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

Ze Yang[1], Ruibo Li[1], Evan Ling[2], Chi Zhang[1], Yiming Wang[1],
Dezhao Huang[2], Keng Teck Ma[2], Minhoe Hur[3], Guosheng Lin[1*]
[1]School of Computer Science and Engineering, Nanyang Technological University, Singapore
[2]Hyundai Motor Group Innovation Center in Singapore (HMGICS)
[3]AIRS Company, Hyundai Motor Group
ze001@e.ntu.edu.sg, evan.ling@hmgics.com, minhoe.hur@hyundai.com, gslin@ntu.edu.sg

Abstract

Continual semantic segmentation (CSS) aims to extend an existing model to tackle unseen tasks while retaining its old knowledge. Naively fine-tuning the old model on new data leads to catastrophic forgetting. A common solution is knowledge distillation (KD), where the output distribution of the new model is regularized to be similar to that of the old model. However, in CSS, this is challenging because of the background shift issue. Existing KD-based CSS methods continue to suffer from confusion between the background and novel classes since they fail to establish a reliable class correspondence for distillation. To address this issue, we propose a new label-guided knowledge distillation (LGKD) loss, where the old model output is expanded and transplanted (with the guidance of the ground truth label) to form a semantically appropriate class correspondence with the new model output. Consequently, the useful knowledge from the old model can be effectively distilled into the new model without causing confusion. We conduct extensive experiments on two prevailing CSS benchmarks, Pascal-VOC and ADE20K, where our LGKD significantly boosts the performance of three competing methods, especially on novel mIoU by up to +76%, setting new state-of-the-art. Finally, to further demonstrate its generalization ability, we introduce the first CSS benchmark for 3D point cloud based on ScanNet, along with several re-implemented baselines for comparison. Experiments show that LGKD is versatile in both 2D and 3D modalities without requiring ad hoc design. Codes are available at https://github.com/Ze-Yang/LGKD.

1. Introduction

Fully supervised semantic segmentation has witnessed tremendous success [33, 63, 27, 8, 16, 22, 47, 65, 9] in recent years. These algorithms generally assume a fixed number of classes to be learned. However, in real-world applications, it is often expected that a deployed model can be continuously generalized to handle new classes while not forgetting the old ones. A simple solution is to expand the original dataset with newly available samples and retrain a new model from scratch, dubbed as *Joint Training*. Obviously, this is computationally expensive and requires an increasing amount of space to store the old data over time.

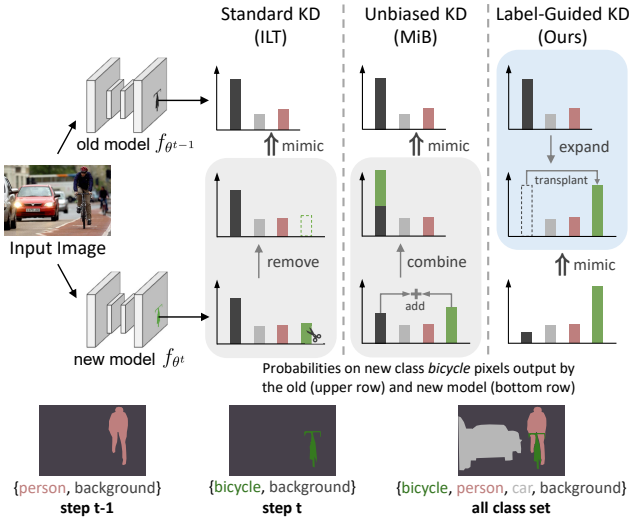

Figure 1. Illustration of knowledge distillation (KD) strategies in continual semantic segmentation. Standard KD and unbiased KD reduce the new class (bicycle) dimension(s) of the probabilities output by the new model via *remove* and *combine* respectively, which collapses the class correspondence across incremental steps and raises confusion between background and new classes (see Sec. 1). In contrast, our label-guided KD uses the ground truth label as guidance to *expand* the probabilities predicted by the old model. It builds a reliable class correspondence across different learning steps without discrepancy (ILT) or entanglement (MiB) (see Sec. 3.2). A $\xRightarrow{\text{mimic}}$ B: encourage A to be similar to B.

*Corresponding author: Guosheng Lin

Further, it may raise privacy issues in some circumstances, *e.g.*, medical images and face data.

To address this problem, continual semantic segmentation (CSS) has been proposed by [35] as an emerging research direction, where the training scheme is separated into several steps with each step tackling a set of unseen classes. Specifically, given the old model and new training data (only new classes are labeled while others are treated as background), the new model is supposed to recognize both old and new classes. Under this scenario, naively fine-tuning the old model on new data tends to suffer from *catastrophic forgetting* [17, 26], where the recognition capability of old classes is quickly lost.

Knowledge distillation (KD), first proposed in image classification [19], has recently been introduced to CSS [35, 4] to mitigate the forgetting issue. As Fig. 1 shows, *ILT* [35] removes new class probabilities (green bar) predicted by the new model, and simply distill old classes and background accordingly. However, they ignore the background shift problem [4], where the new class *bicycle* at current step t was labeled as background at last step $t-1$. As a result, the old model, which regards *bicycle* as background, will output a high background score for the new class *bicycle* pixels. Via KD, it will mislead the new model to *misclassify new classes as background*. Obviously, this KD strategy hinders the learning of new classes because the class correspondence for distillation is corrupted, *i.e.*, naively mapping the new background to the old background.

This issue was highlighted in *MiB* [4] where they proposed a new class correspondence by combination. Concretely, they combined new classes with the new background via probability summation (Fig. 1) to form a pseudo class, which was treated as the counterpart to the old background for distillation. However, this strategy, though alleviating class mis-correspondence, entangles new classes with the new background and tends to *misclassify background as new classes*, as detailed in Sec. 3.2. In this paper, we term the error of mistaking background for new (novel) classes or vice-versa as *novel-background confusion*.

The key insight to overcome the novel-background confusion is to build a reliable class correspondence across different learning steps without corruption or entanglement. To this end, we devise a novel Label-Guided Knowledge Distillation (LGKD) loss, where the class probabilities predicted by the old model are *expanded* to have the same dimension as the output of the new model (see the blue block in Fig. 1). The background probability is then transplanted to the corresponding ground truth label (class) of the input pixels. In this way, the knowledge from the old background at the last step can be correctly distilled into its corresponding semantic class at the current step, *i.e.*, either the new background or a novel class. Note that our LGKD is a generic regularization term with negligible computational cost and can be easily incorporated into existing arts. We validate its effectiveness on two prevailing CSS benchmarks Pascal-VOC and ADE20K, where our LGKD consistently yields promising improvements upon three competing methods, especially on novel (new class) mIoU (up to +76%), setting new state-of-the-art. To further demonstrate the generalization ability of our approach, we establish a challenging CSS benchmark based on ScanNet for 3D point cloud and re-implement multiple baselines for comparison. *To our best knowledge, this is the first work to conduct CSS on 3D point cloud.* Extensive experiments showcase that our LGKD loss is capable to handle both 2D and 3D modalities with no ad hoc design.

The main contributions of this paper can be summarized as:

- We propose a new label-guided knowledge distillation (LGKD) loss for CSS, which builds a reliable class correspondence across incremental steps and alleviates novel-background confusion.
- LGKD is a generic regularization term with negligible computational cost, which can be readily combined with existing methods. Extensive experiments on two prevailing CSS benchmarks, Pascal-VOC and ADE20K, showcase that LGKD significantly improves three competitive methods, particularly on novel mIoU by up to +76%, *setting new state-of-the-art*.
- To further demonstrate its generalization capability, we establish the *first* CSS benchmark for 3D point cloud based on ScanNet and re-implement multiple baselines for comparison. Experiments illustrate that our LGKD is versatile in both 2D and 3D modalities without any ad hoc design.

2. Related Works

Semantic Segmentation, aiming to perform pixel-level classification, has witnessed great advancement [63, 6, 7, 53, 22, 16] in recent years since the pioneering work Fully Convolutional Network (FCN) [33]. Among the above methods, the DeepLab series [6, 7] is well-known for their effective designs, *e.g.*, atrous spatial pyramid pooling (ASPP) [6], atrous convolution in cascade [7]. DeepLabv3 [7] is commonly used as the segmentation framework in prior CSS works [4, 13, 61].

In 3D point cloud modality, PointNet [38] is a pioneering work to directly process unstructured point cloud. Subsequent work PointNet++ [39] learns to capture local structures and recognize fine-grained patterns via a hierarchical neural network that applies PointNet recursively. While more complicated techniques have been proposed in recent works [46, 45, 62, 68, 10], we employ PointNet++ as our 3D segmentation model for its simplicity and efficiency.

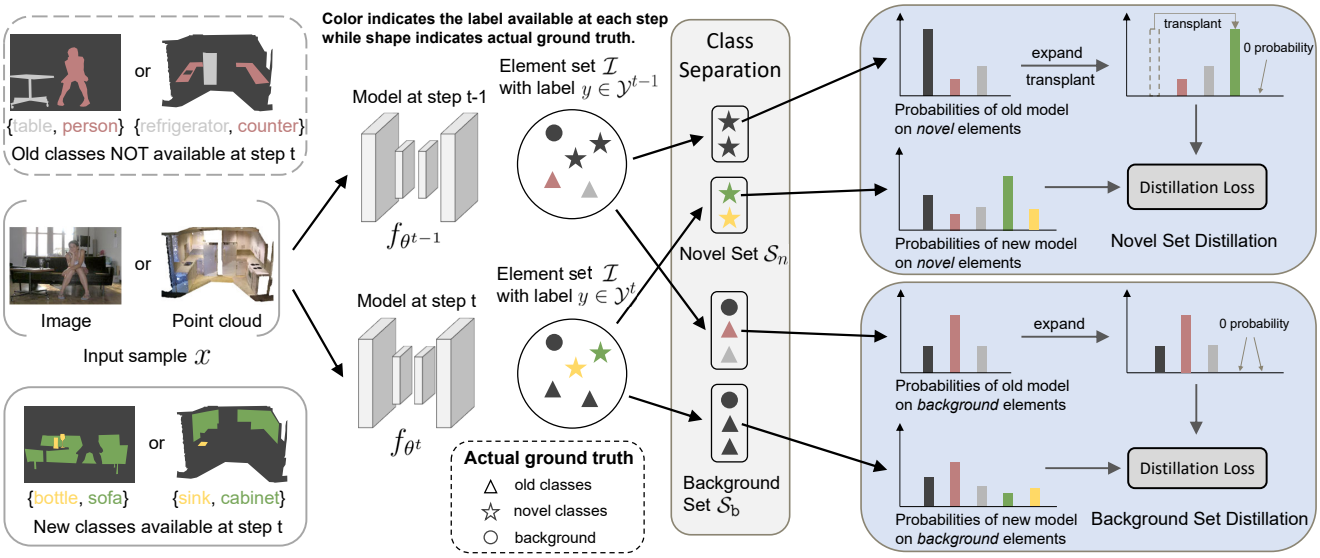

Figure 2. Illustration of our LGKD loss. At learning step t, an input sample x is fed into the old (top) and new (bottom) models to obtain the class probabilities for each element $i \in \mathcal{I}$ (pixel or point). These elements are then *separated* into the novel and background sets according to their ground truth label available at the current step t (yellow and green as novel set). Finally, to build up a reliable class correspondence, we extend the old probabilities (output by the old model) to have the same dimension as the new ones by *expand* and *transplant* (for novel set only), such that useful prior knowledge from the old model can be effectively distilled into the new model to facilitate learning new classes. We highlight that our LGKD can effectively alleviate the *novel-background confusion* issue as occurs in [35, 4].

Continual Learning aims at training a deep neural network to learn multiple tasks in sequence continually. Distinct from one-step incremental learning [50, 51, 55, 56, 59, 58, 30, 60, 29, 57, 31, 44], continual learning generally confronts more severe forgetting issues during multiple-step learning. Numerous works tackle the catastrophic forgetting issue and they can be classified into three categories. The *model growing* category [2, 48, 52] tackles this problem by dynamically extending model capacity. The *memory replay* approaches retain the old-class knowledge by keeping a small amount of old-class data in current training. Those data can be raw data [41, 40, 3], features [18, 23] or generated data [24, 42, 32]. The *regularization* approaches constrain the model by regularizing the parameters [26, 1, 54], the gradients [21, 5], logits [40, 3, 4] or the features [20, 12, 14] of the model.

Continual Semantic Segmentation. Existing CSS methods can be divided into two categories: 1) *replay-based*. RECALL [34] rehearses old-class data obtained by web-crawling or generative models. Similarly, [49] utilizes an iterative relabeling strategy with rehearsal-based incremental learning. Following [4], replay-based methods are out of the scope of this paper since accessing old data violates the conventional CSS assumption. 2) *regularization-based*. ILT [35] is the first work to introduce KD into CSS. MiB [4] raises the background shift problem in CSS and proposes an unbiased KD to tackle it. PLOP [13] distills both short- and long-range spatial relations via local pooled output distillation. SDR [36] leverages prototype matching, feature sparsification and contrastive learning to enforce feature consistency and discrimination. Most recently, class similarity KD, representation compensation, structure-preserving loss and feature projection, and the biased context are investigated by REMINDER [37], RCIL [61], SPPA [28] and RBC [64] respectively. However, the above approaches tend to suffer from the novel-background confusion problem since their distillation term lacks an appropriate class correspondence. Therefore, we propose a generic distillation term, dubbed LGKD, with a reliable class correspondence to alleviate this confusion issue.

3. Method

3.1. Problem Definition and Setups

Before introducing the continual learning setting, we first present the setups of the standard semantic segmentation task. To ensure the generality, let $\mathcal{X} \in \mathbb{R}^{N \times C_{\text{in}}}$ denote the input space (*e.g.*, image, point cloud, etc.), where N is the number of input elements and C_{in} is the number of input channels (typically $C_{\text{in}} = 3$, *i.e.*, RGB, for image and $C_{\text{in}} >= 3$, *e.g.*, XYZ, RGB, normals, for point cloud). Let $\mathcal{Y} \in \mathcal{C}^N$ denote the output label space, where $\mathcal{C}$ is a label set that contains all classes including the background class $\mathtt{b} \in \mathcal{C}$. Note that each sample $x \in \mathcal{X}$ comprises a set of elements $\mathcal{I}$ (*e.g.*, pixels, points, etc.) of corresponding label $y \in \mathcal{Y}$ with constant cardinality $|\mathcal{I}| = N$.

Given a training set $\mathcal{T} \subset \mathcal{X} \times \mathcal{Y}$, the target of semantic segmentation is to learn a model f_θ parametrized by θ that maps the input space $\mathcal{X}$ to a set of class probability vectors $f_\theta : \mathcal{X} \mapsto \mathbb{R}^{N \times |\mathcal{C}|}$. The output segmentation mask is then computed as $\hat{y} = \{\arg\max_{c \in \mathcal{C}} f_\theta(x)[i,c]\}_{i=1}^{N}$, where $f_\theta(x)[i,c]$ is the probability of element i belonging to category c.

As opposed to the standard training scheme where the model can access and learn from the full training set $\mathcal{T} \subset \mathcal{X} \times \mathcal{C}^N$ all at once, continual learning, instead, involves several learning steps $\{t\}_{t=0}^{T}$ each with a subset of label $\mathcal{C}^t$. More specifically, at training step t, the previous label set $\mathcal{C}^{0:t-1}$ will be expanded with a set of unseen classes $\mathcal{C}^t$ to form a new label set $\mathcal{C}^{0:t} = \mathcal{C}^{0:t-1} \cup \mathcal{C}^t$. Then a training set $\mathcal{T}^t \subset \mathcal{X} \times (\mathcal{C}^t)^N$ will be provided along with the last model $f_{\theta^{t-1}} : \mathcal{X} \mapsto \mathbb{R}^{N \times |\mathcal{C}^{0:t-1}|}$ to obtain an updated model $f_{\theta^t} : \mathcal{X} \mapsto \mathbb{R}^{N \times |\mathcal{C}^{0:t}|}$. Note that the training set $\mathcal{T}^t$ may contain element $i \in \mathcal{I}$ from either previous classes $\mathcal{C}^{0:t-1}$ or future classes $\mathcal{C}^{t+1:T}$, though the annotations are collapsed into the background class $\mathtt{b}$ at the current step t. This is referred to as *overlapped* [4] setting. Following the standard incremental setup, we assume that the label sets of each step are disjoint from each other except for the special background class, *i.e.*, $\mathcal{C}^i \cap \mathcal{C}^j = \mathtt{b}$ $(i \neq j)$.

3.2. Revisiting Knowledge Distillation Loss

A naive way to tackle the continual learning problem is to initialize a new model f_{θ^t} with the weights of the last model $f_{\theta^{t-1}}$ and then optimize the network with the training set $\mathcal{T}^t$. However, this will lead to the catastrophic forgetting of old classes $\mathcal{C}^{0:t-1}$. since no old class samples with annotation are available at the current step.

Standard knowledge distillation. To alleviate the forgetting problem, knowledge distillation [19] is a common solution to preserve old knowledge by regularizing the output class probability distribution of the new model f_{θ^t} to be close to that of the old model $f_{\theta^{t-1}}$ given the same input sample. Formally, the standard knowledge distillation loss ℓ_{kd} adopted in CSS by ILT [35] can be formulated as:

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

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:

$$\hat{q}_x^t(i,c) = q_x^t(i,c) / \sum\nolimits_{k \in \mathcal{C}^{0:t-1}} q_x^t(i,k)\,. \quad (2)$$

However, there exists a *discrepancy* issue in the standard distillation loss. Concretely, it fails to leverage prior knowledge to cope with the new classes, as the new class probabilities $q_x^t(i,c)$ $(c \in \mathcal{C}^t \setminus \mathtt{b})$ are not incorporated in distillation (Eq. 1). Moreover, it wrongly enhances the background probability for novel class elements due to the background shift [4]. Hence, it tends to mistake the novel classes for the background (refer to supp. material for detailed discussion).

Unbiased knowledge distillation. To overcome the above issue, MiB [4] proposes an unbiased knowledge distillation loss for CSS by revising $\hat{q}_x^t(i,c)$ in Eq. 2 as:

$$\hat{q}_x^t(i,c) = \begin{cases} q_x^t(i,c) & \text{if } c \neq \mathtt{b} \\ \sum_{k \in \mathcal{C}^t} q_x^t(i,k) & \text{if } c = \mathtt{b}\,, \end{cases} \quad (3)$$

where they compare the old background probability $q_x^{t-1}(i,\mathtt{b})$ with the probability of being either a novel class or the new background, *i.e.*, $\sum_{k \in \mathcal{C}^t} q_x^t(i,k)$, rather than directly with its counterpart $q_x^t(i,\mathtt{b})$. Such class probability combination mechanism is meant to facilitate distilling the knowledge from the old background class defined at step $t-1$ to the novel classes at step t. Nevertheless, it *entangles* the novel classes with the new background due to the probability combination. As a consequence of entanglement, the probabilities of novel classes will be undesirably enhanced when the distillation loss only intends to improve the probabilities of the new background, and vice versa. As opposed to ILT [35], this leads to the other confusion — misclassifying background as novel classes.

3.3. Label-Guided Knowledge Distillation

Class separation. Unlike standard KD and unbiased KD, which apply the same KD strategy to all input elements, we propose to separate the input elements into two sets according to *the ground truth label available at the current step*, namely the background set $\mathcal{S}_\mathtt{b} = \{i | i \in \mathcal{I}, y_i = \mathtt{b}\}$ and the novel set $\mathcal{S}_n = \{i | i \in \mathcal{I}, y_i \in \mathcal{C}^t \setminus \mathtt{b}\}$. In this way, we can customize the distillation strategy for each class set and establish a more accurate class correspondence, which facilitates more effective knowledge distillation.

Class correspondence. As opposed to revising the class probabilities of the new model $q_x^t(i,c)$ as in [4], we instead resort to correcting the class probabilities of the old model $q_x^{t-1}(i,c)$ according to the ground truth label y_i. Our design is inspired by the wireless communication system. In order to improve the useful information captured by the receiver, one should ensure that the signal released by the transmitter is correct and that both transmitter and receiver share the same communication protocol. Likewise, in our case, the objective is to assure the class probabilities (signal) predicted by the old model $f_{\theta^{t-1}}$ (transmitter) contain correct class semantics (after background shift) and are well-aligned with (protocol) the output class space of the new model f_{θ^t} (receiver). **Protocol:** For the elements from the novel set $\mathcal{S}_n$, we *expand* the output of the old model $q_x^{t-1}(i,c)$ with zero probability assigned to the extra novel classes $c \in \mathcal{C}^t \setminus \mathtt{b}$ such that it has the same output class space dimensions with the new model. **Signal:** Due to the

background shift, we *transplant* the probability of the background to the corresponding ground truth novel class and set the background probability to zero while maintaining the others unchanged. The final corrected probability distribution to be distilled from can be formulated as:

$$\bar{q}_x^{t-1}(i,c) = \begin{cases} 0 & \text{if } c \in \mathcal{C}^t \text{ and } c \neq y_i \\ q_x^{t-1}(i,\mathtt{b}) & \text{if } c = y_i \\ q_x^{t-1}(i,c) & \text{otherwise}\,, \end{cases} \tag{4}$$

where $i \in \mathcal{S}_n$. We highlight that, this novel set distillation strategy can preserve the knowledge from $f_{\theta^{t-1}}$ for the old classes while also correcting the class probabilities from the old model for the new classes given the ground truth label y_i as guidance. As for the case of the background set $i \in \mathcal{S}_\mathtt{b}$, it can be treated in a similar yet simpler way by merely expanding zero probability for the extra novel classes as:

$$\bar{q}_x^{t-1}(i,c) = \begin{cases} 0 & \text{if } c \in \mathcal{C}^t \setminus \mathtt{b} \\ q_x^{t-1}(i,c) & \text{otherwise}\,, \end{cases} \tag{5}$$

In summary, our distillation loss can be given as:

$$\ell_{\text{kd}}^{\theta^t}(x,y) = \lambda_n \cdot \bar{\ell}_{\text{kd}}^{\theta^t}(x,y,\mathcal{S}_n) + \lambda_\mathtt{b} \cdot \bar{\ell}_{\text{kd}}^{\theta^t}(x,y,\mathcal{S}_\mathtt{b})\,, \tag{6}$$

where λ_n and $\lambda_\mathtt{b}$ balances the contribution of the novel set $\mathcal{S}_n$ and background set $\mathcal{S}_\mathtt{b}$; $\bar{\ell}_{\text{kd}}^{\theta^t}(x,y,\mathcal{S})$ is the distillation loss defined within set $\mathcal{S}$ as:

$$\bar{\ell}_{\text{kd}}^{\theta^t}(x,y,\mathcal{S}) = -\frac{1}{N}\sum_{i\in\mathcal{S}}\sum_{c\in\mathcal{C}^{0:t}} \bar{q}_x^{t-1}(i,c)\log q_x^t(i,c)\,. \tag{7}$$

The overall training objective is then computed as:

$$\mathcal{L}(\theta^t) = \frac{1}{|\mathcal{T}^t|}\sum_{(x,y)\in\mathcal{T}^t}\left(\lambda\ell_{ce}^{\theta^t}(x,y) + \ell_{\text{kd}}^{\theta^t}(x,y)\right)\,, \tag{8}$$

where $\ell_{ce}^{\theta^t}(x,y)$ is the cross-entropy loss. λ is the hyperparameter to balance the importance of the two terms, and is set as $\lambda = 1$ in experiments.

4. Experiment

4.1. Benchmarking 3D Continual Segmentation

Existing CSS methods have mainly been evaluated on 2D CSS benchmarks, leaving a significant gap in the exploration of 3D CSS. In order to validate the versatility of LGKD across different modalities, we introduce the first 3D CSS benchmark based on ScanNet [11] and present several baseline methods for comparison.

ScanNet contains 1201 and 312 indoor scenes in the training and validation set respectively with 20 classes in total, including a special class for "other furniture", *i.e.* background in the standard segmentation task. In Fig. 3, we provide the scene-wise class frequency statistics of ScanNet [11], including both train and validation sets. The scene-wise class frequency indicates the number of scenes that a certain class appears in. Surprisingly, we find that ScanNet exhibits a challenging long-tail category distribution. Additionally, we provide the point-wise class frequency statistics and demonstrate the difficulty of our proposed benchmark in the supp. material. For a challenging CSS benchmark, we split the entire class set into multiple subsets according to the descending order of scene-wise class frequency as in Fig. 3, with the tail classes constituting the novel sets for continual learning. 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. Following the Pascal-VOC 2012 CSS benchmark [35, 4], we define three different incremental scenarios: adding one class (*18-1*), adding five classes simultaneously (*14-5*) and sequentially (*14-1*). 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.

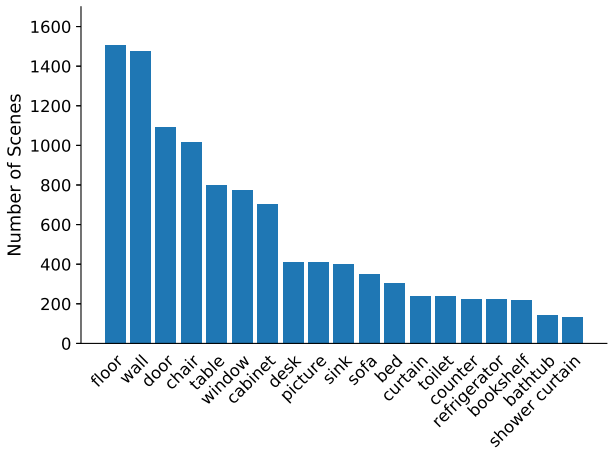

Figure 3. The scene-wise class frequency statistics of our proposed ScanNet benchmark for 3D continual semantic segmentation.

We form the training set by including all the scenes that contain at least one point from the novel classes at the current step, with the others (either previous or future classes) annotated as "other furniture". The validation set is established in a similar way except that the label of the previous classes $\mathcal{C}^{0:t-1}$ are maintained, as the model is required to predict all seen classes up to the current step. Finally, we report the *old-*, *new-* and *all-class* mIoU metrics as well as the background IoU at the end of all incremental steps.

4.2. Baselines and Setups

To validate the effectiveness of our approach, we conduct thorough comparison with the latest state-of-the-art CSS methods as well as baselines. FT is a straightforward

Method	**19-1** (2 steps)				**15-5** (2 steps)				**15-1** (6 steps)			
	0	*1-19*	*20*	*all*	*0*	*1-15*	*16-20*	*all*	*0*	*1-15*	*16-20*	*all*
FT†	-	6.8	12.9	-	-	2.1	33.1	-	-	0.2	1.8	-
EWC† [26]	-	26.9	14.0	-	-	24.3	35.5	-	-	0.3	4.3	-
LwF-MC† [40]	-	64.4	13.3	-	-	58.1	35.0	-	-	6.4	8.4	-
ILT† [35]	-	67.1	12.3	-	-	66.3	40.6	-	-	4.9	7.8	-
ILT* [35]	88.6	66.2	8.3	64.5	89.2	65.2	38.1	59.9	77.7	3.7	7.9	8.2
SDR* [36]	90.0	68.9	24.2	67.8	90.1	76.6	50.9	71.1	83.9	34.1	13.0	31.5
REMINDER* [37]	91.9	75.6	33.9	74.4	90.2	75.1	49.1	69.6	83.3	65.2	27.0	57.0
SPPA [28]	-	-	36.2	74.6	-	-	52.9	72.1	-	-	23.3	56.0
MiB† [4]	-	70.2	22.1	-	-	75.5	49.4	-	-	35.1	13.5	-
MiB* [4]	90.3	71.5	23.2	70.1	89.6	74.2	46.3	68.3	83.2	38.0	13.9	34.5
LGKD+MiB (Ours)	91.4	72.1	40.4	71.5	**91.6**	75.2	54.8	71.1	79.7	39.3	17.0	35.9
PLOP* [13]	92.1	75.3	36.1	74.2	89.7	74.2	47.5	68.6	84.7	64.6	21.0	55.2
LGKD+PLOP (Ours)	**92.9**	**76.5**	**42.9**	**75.7**	91.4	**78.7**	**56.1**	**73.9**	89.3	**69.3**	**30.9**	**61.1**
RCIL* [61]	90.1	73.6	24.5	72.1	88.9	75.9	48.4	70.0	83.9	67.9	23.1	58.0
LGKD+RCIL (Ours)	92.8	**76.5**	37.5	75.5	91.4	77.6	54.3	72.7	**89.4**	69.0	29.1	60.5
Joint	93.7	77.6	78.1	78.4	93.7	79.0	73.6	78.4	93.7	79.0	73.6	78.4

Table 1. Continual Semantic Segmentation performance (mIoU) on Pascal-VOC 2012 under different incremental scenarios. † indicates the results are excerpted from [4]. * suggests the results are reproduced with the official codes. *0* stands for the background class and *all* represents the mIoU over all classes including the background. Best results are highlighted in **Red** while runner-up in Blue.

lower-bound baseline, which simply finetunes the model with the newly available data at each step. On the contrary, Joint, training a standard segmentation model on all classes within a single step, may serve as an upper bound for CSS approaches, though not always true if the CSS benchmark shows a long-tail nature with tail classes as new classes (*e.g.*, ScanNet). For completeness, we include two general continual learning methods [26, 40] that are not specifically designed for semantic segmentation. Further, we mainly benchmark against the recent state-of-the-art methods taylored for segmentation [35, 4, 13, 36, 37, 61]. Following [4], we do not consider replay-based methods, *e.g.*, [40]) in our comparison as the access to previous annotated data violates the standard class-incremental learning assumption. For the sake of generalization, we conduct experiments on both 2D image modality, *i.e.*, Pascal-VOC 2012 [15] and ADE20K [66], and 3D point cloud modality, *i.e.* ScanNet [11]. Besides the *old-*, *new-* and *all-class* (including the background class) mIoU, we additionally report the background IoU because it is a special category under CSS with its semantic content varying across different learning steps.

CSS protocols. As for the label masking policy for the incremental steps, [4] follows and adapts two different CSS protocols, namely *disjoint* [35] and *overlapped* [43]. Both protocols assume that only the novel classes at the current step are labeled while the others are masked out as background. However, in the disjoint protocol, input samples for step t should only contain elements (*i.e.*, pixels or points) that belong to either current or previous classes, *i.e.*, $\mathcal{C}^{0:t}$. On the contrary, the overlapped protocol allows the input samples to contain either previous, current or future classes, *i.e.*, $\mathcal{C}^{0:T}$. The disjoint protocol makes a strong assumption that the current training data should not contain any samples of the class we would like to learn in the future, which is impracticable. Therefore, we adopt the more realistic overlapped protocol throughout our experiments.

4.3. Implementation Details

In 2D image modality, we follow the framework DeepLabv3 [7] and training hyperparameters of the baselines when building our LGKD upon them, except that we adopt synchronized batch normalization with different batch size and learning rate settings. In 3D point cloud modality, we use PointNet++ [39] with multi-scale grouping as our base model. We use Adam [25] 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. We remove possible duplicated points before calculating the IoU metric. For all benchmarks, we report the mIoU results on the standard validation set. We run our experiments on two NVIDIA RTX 3090 GPUs for Pascal VOC 2012 and ADE20K while one for ScanNet. More details can be found in the supp. material.

4.4. Performance

Pascal-VOC 2012 consists of 10,582 images for training and 1449 images for validation with 20 foreground categories. Following [4], we conduct three different continual learning scenarios, *i.e.*, adding one class (*19-1*), five classes simultaneously (*15-5*) and sequentially (*15-1*). The split of the class set for each class-incremental step follows the alphabetical order. Table 1 shows comprehensive results of our method along with other baselines. Clearly, FT performs poorly across all the incremental settings as the old

Method	**100-50** (2 steps)				**100-10** (6 steps)				**50-50** (3 steps)			
	0	*1-100*	*101-150*	*all*	*0*	*1-100*	*101-150*	*all*	*0*	*1-50*	*51-150*	*all*
ILT† [35]	10.9	18.4	14.4	17.0	8.5	0.0	3.1	1.1	8.6	3.4	12.9	9.7
MiB† [4]	19.7	40.7	17.2	32.8	0.0	38.6	11.1	29.2	0.0	46.5	21.0	29.3
REMINDER* [37]	26.6	42.1	16.1	33.4	24.0	38.9	20.3	32.6	24.2	48.0	20.1	29.4
SPPA [28]	-	42.9	19.9	-	-	41.0	12.5	-	-	**49.8**	23.9	-
PLOP* [13]	26.7	42.1	14.6	32.9	23.9	39.4	14.5	31.1	24.5	48.2	20.7	29.8
LGKD+PLOP (Ours)	**27.4**	**43.6**	**25.7**	**37.5**	**26.0**	42.1	**22.0**	**35.4**	25.3	49.4	**29.4**	**36.0**
RCIL* [61]	26.3	41.7	17.3	33.5	0.0	37.2	14.9	29.5	21.2	47.8	23.0	31.2
LGKD+RCIL (Ours)	26.7	43.3	25.1	37.2	24.0	**42.2**	20.4	34.9	**25.7**	49.1	27.2	34.4
Joint	29.1	42.6	28.2	37.7	29.1	42.6	28.2	37.7	29.1	49.2	32.1	37.7

Table 2. The mIoU(%) of the last step on the ADE20K dataset for different overlapped continual learning scenarios. † indicates the results are excerpted from [13]. * suggests the results are reproduced with the official codes. Best results are highlighted in **Red** while runner-up in Blue.

Method	**18-1** (2 steps)				**14-5** (2 steps)				**14-1** (6 steps)			
	0	*1-18*	*19*	*all*	*0*	*1-14*	*15-19*	*all*	*0*	*1-14*	*15-19*	*all*
FT	19.2	0.0	28.6	2.4	18.6	0.0	44.3	12.0	19.2	0.0	7.8	2.9
ILT [35]	23.8	14.9	25.8	15.9	26.7	22.2	43.4	27.7	19.3	0.0	18.5	5.6
MiB [4]	55.1	55.0	32.2	53.8	52.9	**57.8**	45.5	54.5	45.5	**56.5**	29.7	49.2
LGKD+MiB (Ours)	**55.2**	**55.5**	**40.8**	**54.8**	**54.3**	57.6	**49.0**	**55.3**	**51.9**	55.7	**38.4**	**51.2**
Joint	53.2	55.0	40.9	54.2	53.2	56.7	47.5	54.2	53.2	56.7	47.5	54.2

Table 3. Continual Semantic Segmentation performance (mIoU) on ScanNet under different incremental scenarios. *0* stands for the background class and *all* represents the mIoU over all classes including the background. All results are obtained with our implementation.

class pixels are masked as background under CSS setup, and the model is explicitly taught to wrongly predict the old classes as background. EWC [26] yields similar results for novel classes while fairly outperforming the FT baseline in terms of base mIoU. Significant improvement on base classes across all settings can be observed when knowledge distillation is introduced in [40, 35]. MiB [4] obtains consistent improvement on both base and novel performance across different settings, due to addressing the background shift problem in CSS. Recent approaches [14, 61] obtain drastically higher performance against MiB when multiple incremental steps are involved (*15-1*). Despite their promising performance, our LGKD is able to further boost their performance by a large margin, especially on novel performance. For instance, LGKD+MiB achieves +74% (40.4 vs. 23.2) improvement against MiB on novel mIoU under *19-1* setting. These promising novel class results along with the improved background performance prove that our LGKD loss can effectively alleviate the novel-background confusion problem.

ADE20K [67] is a challenging dataset consisting of 20,210 training and 2000 testing images of complex scenes. Similar to VOC, we conducted three different CSS scenarios, *i.e.*, *100-50*, *100-10* and *50-50*, as shown in Table 2. Specifically, we build our LGKD upon two competitive methods, *i.e.*, PLOP [13] and RCIL [61], where LGKD achieves considerable improvement, especially on novel mIoU (*up to +76% from 14.6 to 25.7 with PLOP under 100-50 setting*), setting new state-of-the-art across all incremental scenarios.

ScanNet. As shown in Table 3, we compare our method with FT baseline, ILT [35] and MiB [4] on our proposed ScanNet benchmark to illustrate the generality of LGKD in 3D modality. Similar to their performance in 2D modality, FT and ILT suffer from catastrophic forgetting when learning new classes. However, MiB significantly outperforms the former methods across all three settings and metrics, in particular, the base class performance, which once again demonstrates the effectiveness of tackling the background shift in CSS. Compared with MiB, our LGKD further improves the novel performance by a large margin: +27% (40.8 vs. 32.2) and +29% (38.4 vs. 29.7) relative improvement of novel mIoU against MiB are obtained under *18-1* and *14-1* settings respectively, which proves the considerable importance of overcoming the background shift with correct class correspondence. It is also noteworthy that our approach also outperforms joint training, which shows that LGKD is also effective in alleviating long-tail issues.

4.5. Ablation Study

Unless otherwise stated, all ablative studies are conducted upon VOC *15-5* setting with PLOP [13] as the baseline. Refer to the supp. material for the rationale.

Expand & Transplant. We compare three different designs for novel set distillation in Table 5. Firstly, *One-hot* is a hard-label distillation with all the class probabilities transplanted to the ground truth novel class, making the probability of the ground truth class equal to one. In this case, the distillation loss is equivalent to the standard one-hot cross-

	VOC 15-5 (2 steps)				VOC 15-1 (6 steps)			
	0	*1-15*	*16-20*	*all*	*0*	*1-15*	*16-20*	*all*
PLOP [13]	89.7	74.2	47.5	68.6	84.7	64.6	21.0	55.2
+Standard KD [35]	90.7	74.4	49.2	69.2	88.4	66.4	13.4	54.8
+Unbiased KD [4]	90.2	77.1	51.8	71.7	86.3	67.1	22.6	57.4
+LGKD (Ours)	**91.4**	**78.7**	**56.1**	**73.9**	**89.3**	**69.3**	**30.9**	**61.1**
	ADE20k 100-50 (2 steps)				**ADE20k 100-10** (6 steps)			
	0	*1-100*	*101-150*	*all*	*0*	*1-100*	*101-150*	*all*
PLOP [13]	26.7	42.1	14.6	32.9	23.9	39.4	14.5	31.1
+Standard KD [35]	**27.4**	42.5	13.6	32.8	23.2	39.8	13.5	31.0
+Unbiased KD [4]	27.0	42.6	14.5	33.2	23.7	39.8	15.1	31.5
+LGKD (Ours)	**27.4**	**43.6**	**25.7**	**37.5**	**26.0**	**42.1**	**22.0**	**35.4**

Table 4. Effect of different knowledge distillation terms. Standard KD is adopted in ILT [35] while unbiased KD is adopted in MiB [4].

	0	*1-15*	*16-20*	*all*
One-hot	**91.4**	76.7	51.9	71.5
Expand only	90.4	78.5	30.9	67.7
Expand & Transplant	**91.4**	**78.7**	**56.1**	**73.9**

Table 5. Effect of our novel set distillation design including expand and transplant operations.

λ_b	λ_n	*0*	*1-15*	*16-20*	*all*
0	0	90.5	75.5	53.6	71.0
1	0	91.3	76.9	53.1	71.9
2	0	91.6	77.3	53.1	72.2
5	0	**91.8**	77.9	52.0	72.4
10	0	**91.8**	78.3	49.8	72.2
5	0.5	91.6	78.6	55.7	73.8
5	1	91.4	**78.7**	**56.1**	**73.9**
5	2	91.0	78.6	55.5	73.7

Table 6. Effect of the weighted factors λ_n and λ_b.

entropy loss, discarding all the old class probabilities predicted by the old model. As expected, the base performance is slightly compromised compared with *Expand & Transplant* since the old class knowledge is not distilled. Nevertheless, the novel performance also drops since using hard labels for the novel class pixels will develop a bias toward novel classes, resulting in false positive predictions. This illustrates the significance of retaining the old class probabilities during transplant. Secondly, *Expand only* will encourage the new model to misclassify a novel class into either background or an old class. Thus, a sharp drop in novel mIoU from 56.1 to 30.9 is observed. This uncovers the importance of correcting the old model output by transplant to handle the background shift. Finally, transplanting the background probability into the corresponding ground truth novel class while retaining the old class probabilities can establish an appropriate class correspondence and facilitate learning new classes, hence yielding the best novel performance.

Weighted factor. As table 6 shows, increasing λ_b leads to higher base performance, but it will compromise the novel performance when λ_b becomes too large, *e.g.*, $\lambda_b = 10$. It indicates that the background set distillation is mainly responsible for retaining old knowledge. We set $\lambda_b = 5$ as the optimal solution as it obtains promising base mIoU while not sacrificing the novel mIoU severely. With the novel set distillation kicking in, the novel mIoU is significantly improved from 52.0 to 56.1 with $\lambda_n = 1$. Meanwhile, it also boosts the base mIoU from 77.9 to 78.7. This suggests that the novel set distillation facilitates learning new classes while maintaining the old class knowledge distilled from the old model.

Effect of different KDs. To validate the superiority of LGKD, we compare our LGKD with other knowledge distillation (KD) terms on both VOC and ADE20k upon the same baseline PLOP [13] (Fig. 4), since PLOP does not adopt class probability-wise KD. Firstly, PLOP can hardly benefit from adding the standard KD [35], probably due to its tendency to misclassify new classes as the background (see Sec. 3.2). Secondly, though unbiased KD [4] yields a moderate boost on VOC, it makes almost no difference on the more challenging benchmark ADE20k. In contrast, LGKD consistently surpasses standard KD and unbiased KD by a significant margin, especially in terms of novel mIoU. Concretely, LGKD yields +8.3 / +11.2 / +6.9 points improvement against unbiased KD on *15-1* / *100-50* / *100-10* setting.

4.6. Error Analysis

We carry out comprehensive error analysis on both Pascal-VOC 2012 and ScanNet datasets to validate the effectiveness of our LGKD in mitigating the novel-background confusion problem.

2D image. As Fig. 4 shows, under the VOC *15-5* incremental scenario, naive finetuning completely forgets the old class bird while also mistaking the background for the novel class potted plant. Thanks to KD mechanism, ILT [35], REMINDER [37], MiB [4], PLOP [13] and RCIL [61] are capable to retain the old knowledge (bird and chair). Nev-

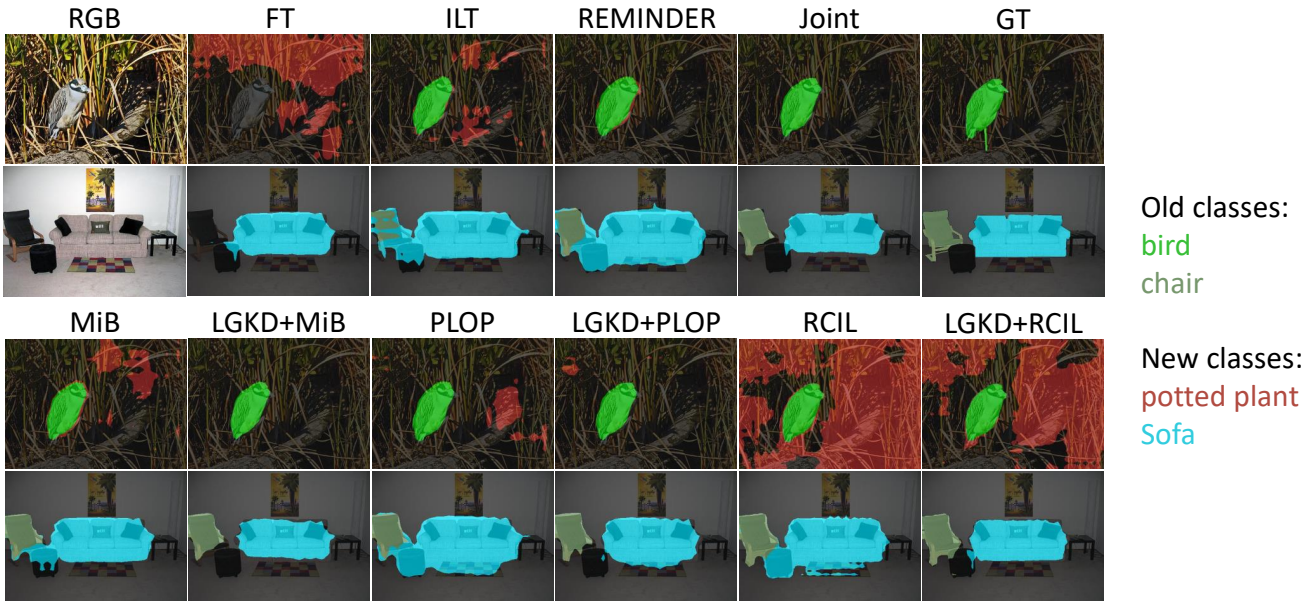

Figure 4. Error analysis of baselines along with our LGKD under VOC 15-5 setting. MiB [4], PLOP [13] and RCIL [61] tend to mistake weed and suitcase (background—things of no interest) for potted plant and sofa (novel classes) respectively. Such confusion is effectively alleviated when equipped with our LGKD, except for the bird image with RCIL, where the confusion is slightly mitigated.

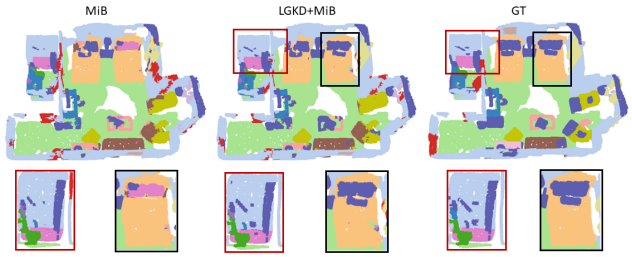

Figure 5. Error analysis of MiB [4] vs. LGKD+MiB on step 4 (with bathtub depicted in pink as the novel class) of the ScanNet *14-1* incremental scenario. Our LGKD effectively alleviates the confusion made by MiB (see the black boxes) — mistaking the background (dark purple) for the novel class bathtub. Meanwhile, LGKD+MiB achieves decent accuracy on the novel class bathtub (see the red boxes).

ertheless, these methods are consistently prone to misclassifying the background (weed and suitcase—things of no interest) as novel classes (potted plant and sofa). On the contrary, when incorporating our LGKD into MiB, PLOP and RCIL, the above novel-background confusion is effectively suppressed, except for the bird image with RCIL, where the confusion is slightly alleviated. The above examples demonstrate the effectiveness of LGKD in tackling the novel-background confusion issue. More visualization results can be found in the supp. material.

3D point cloud. We show the error results in Fig. 5 for step 4 of the ScanNet *14-1* setting, where the current novel class is bathtub depicted in pink. Clearly, our method effectively corrects the false prediction made by MiB [4] (see the black boxes), where the background points (highlighted in dark purple) are mistaken for the novel class bathtub. Meanwhile, our LGKD does not sacrifice the capability to recognize the true bathtub points (see red boxes).

5. Conclusion

In this work, we identify a key issue in continual semantic segmentation (CSS)—novel background confusion. This is generally caused by the corrupted class correspondence during knowledge distillation. To alleviate such confusion, we propose a novel solution termed Label-Guided Knowledge Distillation (LGKD), which can establish an appropriate class correspondence between the output of the old and the new model for distillation, guided by the ground truth label. We validate its effectiveness on both existing 2D benchmarks and our proposed 3D CSS benchmark, where LGKD can consistently outperform competing methods by a large margin, especially on novel mIoU, setting new state-of-the-art. Finally, we hope this work will inspire further in-depth exploration of 3D CSS in the future.

6. Acknowledgement

This work is supported by the Hyundai research grant (04OIS000252C130).

References

[1] Rahaf Aljundi, Francesca Babiloni, Mohamed Elhoseiny, Marcus Rohrbach, and Tinne Tuytelaars. Memory aware

synapses: Learning what (not) to forget. In *ECCV*, 2018. 3

[2] Rahaf Aljundi, Punarjay Chakravarty, and Tinne Tuytelaars. Expert gate: Lifelong learning with a network of experts. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3366–3375, 2017. 3

[3] Francisco M Castro, Manuel J Marín-Jiménez, Nicolás Guil, Cordelia Schmid, and Karteek Alahari. End-to-end incremental learning. In *ECCV*, 2018. 3

[4] 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, 3, 4, 5, 6, 7, 8, 9

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

[6] 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 T-PAMI*, 40(4):834–848, 2017. 2

[7] Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. Rethinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv:1706.05587*, 2017. 2, 6

[8] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *ECCV*, 2018. 1

[9] 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

[10] Ran Cheng, Ryan Razani, Ehsan Taghavi, Enxu Li, and Bingbing Liu. 2-s3net: Attentive feature fusion with adaptive feature selection for sparse semantic segmentation network. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12547–12556, 2021. 2

[11] Angela Dai, Angel X. Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proc. Computer Vision and Pattern Recognition (CVPR), IEEE*, 2017. 5, 6

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

[13] 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. 2, 3, 6, 7, 8, 9

[14] Arthur Douillard, Matthieu Cord, Charles Ollion, Thomas Robert, and Eduardo Valle. Podnet: Pooled outputs distillation for small-tasks incremental learning. In *European Conference on Computer Vision*, pages 86–102. Springer, 2020. 3, 7

[15] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge 2012 (VOC2012) Results. http://www.pascal-network.org/challenges/VOC/voc2012/workshop/index.html, 2012. 6

[16] Jun Fu, Jing Liu, Haijie Tian, Yong Li, Yongjun Bao, Zhiwei Fang, and Hanqing Lu. Dual attention network for scene segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3146–3154, 2019. 1, 2

[17] Ian J Goodfellow, Mehdi Mirza, Da Xiao, Aaron Courville, and Yoshua Bengio. An empirical investigation of catastrophic forgetting in gradient-based neural networks. *arXiv preprint arXiv:1312.6211*, 2013. 2

[18] Tyler L Hayes, Kushal Kafle, Robik Shrestha, Manoj Acharya, and Christopher Kanan. Remind your neural network to prevent catastrophic forgetting. In *European Conference on Computer Vision*, pages 466–483. Springer, 2020. 3

[19] Geoffrey Hinton, Oriol Vinyals, Jeff Dean, et al. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2(7), 2015. 2, 4

[20] Saihui Hou, Xinyu Pan, Chen Change Loy, Zilei Wang, and Dahua Lin. Learning a unified classifier incrementally via rebalancing. In *CVPR*, 2019. 3

[21] Guannan Hu, Wu Zhang, Hu Ding, and Wenhao Zhu. Gradient episodic memory with a soft constraint for continual learning. *arXiv preprint arXiv:2011.07801*, 2020. 3

[22] 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. 1, 2

[23] Ahmet Iscen, Jeffrey Zhang, Svetlana Lazebnik, and Cordelia Schmid. Memory-efficient incremental learning through feature adaptation. In *European conference on computer vision*, pages 699–715. Springer, 2020. 3

[24] Ronald Kemker and Christopher Kanan. Fearnet: Brain-inspired model for incremental learning. *arXiv preprint arXiv:1711.10563*, 2017. 3

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

[26] James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, et al. Overcoming catastrophic forgetting in neural networks. *Proceedings of the national academy of sciences*, 114(13):3521–3526, 2017. 2, 3, 6, 7

[27] Guosheng Lin, Anton Milan, Chunhua Shen, and Ian Reid. Refinenet: Multi-path refinement networks for high-resolution semantic segmentation. In *CVPR*, 2017. 1

[28] Zihan Lin, Zilei Wang, and Yixin Zhang. Continual semantic segmentation via structure preserving and projected feature

alignment. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXIX*, pages 345–361. Springer, 2022. 3, 6, 7

[29] Weide Liu, Chi Zhang, Henghui Ding, Tzu-Yi Hung, and Guosheng Lin. Few-shot segmentation with optimal transport matching and message flow. *arXiv preprint arXiv:2108.08518*, 2021. 3

[30] Weide Liu, Chi Zhang, Guosheng Lin, and Fayao Liu. Crnet: Cross-reference networks for few-shot segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4165–4173, 2020. 3

[31] Weide Liu, Chi Zhang, Guosheng Lin, and Fayao Liu. Crcnet: Few-shot segmentation with cross-reference and region–global conditional networks. *International Journal of Computer Vision*, 130(12):3140–3157, 2022. 3

[32] Yaoyao Liu, Yuting Su, An-An Liu, Bernt Schiele, and Qianru Sun. Mnemonics training: Multi-class incremental learning without forgetting. In *Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition*, pages 12245–12254, 2020. 3

[33] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *CVPR*, 2015. 1, 2

[34] Andrea Maracani, Umberto Michieli, Marco Toldo, and Pietro Zanuttigh. Recall: Replay-based continual learning in semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7026–7035, 2021. 3

[35] Umberto Michieli and Pietro Zanuttigh. Incremental learning techniques for semantic segmentation. In *ICCV-WS*, pages 0–0, 2019. 2, 3, 4, 5, 6, 7, 8

[36] 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, 6

[37] 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, 6, 7, 8

[38] Charles R Qi, Hao Su, Kaichun Mo, and Leonidas J Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 652–660, 2017. 2

[39] 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. 2, 6

[40] Sylvestre-Alvise Rebuffi, Alexander Kolesnikov, Georg Sperl, and Christoph H Lampert. icarl: Incremental classifier and representation learning. In *CVPR*, 2017. 3, 6, 7

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

[42] Hanul Shin, Jung Kwon Lee, Jaehong Kim, and Jiwon Kim. Continual learning with deep generative replay. In *NeurIPS*, 2017. 3

[43] Konstantin Shmelkov, Cordelia Schmid, and Karteek Alahari. Incremental learning of object detectors without catastrophic forgetting. In *ICCV*, 2017. 6

[44] Nan Song, Chi Zhang, and Guosheng Lin. Few-shot open-set recognition using background as unknowns. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 5970–5979, 2022. 3

[45] Haotian Tang, Zhijian Liu, Shengyu Zhao, Yujun Lin, Ji Lin, Hanrui Wang, and Song Han. Searching efficient 3d architectures with sparse point-voxel convolution. In *European conference on computer vision*, pages 685–702. Springer, 2020. 2

[46] Hugues Thomas, Charles R Qi, Jean-Emmanuel Deschaud, Beatriz Marcotegui, François Goulette, and Leonidas J Guibas. Kpconv: Flexible and deformable convolution for point clouds. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6411–6420, 2019. 2

[47] Huiyu Wang, Yukun Zhu, Hartwig Adam, Alan Yuille, and Liang-Chieh Chen. Max-deeplab: End-to-end panoptic segmentation with mask transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5463–5474, 2021. 1

[48] Ju Xu and Zhanxing Zhu. Reinforced continual learning. *Advances in Neural Information Processing Systems*, 31, 2018. 3

[49] Shipeng Yan, Jiale Zhou, Jiangwei Xie, Songyang Zhang, and Xuming He. An em framework for online incremental learning of semantic segmentation. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 3052–3060, 2021. 3

[50] Ze Yang, Yali Wang, Xianyu Chen, Jianzhuang Liu, and Yu Qiao. Context-transformer: Tackling object confusion for few-shot detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 12653–12660, 2020. 3

[51] Ze Yang, Chi Zhang, Ruibo Li, Yi Xu, and Guosheng Lin. Efficient few-shot object detection via knowledge inheritance. *IEEE Transactions on Image Processing*, 32:321–334, 2022. 3

[52] Jaehong Yoon, Eunho Yang, Jeongtae Lee, and Sung Ju Hwang. Lifelong learning with dynamically expandable networks. *arXiv preprint arXiv:1708.01547*, 2017. 3

[53] 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. 2

[54] Friedemann Zenke, Ben Poole, and Surya Ganguli. Continual learning through synaptic intelligence. In *ICML*, 2017. 3

[55] Chi Zhang, Yujun Cai, Guosheng Lin, and Chunhua Shen. Deepemd: Few-shot image classification with differentiable earth mover's distance and structured classifiers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12203–12213, 2020. 3

[56] Chi Zhang, Yujun Cai, Guosheng Lin, and Chunhua Shen. Deepemd: Differentiable earth mover's distance for few-shot

learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(5):5632–5648, 2022. 3

[57] Chi Zhang, Henghui Ding, Guosheng Lin, Ruibo Li, Changhu Wang, and Chunhua Shen. Meta navigator: Search for a good adaptation policy for few-shot learning. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9435–9444, 2021. 3

[58] Chi Zhang, Guosheng Lin, Fayao Liu, Jiushuang Guo, Qingyao Wu, and Rui Yao. Pyramid graph networks with connection attentions for region-based one-shot semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9587–9595, 2019. 3

[59] Chi Zhang, Guosheng Lin, Fayao Liu, Rui Yao, and Chunhua Shen. Canet: Class-agnostic segmentation networks with iterative refinement and attentive few-shot learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5217–5226, 2019. 3

[60] Chi Zhang, Nan Song, Guosheng Lin, Yun Zheng, Pan Pan, and Yinghui Xu. Few-shot incremental learning with continually evolved classifiers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12455–12464, 2021. 3

[61] 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. 2, 3, 6, 7, 8, 9

[62] Yang Zhang, Zixiang Zhou, Philip David, Xiangyu Yue, Zerong Xi, Boqing Gong, and Hassan Foroosh. Polarnet: An improved grid representation for online lidar point clouds semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9601–9610, 2020. 2

[63] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *CVPR*, 2017. 1, 2

[64] Hanbin Zhao, Fengyu Yang, Xinghe Fu, and Xi Li. Rbc: Rectifying the biased context in continual semantic segmentation. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXIV*, pages 55–72. Springer, 2022. 3

[65] Sixiao Zheng, Jiachen Lu, Hengshuang Zhao, Xiatian Zhu, Zekun Luo, Yabiao Wang, Yanwei Fu, Jianfeng Feng, Tao Xiang, Philip HS Torr, et al. Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6881–6890, 2021. 1

[66] Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ade20k dataset. In *CVPR*, 2017. 6

[67] 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. 7

[68] Xinge Zhu, Hui Zhou, Tai Wang, Fangzhou Hong, Yuexin Ma, Wei Li, Hongsheng Li, and Dahua Lin. Cylindrical and asymmetrical 3d convolution networks for lidar segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9939–9948, 2021. 2