SCOPE: Scene-Contextualized Incremental Few-Shot 3D Segmentation

Vishal Thengane[1,2] Zhaochong An[3] Tianjin Huang[4] Son Lam Phung[2]
Abdesselam Bouzerdoum[2] Lu Yin[1] Na Zhao[5,*] Xiatian Zhu[1,*]
[1]University of Surrey, UK [2]University of Wollongong, Australia
[3]University of Copenhagen, Denmark [4]University of Exeter, UK
[5]Singapore University of Technology and Design, Singapore

Abstract

*Incremental Few-Shot (IFS) segmentation aims to learn new categories over time from only a few annotations. Although widely studied in 2D, it remains underexplored for 3D point clouds. Existing methods suffer from catastrophic forgetting or fail to learn discriminative prototypes under sparse supervision, and often overlook a key cue: novel categories frequently appear as unlabelled background in base-training scenes. We introduce **SCOPE** (**S**cene-**CO**ntextualised **P**rototype **E**nrichment), a plug-and-play background-guided prototype enrichment framework that integrates with any prototype-based 3D segmentation method. After base training, a class-agnostic segmentation model extracts high-confidence pseudo-instances from background regions to build a prototype pool. When novel classes arrive with few labelled samples, relevant background prototypes are retrieved and fused with few-shot prototypes to form enriched representations without retraining the model or adding new parameters. Experiments on ScanNet and S3DIS show that SCOPE achieves SOTA performance, improving novel-class IoU by up to 6.98% and 3.61%, and mean IoU by 2.25% and 1.70%, respectively, while maintaining low forgetting. Code is available here.*

1. Introduction

Point Cloud Segmentation (PCS) forms the foundation of embodied perception tasks in indoor environments, including robotics, autonomous driving, and AR/VR [11, 14, 27, 29, 33]. While Fully Supervised (FS) methods achieve strong performance given abundant annotations [12, 36, 52], practical deployments face two key constraints: (i) *novel categories emerge over time* as systems encounter new environments, and (ii) *only a handful of annotations* are available when such categories first appear.

Existing approaches address these challenges independently. Few-shot segmentation methods [4, 53] learn from limited examples but fail to retain previously learned knowledge. Generalised Few-Shot 3D PCS (GFS-PCS) methods [3, 40, 46] recognise both base and novel classes, yet allow only a single update and assume prior knowledge of future categories–an unrealistic constraint in open-world scenarios. In contrast, Class-Incremental 3D PCS (CI-PCS) approaches [21, 24, 39, 49] support multiple updates but require abundant supervision and degrade sharply under limited annotations. Incremental Few-Shot 3D PCS (IFS-PCS) unifies these paradigms by enabling sequential learning of new classes from few-shot examples while preserving prior knowledge; however, despite progress in 2D [8, 15], it remains underexplored for 3D point clouds. A brief comparison of these paradigms is presented in Tab. 1.

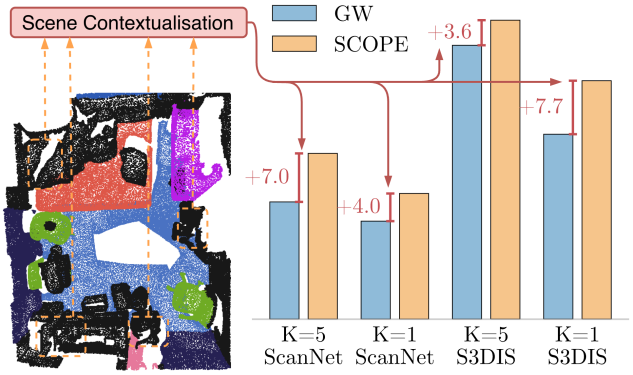

Figure 1. Compared with GW [46], SCOPE improves performance on novel classes by enriching prototypes with background context, without additional computational overhead.

Directly adopting these approaches proves suboptimal (Tab. 2), as they struggle to generalise to novel classes under scarce supervision and often forget previously learned categories due to overfitting. Conversely, GFS methods [40, 46] retain base-class knowledge but show limited adaptability when learning new categories. Similarly, a recent hyperbolic-prototype approach for IFS-PCS [37] exhibits comparable limitations and does not yet match the

*Corresponding authors: Na Zhao and Xiatian Zhu

Table 1. Comparison of PCS paradigms across future knowledge, few-shot, base-novel generalisation, and multi-step learning.

PCS Paradigms	Future Knowledge	Few-Shot	Base + Novel	Multi-Step
Fully Supervised	–	✗	–	–
Few-Shot	✗	✓	✗	✗
Generalized Few-Shot	✗	✓	✓	✗
Class-Incremental	✓	✗	✓	✓
Incremental Few-Shot	✓	✓	✓	✓

strongest GFS baselines. We attribute these limitations to the underutilisation of background regions in base scenes, which often contain object-like structures indicative of future classes. Leveraging this contextual information is therefore essential for robust IFS-PCS.

To this end, we propose **SCOPE** (***S**cene-**CO**ntextualised **P**rototype **E**nrichment*), a background-guided prototype enrichment framework. Our key observation is that background regions–typically collapsed into a single coarse label–contain object-like structures that the backbone cannot disentangle during base training but often correspond to future classes. In the IFS-PCS setting, future class identities are unknown during the base stage; thus class-specific prototypes cannot be constructed in advance. We therefore employ an off-the-shelf class-agnostic segmentation model to extract high-confidence pseudo-instances and store them in an *Instance Prototype Bank* (**IPB**), forming a reservoir of transferable object-level cues. For each novel class, aligned context is retrieved from the IPB via the *Contextual Prototype Retrieval* (**CPR**) module. Since retrieved instances vary in relevance, an *Attention-Based Prototype Enrichment* (**APE**) mechanism selectively weights and fuses them with few-shot prototypes. This produces context-aware and discriminative prototypes without modifying the backbone or introducing additional parameters, satisfying the minimal-adaptation principle of few-shot learning.

Compared with methods from related paradigms [37, 39, 40], SCOPE achieves consistent gains across benchmarks, demonstrating stronger novel-class adaptation and improved retention of prior knowledge. In summary, our contributions are threefold: (1) we propose a plug-and-play framework that mines contextual cues from base scenes and constructs an IPB using an off-the-shelf class-agnostic segmentation model with no additional computational overhead; (2) we introduce a CPR module to retrieve relevant background cues without future class knowledge and an APE mechanism for prototype enrichment; and (3) we establish new SOTA performance across multiple IFS-PCS settings on standard 3D segmentation benchmarks.

2. Related Work

2.1. 3D Scene Understanding

Fully supervised 3D semantic segmentation has been extensively studied, typically requiring dense point-wise labels [18, 23, 30, 31, 44, 52]. Early models such as PointNet [30] processed raw point clouds but struggled to capture local structures, while subsequent architectures including PointNet++ [31], DGCNN [44], and transformer-based models [52] progressively enhanced geometric and contextual reasoning. Recently, foundation-style models [10, 20, 43, 51, 54] have explored cross-modal representations and open-vocabulary objectness without explicit class labels [19, 28, 38, 48]. However, these approaches rely on large-scale annotations and fixed label spaces, limiting scalability in open-world environments.

2.2. Few-Shot 3D Segmentation

Few-shot 3D segmentation reduces reliance on large-scale annotated data. Zhao et al. [53] proposed the first few-shot 3D segmentation method, enabling recognition of unseen classes from limited labelled samples. Existing approaches primarily refine prototypes or query embeddings via non-parametric optimisation [16, 26, 42, 45, 47, 50, 55], or explicitly model support-query correlations [2, 4], as demonstrated by COSeg [2]. Extending this paradigm, generalised few-shot methods jointly model base and novel classes. Representative examples include PIFS [8], which refines prototypes via distillation; CAPL [40], which incorporates co-occurrence priors; and GW [46], which extends the paradigm to 3D via geometric cues. Tsai *et al*. [41] mine background regions to construct pseudo-class prototypes via language-guided pre-clustering. GFS-VL [3] leverages 3D vision-language models (VLMs) to effectively handle sparse novel samples. In contrast, our method learns future classes across stages without additional supervision, storing reusable background knowledge in a prototype pool.

2.3. Incremental 3D Segmentation

Incremental learning acquires new classes sequentially while mitigating catastrophic forgetting. This is typically achieved through replay [32], knowledge distillation [17], regularisation [24], or consolidation strategies [1, 7, 9, 21, 34]. Su et al. [35] and subsequent works [6, 22] adapt these strategies to PCS, demonstrating promising results but often relying on large memory buffers and repeated retraining cycles. Thengane *et al*. [39] address long-tail classes via object-count priors, yet their approach still depends on abundant supervision for stable adaptation. As these methods are primarily designed for large-scale data regimes, they remain suboptimal under few-shot conditions.

2.4. Incremental Few-Shot 3D Segmentation

Incremental Few-Shot (IFS) learning aims to learn new classes sequentially from limited supervision while retaining prior knowledge. Although extensively studied in 2D vision [8, 15, 25], its extension to 3D remains underexplored. In 3D, Sur et al. [37] introduce hyperbolic prototypes for IFS segmentation but remain behind GFS-PCS

baselines in performance. Our method addresses this gap by leveraging background context as transferable knowledge, achieving stronger performance for novel classes in the IFS-PCS setting than methods from other paradigms.

3. Methodology

3.1. Problem Formulation

The goal of the IFS-PCS is to learn a model Φ that assigns each point in a point cloud to its semantic label in a set $\boldsymbol{C}$, where $\boldsymbol{C}$ is progressively expanded using a small number of annotated samples for newly encountered classes.

GFS-PCS. Let $\boldsymbol{C}^{\mathrm{b}}$ and $\boldsymbol{C}^{\mathrm{n}}$ denote disjoint sets of base and novel classes, respectively, such that $\boldsymbol{C} = \boldsymbol{C}^{\mathrm{b}} \cup \boldsymbol{C}^{\mathrm{n}}$ and $|\boldsymbol{C}^{\mathrm{b}}| = N^b$, $|\boldsymbol{C}^{\mathrm{n}}| = N^n$, giving $N = N^b + N^n$ classes. The model is first trained on the base classes using $\mathbf{D}^{\mathrm{b}} = \{(\mathbf{X}, \boldsymbol{y}^{\mathrm{b}}) \mid \mathbf{X} \in \mathcal{X}, \boldsymbol{y}^{\mathrm{b}} \in \mathcal{Y}^{\mathrm{b}}\}$, where $\boldsymbol{y}^{\mathrm{b}}$ assigns each point a label from $\boldsymbol{C}^{\mathrm{b}}$ or background (-1). Novel classes are introduced in a single batch, each with K labelled support examples, forming the few-shot dataset $\mathbf{D}^{\mathrm{n}} = \left\{\{(\mathbf{X}, \boldsymbol{y}_k^{\mathrm{c}})\}_{k=1}^{K}\right\}_{c \in \boldsymbol{C}^{\mathrm{n}}}$. Each support point cloud $\mathbf{X}_k \in \mathbb{R}^{M \times d_0}$ contains M points with d_0-dimensional features, and $\boldsymbol{y}_k^{\mathrm{c}}$ provides point-wise annotations for class c with all other points labelled as background. Evaluation is performed on $\mathbf{D}^{\mathrm{test}}$ across all classes in $\boldsymbol{C}$.

IFS-PCS. In contrast, in the IFS-PCS setting, novel classes are registered over multiple stages, reflecting a naturally evolving environment. Let $\boldsymbol{C}^{\mathrm{t}}$ denote the set of semantic classes known after stage t, where each stage expands the model's output space. During learning, the model sequentially receives datasets $\{\mathbf{D}^0, \ldots, \mathbf{D}^{\mathrm{T}}\}$, where $\mathbf{D}^{\mathrm{t}} = \{(\mathbf{X}, \boldsymbol{y}^{\mathrm{t}}) \mid \mathbf{X} \in \mathcal{X}, \boldsymbol{y}^{\mathrm{t}} \in \mathcal{Y}^{\mathrm{t}}\}$. In each $\mathbf{D}^{\mathrm{t}}$, $\mathbf{X} \in \mathbb{R}^{M \times d_0}$ denotes a point cloud with M points and d_0-dimensional features, while $\boldsymbol{y}^{\mathrm{t}} \in \mathbb{R}^M$ assigns one class index per point, $\mathcal{Y}^{\mathrm{t}} \subseteq \boldsymbol{C}^{\mathrm{t}}$. The dataset $\mathbf{D}^0$ contains abundant base-class examples, analogous to $\mathbf{D}^{\mathrm{b}}$ in GFS-PCS. Subsequent stages introduce novel classes incrementally via few-shot support datasets, $\mathbf{D}^{\mathrm{n}} = \bigcup_{t=1}^{T} \mathbf{D}^{\mathrm{t}}$, each extending the known class set. Although the same point cloud $\mathbf{X}$ may reappear across stages, only labels corresponding to the currently known classes $\boldsymbol{C}^{\mathrm{t}}$ are available, *i.e.*, supervision is restricted to classes introduced at that stage[1].

At $t = 0$, the model is trained on the large base dataset $\mathbf{D}^0 = \{(\mathbf{X}_i, \boldsymbol{y}_i^0)\}_{i=1}^{|\mathbf{D}^0|}$ with labels drawn from $\boldsymbol{C}^0$. For each subsequent stage $t \geq 1$, the model receives a small few-shot support set $\mathbf{D}^{\mathrm{t}}$ introducing previously unseen classes from $\boldsymbol{C}^{\mathrm{t}}$. Specifically, each novel class $c \in \boldsymbol{C}^{\mathrm{t}}$ is supported by exactly K labelled point clouds:

$$\left\{(\mathbf{X}_k, \boldsymbol{y}_k^{\mathrm{c}})\right\}_{k=1}^{K}, \quad \mathbf{X}_k \in \mathbb{R}^{M \times d_0},\ \boldsymbol{y}_k^{\mathrm{c}} \in \mathbb{R}^M. \tag{1}$$

[1]Henceforth, we use $t{=}0$ and $t{=}b$ interchangeably to denote the 0-th stage of the IFS-PCS setting.

Each $\boldsymbol{y}_k^{\mathrm{c}}$ assigns label c to points of that class and background to all others, *i.e.*, $\boldsymbol{y}_k^{\mathrm{c}}(x) \in \{c, -1\}, \forall\, x \in \mathbf{X}_k$, where -1 denotes background or unknown regions. Evaluation is conducted on a test set $\mathbf{D}^{\mathrm{test}}$ covering all categories observed so far, *i.e.*, $\mathbf{Y}^{\mathrm{test}} \subseteq \boldsymbol{C}^{\mathrm{t}'}$, where $\boldsymbol{C}^{\mathrm{t}'} = \bigcup_{i=0}^{t} \boldsymbol{C}^{\mathrm{i}}$.

3.2. Method Overview

Our Idea. The core idea of SCOPE is to exploit contextual cues in base scenes to improve incremental few-shot point cloud segmentation. Although base training focuses only on known categories, background regions contain rich geometric and semantic signals corresponding to objects that later emerge as novel classes. However, directly extracting background features with the encoder Φ yields coarse, non-discriminative embeddings, as all unseen regions are collapsed into a single background class. To address this, we incorporate a class-agnostic segmentation model Θ [19] to detect object-like regions in the background, mining latent structure to produce transferable contextual cues that enrich few-shot representations in later stages. This allows adaptation to new classes without introducing additional parameters or retraining the encoder, effectively bridging few-shot generalisation and incremental learning in PCS.

Overview. SCOPE is implemented as a three-stage pipeline that operationalises this idea through background mining and prototype refinement. As illustrated in Fig. 2, the framework consists of: (i) Base Training, (ii) Scene Contextualisation, and (iii) Incremental Class Registration. During *Base Training*, the encoder Φ is learned together with base prototypes $\mathbf{P}^{\mathrm{b}}$ on fully labelled base data. The *Scene Contextualisation* stage applies the class-agnostic model Θ to background regions and converts the resulting masks into an *Instance Prototype Bank* (IPB), denoted $\mathcal{P}$. During *Incremental Class Registration*, each novel class $c \in \boldsymbol{C}^{\mathrm{t}}$ is initialised with a few-shot prototype $\boldsymbol{p}^{\mathrm{c}}$. Our *Contextual Prototype Retrieval* (CPR) module retrieves a class-specific subset $\mathcal{B}^{\mathrm{c}} \subset \mathcal{P}$ of semantically aligned prototypes, which are fused via the *Attention-based Prototype Enrichment* (APE) module to obtain enriched representations $\tilde{\boldsymbol{p}}^{\mathrm{c}}$. This enables expansion to new categories without modifying the encoder or introducing additional learnable parameters. The framework remains fully plug-and-play and can be seamlessly applied to any prototype-based segmenter without altering the backbone or training schedule.

3.3. Base Training

During the base stage, the model learns to encode geometric and semantic cues from the fully labelled base dataset $\mathbf{D}^{\mathrm{b}}$, building a discriminative embedding space. It jointly trains a backbone network Φ' and a projection head $\mathcal{H}$, forming the overall encoder $\Phi = \mathcal{H} \circ \Phi'$. Given an input point cloud

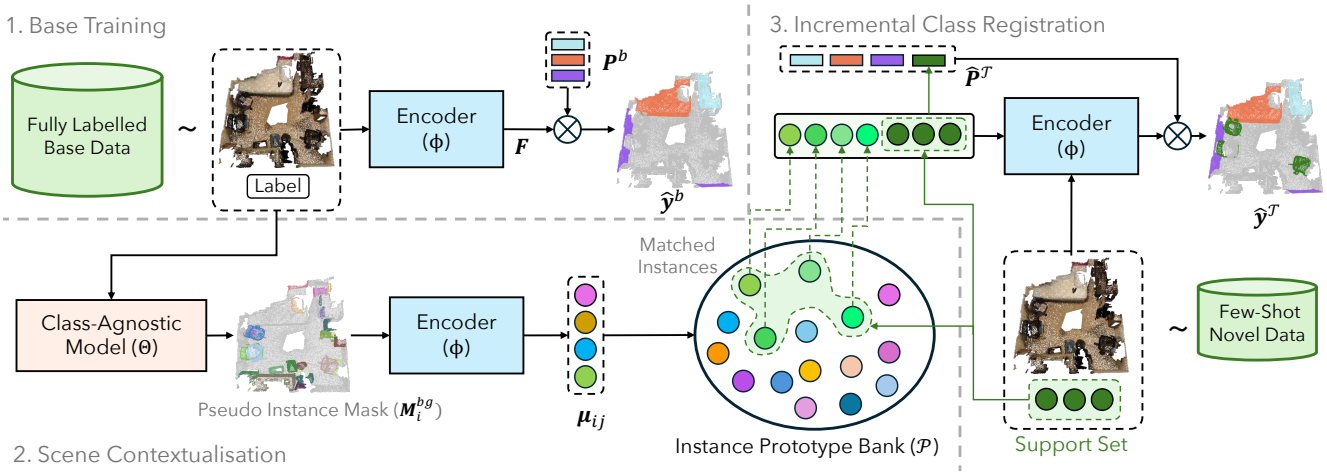

Figure 2. **Overview of SCOPE**. The framework comprises three stages: (1) *Base Training*, where the encoder Φ is trained on labelled base data; (2) *Scene Contextualisation*, which extracts background regions to build an Instance Prototype Bank (IPB); and (3) *Incremental Class Registration*, where retrieved prototypes are fused with few-shot prototypes via attention to yield refined novel-class representations.

$\mathbf{X}_i$, the encoder produces point-wise embeddings:

$$\mathbf{F}_i = \Phi(\mathbf{X}_i) = \mathcal{H}\big(\Phi'(\mathbf{X}_i)\big), \quad \mathbf{F}_i \in \mathbb{R}^{M\times D}, \tag{2}$$

where M denotes the number of points in $\mathbf{X}_i$ and D is the embedding dimension. To associate these embeddings with semantic meaning, the model learns a set of base prototypes $\mathbf{P}^{\mathrm{b}} = \{\boldsymbol{p}^{\mathrm{c}}\}_{c\in\boldsymbol{C}^{\mathrm{b}}} \in \mathbb{R}^{N^b\times D}$, where each $\boldsymbol{p}^c$ acts as a class-level representative in the embedding space. Per-point logits are obtained by measuring similarity between embeddings and prototypes:

$$\hat{\boldsymbol{y}}_i^{\mathrm{b}} = \underset{c\in\boldsymbol{C}^{\mathrm{b}}}{\arg\max}\,\big(\mathbf{F}_i\cdot(\mathbf{P}^{\mathrm{b}})^{\top}\big), \quad \hat{\boldsymbol{y}}_i^{\mathrm{b}} \in \mathbb{R}^{M\times N^b}. \tag{3}$$

Unlike prior methods that discard background features, our approach retains and reuses them to build transferable contextual prototypes for incremental adaptation.

3.4. Scene Contextualisation

After base training, the model predicts labels only from the base class set $\boldsymbol{C}^{\mathrm{b}}$, assigning all other points to the background. However, these background points often belong to unlabelled objects or structural patterns that may correspond to future novel classes. Because all unseen content is collapsed into a single background label, the encoder Φ cannot delineate object boundaries, resulting in coarse and non-discriminative representations. To address this limitation, we introduce the *Scene Contextualisation Module* (**SCM**), which mines object-like regions using an off-the-shelf class-agnostic model [19]. The model is applied once offline and subsequently discarded, and the resulting regions are converted into transferable prototypes.

Pseudo-Mask Generation. For each input scene $\mathbf{X}_i$ from the base dataset $\mathbf{D}^{\mathrm{b}}$, the class-agnostic segmentation model Θ predicts, in an offline manner, a set of pseudo-instance masks with associated confidence scores:

$$\Theta(\mathbf{X}_i) = \big\{\,(\hat{\mathbf{M}}_{i,j},\, \boldsymbol{s}_{i,j})\,\big\}_{j=1}^{Q_i}, \tag{4}$$

where $\hat{\mathbf{M}}_{i,j} \in \{0,1\}^M$ denotes the j-th binary mask predicted for scene $\mathbf{X}_i$, $\boldsymbol{s}_{i,j} \in [0,1]$ represents its corresponding confidence score, and Q_i is the total number of predicted masks for that scene. Only masks corresponding to background points and having confidence above a threshold τ are retained:

$$\mathbf{M}_i^{\mathrm{bg}} = \big\{\,\hat{\mathbf{M}}_{i,j} \;\big|\; \hat{\mathbf{M}}_{i,j} \subseteq \mathbf{X}_i[\boldsymbol{y}_i^{\mathrm{b}} = -1],\; \boldsymbol{s}_{i,j} > \tau\,\big\}, \tag{5}$$

where $\mathbf{X}_i[\boldsymbol{y}_i^{\mathrm{b}} = -1]$ denotes the subset of points in scene $\mathbf{X}_i$ labelled as background in the base dataset. Each resulting background mask $\mathbf{M}_i^{\mathrm{bg}}$ delineates high-confidence object-like regions within the background, which are treated as potential instances of unseen or novel classes.

Instance Prototype Bank (IPB). The IPB aims to construct robust feature prototypes from the pseudo background masks $\mathbf{M}_i^{\mathrm{bg}}$ predicted by the class-agnostic model Θ, providing transferable cues that facilitate subsequent novel-class registration. Each input scene $\mathbf{X}_i$ is processed by the encoder Φ to extract point-wise features $\mathbf{F}_i$ (Eq. (2)). These features are then aggregated into instance-level prototypes for each pseudo mask $\hat{\mathbf{M}}_{i,j} \in \mathbf{M}_i^{\mathrm{bg}}$ using a masked average-pooling operator $\mathcal{F}_{\mathrm{Pool}}$:

$$\boldsymbol{\mu}_{i,j} = \mathcal{F}_{\mathrm{Pool}}(\mathbf{F}_i,\, \hat{\mathbf{M}}_{i,j}), \qquad \hat{\mathbf{M}}_{i,j} \in \{0,1\}^M. \tag{6}$$

The resulting vector $\boldsymbol{\mu}_{i,j} \in \mathbb{R}^D$ serves as the *instance prototype* for the j-th pseudo background region of scene i. Since the novel classes are unknown during base training,

all such prototypes are collected across the full dataset to form the *Instance Prototype Bank* (IPB):

$$\mathcal{P} = \bigcup_i \bigcup_j \{\, \boldsymbol{\mu}_{i,j} \,\}. \tag{7}$$

For notational convenience, we collapse the scene-instance indices (i, j) into a single bank index b, and denote the IPB elements as $\{\boldsymbol{\mu}_b\}_{b=1}^{|\mathcal{P}|}$. This bank provides a rich reservoir of object-like background patterns, enabling initialisation and refinement of novel-class representations during subsequent incremental stages. Importantly, the IPB is constructed once after base training and then frozen throughout all incremental stages, introducing no additional optimisation or memory overhead during incremental adaptation.

3.5. Incremental Class Registration

At each incremental stage t ($t \geq 1$), the few-shot dataset $\mathbf{D}^{\mathrm{t}}$ introduces a set of novel categories $\boldsymbol{C}^{\mathrm{t}}$, each provided with K annotated support point clouds $\{(\mathbf{X}_k,\, \boldsymbol{y}_k^{\mathrm{c}})\}_{k=1}^{K}$. For every novel class $c \in \boldsymbol{C}^{\mathrm{t}}$, an initial class prototype $\boldsymbol{p}^{\mathrm{c}}$ is obtained by extracting encoder features from the support examples and aggregating those corresponding to class c:

$$\boldsymbol{p}^{\mathrm{c}} = \frac{1}{K} \sum_{k=1}^{K} \mathcal{F}_{\mathrm{Pool}}\Big(\mathbf{F}_k,\, \mathbf{1}[\boldsymbol{y}_k{=}c]\Big), \tag{8}$$

where $\mathbf{F}_k$ denotes the features of the k-th support example, and $\mathbf{1}[\boldsymbol{y}_k{=}c]$ is a binary mask selecting points labelled as class c. The resulting set of few-shot prototypes for stage t is denoted as $\mathbf{P}^{\mathrm{t}} = \{\boldsymbol{p}^{\mathrm{c}} \mid c \in \boldsymbol{C}^{\mathrm{t}}\} \in \mathbb{R}^{|\boldsymbol{C}^{\mathrm{t}}| \times D}$, and provides the initial representation for each novel category before contextual refinement.

Contextual Prototype Retrieval (CPR). Few-shot prototypes $\boldsymbol{p}^{\mathrm{c}}$, derived from limited annotations, often lack semantic diversity and do not fully capture the underlying class structure. To compensate for this limitation, the *Contextual Prototype Retrieval* (CPR) module identifies semantically aligned background embeddings from the IPB ($\mathcal{P}$).

For each novel class $c \in \boldsymbol{C}^{\mathrm{t}}$, we compute the cosine similarity between its few-shot prototype $\boldsymbol{p}^{\mathrm{c}}$ and every background prototype $\boldsymbol{\mu}_b$ in the IPB:

$$\boldsymbol{\sigma}_b^{\mathrm{c}} = \frac{(\boldsymbol{p}^{\mathrm{c}})^{\top} \boldsymbol{\mu}_b}{\|\boldsymbol{p}^{\mathrm{c}}\|_2 \, \|\boldsymbol{\mu}_b\|_2}. \tag{9}$$

The top-R most relevant background prototypes are then selected to construct a class-specific context pool:

$$\mathcal{B}^{\mathrm{c}} = \big\{\boldsymbol{\mu}_b \;\big|\; b \in \mathrm{TopR}(\boldsymbol{\sigma}_{\mathrm{b}}^{\mathrm{c}})\big\} = \{\boldsymbol{\mu}_{\mathrm{r}}^{\mathrm{c}}\}_{\mathrm{r=1}}^{\mathrm{R}}. \tag{10}$$

This retrieval step yields a compact and semantically aligned set of background pseudo-instance prototypes that provide auxiliary structural cues for refining the initial few-shot prototypes for robust adaptation. However, not all retrieved background prototypes are equally informative; some may be noisy or lack strong objectness. Therefore, a dedicated enrichment mechanism is introduced to selectively integrate the most relevant contextual signals.

Attention-Based Prototype Enrichment (APE). During incremental stage t, the retrieved contextual set for class c is $\mathcal{B}^{\mathrm{c}} = \{\boldsymbol{\mu}_r^{\mathrm{c}}\}_{r=1}^{R}$. Given the few-shot prototype $\boldsymbol{p}^{\mathrm{c}}$ for $c \in \boldsymbol{C}^{\mathrm{t}}$, we refine it through a parameter-free, attention-driven mechanism that fuses it with the these retrieved prototypes.

Both the few-shot and contextual prototypes are ℓ_2-normalised:

$$\bar{\boldsymbol{p}}^{\mathrm{c}} = \mathcal{F}_{\mathrm{Norm}}(\boldsymbol{p}^{\mathrm{c}}), \qquad \bar{\mathcal{B}}^{\mathrm{c}} = \big\{\mathcal{F}_{\mathrm{Norm}}(\boldsymbol{\mu}_r^{\mathrm{c}}) \,\big|\, \boldsymbol{\mu}_r^{\mathrm{c}} \in \mathcal{B}^{\mathrm{c}}\big\}.$$

A scaled dot-product cross-attention operation then uses the few-shot prototype as the query and the contextual prototypes as keys and values, without introducing any learnable parameters or projection heads. This produces a set of scalar attention weights, one per retrieved prototype, quantifying its relevance to class c. The context-enhanced representation is computed as:

$$\boldsymbol{h}^{\mathrm{c}} = \sum_{r=1}^{R} \mathrm{CrossAttention}\big(\bar{\boldsymbol{p}}^{\mathrm{c}},\, \bar{\mathcal{B}}^{\mathrm{c}}\big)_r \; \bar{\boldsymbol{\mu}}_r^{\mathrm{c}}, \tag{11}$$

The enriched class prototype is obtained by combining the original few-shot prototype with the attention-weighted contextual prototypes:

$$\tilde{\boldsymbol{p}}^{\mathrm{c}} = \lambda\, \boldsymbol{p}^{\mathrm{c}} + (1 - \lambda)\, \boldsymbol{h}^{\mathrm{c}}, \qquad \lambda \in [0, 1]. \tag{12}$$

The parameter-free attention mechanism suppresses noisy context while preserving transferable structural cues for prototype refinement. The refined prototypes for all classes observed up to stage t are assembled as

$$\tilde{\mathbf{P}}^{\leq \mathrm{t}} = \big[\, \mathbf{P}^{\mathrm{b}},\, \ldots,\, \tilde{\mathbf{P}}^{\mathrm{t}} \,\big], \qquad \tilde{\mathbf{P}}^{\mathrm{t}} = \big\{\tilde{\boldsymbol{p}}^{\mathrm{c}} \;\big|\; c \in \boldsymbol{C}^{\mathrm{t}}\big\}, \tag{13}$$

where $\mathbf{P}^{\mathrm{b}}$ denotes the base prototypes and $\boldsymbol{C}^{\mathrm{t}}$ the novel classes introduced at stage t. The cumulative class set up to stage t is $\boldsymbol{C}^{\leq \mathrm{t}} = \bigcup_{i=0}^{t} \boldsymbol{C}^{\mathrm{i}}$, and the prototype matrix $\tilde{\mathbf{P}}^{\leq \mathrm{t}}$ serves as the unified classifier for segmentation. The final point-wise predictions are obtained via:

$$\hat{\boldsymbol{y}}_i^{\leq \mathrm{t}} = \underset{c \in \boldsymbol{C}^{\leq \mathrm{t}}}{\arg\max} \left(\mathbf{F}_i \cdot (\tilde{\mathbf{P}}^{\leq \mathrm{t}})^{\top}\right) \in \mathbb{R}^{M \times |\boldsymbol{C}^{\leq \mathrm{t}}|}. \tag{14}$$

4. Experiments

We evaluate SCOPE on two benchmarks against representative incremental, few-shot, and generalised few-shot baselines. We first present the experimental setup, followed by quantitative and qualitative results, ablations, and a discussion of efficiency and limitations (see ***Supp.*** for details).

Table 2. Comparison of baseline methods and SCOPE on the *ScanNet* dataset under the IFS-PCS setting with K=5 and K=1. We report mIoU, mIoU-B, mIoU-N, HM, mIoU-I, and FPP. All metrics are higher is better except FPP. Best results are highlighted in **bold**.

Methods	Venue	K=5						K=1					
		mIoU	**mIoU-B**	**mIoU-N**	**HM**	**mIOU-I**	**FPP**	**mIoU**	**mIoU-B**	**mIoU-N**	**HM**	**mIOU-I**	**FPP**
JT (Oracle)	–	45.34	48.68	36.97	42.03	–	–	45.34	48.68	36.97	42.03	–	–
Finetuning	–	0.38	0.48	0.15	0.23	12.28	34.25	0.42	0.49	0.26	0.34	10.93	40.67
LwF [24]	TPAMI'17	4.12	5.77	0.27	0.51	14.75	35.39	0.37	0.33	0.47	0.39	12.06	40.84
CLIMB-3D [39]	BMVC'25	4.23	5.92	0.02	0.04	15.98	35.25	0.02	0.00	0.06	0.00	10.36	41.16
EWC [21]	PNAS'17	3.87	5.41	0.03	0.06	17.66	35.76	2.33	4.42	0.02	0.05	16.57	35.22
GUA [49]	CVPR'23	13.43	18.01	3.14	5.35	24.59	25.55	8.97	11.69	1.59	2.80	23.75	31.87
AttMPTI [53]	CVPR'21	12.42	17.93	3.26	5.52	25.16	25.50	9.24	12.43	1.68	2.96	23.82	31.00
HIPO [37]	CVPR'25	14.95	25.34	7.44	11.50	27.63	17.60	11.94	14.63	2.91	4.86	25.94	28.31
PIFS [8]	BMVC'21	25.39	34.81	3.43	6.24	33.11	8.74	24.61	33.66	3.49	6.32	33.68	9.88
CAPL [40]	CVPR'22	31.73	39.01	14.75	21.36	34.55	-0.65	30.48	39.10	10.38	16.28	34.00	-0.74
GW [46]	ICCV'23	34.27	41.72	16.88	23.94	37.67	1.49	33.53	41.85	14.11	20.99	37.38	1.36
SCOPE (Ours)	–	**36.52**	**41.94**	**23.86**	**30.38**	**38.91**	**1.27**	**34.78**	**41.94**	**18.09**	**25.12**	**38.46**	**1.27**

Table 3. Comparison of baseline methods and SCOPE on the *S3DIS* dataset under the IFS-PCS setting with K=5 and K=1. We report mIoU, mIoU-B, mIoU-N, HM, mIoU-I, and FPP. All metrics are higher is better except FPP. Best results are highlighted in **bold**.

Methods	Venue	K=5						K=1					
		mIoU	**mIoU-B**	**mIoU-N**	**HM**	**mIOU-I**	**FPP**	**mIoU**	**mIoU-B**	**mIoU-N**	**HM**	**mIOU-I**	**FPP**
JT (Oracle)	–	73.62	81.57	64.34	71.94	–	–	73.62	81.57	64.34	71.94	–	–
Finetuning	–	2.25	3.26	1.07	1.61	24.57	65.55	0.52	0.06	1.06	0.11	22.88	75.58
LwF [24]	TPAMI'17	12.66	20.29	3.76	6.35	37.55	55.36	7.89	12.61	2.39	4.02	28.44	63.04
CLIMB-3D [39]	BMVC'25	17.71	31.37	1.76	3.32	41.72	44.27	7.24	12.77	0.79	1.49	30.98	68.40
EWC [21]	PNAS'17	19.48	35.56	0.73	1.44	44.78	40.09	13.74	20.66	1.64	3.00	37.85	54.98
GUA [49]	CVPR'23	20.18	29.30	11.43	16.45	36.61	44.53	9.84	15.41	5.57	8.19	29.52	58.42
AttMPTI [53]	CVPR'21	25.83	34.82	15.75	21.70	40.70	38.59	18.98	25.02	12.49	16.66	37.67	48.81
HIPO [37]	CVPR'25	27.73	38.00	18.36	24.76	42.01	35.96	23.34	30.38	16.34	21.25	39.80	42.57
PIFS [8]	BMVC'21	40.68	49.96	29.86	37.38	53.84	24.93	36.13	50.50	19.36	27.99	52.19	24.38
CAPL [40]	CVPR'22	55.52	73.11	35.01	47.27	63.69	0.64	49.16	73.09	21.25	32.79	60.89	0.64
GW [46]	ICCV'23	57.71	73.38	39.42	51.29	63.69	0.04	51.73	73.25	26.62	39.02	61.15	0.17
SCOPE (Ours)	–	**59.41**	**73.44**	**43.03**	**54.25**	**65.24**	**-0.03**	**55.36**	**73.39**	**34.32**	**46.73**	**63.02**	**0.02**

4.1. Experimental Setup

Datasets. We evaluate on two standardindoor benchmarks: **S3DIS** [5] and **ScanNet** [13]. S3DIS contains 272 scenes across six areas with 13 classes, while **ScanNet** comprises 1,513 scenes with 20 categories. For fair comparison, following [46], the six least-represented classes form the novel set $\boldsymbol{C}^{\mathrm{n}}$ introduced incrementally for $t{\geq}1$, while the remaining classes constitute the base set $\boldsymbol{C}^{\mathrm{b}}$, reflecting a long-tailed distribution. Additional preprocessing and incremental-stage details are provided in the ***Supp.***.

Evaluation Metrics. As incremental few-shot segmentation bridges generalised few-shot and incremental learning, we adopt a unified evaluation suite used in both paradigms [39, 46]. Following the GFS protocol, we report *mIoU-B*, *mIoU-N*, and *mIoU*, measuring segmentation performance on base, novel, and all classes, respectively. In addition, we report the harmonic mean (*HM*) of *mIoU-B* and *mIoU-N* to assess the balance between base and novel recognition. From the incremental perspective, we compute the average incremental mIoU (*mIoU-I*), defined as the average *mIoU* across all stages, and the forgetting percentage points (*FPP*), measured as the drop in *mIoU-B* from the base stage (t=0) to the final stage T.

Implementation. For the base stage, all baselines adopt the training protocol of Xu *et al*. [46]. During incremental stages, the backbone remains frozen, and only class-specific prototypes are updated from few-shot support samples. Additional implementation details are provided in the ***Supp.***.

4.2. Results

We evaluate SCOPE on S3DIS and ScanNet against representative methods from multiple paradigms. Quantitative results are reported in Tabs. 2 and 3, with per-task performance in Fig. 3 and qualitative comparisons in Fig. 4.

Results on ScanNet. Tab. 2 presents results on ScanNet

under both K=5 and K=1 settings. SCOPE consistently outperforms all baselines across metrics in both regimes. Compared with the GFS-PCS baselines CAPL and GW under K=5, our method substantially improves novel-class recognition, increasing mIoU-N from 14.75% and 16.88% to **23.86%**, and HM from 21.36% and 23.94% to **30.86%**. mIoU-I also improves from 37.67% (GW) to **38.91%**, indicating stronger stability across incremental stages. Under the more challenging K=1 setting, SCOPE maintains clear advantages, surpassing GW by **+3.98** and **+4.13** percentage points in mIoU-N and HM, respectively. Relative to the recent IFS-PCS method HIPO, our framework achieves large gains of **+16.43**, **+18.88**, and **+11.28** in mIoU-N, HM, and mIoU-I under K=5, with similarly consistent improvements of **+15.17**, **+20.26**, and **+12.52** under K=1.

Results on S3DIS. Tab. 3 reports results on S3DIS under K=5 and K=1. Consistent with ScanNet, SCOPE achieves state-of-the-art performance while exhibiting minimal forgetting. Under K=5, it surpasses the strongest GFS-PCS baselines (GW and CAPL), improving mIoU-N from 35.01% and 39.42% to **43.03%**, HM from 47.27% and 51.29% to **54.25%**, and mIoU-I from 63.69% to **65.24%**. In the K=1 setting, SCOPE remains robust, achieving **34.32%** mIoU-N and **46.73%** HM, substantially outperforming GW (26.62% and 39.02%). Compared with HIPO, our approach delivers significant gains of **+24.67** (mIoU-N), **+29.49** (HM), and **+23.23** (mIoU-I) under K=5, with further improvements of **+17.98** (mIoU-N) and **+23.22** (mIoU-I) under K=1.

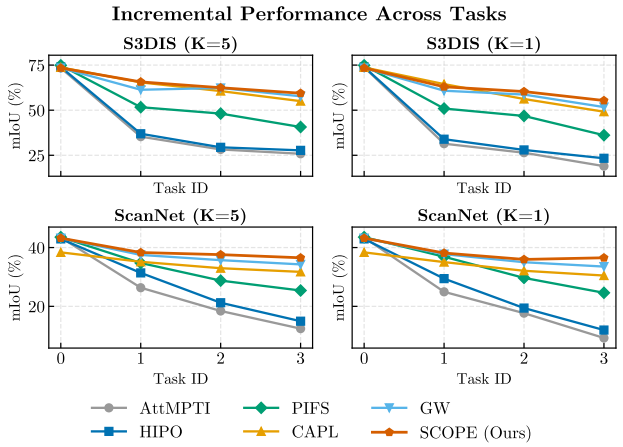

Figure 3. Incremental performance on S3DIS and ScanNet for t=0 to t=3 under K=5 (left) and K=1 (right), respectively. Curves show the evolution of mIoU across incremental stages.

Performance Across Tasks. Fig. 3 illustrates mIoU progression over incremental tasks ($t \in \{0, 1, 2, 3\}$) on S3DIS and ScanNet under both K=5 (left) and K=1 (right) settings. Higher trajectories at later stages indicate stronger knowledge retention and more effective adaptation to newly introduced classes. Across all settings, SCOPE consistently maintains a superior trajectory compared to competing approaches, with the performance gap widening as tasks accumulate–particularly under the challenging K=1 regime. This behaviour reflects a favourable stability–plasticity balance: base-class performance is preserved while novel categories are integrated with minimal degradation. Among baselines, AttMPTI [53] exhibits a pronounced decline across stages, consistent with its design for static few-shot learning rather than incremental updates. HIPO [37], despite targeting the IFS-PCS setting, also shows noticeable performance drop-off, underscoring the difficulty of maintaining stability under limited supervision. In contrast, SCOPE demonstrates smoother and more stable progression across stages, evidencing the effectiveness of background-guided contextual refinement in reducing forgetting and enabling reliable incremental adaptation.

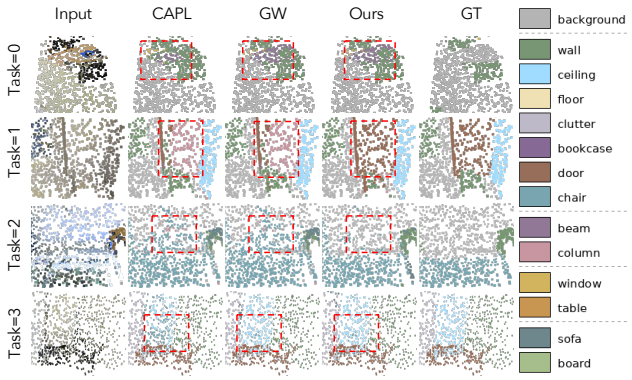

Figure 4. Qualitative comparison with competing methods from t=0 to t=3. The colour palette (right) denotes semantic classes, and dotted separators indicate newly introduced classes at each incremental stage (top to bottom).

Qualitative Comparison. Qualitative results from t=0 to t=3 are illustrated in Fig. 4, with key differences highlighted by red dashed boxes. At t=0, our method and GW produce identical predictions, as both originate from the same base model, whereas CAPL–trained under a different paradigm–exhibits visibly different outputs. At t=1, distinctions emerge: CAPL and GW misclassify parts of the *door* as *column*, while our method yields more accurate segmentation with only minor errors. This trend persists at t=2, where baselines produce increasingly fragmented masks and occasionally hallucinate objects in background regions. In contrast, our predictions maintain stronger structural coherence and semantic stability. By t=3, baseline inconsistencies become more pronounced (e.g., in the *ceiling*), whereas our outputs remain better aligned with the ground truth. Although newly introduced classes are not explicitly marked, detailed per-task visualisations and additional results–including plug-and-play demonstrations showing seamless integration with other prototype-based

Table 4. Comprehensive ablation study on ScanNet ($K = 5$).

Variant	mIoU	mIoU-N	HM	mIoU-I	FPP ↓
Support Set Only (GW [46])	34.27	16.88	23.94	37.67	1.49
+ CPR (mean)	35.68	22.12	28.91	38.02	1.50
+ APE (full)	**36.52**	**23.86**	**30.38**	**38.91**	**1.27**

baselines–are provided in the ***Supp.***.

Across datasets and low-shot settings, SCOPE consistently preserves base knowledge while improving novel-class segmentation. Qualitative results show fewer artefacts and more coherent predictions across tasks, highlighting contextual cues as a key driver of incremental performance.

4.3. Ablation Study

Below, we present ablation studies analysing the effect of CPR, APE, and sensitivity of key hyperparameters on the ScanNet dataset, following the setup described in Sec. 4.1.

Effect of CPR. Tab. 4 shows that introducing CPR on top of the GW [46] baseline with mean aggregation yields substantial gains on novel classes, improving mIoU-N by +5.24 and HM by +4.97 percentage points. mIoU increases from 34.27% to 35.68%, and mIoU-I from 37.67% to 38.02%, confirming the benefit of semantically aligned background prototypes for enhancing few-shot representations.

Effect of APE. Building upon CPR, APE further improves novel and incremental performance, increasing mIoU-N and HM by +1.74 and +1.47 points, raising mIoU-I to 38.91%, and reducing forgetting from 1.50 to 1.27. This confirms that adaptive attention weighting suppresses noisy pseudo-instances and weak retrieval, yielding more discriminative and stable incremental learning.

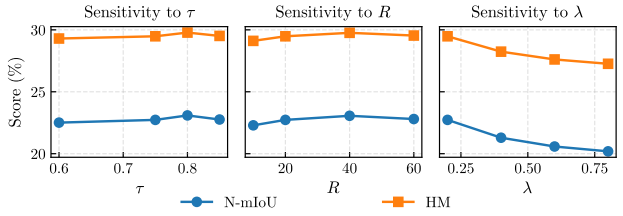

Figure 5. Hyperparameter sensitivity on ScanNet in terms of N-IoU and HM, obtained by varying one parameter at a time while keeping the others fixed ($\tau = 0.75$, $R = 20$, $\lambda = 0.5$).

Effect of Hyperparameters. We analyse three key hyperparameters: the confidence threshold τ for IPB construction, the number of retrieved prototypes R, and the fusion weight λ. As shown in Fig. 5, performance remains stable across reasonable ranges. Increasing τ yields improvements up to $\tau = 0.8$, after which performance slightly declines, likely due to overly aggressive filtering of informative background structures. Moderate R values (around $R = 40$) offer the best balance between contextual diversity and noise robustness, whereas larger R may introduce less relevant prototypes. For λ, smaller values, corresponding to stronger background integration, achieve better results; as λ increases, reduced contextual contribution leads to performance degradation, indicating that few-shot prototypes alone are insufficient for reliable IFS-PCS.

4.4. Discussion

Error Propagation. To assess the noise introduced into the IPB by the class-agnostic segmentation model, we construct an alternative IPB using ground-truth masks and compare it with the pseudo-mask version. The performance gap is small (24.77 vs. 23.86 mIoU-N and 31.20 vs. 30.38 HM for GT and pseudo masks, respectively), indicating that pseudo-mask noise has limited impact. This is largely because low-confidence instances are filtered during IPB construction, and APE further mitigates residual noise through selective prototype weighting (Sec. 4.3).

Computational Efficiency. Compared with the strongest baseline, GW [46], SCOPE incurs negligible computational overhead during incremental learning. The instance prototype bank is constructed once offline after base training, requires no further optimisation, and adds less than 1 MB of memory storage. During incremental stages, the only additional operations are non-parametric CPR and APE, both parameter- and training-free. As a result, the per-task computational cost remains virtually unchanged, with runtime comparable to GW (18.60 s vs. 18.58 s per task).

5. Conclusion

We introduced SCOPE, a lightweight plug-and-play framework for IFS-PCS. Our key insight is that background regions in base scenes encode reusable object-level structure useful for future novel classes. We exploit this by mining high-confidence pseudo-instances with a class-agnostic segmenter to build an IPB, which is used through CPR and APE to refine sparse novel-class prototypes. This design enables adaptation under limited supervision without backbone retraining or additional parameters, achieving a strong stability–plasticity balance. Experiments on standard 3D benchmarks show consistent improvements in novel-class accuracy and reduced forgetting over representative baselines. Overall, proposed framework demonstrates that background-guided prototype enrichment is an effective and scalable strategy for low-shot continual 3D scene understanding. Future work will extend the framework to large-scale outdoor and multi-modal settings while reducing reliance on class-agnostic segmentation models.

Acknowledgements. This research was supported in part by the UKRI-AHRC CoSTAR National Lab for Creative Industries Research and Development (AH/Y001060/1), the Alan Turing Mobility Grant (Academic Year 2024–26), and the Ministry of Education, Singapore, under the MOE Academic Research Fund Tier 2 (MOE-T2EP20124-0013).

References

[1] Rahaf Aljundi, Francesca Babiloni, Mohamed Elhoseiny, Marcus Rohrbach, and Tinne Tuytelaars. Memory aware synapses: Learning what (not) to forget. In *Eur. Conf. Comput. Vis.*, pages 139–154, 2018. 2

[2] Zhaochong An, Guolei Sun, Yun Liu, Fayao Liu, Zongwei Wu, Dan Wang, Luc Van Gool, and Serge Belongie. Rethinking few-shot 3D point cloud semantic segmentation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 3996–4006, 2024. 2

[3] Zhaochong An, Guolei Sun, Yun Liu, Runjia Li, Junlin Han, Ender Konukoglu, and Serge Belongie. Generalized few-shot 3d point cloud segmentation with vision-language model. In *CVPR*, pages 16997–17007, 2025. 1, 2

[4] Zhaochong An, Guolei Sun, Yun Liu, Runjia Li, Min Wu, Ming-Ming Cheng, Ender Konukoglu, and Serge Belongie. Multimodality helps few-shot 3d point cloud semantic segmentation. *ICLR*, 2025. 1, 2

[5] Iro Armeni, Ozan Sener, Amir R Zamir, Helen Jiang, Ioannis Brilakis, Martin Fischer, and Silvio Savarese. 3D semantic parsing of large-scale indoor spaces. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 1534–1543, 2016. 6

[6] Mohamed El Amine Boudjoghra, Salwa Al Khatib, Jean Lahoud, Hisham Cholakkal, Rao Anwer, Salman H Khan, and Fahad Shahbaz Khan. 3d indoor instance segmentation in an open-world. *Adv. Neural Inform. Process. Syst.*, 36, 2024. 2

[7] Pietro Buzzega, Matteo Boschini, Angelo Porrello, Davide Abati, and Simone Calderara. Dark experience for general continual learning: a strong, simple baseline. *Adv. Neural Inform. Process. Syst.*, 33:15920–15930, 2020. 2

[8] Fabio Cermelli, Massimiliano Mancini, Yongqin Xian, Zeynep Akata, and Barbara Caputo. Prototype-based incremental few-shot segmentation. In *The 32nd British Machine Vision Conference*. BMVA Press, 2021. 1, 2, 6

[9] Hyuntak Cha, Jaeho Lee, and Jinwoo Shin. Co2l: Contrastive continual learning. In *Int. Conf. Comput. Vis.*, pages 9516–9525, 2021. 2

[10] Runnan Chen, Youquan Liu, Lingdong Kong, Xinge Zhu, Yuexin Ma, Yikang Li, Yuenan Hou, Yu Qiao, and Wenping Wang. CLIP2Scene: Towards label-efficient 3D scene understanding by CLIP. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 7020–7030, 2023. 2

[11] Xiaozhi Chen, Huimin Ma, Ji Wan, Bo Li, and Tian Xia. Multi-view 3d object detection network for autonomous driving. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2017. 1

[12] Christopher Choy, JunYoung Gwak, and Silvio Savarese. 4d spatio-temporal convnets: Minkowski convolutional neural networks. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 3075–3084, 2019. 1

[13] Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. ScanNet: Richly-annotated 3D reconstructions of indoor scenes. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 5828–5839, 2017. 6

[14] Martin Engelcke, Dushyant Rao, Dominic Zeng Wang, Chi Hay Tong, and Ingmar Posner. Vote3deep: Fast object detection in 3d point clouds using efficient convolutional neural networks. In *Int. Conf. Robot. Autom.*, pages 1355–1361. IEEE, 2017. 1

[15] Dan Andrei Ganea, Bas Boom, and Ronald Poppe. Incremental few-shot instance segmentation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 1185–1194, 2021. 1, 2

[16] Shuting He, Xudong Jiang, Wei Jiang, and Henghui Ding. Prototype adaption and projection for few-and zero-shot 3D point cloud semantic segmentation. *IEEE Trans. Image Process.*, 32:3199–3211, 2023. 2

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

[18] Qingyong Hu, Bo Yang, Linhai Xie, Stefano Rosa, Yulan Guo, Zhihua Wang, Niki Trigoni, and Andrew Markham. RandLA-Net: Efficient semantic segmentation of large-scale point clouds. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 11108–11117, 2020. 2

[19] Rui Huang, Songyou Peng, Ayca Takmaz, Federico Tombari, Marc Pollefeys, Shiji Song, Gao Huang, and Francis Engelmann. Segment3d: Learning fine-grained class-agnostic 3d segmentation without manual labels. In *Eur. Conf. Comput. Vis.*, pages 278–295. Springer, 2024. 2, 3, 4

[20] Tianyu Huang, Bowen Dong, Yunhan Yang, Xiaoshui Huang, Rynson WH Lau, Wanli Ouyang, and Wangmeng Zuo. CLIP2Point: Transfer CLIP to point cloud classification with image-depth pre-training. In *Int. Conf. Comput. Vis.*, pages 22157–22167, 2023. 2

[21] 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. *NAS*, 114(13):3521–3526, 2017. 1, 2, 6

[22] Theodora Kontogianni, Yuanwen Yue, Siyu Tang, and Konrad Schindler. Is continual learning ready for real-world challenges? *arXiv preprint*, 2024. 2

[23] Yangyan Li, Rui Bu, Mingchao Sun, Wei Wu, Xinhan Di, and Baoquan Chen. PointCNN: Convolution on X-transformed points. In *Adv. Neural Inform. Process. Syst.*, pages 828–838, 2018. 2

[24] Zhizhong Li and Derek Hoiem. Learning without forgetting. *IEEE Trans. Pattern Anal. Mach. Intell.*, 40(12):2935–2947, 2017. 1, 2, 6

[25] Sun-Ao Liu, Yiheng Zhang, Zhaofan Qiu, Hongtao Xie, Yongdong Zhang, and Ting Yao. Learning orthogonal prototypes for generalized few-shot semantic segmentation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 11319–11328, 2023. 2

[26] Yongqiang Mao, Zonghao Guo, LU Xiaonan, Zhiqiang Yuan, and Haowen Guo. Bidirectional feature globalization for few-shot semantic segmentation of 3D point cloud scenes. In *3DV*, pages 505–514, 2022. 2

[27] Andres Milioto, Ignacio Vizzo, Jens Behley, and Cyrill Stachniss. Rangenet++: Fast and accurate lidar semantic segmentation. In *IROS*, pages 4213–4220. IEEE, 2019. 1

[28] Songyou Peng, Kyle Genova, Chiyu Jiang, Andrea Tagliasacchi, Marc Pollefeys, Thomas Funkhouser, et al.

OpenScene: 3D scene understanding with open vocabularies. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 815–824, 2023. 2

[29] Roberto Pierdicca, Marina Paolanti, Francesca Matrone, Massimo Martini, Christian Morbidoni, Eva Savina Malinverni, Emanuele Frontoni, and Andrea Maria Lingua. Point cloud semantic segmentation using a deep learning framework for cultural heritage. *Remote Sensing*, 12(6):1005, 2020. 1

[30] Charles R Qi, Hao Su, Kaichun Mo, and Leonidas J Guibas. PointNet: Deep learning on point sets for 3D classification and segmentation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 652–660, 2017. 2

[31] Charles Ruizhongtai Qi, Li Yi, Hao Su, and Leonidas J Guibas. PointNet++: Deep hierarchical feature learning on point sets in a metric space. In *Adv. Neural Inform. Process. Syst.*, pages 5099–5108, 2017. 2

[32] Sylvestre-Alvise Rebuffi, Alexander Kolesnikov, Georg Sperl, and Christoph H Lampert. icarl: Incremental classifier and representation learning. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 2001–2010, 2017. 2

[33] Jonas Schult, Francis Engelmann, Alexander Hermans, Or Litany, Siyu Tang, and Bastian Leibe. Mask3D: Mask Transformer for 3D Semantic Instance Segmentation. In *Int. Conf. Robot. Autom.*, 2023. 1

[34] Joan Serra, Didac Suris, Marius Miron, and Alexandros Karatzoglou. Overcoming catastrophic forgetting with hard attention to the task. In *Int. Conf. Mach. Learn.*, pages 4548–4557. PMLR, 2018. 2

[35] Yuanzhi Su, Siyuan Chen, and Yuan-Gen Wang. Balanced residual distillation learning for 3d point cloud class-incremental semantic segmentation. *arXiv preprint*, 2024. 2

[36] Jiahao Sun, Chunmei Qing, Junpeng Tan, and Xiangmin Xu. Superpoint transformer for 3d scene instance segmentation. *arXiv preprint*, 2022. 1

[37] Tanuj Sur, Samrat Mukherjee, Kaizer Rahaman, Subhasis Chaudhuri, Muhammad Haris Khan, and Biplab Banerjee. Hyperbolic uncertainty-aware few-shot incremental point cloud segmentation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 11810–11821, 2025. 1, 2, 6, 7

[38] Ayça Takmaz, Elisabetta Fedele, Robert W Sumner, Marc Pollefeys, Federico Tombari, and Francis Engelmann. OpenMask3D: Open-vocabulary 3D instance segmentation. In *Adv. Neural Inform. Process. Syst.*, pages 68367–68390, 2023. 2

[39] Vishal Thengane, Jean Lahoud, Hisham Cholakkal, Rao Muhammad Anwer, Lu Yin, Xiatian Zhu, and Salman Khan. Climb-3d: Continual learning for imbalanced 3d instance segmentation. In *Brit. Mach. Vis. Conf.*, 2025. 1, 2, 6

[40] Zhuotao Tian, Xin Lai, Li Jiang, Shu Liu, Michelle Shu, Hengshuang Zhao, and Jiaya Jia. Generalized few-shot semantic segmentation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 11563–11572, 2022. 1, 2, 6

[41] Chih-Jung Tsai, Hwann-Tzong Chen, and Tyng-Luh Liu. Pseudo-embedding for generalized few-shot 3D segmentation. In *Eur. Conf. Comput. Vis.*, pages 383–400, 2024. 2

[42] Jiahui Wang, Haiyue Zhu, Haoren Guo, Abdullah Al Mamun, Cheng Xiang, and Tong Heng Lee. Few-shot point cloud semantic segmentation via contrastive self-supervision and multi-resolution attention. In *Int. Conf. Robot. Autom.*, pages 2811–2817, 2023. 2

[43] Peng Wang, An Yang, Rui Men, Junyang Lin, Shuai Bai, Zhikang Li, Jianxin Ma, Chang Zhou, Jingren Zhou, and Hongxia Yang. OFA: Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework. In *Int. Conf. Mach. Learn.*, pages 23318–23340, 2022. 2

[44] Yue Wang, Yongbin Sun, Ziwei Liu, Sanjay E Sarma, Michael M Bronstein, and Justin M Solomon. Dynamic graph CNN for learning on point clouds. *ACM Trans. Graph.*, 38(5):1–12, 2019. 2

[45] Lili Wei, Congyan Lang, Ziyi Chen, Tao Wang, Yidong Li, and Jun Liu. Generated and pseudo content guided prototype refinement for few-shot point cloud segmentation. In *Adv. Neural Inform. Process. Syst.*, pages 31103–31123, 2024. 2

[46] Yating Xu, Conghui Hu, Na Zhao, and Gim Hee Lee. Generalized few-shot point cloud segmentation via geometric words. In *Int. Conf. Comput. Vis.*, pages 21506–21515, 2023. 1, 2, 6, 8

[47] Yating Xu, Na Zhao, and Gim Hee Lee. Towards robust few-shot point cloud semantic segmentation. In *The 34th British Machine Vision Conference*. BMVA Press, 2023. 2

[48] Jihan Yang, Runyu Ding, Weipeng Deng, Zhe Wang, and Xiaojuan Qi. RegionPLC: Regional point-language contrastive learning for open-world 3D scene understanding. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 19823–19832, 2024. 2

[49] Yuwei Yang, Munawar Hayat, Zhao Jin, Chao Ren, and Yinjie Lei. Geometry and uncertainty-aware 3d point cloud class-incremental semantic segmentation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 21759–21768, 2023. 1, 6

[50] Canyu Zhang, Zhenyao Wu, Xinyi Wu, Ziyu Zhao, and Song Wang. Few-shot 3D point cloud semantic segmentation via stratified class-specific attention based transformer network. In *AAAI*, pages 3410–3417, 2023. 2

[51] Junbo Zhang, Runpei Dong, and Kaisheng Ma. CLIP-FO3D: Learning free open-world 3D scene representations from 2D dense CLIP. In *Int. Conf. Comput. Vis.*, pages 2048–2059, 2023. 2

[52] Hengshuang Zhao, Li Jiang, Jiaya Jia, Philip Torr, and Vladlen Koltun. Point transformer, 2021. 1, 2

[53] Na Zhao, Tat-Seng Chua, and Gim Hee Lee. Few-shot 3d point cloud semantic segmentation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 8873–8882, 2021. 1, 2, 6, 7

[54] Yuchen Zhou, Jiayuan Gu, Tung Yen Chiang, Fanbo Xiang, and Hao Su. Point-sam: Promptable 3D segmentation model for point clouds. In *Int. Conf. Learn. Represent.*, 2025. 2

[55] Xiangyang Zhu, Renrui Zhang, Bowei He, Ziyu Guo, Jiaming Liu, Han Xiao, Chaoyou Fu, Hao Dong, and Peng Gao. No time to train: Empowering non-parametric networks for few-shot 3D scene segmentation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2024. 2