Supplementary

This supplement contains more ablation studies, time complexities, and visualizations that could not fit in the study. In particular, we include (1) extra examples on divergence vs distance threshold, (2) time complexity for different methods, (3) discussions on SLAM methods, and (4) more visualizations of selected keyframes.

A. Extra examples on divergence vs distance threshold

We present additional examples comparing divergence against distance threshold, comprising 4 examples on HabitatSim and 3 examples on KITTI. Overall, this follows the previous experiments conducted in Tab. I and Tab. II, where NetVLAD+SceneSum demonstrated the best results, followed by PCL+SceneSum.

B. Time complexity for different methods

We present the time taken by both the baseline methods and our approach in Tab. I and Tab. II. For non-learning-based methods, self-supervised, and supervised methods, we only consider the inference time for the test set. In contrast, for autolabeling, the training time is included in the total time calculation, as training is a crucial part of the optimization process.

Table I. Inference Time (seconds) for Habitat-Sim Dataset. (s) represents the methods supervised by ground truth. (a) represents the methods with autolabelling.

Scene	Goffs	Micanopy	Spotswood	Springhill	Stilwell	Stokes
DR-DSN	63.376	27.983	30.089	37.623	38.099	22.257
CA-SUM	139.969	51.596	51.098	64.564	67.103	46.769
PySceneDetect	17.563	6.764	7.516	9.484	10.809	6.074
VSUMM NetVLAD	23.475	8.893	9.211	11.409	13.39	6.496
VSUMM PCL	20.03	9.702	9.85	13.002	12.156	7.713
NetVLAD+SceneSum	183.305	206.311	89.960	293.312	97.829	64.503
PCL+SceneSum	166.037	207.957	75.142	292.857	95.879	56.893
NetVLAD+SceneSum(S)	194.999	166.694	128.187	383.043	95.815	58.244
PCL+SceneSum(S)	185.233	172.785	147.617	394.343	84.859	60.510
NetVLAD+SceneSum(A)	148.894	149.157	81.360	304.567	78.026	54.432
PCL+SceneSum(A)	151.797	138.924	74.393	309.289	75.936	57.571

Table II. Inference Time (seconds) for KITTI Dataset. Other abbreviations follow by Tab. I

Scene	KITTI(0018)	KITTI(0027)	KITTI(0028)
DR-DSN	16.775	16.017	18.354
CA-SUM	52.493	34.428	43.639
PySceneDetect	1.92	4.015	4.48
VSUMM NetVLAD	1.78	3.098	3.461
VSUMM PCL	1.462	2.432	2.898
NetVLAD+SceneSum	26.395	43.996	60.562
PCL+SceneSum	22.730	37.828	56.860
NetVLAD+Supervised	19.084	38.316	56.531
PCL+Supervised	18.183	34.909	52.305
NetVLAD+Autolabelling	17.350	34.002	50.562
PCL+Autolabelling	17.442	33.426	46.654

C. Discussions on SLAM methods

Some may argue that SLAM+KNN could easily solve this problem. However, attempts to map entire scenes using SLAM methods like OpenVSLAM were unsuccessful in Habitat environments at a sampling rate of 1 frame per second. In our study, OpenVSLAM faced a major challenge with tracking loss, preventing the generation of a complete map, as shown in Fig. I.

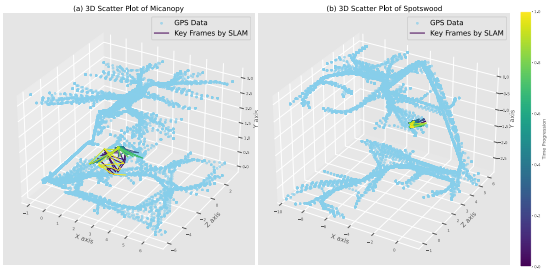


Figure I. **OpenVSLAM fails to create a complete map in Scene Micanopy and Spotswood**

We employ OpenVSLAM as one of our SLAM (Simultaneous Localization and Mapping) methods, but it frequently experiences tracking loss during the map generation phase, leading to disorganized and unstructured maps. This recurring issue highlights a fundamental limitation of our current SLAM approach, suggesting the need for further investigation and refinement to achieve reliable and well-structured map generation.

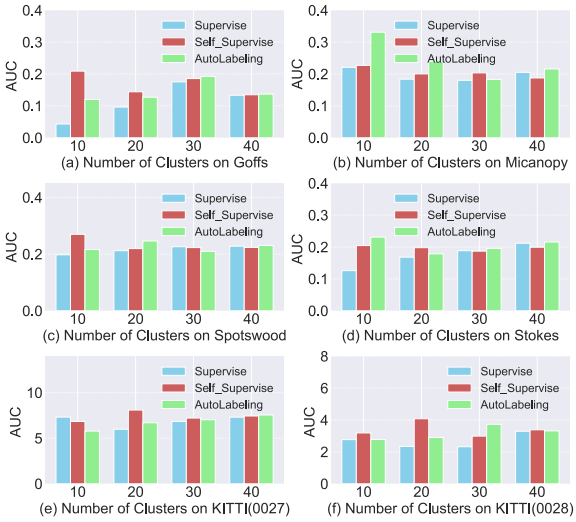
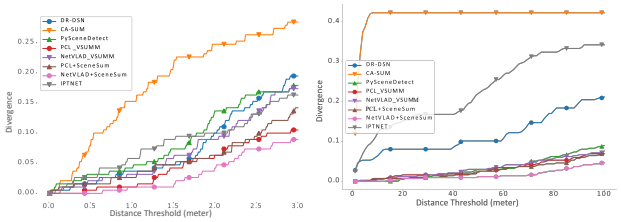


Figure II. Performance Analysis of Supervised Method, Self-Supervised Method and AutoLabeling under Different Number of Clusters in (a) Goffs (b) Micanopy (c) Spotswood (d) Springhill (e) KITTI(0027) (f) KITTI(0028)

C.1. Ablation Study

Supervised vs Self-supervised. This section aims to determine if introducing supervision loss results in significant performance changes in the model. Fig. II presents the average performance of supervised and self-supervised methods



(a) Habitat-Sim (Stilwell) at 20 Summarized Frames

(b) KITTI (0028) at 30 Summarized Frames.

Figure III. Comparison of Divergence vs Distance Threshold for Habitat-Sim and KITTI dataset

on the Habitat-Sim and KITTI environments. It compares these two models across six scenarios at cluster sizes of 10, 20, 30, and 40, using the AUC metric. Generally, the supervised model slightly outperforms the self-supervised model. Notably, with 10 clusters, the supervised approach shows a substantial improvement over the self-supervised one. For other cluster sizes, incorporating ground truth data during training offers only a slight performance boost. This highlights the robustness and effectiveness of self-supervised SceneSum across various contexts, proving it to be a adaptable tool, especially in situations where ground truth data is unavailable.

More VPR vs Contrastive-based clustering comparison. As discussed in Sec. 4.2, Tab. 1 and Tab. 3 again show that SceneSum, when combined with VPR-based clustering, significantly outperforms contrastive-based clustering in most scenes. This advantage is due to VPR's ability to capture distinctive information about locations and places, focusing on scene context and spatial relationships. In contrast, the contrastive-based clustering method captures only visual information, making VPR-based clustering more suitable for scene summarization.

Divergence comparisons. Fig. IIIa and IIIb highlight a significant trend in the performance of SceneSum approaches for scene summarization: as the distance threshold increases, the divergence also rises across all methods. NetVLAD+SceneSum consistently outperforms all baseline methods at every distance threshold, showing stable performance regardless of the distance evaluated. Additional results for more scenes are presented in Fig. IV

Table III. Best AUC results for different clustering on Habitat-sim

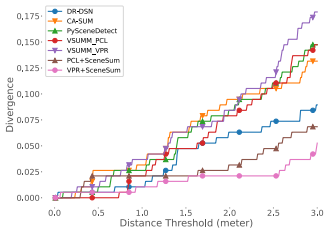
Scene / Clustering	Goffs	Micanopy	Spotswood	Springhill	Stilwell	Stokes	AVG.	SD.
NetVLAD	0.050	0.117	0.146	0.092	0.111	0.135	0.109	0.031
MixVPR	0.084	0.179	0.194	0.094	0.084	0.140	0.129	0.045
Patch	0.053	0.098	0.121	0.102	0.073	0.119	0.094	0.024

Sec. 4.3, 4.4, and C.1 reveal a key finding: self-supervised SceneSum exhibits strong summarization performance when applied zero-shot to a new scene. In this section, we explore the model's auto-labeling capability. If the time required for scene summarization is not a concern, SceneSum can function as a network summarizer, summarizing the scene video while simultaneously training on it.

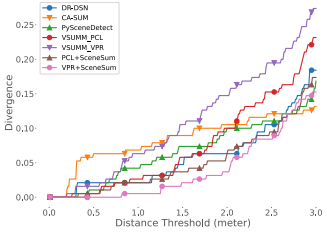
Fig. II shows that the auto-labeling results closely match those of the self-supervised approach, confirming the effectiveness of self-supervised inference. Even when comparing self-supervised SceneSum to a version that overfits the test dataset by training on it, the outcomes remain similar.

D. More visualizations on selected keyframes

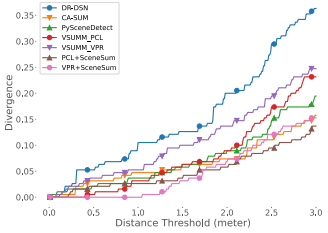
We have presented more visualizations on selected keyframes, including 5 Habitat scenes and 2 KITTI scenes, as shown in Figure. V, Figure. VI, Figure. VII, Figure. VIII, Figure. IX, Figure. X, and Figure. XI.



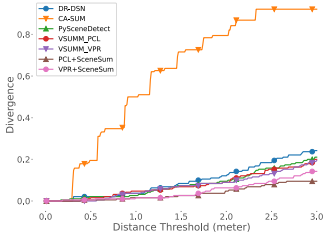
(a) Habitat-Sim (Goffs)



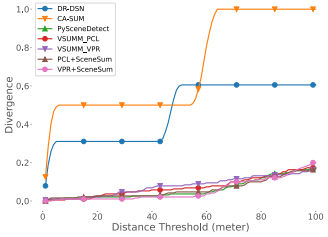
(b) Habitat-Sim (Micanopy)



(c) Habitat-Sim (Spotswood)



(d) Habitat-Sim (Stokes)



(e) KITTI (0018)

Figure IV. Divergence vs Distance Threshold at 20 Summarized Frames

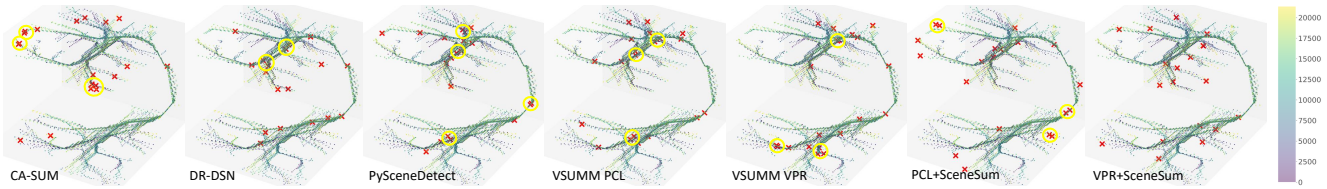


Figure V. **Selected keyframes in Habitat-Sim Dataset.** We summarize 20 keyframes of 7 baselines on scene *Goffs*. All frames are color-coded by temporal order. Summarized keyframes are marked with red crosses. Groups of frames that are geographically close to each other are circled in yellow.

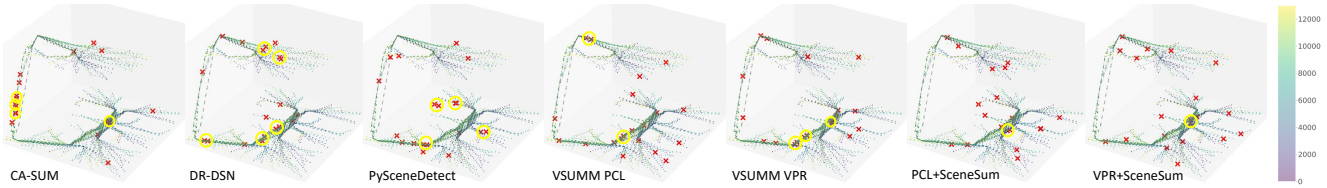


Figure VI. **Selected keyframes in Habitat-Sim Dataset.** We summarize 20 keyframes of 7 baselines on scene *Stilwell*. The baselines and annotations follow Fig. V

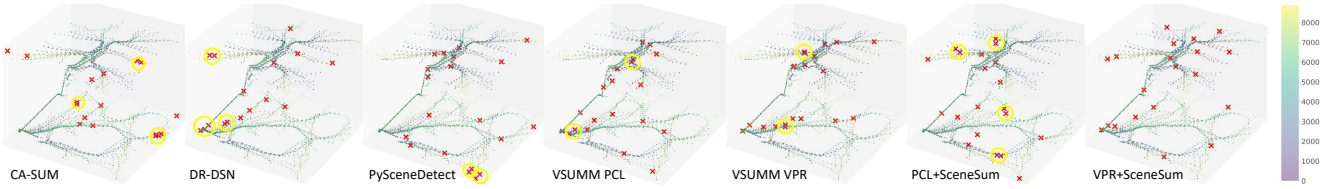


Figure VII. **Selected keyframes in Habitat-Sim Dataset.** We summarize 20 keyframes of 7 baselines on scene *Micanopy*. The baselines and annotations follow Fig. V

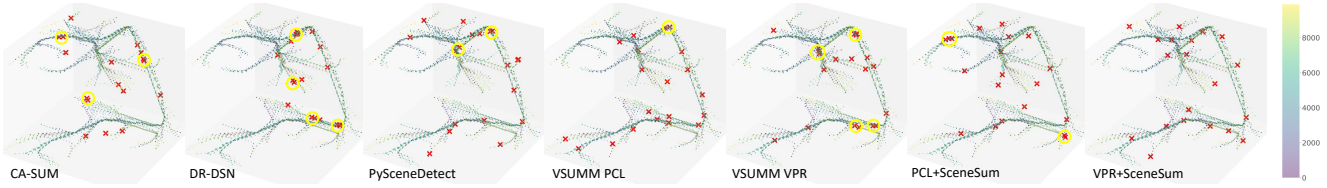


Figure VIII. **Selected keyframes in Habitat-Sim Dataset.** We summarize 20 keyframes of 7 baselines on scene *Spotswood*. The baselines and annotations follow Fig. V

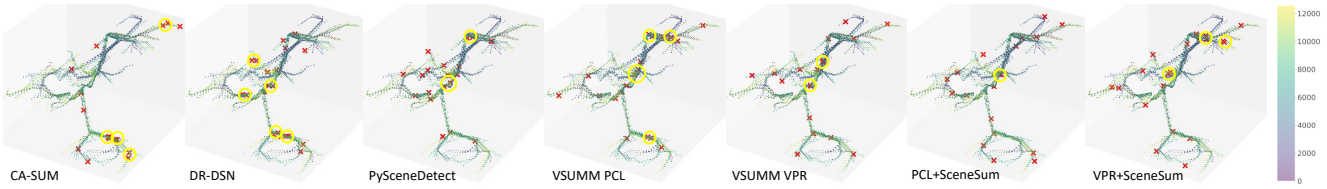


Figure IX. **Selected keyframes in Habitat-Sim Dataset.** We summarize 20 keyframes of 7 baselines on scene *Springhill*. The baselines and annotations follow Fig. V

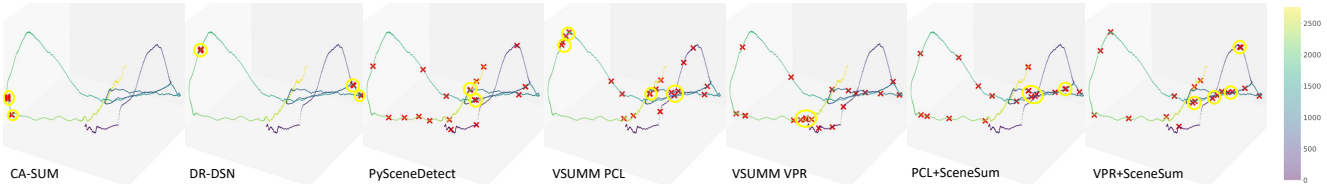


Figure X. **Selected keyframes in KITTI Dataset.** We summarize 20 keyframes of 7 baselines on scene *0018*. The baselines and annotations follow Fig. V

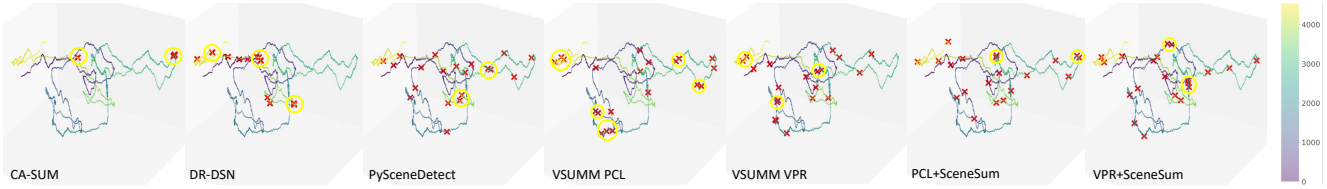


Figure XI. **Selected keyframes in KITTI Dataset.** We summarize 20 keyframes of 7 baselines on scene *0027*. The baselines and annotations follow Fig. V