.6	15.6	32.3	37.4
BalConpas	47.3	24.2	38.6	43.6
ECLIPSE	43.4	17.4	34.6	-
CoMasTRe	42.3	18.4	34.4	38.4
SimCIS	49.7	27.4	42.3	49.2
LandCIS	**51.0**	**30.6**	**43.7**	**50.8**
joint	*51.2*			

3. Experiments

Setup. We evaluate on ADE20K [15, 16] under continual panoptic segmentation (CPS, reported in PQ) and continual semantic segmentation (CSS, reported in mIoU). We adopt class-incremental settings A-B, where A denotes the number of base classes and B the number of new classes introduced per stage. For CPS, we use Mask2Former with a ResNet-50 backbone; for CSS, we use ResNet-101. The number of concept queries is set to 80. LandCIS extends the SimCIS [17] baseline under the same training protocol.

3.1. Continual Panoptic Segmentation

Table 1 summarizes the CPS results. LandCIS consistently outperforms prior methods across all three settings. The improvements are especially pronounced for newly introduced classes (e.g., +1.8 PQ on 100-5 new and +2.3 PQ on 100-10 new over SimCIS), suggesting that hierarchical anchoring provides stronger initialization for emerging semantics.

3.2. Continual Semantic Segmentation

Table 2 reports CSS results under the 100-10 setting. LandCIS achieves 43.7 mIoU over all classes, improving over SimCIS by +1.4 and moving closer to the joint-training upper bound of 51.2. The gain on new classes (+3.2) is again larger than that on base classes (+1.3), consistent with stronger concept adaptation.

3.3. Component Analysis

Table 3 shows that each component contributes positively. HSQA improves concept initialization, CSC stabilizes se-

Table 3. Ablation of core components on ADE20K CPS 100-5, reported in PQ (%).

Config	HSQA	CSC	ASMR	0-100	new	all	avg
Baseline				42.1	21.9	35.4	38.7
+HSQA	✓			42.5	22.8	35.8	39.2
+CSC	✓	✓		42.9	23.2	36.2	39.6
Full	✓	✓	✓	**43.2**	**23.7**	**36.8**	**40.1**

Table 4. Hyperparameter sensitivity on ADE20K CPS 100-5, reported in PQ (%) for all classes.

α	PQ	η	PQ	L	PQ
0.01	36.1	0.1	36.3	1	36.0
0.1	36.5	0.5	36.6	2	36.4
0.5	**36.8**	**1.0**	**36.8**	**3**	**36.8**
1.0	36.6	2.0	36.5	4	36.7
2.0	36.0	5.0	35.9	5	36.5

Table 5. Average Concept Purity on ADE20K CPS 100-10. Higher indicates more specialized query roles.

Method	Intermediate	Final
SimCIS	0.58	0.52
LandCIS	**0.67**	**0.63**

mantic assignments across stages, and ASMR preserves concept-level information from earlier stages.

Hyperparameter Sensitivity. We vary α (CSC), η (ASMR), and the number of hierarchy levels L on CPS 100-5 (Table 4). Performance peaks at α=0.5, η=1.0, and L=3. Smaller values weaken regularization, while larger ones limit plasticity. Using a single level (L=1) reduces performance by 0.8 PQ, confirming the benefit of multi-level anchoring. Increasing L beyond 3 brings no further gain.

3.4. Concept Quality Analysis

Beyond task accuracy, we examine whether the proposed framework yields higher-quality concept representations.

Concept Purity. We define concept purity to measure whether each query develops a stable semantic role. For each query index q_i, we collect its matched class distribution $P(c|q_i)$ over the validation set and compute:

$$\text{Purity}(q_i) = 1 - \frac{H(P(c|q_i))}{\log C}, \tag{6}$$

where $H(\cdot)$ denotes Shannon entropy and C is the number of seen classes. Higher purity indicates that a query is more consistently associated with the same semantic concept. Table 5 reports the average purity at both an intermediate stage and the final stage. LandCIS achieves substantially higher purity in both cases, confirming that hierarchical anchoring and concept-level regularization encourage queries to develop more stable and specialized semantic roles.

Robustness to Task Order. We evaluate three random orderings of the 100-10 CPS protocol (Table 6). LandCIS exhibits lower variance across orderings, indicating that structured concept representations reduce sensitivity to the order in which concepts are introduced.

Table 6. Robustness to three random task orderings on ADE20K CPS 100-10, reported in PQ (%) for all classes. Best in **bold**, second best underlined.

Method	PQ (all) ↑
BalConpas	34.7 ± 0.18
SimCIS	38.1 ± 0.22
LandCIS	$\mathbf{39.4 \pm 0.15}$

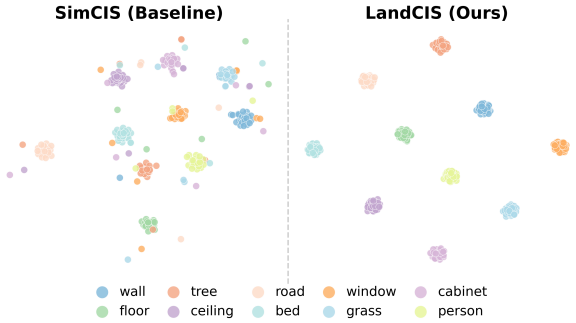

Figure 2. t-SNE visualization of query embeddings at the final stage of CPS 100-10. **Left:** SimCIS. **Right:** LandCIS. Each color represents the dominant matched class of a query. LandCIS shows tighter, better-separated concept clusters.

Query Embedding Visualization. Figure 2 visualizes decoder query embeddings at the final stage using t-SNE, with points colored by their dominant matched class. Compared with SimCIS, LandCIS produces tighter and more clearly separated clusters, providing visual evidence that hierarchical anchoring encourages distinct and stable concept roles.

4. Conclusion

We presented LandCIS, a framework that models decoder queries as structured concept carriers for continual segmentation. By combining hierarchical anchoring, cross-stage consistency, and semantic memory replay, LandCIS improves both segmentation performance and concept stability. Further analyses reveal more specialized query roles and stronger robustness, supporting continual learning as a meaningful setting for studying concept formation and reuse.

Limitations. Our study focuses on continual segmentation and does not yet examine whether the learned concepts transfer to other visual tasks such as detection or image generation. Moreover, although the hierarchy levels are grounded in backbone depth, the semantic meaning of individual anchors within each level emerges from unsupervised clustering rather than explicit supervision, which limits direct interpretability. Richer evaluations of concept compositionality and cross-modal grounding with language remain important directions for future work.

References

[1] David Bau, Bolei Zhou, Aditya Khosla, Aude Oliva, and Antonio Torralba. Network dissection: Quantifying interpretability of deep visual representations. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6541–6549, 2017. 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. 1

[3] 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. 1

[4] Bowen Cheng, Anwesa Choudhuri, Ishan Misra, Alexander Kirillov, Rohit Girdhar, and Alexander G Schwing. Mask2former for video instance segmentation. *arXiv preprint arXiv:2112.10764*, 2021. 1

[5] 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

[6] Matthias De Lange, Rahaf Aljundi, Marc Masana, Sarah Parisot, Xu Jia, Aleš Leonardis, Gregory Slabaugh, and Tinne Tuytelaars. A continual learning survey: Defying forgetting in classification tasks. *IEEE transactions on pattern analysis and machine intelligence*, 44(7):3366–3385, 2021. 1

[7] 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. 1

[8] Amirata Ghorbani, James Wexler, James Y Zou, and Been Kim. Towards automatic concept-based explanations. *Advances in neural information processing systems*, 32, 2019. 1

[9] 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

[10] 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

[11] 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*, pages 3346–3356, 2024. 1

[12] Pang Wei Koh, Thao Nguyen, Yew Siang Tang, Stephen Mussmann, Emma Pierson, Been Kim, and Percy Liang. Concept bottleneck models. In *International conference on machine learning*, pages 5338–5348. PMLR, 2020. 1

[13] Zihan Lin, Zilei Wang, and Xu Wang. Towards continual universal segmentation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 29417–29427, 2025. 1

[14] Francesco Locatello, Dirk Weissenborn, Thomas Unterthiner, Aravindh Mahendran, Georg Heigold, Jakob Uszkoreit, Alexey Dosovitskiy, and Thomas Kipf. Object-centric learning with slot attention. *Advances in neural information processing systems*, 33:11525–11538, 2020. 1

[15] 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. 3

[16] Bolei Zhou, Hang Zhao, Xavier Puig, Tete Xiao, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Semantic understanding of scenes through the ade20k dataset. *International journal of computer vision*, 127(3):302–321, 2019. 3

[17] Yuchen Zhu, Cheng Shi, Dingyou Wang, Jiajin Tang, Zhengxuan Wei, Yu Wu, Guanbin Li, and Sibei Yang. Rethinking query-based transformer for continual image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4595–4606, 2025. 1, 3