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

Supplementary Material

Overview. This supplementary document provides additional technical details and extended analyses that complement the main paper. We first describe the datasets and preprocessing pipeline used in our experiments, including scene statistics, class distributions, and the construction of incremental tasks (Sec. B). We then provide additional implementation details of our framework in Sec. C. Next, Sec. D demonstrates the flexibility of SCOPE by showing how the proposed Scene Contextualisation (SC) module can be integrated in a plug-and-play manner with existing prototype-based incremental learners. Next, Sec. E reports detailed results for each incremental task, including qualitative visualisations on ScanNet and S3DIS that highlight improvements in localisation, boundary accuracy, and overall structural consistency. Finally, Sec. F evaluates the robustness of SCOPE under a long-term incremental setting spanning six stages on ScanNet, demonstrating its effectiveness in sustained incremental learning over extended task sequences. We also discuss the limitations of SCOPE in Sec. G.

A. Competitors Comparison

Tab. A1 summarises the competitor methods considered in our evaluation. These approaches span several related paradigms, including incremental learning (specifically class-incremental learning), few-shot segmentation, generalised few-shot segmentation, and incremental few-shot segmentation. This diversity enables a comprehensive comparison under the combined challenges of data scarcity, incremental adaptation, and knowledge retention.

B. Additional Dataset Details

This section extends the dataset description provided in Sec. 4.1 of the main paper. We provide further details on dataset preprocessing, dataset characteristics, class partitions, and the construction of incremental learning tasks used in our experiments.

B.1. Preprocessing

We follow the preprocessing pipeline proposed in [53]. Each scene is partitioned into non-overlapping $1\,\mathrm{m} \times 1\,\mathrm{m}$ blocks on the xy-plane, and $M{=}2048$ points are sampled from each block. Each point is represented by a 9-dimensional feature vector ($d_0{=}9$) consisting of the `XYZ` spatial coordinates, `RGB` colour values, and block-normalised $\overline{\texttt{XYZ}}$ coordinates.

B.2. Base-Novel Distribution

S3DIS. We follow the standard evaluation protocol where Area 6 is reserved for testing, while the remaining five areas provide training data for both the base and incremental stages. In addition to the overview presented in the main paper, this appendix reports the distribution of labelled points across all 13 semantic classes, summarised in Tab. B2. Although S3DIS is not strongly imbalanced, several categories such as *beam*, *column*, and *board* appear less frequently than dominant structural classes like *wall* and *floor*. Following the protocol of [46], these lower-frequency classes form the novel set $\boldsymbol{C}^{\mathrm{n}}$ introduced during incremental stages.

ScanNet. We adopt the official split of 1,201 training scenes and 312 testing scenes. Compared with S3DIS, ScanNet exhibits greater variability in object occurrence across scenes, with several categories appearing only sporadically. The distribution of labelled points across all 20 semantic classes is also reported in Tab. B2. Following the class partitioning strategy of [46], the least frequent categories are grouped into the novel set $\boldsymbol{C}^{\mathrm{n}}$ and introduced progressively during incremental learning.

B.3. Incremental Task Construction

In addition to the class partitions, we detail the construction of incremental tasks for both datasets. Novel classes are introduced sequentially across multiple stages following the protocol described in the main paper. Each stage introduces a small set of novel classes together with their few-shot annotated samples, simulating the gradual emergence of new object categories in dynamic indoor environments.

For clarity, we use the notation `XB-YI` to denote an incremental configuration, where `X` represents the number of base classes used during the initial training stage and `Y` denotes the number of novel classes introduced at each incremental step. For example, `15B-2I` indicates a configuration with 15 base classes and two novel classes added at every incremental stage.

C. Additional Implementation Details

The main experiments reported in Tab. 2 and Tab. 3 follow a four-stage learning protocol, including the base stage. All reported results are averaged over five runs using different few-shot support samples to ensure statistical robustness.

All experiments follow the base-stage training protocol of Xu *et al*. [46] ($t{=}0$). The encoder Φ is trained with full supervision on $\mathbf{D}^{\mathrm{b}}$. For incremental stages ($t{\geq}1$), the backbone remains frozen and only class-specific prototypes

Table A1. Overview of competitor methods used in our evaluation. Methods span, few-shot learning (FS), generalised few-shot learning (GFS), class-incremental learning (CI), and incremental few-shot learning (IFS) paradigms.

Method	Paradigm	Few-shot	Incremental	Prototype-based	Key Idea
FT (Naïve Finetuning)	–	✗	✓	✗	Direct fine-tuning on new tasks without forgetting control
LwF [24]	CI	✗	✓	✗	Knowledge distillation to preserve previous predictions
EWC [21]	CI	✗	✓	✗	Fisher-information regularisation to protect important weights
GUA [49]	CI	✗	✓	✗	Geometry and uncertainty-aware regularisation for PCS
CLIMB-3D [39]	CI	✗	✓	✗	Handles long-tail distributions in incremental 3D segmentation
AttMPTI [53]	FS	✓	✗	✓	Multi-prototype inference with label propagation
CAPL [40]	GFS	✓	✗	✓	Prototype learning with co-occurrence priors
GW [46]	GFS	✓	✗	✓	Geometry-guided prototype learning for 3D segmentation
PIFS [8]	IFS	✓	✓	✓	Prototype learning with knowledge distillation
HIPO [37]	IFS	✓	✓	✓	Hyperbolic prototype embeddings for incremental few-shot learning

Table B2. Overview of the base and novel class partitions for S3DIS and ScanNet. Following the protocol in the main paper, the majority classes form the base set $\boldsymbol{C}^{\mathrm{b}}$, while the six least-represented categories constitute the novel set $\boldsymbol{C}^{\mathrm{n}}$ introduced incrementally.

Datasets	Base Classes ($\boldsymbol{C}^{\mathrm{b}}$)	Novel Classes ($\boldsymbol{C}^{\mathrm{n}}$)
S3DIS	Wall, Ceiling, Floor, Clutter, Bookcase, Door, Chair	Beam, Column, Window, Table, Sofa, Board
ScanNet	Refrigerator, Desk, Curtain, Sofa, Bookshelf, Bed, Table, Otherfurniture, Window, Cabinet, Door, Chair, Unannotated, Floor, Wall	Sink, Toilet, Bathtub, Shower Curtain, Picture, Counter

are updated using few-shot support sets $\mathbf{D}^{\mathrm{t}}$. For the point cloud encoder Φ', we adopt DGCNN [44] as the backbone. The network is first pre-trained for 100 epochs and then fine-tuned during the base stage for 150 epochs. Note that our method is model-agnostic and can be readily applied to alternative backbones such as Point Transformer [52], as demonstrated in [46, 49]. For the scene contextualisation module, the pseudo-instance filtering threshold τ is set to 0.75, the number of retrieved prototypes to $R{=}50$, and the fusion weight to $\lambda{=}0.5$. Although this stage is model-agnostic, we employ Segment3D [19] as the class-agnostic segmenter because it is trained without 3D ground-truth supervision—a requirement that, to the best of our knowledge, is not satisfied by other currently available models. During the novel-class registration phase, no backpropagation is performed; instead, new class prototypes are computed and integrated using the procedures described in the main paper. Other implementation details follow those of [46].

D. SCOPE as Plug-and-Play

Tab. D3 reports the effect of adding our Scene Contextualisation (SC) module to existing prototype-based IFS-PCS methods on ScanNet under the $k{=}5$ shot setting. SC acts as a lightweight plug-and-play refinement that enriches prototype formation using contextual cues from background regions in base scenes, without modifying the backbone or optimisation pipeline.

Across all baselines, SC consistently improves novel-class learning while maintaining a balanced performance between base and incremental stages. For **PIFS**, SC increases mIoU-N from 3.43 to **4.93** and HM from 6.24 to **8.64**, indicating improved few-shot adaptation. **CAPL** also benefits from SC, with mIoU-N improving from 14.75 to **18.70** and HM from 21.36 to **25.25**, reflecting more stable integration of base and novel representations. Applying SC to **GW** yields our final model, **SCOPE**, which achieves the strongest overall performance: mIoU improves from 34.27 to **36.52**, mIoU-N from 16.88 to **23.86**, and HM from 23.94 to **30.38**, while maintaining controlled forgetting (FPP decreases from 1.49 to **1.27**). The mIoU-I also increases from 37.67 to **38.91**, indicating improved balance between knowledge retention and novel-class adaptation.

Overall, these results demonstrate that SC consistently enhances prototype-based methods, and its integration with GW produces the best-performing system. This confirms scene contextualisation as an effective and general plug-and-play component for prototype-based IFS-PCS.

E. Per Task Breakdown

Incremental Performance. Tab. E4 presents results across all incremental stages in the ScanNet 15B–2I setting. SCOPE consistently outperforms the previous top-performing approach, GW [46]. Since SCOPE focuses on enhancing novel-class learning through prototype refinement without updating the backbone, the base-class performance remains comparable to GW across all tasks. As expected, the improvements are most pronounced on the

Table D3. Effect of adding Scene Contextualisation (SC) to prototype-based IFS-PCS methods on ScanNet (5-shot). SC consistently improves novel-class performance (mIoU-N), incremental-stage accuracy (mIOU-I), and harmonic balance (HM) across all baselines. Applying SC to GW yields our full model (**Ours**), which achieves the best overall performance.

Method	**mIoU**	**mIoU-B**	**mIoU-N**	**HM**	**mIOU-I**	**FPP**
PIFS	25.39	34.81	3.43	6.24	33.11	8.74
+ SC	27.36	34.94	4.93	8.64	34.03	8.61
CAPL	31.73	39.01	14.75	21.36	34.55	-0.65
+ SC	32.97	39.08	18.70	25.25	35.19	-0.72
GW	34.27	41.72	16.88	23.94	37.67	1.49
Ours	36.52	41.94	23.86	30.38	38.91	1.27

Table E4. Per-class mIoU across incremental stages on the ScanNet 15B–2I setting. "Mean" denotes the average over base or novel classes. Dashes indicate classes not yet introduced.

Method	**Task**	**Base Mean**	**Novel Mean**	**Novel Classes (C^n)**					
				Sink	Toilet	Bathtub	Shower Curtain	Picture	Counter
GW [46]	1	40.92	13.45	8.96	17.94	–	–	–	–
	2	41.01	17.16	9.38	21.63	16.64	21.00	–	–
	3	41.72	16.88	9.78	22.54	18.16	22.77	13.78	14.23
SCOPE (Ours)	1	41.02	19.29	12.70	25.88	–	–	–	–
	2	41.15	25.10	12.90	32.33	26.05	29.13	–	–
	3	41.94	23.86	17.88	32.59	25.80	30.13	17.27	19.51

newly introduced categories. In Task 1, SCOPE increases the novel mean mIoU from 13.45 to 19.29, with notable gains on *sink* (+3.74) and *toilet* (+7.94). In Task 2, the novel mean further improves from 17.16 to 25.10, outperforming GW on all four newly introduced categories, including substantial improvements on *bathtub* (+9.41) and *shower curtain* (+8.13). By Task 3, after all six novel categories have appeared, SCOPE maintains a clear advantage with a novel mean of 23.86 compared to 16.88 for GW.

Additionally, certain categories such as *sink* and *toilet* continue to benefit as more novel classes are introduced. For instance, the mIoU for *sink* increases from 12.70 in Task 1 to 17.88 in Task 3, while *toilet* improves from 25.88 to 32.59 over the same period. This trend suggests that the proposed scene-contextual prototype refinement becomes more effective as the model accumulates richer contextual cues across incremental stages.

Overall, these results show that SCOPE not only preserves performance on base classes but also delivers significantly stronger few-shot generalisation to novel categories, enabling more robust incremental adaptation across all stages.

Qualitative Performance (S3DIS, 5-shot). As shown in Fig. E1, under the 5-shot setting on S3DIS, our method consistently improves as new classes are introduced, outperforming existing baselines. For example, in *Task 1*, when the *beam* class is introduced, CAPL frequently confuses columns with beams, while both GW and our method produce more stable predictions; however, our masks are noticeably cleaner and better aligned with the ground truth. In *Task 2*, after introducing the *table* class, both GW and CAPL fail to identify table regions reliably, whereas our method produces markedly more accurate and complete masks. A similar trend is observed when the *board* class is added: the predictions generated by our model remain significantly closer to the ground truth, demonstrating better generalisation and reduced confusion across related categories.

F. Long-Term Scalability

We further evaluate the robustness of SCOPE under a long-term incremental setting spanning six stages on ScanNet. In this challenging scenario, SCOPE achieves **19.75/26.79** (mIoU-N/HM), outperforming GW which attains 15.64/22.74. These results indicate that the proposed scene-contextual prototype enrichment remains effective as new classes are introduced over multiple stages, enabling stable accumulation of knowledge while maintaining competitive base-class performance. The improvements suggest that leveraging contextual cues from background regions helps the model form more reliable prototypes for novel classes, thereby supporting sustained incremental learning

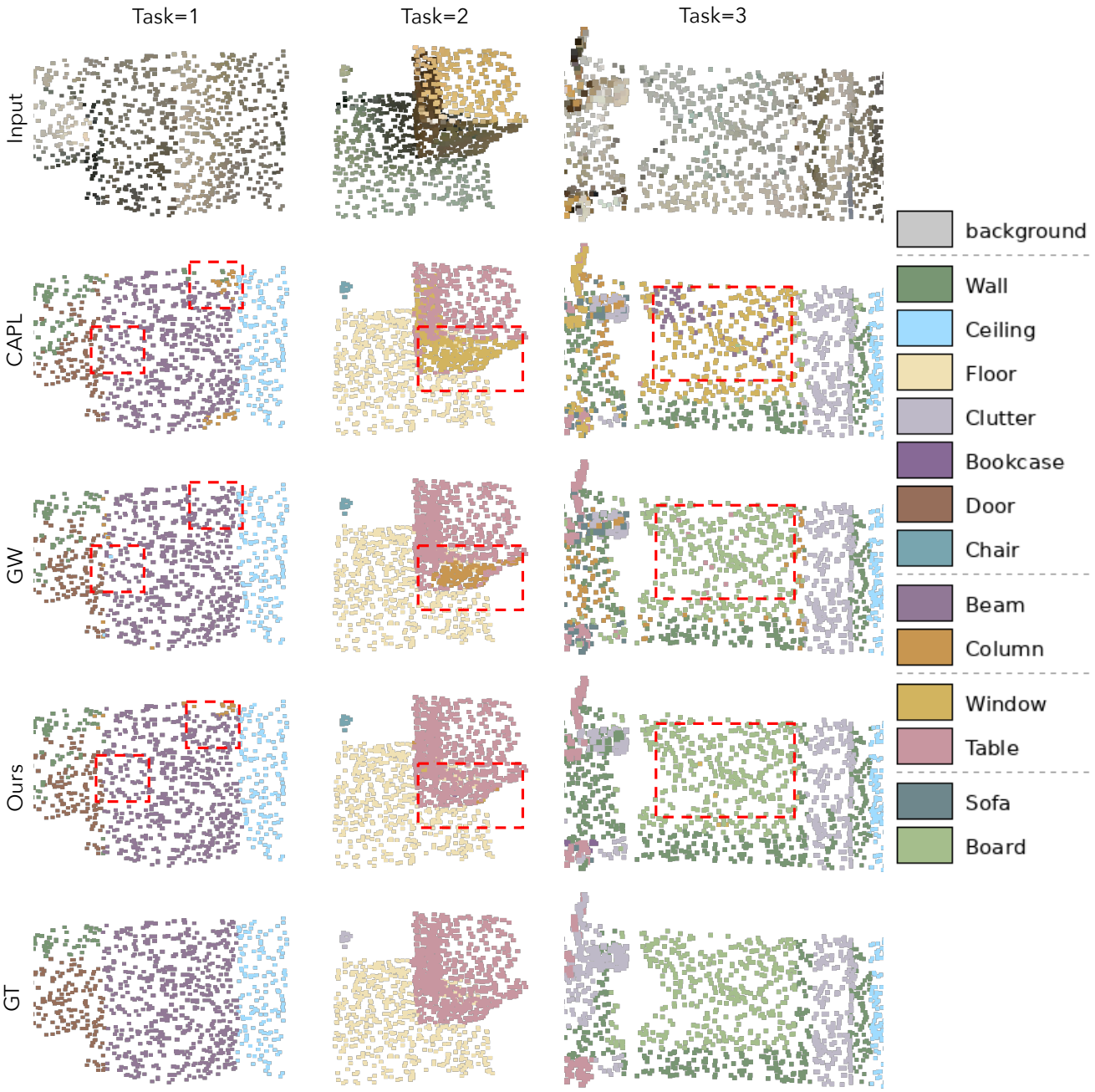

Figure E1. Qualitative comparison of SCOPE with competing baselines from task t=1 to t=3. The colour palette (right) denotes semantic classes, while dotted separators indicate the introduction of new classes at each incremental stage.

over extended task sequences.

G. Limitations

Although our method achieves strong performance and is relatively insensitive to hyperparameters, its effectiveness depends on the quality of pseudo-instance masks extracted during scene contextualisation. Performance may degrade if the class-agnostic model produces inaccurate or fragmented proposals. Moreover, the limited availability of class-agnostic models that do not require 3D ground-truth supervision constrains adoption, as such architectures remain scarce.