# Memory-Augmented Re-Completion for 3D Semantic Scene Completion

**Yu-Wen Tseng[1], Sheng-Ping Yang[1], Jhih-Ciang Wu[1,3], I-Bin Liao[4], Yung-Hui Li[4], Hong-Han Shuai[2], Wen-Huang Cheng[1]***

[1] National Taiwan University, Taiwan
[2] National Yang Ming Chiao Tung University, Taiwan
[3] National Taiwan Normal University, Taiwan
[4] Hon Hai Research Institute, Taiwan

## Abstract

Semantic Scene Completion (SSC) aims to reconstruct a 3D voxel representation occupied by semantic classes based on ordinary inputs such as 2D RGB images, depth maps, or point clouds. Given the cost-effective and promising applications in autonomous driving, camera-based SSC has attracted considerable attention to developing various approaches. However, current methods mainly focus on precise 2D-to-3D projection while overlooking the challenge of completing *invisible regions*, leading to numerous false negatives and suboptimal SSC performance. To address this issue, we propose a novel architecture, *Memory-augmented Re-completion (MARE)*, designed to enhance completion capability. Our MARE model encapsulates regional relationships by incorporating a memory bank that stores vital region-tokens while two protocols concerning diversity and age are adopted to optimize the bank adversarially. Additionally, we introduce a Re-completion pipeline incorporated with an Information Spreading module to progressively complete the invisible regions while bridging the scale gap between region-level and voxel-level information. Extensive experiments conducted on the SSCBench-KITTI-360 and SemanticKITTI datasets validate the effectiveness of our approach.

**Code** — https://github.com/ywtseng0226/MARE

## Introduction

Learning to thoroughly comprehend the environment is crucial for autonomous vehicles, encompassing the recognition and reaction to various elements such as road conditions (Singh et al. 2022), pedestrian movement (Duan et al. 2022), and other vehicles (Guériau et al. 2015; Xu et al. 2023). With a satisfactory understanding of the scene, the perception systems can provide precise navigation and effectively avoid collisions, ensuring the safety of both passengers and others on the road. Consequently, the development of advanced methods that enhance AI's ability to accurately perceive and interact with the real world is vital for advancing the capabilities of autonomous driving systems.

To build reliable strategies for self-driving automobiles, one of the related tasks is 3D Semantic Scene Completion (SSC), which seeks to infer both the occupancy and semantic state of each voxel in the 3D space. The challenge of SSC lies in understanding the environment using relatively coarse inputs, such as monocular RGB images or sparse point clouds, while the fine-grained dense volumes are expected to describe the surroundings as the vehicle moves across streets. Although recent advancements (Cao and De Charette 2022; Li et al. 2023c; Jiang et al. 2024) have facilitated the evolution of SSC by introducing various techniques to transfer knowledge between 2D and 3D representations, a fundamental issue, termed *insufficient completion*, persists. As shown in Figure 1, the 3D scene volume generated by the SOTA method (Jiang et al. 2024) contains numerous false negatives (FNs) colored in light gray.

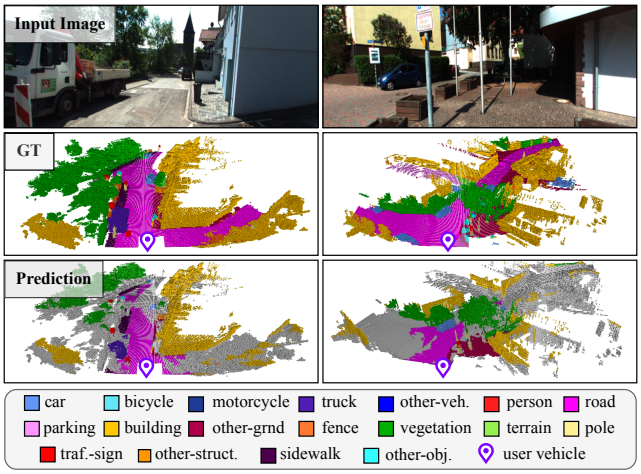


Figure 1: **Insufficient completion problem.** We illustrate the input image, ground truth, and prediction from Symphonies on SSCBench-KITTI-360. The voxels highlighted in light gray within the prediction represent false negatives.

These FNs are primarily found in invisible regions, which include two scenarios: occlusion, exemplified by the area occluded by the truck in the left example, and regions outside the field-of-view (FOV), such as the road sections on either side that are absent from the 2D image. This observation highlights the significant challenges in existing models: completing the invisible regions solely based on input information is insufficient, which results in a high empty rate of

*Corresponding author. wenhuang@csie.ntu.edu.tw

prediction. Moreover, current methods notably lack a mechanism to adequately fill these areas. Successfully completing these regions to preserve the integrity of object shapes could significantly enhance the safety and accuracy of decisions made in autonomous driving systems.

In this paper, we introduce a novel architecture, Memory-augmented Re-completion (MARE), designed to address the incomplete predictions in invisible regions. MARE aims to enhance the completion of these regions by leveraging regional information derived from visible areas. To achieve this, we incorporate a token-based SSC Transformer, employing cross-attention modules to facilitate the transfer of 2D information into 3D volume features, with token features serving as intermediate mediators. During this process, tokens encapsulating regional-level information, referred to as region-tokens, are generated. We establish a Regional Memory Bank using the generated region-tokens, which effectively aggregates representative features. This process considers the diversity and age of the storage units during memory updates, enabling the model to reference stored features during voxel prediction and mitigating the issue of insufficient information in invisible regions. Additionally, we propose the Re-completion pipeline to bridge prediction gaps across varying visibility levels by injecting relevant token features into invisible areas. This approach is complemented by the Information Spreading module, integrating the scale gap between regional and voxel-level representations. The combined strategy enhances the accuracy and completeness of 3D scene reconstruction by existing SSC models, particularly in areas occluded or outside the FOV.

To evaluate the effectiveness of our proposed MARE framework, we conducted extensive experiments on two challenging large-scale benchmarks, including SSCBench-KITTI-360 and SemanticKITTI. Our MARE outperforms SOTAs and sets new art by achieving mIoU scores of 18.84 and 15.39, respectively. Beyond the competitive mIoU scores, MARE demonstrates exceptional robustness in completion ability, as evidenced by the impressive recall scores of 57.09 and 65.38 on these datasets. These results highlight remarkable improvements over other SSC approaches.

Our contributions can be summarized as follows:

- Through a visual analysis of false negative voxels, we found that current SSC methods, relying solely on visible information from the input, are insufficient for fully completing the invisible regions. Additionally, these methods lack mechanisms for effectively filling in these regions.
- We propose MARE, a novel SSC paradigm designed to improve the completion of invisible regions. By leveraging a token-based SSC Transformer, MARE constructs a Regional Memory Bank as a feature repository guided by two principles. The Re-completion pipeline subsequently uses pivots from the memory bank to enrich incomplete regions, enabling a more accurate scene reconstruction.
- Our method outperforms existing approaches in challenging SSC benchmarks, including SSCBench-KITTI-360 and SemanticKITTI, with significant improvements, highlighting MARE's robustness in accurately predicting invisible voxels that were initially classified as empty.

# Related Work

## 3D Semantic Scene Completion

3D SSC, first proposed by SSCNet (Song et al. 2017), aims to predict the occupancy and semantic status of each voxel within the 3D scene using insufficient sensor information, such as LiDAR point clouds, depth maps, and 2D RGB images. Subsequently, some methods have used LiDAR point clouds as the primary input modality (Garbade et al. 2019; Rist et al. 2021; Xia et al. 2023), leveraging their innate 3D depth information. However, considering the cost-efficiency and compatibility of sensors, recent researchers have shifted their focus to camera-based SSC (Cao and De Charette 2022; Huang et al. 2023; Li et al. 2023c; Wang and Tong 2024; Li et al. 2023a; Jiang et al. 2024). MonoScene (Cao and De Charette 2022) introduces the first camara-based solution for SSC. TPVFormer (Huang et al. 2023) utilizes a tri-perspective view (TPV) to represent the 3D scene, serving as a trade-off between 3D dense volume and bird's eye view (BEV). VoxFormer (Li et al. 2023c) introduces a two-stage pipeline, allowing the model to first propose a class-agnostic scene and then further infer the semantic status. H2GFormer (Wang and Tong 2024) leverages horizontal-to-global attention and Internal-External Position Awareness Loss to achieve significant performance improvements in SSC. Symphonies (Jiang et al. 2024) uses region-level queries as an information bridge between 2D and 3D, thereby expanding the receptive field of the features. Although current camera-based SSC methods improve the accuracy of transferring information from 2D to 3D spaces, the information scarcity for invisible regions remains unresolved. To overcome this limitation, we introduce MARE, a paradigm that leverages memory-augmented information to effectively bridge these gaps.

## Memory-augmented Representation Learning

Memory-augmented models in learning-based investigations are designed to enhance learning by incorporating external memory modules. These modules allow models to store, retrieve, and update information dynamically, enabling them to perform tasks that require reasoning over long sequences or retaining important information over learning periods. This technique has a broad spectrum of applications, spanning Natural Language Processing (Dai et al. 2019; Borgeaud et al. 2022), Computer Vision (Oh et al. 2019; Gao and Wang 2023; He et al. 2024), Reinforcement Learning (Jeddi, Dehghani, and Shafieezadeh 2023; Morad et al. 2024), and Large Language Model (Wang et al. 2024). A common challenge these methods address is the lack of sufficient information, which is often alleviated by utilizing additional resources to create a memory bank for improved contextual understanding. This challenge is analogous to the problem encountered in the camera-based SSC, where 2D images only provide information for visible regions, leaving occluded or out-of-view areas as invisible ones. To overcome this limitation, we introduce a Regional Memory Bank that stores essential token features and learns a general representation of urban environments from visible regions, thereby effectively filling in the invisible areas.

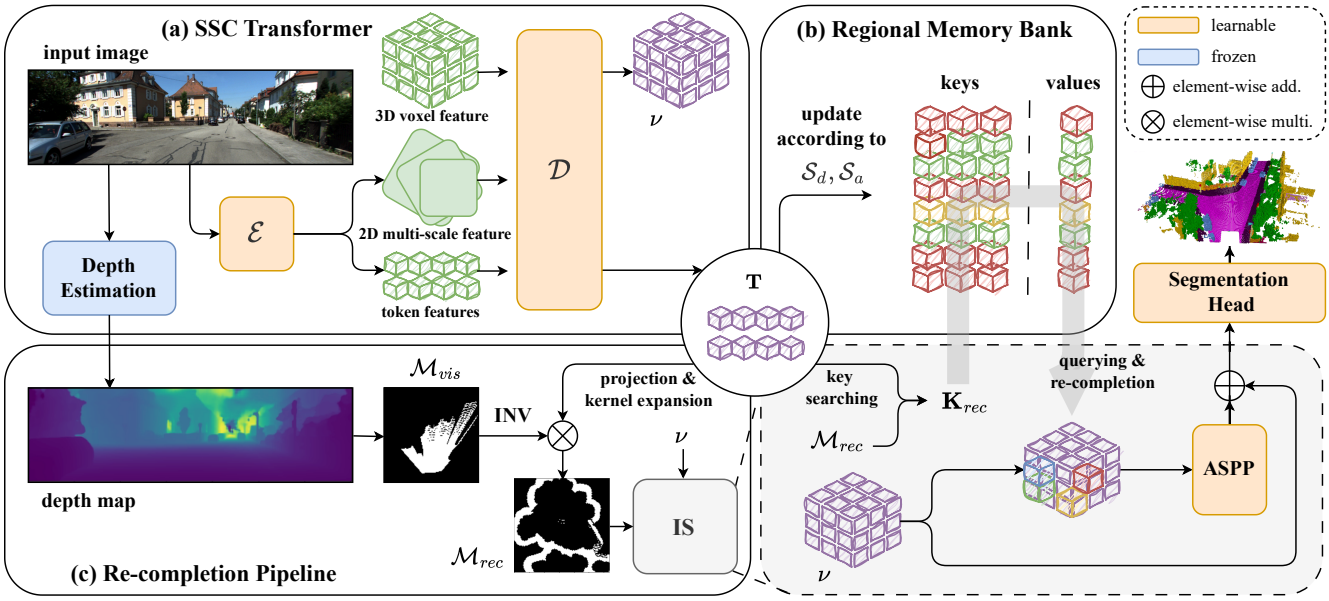


Figure 2: **Overview of MARE.** The MARE paradigm is composed of three key components: (a) the SSC Transformer, which performs initial information extraction and understanding of visible regions; (b) the Regional Memory Bank, which stores the region-tokens generated by the SSC Transformer and updates key-value pairs to serve as a comprehensive repository representing the urban view; and (c) the Re-completion pipeline, which identifies and progressively re-completes invisible areas using region-tokens, with the Information Spreading module bridging the scale gap between region-level and voxel-level feature.

# Methodology

## Problem Formulation

Utilizing a 2D image $I$ captured by a camera mounted on a vehicle, the primary objective of the SSC task is to predict a 3D volume $\hat{y} \in \mathbb{R}^{H \times W \times D}$, which segments the observed scene. Each voxel within this predicted volume is associated with a class $c_i \in \mathcal{C} = \{c_0, c_1, \cdots, c_N\}$, where $c_0$ specifically denotes empty, and the other $N$ classes correspond to various semantic classes. Unlike prior methods (Li et al. 2023c; Wang and Tong 2024), which incorporate data from previous frames, our method exclusively leverages the current frame to train the SSC model $\mathcal{F}$. By optimizing the model parameters $\theta$, we aim to generate predictions $\hat{y} = \mathcal{F}(I; \theta)$ that closely match the actual ground truth $y$.

## Memory-augmented Re-completion

Conventional SSC methods mainly focus on completing visible regions while neglecting the invisible areas within the scene. To address this limitation, we propose the Memory-augmented Re-completion (MARE) framework, which enhances the completion of these neglected regions by gradually injecting tokens to voxel space. As depicted in Figure 2, MARE comprises three key components: the SSC Transformer, the Regional Memory Bank, and the Re-completion pipeline, which we refer to by their abbreviations, $\mathcal{T}$, $\mathcal{B}$, and $\mathcal{R}$, respectively. The SSC Transformer processes the input to generate region-tokens and a 3D representation that reflects the understanding of the visible surroundings. This initial module $\mathcal{T}$ ensures that the model effectively captures both the semantic and spatial information from the visible regions. The Regional Memory Bank then stores region-tokens and updates based on the designed criteria, *i.e.*, diversity and age of the token, to maintain a robust and representative knowledge base of urban 3D views in $\mathcal{B}$. Eventually, the Re-completion pipeline fills in the invisible regions by accessing the stored tokens, following an Information Spreading module to alleviate the scale gap between region-tokens and voxel features. We elaborate on the details of each component in the following sections.

**SSC Transformer.** In contrast to previous methods (Cao and De Charette 2022; Li et al. 2023c) that focus solely on aligning 2D pixels with 3D voxels during the transfer of information from 2D to 3D, we build upon the token-based approach (Jiang et al. 2024) that utilizes region-tokens as intermediaries. The token-based architecture provides more flexible and abstract scene representations, enabling more effective knowledge transformation between the 2D and 3D domains. As illustrated in Figure 2 (a), the 2D image $I$ is processed by the encoder $\mathcal{E}$ (Li et al. 2023b), resulting in 2D multi-scale features and token features. For simplicity, we omit the backbone (He et al. 2016), which serves as a feature extractor and precedes the encoder. These token-based representations and the 3D voxel feature are then decoded by $\mathcal{D}$, which employs cross-attention mechanisms to generate 3D features. The entire process can be formulated as:

$$\mathbf{T}, \nu = \mathcal{D}(\mathcal{E}(I; \theta)), \tag{1}$$

where $\mathbf{T}$ and $\nu$ represent the decoded token features and voxel features, respectively, while $\theta$ denotes the model parameters that need to be updated during the process. More

Algorithm 1: Memory Updating

**Input:** new region-token $t_i \in \mathbf{T}$ and current $t_k \in \mathcal{B}$
**Output:** updated region-tokens in memory bank
$\Omega = \mathcal{B} \cup \mathbf{T}$
Set initial age $\mathcal{S}_a(t_i) = 0$ for all new tokens $t_i$
**Start memory Updating:**
**for** $i = 1$ to $|\Omega|$ **do**
    $\mathcal{S}_d(t_i) \leftarrow -\sum_{j=1}^{|\Omega|} \frac{t_i \cdot t_j}{\|t_i\| \, \|t_j\|}, \quad i \neq j$
**end for**
$\mathcal{S} \leftarrow \mathcal{S}_d - \mathcal{S}_a$
Select highest $|\mathcal{B}|$ score and update the bank
Increase the age of all tokens in $\mathcal{B}$

precisely, leveraging $\mathcal{T}$ allows the receptive field of token features to extend beyond the voxel level, facilitating a more comprehensive integration of information across different scales, which leads to holistic 3D scene reconstruction.

**Regional Memory Bank.** The principle goal of $\mathcal{B}$ is to gather representations containing the characteristics that preserve the relationships between regions within the scene. As shown in Figure 2 (b), we choose to arrange a bank that collects the region-tokens derived from $\mathcal{D}$ for two advantageous reasons: computational efficiency and manageability. Firstly, storing region-tokens reduces storage costs significantly compared to alternatives like vanilla multi-scale features (Liu and Mukhopadhyay 2018) or learnable atoms (Liu, Wang, and Cai 2024). Additionally, using region-tokens simplifies the modeling of complex relationships, as these tokens act as intermediaries between 2D regions and voxel-based representations.

Given a 2D RGB image $I$, the transformer $\mathcal{T}$ yields tokens $\mathbf{T}$ described region-level information for visible regions, associating with the corresponding reference points $\mathbf{P}$ mapped to the 3D space. For each region-token $t_i \in \mathbf{T}$, we search $k$ nearest neighbors to construct the key set, composing the key-value pair $\{\mathbf{K}_i, t_i\}$. Formally, the formulation for obtaining element $k_n \in \mathbf{K}_i$ is defined as

$$k_n = \arg\min_{t_j}[d(t_i, t_j)] \quad \forall i \neq j, \tag{2}$$

where $1 \leq j \leq |\mathbf{T}|$ represents the index for retrieving and $d(\cdot)$ denotes the distance function for two region-tokens.

The key-value pairs derived from (2) can be utilized in the upcoming $\mathcal{R}$, restricting the model focus on relevant 3D points corresponding to regions in the 2D image, particularly in areas that are not directly visible. To ensure that the attention mechanism accurately identifies and prioritizes these relevant 3D points, the distance between region-tokens is measured using the Euclidean distance between their corresponding reference points in the 3D space, which is formulated as

$$d(t_i, t_j) = \|p_i - p_j\|_2, \tag{3}$$

where $p_i$ and $p_j \in \mathbf{P}$ are reference points that are homologous to region-tokens $t_i$ and $t_j$, respectively.

During the training phase, key-value pairs are progressively embedded into $\mathcal{B}$ along with each mini-batch. As the memory capacity exceeds the predefined size $|\mathcal{B}|$, *i.e.*, the cardinality of $\mathcal{B}$, we update the bank based on our planned protocols, considering both the *diversity* and *age* of tokens to determine whether they should be retained or removed adversarially. The diversity criterion confirms that the contents in $\mathcal{B}$ remain varied while effectively capturing regional information across scenes. To achieve this, we estimate each candidate based on cosine similarity, which is defined as

$$\mathcal{S}_d(t_i) = \sum_{i \neq j}^{|\Omega|} \frac{t_i \cdot t_j}{\|t_i\| \, \|t_j\|}, \tag{4}$$

where $\Omega = \mathcal{B} \cup \mathbf{T}$, indicating that this property is considered within both $\mathcal{B}$ and $\mathbf{T}$. More precisely, the similarity score in (4) is calculated using indices from both the current bank and the dominant region-tokens.

The other protocol concerning liquidity aims to prevent the accumulation of outdated tokens in $\mathcal{B}$. Since the scene obtained by the vehicle continually changes, older tokens in $\mathcal{B}$ may lose their generalization capability and could even impair the model's performance. Inspired by the concept of (Yuan, Xie, and Li 2023), we encourage the inclusion of new tokens to prevent the potential domain gap between outdated and current content. In practice, we denote $\mathcal{S}_a(t_i)$ to symbolize the age of the token, which is initialized at 0 and increases with each mining. The overall score for determination can be formulated as

$$\mathcal{S}(t_i) = \mathcal{S}_d(t_i) - \mathcal{S}_a(t_i). \tag{5}$$

Based on the total scores computed in (5), the key-value pairs in $\mathcal{B}$ are retained through an adversarial selection, confirming that the stored representations remain relevant and effective. The detailed procedure for memory updating is offered in Algorithm 1.

**Re-completion Pipeline.** Upon obtaining representative region-tokens stored in the memory bank, as depicted in Figure 2 (c), we utilize them to enhance the predicted voxel features, particularly for the invisible regions. To determine the visibility of these regions, we follow previous methods (Li et al. 2023c; Jiang et al. 2024; Wang and Tong 2024) that estimate the depth map using an off-the-shelf model (Shamsafar et al. 2022). The generated depth map $Z$ is then projected onto 3D coordinates, which can be formulated as

$$x = \frac{(u - \omega_u)}{f_u} z, y = \frac{(v - \omega_u)}{f_v} z, z = Z(u, v), \tag{6}$$

where $\omega_u, \omega_v, f_u, f_v$ are camera positional information and configurations, representing the the center and focal length for horizontal and vertical, respectively.

As noted in (Li et al. 2023c), the 3D point cloud generated often exhibits a weak representation due to the highly inconsistent depth at the horizon. This issue is particularly pronounced in long-range areas, where only a limited number of pixels are available to determine the depth of a vast region. Moreover, we observe that the projected points are predominantly concentrated on the lowest plane within the 3D space. Based on these observations, we eliminate the height dimension of all visible points, projecting them onto the BEV.

| Method | IoU | Prec. | Rec. | mIoU | car | bicycle | motorcycle | truck | other-veh. | person | road | parking | sidewalk | other-grnd. | building | fence | vegetation | terrain | pole | traf.-sign | other-struct. | other-obj. |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| MonoScene | 37.87 | 56.73 | 53.26 | 12.31 | 19.34 | 0.43 | 0.58 | 8.02 | 2.03 | 0.86 | 48.35 | 11.38 | 28.13 | 3.32 | 32.89 | 3.53 | 26.15 | 16.75 | 6.92 | 5.67 | 4.20 | 3.09 |
| TPVFormer | 40.22 | 59.32 | 55.54 | 13.64 | 21.56 | 1.09 | 1.37 | 8.06 | 2.57 | 2.38 | 52.99 | 11.99 | 31.07 | 3.78 | 34.83 | 4.80 | 30.08 | 17.52 | 7.46 | 5.86 | 5.48 | 2.70 |
| VoxFormer | 38.76 | 58.52 | 53.44 | 11.91 | 17.84 | 1.16 | 0.89 | 4.56 | 2.06 | 1.63 | 47.01 | 9.67 | 27.21 | 2.89 | 31.18 | 4.97 | 28.99 | 14.69 | 6.51 | 6.92 | 3.79 | 2.43 |
| OccFormer | 40.27 | 59.70 | 55.31 | 13.81 | 22.58 | 0.66 | 0.26 | 9.89 | 3.82 | 2.77 | 54.30 | 13.44 | 31.53 | 3.55 | 36.42 | 4.80 | 31.00 | 19.51 | 7.77 | 8.51 | 6.95 | 4.60 |
| Symphonies | 44.12 | 69.24 | 54.88 | 18.58 | **30.02** | 1.85 | 5.90 | **25.07** | **12.06** | **8.20** | 54.94 | 13.83 | 32.76 | **6.93** | 35.11 | **8.58** | **38.33** | 11.52 | **14.01** | 9.57 | **14.44** | **11.28** |
| MARE (Ours) | **46.10** | **70.53** | **57.09** | **18.84** | 28.37 | **2.86** | **7.22** | 15.85 | 7.27 | 6.74 | **60.14** | **15.83** | **37.98** | 4.56 | **41.66** | 7.75 | 37.09 | **21.47** | 13.90 | **15.42** | 9.56 | 5.53 |

Table 1: **Quantitative results on SSCBench-KITTI-360 test.** The best results are in **bold**. MARE surpasses all previous methods across multiple metrics and shows substantial improvements in specific categories, *i.e.*, road, sidewalk, and building.

This projection results in a visible binary mask, denoted as $\mathcal{M}_{vis}$. Conversely, the corresponding invisible mask $\mathcal{M}_{inv}$ can be obtained by reversing $\mathcal{M}_{vis}$. Since our re-completion method is designed explicitly for invisible regions and aims to distill information from visible areas, it is intuitive to prioritize filling areas adjacent to the region-tokens. To accomplish this, we apply a kernel expansion technique to identify the neighboring areas around $\mathbf{P}$, which are then combined with $\mathcal{M}_{inv}$ to acquire the re-complete mask $\mathcal{M}_{rec}$.

In the Information Spreading module, each invisible position in $\mathcal{M}_{rec}$, we carry out Re-completion by reuse (2), resulting in $k$ neighbors serving as the keys, denoted as $\mathbf{K}_{rec}$. These keys subsequently go through the $\mathcal{B}$ to query the tokens injected into the 3D voxel features. We follow the strategy (Van Den Oord, Vinyals et al. 2017), where the memory bank is treated as a codebook to implement the querying and injecting process. After computing the similarity between the keys, the corresponding value is selected using a one-hot vector, allowing the entire process to be seamlessly integrated into our end-to-end training framework. While the Re-completion process injects tokens encapsulating region-level information, a scale discrepancy arises when equipping the voxel features with these tokens. To settle this problem, we adopt the Atrous Spatial Pyramid Pooling (ASPP) (Chen et al. 2017), which contains multi-scale dilated convolutions while a skip connection stabilizes gradient updates. Finally, the voxel features are passed through a segmentation head to obtain the semantic prediction.

## Loss Function

The proposed MARE uses a similar optimization design in MonoScene (Cao and De Charette 2022). The overall loss can be formulated as:

$$\mathcal{L} = \mathcal{L}_{ce} + \mathcal{L}_{scal}^{geo} + \mathcal{L}_{scal}^{sem}, \tag{7}$$

where the cross-entropy loss $\mathcal{L}_{ce}$ optimize the semantic classification results for each voxels. The Scene-Class Affinity Loss $\mathcal{L}_{scal}$ optimize the overall score, *i.e.*, precision, call, and specificity. It can be further divided into class-agnostic and semantic versions as $\mathcal{L}_{scal}^{geo}$ and $\mathcal{L}_{scal}^{sem}$.

# Experiments

## Experimental Setup

**Datasets and Evaluation Metrics.** The evaluation is performed on SSCBench-KITTI-360 (Li et al. 2024) and SemanticKITTI (Behley et al. 2019) datasets, which provide densely annotated urban driving scene sequences from the KITTI Odometry Benchmark (Geiger, Lenz, and Urtasun 2012). A target 3D scene of size 51.2m×51.2m×6.4m is divided into 256× 256×32 voxel grids, with each voxel grid having a size of 0.2m×0.2m×0.2m. SSCBench-KITTI-360 provides 9 video sequences, with 7 for training, 1 for validation, and 1 for testing. SemanticKITTI provides 20 video sequences, with 9 for training, 1 for validation, and 11 for testing. Following previous works (Cao and De Charette 2022; Li et al. 2023c; Jiang et al. 2024), we use mean IoU as the evaluation metric for semantic scene completion and IoU as the evaluation metric for class-agnostic scene completion.

**Implementation Details.** Based on the region-token-based SSC method with our proposed MARE paradigm, the model is trained in an end-to-end manner with two NVIDIA V100 GPUs for 30 epochs, with a batch size of two images. In line with Symphonies (Jiang et al. 2024), We employ the AdamW (Loshchilov and Hutter 2017) as the optimizer, ResNet-50 (He et al. 2016) as the backbone, and the pre-trained weights of MaskDINO (Li et al. 2023b) as the token-based Encoder. In the Regional Memory Bank, we set the size $|\mathcal{B}|$ as 1024 and the number of neighbor tokens $k$ as 3. In the Re-completion pipeline, we re-complete the scene for two iterations. These configurations in our proposed MARE paradigm are based on experimental observations.

**Baseline Methods.** Within the scope of the SSC task, there has been growing interest in camera-based SSC approaches due to their cost-effectiveness and portability. Accordingly, this section focuses on comparisons of camera-based methods. Specifically, we compare our approach with the region-token based method Symphonies (Jiang et al. 2024), the pioneering MonoScene (Cao and De Charette 2022), and other state-of-the-art methods such as TPVFormer (Huang et al. 2023), VoxFormer (Li et al. 2023c), OccFormer (Zhang, Zhu, and Du 2023), NDC-Scene (Yao et al. 2023), and H2GFormer (Wang and Tong 2024).

| Method | IoU | Prec. | Rec. | mIoU | road | sidewalk | parking | other-grnd. | building | car | truck | bicycle | motorcycle | other-veh. | vegetation | trunk | terrain | person | bicyclist | motorcyclist | fence | pole | traf.-sign |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| MonoScene* | 36.86 | 52.19 | 55.50 | 11.08 | 56.52 | 26.72 | 14.27 | 0.46 | 14.09 | 23.26 | 6.98 | 0.61 | 0.45 | 1.48 | 17.89 | 2.81 | 29.64 | 1.86 | 1.20 | 0.00 | 5.84 | 4.14 | 2.25 |
| TPVFormer | 35.61 | - | - | 11.36 | 56.50 | 25.87 | 20.60 | 0.85 | 13.88 | 23.81 | 8.08 | 0.36 | 0.05 | 4.35 | 16.92 | 2.26 | 30.38 | 0.51 | 0.89 | 0.00 | 5.94 | 3.14 | 1.52 |
| VoxFormer-S | 44.02 | **62.32** | 59.99 | 12.35 | 54.76 | 26.35 | 15.50 | 0.70 | 17.65 | 25.79 | 5.63 | 0.59 | 0.51 | 3.77 | 24.39 | 5.08 | 29.96 | 1.78 | **3.32** | 0.00 | 7.64 | 7.11 | 4.18 |
| OccFormer | 36.50 | - | - | 13.46 | 58.85 | 26.88 | 19.61 | 0.31 | 14.40 | 25.09 | **25.53** | 0.81 | 1.19 | 8.52 | 19.63 | 3.93 | 32.62 | 2.78 | 2.82 | 0.00 | 5.61 | 4.26 | 2.86 |
| NDC-Scene | 37.24 | - | - | 12.70 | **59.20** | 28.24 | **21.42** | **1.67** | 14.94 | 26.26 | 14.75 | 1.67 | 2.37 | 7.73 | 19.09 | 3.51 | 31.04 | **3.60** | 2.74 | 0.00 | 6.65 | 4.53 | 2.73 |
| H2GFormer-S | **44.57** | 62.17 | 61.16 | 13.73 | 56.08 | **29.12** | 17.83 | 0.45 | 19.74 | 27.60 | 10.00 | 0.50 | 0.47 | 7.39 | 26.25 | 7.80 | 34.42 | 1.54 | 2.88 | 0.00 | 7.24 | 7.88 | 4.68 |
| Symphonies | 41.92 | - | - | 14.89 | 56.37 | 27.58 | 15.28 | 0.95 | 21.64 | 28.68 | 20.44 | 2.54 | 2.82 | **13.89** | 25.72 | 6.60 | 30.87 | 3.52 | 2.24 | 0.00 | 8.40 | **9.57** | 5.76 |
| MARE (Ours) | 43.20 | 56.01 | **65.38** | **15.39** | 56.26 | 25.91 | 19.46 | 0.65 | **23.36** | **29.06** | 17.82 | **2.83** | **3.63** | 13.54 | **26.80** | **8.31** | **33.80** | 3.20 | 2.10 | 0.00 | **10.38** | 9.39 | **5.95** |

Table 2: **Quantitative results on SemanticKITTI `val`.** * represents the reproduced results in (Huang et al. 2023). The best results are in **bold**. MARE demonstrates superior performance compared to all other methods, particularly in the recall score.

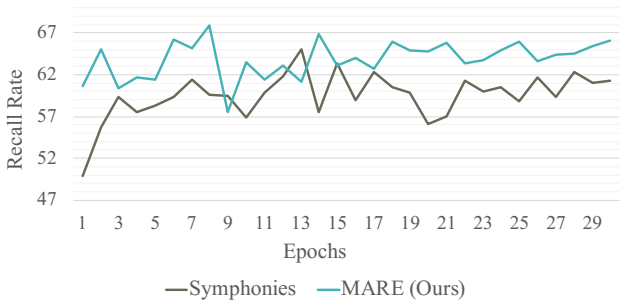


Figure 3: **Recall tendency on SemanticKITTI.** Throughout the training process, MARE approximately exhibits better completion performance compared to Symphonies.

## Quantitative Results

**Results on SSCBench-KITTI-360 test.** As shown in Table 1, indicate that MARE achieves superior performance, surpassing all competing methods. Specifically, MARE demonstrates notable performance gains over Symphonies, with improvements of 2.02 in IoU, 1.31 in precision, 2.21 in recall, and 0.26 in mIoU. Importantly, the significant enhancements in class-agnostic metrics such as IoU and recall highlight MARE's efficacy in accurately re-completing voxels initially misclassified as empty. Furthermore, significant enhancements in IoU scores were observed for classes frequently found around visible regions. Specifically, the IoU score for road improved from 54.94 to 60.14, for sidewalk from 32.76 to 37.98, and for building from 35.11 to 41.66. For categories that exhibited a decline in performance, such as truck and other-vehicle, we hypothesize that this degradation may be linked to our injection mechanism, which could slightly disrupt the fine feature representation needed at the edges of visible regions. In contrast, categories characterized by more coarse features, such as road and building, have demonstrated significant improvements. This identified limitation presents an opportunity for future work to enhance the performance of our designed paradigm.

**Results on SemanticKITTI val.** As shown in Table 2, our proposed MARE method surpasses previous camera-based SSC approaches, particularly in terms of recall score, where a significant improvement is observed. This substantial increase underscores the efficacy of our approach in accurately re-completing voxels that were previously misclassified as empty, further validating the robustness of our method. Nevertheless, the decline in the score of precision deserves attention. Upon examining the dataset, we identified that, despite the widespread use of SemanticKITTI in SSC tasks, inconsistencies in labeling, particularly for categories such as moving objects, are prevalent. These inconsistencies likely disrupt the learning of token representations, which we hypothesize as a contributing factor to the comparatively smaller improvement of MARE on SemanticKITTI relative to SSCBench-KITTI-360.

**Recall Tendency.** To assess the efficacy of our proposed MARE method in re-completing voxels that might otherwise be misclassified as empty, recall is a pivotal performance metric. The quantitative results across both datasets have demonstrated substantial improvements in recall scores. In this analysis, we further examine the progression of recall scores throughout the training process. As illustrated in Figure 3, MARE not only achieves higher recall scores from the initial epoch but also consistently outperforms Symphonies throughout the training process. This trend indicates that our proposed Re-completion pipeline effectively drives the model to identify and predict more occupied voxels during learning, thereby mitigating the tendency to classify invisible regions as empty.

## Qualitative Results

The qualitative results comparing MARE and Symphonies on the SSCBench-KITTI-360 dataset are depicted in Figure 4. We present three examples to illustrate the effectiveness of MARE in re-completing invisible regions. In the second and last rows, Symphonies struggles to detect certain classes outside of FOV areas, *i.e.*, road, parking, and terrain. MARE, on the other hand, successfully re-completes these regions, providing accurate predictions for the correct classes. In the final row, MARE not only completes the side

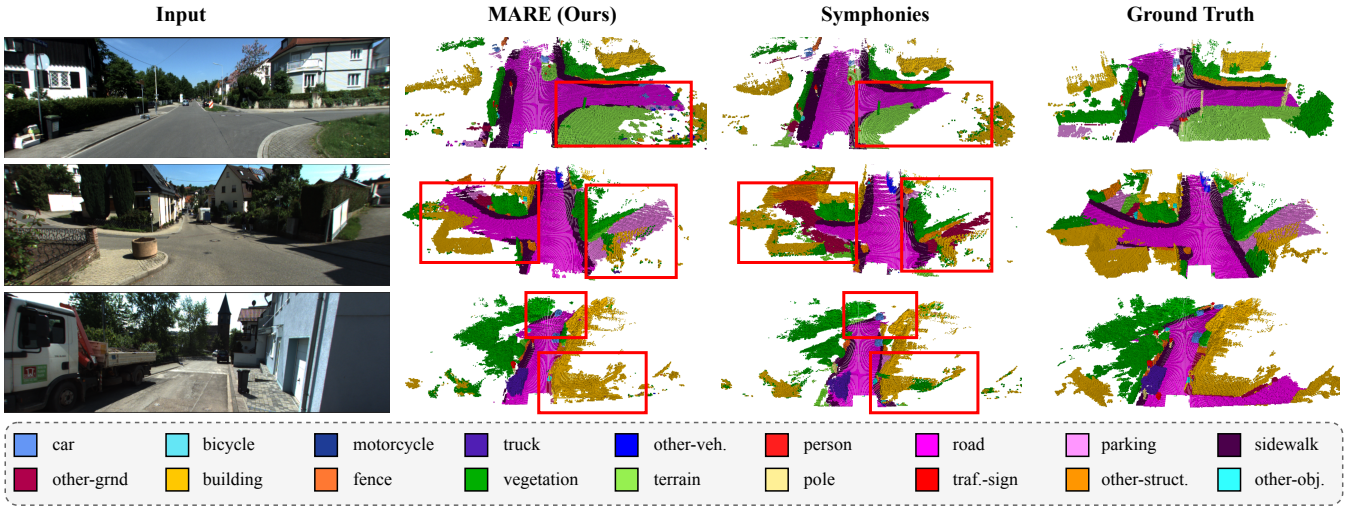


Figure 4: **Qualitative results on SSCBench-KITTI-360.** The regions marked with red boxes in the figure illustrate that MARE is capable of accurately completing invisible areas, including those outside of FOV and regions that are occluded.

| $\mathcal{T}$ | $\mathcal{R}$ | $\mathcal{B}$ | IoU | Prec. | Rec. | mIoU |
|---|---|---|---|---|---|---|
| ✓ | | | 43.80 | 68.10 | 55.10 | 18.25 |
| ✓ | ✓ | | 45.99 | 70.43 | 56.99 | 18.60 |
| ✓ | ✓ | ✓ | **46.10** | **70.53** | **57.09** | **18.84** |

Table 3: **Ablation study on components in MARE.** $\mathcal{T}$, $\mathcal{R}$, and $\mathcal{B}$ denote the SSC Transformer, Re-completion pipeline, and Regional Memory Bank respectively.

| Near | Medium | Far | IoU | Prec. | Rec. | mIoU |
|---|---|---|---|---|---|---|
| ✓ | | | 42.95 | 59.20 | 61.00 | 15.06 |
| ✓ | ✓ | | **43.20** | 56.01 | **65.38** | **15.39** |
| ✓ | ✓ | ✓ | 43.03 | **60.05** | 60.30 | 14.01 |

Table 4: **Ablation study on iteration for Re-complete.** Near, medium, and far respectively represent the distance levels after the first, second, and third re-completion.

building but also accurately predicts the road occluded by a vehicle in the distance. The examples above demonstrate that MARE can effectively complete invisible regions, allowing for a more comprehensive and accurate understanding of the surrounding scene. In the context of autonomous driving applications, this provides a more complete perception result for subsequent decision-making processes.

## Ablation Study

We conduct an in-depth analysis of the key components of MARE on the SSCBench-KITTI-360 dataset, while the iteration process of the Re-completion pipeline is specifically examined on the SemanticKITTI dataset. This analysis provides insights into the performance enhancements attributed to each component and process in MARE.

**Effectiveness of each module in MARE.** As shown in Table 3, the first row denotes the performance of the $\mathcal{T}$ alone. In the second row, where $\mathcal{R}$ is added without memory-augmented information, the nearest region-token from the visible area is used as the injected token. Initial performance enhancements are observed with the re-completion process, with the most significant improvements occurring in the final row after integrating the memory bank $\mathcal{B}$.

**Iterations for Re-completion pipeline.** In the Re-completion pipeline, the filling process can iterate multiple times, adopting a near-to-far re-completion strategy. After each iteration, the injected tokens are integrated into the visible information, forming the basis for the next iteration. Results in Table 4 show that the second iteration yields the best performance, as the third iteration reaches the scene's less critical edges, often unlabeled.

## Conclusion

We have presented Memory-augmented Re-completion (MARE), an architecture designed to address the challenges of camera-based Semantic Scene Completion (SSC), particularly the accurate completion of invisible regions. By leveraging a Regional Memory Bank and Re-completion pipeline, MARE effectively captures and reuses vital region-level information, bridging the perception gap between visible and invisible areas within urban view in autonomous driving scenarios. Our approach enhances the understanding of the 3D environment and significantly improves performance, as demonstrated by extensive experiments on two datasets. The experimental results highlight MARE's ability to achieve superior mIoU and recall scores compared to existing methods, establishing it as a robust solution for SSC tasks in autonomous driving scenarios. Future work can extend this framework to other domains and refine the model to carry out even more complex environments effectively.

## Acknowledgments

This work is partially supported by the National Science and Technology Council, Taiwan, under Grants: NSTC-112-2628-E-002-033-MY4, NSTC-112-2634-F-002-002-MBK, NSTC-112-2221-E-A49-059-MY3 and NSTC-112-2221-E-A49-094-MY3, and was financially supported in part by the Center of Data Intelligence: Technologies, Applications, and Systems, National Taiwan University (Grants: 114L900901/114L900902/114L900903), from the Featured Areas Research Center Program within the framework of the Higher Education Sprout Project by the Ministry of Education, Taiwan.

## References

Behley, J.; Garbade, M.; Milioto, A.; Quenzel, J.; Behnke, S.; Stachniss, C.; and Gall, J. 2019. Semantickitti: A dataset for semantic scene understanding of lidar sequences. In *ICCV*.

Borgeaud, S.; Mensch, A.; Hoffmann, J.; Cai, T.; Rutherford, E.; Millican, K.; Van Den Driessche, G. B.; Lespiau, J.-B.; Damoc, B.; Clark, A.; et al. 2022. Improving language models by retrieving from trillions of tokens. In *ICML*.

Cao, A.-Q.; and De Charette, R. 2022. Monoscene: Monocular 3d semantic scene completion. In *CVPR*.

Chen, L.-C.; Papandreou, G.; Kokkinos, I.; Murphy, K.; and Yuille, A. L. 2017. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE TPAMI*.

Dai, Z.; Yang, Z.; Yang, Y.; Carbonell, J.; Le, Q. V.; and Salakhutdinov, R. 2019. Transformer-xl: Attentive language models beyond a fixed-length context. *arXiv preprint arXiv:1901.02860*.

Duan, J.; Wang, L.; Long, C.; Zhou, S.; Zheng, F.; Shi, L.; and Hua, G. 2022. Complementary attention gated network for pedestrian trajectory prediction. In *AAAI*.

Gao, R.; and Wang, L. 2023. MeMOTR: Long-term memory-augmented transformer for multi-object tracking. In *ICCV*.

Garbade, M.; Chen, Y.-T.; Sawatzky, J.; and Gall, J. 2019. Two stream 3d semantic scene completion. In *CVPR Workshop*.

Geiger, A.; Lenz, P.; and Urtasun, R. 2012. Are we ready for autonomous driving? the kitti vision benchmark suite. In *CVPR*.

Guériau, M.; Billot, R.; El Faouzi, N.-E.; Hassas, S.; and Armetta, F. 2015. Multi-agent dynamic coupling for cooperative vehicles modeling. In *AAAI*.

He, B.; Li, H.; Jang, Y. K.; Jia, M.; Cao, X.; Shah, A.; Shrivastava, A.; and Lim, S.-N. 2024. Ma-lmm: Memory-augmented large multimodal model for long-term video understanding. In *CVPR*.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *CVPR*.

Huang, Y.; Zheng, W.; Zhang, Y.; Zhou, J.; and Lu, J. 2023. Tri-perspective view for vision-based 3d semantic occupancy prediction. In *CVPR*.

Jeddi, A. B.; Dehghani, N. L.; and Shafieezadeh, A. 2023. Memory-augmented Lyapunov-based safe reinforcement learning: end-to-end safety under uncertainty. *IEEE TAI*.

Jiang, H.; Cheng, T.; Gao, N.; Zhang, H.; Lin, T.; Liu, W.; and Wang, X. 2024. Symphonize 3d semantic scene completion with contextual instance queries. In *CVPR*.

Li, B.; Sun, Y.; Liang, Z.; Du, D.; Zhang, Z.; Wang, X.; Wang, Y.; Jin, X.; and Zeng, W. 2023a. Bridging stereo geometry and BEV representation with reliable mutual interaction for semantic scene completion. *arXiv preprint arXiv:2303.13959*.

Li, F.; Zhang, H.; Xu, H.; Liu, S.; Zhang, L.; Ni, L. M.; and Shum, H.-Y. 2023b. Mask dino: Towards a unified transformer-based framework for object detection and segmentation. In *CVPR*.

Li, Y.; Li, S.; Liu, X.; Gong, M.; Li, K.; Chen, N.; Wang, Z.; Li, Z.; Jiang, T.; Yu, F.; et al. 2024. Sscbench: A large-scale 3d semantic scene completion benchmark for autonomous driving. In *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 13333–13340. IEEE.

Li, Y.; Yu, Z.; Choy, C.; Xiao, C.; Alvarez, J. M.; Fidler, S.; Feng, C.; and Anandkumar, A. 2023c. Voxformer: Sparse voxel transformer for camera-based 3d semantic scene completion. In *CVPR*.

Liu, L.; Wang, W. Y.; and Cai, P. 2024. Point Cloud Classification via Learnable Memory Bank. In *MMM*.

Liu, Q.; and Mukhopadhyay, S. 2018. Unsupervised learning using pretrained CNN and associative memory bank. In *IJCNN*.

Loshchilov, I.; and Hutter, F. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Morad, S.; Kortvelesy, R.; Liwicki, S.; and Prorok, A. 2024. Reinforcement learning with fast and forgetful memory. *NeurIPS*.

Oh, S. W.; Lee, J.-Y.; Xu, N.; and Kim, S. J. 2019. Video object segmentation using space-time memory networks. In *ICCV*.

Rist, C. B.; Emmerichs, D.; Enzweiler, M.; and Gavrila, D. M. 2021. Semantic scene completion using local deep implicit functions on lidar data. *IEEE TPAMI*.

Shamsafar, F.; Woerz, S.; Rahim, R.; and Zell, A. 2022. Mobilestereonet: Towards lightweight deep networks for stereo matching. In *WACV*.

Singh, G.; Akrigg, S.; Di Maio, M.; Fontana, V.; Alitappeh, R. J.; Khan, S.; Saha, S.; Jeddisaravi, K.; Yousefi, F.; Culley, J.; et al. 2022. Road: The road event awareness dataset for autonomous driving. *IEEE TPAMI*.

Song, S.; Yu, F.; Zeng, A.; Chang, A. X.; Savva, M.; and Funkhouser, T. 2017. Semantic scene completion from a single depth image. In *CVPR*.

Van Den Oord, A.; Vinyals, O.; et al. 2017. Neural discrete representation learning. *NeurIPS*.

Wang, W.; Dong, L.; Cheng, H.; Liu, X.; Yan, X.; Gao, J.; and Wei, F. 2024. Augmenting language models with long-term memory. *NeurIPS*.

Wang, Y.; and Tong, C. 2024. H2gformer: Horizontal-to-global voxel transformer for 3d semantic scene completion. In *AAAI*.

Xia, Z.; Liu, Y.; Li, X.; Zhu, X.; Ma, Y.; Li, Y.; Hou, Y.; and Qiao, Y. 2023. Scpnet: Semantic scene completion on point cloud. In *CVPR*.

Xu, R.; Xia, X.; Li, J.; Li, H.; Zhang, S.; Tu, Z.; Meng, Z.; Xiang, H.; Dong, X.; Song, R.; et al. 2023. V2v4real: A real-world large-scale dataset for vehicle-to-vehicle cooperative perception. In *CVPR*.

Yao, J.; Li, C.; Sun, K.; Cai, Y.; Li, H.; Ouyang, W.; and Li, H. 2023. Ndc-scene: Boost monocular 3d semantic scene completion in normalized device coordinates space. In *ICCV*.

Yuan, L.; Xie, B.; and Li, S. 2023. Robust test-time adaptation in dynamic scenarios. In *CVPR*.

Zhang, Y.; Zhu, Z.; and Du, D. 2023. Occformer: Dual-path transformer for vision-based 3d semantic occupancy prediction. In *ICCV*.