CoMFormer: Continual Learning in Semantic and Panoptic Segmentation

Supplementary Material

Anonymous CVPR submission

Paper ID 7744

1. Additional quantitative results

Impact of $\lambda_D$. To demonstrate the robustness of our approach to the hyper-parameter choice, we report an ablation study on the impact of $\lambda_D$ in Tab. 1. Setting $\lambda_D$ between 5 and 50 maintains stable results on both CSS and CPS while obtaining different trade-offs between new and old classes. In particular, when setting $\lambda_D = 5$, we achieve higher results on the novel classes at the cost of losing performance on the old ones. Differently, setting $\lambda_D = 50$ increases the regularization and reduces forgetting, improving the performance of old classes while decreasing it on the new classes. Setting $\lambda_D = 10$ achieves the best trade-off between learning and forgetting on both semantic and panoptic segmentation. In the paper, we reported results for $\lambda_D = 10$.

50-50 in Continual Panoptic Segmentation. In Tab. 2 we report additional experiments on Continual Panoptic Segmentation on the 50-50 setting where we perform three tasks of 50 classes. CoMFormer outperforms all the baselines, obtaining the best results on both old and new classes. In particular, we can see that it exceeds the best competitor, PLOP, by 0.5 PQ in the old classes and 0.2 PQ in the new ones. When comparing with MiB, however, we can see that the gap is more relevant: +11.6 PQ on old classes and +10.2 on the new ones. Finally, we can see that CoMFormer obtains a small performance gap with the Joint baselines, which is more relevant for the new classes (-7.6 PQ).

50-50 in Continual Semantic Segmentation. Tab. 3 reports the additional results on the Continual Semantic Segmentation benchmark on the 50-50 setting in mIoU, comparing CoMFormer with previous works based on DeepLab [2] and our re-implementation based on the CoMFormer architecture, both in Per-Pixel and Mask-Based fashion. We observe that CoMFormer achieves a new state of the art. In particular, when comparing it with previous works, we can see that it outperforms the best baseline (RCIL) on both old (+0.9 mIoU) and new classes (+1.6 mIoU), for an overall improvement of 1.6 mIoU. Furthermore, CoMFormer also outperforms the baselines implemented on the same architecture: w.r.t. to Per-Pixel baselines, there is a relevant performance gap, especially regarding the old classes (CoMFormer 49.2 vs MiB 44.9 mIoU). Considering the Mask-Based baselines, CoMFormer shows the best performance, improving PLOP by 1.1 mIoU on the old classes and by 0.3 mIoU on all.

| $\lambda_D$ | 1-100 | 101-150 | All | 1-100 | 101-150 | All |
|-------------|-------|---------|-----|-------|---------|-----|
| 5           | 38.8  | 17.5    | 31.7| 35.3  | 18.2    | 29.6|
| 10          | 40.6  | 15.6    | 32.3| 36.0  | 17.1    | 29.7|
| 50          | 41.0  | 13.8    | 31.9| 38.4  | 12.6    | 29.8|

Table 1. Impact of $\lambda_D$ in CPS and CSS in ADE20K 100-10.

| Method   | 1-50 | 51-150 | avg | all |
|----------|------|--------|-----|-----|
| FT       | 0.0  | 14.3   | 23.1| 9.5 |
| MiB      | 33.6 | 16.3   | 31.8| 22.1|
| PLOP     | 44.7 | 26.3   | 37.9| 32.4|
| CoMFormer| 45.2 | 26.5   | 37.9| 32.7|
| Joint    | 50.2 | 34.1   | —   | 39.5|

Table 2. Continual Panoptic Segmentation results on ADE20K dataset on 50-50 setting in PQ.

| Architecture | Method | 1-50 | 51-150 | avg | all |
|--------------|--------|------|--------|-----|-----|
| DeepLab-v3   | MiB [1]| 45.3 | 21.6   | 38.9| 29.3|
|              | PLOP   | 48.6 | 21.6   | 39.4| 30.4|
|              | RCIL [4]| 48.3 | 25.0   | —   | 32.5|
| Per-Pixel    | MiB    | 44.9 | 25.4   | 35.0| 31.9|
|              | PLOP   | 43.2 | 24.7   | 34.6| 30.9|
| Mask-Based   | MiB    | 24.6 | 19.4   | 25.8| 21.1|
|              | PLOP   | 48.1 | 26.6   | 36.5| 33.8|
|              | CoMFormer| 49.2 | 26.6   | 36.6| 34.1|
|              | Joint  | 53.4 | 38.0   | —   | 43.1|

Table 3. Continual Semantic Segmentation results on ADE20K dataset on 50-50 setting in mIoU.
Figure 1. **Qualitative results** of CoMFormer v.s. MiB and PLOP on the 100-50 continual panoptic segmentation setting on ADE20K.

Figure 2. **Qualitative results** of CoMFormer v.s. MiB and PLOP on the 100-5 continual panoptic segmentation setting on ADE20K.
Figure 3. Qualitative results of CoMFormer v.s. MiB and PLOP on multiple settings of the continual semantic segmentation benchmark on ADE20K.

2. Additional qualitative results

Continual Panoptic Segmentation. Fig. 1 and Fig. 2 report additional qualitative results on, respectively, the 100-50 and 100-5 settings in continual panoptic segmentation, comparing CoMFormer with MiB and PLOP using images randomly sampled from the validation set. Considering the 100-50, we can see that PLOP and CoMFormer achieve visually similar results, while MiB struggles in segmenting every image object (for example, the clock in the third row). Differently, on the 100-5, CoMFormer visually outperforms the other baselines being able to correctly segment all the objects in the image (e.g. the tent in the first row, the table in the third row, and the rug in the fourth row). However, we note a common error across all the methods: some classes are correctly segmented but misclassified (e.g. grass instead of earth in the first row and window instead of door in the third). This error is less present in CoMFormer w.r.t. PLOP.
and MiB, as can be seen from the chandelier in the last row and the armchair in the third.

**Continual Semantic Segmentation.** Fig. 3 reports the qualitative results for the 100-50, 100-10, and 100-5 settings of the continual semantic segmentation benchmark comparing CoMFormer with MiB and PLOP on images randomly sampled from the validation set. Considering the 100-50 setting, MiB is far worse than other baselines: it is not able to correctly segment the object in the image, achieving low performance. PLOP and CoMFormer achieve similar results, being able to segment all the objects in the images. Differently, on the 100-10 setting, the difference among methods becomes more evident: considering the second row, CoMFormer correctly segments the rock and the wall, while misclassifying the road with earth. However, both PLOP and MiB are not able to segment the image: the former is not able to report any segment in that area, while the latter segments incorrectly the area as mount. Finally, considering the 100-5 setting, we note that MiB achieves poor performance on both images, being unable to finely segment the image pixels. Comparing CoMFormer with PLOP, our model CoMFormer is able to segment more classes (e.g. the truck in the fifth row and the bag in the last row), obtaining better performances. Overall, the qualitative results confirm the quantitative findings, where CoMFormer outperforms the other methods, especially considering settings where multiple learning steps are performed. Those longer continual settings are more realistic and allows us to benchmark more efficiently what a truly lifelong learning agent should be.

### 3. Class Ordering

In Tab. 4 we report the class ordering of ADE20K that we used for all the reported experiments, following the previous benchmarks [1, 3]. Considering the 100-50, 100-10, and 100-5 settings, reported in the main paper, we note that 44 of the new classes are “things”, while the other 6 are “stuff”. While there is no difference between “things” and “stuff” in semantic segmentation, it is especially relevant in the panoptic segmentation task, where the goal is to separate in different segments multiple instances of the “things” classes, since it introduces additional challenges.

**References**

[1] Fabio Cermelli, Massimiliano Mancini, Samuel Rota Bulò, Elisa Ricci, and Barbara Caputo. Modeling the background for incremental learning in semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. (pages 1, 4).

[2] Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. Rethinking atrous convolution for semantic image segmentation. In *arXiv preprint library*, 2017. (page 1).

[3] Arthur Douillard, Yifu Chen, Arnaud Dapogny, and Matthieu Cord. Plop: Learning without forgetting for continual semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. (pages 1, 4).

[4] 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. (page 1).

[5] 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 (CVPR)*, 2017. (page 5).
| Idx | Name       | Thing | Idx | Name       | Thing | Idx | Name       | Thing |
|-----|------------|-------|-----|------------|-------|-----|------------|-------|
| 1   | wall       | ✓     | 51  | refrigerator✓ |       | 101 | poster✓   |       |
| 2   | building   |       | 52  | grandstand✓ |       | 102 | stage✓    |       |
| 3   | sky        | ✓     | 53  | path       | ✓     | 103 | van       | ✓     |
| 4   | floor      | ✓     | 54  | stairs✓    |       | 104 | ship✓     |       |
| 5   | tree       | ✓     | 55  | runway✓    |       | 105 | fountain✓ |       |
| 6   | ceiling    | ✓     | 56  | case✓      |       | 106 | conveyer✓ |       |
| 7   | road       |       | 57  | pool       | ✓     | 107 | canopy✓   |       |
| 8   | bed        | ✓     | 58  | pillow✓    |       | 108 | washer✓   |       |
| 9   | windowpane | ✓     | 59  | screen✓    |       | 109 | plaything✓|       |
| 10  | grass      | ✓     | 60  | stairway✓  |       | 110 | swimming✓ |       |
| 11  | cabinet    | ✓     | 61  | river✓     |       | 111 | stool✓    |       |
| 12  | sidewalk   |       | 62  | bridge✓    |       | 112 | barrel✓   |       |
| 13  | person     | ✓     | 63  | bookcase✓  |       | 113 | basket✓  |       |
| 14  | earth      | ✓     | 64  | blind✓     |       | 114 | waterfall✓|       |
| 15  | door       | ✓     | 65  | coffee✓    |       | 115 | tent✓    |       |
| 16  | table      | ✓     | 66  | toilet✓    |       | 116 | bag✓     |       |
| 17  | mountain   |       | 67  | flower✓    |       | 117 | minibike✓|       |
| 18  | plant      | ✓     | 68  | book✓      |       | 118 | cradle✓  |       |
| 19  | curtain    | ✓     | 69  | hill✓      |       | 119 | oven✓    |       |
| 20  | chair      | ✓     | 70  | bench✓     |       | 120 | ball✓    |       |
| 21  | car        | ✓     | 71  | countertop✓|       | 121 | food✓    |       |
| 22  | water      | ✓     | 72  | stove✓     |       | 122 | step✓    |       |
| 23  | painting   | ✓     | 73  | palm✓      |       | 123 | tank✓    |       |
| 24  | sofa       | ✓     | 74  | kitchen✓  |       | 124 | trade✓   |       |
| 25  | shelf      | ✓     | 75  | computer✓ |       | 125 | microwave✓|       |
| 26  | house      | ✓     | 76  | swivel✓   |       | 126 | pot✓     |       |
| 27  | sea        |       | 77  | boat✓     |       | 127 | animal✓ |       |
| 28  | mirror     | ✓     | 78  | bar✓      |       | 128 | bicycle✓ |       |
| 29  | rug        | ✓     | 79  | arcade✓   |       | 129 | lake✓    |       |
| 30  | field      | ✓     | 80  | hovel✓    |       | 130 | dishwasher✓|       |
| 31  | armchair   | ✓     | 81  | bus✓      |       | 131 | screen✓ |       |
| 32  | seat       | ✓     | 82  | towel✓    |       | 132 | blanket✓|       |
| 33  | fence      | ✓     | 83  | light✓    |       | 133 | sculpture✓|       |
| 34  | desk       | ✓     | 84  | truck✓   |       | 134 | hood✓    |       |
| 35  | rock       | ✓     | 85  | tower✓   |       | 135 | scone✓   |       |
| 36  | wardrobe   | ✓     | 86  | chandelier✓|       | 136 | vase✓    |       |
| 37  | lamp       | ✓     | 87  | awning✓  |       | 137 | traffic✓ |       |
| 38  | bathtub    | ✓     | 88  | streetlight✓|       | 138 | tray✓    |       |
| 39  | railing     | ✓     | 89  | booth✓   |       | 139 | ashcan✓ |       |
| 40  | cushion    | ✓     | 90  | television✓|       | 140 | fan✓    |       |
| 41  | base       | ✓     | 91  | airplane✓ |       | 141 | pier✓    |       |
| 42  | box        | ✓     | 92  | dirt✓    |       | 142 | crt✓    |       |
| 43  | column     | ✓     | 93  | apparel✓ |       | 143 | plate✓ |       |
| 44  | signboard  | ✓     | 94  | pole✓    |       | 144 | monitor✓ |       |
| 45  | chest      | ✓     | 95  | land✓    |       | 145 | bulletin✓|       |
| 46  | counter    | ✓     | 96  | bannister✓|       | 146 | shower✓ |       |
| 47  | sand       | ✓     | 97  | escalator✓|       | 147 | radiatoer✓|       |
| 48  | sink       | ✓     | 98  | ottoman✓ |       | 148 | glass✓ |       |
| 49  | skyscraper | ✓     | 99  | bottle✓ |       | 149 | clock✓ |       |
| 50  | fireplace  | ✓     | 100 | buffet✓ |       | 150 | flag✓ |       |

Table 4. Class ordering of ADE20K [5] used in all reported experiments.