DREAM: Dual-Memory Reasoning for Anticipatory 3D Perception in Autonomous Driving

Anonymous CVPR submission

Paper ID 12

Abstract

*Autonomous driving systems must not only interpret the current scene but also anticipate perceptual challenges such as occlusions, sparse observations, and recurring failure patterns. However, most existing LiDAR semantic segmentation methods operate in a frame-centric manner, processing each scan independently and discarding past contextual experience, which limits robustness under dynamic scene changes. We propose **DREAM**, a **D**ual-memory **REA**soning framework for continual LiDAR se**M**antic segmentation that reframes perception as an experience-driven and anticipatory process. DREAM maintains two complementary memory banks in a strictly online setting: a latent memory that stores compact semantic abstractions of previously observed scenes, and an error memory that records representations associated with uncertain predictions. At each timestep, relevant memory entries are retrieved via cosine similarity and integrated into the feature space through a lightweight modulation mechanism, enabling the model to reinforce consistent semantic patterns while suppressing recurring failure modes. The backbone remains frozen and no past scans are replayed, ensuring computational efficiency and bounded memory growth. Extensive experiments on multiple large-scale LiDAR benchmarks demonstrate that DREAM achieves state-of-the-art performance, with consistent improvements on dynamic and small-scale objects, highlighting the effectiveness of persistent and error-aware memory for robust long-horizon perception.*

1. Introduction

Semantic scene understanding is a core capability for autonomous systems operating in complex environments such as autonomous driving. Among perception tasks, semantic segmentation provides dense structured representations of the surrounding scene that are critical for downstream planning and decision-making. However, reliable autonomy requires perception systems that remain robust under challenging sensing conditions such as occlusions, sparse measurements, and dynamic scene changes. Despite strong progress in deep learning–based perception, most LiDAR semantic segmentation methods remain fundamentally reactive, processing each scan independently or within short temporal windows [7]. This frame-centric formulation prevents perception systems from leveraging past experience to anticipate recurring perceptual challenges.

In real-world deployment, LiDAR observations arrive as a continuous stream where each scan provides only a partial and noisy view of the environment. Measurements are often affected by occlusions, sensor noise, and long-range sparsity. Under such conditions, treating frames independently [4, 13, 21, 27, 29, 32] leads to brittle predictions and temporal inconsistencies, particularly for distant or rarely observed structures. In contrast, human perception relies on accumulated experience: past observations inform future interpretation and help anticipate recurring patterns of uncertainty [14]. Enabling perception models to accumulate and exploit such experience remains a key challenge for robust long-horizon scene understanding.

Three-dimensional representations provide a natural foundation for storing perceptual experience. Unlike image projections, 3D space preserves metric structure, allowing memory to maintain consistent object scales and spatial relationships across viewpoints. This enables stable semantic priors, more uniform representation across near and far regions, and improved reasoning about occlusions and motion. Regions that become temporarily unobserved due to occlusions can therefore be retained in memory without interference from newly observed surfaces. However, maintaining spatial memory alone does not fully address a key limitation of deployed perception systems: the tendency to make similar errors under recurring sensing conditions repeatedly.

Recent work has begun exploring memory-augmented perception for LiDAR segmentation. Earlier approaches, such as Cylinder3D [32] and Meta-RangeSeg [25], primarily focus on improving feature representations and spatial-temporal aggregation for semantic segmentation, but they

do not explicitly maintain persistent memory across observations. MemorySeg [12], for example, maintains a persistent voxel memory that aggregates information across streaming LiDAR scans and improves understanding of long-range and occluded regions. While effective, such approaches primarily accumulate spatial context and treat memory as a passive repository of past observations. In practice, perception systems frequently encounter recurring failure patterns, such as systematic uncertainty under severe sparsity or occlusion. Conventional architectures lack mechanisms to explicitly represent these failure modes and use them to guide future predictions.

To address this limitation, we propose **DREAM**, a dual-memory framework for continual LiDAR semantic segmentation that augments perception with both perceptual and error-aware memory. DREAM maintains two compact latent memory banks operating in a strictly online setting. A *latent perceptual memory* stores compressed semantic descriptors of previously observed scenes, while an *error memory* stores latent representations associated with uncertain predictions. At each timestep, a global latent descriptor extracted from the current scan retrieves relevant entries from both memory banks using cosine-similarity addressing. The retrieved latent and error memories are projected into the dense feature space and integrated through a lightweight modulation mechanism prior to classification. This allows the network to reinforce reliable semantic patterns while suppressing conditions that previously produced low-confidence predictions.

The proposed framework DREAM, processes one scan at a time, maintains bounded memory capacity, and avoids computationally expensive attention mechanisms or long-horizon backpropagation. The backbone encoder remains frozen, while lightweight memory modulation layers condition perception on retrieved experience. This design enables temporally aware semantic segmentation while maintaining efficiency suitable for real-time autonomous driving scenarios. The effectiveness of the proposed DREAM architecture is substantiated through comprehensive evaluations on three benchmark datasets: SemanticKITTI[2], nuScenes[3], and PandaSet[26].

In summary, the key contributions of this work are as follows:

- Proposal of DREAM, a Dual-Memory Reasoning for Anticipatory 3D Perception for Autonomous Driving.
- We introduce an error-aware memory mechanism that explicitly models recurring uncertainty patterns and helps mitigate future perceptual failures
- Proposal of a Dual Memory Modulation scheme for continual LiDAR semantic segmentation that integrates perceptual and error-aware memories to enable experience-conditioned perception, one of its first kinds.

2. Related Works

In this section, we review existing LiDAR semantic segmentation methods and discuss approaches that incorporate temporal reasoning for streaming LiDAR perception.

2.1. LiDAR Semantic Segmentation

LiDAR semantic segmentation methods can be broadly categorized based on the underlying data representation: point-based [11, 15, 16, 24], projection-based [6, 27, 31], voxel-based [4, 10, 32], and hybrid approaches that combine multiple representations [21, 28].

Point-based methods [11, 15, 16, 24] operate directly on raw point clouds and learn point-wise features using permutation-invariant networks. While effective, many of these approaches rely on aggressive downsampling to maintain computational efficiency. For example, KPConv [24] introduces deformable point convolutions that significantly improve segmentation performance, but at the cost of increased computational complexity.

Projection-based methods [6, 27, 31] project 3D point clouds into structured 2D representations such as range-view or bird's-eye-view images. These approaches benefit from efficient 2D convolutional networks and often achieve real-time inference. However, the projection process introduces geometric distortions and information loss, which can limit segmentation accuracy.

Voxel-based approaches [4, 10, 32] convert point clouds into structured 3D grids and process them using sparse convolutional networks [5, 22]. These methods have demonstrated strong performance by preserving spatial structure while enabling efficient computation. Nevertheless, discretization into voxels may lead to loss of fine geometric details when voxel resolution is reduced.

To leverage the strengths of different representations, recent works combine multiple modalities [21, 28–30]. For instance, SPVNAS [21] integrates point-based and voxel-based features within a unified architecture, while RPVNet [28] employs a tri-branch network that jointly processes range-view, point-wise, and voxel features. These hybrid approaches improve segmentation accuracy by exploiting complementary geometric representations.

2.2. Temporal LiDAR Reasoning

To incorporate temporal context, several works attempt to aggregate information from multiple LiDAR frames. Early approaches [18, 20, 25] integrate a small number of past frames to enhance segmentation accuracy. For example, SpSequenceNet [20] introduces cross-frame attention and interpolation mechanisms to combine features from consecutive scans, while MetaRangeSeg [25] incorporates residual depth information from previous frames. Although these methods improve short-term consistency, they oper-

ate within limited temporal windows and discard historical context beyond a few frames.

Other works explore recurrent formulations for temporal state maintenance. Duerr et al. [8] propose a recurrent model that aligns range-view features across streaming scans, while MemorySeg [12] introduces a persistent latent voxel memory for improved segmentation in long-range and occluded regions. StrObe [9] similarly maintains a multi-scale BEV memory for incremental object detection. However, range-view and BEV representations are prone to projection artifacts and occlusion conflicts, and none of these approaches explicitly model recurring perceptual failures.

In contrast, DREAM augments LiDAR segmentation with persistent latent memory designed for long-horizon perception. Rather than storing raw observations or geometric maps, our framework maintains compact representations capturing both semantic experience and historical uncertainty. Retrieving relevant past contexts while suppressing recurring failure patterns enables robust, anticipatory perception, making DREAM one of the first approaches to handle LiDAR segmentation entirely within the latent space.

3. Methodology

Most LiDAR segmentation networks process each scan independently, ignoring temporal context critical for handling occlusions, sparsity, and recurring scene configurations. To address this, we introduce **DREAM** (Fig. 1), a dual-memory reasoning framework for continual LiDAR semantic segmentation. DREAM treats segmentation as an experience-driven process where structured latent memories evolve, enabling anticipatory reasoning over streaming data in a strictly online regime.

3.1. Problem Formulation

Refer to Fig. 1 for the detailed architecture diagram of the proposed DREAM framework. We model segmentation as a recurrent reasoning process. At each timestep t, the model receives a LiDAR scan X_t and produces semantic predictions $\hat{Y}_t$ conditioned on both the current input and an accumulated memory state $\mathcal{M}_{t-1}$. The system operates online: past scans are never revisited, and only compact latent representations are retained.

Given a temporally ordered sequence $\{X_t\}_{t=1}^{T}$, the model defines the mapping

$$\hat{Y}_t = f(X_t, \mathcal{M}_{t-1}; \theta), \tag{1}$$

where θ denotes learnable parameters. After prediction, the memory is updated as

$$\mathcal{M}_t = \mathcal{U}(\mathcal{M}_{t-1}, X_t, \hat{Y}_t). \tag{2}$$

The central challenge is to design a compact yet expressive representation for $\mathcal{M}_t$, along with mechanisms for retrieval and integration into perception.

3.2. Perceptual Encoding

Each raw LiDAR scan X_t is projected into a range-view encoder $I_t \in \mathbb{R}^{5\times H\times W}$ containing range, spatial coordinates (x, y, z), and intensity. A convolutional backbone $g(\cdot)$ extracts hierarchical features:

$$F_t, Z_t = g(I_t), \tag{3}$$

where $F_t \in \mathbb{R}^{C\times H'\times W'}$ are dense spatial features for segmentation and $Z_t \in \mathbb{R}^{D\times h\times w}$ are higher-level latent features. The backbone is frozen to stabilize representations and decouple long-term reasoning from low-level feature drift.

We summarize the latent representation via global average pooling:

$$\bar{z}_t = \mathrm{GAP}(Z_t) \in \mathbb{R}^D, \tag{4}$$

which serves as the interface between perception and memory.

3.3. Latent Memory Representation

LiDAR observations are inherently partial, with objects often sparsely sampled, occluded, or incomplete across individual scans. Relying solely on the current frame leads to unstable predictions, particularly for distant or rarely observed objects. To address this, we introduce a latent perceptual memory that stores compressed semantic descriptors from past observations, enabling the model to retrieve relevant contextual information during inference.

Formally, we maintain a fixed-capacity latent memory bank

$$\mathcal{M}_t^L = \{m_i^L \in \mathbb{R}^D\}_{i=1}^{K_L}, \tag{5}$$

where each entry encodes a compressed semantic abstraction of a previously observed scene. Given the global descriptor $\bar{z}_t$ extracted from the current scan, the system retrieves relevant past contexts through content-based addressing using cosine similarity:

$$s_i = \frac{\bar{z}_t^\top m_i^L}{\|\bar{z}_t\|\|m_i^L\|}. \tag{6}$$

The top-k most similar entries are selected to form the retrieved latent context

$$M_t^L \subset \mathcal{M}_{t-1}^L. \tag{7}$$

By retrieving semantically related past descriptors, the model can incorporate previously observed contextual information when interpreting the current scan. This mechanism provides a compact form of long-term contextual memory that improves robustness to sparse observations, partial visibility, and recurring scene configurations.

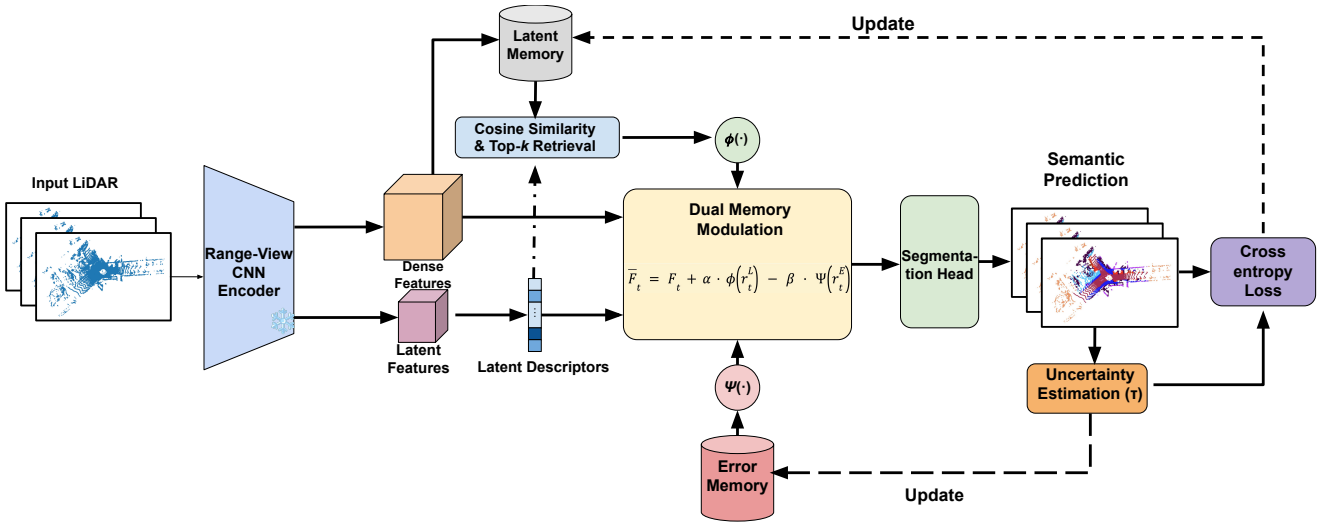

Figure 1. Overview of the proposed DREAM architecture. The network processes each LiDAR scan through hierarchical feature encoding and dual-memory reasoning modules that leverage past experience to guide current perception.

3.4. Error Memory Representation

Complementary to perceptual memory, we introduce an error memory bank

$$\mathcal{M}_t^E = \{m_j^E \in \mathbb{R}^D\}_{j=1}^{K_E}, \tag{8}$$

which explicitly encodes latent representations associated with unreliable predictions. Let $p_t(c|u)$ denote the predicted class probability at location u. Uncertain regions are identified using a confidence threshold τ:

$$\mathcal{E}_t = \{Z_t(u) \mid \max_c p_t(c|u) < \tau\}. \tag{9}$$

Latent vectors from $\mathcal{E}_t$ are aggregated using mean pooling to produce a compact descriptor $\bar{e}_t \in \mathbb{R}^D$, which is then inserted into $\mathcal{M}_t^E$. While $\mathcal{M}_t^L$ captures stable semantic abstractions, $\mathcal{M}_t^E$ accumulates representations of recurring perceptual failures. By explicitly modeling these patterns, the system can anticipate similar uncertainty conditions in future observations and adjust its predictions accordingly.

3.5. Dual Memory Modulation

The retrieved latent memories M_t^L and error memories M_t^E are integrated into the dense feature space prior to classification. Error memory retrieval follows the same content-based addressing mechanism as the latent memory, where cosine similarity between the global descriptor $\bar{z}_t$ and stored error memory entries is computed and the most similar entry is selected using top-1 retrieval. Let $\phi(\cdot)$ and $\psi(\cdot)$ denote learnable projection functions mapping latent vectors to the dense feature space. We define the modulated features as

$$\tilde{F}_t = F_t + \alpha \cdot \phi(M_t^L) - \beta \cdot \psi(M_t^E), \tag{10}$$

where α and β control the strength of positive reinforcement and error suppression, respectively. In practice, α and β are treated as fixed hyperparameters selected via validation to balance contextual enhancement against over-suppression. Empirically, moderate values ensure stable training and consistent gains across datasets. The additive formulation preserves spatial alignment while injecting global semantic priors derived from past experience. Crucially, modulation occurs before classification, allowing retrieved memories to influence perception before predictions are produced.

3.6. Segmentation and Optimization

The modulated features $\tilde{F}_t$ are fed to a lightweight segmentation head $h(\cdot)$:

$$\hat{Y}_t = h(\tilde{F}_t). \tag{11}$$

Training minimizes a class-weighted cross-entropy loss

$$\mathcal{L}_t = \mathrm{CE}(\hat{Y}_t, Y_t), \tag{12}$$

where only the segmentation head and modulation layers are optimized, while the backbone remains frozen.

3.7. Online Memory Update

After prediction, the global descriptor $\bar{z}_t$ is inserted into $\mathcal{M}_t^L$, and the aggregated error descriptor $\bar{e}_t$ derived from uncertain latents $\mathcal{E}_t$ is inserted into $\mathcal{M}_t^E$. Both memories operate with fixed capacity using a replacement strategy (e.g., FIFO), ensuring bounded memory growth while retaining long-term experience in compressed form.

3.8. Memory-Conditioned Perception

DREAM formulates LiDAR semantic segmentation as a perception problem conditioned on accumulated context

from streaming observations. Rather than processing each scan independently, the model maintains two latent memory banks, $\mathcal{M}_t^L$ and $\mathcal{M}_t^E$, which store compact representations of previously observed semantic structures and historical low-confidence predictions. At each timestep, a global latent descriptor retrieves relevant entries through content-based addressing, and the retrieved memories modulate the dense feature representation prior to classification. This mechanism allows the model to incorporate contextual information from past observations while maintaining a feed-forward inference pipeline. As a result, DREAM integrates long-horizon contextual cues into LiDAR perception in a computationally efficient manner, enabling temporally aware semantic segmentation in streaming autonomous driving scenarios.

4. Experimental Setup

We evaluate the proposed framework on three large-scale autonomous driving datasets: SemanticKITTI [2], nuScenes [3], and PandaSet [26]. These datasets are collected using different LiDAR sensors and capture diverse urban environments, making them suitable for evaluating perception systems operating in streaming conditions. Our experiments focus on assessing the effectiveness of persistent memory for long-horizon perception and its ability to improve robustness under sparse observations, occlusions, and dynamic scene changes. In addition to comparisons with existing methods, we perform ablation studies to analyze the contribution of each component of the proposed dual-memory design.

4.1. Datasets

We evaluate on three large-scale autonomous driving datasets. **SemanticKITTI** [2] comprises 22 sequences captured with a Velodyne HDL-64E in Germany, with sequences 0–10 for training and 11–21 for testing. Labels are mapped to 19 classes (single-scan) and 25 classes (multi-scan), where we primarily report multi-scan results as our method models temporal context. **nuScenes** [3] contains 1000 driving scenes from Boston and Singapore (700/150/150 train/val/test split), captured with a Velodyne HDL-32E at 2 Hz, with 31 classes mapped to 16 categories. **PandaSet** [26] provides temporally continuous sequences recorded in Silicon Valley using a Hesai Pandar64, with 103 sequences and fine-grained classes grouped into 14 categories following [8].

4.2. Evaluation Protocol

During evaluation, each sequence is processed in temporal order from the first frame to the last. The memory state is updated online during inference, allowing the model to accumulate experience across the sequence. Performance is evaluated using Intersection-over-Union (IoU) [17], with mean IoU (mIoU) across all semantic classes reported as the primary metric. In addition, we report seen-class mIoU, which averages IoU only over classes present in the evaluated sequence, as well as class-wise IoU scores for detailed analysis.

4.3. Implementation Details

Training proceeds in two stages: the segmentation head is first trained for 50 epochs in a single-scan setting, after which the dual-memory mechanism and modulation layers are fine-tuned for 20 additional epochs in an online recurrent manner. The backbone remains frozen throughout. Each LiDAR scan is projected into a 64×1024 range-view with five channels (range, x, y, z, intensity). A pre-trained ResNet-18 extracts dense features from the third residual stage and a 512-dimensional global descriptor from the final stage for memory retrieval. The latent and error memory banks store K_L=64 and K_E=32 vectors (D=512) respectively, using cosine similarity for top-1 retrieval and FIFO replacement. The error memory is updated via low-confidence regions. Optimization uses Adam (lr $= 5 \times 10^{-4}$) with class-weighted cross-entropy and mixed precision training. Augmentations include global scaling, random translation, and vertical-axis rotation. The model is trained on 8 NVIDIA A100 GPUs (80GB) using distributed data parallelism.

4.4. Quantitative Analysis

We evaluate DREAM on SemanticKITTI, nuScenes, and PandaSet, with results summarized in Tables 1, 2, and 3.

On SemanticKITTI (Tab. 1), DREAM achieves 61.7% mIoU, outperforming MemorySeg (58.3%) and MetaRangeSeg (49.7%). Gains are most pronounced in dynamic categories (moving cars, bicyclists, pedestrians, motorcyclists) and structural classes (building, trunk, terrain), reflecting the benefit of temporal reasoning via the dual-memory design.

On nuScenes (Tab. 2), DREAM achieves **81.8%** mIoU, with notable improvements in barrier, pedestrian, and traffic cone categories, suggesting the memory mechanism effectively addresses sparsity and scale challenges.

On PandaSet (Tab. 3), DREAM achieves **71.0%** mIoU, performing consistently across foreground classes (truck, person, traffic sign) and background classes (sidewalk, terrain), confirming generalization across sensor configurations and urban environments.

Across all benchmarks, persistent latent memory enables robust segmentation under sparse observations, occlusions, and dynamic scene changes.

4.5. Comparison Against State-of-the-Art

We compare DREAM with prior LiDAR-only methods across all three benchmarks (Tab. 1–3).

Method	mIoU	Car	Bicycle	Motorcycle	Truck	Other Vehicle	Person	Bicyclist	Motorcyclist	Road	Parking	Sidewalk	Other Ground	Building	Fence	Vegetation	Trunk	Terrain	Pole	Traffic Sign	Car (m)	Bicyclist (m)	Person (m)	Motorcyclist (m)	Other Vehicle (m)	Truck (m)
TangentConv [23]	34.1	84.9	2.0	18.2	21.1	18.5	1.6	0.0	0.0	83.9	38.3	64.0	15.3	85.8	49.1	79.5	43.2	56.7	36.4	31.2	40.3	1.1	6.4	1.9	30.1	42.2
DarkNet53Seg [1]	41.6	84.1	30.4	32.9	20.2	20.7	7.5	0.0	0.0	91.6	64.9	75.3	27.5	85.2	56.5	78.4	50.7	64.8	38.1	53.3	61.5	14.1	15.2	0.2	28.9	37.8
KPConv [24]	51.2	93.7	44.9	47.2	42.5	38.6	21.6	0.0	0.0	86.5	58.4	70.5	26.7	90.8	64.5	84.6	70.3	66.0	57.0	53.9	69.4	67.4	67.5	47.2	4.7	5.8
Cylinder3D [32]	52.5	94.6	67.6	63.8	41.3	38.8	12.5	1.7	0.2	90.7	**65.0**	74.5	**32.3**	92.6	66.0	**85.8**	72.0	68.9	63.1	61.4	74.9	68.3	65.7	11.9	0.1	0.0
SpSequenceNet [20]	43.1	88.5	24.0	26.2	29.2	22.7	6.3	0.0	0.0	90.1	57.6	73.9	27.1	91.2	66.8	84.0	66.0	65.7	50.8	48.7	53.2	41.2	26.2	36.2	2.3	0.1
TemporalLidarSeg [8]	47.0	92.1	47.7	40.9	39.2	35.0	14.4	0.0	0.0	91.8	59.6	**75.8**	23.2	89.8	63.8	82.3	62.5	64.7	52.6	60.4	68.2	42.8	40.4	12.9	12.4	2.1
TemporalLatticeNet [19]	47.1	91.6	35.4	36.1	26.9	23.0	9.4	0.0	0.0	91.5	59.3	75.3	27.5	89.6	65.3	84.6	66.7	70.4	57.2	60.4	59.7	41.7	51.0	48.8	5.9	0.0
Meta-RangeSeg [25]	49.7	90.8	50.0	49.5	29.5	34.8	16.6	0.0	0.0	90.8	62.9	74.8	26.5	89.8	62.1	82.8	65.7	66.5	56.2	64.5	69.0	60.4	57.9	22.0	16.6	2.6
MemorySeg [12]	58.3	94.0	68.3	68.8	51.3	40.9	27.0	0.3	2.8	89.9	64.3	74.8	29.2	92.2	69.3	84.8	75.1	70.1	65.5	**68.5**	71.7	**74.4**	**71.7**	73.9	15.1	13.6
DREAM (Ours)	**61.7**	**97.6**	**69.1**	**68.9**	**55.7**	**43.1**	**29.0**	**1.7**	**3.2**	**93.5**	62.6	74.8	29.2	92.2	69.3	84.8	**77.3**	**71.0**	**65.6**	68.4	71.5	70.3	71.5	**74.3**	**37.6**	**49.7**

Table 1. Comparison with state-of-the-art LiDAR semantic segmentation methods on the **SemanticKITTI** multi-scan test benchmark. (m) denotes moving classes. Metrics are reported in mIoU (%).

Method	mIoU	FW-mIoU	Barrier	Bicycle	Bus	Car	Construction	Motorcycle	Pedestrian	Traffic Cone	Trailer	Truck	Drivable	Other Flat	Sidewalk	Terrain	Manmade	Vegetation
PolarNet [31]	69.4	87.4	72.2	16.8	77.0	86.5	51.1	69.7	64.8	54.1	69.7	63.5	96.6	67.1	77.7	72.1	87.1	84.5
Cylinder3D [32]	77.2	89.9	82.8	29.8	84.3	89.4	63.0	**79.3**	77.2	73.4	84.6	69.1	97.7	**70.2**	80.3	75.5	90.4	87.6
SPVCNN [21]	77.4	89.7	80.0	30.0	91.9	90.8	64.7	79.0	75.6	70.9	81.0	74.6	97.4	69.2	80.0	76.1	89.3	87.1
$(AF)^2$-S3Net [4]	78.3	88.5	78.9	**52.2**	89.9	84.2	**77.4**	74.3	77.3	72.0	83.9	73.8	97.1	66.5	77.5	74.0	87.7	86.8
MemorySeg [12]	80.6	91.4	84.9	40.2	91.2	92.4	71.2	73.5	**85.9**	77.8	88.0	76.4	**97.9**	69.0	81.2	77.6	92.6	89.7
DREAM (Ours)	**81.8**	**91.4**	**86.3**	40.2	**92.3**	**94.5**	**79.4**	73.9	84.7	**78.0**	**88.4**	**79.1**	96.2	68.3	**81.9**	76.3	**93.5**	**90.4**

Table 2. Comparison with state-of-the-art LiDAR-only semantic segmentation methods on the **nuScenes** test set [3]. Metrics are reported in mIoU (%). Best results are highlighted in **bold**.

On SemanticKITTI (Tab. 1), DREAM outperforms Cylinder3D (52.5%), Meta-RangeSeg (49.7%), and MemorySeg (58.3%), with the largest gains in dynamic classes (moving cars, bicyclists, pedestrians, motorcyclists), demonstrating effective temporal modeling via dual memory.

On nuScenes (Tab. 2), DREAM sets a new state-of-the-art among LiDAR-only methods, surpassing MemorySeg (80.6%) and $(AF)^2$-S3Net (78.3%), with notable improvements in small-object categories such as barrier, pedestrian, and traffic cone.

On PandaSet (Tab. 3), DREAM achieves the best overall performance over MemorySeg (70.3%) and SPVCNN (64.7%), with consistent gains across foreground and background categories including truck, person, and sidewalk.

Across all datasets, persistent latent memory reasoning consistently outperforms both single-frame and short-term temporal fusion approaches.

4.6. Qualitative Analysis

4.6.1. Temporal Consistency Analysis

Fig. 3 presents qualitative results across four consecutive LiDAR frames ($t-2, t-1, t, t+1$). The top row shows the input LiDAR intensity visualization, while the middle and bottom rows illustrate the ground-truth semantic labels and predictions from DREAM, respectively.

Despite the sparse and partially observed nature of individual LiDAR scans, the predicted semantic structure remains stable across time. In particular, large structural classes such as road and vegetation are consistently segmented across frames. This temporal stability demonstrates the benefit of the proposed memory-based reasoning mechanism, which allows the model to leverage historical context when interpreting the current observation.

4.6.2. Occlusion Recovery

Fig. 4 illustrates the behavior of the DREAM framework under partial occlusions. The top row shows the input LiDAR intensity visualization, where the highlighted region exhibits varying point density across consecutive frames due to occlusion and sparse observations.

The middle row shows the ground-truth semantic labels, while the bottom row presents predictions produced by DREAM. Despite incomplete observations in individual LiDAR scans, DREAM maintains temporally consistent semantic predictions with 61.7% mIoU and 81.8% mIoU on SemanticKITTI and nuScenes datasets, respectively, across frames, demonstrating the proposed memory mechanism's

Method	mIoU	Car	Bicycle	Motorcycle	Truck	Other Vehicle	Person	Road	Road Barriers	Sidewalk	Building	Vegetation	Terrain	Background	Traffic Sign
SqueezeSegv3 [27]	55.7	92.8	24.1	18.0	36.5	54.3	63.0	91.1	11.9	71.3	86.2	85.0	61.3	63.2	20.6
SalsaNext [6]	57.8	92.1	40.7	31.7	28.7	56.2	69.0	90.0	22.6	67.1	85.6	83.4	58.5	63.3	20.6
TemporalLidarSeg [8]	60.0	93.7	33.6	38.0	37.1	59.9	72.0	91.1	14.6	70.6	88.2	88.4	63.8	68.4	20.7
SPVCNN [21]	64.7	95.8	38.1	46.3	44.0	74.1	78.6	91.2	**28.3**	70.3	87.2	87.5	61.5	67.5	35.6
MemorySeg [12]	70.3	97.2	60.2	58.4	62.9	74.3	82.6	92.1	27.7	74.1	89.4	90.7	**64.9**	72.8	36.4
DREAM (Ours)	**71.0**	**97.8**	**60.5**	**59.1**	**64.0**	**75.4**	**82.9**	**93.6**	28.1	**75.4**	**90.7**	**89.2**	64.5	**74.6**	**37.1**

Table 3. Comparison with state-of-the-art LiDAR semantic segmentation methods on the PandaSet test set [26]. Metrics are reported in mIoU (%). Best results are highlighted in **bold**.

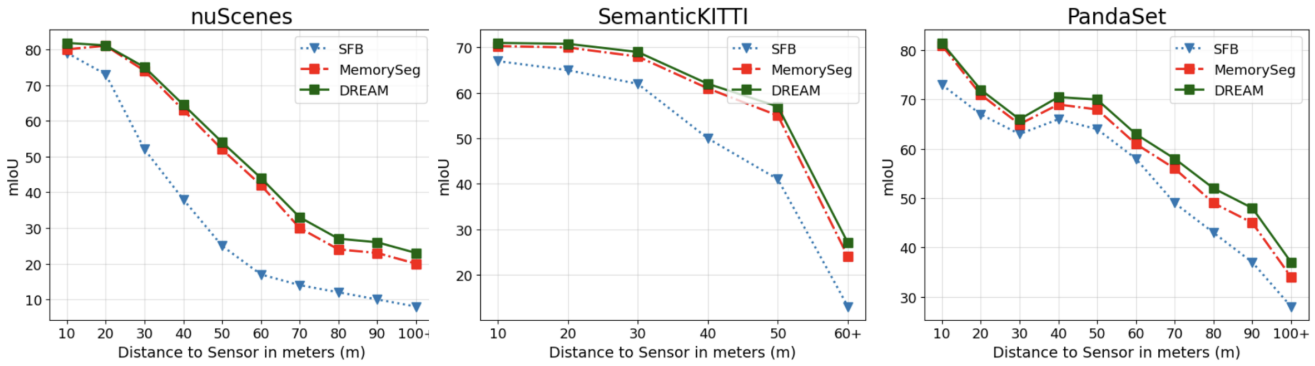

Figure 2. Distance-based performance comparison of SFB (dooted blue), MemorySeg (solid red), and DREAM (solid green) across nuScenes, SemanticKITTI, and PandaSet. DREAM demonstrates consistently improved robustness at mid- and long-range distances, with gains increasing as point sparsity and occlusion become more severe.

Method	Params (M)	FPS
SFB	14.2	29
DREAM (Full)	16.8	25

Table 4. Parameter count and inference speed comparison on SemanticKITTI. DREAM introduces a modest increase in the number of parameters and runtime while maintaining real-time performance.

ability to recover semantic structure in occluded or sparsely observed regions.

4.7. Ablation Studies

4.7.1. Importance of the Memory

To evaluate the impact of persistent memory, we compare DREAM against a single-frame baseline (SFB) that removes both latent perceptual memory and error memory components. The comparison is performed using the distance-based evaluation shown in Fig. 2. Results indicate that incorporating memory significantly improves segmentation performance, particularly at larger distances where LiDAR measurements are sparse and noisy. This demonstrates that maintaining a persistent latent representation of past observations helps the model recover semantic context that may be missing in individual frames.

4.7.2. Distance-based Analysis

Fig. 2 highlights the benefit of persistent dual memory at longer ranges, where LiDAR point clouds become increasingly sparse. We compare DREAM against a single-frame baseline (SFB) that removes both memory modules.

DREAM consistently outperforms the SFB across all distance bins, with the most pronounced gains at 80m: 26.47% mIoU on nuScenes and 52.33% mIoU on PandaSet, where point density is lowest. The progressive improvement with distance confirms that persistent memory becomes more beneficial as perceptual difficulty increases, enhancing semantic consistency through compact scene representations and explicit uncertainty modeling.

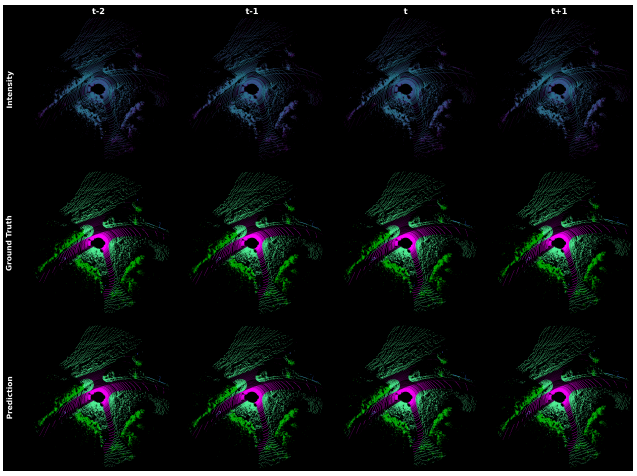

Figure 3. Temporal qualitative results across consecutive LiDAR frames.Columns correspond to four consecutive frames ($t-2$, $t-1$, t, $t+1$). The top row shows the input LiDAR intensity visualization, the middle row shows the ground-truth semantic labels, and the bottom row shows the predictions produced by the proposed DREAM framework.

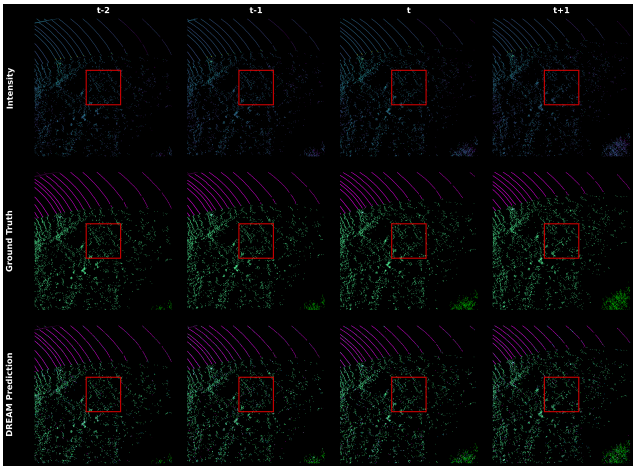

Figure 4. Qualitative example illustrating occlusion recovery across consecutive frames. As objects become partially occluded, DREAM maintains stable semantic predictions by leveraging temporal context stored in latent memory.

4.7.3. Effect of Error Memory

We evaluate the contribution of the proposed error memory on the SemanticKITTI multi-scan validation set (Tab. 5). Compared to the single-frame baseline (SFB), introducing latent perceptual memory improves mIoU from 78.6% to 81.0%, highlighting the benefit of persistent semantic aggregation. Adding the error memory further increases performance to 81.8%, with more pronounced gains in long-range regions, where mIoU improves from 24.8% to 27.3%. These results indicate that explicitly modeling recurrent low-confidence patterns complements spatial memory and enhances robustness under sparse and occluded observations.

Method	mIoU (%)
SFB	58.4
DREAM w/o Error Memory	60.9
DREAM (Full)	**61.5**

Table 5. Ablation study on the SemanticKITTI multi-scan validation set. Latent perceptual memory accounts for the majority of the improvement over the single-frame baseline (SFB). The proposed error memory further enhances performance, particularly in long-range regions (60m+), where point sparsity and occlusions are more severe.

4.7.4. Parameter Count.

DREAM introduces only a modest computational overhead compared to the single-frame baseline (SFB). As shown in Tab. 4, the proposed dual-memory design increases the parameter count from 14.2M to 16.8M (approximately 18%) and reduces inference speed from 29 FPS to 25 FPS (approximately 15%), while maintaining real-time performance.

5. Conclusion

We presented DREAM, a dual-memory reasoning framework for continual LiDAR semantic segmentation that reformulates perception as a temporally grounded process. DREAM maintains two complementary memory banks: a perceptual memory that accumulates stable semantic abstractions and an error memory that records recurring failure patterns. Retrieved memories modulate dense features before classification, reinforcing consistent semantics while suppressing previously observed errors. The framework operates strictly online, avoids scan replay, and keeps the backbone frozen, yielding a bounded-memory design suitable for real-world deployment. Experiments on SemanticKITTI, nuScenes, and PandaSet demonstrate state-of-the-art performance, with the most significant gains in dynamic and small-object categories. DREAM advances LiDAR segmentation from stateless prediction to memory-augmented reasoning, offering a scalable foundation for long-horizon autonomous perception.

References

[1] Jens Behley, Martin Garbade, Andres Milioto, Jan Quenzel, Sven Behnke, C. Stachniss, and Juergen Gall. A dataset for semantic segmentation of point cloud sequences. *ArXiv*, abs/1904.01416, 2019. 6

[2] J. Behley, M. Garbade, A. Milioto, J. Quenzel, S. Behnke, C. Stachniss, and J. Gall. SemanticKITTI: A Dataset for Semantic Scene Understanding of LiDAR Sequences. In *Proc.*

of the IEEE/CVF International Conf. on Computer Vision (ICCV), 2019. 2, 5

[3] Holger Caesar, Varun Bankiti, Alex H. Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yuxin Pan, Giancarlo Baldan, and Oscar Beijbom. nuscenes: A multimodal dataset for autonomous driving. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11618–11628, 2019. 2, 5, 6

[4] Ran Cheng, Ryan Razani, Ehsan Moeen Taghavi, Enxu Li, and Bingbing Liu. (af)2-s3net: Attentive feature fusion with adaptive feature selection for sparse semantic segmentation network. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12542–12551, 2021. 1, 2, 6

[5] Christopher Bongsoo Choy, JunYoung Gwak, and Silvio Savarese. 4d spatio-temporal convnets: Minkowski convolutional neural networks. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3070–3079, 2019. 2

[6] Tiago Cortinhal, George Tzelepis, and Eren Erdal Aksoy. Salsanext: Fast semantic segmentation of lidar point clouds for autonomous driving. *ArXiv*, abs/2003.03653, 2020. 2, 7

[7] Ayush Dewan and Wolfram Burgard. Deeptemporalseg: Temporally consistent semantic segmentation of 3d lidar scans. *2020 IEEE International Conference on Robotics and Automation (ICRA)*, pages 2624–2630, 2019. 1

[8] Fabian Duerr, Mario Pfaller, Hendrik Weigel, and Jürgen Beyerer. Lidar-based recurrent 3d semantic segmentation with temporal memory alignment. *2020 International Conference on 3D Vision (3DV)*, pages 781–790, 2020. 3, 5, 6, 7

[9] Davi Frossard, Simon Suo, Sergio Casas, James Tu, Rui Hu, and Raquel Urtasun. Strobe: Streaming object detection from lidar packets. *ArXiv*, abs/2011.06425, 2020. 3

[10] Yuenan Hou, Xinge Zhu, Yuexin Ma, Chen Change Loy, and Yikang Li. Point-to-voxel knowledge distillation for lidar semantic segmentation. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8469–8478, 2022. 2

[11] Qingyong Hu, Bo Yang, Linhai Xie, Stefano Rosa, Yulan Guo, Zhihua Wang, Agathoniki Trigoni, and A. Markham. Randla-net: Efficient semantic segmentation of large-scale point clouds. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11105–11114, 2019. 2

[12] Enxu Li, Sergio Casas, and Raquel Urtasun. Memoryseg: Online lidar semantic segmentation with a latent memory. *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 745–754, 2023. 2, 3, 6, 7

[13] Andres Milioto, Ignacio Vizzo, Jens Behley, and C. Stachniss. Rangenet ++: Fast and accurate lidar semantic segmentation. *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 4213–4220, 2019. 1

[14] Alireza Modirshanechi, Sophia Becker, Johanni Brea, and Wulfram Gerstner. Surprise and novelty in the brain. *Current Opinion in Neurobiology*, 82:102758, 2023. 1

[15] C. Qi, Hao Su, Kaichun Mo, and Leonidas J. Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 77–85, 2016. 2

[16] C. Qi, L. Yi, Hao Su, and Leonidas J. Guibas. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. *ArXiv*, abs/1706.02413, 2017. 2

[17] Seyed Hamid Rezatofighi, Nathan Tsoi, JunYoung Gwak, Amir Sadeghian, Ian D. Reid, and Silvio Savarese. Generalized intersection over union: A metric and a loss for bounding box regression. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 658–666, 2019. 5

[18] Peer Schütt, Radu Alexandru Rosu, and Sven Behnke. Abstract flow for temporal semantic segmentation on the permutohedral lattice. *2022 International Conference on Robotics and Automation (ICRA)*, pages 5139–5145, 2022. 2

[19] Peer Schutt, Radu Alexandru Rosu, and Sven Behnke. Abstract flow for temporal semantic segmentation on the permutohedral lattice. In *2022 International Conference on Robotics and Automation (ICRA)*, page 5139–5145. IEEE Press, 2022. 6

[20] Hanyu Shi, Guosheng Lin, Hao Wang, Tzu-Yi Hung, and Zhenhua Wang. Spsequencenet: Semantic segmentation network on 4d point clouds. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 2, 6

[21] Haotian Tang, Zhijian Liu, Shengyu Zhao, Yujun Lin, Ji Lin, Hanrui Wang, and Song Han. Searching efficient 3d architectures with sparse point-voxel convolution. *ArXiv*, abs/2007.16100, 2020. 1, 2, 6, 7

[22] Haotian Tang, Zhijian Liu, Xiuyu Li, Yujun Lin, and Song Han. Torchsparse: Efficient point cloud inference engine. *ArXiv*, abs/2204.10319, 2022. 2

[23] Maxim Tatarchenko, Jaesik Park, Vladlen Koltun, and Qian-Yi Zhou. Tangent convolutions for dense prediction in 3d. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3887–3896, 2018. 6

[24] Hugues Thomas, C. Qi, Jean-Emmanuel Deschaud, Beatriz Marcotegui, François Goulette, and Leonidas J. Guibas. Kpconv: Flexible and deformable convolution for point clouds. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 6410–6419, 2019. 2, 6

[25] Song Wang, Jianke Zhu, and Ruixiang Zhang. Meta-rangeseg: Lidar sequence semantic segmentation using multiple feature aggregation. *IEEE Robotics and Automation Letters*, 7:9739–9746, 2022. 1, 2, 6

[26] Pengchuan Xiao, Zhenlei Shao, Steven Hao, Zishuo Zhang, Xiaolin Chai, Judy Jiao, Zesong Li, Jian Wu, Kai Sun, Kun Jiang, Yunlong Wang, and Diange Yang. Pandaset: Advanced sensor suite dataset for autonomous driving. *2021 IEEE International Intelligent Transportation Systems Conference (ITSC)*, pages 3095–3101, 2021. 2, 5, 7

[27] Chenfeng Xu, Bichen Wu, Zining Wang, Wei Zhan, Péter Vajda, Kurt Keutzer, and Masayoshi Tomizuka. Squeezesegv3: Spatially-adaptive convolution for efficient point-

cloud segmentation. In *European Conference on Computer Vision*, 2020. 1, 2, 7

[28] Jianyun Xu, Ruixiang Zhang, Jian Dou, Yushi Zhu, Jie Sun, and Shiliang Pu. Rpvnet: A deep and efficient range-point-voxel fusion network for lidar point cloud segmentation. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 16004–16013, 2021. 2

[29] Xu Yan, Jiantao Gao, Chaoda Zheng, Chaoda Zheng, Ruimao Zhang, Shenghui Cui, and Zhen Li. 2dpass: 2d priors assisted semantic segmentation on lidar point clouds. In *European Conference on Computer Vision*, 2022. 1

[30] Maosheng Ye, Shuangjie Xu, Tongyi Cao, and Qifeng Chen. Drinet: A dual-representation iterative learning network for point cloud segmentation. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 7427–7436, 2021. 2

[31] Yang Zhang, Zixiang Zhou, Philip David, Xiangyu Yue, Zerong Xi, and Hassan Foroosh. Polarnet: An improved grid representation for online lidar point clouds semantic segmentation. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9598–9607, 2020. 2, 6

[32] Xinge Zhu, Hui Zhou, Tai Wang, Fangzhou Hong, Wei Li, Yuexin Ma, Hongsheng Li, Ruigang Yang, and Da Lin. Cylindrical and asymmetrical 3d convolution networks for lidar-based perception. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44:6807–6822, 2021. 1, 2, 6