Label-Guided Knowledge Distillation for Continual Semantic Segmentation on 2D Images and 3D Point Clouds—Supplementary Material

A. Introduction

In the supplementary material, we provide more analysis, details and visualization results, which can be summarized as:

- We analyze why the standard KD suffers from the novel-background confusion.
- We demonstrate the difficulty of our proposed ScanNet benchmark for CSS with the class frequency statistics.
- We present qualitative visualization results to show the superiority of our method against the other competing approaches.
- We provide more implementation details to ensure reproducibility.

B. Analysis and Discussion

B.1. Standard knowledge distillation

Despite its great success in the context of image classification [3, 7], the standard distillation loss indeed has a critical drawback when a special class *background* gets involved, *e.g.*, object detection [12], segmentation [8]. To recap, the standard knowledge distillation loss ℓ_{kd} adopted in CSS by ILT [8] can be formulated as:

$$\ell_{\text{kd}}^{\theta^t}(x,y) = -\frac{1}{|\mathcal{I}|}\sum_{i\in\mathcal{I}}\sum_{c\in\mathcal{C}^{0:t-1}} q_x^{t-1}(i,c)\log\hat{q}_x^t(i,c)\,, \quad \text{(A)}$$

where $\hat{q}_x^t(i,c)$ refers to the probability of class c for element i predicted by f_{θ^t} but re-normalized over all old classes:

$$\hat{q}_x^t(i,c) = q_x^t(i,c)/\sum\nolimits_{k\in\mathcal{C}^{0:t-1}} q_x^t(i,k)\,. \quad \text{(B)}$$

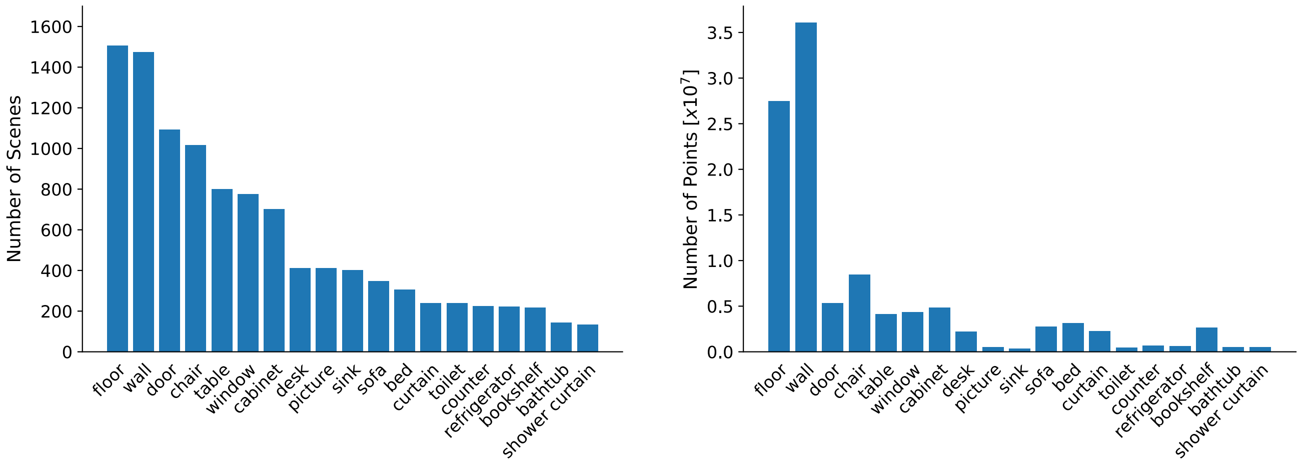
The above formulation completely ignores the semantic shift of the background class across different incremental steps: an element x_i assigned with background label $\mathtt{b}$ at the last step $t-1$ might become a new (novel) class $y_i \in \mathcal{C}^t\setminus\mathtt{b}$ at the current step t. For such new class elements, the old model in fact outputs a high background score as they are masked as the background at the last step. Through knowledge distillation, the new model is encouraged to similarly predict a high background score for these, however, new class elements $\{x_i|y_i\in\mathcal{C}^t\setminus\mathtt{b}\}$. Consequently, it will mislead the new model to wrongly classify the novel class elements into the background, which hinders the learning of novel classes and causes novel-background confusion.

C. More Statistics for the ScanNet Benchmark

For a better view of ScanNet, we provide the class frequency statistics of the whole dataset, including both train and validation sets, in Fig A. Specifically, the scene-wise frequency indicates the number of scenes that a certain class appears in, while the point-wise frequency indicates the total number of points belonging to a certain class across all scenes. Notably, we define the class order for CSS according to the descending order of the scene-wise frequency, *i.e.*, {other furniture, floor, wall, door, chair, table, window, cabinet, desk, picture, sink, sofa, bed, curtain, toilet, counter, refrigerator, bookshelf, bathtub, shower curtain}. This means that the novel classes to be learned in each step are set to be those rare ones, a.k.a. tail classes in the long-tail field. Take *14-5* setting for instance, the last five rare classes, i.e., counter, refrigerator, bookshelf, bathtub, shower curtain, form the novel class set, and likewise for the other settings. Meanwhile, the point-wise frequency for four out of five potential novel classes (last five classes) happens to be rare as well, *i.e.*, counter, refrigerator, bathtub and shower curtain. This suggests that such objects are either small or sparse in the captured scenes, causing additional learning difficulty especially under the CSS scenario. In summary, the two difficulties mentioned above ensure that our proposed benchmark is highly challenging.

D. Rationale for Ablation Study Setups

We conduct the ablative studies upon VOC *15-5* setting with PLOP [4] as the baseline. This is because PLOP is a significant milestone and has served as the basis for several subsequent works [9, 13]. Moreover, the VOC *15-5* setting involves only one incremental step with class 16 to class 20 consistently being new classes. In contrast, in the VOC *15-1* setting that includes multiple steps, class 16 to class 19 could be either a new or an old class, depending on the



(a) Statistics of scene-wise frequency.

(b) Statistics of point-wise frequency.

Figure A: The class frequency statistics of our proposed ScanNet benchmark for 3D continual semantic segmentation.

current learning step. However, in the ablation study, we require consistent metrics to observe both the forgetting issue (*1-15*) and generalization capacity (*16-20*). Therefore, to ensure that the metric *16-20* consistently reflects novel performance while the metric *1-15* indicates base performance, we choose the VOC *15-5* setting for the ablative studies.

E. Visualization

We provide qualitative visualization results in Fig. B (on the next page) to show the superiority of our LGKD against the other competing approaches under VOC *15-5* setting. To better compare the generalization ability and old knowledge preservation capacity across different methods, we construct three groups of samples containing *old classes only* to inspect the forgetting issue, *new classes only* to verify the generalization ability and *old & new classes* to showcase how well the model can strike a balance between generalization (plasticity) and old knowledge preservation (rigidity). The top half of the figure shows the results for the baseline methods. Specifically, FT performs well in novel classes while suffering from catastrophic forgetting in old classes. With the standard KD, ILT [8] can better preserve old knowledge, *e.g.*, the bird (green) on the second row is remembered. The recent method REMINDER [9] obtains promising results except for the tv monitor on the fourth row. As expected, Joint training yields the best among these baselines as it does not suffer from the nature of continual learning — catastrophic forgetting.

Additionally, the bottom half vividly demonstrates how our LGKD addresses the novel-background confusion and improves three existing state-of-the-art methods. Concretely, MiB [1] and PLOP [4] develop a bias toward the new classes and tend to misclassify the background as a novel class. For instance, the background weed (row 2) and stone (row 3) are classified as the new classes potted plant (row 2) and sofa (row 3) respectively. In the last row, the suitcase (background) is recognized as the new class sofa. Obviously, our LGKD can effectively alleviate the above novel-background confusion. For RCIL [13], it suffers more seriously from the novel-background confusion issue. For instance, it severely misclassifies the background weed into the new class potted plant (row 2). Additionally, it mistakes the bench (background) for the new class train (row 5). By adding our LGKD, such confusion is mostly alleviated, though with a failure case, *i.e.*, row 3, and an imperfect case, *i.e.*, row 2.

F. More Implementation Details

2D image. Following the state-of-the-art methods [1, 4, 9, 13], we adopt the Deeplab-v3 [2] model with ResNet-101 [5] as backbone and choose the output stride of 16. For fairness, we follow the training details of our competitors when comparing against them, except that we adopt synchronized batch normalization, which we found beneficial for CSS. The backbone is initialized with the ImageNet pre-trained model [11]. The initial learning rate is set to 0.01 / 0.02 for the first learning step, and 0.001 / 0.002 for subsequent steps on Pascal-VOC / ADE20k. Under all Pascal-VOC settings, the batch size is 16 for the initial learning step and 24 for subsequent steps. On ADE20k, the batch size, consistent across all learning steps, is set to 12 for *100-50* and *50-50* settings, and 16 for *100-10* setting. We train the models for 30 / 60 epochs on two NVIDIA RTX 3090 GPUs for Pascal VOC 2012 / ADE20k, while [9] trains for 70 epochs with a batch size of 10 on ADE20k with a single GPU. We notice that their prototype computation is not synchronized across multiple GPU devices, thus it does not support training on multiple GPUs. We crop the images to 512×512 during both training and testing, and apply the same data augmentation strategies as [2]. For all benchmarks, we report the mIoU results on the standard validation set.

3D point cloud. We use PointNet++ [10] with multi-scale grouping as our base model. For each input scene, we randomly sample 8192 points with replacement from a randomly chosen $1.5m \times 1.5m$ area. Data augmentation strategies (*e.g*., translation, rotation, re-scaling) are randomly applied during training. For testing, we split the whole scene into $1.5m \times 1.5m$ grids and apply the same point sampling strategy as training. We use Adam [6] with an initial learning rate 10^{-2} for the first step and 5×10^{-3} for the subsequent steps and set the weight decay to 0. The learning rate is decayed by 0.7 every 100 epochs. We train our network for 500 epochs with a batch size of 32 on a single NVIDIA RTX 3090 GPU. We remove possible duplicated points before calculating the IoU metric. Finally, we report the mIoU results on the standard validation set.

References

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

[2] Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. Rethinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv:1706.05587*, 2017. 2, 3

[3] Prithviraj Dhar, Rajat Vikram Singh, Kuan-Chuan Peng, Ziyan Wu, and Rama Chellappa. Learning without memorizing. In *CVPR*, 2019. 1

[4] 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, 2

[5] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 2

[6] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 3

[7] Zhizhong Li and Derek Hoiem. Learning without forgetting. *IEEE T-PAMI*, 40(12):2935–2947, 2017. 1

[8] Umberto Michieli and Pietro Zanuttigh. Incremental learning techniques for semantic segmentation. In *ICCV-WS*, pages 0–0, 2019. 1, 2

[9] 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. 1, 2

[10] Charles Ruizhongtai Qi, Li Yi, Hao Su, and Leonidas J Guibas. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. *Advances in neural information processing systems*, 30, 2017. 3

[11] Samuel Rota Bulò, Lorenzo Porzi, and Peter Kontschieder. In-place activated batchnorm for memory-optimized training of dnns. In *CVPR*, 2018. 2

[12] Konstantin Shmelkov, Cordelia Schmid, and Karteek Alahari. Incremental learning of object detectors without catastrophic forgetting. In *ICCV*, 2017. 1

[13] 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, 2

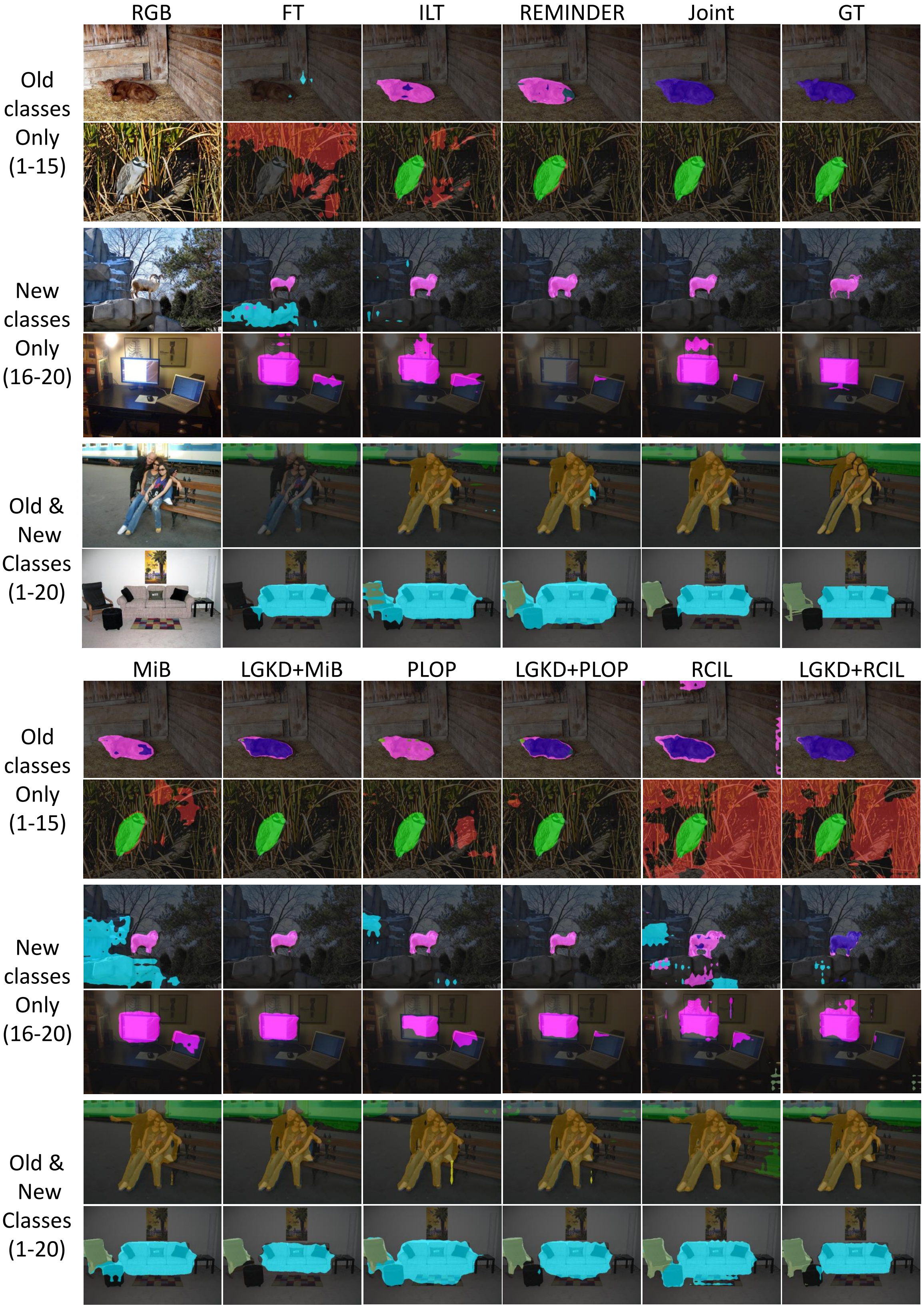


Figure B: Qualitative visualization results of baselines along with our LGKD under the VOC *15-5* setting. Our approach outperforms the others on both base classes, *e.g.*, cow, bird, person, chair, and novel classes, *e.g.*, sheep, tv monitor, train, sofa.