

# Towards Temporal Fusion Beyond the Field of View for Camera-based Semantic Scene Completion

Jongseong Bae\*, Junwoo Ha\*, Jinnyeong Heo\*, Yeongin Lee\*, Ha Young Kim<sup>†</sup>

Yonsei University  
{js.bae,gkwnsdn0402,wlsud132,zeroim,hayoung.kim}@yonsei.ac.kr

## Abstract

Recent camera-based 3D semantic scene completion (SSC) methods have increasingly explored leveraging temporal cues to enrich the features of the current frame. However, while these approaches primarily focus on enhancing in-frame regions, they often struggle to reconstruct critical out-of-frame areas near the sides of the ego-vehicle, although previous frames commonly contain valuable contextual information about these unseen regions. To address this limitation, we propose the Current-Centric Contextual 3D Fusion (C3DFusion) module, which generates hidden region-aware 3D feature geometry by explicitly aligning 3D-lifted point features from both current and historical frames. C3DFusion performs enhanced temporal fusion through two complementary techniques—historical context blurring and current-centric feature densification—which suppress noise from inaccurately warped historical point features by attenuating their scale, and enhance current point features by increasing their volumetric contribution. Simply integrated into standard SSC architectures, C3DFusion demonstrates strong effectiveness, significantly outperforming state-of-the-art methods on the SemanticKITTI and SSCBench-KITTI-360 datasets. Furthermore, it exhibits robust generalization, achieving notable performance gains when applied to other baseline models.

## Introduction

3D semantic scene completion (SSC) (Song et al. 2017) has recently garnered significant attention as a fundamental 3D perception task, particularly in applications such as autonomous driving (Hu et al. 2023). SSC aims to simultaneously reconstruct voxelized 3D geometry and predict semantic labels for each voxel—an inherently challenging task, especially for real-world deployment. Existing SSC methods are typically categorized by input modality. LiDAR-based methods (Xia et al. 2023; Cheng et al. 2021; Yan et al. 2021; Yang et al. 2021) have demonstrated superior performance and thus become the dominant paradigm; however, their widespread adoption is hindered by the high cost and limited scalability of LiDAR sensors. In contrast, camera-based methods have recently gained momentum, with rapid

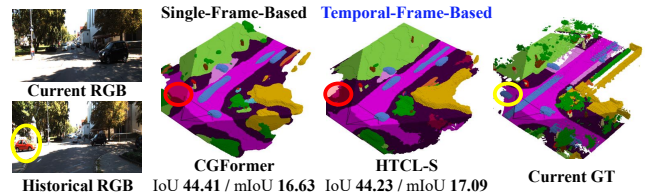


Figure 1: Existing temporal fusion models struggle to complete out-of-frame geometry in the current frame. For example, HTCL-S (Li et al. 2024a), a recent method that performs temporal fusion via 2D feature warping, fails to recover the car on the left side despite its visibility in previous frames, resulting in performance comparable to that of the single-frame-based CGFormer (Yu et al. 2024).

advances significantly narrowing the performance gap compared to their LiDAR-based counterparts.

Boosted by the release of the SemanticKITTI benchmark (Behley et al. 2019), camera-based SSC methods (Li et al. 2023b; Huang et al. 2023; Zhang, Zhu, and Du 2023; Miao et al. 2023; Jiang et al. 2024) have been actively explored, driving continued progress in this field. In particular, given the temporally continuous nature of input in driving scenarios, numerous recent approaches (Li et al. 2024a; Ye et al. 2024; Li et al. 2024b; Wang et al. 2025a; Lu et al. 2025) leverage sequences of past frames to enrich the current frame’s feature representation and improve performance. However, while these methods primarily focus on in-camera-view geometry from the current frame, they tend to overlook out-of-camera-view regions—blind spots that are often located near the ego-vehicle and therefore require especially precise perception for safe navigation. Although past frames provide broader spatial context around the current viewpoint, existing methods still rely heavily on extrapolating from visible geometry to infer these blind spots, often resulting in limited accuracy (see Figure 1).

In this paper, we propose an intuitive yet effective temporal geometry fusion method, Current-Centric Contextual 3D Fusion (C3DFusion), which addresses the aforementioned challenge by explicitly integrating historical and current features in 3D space. Specifically, we map 2D features from all frames directly onto their corresponding lifted 3D points via backprojection, and align the lifted point features from his-

\*These authors contributed equally.

<sup>†</sup>Corresponding author.

torical frames to the current frame’s metric space through geometric warping using known camera poses. To effectively fuse the historical geometry with the current geometry while mitigating geometric discrepancies between them, we introduce two complementary techniques: historical context blurring and current-centric feature densification. The former attenuates the scale of historical point features according to their depth, reducing the influence of potentially inaccurate geometry arising from distance-dependent depth estimation errors. The latter densifies the current lifted points by interpolating over the corresponding 2D feature and depth maps prior to backprojection, thereby enhancing their volumetric contribution to the fused geometric representation.

Our extensive experiments demonstrate the effectiveness of the proposed C3DFusion both quantitatively and qualitatively, particularly against existing camera-based SSC methods that incorporate temporal fusion. Upon integration into a standard camera-based SSC architecture, our method achieves state-of-the-art (SOTA) performance, with IoU and mIoU scores of 47.62 and 18.98 on SemanticKITTI, and 49.28 and 21.74 on SSCBench-KITTI-360 (Li et al. 2024c), respectively, significantly outperforming prior approaches. Furthermore, incorporating C3DFusion into other baseline models yields consistent performance improvements, demonstrating its strong generalization capability.

Our key contributions are as follows:

- To the best of our knowledge, we are the first to primarily address out-of-camera-view completion via temporal cues in camera-based SSC—a long-overlooked yet crucial factor in safety-critical driving applications.
- We propose a simple yet effective temporal geometry fusion method, C3DFusion, that performs perspective alignment between explicitly point-mapped historical and current frame features in the 3D metric space of the current frame.
- To further enhance temporal fusion, we introduce two refinement techniques—historical context blurring and current-centric feature densification—that reduce noise caused by geometric inaccuracies in warped historical features and improve geometric fidelity by emphasizing information from the current frame.
- Built on C3DFusion, our camera-based SSC model achieves strong SOTA performance on the SemanticKITTI and SSCBench-KITTI-360 benchmarks, while demonstrating robust generalization across diverse existing architectures.

## Related Work

**3D Semantic Scene Completion** Since the introduction of the SSC task, early methods (Chen et al. 2020; Li et al. 2020; Zhang et al. 2018) focused primarily on indoor environments, using datasets such as NYUv2 (Silberman et al. 2012). The release of the large-scale SemanticKITTI benchmark subsequently catalyzed research on outdoor SSC, sparking a surge of LiDAR-based approaches (Roldao, De Charette, and Verroust-Blondet 2020; Cheng et al. 2021; Yan et al. 2021; Yang et al. 2021; Xia et al. 2023), which have since become the dominant

paradigm. Recently, camera-based methods have gained momentum due to the affordability and rich contextual information offered by RGB sensors. MonoScene (Cao and De Charette 2022) projects 2D features along optical rays for voxel-wise prediction, while TPVFormer (Huang et al. 2023) lifts features onto multiple planes to capture diverse spatial perspectives. OccFormer (Zhang, Zhu, and Du 2023) adopts a Lift-Splat (LSS) (Phillion and Fidler 2020)-like strategy to construct volumetric context. VoxFormer (Li et al. 2023b) introduces an MAE-like (He et al. 2022) architecture with deformable attention, and Symphonies (Jiang et al. 2024) models instance-level representations via learnable queries. SOTA models such as CGFormer (Yu et al. 2024), ScanSSC (Bae, Ha, and Kim 2025), and L2COcc (Wang et al. 2025b) further combine LSS-style feature lifting with deformable attention to enhance semantic reasoning. Meanwhile, the recent release of the Occ3D-nuScenes benchmark (Tian et al. 2023) has spurred a parallel line of research (Huang et al. 2021; Xu et al. 2024; Ma et al. 2024; Li et al. 2023c; Kim et al. 2025), which leverages multi-camera systems as input. In this work, we focus on the single-view RGB setting and aim to overcome its limited field of view by leveraging temporal cues, while proposing an alternative to the prevalent LSS-style feature lifting paradigm.

**Temporal Fusion for Camera-based 3D Perception** A widely adopted strategy in 3D perception tasks such as object detection is to perform temporal fusion directly in the BEV feature space, either via attention mechanisms (Li et al. 2024d) or by warping and concatenating features across time steps (Huang and Huang 2022; Wang et al. 2024; Yang et al. 2023). In the context of SSC, various temporal fusion techniques have been proposed. VoxFormer-T (Li et al. 2023b) and SGN (Mei et al. 2024) extend attention mechanisms to integrate features across multiple frames. HTCLS (Li et al. 2024a) and Hi-SOP (Li et al. 2024b) utilize contextual pattern affinity to temporally align features from current and past RGB frames within the 2D feature space. FlowScene (Wang et al. 2025a) employs optical flow between adjacent frames to guide occlusion correction and improve voxel refinement. CVT-Occ (Ye et al. 2024), on the other hand, enhances the volumetric representation of the current frame by constructing a cost volume across temporal frames, while CF-SSC (Lu et al. 2025) takes a different approach by synthesizing future frames from past observations to better recover occluded regions. These approaches largely concentrate on regions within the current camera’s field of view and often neglect the reconstruction of out-of-camera-view areas. In contrast, our method explicitly targets these unseen regions by projecting 3D points from both historical and current frames into a unified target space and processing them in a single pass to generate coherent voxel features.

## Method

### Overview

Given a sequence of  $n$  RGB images  $\mathbf{I} = \{\mathbf{I}_i \in \mathbb{R}^{H \times W \times 3} \mid i = t - n + 1, \dots, t\}$  with resolution  $(H, W)$ , we aim to predict voxel-wise semantic class probabilities at time  $t$  within a voxelized 3D space  $\mathbf{V} \in \mathbb{R}^{X \times Y \times Z \times P}$ , where  $(X, Y, Z)$

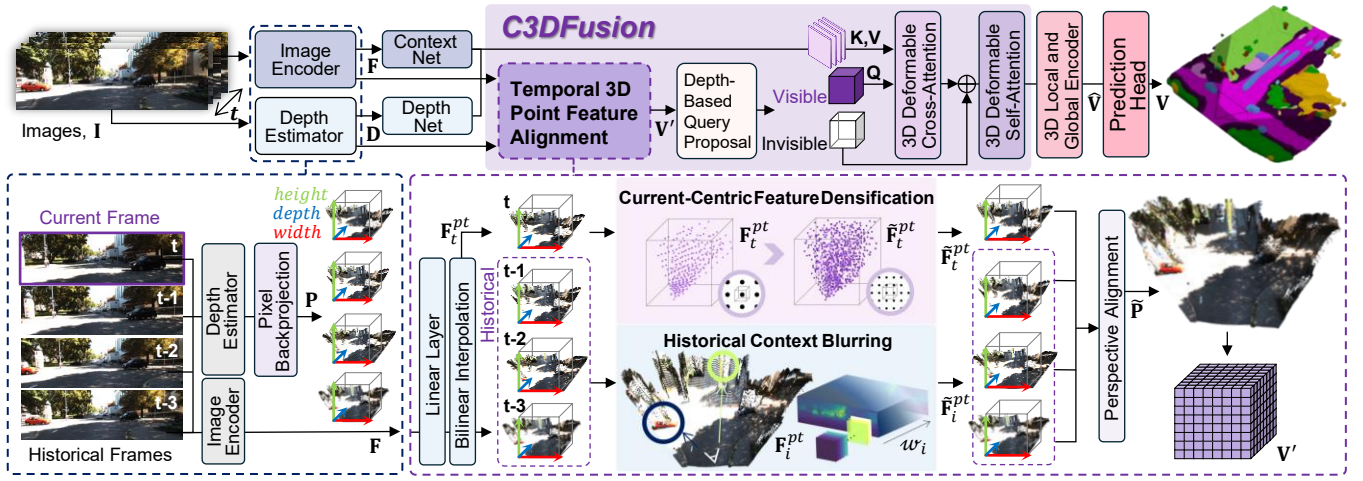


Figure 2: An overview of our model, highlighting the proposed C3DFusion. The symbol ‘ $\oplus$ ’ denotes feature concatenation.

denotes the spatial resolution of the volume and  $P$  is the number of semantic classes, including the empty class. Our model follows the standard architectural paradigm of modern camera-based SSC (Jiang et al. 2024; Yu et al. 2024; Li et al. 2024a; Wang et al. 2025b), consisting of three main components: viewing transformation, voxel processing, and semantic prediction. Among these, the proposed C3DFusion primarily focuses on the viewing transformation stage, which plays a pivotal role in overall SSC performance, as it performs both 2D-to-3D lifting and temporal geometry fusion. An architectural overview is provided in Figure 2.

### C3DFusion

**Temporal 3D Point Feature Alignment** In camera-based SSC, mapping 2D image features into 3D space is a fundamental step. To achieve this, numerous methods adopt backprojection techniques in conjunction with off-the-shelf depth estimators. Early methods (Li et al. 2023b; Jiang et al. 2024) primarily use the resulting 3D points as voxel proposals for cross-attention with 2D image features. More recent methods (Zhang, Zhu, and Du 2023; Yu et al. 2024; Wang et al. 2025b) further advance this process by constructing volumetric geometry using the LSS strategy, which effectively generates a frustum-shaped dense feature volume in 3D space. Since our approach targets temporal alignment in 3D space, a natural extension would be to apply temporal fusion directly to such volumetric features. However, we hypothesize that when extended across multiple frames, these sparsely densified and long-tailed features introduce geometric noise that ultimately degrades semantic prediction accuracy—an effect we quantitatively validate in our experiments. To handle this issue, we instead adopt a strategy that directly maps 2D image features onto 3D lifted point clouds from both historical and current frames, and appropriately aggregates them with a focus on the current perspective.

From the input image sequence  $\mathbf{I}$ , we extract 2D image features  $\mathbf{F} = \{\mathbf{F}_i \in \mathbb{R}^{H' \times W' \times C}\}$  and corresponding depth maps  $\mathbf{D} = \{\mathbf{D}_i \in \mathbb{R}^{H \times W}\}$  using an image encoder and a pretrained depth estimator, where  $(H', W')$  denotes the

spatial resolution of the volume and  $C$  is the feature dimension. Using the depth maps with known camera parameters, each image pixel is backprojected into 3D space to obtain a set of point clouds  $\mathbf{P} = \{\mathbf{P}_i \in \mathbb{R}^{HW \times 3}\}$ , where each point cloud is defined in its corresponding metric 3D coordinate system. To obtain corresponding point features  $\mathbf{F}^{pt} = \{\mathbf{F}_i^{pt} \in \mathbb{R}^{HW \times C'}\}$ , we first apply a linear layer to each feature map  $\mathbf{F}_i$ , followed by bilinear interpolation to match the resolution of the depth maps, as follows:

$$\mathbf{F}_i^{pt} = \text{Flatten}(\text{Bilinear}(\text{Linear}(\mathbf{F}_i), (H, W))), \quad (1)$$

where  $\text{Bilinear}(\cdot, (a, b))$  denotes bilinear interpolation to resolution  $(a, b)$ , and  $\text{Flatten}(\cdot)$  flattens the spatial dimensions of the feature map.

Given the relative poses derived from the extrinsic matrices, the historical points  $\{\mathbf{P}_{t-n+1}, \dots, \mathbf{P}_{t-1}\}$  are warped into the current frame’s coordinate system, resulting in  $\{\tilde{\mathbf{P}}_{t-n+1}, \dots, \tilde{\mathbf{P}}_{t-1}\}$ , for fusion with the current points  $\mathbf{P}_t$ . Instead of directly merging them, we propose two techniques to improve the quality of the fused geometry: historical context blurring and current-centric feature densification.

**Historical Context Blurring** Assuming the ego-vehicle moves continuously forward, the unfiltered warped historical points that remain within the current coordinate system tend to originate from farther regions in the original camera view—since nearer regions have already been passed—and therefore typically have larger depth values. Since we use backprojection based on estimated depth maps, the accuracy of the lifted 3D geometry inherently depends on the quality of the depth estimator, which typically degrades at greater depths (Poggi et al. 2020). To mitigate geometric discrepancies arising from this limitation, we apply a regularization that scales the magnitude of historical point features inversely proportional to their estimated depth values. Specifically, we apply min-max normalization to each historical depth map  $\{\mathbf{D}_{t-n+1}, \dots, \mathbf{D}_{t-1}\}$  independently, subtract the normalized values from 1, and employ the resulting

values as per-point weights  $w_i \in \mathbb{R}^{H \times W}$  in the range  $[0, 1]$ :

$$w_i = 1 - \text{MinMax}(\mathbf{D}_i), \quad (2)$$

where  $\text{MinMax}(\cdot)$  denotes min-max normalization. By element-wise multiplying these weights with the point features  $\mathbf{F}_i^{pt}$ , we obtain the rescaled features  $\tilde{\mathbf{F}}_i^{pt}$ , which are finally associated with their corresponding warped points  $\tilde{\mathbf{P}}_i$ :

$$\tilde{\mathbf{F}}_i^{pt} = w_i \odot \mathbf{F}_i^{pt}, \quad (3)$$

where  $\odot$  denotes element-wise multiplication.

**Current-Centric Feature Densification** For in-camera-view regions of the current frame, there is substantial overlap with preceding frames, resulting in numerous points within these areas in the fused 3D space. However, since each frame contributes a fixed number of points equal to the image grid size  $HW$ , the influence of the current frame’s features may become diluted during temporal aggregation—despite their greater temporal relevance. To emphasize the current frame’s contextual information in these regions, we increase the density of its point cloud by bilinearly interpolating the point feature  $\mathbf{F}_t^{pt}$  and depth map  $\mathbf{D}_t$  prior to backprojection, yielding densified current points  $\tilde{\mathbf{P}}_t$ :

$$\tilde{\mathbf{D}}_t = \text{Bilinear}(\mathbf{D}_t, (\tilde{H}, \tilde{W})), \quad (4)$$

$$\tilde{\mathbf{F}}_t^{pt} = \text{Flatten}(\text{Bilinear}(\mathbf{F}_t^{pt \rightarrow \text{grid}}, (\tilde{H}, \tilde{W}))), \quad (5)$$

where  $\mathbf{F}_t^{pt \rightarrow \text{grid}}$  denotes  $\mathbf{F}_t^{pt}$  rearranged into its original spatial grid form  $(H, W, C')$  before interpolation. By default, we set the upsampling resolution to  $(\tilde{H}, \tilde{W}) = (2H, 2W)$ .

**Voxel Aggregation** Given the temporally aligned point clouds from multiple frames, we define the unified point set as  $\tilde{\mathbf{P}} = \bigcup_{i=t-n+1}^t \tilde{\mathbf{P}}_i$ , and the corresponding set of refined point features as  $\tilde{\mathbf{F}}^{pt} = \bigcup_{i=t-n+1}^t \tilde{\mathbf{F}}_i^{pt}$ . We then discard any points that fall outside the predefined spatial boundaries of the current target voxel grid. The remaining points are voxelized into the volume features  $\mathbf{V}' \in \mathbb{R}^{X' \times Y' \times Z' \times C'}$ . For each voxel at position  $(x, y, z)$  in  $\mathbf{V}'$ , its feature is computed by summing all point features within the voxel and dividing by the number of frames  $n$  as:

$$\mathbf{V}'(x, y, z) = \begin{cases} \frac{1}{n} \sum_{j=1}^v \mathbf{f}_j, & \text{if } v > 0 \\ \mathbf{0}, & \text{if } v = 0 \end{cases} \quad (6)$$

where  $v$  is the number of points falling into the voxel, and  $\{\mathbf{f}_1, \dots, \mathbf{f}_v\} \subset \tilde{\mathbf{F}}^{pt}$  are the associated features.

**MAE-like Voxel Refinement** Following the milestone paradigm of voxel proposal-based refinement (Li et al. 2023b; Jiang et al. 2024; Yu et al. 2024; Wang et al. 2025b), initially occupied voxels in  $\mathbf{V}'$  first undergo cross-attention to supplement lifted features with additional 2D context, while unoccupied voxels are refined via self-attention to extrapolate missing regions. As we use 3D deformable attention (Li et al. 2023a) for the cross-attention stage, current 2D image feature  $\mathbf{F}_t$  and depth map  $\mathbf{D}_t$  are passed through additional context and depth networks (Zhang, Zhu,

and Du 2023; Yu et al. 2024; Wang et al. 2025b), yielding  $\mathbf{F}_{\text{cross}} \in \mathbb{R}^{H' \times W' \times C}$  and  $\mathbf{D}_{\text{cross}} \in \mathbb{R}^{H' \times W' \times B}$ , where  $B$  denotes the number of discretized depth bins. These are then fed into the cross-attention module, whose output is subsequently processed by the self-attention module as follows:

$$\mathbf{V}'_{\text{cross}} = \text{DeformCross}(\mathbf{V}', \mathbf{F}_{\text{cross}}, \mathbf{M}_{\text{cross}}, \mathbf{D}_{\text{cross}}), \quad (7)$$

$$\mathbf{V}'_{\text{self}} = \text{DeformSelf}(\mathbf{V}'_{\text{cross}}, \mathbf{V}'_{\text{cross}}, \mathbf{M}_{\text{self}}), \quad (8)$$

where  $\text{DeformCross}(\cdot, \cdot, \cdot, \cdot)$  and  $\text{DeformSelf}(\cdot, \cdot, \cdot)$  denote 3D deformable cross- and self-attention modules, taking query, key-value, and attention mask as inputs.  $\text{DeformCross}$  additionally leverages discretized depth probability.  $\mathbf{M}_{\text{cross}}$  and  $\mathbf{M}_{\text{self}}$  are binary masks indicating the occupied and unoccupied voxels in  $\mathbf{V}'$ , respectively.

## Voxel Processing

Once the viewing-transformed voxel feature  $\mathbf{V}'_{\text{self}}$  is obtained, we further process it using a voxel backbone network to capture geometric patterns across multiple spatial scales within the voxel space. For this stage, we adopt the voxel processing architecture from CGFormer, a recent approach in single-frame camera-based SSC.

Specifically, given  $\mathbf{V}'_{\text{self}}$ , the voxel processing network is divided into two branches: a voxel-based branch serving as a local encoder implemented with a 3D ResNet (He et al. 2016), and a TPV-based branch serving as a global encoder implemented with a 2D Swin Transformer (Liu et al. 2021):

$$\mathbf{V}'_{\text{vox}} = \text{ResNet3D}(\mathbf{V}'_{\text{self}}), \quad (9)$$

$$\mathbf{V}'_{\text{tpv}} = \{\text{Swin}(\text{Pool}(\mathbf{V}'_{\text{self}}, \text{dim})) \mid \text{dim} \in \{xy, yz, zx\}\}, \quad (10)$$

where  $\text{Pool}(\cdot)$  denotes a max-pooling operation along the specified dimension, resulting in a 2D feature map on the corresponding plane.

The outputs  $\mathbf{V}'_{\text{vox}} \in \mathbb{R}^{X' \times Y' \times Z' \times C'}$  and  $\mathbf{V}'_{\text{tpv}} = \{\mathbf{V}'_{xy} \in \mathbb{R}^{X' \times Y' \times 1 \times C'}, \mathbf{V}'_{yz} \in \mathbb{R}^{1 \times Y' \times Z' \times C'}, \mathbf{V}'_{zx} \in \mathbb{R}^{X' \times 1 \times Z' \times C'}\}$  are aggregated via weighted summation to produce the final voxel feature  $\hat{\mathbf{V}}$ , where the weight  $\mathbf{W} \in \mathbb{R}^{X' \times Y' \times Z' \times 4}$  is generated from  $\mathbf{V}'_{\text{vox}}$  using a linear layer followed by a softmax along the channel dimension:

$$\mathbf{W} = \text{Softmax}(\text{Linear}(\mathbf{V}'_{\text{vox}})), \quad (11)$$

$$\hat{\mathbf{V}} = \sum_{k=1}^4 \mathbf{W}_k \odot \mathbf{V}'_k, \quad (12)$$

where  $\mathbf{W}_k \in \mathbb{R}^{X' \times Y' \times Z' \times 1}$  is the  $k$ -th slice of  $\mathbf{W}$ , and  $\mathbf{V}'_k$  is one of the intermediate voxel features from the set  $\{\mathbf{V}'_{\text{vox}}, \mathbf{V}'_{xy}, \mathbf{V}'_{yz}, \mathbf{V}'_{zx}\}$ .

## Semantic Prediction

The processed 3D feature volume  $\hat{\mathbf{V}}$  is finally fed into the semantic prediction head, which consists of a 3D convolutional layer, followed by normalization and a linear projection. This produces voxel-wise class logits  $\mathbf{V}^{\text{logit}} \in \mathbb{R}^{X' \times Y' \times Z' \times P}$ . To obtain the final semantic prediction volume  $\mathbf{V}$ , the logits  $\mathbf{V}^{\text{logit}}$  are first upsampled via trilinear

interpolation and then passed through a softmax function. This process can be summarized as:

$$\mathbf{V}^{logit} = \text{Linear}(\text{Norm}(\text{Conv3D}(\hat{\mathbf{V}}))), \quad (13)$$

$$\mathbf{V} = \text{Softmax}(\text{Trilinear}(\mathbf{V}^{logit}, (X, Y, Z))), \quad (14)$$

where  $\text{Trilinear}(\cdot, (a, b, c))$  denotes trilinear interpolation to resolution  $(a, b, c)$ .

## Training Loss

Following prior works (Yu et al. 2024; Bae, Ha, and Kim 2025; Wang et al. 2025b), we employ a combination of four losses: cross-entropy loss  $\mathcal{L}_{ce}$ , affinity losses  $\mathcal{L}_{scal}^{geo}$  and  $\mathcal{L}_{scal}^{sem}$ , and depth loss  $\mathcal{L}_d$ . The total loss  $\mathcal{L}$  is defined as:

$$\mathcal{L} = \lambda_{ce}\mathcal{L}_{ce} + \lambda_{scal}^{geo}\mathcal{L}_{scal}^{geo} + \lambda_{scal}^{sem}\mathcal{L}_{scal}^{sem} + \lambda_d\mathcal{L}_d, \quad (15)$$

where  $\lambda_{ce} = \lambda_{scal}^{geo} = \lambda_{scal}^{sem} = 1$ , and  $\lambda_d = 0.001$ .

## Experiments

### Experimental Settings

**Datasets** We conduct experiments on two widely used SSC benchmarks: SemanticKITTI and SSCBench-KITTI360. SemanticKITTI consists of 22 outdoor driving scenes, offering both LiDAR sweeps and stereo RGB images. It is partitioned into 10 training scenes, 1 validation scene, and 11 test scenes. The ground truth voxel grid has a resolution of  $256 \times 256 \times 32$ , covering a spatial extent of  $51.2m \times 51.2m \times 6.4m$ , meaning each voxel corresponds to a volume of  $0.2m \times 0.2m \times 0.2m$ . In total, 21 semantic classes are defined: 19 semantic categories, 1 empty class, and 1 unknown class. SSCBench-KITTI360 shares the same spatial coverage and ground truth voxel grid resolution as SemanticKITTI. The dataset comprises 9 driving scenes, partitioned into 7 for training, 1 for validation, and 1 for testing. In total, 20 semantic classes are defined, including 18 semantic categories, 1 free class, and 1 unknown class.

**Metrics** Following widely adopted practices, we employ Intersection over Union (IoU) and mean IoU (mIoU) as quantitative evaluation metrics: IoU measures class-agnostic scene completion accuracy, while mIoU captures class-specific SSC performance.

### Performance Comparisons

**Quantitative Results** The quantitative performance comparison between our proposed model and existing methods on the SemanticKITTI benchmark is presented in Table 1. The results clearly demonstrate the effectiveness of our approach, which achieves SOTA performance with an IoU of 47.62 and an mIoU of 18.98, significantly surpassing previous methods on both metrics. We attribute this performance gain to the enhanced completion of out-of-camera-view regions—an area that constitutes the core focus of this work. To verify this, we separately evaluate performance on these regions, as shown in the rightmost columns of the table. Our model achieves particularly notable improvements in these challenging areas, reaching an IoU of 44.37 and an mIoU of 17.17, while other methods perform significantly worse on

the same regions. These results validate our core hypothesis that conventional methods underestimate out-of-view (OOV) geometry, which our method successfully completes.

In addition, to demonstrate the generalizability of our method, we report results on the SSCBench-KITTI360 benchmark in Table 2. Our model again achieves strong SOTA performance, with an IoU of 49.28 and an mIoU of 21.74, while also delivering significant improvements in OOV regions, achieving 52.41 IoU and 17.17 mIoU. These results further exhibit the effectiveness of our approach across different datasets.

**Qualitative Results** Figure 3 presents a qualitative comparison between our method and existing open-source camera-based approaches on the SemanticKITTI validation set. In scenarios where critical semantic objects, such as cars and persons, are invisible but located very close to the ego-vehicle in the current frame, temporal-frame-based baseline methods like VoxFormer-T and HTCL-S fail to recover these OOV regions, showing performance comparable to CGFormer, which relies only on a single frame. In contrast, our approach successfully reconstructs the missing structures, demonstrating effective use of temporal cues to accurately capture their spatial context even in the absence of direct visual observations in the current frame.

### Ablation Studies

All ablation experiments are conducted on the validation split of the SemanticKITTI benchmark.

**C3DFusion** Table 3 presents the ablation study of the main components proposed in C3DFusion. For the baseline, we extend CGFormer to process temporal frames by aligning each LSS-style feature volume using perspective warping, followed by a summation of the aligned volumes. Notably, comparing the baseline with variant (a) quantitatively validates our hypothesis from the method section regarding simple temporal LSS feature fusion: replacing LSS feature alignment with point-based geometric temporal fusion results in significant improvements of 0.5 IoU and 1.87 mIoU. Further comparisons between (a) and (b) or (c) highlight the individual contributions of the proposed historical context blurring and current-centric feature densification components to the performance gains, and applying the complete C3DFusion framework ultimately achieves the best IoU and mIoU scores, clearly demonstrating its effectiveness.

Additionally, to assess the practical generalizability of C3DFusion, we conduct further experiments by integrating it with several milestone camera-based SSC methods, as shown in Table 4. For the baseline models, we select originally single-frame methods that adopt the backprojection technique—VoxFormer-S, which uses it solely for voxel proposals, and others that apply it for LSS-based 3D lifting. We extend them to temporal settings by first applying temporal LSS fusion (as in Table 3), then replace it with C3DFusion for performance comparison. The results show that incorporating C3DFusion consistently leads to significant performance improvements across all models. While temporal LSS fusion often shows limited effectiveness in terms of mIoU—particularly with models such as






















				                  																		OOV (val.)		
Method	Input	IoU	mIoU																			IoU	mIoU	
Single-Frame-Based																								
TPVFormer	Mono	34.25	11.26	55.10	27.20	27.40	6.50	14.80	19.20	3.70	1.00	0.50	2.30	13.90	2.60	20.40	1.10	2.40	0.30	11.00	2.90	1.50	30.76	7.87
OccFormer	Mono	34.53	12.32	55.90	30.30	31.50	6.50	15.70	21.60	1.20	1.50	1.70	3.20	16.80	3.90	21.30	2.20	1.10	0.20	11.90	3.80	3.70	30.46	8.68
Symphonies	Stereo	42.19	15.04	58.40	29.30	26.90	11.70	24.70	23.60	3.20	3.60	2.60	5.60	24.20	10.00	23.10	3.20	1.90	2.00	16.10	7.70	8.00	23.48	6.40
CGFormer	Stereo	44.41	16.63	64.30	34.20	34.10	12.10	25.80	26.10	4.30	3.70	1.30	2.70	24.50	11.20	29.30	1.70	3.60	0.40	18.70	8.70	9.30	33.54	9.06
ScanSSC	Stereo	44.54	17.40	66.20	35.90	35.10	12.50	25.30	27.10	3.50	3.50	3.20	6.10	25.20	11.00	30.60	1.80	5.30	0.70	20.50	8.40	8.90	33.60	9.50
L2COcc-D	Stereo	45.37	18.18	68.20	36.90	34.60	16.20	25.80	28.30	4.50	4.90	3.30	7.20	26.20	11.90	32.00	2.10	2.40	0.30	21.60	9.60	9.50	31.85	10.05
Temporal-Frame-Based																								
VoxFormer-T	Stereo	43.21	13.41	54.10	26.90	25.10	7.30	23.50	21.70	3.60	1.90	1.60	4.10	24.40	8.10	24.20	1.60	1.10	0.00	13.10	6.60	5.70	40.21	11.58
HTCL-S	Stereo	44.23	17.09	64.40	34.80	33.80	12.40	25.90	27.30	10.80	1.80	2.20	5.40	25.30	10.80	31.20	1.10	3.10	0.90	21.10	9.00	8.30	33.14	9.04
Hi-SOP	Stereo	44.57	17.49	63.95	34.27	35.85	13.77	25.91	27.35	7.18	2.99	2.59	7.19	26.07	10.35	30.77	1.68	4.81	1.06	20.15	8.70	7.90	-	-
FlowScene	Stereo	45.20	17.70	64.10	35.00	33.70	13.00	27.70	26.40	10.00	4.20	3.10	7.00	26.30	10.00	30.20	3.10	5.10	1.10	20.20	8.90	9.10	-	-
CF-SSC	Stereo	46.21	16.40	61.30	33.30	29.20	11.90	30.40	26.30	4.80	2.60	2.70	6.30	28.50	11.40	28.30	1.50	1.40	0.40	17.70	7.20	6.30	-	-
Ours	Stereo	47.62	18.98	67.00	36.30	33.20	19.30	30.60	29.00	3.30	5.40	4.40	4.70	29.60	14.70	33.80	1.60	2.80	0.30	22.80	11.40	10.40	44.37	17.17

Table 1: Quantitative results on SemanticKITTI hidden test set. ‘OOV’ on the right indicates performance on out-of-view regions, evaluated on the validation set using our implementation. **Bold** / underline highlight the best / second-best, respectively.



















Method	Input	IoU	mIoU																			OOV (test)	
																						IoU	mIoU
Single-Frame-Based																							
TPVFormer	Mono	40.22	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	-	-
OccFormer	Mono	40.27	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	Stereo	44.12	18.58	<u>30.02</u>	1.85	5.90	<b>25.07</b>	<b>12.06</b>	<b>8.20</b>	54.94	13.83	32.76	<b>6.93</b>	35.11	8.58	38.33	11.52	14.01	9.57	<b>14.44</b>	<b>11.28</b>	34.39	11.93
CGFormer	Stereo	48.07	20.05	29.85	3.42	3.96	17.59	6.79	6.63	<b>63.85</b>	17.15	<u>40.72</u>	5.53	42.73	8.22	38.80	24.94	16.24	17.45	10.18	6.77	44.72	15.61
ScanSSC	Stereo	<u>48.29</u>	<u>20.14</u>	29.91	3.78	4.28	14.34	<u>9.08</u>	6.65	62.21	<u>18.16</u>	40.19	5.16	42.68	8.83	<u>38.84</u>	<u>25.50</u>	16.60	<u>19.14</u>	10.30	6.89	45.09	15.44
Temporal-Frame-Based																							
FlowScene	Stereo	46.98	19.12	29.83	<u>4.44</u>	3.78	16.71	8.71	<u>7.77</u>	60.70	16.99	39.59	<u>6.01</u>	<u>43.17</u>	<u>9.45</u>	37.32	25.14	<u>17.35</u>	18.12	10.63	7.56	-	-
CF-SSC	Stereo	45.79	19.10	28.10	3.39	<u>6.87</u>	16.76	7.75	5.68	59.01	16.80	37.60	4.95	42.16	8.26	36.14	21.89	14.73	17.72	9.73	7.14	-	-
Ours	Stereo	<b>49.28</b>	<b>21.74</b>	<b>31.16</b>	<b>5.39</b>	<b>7.01</b>	<u>18.12</u>	8.25	5.66	<u>63.70</u>	<b>19.12</b>	<b>41.64</b>	5.09	<b>43.93</b>	<b>10.43</b>	<b>40.73</b>	<b>27.62</b>	<b>19.30</b>	<b>23.08</b>	<u>12.34</u>	<u>8.74</u>	<b>52.41</b>	<b>23.72</b>

Table 2: Quantitative results on SSCBench-KITTI-360 test set. ‘OOV’ on the right indicates performance on out-of-view regions, evaluated on the test set using our implementation. **Bold** / underline highlight the best / second-best, respectively.

VoxFormer-S and CGFormer—C3DFusion reliably boosts mIoU across various architectures. Remarkably, methods such as OccFormer and ScanSSC, when integrated with C3DFusion, surpass the previous SOTA mIoU of 18.22 achieved by L2COcc. These findings underscore the strong generalizability of C3DFusion across diverse architectures and its practical utility within camera-based SSC pipelines.

	TPFA	HCB	CCFD	IoU	mIoU
Baseline				48.59	16.58
(a)	✓			49.09	18.45
(b)	✓	✓		48.99	18.88
(c)	✓		✓	48.87	18.86
<b>Ours</b>	✓	✓	✓	<b>49.53</b>	<b>19.31</b>

Table 3: Ablation study of C3DFusion. ‘TPFA’, ‘HCB’, and ‘CCFD’ denote temporal 3D point feature alignment, historical context blurring, and current-centric feature densification, respectively.

**Intensity of Current-Centric Feature Densification** Table 5 presents the ablation study on the intensity of current-

Method	IoU	mIoU
VoxFormer-S	44.02	12.35
+ Temporal LSS fusion	44.39 (+0.37)	9.56 (-2.79)
+ C3DFusion	<b>45.98 (+1.96)</b>	<b>15.12 (+2.77)</b>
OccFormer	36.50	13.46
+ Temporal LSS fusion	44.48 (+7.98)	16.91 (+3.45)
+ C3DFusion	<b>44.83 (+8.36)</b>	<b>18.29 (+4.83)</b>
ScanSSC	45.95	17.12
+ Temporal LSS fusion	49.31 (+3.36)	17.67 (+0.55)
+ C3DFusion	<b>49.89 (+3.94)</b>	<b>18.73 (+1.61)</b>
CGFormer	45.99	16.87
+ Temporal LSS fusion	48.59 (+2.60)	16.58 (-0.29)
+ C3DFusion (Ours)	<b>49.53 (+3.54)</b>	<b>19.31 (+2.44)</b>

Table 4: Generalization of C3DFusion.

centric feature densification by varying the target resolution for point feature interpolation. The best performance is achieved with a  $2\times$  interpolation factor, while further increasing the resolution results in a progressive degradation of mIoU. We attribute this decline primarily to geometric inaccuracies arising from the interpolation of the estimated

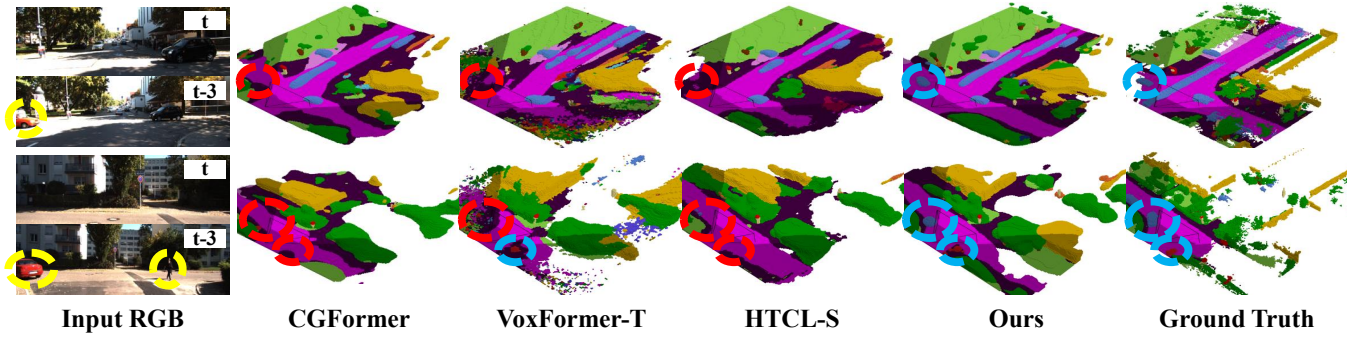


Figure 3: Visual comparison of our model against other recent camera-based methods on the SemanticKITTI validation set.

Interpolation Factor	IoU	mIoU
$\times 1$	48.99	18.88
$\times 2$	<b>49.53</b>	<b>19.31</b>
$\times 3$	48.53	19.08
$\times 4$	49.21	18.86

Table 5: Ablation study on the intensity of current-centric feature densification, controlled by the interpolation factor.

History	Current	IoU	mIoU
✓		<b>49.53</b>	<b>19.31</b>
	✓	48.74	18.29
✓	✓	49.22	18.76

Table 6: Ablation study on historical context blurring.

# of Frames $n$	IoU	mIoU
1	45.99	16.37
2	47.72	18.00
3	48.90	18.62
<b>4</b>	<b>49.53</b>	<b>19.31</b>
5	49.20	18.58

Table 7: Ablation study on the number of input frames  $n$ .

depth map—excessive upsampling amplifies depth estimation errors, leading to less reliable geometric representations and ultimately harming performance.

**Context Blurring to Current Frame** We conduct an ablation study on extending historical context blurring to the current frame, as shown in Table 6. As expected, this extension results in a substantial drop in both IoU and mIoU, underscoring the critical importance of preserving current point features for accurate scene completion—even though these features also inherit geometric inaccuracies arising from the distance-dependent errors in estimated depth.

**Number of Frames in Sequence** Table 7 presents an ablation study on the effect of varying the number of frames in the input sequence. The results show steady improvements in both IoU and mIoU as the number of frames increases, with performance peaking at 4 frames before declining. We infer that additional preceding frames introduce noise rather than useful contextual information. Accordingly, we empirically select 4 frames as the input sequence.

## Efficiency Comparisons

In Table 8, we compare the number of parameters and memory usage of our model with other open-source camera-based methods, particularly the baseline using temporal LSS fusion strategy. Our model demonstrates a strong performance–efficiency trade-off among temporal-frame-based methods such as VoxFormer-T and HTCL-S, significantly outperforming HTCL-S despite using fewer parameters and less memory. Moreover, compared to the temporal LSS fusion strategy, integrating C3DFusion yields substantially better performance with comparable parameter count and memory usage, highlighting the superiority of our temporal fusion approach in terms of efficiency.

Method	# of Params.	Memory	mIoU
VoxFormer-T	58 M	14763 MB	13.35
HTCL-S	182 M	34593 MB	17.13
CGFormer	163 M	16013 MB	16.87
+ Temporal LSS fusion	163 M	22687 MB	16.58
+ C3DFusion (Ours)	160 M	23987 MB	19.31

Table 8: Efficiency comparison with existing methods.

## Conclusion

In this work, we address the challenge of completing regions beyond the current camera’s field of view by leveraging temporal context in camera-based SSC—a critical capability for autonomous driving applications. To this end, we propose C3DFusion, a temporal geometry fusion approach that aligns 3D point features mapped directly from 2D image features via backprojection. We further introduce techniques such as historical context blurring and current-centric feature densification to mitigate noise from warped historical points and strengthen the contribution of current-frame features in the aggregated point cloud. Extensive experiments demonstrate that C3DFusion not only achieves SOTA performance on standard benchmarks, but also consistently improves performance across a variety of architectures. Owing to its simplicity, generalizability, and strong effectiveness, we believe C3DFusion offers valuable insights for camera-based SSC and broader 3D perception tasks, and can serve as a solid foundation for future research in temporal fusion.

## Acknowledgments

This work was supported by the National Research Foundation (NRF) grant funded by Korea Government (MSIT) (No. 2023R1A2C200337911) and the Ministry of Education of the Republic of Korea and the National Research Foundation of Korea (NRF-2024S1A5C3A03046579).

## References

- Bae, J.; Ha, J.; and Kim, H. Y. 2025. Three Cars Approaching within 100m! Enhancing Distant Geometry by Tri-Axis Voxel Scanning for Camera-based Semantic Scene Completion. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 11939–11948.
- 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 *Proceedings of the IEEE/CVF international conference on computer vision*, 9297–9307.
- Cao, A.-Q.; and De Charette, R. 2022. Monoscene: Monocular 3d semantic scene completion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3991–4001.
- Chen, X.; Lin, K.-Y.; Qian, C.; Zeng, G.; and Li, H. 2020. 3d sketch-aware semantic scene completion via semi-supervised structure prior. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4193–4202.
- Cheng, R.; Agia, C.; Ren, Y.; Li, X.; and Bingbing, L. 2021. S3cnet: A sparse semantic scene completion network for lidar point clouds. In *Conference on Robot Learning*, 2148–2161. PMLR.
- He, K.; Chen, X.; Xie, S.; Li, Y.; Dollár, P.; and Girshick, R. 2022. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 16000–16009.
- He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.
- Hu, Y.; Yang, J.; Chen, L.; Li, K.; Sima, C.; Zhu, X.; Chai, S.; Du, S.; Lin, T.; Wang, W.; et al. 2023. Planning-oriented autonomous driving. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 17853–17862.
- Huang, J.; and Huang, G. 2022. Bevdet4d: Exploit temporal cues in multi-camera 3d object detection. *arXiv preprint arXiv:2203.17054*.
- Huang, J.; Huang, G.; Zhu, Z.; Ye, Y.; and Du, D. 2021. Bevdet: High-performance multi-camera 3d object detection in bird-eye-view. *arXiv preprint arXiv:2112.11790*.
- Huang, Y.; Zheng, W.; Zhang, Y.; Zhou, J.; and Lu, J. 2023. Tri-perspective view for vision-based 3d semantic occupancy prediction. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 9223–9232.
- 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 20258–20267.
- Kim, J.; Kang, C.; Lee, D.; Choi, S.; and Choi, J. W. 2025. Protoocc: Accurate, efficient 3d occupancy prediction using dual branch encoder-prototype query decoder. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 4284–4292.
- Li, B.; Deng, J.; Zhang, W.; Liang, Z.; Du, D.; Jin, X.; and Zeng, W. 2024a. Hierarchical Temporal Context Learning for Camera-based Semantic Scene Completion. *arXiv preprint arXiv:2407.02077*.
- Li, B.; Jin, X.; Deng, J.; Sun, Y.; Wang, X.; and Zeng, W. 2024b. Hierarchical Context Alignment with Disentangled Geometric and Temporal Modeling for Semantic Occupancy Prediction. *arXiv preprint arXiv:2412.08243*.
- Li, H.; Zhang, H.; Zeng, Z.; Liu, S.; Li, F.; Ren, T.; and Zhang, L. 2023a. Dfa3d: 3d deformable attention for 2d-to-3d feature lifting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 6684–6693.
- Li, J.; Han, K.; Wang, P.; Liu, Y.; and Yuan, X. 2020. Anisotropic convolutional networks for 3d semantic scene completion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3351–3359.
- Li, Y.; Li, S.; Liu, X.; Gong, M.; Li, K.; Chen, N.; Wang, Z.; Li, Z.; Jiang, T.; Yu, F.; Wang, Y.; Zhao, H.; Yu, Z.; and Feng, C. 2024c. SSCBench: A Large-Scale 3D Semantic Scene Completion Benchmark for Autonomous Driving. *arXiv:2306.09001*.
- Li, Y.; Yu, Z.; Choy, C.; Xiao, C.; Alvarez, J. M.; Fidler, S.; Feng, C.; and Anandkumar, A. 2023b. Voxformer: Sparse voxel transformer for camera-based 3d semantic scene completion. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 9087–9098.
- Li, Z.; Wang, W.; Li, H.; Xie, E.; Sima, C.; Lu, T.; Yu, Q.; and Dai, J. 2024d. Bevformer: learning bird’s-eye-view representation from lidar-camera via spatiotemporal transformers. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.
- Li, Z.; Yu, Z.; Austin, D.; Fang, M.; Lan, S.; Kautz, J.; and Alvarez, J. M. 2023c. Fb-occ: 3d occupancy prediction based on forward-backward view transformation. *arXiv preprint arXiv:2307.01492*.
- Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; and Guo, B. 2021. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, 10012–10022.
- Lu, H.; Su, Y.; Zhang, X.; and Hu, H. 2025. One Step Closer: Creating the Future to Boost Monocular Semantic Scene Completion. *arXiv preprint arXiv:2507.13801*.
- Ma, Q.; Tan, X.; Qu, Y.; Ma, L.; Zhang, Z.; and Xie, Y. 2024. Cotr: Compact occupancy transformer for vision-based 3d occupancy prediction. In *Proceedings of the IEEE/CVF*



*Conference on Computer Vision and Pattern Recognition*, 19936–19945.

Mei, J.; Yang, Y.; Wang, M.; Zhu, J.; Ra, J.; Ma, Y.; Li, L.; and Liu, Y. 2024. Camera-based 3d semantic scene completion with sparse guidance network. *IEEE Transactions on Image Processing*.

Miao, R.; Liu, W.; Chen, M.; Gong, Z.; Xu, W.; Hu, C.; and Zhou, S. 2023. Occdepth: A depth-aware method for 3d semantic scene completion. *arXiv preprint arXiv:2302.13540*.

Philion, J.; and Fidler, S. 2020. Lift, splat, shoot: Encoding images from arbitrary camera rigs by implicitly unprojecting to 3d. In *European conference on computer vision*, 194–210. Springer.

Poggi, M.; Aleotti, F.; Tosi, F.; and Mattoccia, S. 2020. On the uncertainty of self-supervised monocular depth estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 3227–3237.

Roldao, L.; De Charette, R.; and Verroust-Blondet, A. 2020. Lmscnet: Lightweight multiscale 3d semantic completion. In *2020 International Conference on 3D Vision (3DV)*, 111–119. IEEE.

Silberman, N.; Hoiem, D.; Kohli, P.; and Fergus, R. 2012. Indoor segmentation and support inference from rgb-d images. In *European conference on computer vision*, 746–760. Springer.

Song, S.; Yu, F.; Zeng, A.; Chang, A. X.; Savva, M.; and Funkhouser, T. 2017. Semantic scene completion from a single depth image. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1746–1754.

Tian, X.; Jiang, T.; Yun, L.; Mao, Y.; Yang, H.; Wang, Y.; Wang, Y.; and Zhao, H. 2023. Occ3d: A large-scale 3d occupancy prediction benchmark for autonomous driving. *Advances in Neural Information Processing Systems*, 36: 64318–64330.

Wang, M.; Wu, F.; Li, R.; Qin, Y.; Tang, Z.; and Li, K. 2025a. Learning Temporal 3D Semantic Scene Completion via Optical Flow Guidance. *arXiv preprint arXiv:2502.14520*.

Wang, R.; Ma, Y.; Yao, Y.; Tao, S.; Li, H.; Zhu, Z.; Liu, Y.; and Zuo, X. 2025b. L2COcc: Lightweight Camera-Centric Semantic Scene Completion via Distillation of LiDAR Model. *arXiv preprint arXiv:2503.12369*.

Wang, Y.; Chen, Y.; Liao, X.; Fan, L.; and Zhang, Z. 2024. Panoocc: Unified occupancy representation for camera-based 3d panoptic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 17158–17168.

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 *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 17642–17651.

Xu, J.; Peng, L.; Cheng, H.; Xia, L.; Zhou, Q.; Deng, D.; Qian, W.; Wang, W.; and Cai, D. 2024. Regulating intermediate 3d features for vision-centric autonomous driving. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 6306–6314.

Yan, X.; Gao, J.; Li, J.; Zhang, R.; Li, Z.; Huang, R.; and Cui, S. 2021. Sparse single sweep lidar point cloud segmentation via learning contextual shape priors from scene completion. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, 3101–3109.

Yang, C.; Chen, Y.; Tian, H.; Tao, C.; Zhu, X.; Zhang, Z.; Huang, G.; Li, H.; Qiao, Y.; Lu, L.; et al. 2023. Bevformer v2: Adapting modern image backbones to bird’s-eye-view recognition via perspective supervision. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 17830–17839.

Yang, X.; Zou, H.; Kong, X.; Huang, T.; Liu, Y.; Li, W.; Wen, F.; and Zhang, H. 2021. Semantic segmentation-assisted scene completion for lidar point clouds. In *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 3555–3562. IEEE.

Ye, Z.; Jiang, T.; Xu, C.; Li, Y.; and Zhao, H. 2024. Cvt-occ: Cost volume temporal fusion for 3d occupancy prediction. In *European Conference on Computer Vision*, 381–397. Springer.

Yu, Z.; Zhang, R.; Ying, J.; Yu, J.; Hu, X.; Luo, L.; Cao, S.-Y.; and Shen, H.-L. 2024. Context and geometry aware voxel transformer for semantic scene completion. *Advances in Neural Information Processing Systems*, 37: 1531–1555.

Zhang, J.; Zhao, H.; Yao, A.; Chen, Y.; Zhang, L.; and Liao, H. 2018. Efficient semantic scene completion network with spatial group convolution. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 733–749.

Zhang, Y.; Zhu, Z.; and Du, D. 2023. Occformer: Dual-path transformer for vision-based 3d semantic occupancy prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 9433–9443.