Continual Segmentation with Disentangled Objectness Learning and Class Recognition

Supplementary Material

This supplementary material first presents a detailed workflow of the class recognition process. Next, it provides more comparisons with state-of-the-art methods. Finally, more ablation studies and detailed results are reported.

A. Workflow of Stage 2 Class Recognition

In detail, we present the class decoder architecture in Fig. A1. It is a single Transformer decoder block whose key (K) and value (V) are pixel embedding; query (Q) is positional embedding with task query. Note that only one positional embedding (red), together with one task query (blue), is fed through the class decoder once a time. The task embedding (purple) is the corresponding output of the task query, and the other embedding (gray) is discarded.

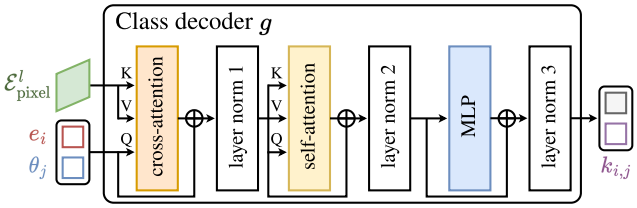

Figure A1. Class decoder architecture.

We also illustrate the stage 2 class recognition process in Algorithm A1. The class decoding process requires positional embeddings, task queries, and pixel embeddings as input and outputs the class probability of all proposals.

B. More Comparisons with State-of-the-arts

We compare CoMasTRe with state-of-the-art continual segmentation methods. Here, we additionally report quantitative results in ADE20K *50-50* setting and qualitative results of PASCAL VOC *15-1* and ADE20K *100-10*.

Quantitative results in ADE20K *50-50* setting. As shown in Tab. A1, we report benchmark results in ADE20K [8] *50-50* setting. The results show that we achieve a new state-of-the-art in this setting with 0.32 percent point (*p.p*) leap on *all* metric compared with CoMFormer [2]. Specifically, our method reaches a competitive new class performance while maintaining better old knowledge (0.63 *p.p* gain compared with CoMFormer). Furthermore, CoMasTRe performs better across all learning steps, surpassing CoMFormer by 3.12 *p.p* and even can beat the highest per-pixel method PLOP [3] by 0.30 *p.p* on *avg*.

Qualitative results on PASCAL VOC 2012. Compared with CoMFormer [2], we visualize the segmentation results under PASCAL VOC *15-1* in Fig. A2. By comparison, our method is more resistant to forgetting old similar classes. For example, CoMFormer starts to forget the *horse* after learning *sheep* class at step 2 (first row) and misrecognizes the *dog* as a *sheep* at step 4 (third row). Additionally, our method is less prone to overconfidence on new classes, *e.g.*, CoMFormer falsely recognizes a *sofa* at step 5 (first row), but our method does not.

Qualitative results on ADE20K. As shown in Fig. A3, we visualize the results in ADE20K *100-10* settings. The visualization suggests that our method preserves better knowledge of previous classes. Compared with CoMFormer [2], our method correctly segments the *field* in the first row, the

Algorithm A1 Class recognition at step t

Input: $\mathcal{E}^{\text{m}}_{\text{pos}} = \{e_1, \dots, e_M\}$: matched positional embeds
$\mathcal{Q}^t_{\text{task}} = \{\theta_1, \dots, \theta_t\}$: task queries at step t
$\Phi^t = \{\phi_1, \dots, \phi_t\}$: classifiers at step t
$\mathcal{E}^l_{\text{pixel}}$: corresponding pixel embeddings
g: class decoder
Output: $\mathcal{P}^t$: class probability at step t

for $i \leftarrow 1, \dots, M$ **do**
  for $j \leftarrow 1, \dots, t$ **do**
    $\mathcal{Q}^{i,j}_{\text{cls}} \leftarrow (e_i, \theta_j)$
    $k_{i,j} \leftarrow g(\mathcal{Q}^{i,j}_{\text{cls}}, \mathcal{E}^l_{\text{pixel}})$
    $z_{i,j} \leftarrow \phi_j(k_{i,j})$
  end for
  $z_i \leftarrow [z_{i,1}, \dots, z_{i,t}]$
  $p_i \leftarrow \text{sigmoid}(z_i)$
end for
$\mathcal{P}^t \leftarrow \{p_1, \dots, p_M\}$

Table A1. Benchmark results in ADE20K *50-50* setting. The 1st and 2nd highest results are marked in **bold** and underline.

Paradigm	Method	**50-50** (3 tasks) *1-50*	*51-150*	*all*	*avg*
Per-Pixel	MiB [1]	45.57	21.01	29.31	38.98
	SDR [5]	45.66	18.76	27.85	34.25
	PLOP [3]	48.83	20.99	30.40	<u>39.42</u>
	REMINDER [6]	47.11	20.35	29.39	39.26
	RCIL [7]	48.30	25.00	32.50	—
	Joint	51.21	32.77	39.00	—
Query	CoMFormer [2]	<u>49.20</u>	**26.60**	<u>34.10</u>	36.60
	Joint	53.40	38.00	43.10	—
	CoMasTRe (ours)	**49.83**	<u>26.56</u>	**34.42**	**39.72**
	Joint	54.09	39.49	44.36	—

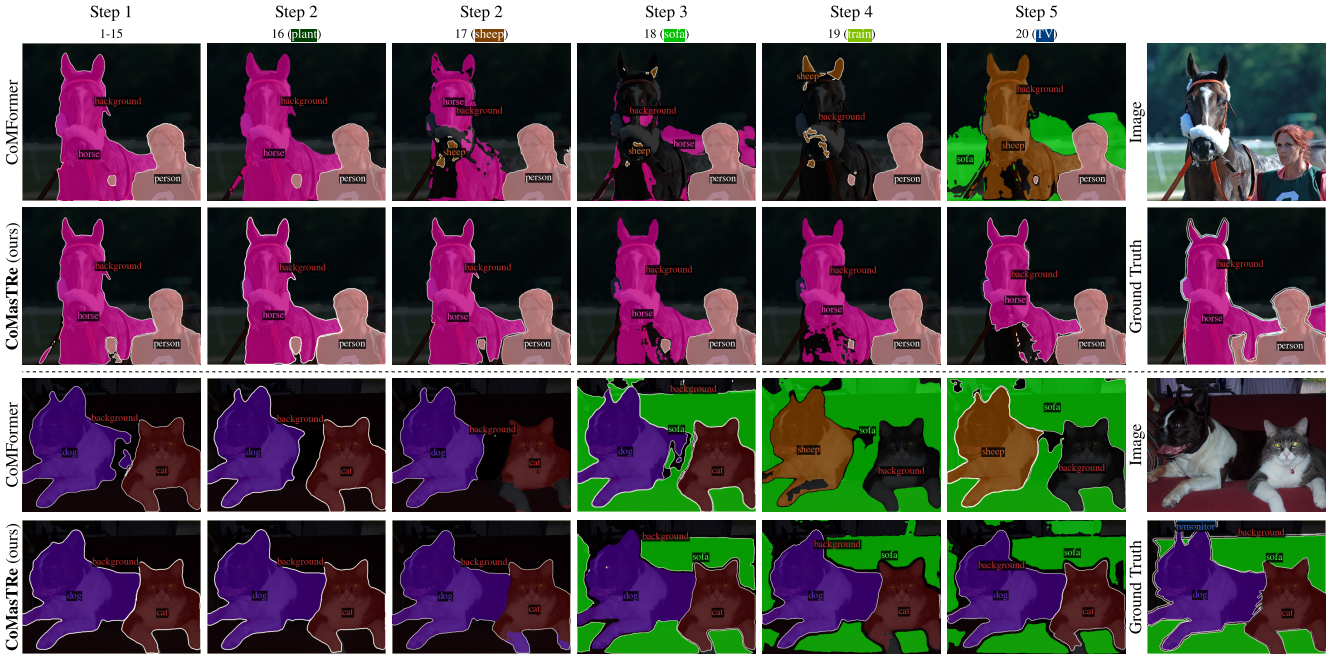

Figure A2. Qualitative results compared with CoMFormer [2] in PASCAL VOC *15-1* setting.

Table A2. Average recall (AR) of mask proposals in PASCAL VOC *15-1* setting, where $\mathcal{L}_{\text{mask-kd}}$ for mask distillation, $\mathcal{L}_{\text{os-kd}}$ for objectness score distillation, $\mathcal{L}_{\text{pe-kd}}$ for position distillation, s for objectness score, and α for high objectness threshold.

#	Case	AR	
		$s > 0$	$s > \alpha$
1	$\mathcal{L}_{\text{mask-kd}}$	55.89	13.75
2	$\mathcal{L}_{\text{os-kd}}$	50.92	47.84
3	$\mathcal{L}_{\text{mask-kd}} + \mathcal{L}_{\text{os-kd}}$	<u>55.63</u>	<u>49.02</u>
4	$\mathcal{L}_{\text{mask-kd}} + \mathcal{L}_{\text{os-kd}} + \mathcal{L}_{\text{pe-kd}}$	<u>55.84</u>	<u>48.83</u>

transporter in the second row, the *mirror* in the third row, the *mountain* in the fifth row, and the *house* in the sixth row. In addition, CoMasTRe also proposes better masks for old classes. For example, our method maintains the knowledge of the *door* in the fourth row (yellow box) and the *boat* in the sixth row (red boxes).

C. More Ablation Studies and Detailed Results

We present more ablation studies on the forgetting of objectness and the effectiveness of objectness score reweighting. In addition, per-class segmentation results on PASCAL VOC are reported.

Objectness forgetting analysis. We analyze the forgetting of objectness by ablating objectness distillation components. The ablation is conducted in the same cases as in Tab. 5 in the main text. For each case, we use average recall (AR) to indicate the performance of mask proposals. Here, s stands for the objectness score, and α is the objectness threshold during inference. As shown in Tab. A2, we report the performance in PASCAL VOC [4] *15-1* setting. By comparing case 1 and 3, without $\mathcal{L}_{\text{os-kd}}$, AR ($s > 0$) remains unchanged, but AR ($s > \alpha$) diminishes (-35.27 *p.p* AR), which means the objectness scores fail to indicate old class objectness. By comparing case 2 and 3, we observe slight forgetting of mask proposals without $\mathcal{L}_{\text{mask-kd}}$ (-4.71 *p.p* AR), showing the forgetting robustness of mask proposals. When comparing case 3 and 4, we find position distillation contributes most to continual classification (see Tab. 5 in the main text), as the AR changes are negligible (underlined in Tab. A2).

Table A3. Ablation results of objectness score reweighting in PASCAL VOC *15-1* setting by varying reweighting strength β. Note that no reweight is applied when $\beta = 0.0$.

#	β	*1-15*	*16-20*	*all*	*avg*
1	0.0	67.47	41.83	61.37	68.47
2	1.0	**69.79**	42.98	63.41	70.35
3	2.0	69.77	**43.62**	**63.54**	**70.63**

Effectiveness of objectness score reweighting. In Tab. A3, we ablate the effectiveness of objectness score reweighting mentioned in Sec. 3.3.1 of the main text. The reweight strength β is set to 0.0, 1.0, and 2.0, respectively. Please note that when β is set to 0.0, the distillation is equivalent

Figure A3. Qualitative results compared with CoMFormer [2] in ADE20K *100-10* setting.

Table A4. Per-class segmentation results on PASCAL VOC 2012 in mIoU (%). Incremented classes (*inc*) are marked in green.

Setting	bg	aero	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	mbike	person	plant	sheep	sofa	train	tv	*base*	*inc*	*all*
19-1 (2 steps)	93.7	91.9	42.8	88.9	65.4	81.1	87.9	80.7	91.9	37.0	80.1	51.3	86.0	82.9	82.0	87.0	56.6	87.3	50.6	77.4	69.5	75.1	69.5	74.9
15-5 (2 steps)	93.5	90.4	43.6	90.8	64.2	82.8	88.2	88.6	94.1	42.8	80.9	70.5	89.4	82.7	84.6	88.4	39.4	59.1	38.9	62.0	60.2	79.7	51.9	73.1
15-1 (6 steps)	88.9	86.4	38.0	82.6	53.2	76.8	76.8	83.2	82.5	36.2	59.3	48.4	80.4	63.1	74.7	85.8	29.9	47.1	33.0	55.4	52.7	69.8	43.6	63.5
Joint	94.3	91.8	43.0	91.1	65.3	85.2	92.2	87.2	93.1	44.2	85.3	69.4	90.5	86.9	85.3	88.4	60.7	83.9	48.1	85.4	68.3	—	—	78.1

to the regular unweighted one. The results show that the reweighting ($\beta = 2.0$) leads to 2.17 *p.p* performance gain on *all* metric compared with the unweighted version.

Per-class results on PASCAL VOC. In Tab. A4, we provide per-class experimental results on PASCAL VOC 2012 in different continual segmentation settings. The results indicate that in addition to learning new classes, the old class performance can be further strengthened in later learning steps compared with *joint* baseline, such as *sheep* class in *19-1* (+3.4 *p.p*) and *car* class in *15-5* (+1.4 *p.p*).

References

[1] Fabio Cermelli, Massimiliano Mancini, Samuel Rota Bulo, Elisa Ricci, and Barbara Caputo. Modeling the Background for Incremental Learning in Semantic Segmentation. In *CVPR*, 2020. 1

[2] Fabio Cermelli, Matthieu Cord, and Arthur Douillard. CoMFormer: Continual Learning in Semantic and Panoptic Segmentation. In *CVPR*, 2023. 1, 2, 3

[3] Arthur Douillard, Yifu Chen, Arnaud Dapogny, and Matthieu Cord. PLOP: Learning Without Forgetting for Continual Semantic Segmentation. In *CVPR*, 2021. 1

[4] Mark Everingham, Luc Van Gool, Christopher K. I. Williams, John Winn, and Andrew Zisserman. The Pascal Visual Object Classes (VOC) Challenge. *IJCV*, 88(2):303–338, 2010. 2

[5] Umberto Michieli and Pietro Zanuttigh. Continual Semantic Segmentation via Repulsion-Attraction of Sparse and Disentangled Latent Representations. In *CVPR*, 2021. 1

[6] Minh Hieu Phan, The-Anh Ta, Son Lam Phung, Long Tran-Thanh, and Abdesselam Bouzerdoum. Class Similarity Weighted Knowledge Distillation for Continual Semantic Segmentation. In *CVPR*, 2022. 1

[7] Chang-Bin Zhang, Jia-Wen Xiao, Xialei Liu, Ying-Cong Chen, and Ming-Ming Cheng. Representation Compensation Networks for Continual Semantic Segmentation. In *CVPR*, 2022. 1

[8] Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene Parsing Through ADE20K Dataset. In *CVPR*, 2017. 1