LAGD: Local Topological-Alignment and Global Semantic-Deconstruction for Incremental 3D Semantic Segmentation

Yumin Zhang[1], Haoran Duan[1], Rui Sun[1], Yue Cheng[2]
Tejal Shah[1], Rajiv Ranjan[1], Bo Wei[1]*
[1]School of Computing, Newcastle University, United Kingdom
[2]School of Software Engineering, Beijing Jiaotong University, China
{y.zhang361, haoran.duan, rui.sun, tejal.shah, raj.ranjan, bo.wei}@newcastle.ac.uk, yuecheng@bjtu.edu.cn

Abstract

Numerous deep learning-based works focusing on 3D semantic segmentation have been proposed and have achieved impressive performance. However, due to catastrophic forgetting, existing methods degrade dramatically in a real-world scenario where new 3D semantic categories are arriving continually. As such, applying typical class-incremental learning methods on 3D data can aggravate forgetting due to their irregular and noisy geometric structure. Aiming to address this realistic challenge, from the perspective of capturing local topological characteristics and mitigating global semantic shift, we propose a unified framework named Local topological Alignment and Global semantic Deconstruction (LAGD) to incrementally learn semantic knowledge of novel 3D categories while maintaining performance on previously learned knowledge. Specifically, we develop a novel Interaction Topological-aware Alignment (ITA) to maintain the learned knowledge efficiently by capturing the local geometric characteristics with interacted adjacent state-specific knowledge. Besides, to mitigate the forgetting caused by the global semantic shift, we deconstruct the logits into positive and negative parts which are distilled separately, achieving an elaborate distillation process in terms of Semantic-knowledge Deconstruction Distillation (SDD). With the cooperation of ITA and SDD, LAGD achieves a state-of-the-art performance, especially in the long-term incremental learning scenario. Extensive experimental results illustrate the superiority of our proposed LAGD.

Introduction

Semantic segmentation is a fundamental problem in computer vision and has achieved dramatic developments in various fields, *e.g.* autonomous driving (Feng et al. 2020), medical analysis (Ronneberger, Fischer, and Brox 2015; Huang et al. 2020), and few-shot learning (Wang et al. 2019a; Zhang et al. 2024). Unlike regular pixels in 2D vision, 3D point cloud semantic segmentation is more challenging due to its unstructured and unordered characteristics (Yang et al. 2023a). Various fully supervised 3D semantic segmentation approaches (Qi et al. 2017a,b; Zhao et al. 2021; Wu et al. 2024) have been proposed to address this issue and have achieved impressive performance in recent years.

*Corresponding author.

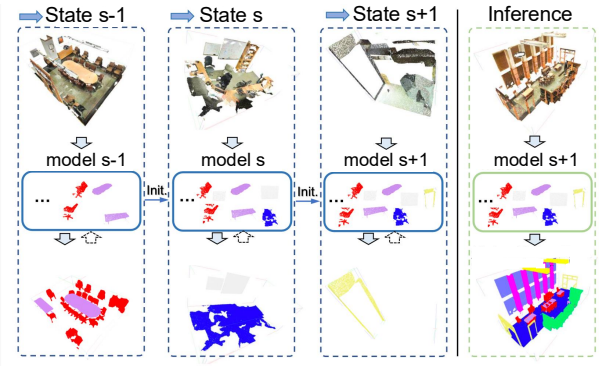

Figure 1: Illustration of incremental 3D semantic segmentation. In each state training state, only ground-truth of current categories is accessible, while in the inference phase, point clouds should be segmented into all semantic categories.

However, existing 3D semantic segmentation methods trained on *close-set* and *static* datasets are not suitable in the real-world class-incremental scenario where the novel 3D categories which are disjoint from previous categories are extended incrementally, as shown in Figure 1. In such a situation, due to catastrophic forgetting (Kirkpatrick et al. 2017), the updating model cannot perform well on previously learned knowledge. A straightforward method to handle it is to retrain the segmentation model with all available data (*i.e.*, joint training), which is restrained by data privacy and the cost of retraining (Zhang et al. 2022). Thus, class incremental learning is proposed to alleviate catastrophic forgetting in updating parameters with available supervised labels only for current categories (Rebuffi et al. 2017).

Nevertheless, most of the class incremental methods focused on 2D vision (Shmelkov, Schmid, and Alahari 2017; Kemker et al. 2018; Zhang et al. 2022), while with less attention to 3D context and mainly focusing on the class-level (Dong et al. 2021; Liu et al. 2021; Chowdhury et al. 2022) simpler than the fine-grind semantic segmentation task. Naively applying current 2D Class-Incremental Semantic Segmentation (CISS) methods to 3D point clouds neglects the precious and unique geometric characteristics features thereby aggravating forgetting (Dong et al. 2021).

Considering these challenges, it is necessary to equip the model with the ability to incrementally explore novel 3D semantic segmentation and maintain the performance without previous training data. Currently, only a few works (Yang et al. 2023a,b) focus on tackling the CISS problem in the 3D context. From the perspective of addressing the background shift, LGKD (Yang et al. 2023b) adjusts the logit scores based on the output from the previously learned model, which neglects the unique 3D geometric knowledge of point clouds. While GUAT (Yang et al. 2023a) constructed a geometric-aware structure to transfer the learned knowledge, it is restricted in the specific edge backbone (Wang et al. 2019c) thus hindering segmentation performance.

Considering these challenges, we propose a unified framework named Local topological-Alignment and Global semantic-Deconstruction (LAGD) to incrementally learn 3D novel categories without catastrophic forgetting. Specifically, from the perspective of capturing local geometric characteristics, we implement the Interaction Topological-aware Alignment (ITA), which consists of two stages: State-specific Knowledge Interaction (SKI) and Topological-aware Structure Alignment (TSA). Due to the previously learned knowledge, maintenance can be viewed as a two-state knowledge interaction process where, to capture the state-specific knowledge exactly, we first interact with representations in SKI. Then, in TSA, we construct topological-aware structures based on the interacted features to model the geometric characteristics embedded in representations. Such structures will be aligned to efficiently transfer the previously learned representation knowledge. However, the global semantic shift also contributes to catastrophic forgetting. Inspired by the success of current decoupled logits techniques, we can efficiently mitigate such a shift by deconstructing the global semantic logits score into optimistic and negative parts. From the intuitive idea that *birds of a feather flock together*, we distill these two parts separately (*i.e.*, Semantic-knowledge Deconstruction Distillation, SDD). Therefore, semantic knowledge is distilled in a fine-grained manner, effectively reducing confused distillation. Comprehensive experiments on two classic 3D semantic segmentation benchmarks: S3DIS (Armeni et al. 2016) and ScanNet (Dai et al. 2017) illustrate the effectiveness of our proposed LAGD. In summary, the main contributions of this work are as follows:

- From the perspective of capturing local geometric characteristics and mitigating global semantic shift, we propose a unified framework LAGD to learn the novel 3D categories incrementally without catastrophic forgetting.
- We develop Interaction Topological-aware Alignment (ITA) to transfer representation knowledge efficiently by aligning the topological-aware structures constructed via interacted state-specific knowledge features.
- Considering the catastrophic forgetting partially caused by the global semantic shift, we deconstruct the logits into optimistic and negative parts which are aligned separately, reducing the confused distillation.
- Comprehensive experimental results on 3D CISS benchmarks illustrate the effectiveness of our method.

Related Work

3D Semantic Segmentation

Voxel-based and point-based methods are the two main streams of the 3D semantic segmentation task. Generally, the voxel-based solutions (Graham, Engelcke, and Van Der Maaten 2018; Choy, Gwak, and Savarese 2019) firstly voxelize original 3D point clouds as regular voxels which are then forwarded into the standard 3D convolutions (Guo et al. 2020). While these methods have achieved proper performance, the inaccurate position information caused by the voxelization process hinders further enhancement (Lai et al. 2022). On the other hand, point-based methods (Qi et al. 2017a,b; Wang et al. 2019b) directly adopt each point feature and corresponding position as input, relying on the powerful characterization capacity of network architecture, achieving impressive success on 3D semantic segmentation tasks. Encouraged by the success of graph networks in capturing geometric information, various graph-based methods (Landrieu and Simonovsky 2018; Wang et al. 2019b) have been proposed. Recently, self-attention mechanism (Lai et al. 2022; Wu et al. 2024) also has been applied to 3D semantic segmentation with excellent performance.

However, these methods are unsuitable for real-world applications where new classes of 3D objects arrive consecutively, requiring the model to learn knowledge incrementally and without catastrophic forgetting.

3D Class-Incremental Learning

Compared to 2D regular pixels, irregular 3D point clouds encompass more sufficient yet complex information, making 3D Class-Incremental Learning (3D-CIL) still a challenging research area. Current 3D-CIL works mainly focus on classification (Liu et al. 2021; Dong et al. 2021; Chowdhury et al. 2022) and object detection (Zhao and Lee 2022; Liang et al. 2023). Specifically, L3DOC (Liu et al. 2021) achieves the goal of lifelong learning on 3D object classification by constructing a distillation based on the point-knowledge memory. In I3DOL (Dong et al. 2021), a geometric attention module is designed to prevent forgetting caused by redundant geometric information. The few-shot 3D classification explored in the FSCIL (Chowdhury et al. 2022). SDCoT (Zhao and Lee 2022) develops a static-dynamic co-teaching module to achieve the balance between alleviating forgetting and consistently learning. Compared to these methods, 3D Class-Incremental Semantic Segmentation (3D-CISS) is a more fine-grained task. In the context of 3D-CISS, the label-guided knowledge distillation (LGKD) (Yang et al. 2023b) is designed to transfer the previously learned knowledge with less confusion, performing well on 2D images and 3D point clouds. Besides, geometry and uncertainty-aware transfer scheme (GUAT) (Yang et al. 2023a) is proposed to effectively transfer the geometric information by aligning the constructed edge-feature structure. Nevertheless, LGKD neglects the unique local characteristics, that hinder performance enhancement for 3D data. As for GUAT, requiring the specific edge network backbone to construct the geometric structure restricts the application.

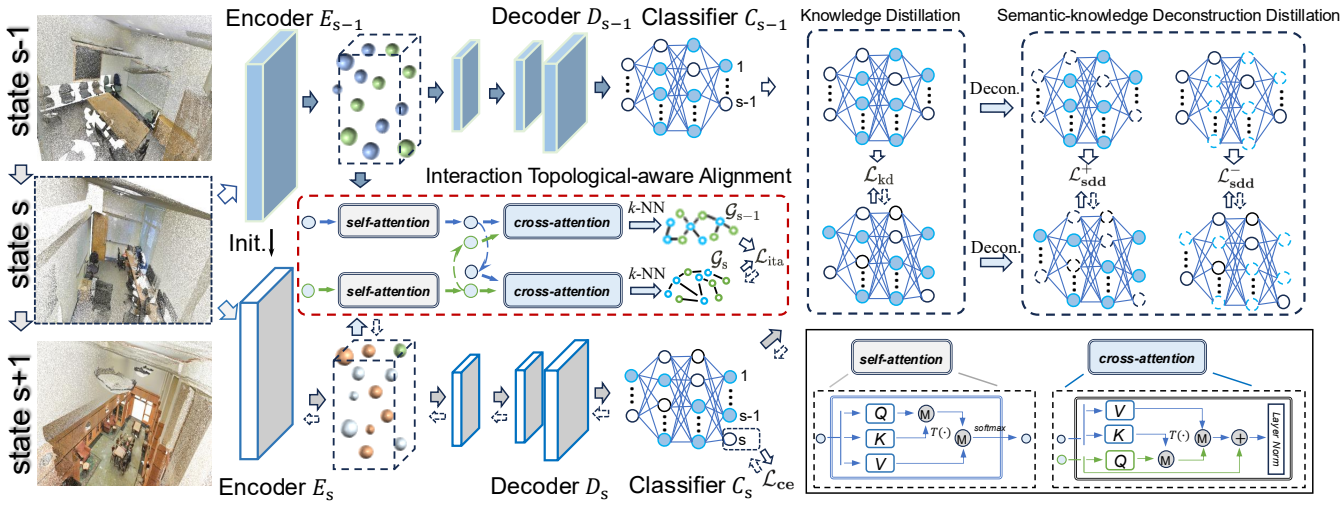

Figure 2: Illustration of LAGD which mainly consists of two parts: ITA and SDD. In the ITA stage, the extracted representation features are interacted with to model the state-specific knowledge and then to construct topological-aware structures. In the SDD stage, the logits are deconstructed into two parts to distill the semantic knowledge in a fine-grained manner.

Proposed Method

Problem Definition and Overview

Problem Definition Our setting follows the previous 3D-CISS works (Yang et al. 2023a,b). Specifically, given an incremental learning task $\mathcal{T}$ consists of $\mathcal{S}$ incremental states:

$$\mathcal{T} = \{\mathcal{T}^1, \mathcal{T}^2, \cdots, \mathcal{T}^{\mathcal{S}}\}. \tag{1}$$

Here, we use $\mathcal{T}^s = \{x_s^i, y_s^i\}_{i=1}^{n^s}$ with n^s points $x_s^i \in \mathbb{R}^{\tau\times(3+f)}$ to denote the s-th $(s = 1, 2, \cdots, \mathcal{S})$ incremental state training task, τ is the number of sample points per point cloud. x_s^i denotes the i-th point data with XYZ coordinates with additional f dimension features (*e.g.*, RGB color) and y_s^i is the corresponding ground-truth. Notably, $\{y_s^i\}$ covers only the novel $\mathcal{C}_s$ classes in current s-th incremental state which are disjoint from previous $\mathcal{C}_p$ classes (*i.e.*, $\sum_{p=0}^{s-1}\mathcal{C}_p \cap \mathcal{C}_s = \varnothing$). Our goal is to learn a model F that can make reliable segmentations on both previous categories $\mathcal{C}_p$ and novel categories $\mathcal{C}_s$, at the s-th incremental learning state. The basic pipeline of 3D-CISS is shown in Figure 1.

Overview The overview of our LAGD framework is shown in Figure 2, which mainly consists of two parts: Interaction Topological-aware Alignment (ITA) and Semantic-knowledge Deconstruction Distillation (SDD). Specifically, in the s-th incremental training state, the current model $F_s = E_s \circ D_s \circ C_s$ is first initialized by the pretrained model F_{s-1}, where E_s, D_s, and C_s represent the encoder, decoder and classifier at s-th incremental state, respectively. For the input x_s, the representation features $\mathcal{I}_{s-1}$ and $\mathcal{I}_s$ are extracted from the encoders E_{s-1} and E_s. Further, to model the state-specific knowledge, the representation features first interacted with the attention mechanism to obtain the interacted features $\widehat{\mathcal{I}}_{s-1|s}$ and $\widehat{\mathcal{I}}_{s|s-1}$, which are utilized to construct the topological-aware structure $\mathcal{G}_{s-1}$ and $\mathcal{G}_s$. Based on this, we can calculate the loss $\mathcal{L}_{\text{ita}}$ which is optimized to achieve the representation alignment and thus maintain previously learned knowledge. Besides, considering the global semantic shift also results in catastrophic forgetting, we deconstruct the logits into positive and negative parts which are distilled separately in SDD. Hence, we can transfer the global semantic in a fine-grained distillation manner by mitigating $\mathcal{L}_{\text{sdd}} = \mathcal{L}_{\text{sdd}}^{+} + \mathcal{L}_{\text{sdd}}^{-}$, achieving an exact global semantic alignment. ITA and SDD are discusssed further in the following subsections.

Supervised Semantic Segmentation Learning

During the s-th incremental state, the supervised labels that we can access only cover the novel categories $\mathcal{C}_s$. Given the input x_s^i, we utilize the corresponding ground-truth y_s^i to calculate the cross-entropy loss $\mathcal{L}_{\text{ce}}$ via:

$$\mathcal{L}_{\text{ce}} = \mathbb{E}_{\{x_s^i, y_s^i\}\in\mathcal{T}^s}[-y_s^i \log F_s(x_s^i; \Theta_{F_s})]. \tag{2}$$

Here, we use Θ_{F_s} to denote the parameters of F_s.

Interaction Topological-aware Alignment

To better capture and transfer the state-specific knowledge, we develop a novel method named Interaction Topological-aware Alignment (ITA) which can be decomposed into two stages: State-specific Knowledge Interaction (SKI) and Topological-aware Structure Alignment (TSA).

State-specific Knowledge Interaction Given the s-th incremental state data $\mathcal{T}^s = \{x_s^i, y_s^i\}$, we first leverage the encoders $E_{s-1/s}$ to extract the representation features $\mathcal{I}_{s-1/s}$, respectively. Since alleviating catastrophic forgetting can be viewed as a cooperative interacted optimization process, we model the state-specific representation knowledge by applying an interaction module to the $\mathcal{I}_{s-1/s}$ to obtain the interacted features $\widehat{\mathcal{I}}_{s-1|s}$ and $\widehat{\mathcal{I}}_{s|s-1}$via:

$$\begin{aligned}\widehat{\mathcal{I}}_{s-1|s} &= \text{InterModule}(\mathcal{I}_{s-1}|\widetilde{\mathcal{I}}_s),\\ \widehat{\mathcal{I}}_{s|s-1} &= \text{InterModule}(\mathcal{I}_s|\widetilde{\mathcal{I}}_{s-1}).\end{aligned} \tag{3}$$

Specifically, to capture the intra-state characteristics, $\mathcal{I}_{s-1/s}$ are first implemented self-attention mechanism:

$$\widetilde{\mathcal{I}}_{s-1} = \text{Softmax}[(\widetilde{\mathcal{W}}_Q \cdot \mathcal{I}_{s-1})(\widetilde{\mathcal{W}}_K \cdot \mathcal{I}_{s-1})^\top](\widetilde{\mathcal{W}}_V \cdot \mathcal{I}_{s-1}). \tag{4}$$

Here, $\widetilde{\mathcal{W}}_{Q,K,V}$ denote the parameters of Q, K and V of self-attention module. Similarily, we can calculate $\widetilde{\mathcal{I}}_s$ by replacing $\mathcal{I}_{s-1}$ with $\mathcal{I}_s$ in Equation (4). Further, to transfer the state-specific knowledge, we exchange the knowledge between $\widetilde{\mathcal{I}}_s$ and $\widetilde{\mathcal{I}}_{s-1}$:

$$\begin{aligned}\widehat{\mathcal{I}}_{s-1|s} = \text{LN}\{&\text{Softmax}[(\widehat{\mathcal{W}}_Q \cdot \widetilde{\mathcal{I}}_{s-1})(\widehat{\mathcal{W}}_K \cdot \widetilde{\mathcal{I}}_s)^\top]\\ &(\widehat{\mathcal{W}}_V \cdot \widetilde{\mathcal{I}}_s) + \widetilde{\mathcal{I}}_{s-1}\},\\ \widehat{\mathcal{I}}_{s|s-1} = \text{LN}\{&\text{Softmax}[(\widehat{\mathcal{W}}_Q \cdot \widetilde{\mathcal{I}}_s)(\widehat{\mathcal{W}}_K \cdot \widetilde{\mathcal{I}}_{s-1})^\top]\\ &(\widehat{\mathcal{W}}_V \cdot \widetilde{\mathcal{I}}_{s-1}) + \widetilde{\mathcal{I}}_s\}.\end{aligned} \tag{5}$$

We use $\widehat{\mathcal{W}}_{Q,K,V}$ to denote the parameters of Q, K and V of the interaction module, and leverage $\text{LN}(\cdot)$ to represent the layer-norm operation.

Topological-aware Structure Alignment After State-specific Knowledge Interaction, we can obtain the representations features $\widehat{\mathcal{I}}_{s-1/s}$ which have exchanged knowledge between the incremental state $s-1$ and s. In addition, the topological-aware characteristics embedded in $\widehat{\mathcal{I}}_{s-1/s}$ are valuable and can be well captured by the topological-aware structure to benefit knowledge transfer.

In particular, given $\widehat{\mathcal{I}}_s = \{\widehat{\mathcal{I}}_s^{(1)}, \widehat{\mathcal{I}}_s^{(2)}, \cdots, \widehat{\mathcal{I}}_s^{(n_j^s)}\}$, we first leverage the k-NN algorithm to sample the k points $\mathcal{P}_s = \{\mathcal{P}_s^{(1)}, \mathcal{P}_s^{(2)}, \cdots, \mathcal{P}_s^{(k)}\}(k < n_j^s)$ from $\widehat{\mathcal{I}}_s$. Then, based on these points, we can obtain the topological-aware structure $\mathcal{G}_s = \{\mathcal{G}_s^{(j,1)}, \mathcal{G}_s^{(j,2)}, \cdots, \mathcal{G}_s^{(j,k)}\}_{j=1}^{n_j^s}$, where

$$\mathcal{G}_s^{(j,k)} = \mathcal{P}_s^{(j)} - \mathcal{P}_s^{(k)}. \tag{6}$$

Similarly, we can obtain the topological-aware structure $\mathcal{G}_{s-1}$ from $\widehat{\mathcal{I}}_{s-1}$. Therefore, the alignment between $\mathcal{G}_s$ and $\mathcal{G}_{s-1}$ can be mitigated by optimizing:

$$\mathcal{L}_{\text{ita}} = \mathbb{E}||\mathcal{G}_s - \mathcal{G}_{s-1}||^2. \tag{7}$$

By aligning the interaction topological-aware structures, the unique geometric knowledge is considered during the knowledge transfer, effectively alleviating the catastrophic forgetting caused by the geometric knowledge loss.

In our implementation, only the representation features extracted from the first layer of the encoder are utilized to construct the topological-aware structure for efficiency.

Semantic-knowledge Deconstruction Distillation

To alleviate the catastrophic forgetting caused by the global semantic shift, it is necessary to distill the semantic knowledge from the pre-trained model to the current model. Due to this reason, Knowledge Distillation (KD) techniques (Hinton, Vinyals, and Dean 2015) are widely used in 2D class-incremental learning (Li and Hoiem 2017; Michieli and Zanuttigh 2019). Specifically in the context of point clouds, for the given input x_s, we can obtain the semantic features $\mathcal{F}_s = D_s(E_s(x^s;\Theta_{E_s});\Theta_{D_s})$. Then, the logit value $\mathcal{W}_s(h,c)$ of a specific 3D class c is calculated by the dot product between the weights of classifier C_s and the semantic features $\mathcal{F}_s$ via:

$$\mathcal{W}_s(h,c) = \Theta_{C_s}(c)^\top \cdot \mathcal{F}_s(h), \tag{8}$$

where $\Theta_{C_s}(c)$ denotes the weights of the classifier for the c-th 3D category, and $\mathcal{F}_s(h)$ represents the semantic features at the position h. For simplification, we omit the bias term of the classifier. Vanilla KD loss $\mathcal{L}_{\text{kd}}$ directly encourages the logits $\mathcal{W}_s$ to approach $\mathcal{W}_{s-1}$ and can be formulized via:

$$\begin{aligned}\mathcal{L}_{\text{kd}} = -\mathbb{E}[&\sum_{c\in\mathcal{C}_p} \sigma_{s-1}(h,c)\log\sigma_s(h,c)\\ &+ (1-\sigma_{s-1}(h,c))\log(1-\sigma_s(h,c))],\end{aligned} \tag{9}$$

where $\sigma_s(h,c) = [1+e^{-\mathcal{W}_s(h,c)}]^{-1}$ with similar definition of $\sigma_{s-1}(h,c)$. However, the vanilla KD is demonstrated as a coupled formulation which limits the potential of logits (Zhao et al. 2022). Considering this, we adopt a deconstructed strategy to effectively mitigate the semantic shift. To some extent, logits reflect how likely and unlikely an input belongs to a specific category (Baek et al. 2022). We can deconstruct the logits into positive and negative parts according to the signal of the classifier parameters Θ_{C_s}:

$$\begin{aligned}\mathcal{W}_s(h,c) &= (\Theta_{C_s}^+ + \Theta_{C_s}^-)\cdot\mathcal{F}_s(h)\\ &= \underbrace{\Theta_{C_s}^+\cdot\mathcal{F}_s(h)}_{\mathcal{W}_s^+(h,c)} + \underbrace{\Theta_{C_s}^-\cdot\mathcal{F}_s(h)}_{\mathcal{W}_s^-(h,c)}.\end{aligned} \tag{10}$$

From a simple yet efficient strategy that *birds of a feather flock together*, we separately distill the two parts of adjacent states. Such a decoupled distillation is called Semantic-knowledge Deconstruction Distillation (SDD), and SDD loss $\mathcal{L}_{\text{sdd}}$ can be formulated as:

$$\mathcal{L}_{\text{sdd}} = \mathcal{L}_{\text{sdd}}^+ + \mathcal{L}_{\text{sdd}}^-, \tag{11}$$

where

$$\begin{aligned}\mathcal{L}_{\text{sdd}}^+ = -[&\mathbb{E}\sum_{c\in\mathcal{C}_p} \Phi_{\text{s-1}}^+(h,c)\log\Phi_{\text{s}}^+(h,c)\\ &+ (1-\Phi_{s-1}^+(h,c))(1-\log\Phi_{\text{s-1}}^+(h,c))].\end{aligned} \tag{12}$$

Here, $\Phi_s^+(h,c) = [1+e^{\mathcal{W}_s^+(h,c)}]^{-1}$. Similarly, we can calculate $\Phi_{s-1}^+(h,c)$. By replacing $\Phi_{s-1}^+(h,c)$ with $\Phi_{s-1}^-(h,c)$, the negative part $\mathcal{L}_{\text{sdd}}^-$ can be obtained. Through optimizing $\mathcal{L}_{\text{sdd}}$, the semantic knowledge from the old model to the novel model is transferred more flexibly and exactly. The difference between KD and SDD is shown in Figure 2.

Overall Optimization and Inference

When the incremental state $s=0$, we name it as the base training state. Suppose the set of base categories is $\mathcal{C}_0$, the base model F_0 is optimized via minimizing Equation (2).

Incremental Training The overall objective function $\mathcal{L}_{\text{obj}}$ of LAGD in incremental states is formulated as:

$$\mathcal{L}_{\text{obj}} = \mathcal{L}_{\text{ce}} + \alpha\mathcal{L}_{\text{kd}} + \beta\mathcal{L}_{\text{ita}} + \gamma\mathcal{L}_{\text{sdd}}, \tag{13}$$

where α, β, and γ are the hyper-parameters. Specifically, in the s-th incremental state, the model F_s learns novel category knowledge by optimizing $\mathcal{L}_{\text{ce}}$. Besides, we utilize $\mathcal{L}_{\text{ita}}$ to transfer the learned geometric characteristics. To alleviate the catastrophic forgetting caused by the semantic knowledge shift, we leverage the $\mathcal{L}_{\text{sdd}}$ to achieve an exact knowledge distillation. Since the model struggles to classify inputs in the early stages of novel class learning, rendering SDD decomposition ineffective, we incorporate $\mathcal{L}_{\text{kd}}$ into the objective function to provide a foundational estimate and ensure model convergence.

Final Inference At the end of each incremental state training, the trained model F_s will be utilized to predict the semantic categories $c \in \{\mathcal{C}_p \cup \mathcal{C}_s\}$.

Experiments

Datasets and Setup

Datasets Two representative 3D semantic segmentation datasets are leveraged to conduct comparison experiments.

- **Stanford 3D Indoor Spaces (S3DIS)** (Armeni et al. 2016) is a large-scale benchmark for indoor scene understanding. It contains 3D scans of six areas (e.g. office, conference room, lobby) including 271 rooms and each room contains about over 106 points. The annotation of S3DIS points corresponds to 13 semantic classes. For each point, we adopt its XYZ coordinates and RGB colors as the input.
- **ScanNet** (Dai et al. 2017) is another challenging 3D segmentation benchmark. It includes 1,513 point clouds of scans collected from 707 unique indoor scenes and consists of 21 different semantic classes.

Following the proposed 3D-CISS methods (Yang et al. 2023a,b), we adopt mean Intersection-over-Union (mIoU) as the evaluation metric for comparison experiments.

Setup In order to evaluate the performance of LAGD, we apply short-term and long-term 3D-CISS settings. Specifically, in the short-term setting, we follow the setting in GUAT (Yang et al. 2023a) which includes $\mathcal{C}_{novel} = \{5, 3, 1\}$ in both S3DIS and ScanNet datasets, and each has a random state split $\mathcal{M}^0$ and an alphabetical split $\mathcal{M}^1$. Besides, we set $\mathcal{C}_{novel} = \{1, 1, \cdots, 1\}$, which consists of total 11 steps on ScanNet to evaluate model performance under long-term incremental states. Comparison results are summarized in Tables 1, 2, and 3, respectively.

Implementation Details and Baselines

Implementation Details We adopt PointNet++ (Qi et al. 2017b) to obtain the semantic feature from the point cloud input. Specifically, for S3DIS, we divide the rooms of S3DIS with a sliding window (Wang et al. 2019c) into 7,547 $1m \times 1m$ blocks. Each block is randomly sampled with 4096 points as the input data. In the training process, we set the batch size as 32, the learning rate as 0.001, and the hyper-parameters $\{\alpha, \beta, \gamma\}$ are set as $\{10, 1, 1\}$. Each incremental state is trained 32 epochs, and utilizes the Adam optimizer (Kingma and Ba 2014) with initial 0.001 learning rate. For ScanNet, we will first generate $1.5m \times 1.5m$ block in each epoch and then randomly sample 4096 points from it as the input. In the training process, we set the batch size as 32, the learning rate is 0.001, and the hyper-parameters $\{\alpha, \beta, \gamma\}$ are set as $\{10, 1, 1\}$, and each incremental state is trained 300 epochs. We leverage Adam optimizer with an initial 0.001 learning rate and the decay factor is set as 0.7 for 50 epochs. Experiments are conducted on Tesla-V100.

Baselines We compare LAGD with the following baselines: **1) Fine-Tune (FT):** The current model F_s is initialized by the last state model F_{s-1}, only the cross-entropy loss is used to optimize the network; **2) Joint-training:** Under each state, the current labels $\mathcal{C}_{\text{s}}$ and previous $\mathcal{C}_{\text{p}}$ are both accessed to train the model F_{s}; **3) Freeze and Add (F&A):** We freeze the backbone and only optimize the classifier added in each state; **4) LwF (Li and Hoiem 2017):** LwF utilizes the calculated distillation loss to maintain the previously learned knowledge; **5) EWC (Kirkpatrick et al. 2017):** EWC mitigates catastrophic forgetting by adding a regularization term to the loss function, preserving the important parameters of previous tasks while learning new ones; and **6) LGKD (Yang et al. 2023b):** LGKD leverages the ground-truth label to construct a reliable class correspondence and reduce the confusion in knowledge distillation.

Quantitative Results

Short-term Experiments on S3DIS The short-term experiments on S3DIS are summarized in Table 1. Our method outperforms all baselines in this setting. Particularly, without any constraint, FT forgets the oldest learned knowledge after incremental learning. Besides, a naive freeze backbone significantly hinders the model from learning new categories. Compared to suboptimal baselines, our method obtains a state-of-the-art (sota) performance with a $2\% \sim 9\%$ increase on average under the split $\mathcal{M}^0$ setting. Notably, changing the order of incremental categories has an impact on the final performance, especially in the $\mathcal{C}_{novel} = 1$ scenario. Compared with the $\mathcal{M}^0$ split, our method decreased by 8.1% and just achieved sub-optimal performance from the average of all categories. However, under the other settings with $\mathcal{M}^1$ split, our method obtains sota performance with an average $0.2\% \sim 7.6\%$ increase.

Short-term Experiments on ScanNet We summarize the short-term experiments on ScanNet in Table 2. Without any measures to alleviate catastrophic forgetting, FT almost overwrites the parameters of the model to fit the novel categories. In $\mathcal{C}_{novel} = 5$ setting, our method achieves sota performance with a $3.9\% \sim 27.3\%$ increase under the split $\mathcal{M}^0$, and with a $3.1\% \sim 27.2\%$ increase under the split $\mathcal{M}^1$. In $\mathcal{C}_{novel} = 3$ and $\mathcal{C}_{novel} = 1$ settings, our method is more flexible in learning novel categories while retaining previous knowledge. On the contrary, F&A and LGKD intensively maintain the previously learned knowledge while failing to incrementally learn novel categories.

Method	$\mathcal{C}_{novel}=5$						$\mathcal{C}_{novel}=3$						$\mathcal{C}_{novel}=1$					
	$\mathcal{M}^0$			$\mathcal{M}^1$			$\mathcal{M}^0$			$\mathcal{M}^1$			$\mathcal{M}^0$			$\mathcal{M}^1$		
	0-7	8-12	all	0-7	8-12	all	0-9	10-12	all	0-9	10-12	all	0-11	12	all	0-11	12	all
BT	51.8	-	-	41.8	-	-	48.1	-	-	43.2	-	-	46.8	-	-	46.2	-	-
FT	7.0	38.7	19.2	1.7	58.3	23.5	3.0	26.6	8.5	1.5	48.1	12.2	3.8	**40.3**	6.6	11.3	1.1	10.5
F&A	45.9	9.3	31.9	39.4	38.1	38.9	45.0	7.2	36.3	44.6	33.1	42.0	46.2	8.84	43.3	**44.9**	0.0	**41.5**
LwF†	43.7	**43.2**	43.5	32.6	**61.6**	43.7	46.2	**46.0**	46.1	**46.1**	44.3	45.9	41.6	35.8	41.1	27.6	20.2	27.0
EWC†	20.6	36.2	26.6	14.1	54.7	29.1	36.8	38.4	37.2	22.1	60.3	30.9	29.7	40.0	30.5	28.1	**37.1**	28.8
LGKD†	25.5	15.9	21.8	25.0	38.8	30.3	19.6	25.0	20.8	18.4	48.5	25.3	12.9	8.5	12.6	0.2	4.2	0.5
LAGD	**51.2**	42.4	**47.8**	**40.7**	57.7	**47.3**	**48.6**	41.9	**47.1**	41.7	**60.4**	**46.0**	**46.3**	35.7	**45.4**	37.5	35.4	37.3
Joint	51.5	45.0	49.0	49.8	41.2	46.5	49.6	47.1	49.0	47.9	41.7	46.5	49.8	39.1	49.0	47.0	40.3	46.5

Table 1: The mIoU (%) of 3D class-incremental segmentation on the S3DIS dataset with $\mathcal{M}^0$ and $\mathcal{M}^1$ split. Except for Joint training, **black bold** denotes the highest results. † means the methods are reproduced for 3D incremental semantic segmentation.

Method	$\mathcal{C}_{novel}=5$						$\mathcal{C}_{novel}=3$						$\mathcal{C}_{novel}=1$					
	$\mathcal{M}^0$			$\mathcal{M}^1$			$\mathcal{M}^0$			$\mathcal{M}^1$			$\mathcal{M}^0$			$\mathcal{M}^1$		
	0-14	15-19	all	0-14	15-19	all	0-16	17-19	all	0-16	17-19	all	0-18	19	all	0-18	19	all
BT	45.2	-	-	44.5	-	-	46.0	-	-	45.7	-	-	47.2	-	-	47.4	-	-
FT	0.0	45.5	11.4	0.0	41.6	10.4	0.0	40.7	6.1	0.0	36.0	5.4	1.1	27.4	2.4	0.6	28.3	2.0
F&A	40.6	5.2	31.7	42.2	0.9	31.9	42.3	0.9	36.0	41.6	1.1	35.6	44.3	1.7	42.2	45.6	28.9	44.8
LwF†	21.5	6.7	17.8	22.9	8.7	19.4	24.0	11.0	22.1	20.7	16.3	20.0	23.9	35.4	24.5	25.7	33.8	26.2
EWC†	7.3	11.6	8.3	0.0	36.3	9.1	3.8	18.5	6.0	5.3	9.9	6.0	3.6	30.7	5.0	2.0	29.5	3.4
LGKD†	**41.4**	0.0	31.0	**44.2**	0.0	33.2	**44.6**	0.0	**48.0**	**43.0**	0.0	**36.5**	**45.6**	0.0	**43.3**	**47.1**	0.0	**44.8**
LAGD	30.7	**50.3**	**35.6**	31.0	**52.3**	**36.3**	33.3	**46.7**	35.3	31.9	**53.2**	35.1	34.1	**38.0**	34.3	35.7	**40.2**	36.0
Joint	50.9	60.0	53.2	47.9	55.1	49.7	51.7	61.2	53.2	48.6	55.9	49.7	53.2	53.1	53.2	49.5	53.7	49.7

Table 2: The mIoU (%) of 3D class-incremental segmentation on the ScanNet dataset with $\mathcal{M}^0$ and $\mathcal{M}^1$ split. Except for Joint training, **black bold** denotes the highest results. † means the methods are reproduced for 3D incremental semantic segmentation.

Incremental State s	0	1	2	3	4	5	6	7	8	9	10	Avg.	$\Delta(\%)$
Num of Classes	10	11	12	13	14	15	16	17	18	19	20		
BT	48.5	-	-	-	-	-	-	-	-	-	-	-	-
FT	48.5	21.7	9.4	4.3	1.0	0.4	0.3	0.2	0.3	0.2	1.5	8.0	⇓ 44.2
F&A	48.5	44.4	40.7	36.1	34.8	34.6	35.4	34.6	33.0	29.2	28.2	36.3	⇓ 15.9
LwF†	48.4	42.4	37.5	32.7	32.2	26.4	26.4	24.0	20.4	20.7	14.6	29.6	⇓ 22.6
EWC†	49.3	35.0	27.4	15.9	15.8	11.2	10.6	7.1	4.9	6.4	6.1	15.8	⇓ 36.4
LGKD†	48.7	45.9	43.5	41.5	38.6	37.1	34.8	33.2	30.1	28.7	24.2	36.9	⇓ 15.3
LAGD	48.8	**47.7**	**47.3**	**46.5**	**46.2**	**44.7**	**43.0**	**42.3**	**42.3**	**39.9**	**31.9**	**43.7**	⇓ 8.5
Joint	51.3	51.3	51.2	51.1	51.1	53.0	52.8	53.2	52.6	53.2	53.2	52.2	-

Table 3: The mIoU(%) of each step on the **ScanNetv2** datasets when the incremental task is **10-1.** Except for Joint training, **black bold** denotes the highest results. † means the methods are reproduced for 3D incremental semantic segmentation.

Long-term Experiments Considering the short-term experiments fail to evaluate the model performance in the long-term scenario, we implemented a long-term experiment on the ScanNet dataset with a total of 11 incremental states. As shown in Table 3, the performance of FT degrades dramatically starting from the first incremental state. On the contrary, our method shows its potential in the long-term incremental learning process, which consistently outperforms baselines in each incremental state and finally achieves sota performance with an improvement of $6.8\% \sim 35.7\%$.

Segmentation Error Analysis

We visualize the segmentation error in Figure 3. As can be seen, in the setting of $\mathcal{C}_{novel} = 5$ on S3DIS ($\mathcal{M}^0$), our method possesses the ability to make a correct segmentation while other baselines fail to predict reasonable results after

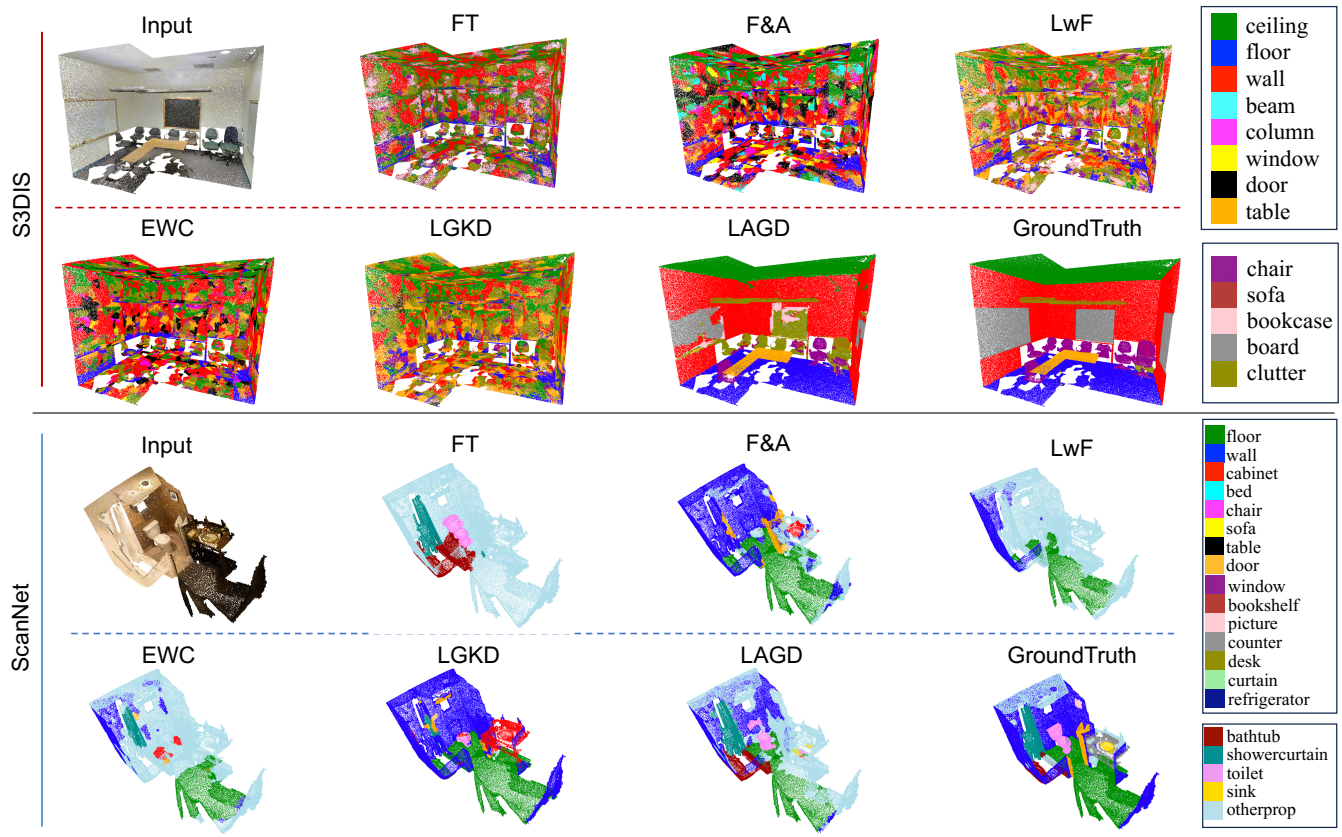

Figure 3: Error Segmentation Analysis of baselines with LAGD on S3DIS and ScanNet datasets of $\mathcal{C}_{novel} = 5$ in $\mathcal{M}^0$ setting. The base and novel classes of different datasets are explained in the discrete legend.

$\mathcal{L}_{\text{kd}}$	$\mathcal{L}_{\text{ita}}$	$\mathcal{L}_{\text{sdd}}$	S3DIS ($\mathcal{M}^0$)			ScanNet ($\mathcal{M}^1$)		
			0-7	8-12	all	0-14	15-19	all
✓	×	×	43.7	**43.2**	43.5	22.9	8.7	19.4
✓	✓	×	42.4	40.5	41.4	26.6	50.5	32.6
✓	×	✓	49.4	43.0	46.9	23.4	47.1	29.3
✓	✓	✓	**51.2**	42.4	**47.8**	**31.0**	**52.3**	**36.3**

Table 4: Ablation experiments on the S3DIS ($\mathcal{M}^0$) and ScanNet ($\mathcal{M}^1$) datasets with the $\mathcal{C}_{novel} = 5$ setting.

incremental segmentic learning. In the setting of $\mathcal{C}_{novel} = 5$ on ScanNet ($\mathcal{M}^1$), compared to baselines, our method enjoys superiority on the fine-grained items segmentation.

Ablation Study

In order to illustrate the effectiveness of each module proposed, we conducted ablation studies on the $\mathcal{C}_{novel} = 5$ on the S3DIS dataset with $\mathcal{M}^0$ split, and on the ScanNet dataset with $\mathcal{M}^1$ split. The experiment results are shown in Table 4 where ✓ and × denote with or without the corresponding module. Specifically, for S3DIS dataset, $\mathcal{L}_{\text{kd}}$ maintains the parameters of the model without changing significantly. $\mathcal{L}_{\text{sdd}}$ effectively reduces the semantic knowledge shift and thus improves the performance of the old categories. $\mathcal{L}_{\text{ita}}$ further enhances the model and achieves sota performance. For ScanNet dataset, catastrophic forgetting happens with only $\mathcal{L}_{\text{kd}}$ to maintain previously learned knowledge. Both $\mathcal{L}_{\text{ita}}$ and $\mathcal{L}_{\text{sdd}}$ are in favor of overcoming forgetting while efficiently learning novel knowledge. With the cooperation of $\mathcal{L}_{\text{ita}}$ and $\mathcal{L}_{\text{sdd}}$, the model achieves sota performance.

Conclusion

In this paper, we explored incremental semantic segmentation in the context of point clouds and proposed LAGD, a unified framework that alleviates catastrophic forgetting from the perspective of capturing local 3D geometric characteristics and mitigates global semantic logits. Specifically, we developed Interaction Topological-aware Alignment (ITA) to model the state-specific knowledge and capture the local geometric characteristics that benefit the transfer of the learned 3D representation knowledge. Considering the catastrophic forgetting partially caused by the global semantic logits shift, from the intuitive idea *birds of a feature flock together*, we proposed the Semantic-knowledge Deconstruction Distillation (SDD) that deconstructs the logits score into positive and negative parts which are distilled separately, achieving a higher performance. With the cooperation of ITA and SDD, LAGD obtained state-of-the-art results, and extensive experiments demonstrate its superiority.

Acknowledgments

This research was partly supported by the National Edge AI Hub for Real Data: Edge Intelligence for Cyber disturbances and Data Quality (EPSRC, EP/Y028813/1).

References

Armeni, I.; Sener, O.; Zamir, A. R.; Jiang, H.; Brilakis, I.; Fischer, M.; and Savarese, S. 2016. 3d semantic parsing of large-scale indoor spaces. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1534–1543.

Baek, D.; Oh, Y.; Lee, S.; Lee, J.; and Ham, B. 2022. Decomposed knowledge distillation for class-incremental semantic segmentation. *Advances in Neural Information Processing Systems*, 35: 10380–10392.

Chowdhury, T.; Cheraghian, A.; Ramasinghe, S.; Ahmadi, S.; Saberi, M.; and Rahman, S. 2022. Few-shot class-incremental learning for 3d point cloud objects. In *European Conference on Computer Vision*, 204–220. Springer.

Choy, C.; Gwak, J.; and Savarese, S. 2019. 4d spatio-temporal convnets: Minkowski convolutional neural networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 3075–3084.

Dai, A.; Chang, A. X.; Savva, M.; Halber, M.; Funkhouser, T.; and Nießner, M. 2017. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 5828–5839.

Dong, J.; Cong, Y.; Sun, G.; Ma, B.; and Wang, L. 2021. I3dol: Incremental 3d object learning without catastrophic forgetting. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, 6066–6074.

Feng, D.; Haase-Schütz, C.; Rosenbaum, L.; Hertlein, H.; Glaeser, C.; Timm, F.; Wiesbeck, W.; and Dietmayer, K. 2020. Deep multi-modal object detection and semantic segmentation for autonomous driving: Datasets, methods, and challenges. *IEEE Transactions on Intelligent Transportation Systems*, 22(3): 1341–1360.

Graham, B.; Engelcke, M.; and Van Der Maaten, L. 2018. 3d semantic segmentation with submanifold sparse convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 9224–9232.

Guo, Y.; Wang, H.; Hu, Q.; Liu, H.; Liu, L.; and Bennamoun, M. 2020. Deep learning for 3d point clouds: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 43(12): 4338–4364.

Hinton, G.; Vinyals, O.; and Dean, J. 2015. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*.

Huang, H.; Lin, L.; Tong, R.; Hu, H.; Zhang, Q.; Iwamoto, Y.; Han, X.; Chen, Y.-W.; and Wu, J. 2020. Unet 3+: A full-scale connected unet for medical image segmentation. In *ICASSP 2020-2020 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, 1055–1059. IEEE.

Kemker, R.; McClure, M.; Abitino, A.; Hayes, T.; and Kanan, C. 2018. Measuring catastrophic forgetting in neural networks. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32.

Kingma, D. P.; and Ba, J. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Kirkpatrick, J.; Pascanu, R.; Rabinowitz, N.; Veness, J.; Desjardins, G.; Rusu, A. A.; Milan, K.; Quan, J.; Ramalho, T.; Grabska-Barwinska, A.; et al. 2017. Overcoming catastrophic forgetting in neural networks. *Proceedings of the national academy of sciences*, 114(13): 3521–3526.

Lai, X.; Liu, J.; Jiang, L.; Wang, L.; Zhao, H.; Liu, S.; Qi, X.; and Jia, J. 2022. Stratified transformer for 3d point cloud segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8500–8509.

Landrieu, L.; and Simonovsky, M. 2018. Large-scale point cloud semantic segmentation with superpoint graphs. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 4558–4567.

Li, Z.; and Hoiem, D. 2017. Learning without forgetting. *IEEE transactions on pattern analysis and machine intelligence*, 40(12): 2935–2947.

Liang, W.; Sun, G.; Liu, C.; Dong, J.; and Wang, K. 2023. I3DOD: Towards Incremental 3D Object Detection via Prompting. In *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 5738–5743. IEEE.

Liu, Y.; Cong, Y.; Sun, G.; Zhang, T.; Dong, J.; and Liu, H. 2021. L3DOC: Lifelong 3D object classification. *IEEE Transactions on Image Processing*, 30: 7486–7498.

Michieli, U.; and Zanuttigh, P. 2019. Incremental learning techniques for semantic segmentation. In *Proceedings of the IEEE/CVF international conference on computer vision workshops*, 0–0.

Qi, C. R.; Su, H.; Mo, K.; and Guibas, L. J. 2017a. Pointnet: Deep learning on point sets for 3d classification and segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 652–660.

Qi, C. R.; Yi, L.; Su, H.; and Guibas, L. J. 2017b. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. *Advances in neural information processing systems*, 30.

Rebuffi, S.-A.; Kolesnikov, A.; Sperl, G.; and Lampert, C. H. 2017. icarl: Incremental classifier and representation learning. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 2001–2010.

Ronneberger, O.; Fischer, P.; and Brox, T. 2015. U-net: Convolutional networks for biomedical image segmentation. In *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*, 234–241. Springer.

Shmelkov, K.; Schmid, C.; and Alahari, K. 2017. Incremental learning of object detectors without catastrophic forgetting. In *Proceedings of the IEEE international conference on computer vision*, 3400–3409.

Wang, K.; Liew, J. H.; Zou, Y.; Zhou, D.; and Feng, J. 2019a. Panet: Few-shot image semantic segmentation with prototype alignment. In *proceedings of the IEEE/CVF international conference on computer vision*, 9197–9206.

Wang, L.; Huang, Y.; Hou, Y.; Zhang, S.; and Shan, J. 2019b. Graph attention convolution for point cloud semantic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 10296–10305.

Wang, Y.; Sun, Y.; Liu, Z.; Sarma, S. E.; Bronstein, M. M.; and Solomon, J. M. 2019c. Dynamic graph cnn for learning on point clouds. *ACM Transactions on Graphics (tog)*, 38(5): 1–12.

Wu, X.; Jiang, L.; Wang, P.-S.; Liu, Z.; Liu, X.; Qiao, Y.; Ouyang, W.; He, T.; and Zhao, H. 2024. Point Transformer V3: Simpler Faster Stronger. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4840–4851.

Yang, Y.; Hayat, M.; Jin, Z.; Ren, C.; and Lei, Y. 2023a. Geometry and uncertainty-aware 3d point cloud class-incremental semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 21759–21768.

Yang, Z.; Li, R.; Ling, E.; Zhang, C.; Wang, Y.; Huang, D.; Ma, K. T.; Hur, M.; and Lin, G. 2023b. Label-guided knowledge distillation for continual semantic segmentation on 2d images and 3d point clouds. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 18601–18612.

Zhang, C.-B.; Xiao, J.-W.; Liu, X.; Chen, Y.-C.; and Cheng, M.-M. 2022. Representation compensation networks for continual semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7053–7064.

Zhang, Y.; Li, H.; Gao, Y.; Duan, H.; Huang, Y.; and Zheng, Y. 2024. Prototype Correlation Matching and Class-Relation Reasoning for Few-Shot Medical Image Segmentation. *IEEE Transactions on Medical Imaging*.

Zhao, B.; Cui, Q.; Song, R.; Qiu, Y.; and Liang, J. 2022. Decoupled knowledge distillation. In *Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition*, 11953–11962.

Zhao, H.; Jiang, L.; Jia, J.; Torr, P. H.; and Koltun, V. 2021. Point transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*, 16259–16268.

Zhao, N.; and Lee, G. H. 2022. Static-dynamic co-teaching for class-incremental 3d object detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, 3436–3445.