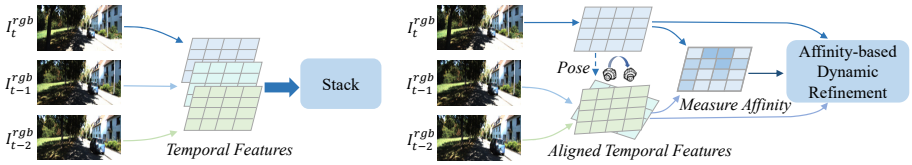
Hierarchical Temporal Context Learning for Camera-based Semantic Scene Completion

Bohan Li[1,2], Jiajun Deng[3], Wenyao Zhang[1,2],
Zhujin Liang[4], Dalong Du[4], Xin Jin[2] ⋆, and Wenjun Zeng[1,2]

[1] Shanghai Jiao Tong University, Shanghai, China
[2] Ningbo Institute of Digital Twin, Eastern Institute of Technology, Ningbo, China
[3] The University of Adelaide, Adelaide, Australia
[4] PhiGent Robotics, Beijing, China
{bohan_li, wy_zhang}@sjtu.edu.cn, jiajun.deng@adelaide.edu.au,
{zhujin.liang, dalong.du}@phigent.ai,
{jinxin, wenjunzengvp}@eitech.edu.cn



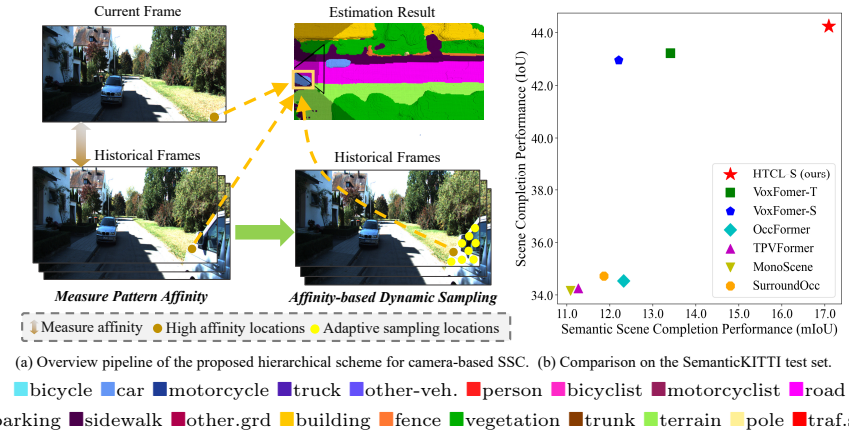
(a) Previous straightforward temporal method. (b) Our hierarchical temporal context learning method.

Fig. 1: Our hierarchical temporal context learning method versus previous straightforward temporal method (VoxFormer-T [22]) in semantic scene completion.

Abstract. Camera-based 3D semantic scene completion (SSC) is pivotal for predicting complicated 3D layouts with limited 2D image observations. The existing mainstream solutions generally leverage temporal information by roughly stacking history frames to supplement the current frame, such straightforward temporal modeling inevitably diminishes valid clues and increases learning difficulty. To address this problem, we present **HTCL**, a novel **H**ierarchical **T**emporal **C**ontext **L**earning paradigm for improving camera-based semantic scene completion. The primary innovation of this work involves decomposing temporal context learning into two hierarchical steps: (a) cross-frame affinity measurement and (b) affinity-based dynamic refinement. Firstly, to separate critical relevant context from redundant information, we introduce the pattern affinity with scale-aware isolation and multiple independent learners for fine-grained contextual correspondence modeling. Subsequently, to dynamically compensate for incomplete observations, we adaptively refine the feature sampling locations based on initially identified locations with high affinity and their neighboring relevant regions. Our method ranks 1^{st} on the SemanticKITTI benchmark and even surpasses LiDAR-based methods in terms of mIoU on the OpenOccupancy benchmark. Our code is available on `https://github.com/Arlo0o/HTCL`.

Keywords: Semantic Scene Completion · Temporal Context Learning

⋆ Corresponding author



(a) Overview pipeline of the proposed hierarchical scheme for camera-based SSC. (b) Comparison on the SemanticKITTI test set.

Fig. 2: (a) Overview pipeline of the proposed method, which measures contextual pattern affinity across temporal frames and dynamically samples relevant context. Our method shows promising performance in comprehending and completing semantic scenes even outside the camera's field of view, as indicated by the car highlighted with the yellow box. (b) Comparison with state-of-the-art camera-based semantic scene completion methods [7, 16, 22, 46, 52] on the SemanticKITTI test set.

1 Introduction

The comprehension of holistic 3D scenes holds paramount importance in autonomous driving systems [16, 22, 46]. This capability directly impacts the planning and obstacle avoidance functionalities of autonomous vehicles, thereby influencing their overall safety and efficiency. However, due to the limitations of real-world sensors such as restricted field of view and measurement noise, this task remains a challenging problem. 3D semantic scene completion (SSC) has been proposed to address the challenges by jointly inferring the geometry and semantics of the scenario from incomplete observations [7, 14, 16, 22, 34, 35, 47, 51].

Given the inherent 3D nature, numerous semantic scene completion (SSC) solutions [14, 34, 35, 47, 51] rely on LiDAR as the primary technique for precise 3D location measurement. Although the LiDAR sensor provides accurate depth information, its deployment introduces a large overhead for both the cost and manual efforts. Therefore, it is necessary to explore an efficient approach for high-fidelity SSC with cost-effective devices. This motivation led to the investigation of camera-based solutions, which are characterized by superior deployment efficiency and the abundance of rich visual context [7, 16, 46, 52].

The early attempts in camera-based SSC methods [7, 36, 39] commonly rely solely on the current frame, which can only provide very limited observation to recover the 3D geometry and semantics. To enrich contextual information, the pioneering work VoxFormer-T [22] proposes to make use of temporal coherence by stacking multiple history frames as supplements to the current frame. Nevertheless, as depicted in Figure 1, this approach assumes that temporal features from different viewpoints are originally corresponded at the pixel level,

thereby just using a straightforward aggregation. However, the shared semantic content undergoes uncertain positional changes across different perspectives. Therefore, directly integrating images from different timestamps may result in blurred predictive information. This ambiguity compromises the stability of semantic occupancy predictions and imposes difficulties on temporal modeling.

To alleviate this issue, we propose a new Hierarchical Temporal Context Learning (HTCL) paradigm, which improves the temporal feature aggregation for accurate 3D semantic scene completion. HTCL takes RGB images from different timestamps to hierarchically infer the 3D semantic occupancy for fine-grained scene comprehension. As depicted in Figure 2 (a), our hierarchical temporal context modeling includes two sequential steps: (1) we explicitly measure the contextual pattern affinity between the current and historical frames, highlighting the most relevant patterns; (2) we adaptively refine the sampling locations based on the preliminary high-affinity locations and their nearby relevant context to dynamically compensate for incomplete observations. Specifically, we first leverage epipolar homo-warping to explicitly align the temporal invariant feature representations and establish temporal feature volumes to fully maintain fine-grained context. Then, we introduce *scale-aware isolation* and *incorporation of diverse independent learning* in the cross-frame affinity measurement to facilitate better affinity distribution modeling in SSC. Subsequently, to dynamically compensate for incomplete observations, we resort to investigating the critical locations with high affinities and the neighboring relevant context. Technically, a multi-level deformable 3D block conditioned with the affinity weights is involved to adaptively refine the sampling locations. Finally, a weighted voxel cross-attention is introduced to aggregate the reliable temporal content.

We conduct extensive experiments to validate the merits of our proposed HTCL. As shown in Figure 2(b), our proposed method achieves significant superiority over existing camera-based methods in terms of geometry (IoU) and semantics (mIoU). Remarkably, our camera-based approach outperforms state-of-the-art VoxFormer-T [22] on the SemanticKITTI benchmark and even surpasses LiDAR-based methods on the OpenOccupancy benchmark in terms of mIoU. Our main contributions are summarized as follows:

- A temporal context learning paradigm with a hierarchical scheme to fully exploit dynamic and dependable 3D semantic scene completion.
- An affinity measurement strategy with scale-aware isolation and multiple independent learners for fine-grained contextual correspondence modeling.
- An affinity-based dynamic refinement schedule to reassemble the temporal content and adaptively compensate for incomplete observations.
- Our method achieves state-of-the-art performance among all camera-based SSC methods on the SemanticKITTI and OpenOccupancy benchmarks.

2 Related Work

2.1 3D Semantic Scene Completion

Semantic Scene Completion (SSC), also known as semantic occupancy prediction, is a dense 3D perception task that jointly addresses semantic segmentation

and scene completion [2, 6]. Numerous previous works leverage LiDAR as the primary input to take advantage of the 3D geometrical information [14, 34, 51]. Due to the cost-effectiveness and portability, camera-based 3D SSC is recently gaining increasing attention [2,6,7,9,16,20,22,34,35,38–40,47,49]. MonoScene [7] first proposes to infer geometry and semantics from a single RGB image with 2D-3D features projection. Inspired by this, a lot of the following works extend the domain of camera-based 3D scene perception [16,19,46,52]. TPVFormer [16] introduces a tri-perspective view to describe the fine-grained representation of a 3D scene. OccFormer [52] introduces a dual-path transformer network to process dense 3D features for semantic occupancy prediction. However, these methods attempt to describe the complicated 3D scene from single-timestep images, which is ineffective for such an inherently ill-posed problem due to incomplete visual cues. In this paper, we advocate delving into reliable temporal content to dynamically aggregate semantic context and compensate for incomplete observations.

2.2 Temporal Information Modeling in 3D Visual Perception

The utilization of temporal information is recently highlighted in temporal 3D object detection [17,21,24–26,29,44,50,50] and video depth estimation [5,27,45] to enhance prediction performance. Temporal 3D object detection solutions focus on coarse-grained predictions at region-level [24, 26], while video depth estimation methods [5, 27] aim to establish matching correspondence from sequential video frames. Consequently, such strategies are insufficient for semantic scene completion, where fine-grained features are essential for dense semantic perception. VoxFormer-T [22] builds the first temporal pipeline for camera-based SSC by simply stacking the features from different frames, while the temporal correspondence modeling for the dense perception task of SSC remains unexplored. In this paper, we propose to explicitly model the temporal context correlation with pattern affinity to aggregate reliable temporal content and compensate for incomplete observations.

3 Methodology

3.1 Preliminary

Given a set of input temporal RGB images $I_{set}^{rgb} = \{I_t^{rgb}, I_{t-1}^{rgb}, \cdots\}$, our objective is to simultaneously estimate the semantic and geometric properties of the 3D scene. Notably, our focus is exclusively on current and historical image frames, omitting consideration of future frames [22] to formulate a more practical scheme for real-world applications. The scene is depicted as a voxel grid $\mathbf{V}$ with dimensions $\mathbb{R}^{H \times W \times Z}$, where H, W, Z denote the height, width and depth of the voxel grid, respectively. Each voxel within the grid is associated with a distinct semantic class represented by C, where C takes values from the set $\{c_0, c_1, \ldots, c_N\}$. The voxel can either correspond to empty space denoted as c_0 or to a particular semantic class from the set $\{c_1, c_2, \ldots, c_N\}$. Here, N represents the total count of

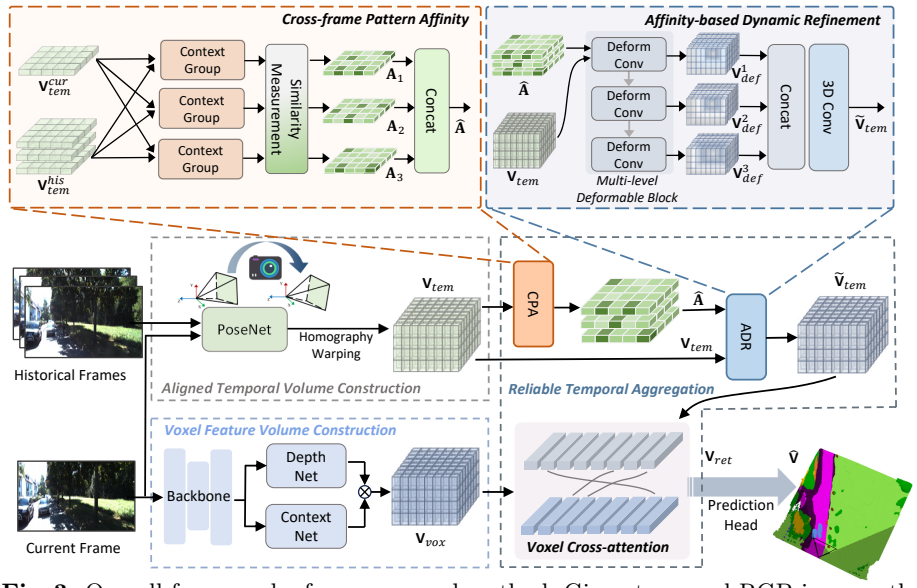


Fig. 3: Overall framework of our proposed method. Given temporal RGB images, the Aligned Temporal Volume is constructed with explicit epipolar homograph warping, while the Voxel Feature Volume is built by extending the LSS paradigm. Afterward, the Reliable Temporal Aggregation is introduced to dynamically aggregate reliable relevant temporal content for fine-grained semantic scene prediction.

available semantic classes. Our objective is to leverage the proposed framework, denoted as Θ, to learn a transformation:

$$\widehat{\mathbf{V}} = \Theta(I_t^{rgb}, I_{t-1}^{rgb}, \cdots), \tag{1}$$

where $\widehat{\mathbf{V}}$ denotes the estimated 3D semantic voxel grid, aiming to approximate the ground truth voxel grid $\mathbf{V}$.

3.2 Overall Framework

As depicted in Figure 3, the overall framework of our proposed method mainly consists of three components: Aligned Temporal Volume Construction in the upper branch, Voxel Feature Volume Construction in the lower branch, and Reliable Temporal Aggregation for fine-grained SSC prediction.

Aligned Temporal Volume Construction. To construct the temporal feature volume $\mathbf{V}_{tem}$, we feed the current and historical frames into a lightweight PoseNet [5,15] to generate the temporal contextual volume $\mathbf{V}_{tem}$ with homography warping. Different from computing matching costs in typical temporal depth estimation solutions [5,27,45], we advocate to maintain the context features with $\mathbf{V}_{tem}$, further details are introduced in Section 3.3.

Voxel Feature Volume Construction. To construct the voxel feature volume $\mathbf{V}_{vox}$, a UNet backbone based on pre-trained EfficientNetB7 [41] is firstly employed to generate features with a spatial dimension of $\mathbb{R}^{H/4\times W/4}$. Next, we

extend the LSS [23, 31] paradigm following recent studies [19, 52] to build the voxel feature volume $\mathbf{V}_{vox}$ from the outer product of the contextual information and the depth distribution. To model the depth distribution, off-the-shelf monocular [3] or stereo [10] depth estimation networks are utilized with depth hypothesis planes of 192. By default, we employ the stereo depth estimation to form a stereo-based pipeline of *HTCL-S*. Moreover, we construct another monocular-based pipeline of *HTCL-M* with the monocular depth estimation, broadening the applicability to scenarios without stereo inputs.

Reliable Temporal Aggregation. We leverage the temporal volume $\mathbf{V}_{tem}$ to form the cross-frame affinity $\hat{\mathbf{A}}$, quantifying the contextual correspondence between current and historical features. Subsequently, we employ the cross-frame affinity to reassemble the temporal content and dynamically refine the sampling locations, yielding the reliable temporal volume $\widetilde{\mathbf{V}}_{tem}$. Further details on the Cross-frame Pattern Affinity (CPA) are introduced in Section 3.4, while the Affinity-based Dynamic Refinement (ADR) is presented in Section 3.5.

A Weighted Voxel Attention (WVA) is employed to aggregate reliable temporal content. Given $\mathbf{V}_{vox}$ and $\widetilde{\mathbf{V}}_{tem}$ as inputs, the query Q is generated from $\mathbf{V}_{vox}$, the key K and the value V are generated from $\widetilde{\mathbf{V}}_{tem}$. During the early training phase, unregulated temporal information could impair the learning of the voxel feature volume. To mitigate this, a flexible learning mechanism involving weighted voxel cross-attention is leveraged in the aggregation process:

$$\mathbf{V}_{ret} = \alpha \cdot \texttt{CrossAtt}(Q, K, V) + \mathbf{V}_{vox}, \tag{2}$$

where the learnable coefficient α is initialized with 0 and gradually increases during training. $\mathbf{V}_{ret}$ represents the aggregated volume, which is fed into an SSC head with upsampling and a softmax layer for SSC prediction $\widehat{\mathbf{V}}$ following [7,52].

3.3 Temporal Content Alignment

Given the fine-grained nature of the Semantic Scene Completion (SSC) task, constructing a dense temporal-aligned feature representation is crucial for accurate and robust perception. Rather than simply stacking the input images from different viewpoints [22], we propose to first align the temporal invariant content with explicit homography transformation.

As shown in Figure 3, the current and historical frames are first fed into a lightweight PoseNet following video depth estimation [15, 45] to generate the relative camera pose for photometric reprojection. Next, we leverage the current and historical frames to generate the current feature map F_t and historical feature maps $\{F_{t-1}, \cdots, F_{t-n}\}$. Following [5, 45], we construct the warped historical features through homography warping with the relative camera pose and alternative depth hypothesis planes, which is formed as:

$$\text{Warp}(\mathbf{p}) = \mathbf{K}_i \cdot \left(\mathbf{R}_{0,i} \cdot \left({\mathbf{K}_0}^{-1} \cdot \mathbf{p} \cdot d_j\right) + \mathbf{t}_{0,i}\right), \tag{3}$$

where $\{\mathbf{K}_i\}_{i=0}^{N-1}$ and $\{[\mathbf{R}_{0,i} \mid \mathbf{t}_{0,i}]\}_{i=1}^{N-1}$ denote camera intrinsic parameters and extrinsic parameters, respectively. d_j denotes j^{th} hypothesized depth of pixel $\mathbf{p}$ in F_t. Following that, we build a historical feature volume $\mathbf{V}_{tem}^{his}$ by aggregating all warped historical features in the canonical space. The historical feature volume contains geometric compatibility with different depth values between the current and historical frames. Next, we lift F_t along the depth dimension as [30, 45] to generate the current feature volume $\mathbf{V}_{tem}^{cur}$. We concatenate $\mathbf{V}_{tem}^{his}$ and $\mathbf{V}_{tem}^{cur}$ following [45] to construct the temporal feature volume $\mathbf{V}_{tem}$:

$$\begin{aligned}\mathbf{V}_{tem} &= \texttt{Concat}\left\{(\mathbf{V}_{tem}^{cur}, \mathbf{V}_{tem}^{his}), \dim = \mathbb{C}\right\} \\ &= \texttt{Concat}\left\{\texttt{Lift}(F_t), \texttt{Warp}(F_{t-1}, \cdots, F_{t-n})\right\}.\end{aligned} \tag{4}$$

The temporal volume $\mathbf{V}_{tem}$ benefits the semantic scene modeling by explicitly aligning the contextual features across different timesteps. In the following section, we elaborate on fully exploiting reliable information according to contextual correspondence with $\mathbf{V}_{tem}$.

Why feature volume instead of cost volume? The key distinction arises from the nature of camera-based Semantic Scene Completion, which is fundamentally not a matching task but rather a dense perception and reconstruction problem. Consequently, instead of directly computing matching costs within the temporal feature volume, our approach prioritizes the maintenance of fine-grained feature context. Moreover, to quantify the relevance of regional patterns within the temporal information, we construct auxiliary pattern affinity between the current and historical features.

3.4 Cross-frame Pattern Affinity Measurement

Although the temporal volume is explicitly aligned, it mixes redundant context from different frames, making it insufficient to directly model the scene representations corresponding to the current frame. Therefore, we propose to construct Cross-frame Pattern Affinity (CPA) to measure the regional contextual correspondence between the historical feature volume $\mathbf{V}_{his}$ and current feature volume $\mathbf{V}_{cur}$.

Similarity Measurement. As a classic similarity metric, cosine similarity is commonly used in semantic analysis [12, 13, 33] and information retrieval [18, 32] for correlation measurement. Given two vectors of α and β, the cosine similarity is calculated as:

$$\texttt{sim}(\alpha, \beta) = \texttt{cos}(\boldsymbol{\alpha}, \boldsymbol{\beta}) = \frac{\boldsymbol{\alpha} \cdot \boldsymbol{\beta}}{||\boldsymbol{\alpha}|| * ||\boldsymbol{\beta}||}. \tag{5}$$

However, the original cosine similarity may yield high similarity scores with two dissimilar vectors [37]. This limitation is acknowledged and rectified through the scale-aware isolation [1], which takes into consideration different pattern scales. Nevertheless, these solutions tend to emphasize vector orientations and encounter challenges when assessing similarity within dense distributions. To overcome these drawbacks, ensemble learning techniques, as discussed in [48],

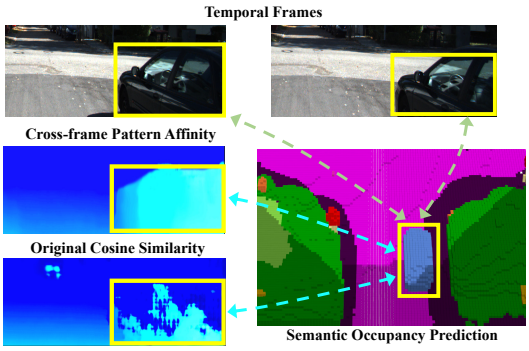


Fig. 4: Visualization of the heat maps from our proposed Cross-frame Pattern Affinity (CPA) and the original cosine similarity.

leverage a diverse set of independent learners to address the aforementioned undesirable properties, thereby enhancing the effectiveness of dense similarity measurements.

Given these concerns, we identify the criteria of an optimal similarity measurement strategy for fine-grained representations in SSC: *incorporation of diverse independent learning* and *scale-aware isolation*. In pursuit of this objective, we advocate employing the scale-aware isolated cosine similarity and taking multi-group context as inputs for affinity computation with dense distributions. Our strategy is implemented through two key steps:

- Incorporate different pattern scales from multi-group context to enable diverse independent similarity learning for fine-grained representations in SSC.
- Generate cosine similarities with scale-aware isolation and aggregate them for reliable pattern affinity measurement.

Multi-group Context Generation. To facilitate diverse independent similarity learning, 3D atrous convolutions with different dilation rates are employed to construct multi-group contextual features. Specifically, the historical feature volume $\mathbf{V}_{tem}^{his}$ is processed by a set of atrous convolutions to generate historical multi-group context $\mathbf{H}_i$ $(i \in (1, 2, 3))$:

$$\mathbf{H}_i = \texttt{GN}\left(\delta\left(\texttt{Atrous}_i(\mathbf{V}_{tem}^{his})\right)\right), \tag{6}$$

where $\texttt{GN}$ and δ denote group normalization and GELU activation, respectively. The atrous convolutions are employed in parallel with dilation rates of 1, 2, and 4, respectively. Note that the current multi-group context $\mathbf{C}_i$ is generated in a symmetrical manner from the current feature volume $\mathbf{V}_{tem}^{cur}$:

$$\mathbf{C}_i = \texttt{GN}\left(\delta\left(\texttt{Atrous}_i(\mathbf{V}_{tem}^{cur})\right)\right). \tag{7}$$

Measure Pattern Affinity for Dense SSC. We upgrade Equation 5 with two primary modifications to formulate the pattern affinity measurement for fine-grained contextual correspondence modeling in SSC. Firstly, we consider different pattern scales of the multi-group context and compute the pattern

affinity $\mathbf{A}_i$ with each scale i. These independent group-scale affinity matrices are further aggregated along the channel dimension. Secondly, we subtract the respective averages during the affinity computation within each group scale to achieve scale-aware isolation. The formulas are represented as:

$$\begin{aligned}\mathbf{A}_i &= \texttt{sim}(\mathbf{C}_i, \mathbf{H}_i) \\ &= \frac{\sum_{j=0}^{C}(\mathbf{C}_i^j - \overline{\mathbf{C}}_i)(\mathbf{H}_i^j - \overline{\mathbf{H}}_i)}{\sqrt{\sum_{j=0}^{C}(\mathbf{C}_i^j - \overline{\mathbf{C}}_i)^2}\sqrt{\sum_{j=0}^{C}(\mathbf{H}_i^j - \overline{\mathbf{H}}_i)^2}},\end{aligned} \tag{8}$$

$$\hat{\mathbf{A}} = \texttt{Concat}\left\{(\mathbf{A}_1, \mathbf{A}_2, \mathbf{A}_3), \dim = \mathbb{C}\right\}, \tag{9}$$

where the affinity matrices $\mathbf{A}_i$ of different group scales are concatenated along the channel dimension to get the cross-frame pattern affinity $\hat{\mathbf{A}}$. The input context matrices $\mathbf{C}_i$ and $\mathbf{H}_i$ are taken as high-dimension vectors with different group scales. $\overline{\mathbf{C}}_i$ and $\overline{\mathbf{H}}_i$ represent averaged context matrices of each group scale. As illustrated in Figure 4, the affinity map from Cross-frame Pattern Affinity (CPA) effectively signifies contextual correspondence within the temporal content.

3.5 Affinity-based Dynamic Refinement

Given our objective of completing and comprehending the 3D scene corresponding to the current frame, it is essential to assign greater weights to the most relevant locations. Simultaneously, investigating their neighboring relevant context is also critical to compensate for incomplete observations.

To this end, we propose to adaptively refine the feature sampling locations based on the obtained identified high-affinity locations and their neighboring relevant regions. We implement the above ideas with 3D deformable convolutions [11, 42]. Specifically, we accomplish the dynamic refinement by introducing the affinity-based correspondence weights and deformable position offsets. In the context of a sampling grid window K_w, the formula is expressed as:

$$\mathbf{V}_{def} = \sum_{k=1}^{K_w} w_k \cdot \mathbf{V}_{tem}(\mathbf{p} + \mathbf{p}_k + \Delta\mathbf{p}_k) \cdot a_k, \tag{10}$$

where K_w represents the number of points in the sampling process. $\Delta\mathbf{p}_k$ denotes the additional offset in the sampling grid. w_k denotes the spatial feature weight and a_k represents the affinity weight from the cross-frame pattern affinity $\hat{\mathbf{A}}$.

To further reason about dynamic modeling through hierarchical context, we optimize the refinement process by considering contextual information from different feature levels. As illustrated in Figure 3, we construct a multi-level deformable block with three cascade 3D deformable convolutions. The output features are aggregated to generate the reliable temporal volume $\widetilde{\mathbf{V}}_{tem}$:

$$\widetilde{\mathbf{V}}_{tem} = \mathbf{W}\left(\texttt{Concat}\left\{(\mathbf{V}_{def}^1, \mathbf{V}_{def}^2, \mathbf{V}_{def}^3), \dim = \mathbb{C}\right\}\right). \tag{11}$$

where the multi-level deformable temporal volumes $\mathbf{V}^i_{def}\,(i \in (1,2,3))$ are concatenated along the channel dimension and processed with a 3D convolution layer $\mathbf{W}$ for dimension reduction.

4 Experiment

4.1 Datasets and Metrics

SemanticKITTI. The SemanticKITTI [2] dataset comprises 22 outdoor scenes with LiDAR scans and stereo images. The ground truth is voxelized as $256\times256\times32$ grids. Each voxel grid has a size of ($0.2m$, $0.2m$, $0.2m$) and is annotated with 21 semantic classes (19 semantics, 1 free and 1 unknown). Following [7, 22], we divide the 22 outdoor scenes into 10 training scenes, 1 validation scene, and 1 test scene. Our proposed HTCL is evaluated on the SemanticKITTI with both temporal stereo images (HTCL-S) and monocular images (HTCL-M).
OpenOccupancy. The OpenOccupancy [43] dataset extends the nuScene [4] dataset by providing dense semantic occupancy annotations. The dataset holds 850 scenes of 34K keyframes with 360-degree LiDAR scans. We split the whole dataset into 28130 training frames and 6019 validation frames following [43]. Each frame holds 400K occupied voxels with 17 semantic labels. Note that we exclusively apply monocular-based HTCL-M on the OpenOccupancy dataset due to the unavailability of stereo images.
Evaluation Metrics. Following the previous works [7, 22], we utilize the mean Intersection over Union **(mIoU)** as our primary metric to assess the performance of the semantic scene completion (SSC) task. Additionally, we report the Intersection over Union **(IoU)** to evaluate the performance of the class-agnostic scene completion (SC) task.

4.2 Experimental Setup

We follow the common practice [7, 22, 52] to initialize the encoder part of our UNet with the pretrained weight of EfficientNetB7 [41]. By default, our model takes the current frame and the previous 3 image frames as inputs. We implemented our model on PyTorch with a batch size of 4. The model is trained 24 epochs with the AdamW optimizer [28]. The learning rate is set to 1×10^{-4} with a weight decay of 0.01.

4.3 Performance.

Quantitative Comparison. As reported in Table 1, we compare our HTCL with recent public methods on the SemanticKITTI dataset, including VoxFormer [22], OccFormer [52], SurroundOcc [46], TPVFormer [16] and MonoScene [7]. VoxFomer-T is a temporal baseline with current and historical 4 images as inputs. We can observe that our proposed method outperforms all the other methods significantly. Compared to VoxFomer-T, our method achieves a remarkable relative

Table 1: Quantitative results on the SemanticKITTI test set with the state-of-the-art SSC methods. The "S-T", "S" and "M" denote temporal stereo images, single-frame stereo images, and single-frame monocular images, respectively.

Methods	HTCL-S (ours)	VoxFormer-T	VoxFormer-S	OccFormer	SurroundOcc	TPVFormer	MonoScene
Input	S-T	S-T	S	M	M	M	M
IoU	**44.23**	43.21	42.95	34.53	34.72	34.25	34.16
mIoU	**17.09**	13.41	12.20	12.32	11.86	11.26	11.08
car	**27.30**	21.70	20.80	21.60	20.60	19.20	18.80
bicycle	1.80	**1.90**	1.00	1.50	1.60	1.00	0.50
motorcycle	**2.20**	1.60	0.70	1.70	1.20	0.50	0.70
truck	**5.70**	3.60	3.50	1.20	1.40	3.70	3.30
other-veh.	**5.40**	4.10	3.70	3.20	4.40	2.30	4.40
person	1.10	1.60	1.40	**2.20**	1.40	1.10	1.00
bicyclist	**3.10**	1.10	2.60	1.10	2.00	2.40	1.40
motorcyclist	**0.90**	0.00	0.20	0.20	0.10	0.30	0.40
road	**64.40**	54.10	53.90	55.90	56.90	55.10	54.70
parking	**33.80**	25.10	21.10	31.50	30.20	27.40	24.80
sidewalk	**34.80**	26.90	25.30	30.30	28.30	27.20	27.10
other.grd	**12.40**	7.30	5.60	6.50	6.80	6.50	5.70
building	**25.90**	23.50	19.80	15.70	15.20	14.80	14.40
fence	**21.10**	13.10	11.10	11.90	11.30	11.00	11.10
vegetation	**25.30**	24.40	22.40	16.80	14.90	13.90	14.90
trunk	**10.80**	8.10	7.50	3.90	3.40	2.60	2.40
terrain	**31.20**	24.20	21.30	21.30	19.30	20.40	19.50
pole	**9.00**	6.60	5.10	3.80	3.90	2.90	3.30
traf.sign	**8.30**	5.70	4.90	3.70	2.40	1.50	2.10

Table 2: Quantitative results on the OpenOccupancy validation set with the state-of-the-art SSC methods. The "L", "M", "M-D" and "M-T" denote LiDAR inputs, monocular images, monocular images with depth maps and temporal monocular images, respectively. The LiDAR points are projected and densified to generate the depth maps.

Methods	HTCL-M (ours)	JS3C-Net	LMSCNet	3DSketch	AICNet	TPVFormer	MonoScene
Input	M-T	L	L	M-D	M-D	M	M
IoU	21.4	**30.2**	27.3	25.6	23.8	15.3	18.4
mIoU	**14.1**	12.5	11.5	10.7	10.6	7.8	6.9
barrier	**14.8**	14.2	12.4	12.0	11.5	9.3	7.1
bicycle	**10.2**	3.4	4.2	5.1	4.0	4.1	3.9
bus	**14.8**	13.6	12.8	10.7	11.8	11.3	9.3
car	**18.9**	12.0	12.1	12.4	12.3	10.1	7.2
const. veh.	**7.6**	7.2	6.2	6.5	5.1	5.2	5.6
motorcycle	**11.3**	4.3	4.7	4.0	3.8	4.3	3.0
pedestrian	**12.3**	7.3	6.2	5.0	6.2	5.9	5.9
traffic cone	**9.6**	6.8	6.3	6.3	6.0	5.3	4.4
trailer	5.5	**9.2**	8.8	8.0	8.2	6.8	4.9
truck	**13.5**	9.1	7.2	7.2	7.5	6.5	4.2
drive. suf.	**32.5**	27.9	24.2	21.8	24.1	13.6	14.9
other flat	**21.7**	15.3	12.3	14.8	13.0	9.0	6.3
sidewalk	**20.7**	14.9	16.6	13.0	12.8	8.3	7.9
terrain	**17.7**	16.2	14.1	11.8	11.5	8.0	7.4
manmade	5.8	**14.0**	13.9	12.0	11.6	9.2	10.0
vegetation	8.5	**24.9**	22.2	21.2	20.2	8.2	7.6

improvement in mIoU even with fewer historical inputs (3 vs. 4). We also report quantitative results on the OpenOccupancy validation set in Table 2. To provide depth maps for AICNet [20] and 3DSketch [8], LiDAR points are projected and densified following OpenOccupancy [43]. Despite LiDAR's inherent advantage in terms of IoU due to more accurate 3D geometric measurement, our HTCL surpasses all the other methods (including LiDAR-based LMSCNet [35] and JS3C-Net [49]) in terms of mIoU, which demonstrates the effectiveness of our method for semantic scene completion.

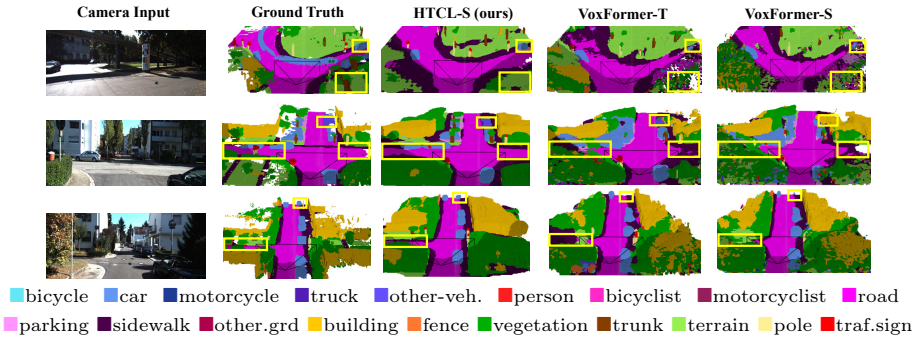


Fig. 5: Qualitative results on the SemanticKITTI validation set. Our HTCL-S captures more complete and accurate scenery layouts compared with VoxFormer.

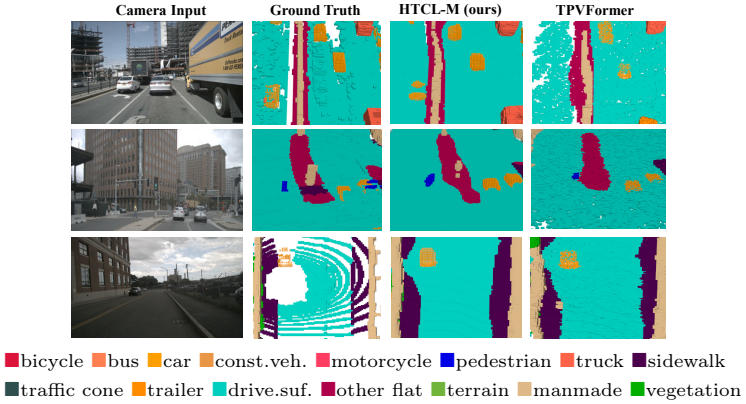


Fig. 6: Qualitative results on the OpenOccupancy validation set. Our HTCL-M generates more complete and comprehensive scenes compared with ground truth.

Qualitative Comparison. Figure 5 provides a qualitative comparison between our proposed method and VoxFormer on the SemanticKITTI validation set. We can observe that the real-world scenes are complex and the annotated ground truth is relatively sparse, which poses challenges for completely reconstructing the semantic scenes from limited visual cues. Compared with VoxFormer, our method captures a more complete and accurate scenery layout (e.g., crossroads in the second and third rows). Moreover, our method effectively hallucinates more proper scenery outside of the camera field of view (e.g., shadow areas in the first and second rows) and demonstrates significant superiority over moving objects (e.g., trucks in the second row). We also visualize the prediction results of our proposed method on the OpenOccupancy validation set as shown in Figure 6. Our proposed method generates much denser and more realistic results compared with the ground truth.

Temporal Stereo Variants Evaluation. To ensure a fair and comprehensive comparison, we implement temporal stereo variants of the baselines as shown in Table 3. Following VoxFormer-T, we employ stacked temporal stereo images as inputs to get variants of MonoScene‡, TPVFormer‡ and OccFormer‡. Note that VoxFormer-T originally employs 4 previous frames, while the other stereo

Table 3: Evaluation results of temporal stereo variants. For MonoScene‡, TPVFormer‡ and OccFormer‡, we employ stacked temporal stereo images as inputs following VoxFormer-T.

Methods	Input	mIoU (%) ↑	Time (s) ↓
MonoScene‡ (2022)	Stereo-T	12.96	**0.281**
TPVFormer‡ (2023)	Stereo-T	13.21	0.324
OccFormer‡ (2023)	Stereo-T	13.57	0.348
VoxFormer-T (2023)	Stereo-T	13.35	0.307
HTCL-M (ours)	Mono-T	16.11	0.289
HTCL-S (ours)	Stereo-T	**17.13**	0.297

Table 4: Ablation study for different architectural components on the SemanticKITTI validation set. The full names of different components are in Sec 4.4.

TCA		CPA		ADR		WVA		IoU (%)	mIoU (%)
Feature Volume	Cost Volume	Scale-aware Isolation	Multi-group	Affinity	Deformable	Coefficient	CrossAtt		
	✓	✓	✓	✓	✓	✓	✓	44.01	16.02
✓			✓	✓	✓	✓	✓	43.07	15.18
✓		✓		✓	✓	✓	✓	43.15	15.26
✓		✓	✓		✓	✓	✓	42.79	14.65
✓		✓	✓	✓		✓	✓	42.96	15.14
✓		✓	✓	✓	✓		✓	43.97	15.85
✓		✓	✓	✓	✓	✓		44.13	15.98
✓		✓	✓	✓	✓	✓	✓	**45.51**	**17.13**

variants employ 3 previous frames as ours. As shown in the table, our method efficiently achieves superior performance with the same temporal inputs.

4.4 Ablation Study

We conduct adequate ablations for our proposed method on the SemanticKITTI validation set. Specifically, we analyze the impact of different architectural components in Table 4 and study the influence of temporal inputs in Table 5.

Temporal Content Alignment (TCA). The ablation study for Temporal Content Alignment (TCA) is reported in the second row of Table 4. We can observe that replacing the cost volume with the feature volume yields obvious performance gains, improving the IoU and mIoU by 1.50 and 1.11, respectively. We attribute this enhancement to the fine-grained feature context preservation.

Cross-frame Pattern Affinity (CPA). The ablation of Cross-frame Pattern Affinity (CPA) is detailed in the third row of Table 4. As we can see, equipping the original cosine similarity with the scale-aware isolation and introducing the multi-group context generation lead to significant performance enhancement, improving the mIoU by 1.95 and 1.87, respectively.

Affinity-based Dynamic Refinement (ADR). The ablation study for the Affinity-based Dynamic Refinement (ADR) is conducted by removing the affinity weights and replacing the deformable convolutions with normal convolutions, as detailed in the fourth row of Table 4. Leveraging the affinity information is effective in modeling contextual correspondence, as the procedure results in a noticeable performance gain of 2.72 IoU and 2.48 mIoU. Furthermore, dynamic refinement with the deformable convolutions offers efficient and flexible contextual modeling, improving IoU and mIoU by 2.55 and 1.99, respectively.

Weighted Voxel Attention (WVA). The ablation study about Weighted Voxel Attention (WVA) is shown in Figure 7 and the fifth row of Table 4. We remove the learnable coefficient and replace the voxel cross-attention with naive concatenation for comparison. It is evident in Figure 7 that the adoption of our proposed strategy leads to a more rapid and stable convergence of the entire model. Additionally, Table 4 shows noteworthy improvements through the incorporation of the learnable coefficient and the voxel cross-attention, enhancing the mIoU by 1.28 and 1.15, respectively.

Table 5: Effect of using a different number of temporal frames. These models are evaluated on the SemanticKITTI validation set.

Temporal Inputs					mIoU (%) ↑	Time (s) ↓
I_{t-1}^{rgb}	I_{t-2}^{rgb}	I_{t-3}^{rgb}	I_{t-4}^{rgb}	I_{t-5}^{rgb}		
✓					15.08	0.268
✓	✓				16.43	0.283
✓	✓	✓			17.13	0.297
✓	✓	✓	✓		17.31	0.311
✓	✓	✓	✓	✓	17.42	0.324

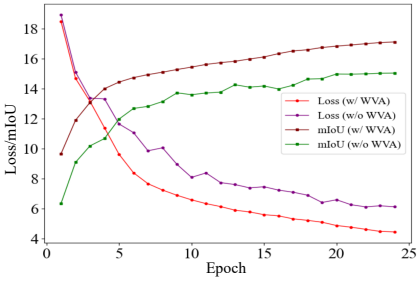


Fig. 7: The convergence and performance curves of WVA.

Temporal Inputs. We report the semantic scene completion performance and running time with temporal inputs of different frame numbers, which are detailed in Table 5. As we can observe, the effectiveness gain of more than 3 previous frames is relatively slight with more running time, thus we adopt 3 frames as the default setting to balance between efficiency and effectiveness.

5 Conclusion

In this paper, we introduce HTCL, an innovative hierarchical temporal in-context learning paradigm for semantic scene completion (SSC). To highlight the most relevant patterns, we introduce the pattern affinity to measure the contextual correspondence between the current and historical frames. Subsequently, to dynamically compensate for incomplete observations, we propose to adaptively refine the feature sampling locations based on the initially high-affinity locations and their neighboring relevant regions. Our method outperforms state-of-the-art camera-based methods and even surpasses LiDAR-based methods for semantic scene completion. We hope HTCL could inspire further research in camera-based temporal modeling for SSC and its applications in 3D visual perception.

Acknowledgments

This work was supported in part by NSFC 62302246 and ZJNSFC under Grant LQ23F010008, and supported by High Performance Computing Center at Eastern Institute of Technology, Ningbo, and Ningbo Institute of Digital Twin.

References

1. Anastasiu, D.C., Karypis, G.: L2ap: Fast cosine similarity search with prefix l-2 norm bounds. In: International Conference on Data Engineering. IEEE (2014)
2. Behley, J., Garbade, M., Milioto, A., Quenzel, J., Behnke, S., Stachniss, C., Gall, J.: Semantickitti: A dataset for semantic scene understanding of lidar sequences. In: ICCV (2019)
3. Bhat, S.F., Alhashim, I., Wonka, P.: Adabins: Depth estimation using adaptive bins. In: CVPR (2021)
4. Caesar, H., Bankiti, V., Lang, A.H., Vora, S., Liong, V.E., Xu, Q., Krishnan, A., Pan, Y., Baldan, G., Beijbom, O.: nuscenes: A multimodal dataset for autonomous driving. In: CVPR (2020)
5. Cai, C., Ji, P., Yan, Q., Xu, Y.: Riav-mvs: Recurrent-indexing an asymmetric volume for multi-view stereo. In: CVPR (2023)
6. Cai, Y., Chen, X., Zhang, C., Lin, K.Y., Wang, X., Li, H.: Semantic scene completion via integrating instances and scene in-the-loop. In: CVPR (2021)
7. Cao, A.Q., de Charette, R.: Monoscene: Monocular 3d semantic scene completion. In: CVPR (2022)
8. Chen, X., Lin, K.Y., Qian, C., Zeng, G., Li, H.: 3d sketch-aware semantic scene completion via semi-supervised structure prior. In: CVPR (2020)
9. Cheng, R., Agia, C., Ren, Y., Li, X., Bingbing, L.: S3cnet: A sparse semantic scene completion network for lidar point clouds. In: Conference on Robot Learning (2021)
10. Cheng, X., Zhong, Y., Harandi, M., Dai, Y., Chang, X., Li, H., Drummond, T., Ge, Z.: Hierarchical neural architecture search for deep stereo matching. NeurIPS **33** (2020)
11. Dai, J., Qi, H., Xiong, Y., Li, Y., Zhang, G., Hu, H., Wei, Y.: Deformable convolutional networks. In: ICCV (2017)
12. Evangelopoulos, N., Zhang, X., Prybutok, V.R.: Latent semantic analysis: five methodological recommendations. European Journal of Information Systems **21**(1), 70–86 (2012)
13. Evangelopoulos, N.E.: Latent semantic analysis. Wiley Interdisciplinary Reviews: Cognitive Science **4**(6) (2013)
14. Garbade, M., Chen, Y.T., Sawatzky, J., Gall, J.: Two stream 3d semantic scene completion. In: CVPRW (2019)
15. Guizilini, V., Ambrus, R., Pillai, S., Raventos, A., Gaidon, A.: 3d packing for self-supervised monocular depth estimation. In: CVPR (2020)
16. Huang, Y., Zheng, W., Zhang, Y., Zhou, J., Lu, J.: Tri-perspective view for vision-based 3d semantic occupancy prediction. In: CVPR (2023)
17. Kopf, J., Rong, X., Huang, J.B.: Robust consistent video depth estimation. In: CVPR (2021)
18. Korenius, T., Laurikkala, J., Juhola, M.: On principal component analysis, cosine and euclidean measures in information retrieval. Information Sciences **177**(22), 4893–4905 (2007)
19. Li, B., Sun, Y., Jin, X., Zeng, W., Zhu, Z., Wang, X., Zhang, Y., Okae, J., Xiao, H., Du, D.: Stereoscene: Bev-assisted stereo matching empowers 3d semantic scene completion (2023)
20. Li, J., Han, K., Wang, P., Liu, Y., Yuan, X.: Anisotropic convolutional networks for 3d semantic scene completion. In: CVPR (2020)

21. Li, S., Luo, Y., Zhu, Y., Zhao, X., Li, Y., Shan, Y.: Enforcing temporal consistency in video depth estimation. In: ICCV (2021)
22. Li, Y., Yu, Z., Choy, C., Xiao, C., Alvarez, J.M., Fidler, S., Feng, C., Anandkumar, A.: Voxformer: Sparse voxel transformer for camera-based 3d semantic scene completion. In: CVPR (2023)
23. Li, Y., Ge, Z., Yu, G., Yang, J., Wang, Z., Shi, Y., Sun, J., Li, Z.: Bevdepth: Acquisition of reliable depth for multi-view 3d object detection. In: AAAI (2023)
24. Lin, X., Lin, T., Pei, Z., Huang, L., Su, Z.: Sparse4d: Multi-view 3d object detection with sparse spatial-temporal fusion. arXiv preprint arXiv:2211.10581 (2022)
25. Liu, Y., Wang, T., Zhang, X., Sun, J.: Petr: Position embedding transformation for multi-view 3d object detection. In: ECCV. Springer (2022)
26. Liu, Y., Yan, J., Jia, F., Li, S., Gao, A., Wang, T., Zhang, X.: Petrv2: A unified framework for 3d perception from multi-camera images. In: ICCV (2023)
27. Long, X., Liu, L., Li, W., Theobalt, C., Wang, W.: Multi-view depth estimation using epipolar spatio-temporal networks. In: CVPR (2021)
28. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101 (2017)
29. Luo, X., Huang, J.B., Szeliski, R., Matzen, K., Kopf, J.: Consistent video depth estimation. ACM Transactions on Graphics (ToG) **39** (2020)
30. Newcombe, R.A., Lovegrove, S.J., Davison, A.J.: Dtam: Dense tracking and mapping in real-time. In: ICCV (2011)
31. Philion, J., Fidler, S.: Lift, splat, shoot: Encoding images from arbitrary camera rigs by implicitly unprojecting to 3d. In: ECCV (2020)
32. Rahutomo, F., Kitasuka, T., Aritsugi, M.: Semantic cosine similarity. In: ICAST (2012)
33. Ramachandran, L., Gehringer, E.F.: Automated assessment of review quality using latent semantic analysis. In: ICALT. IEEE (2011)
34. Rist, C.B., Emmerichs, D., Enzweiler, M., Gavrila, D.M.: Semantic scene completion using local deep implicit functions on lidar data. IEEE transactions on pattern analysis and machine intelligence **44**(10) (2021)
35. Roldao, L., de Charette, R., Verroust-Blondet, A.: Lmscnet: Lightweight multiscale 3d semantic completion. In: 3DV (2020)
36. Roldao, L., De Charette, R., Verroust-Blondet, A.: 3d semantic scene completion: A survey. International Journal of Computer Vision **130**(8) (2022)
37. Sarwar, B., Karypis, G., Konstan, J., Riedl, J.: Item-based collaborative filtering recommendation algorithms. In: WWW (2001)
38. Silberman, N., Hoiem, D., Kohli, P., Fergus, R.: Indoor segmentation and support inference from rgbd images. ECCV (2012)
39. Song, S., Yu, F., Zeng, A., Chang, A.X., Savva, M., Funkhouser, T.: Semantic scene completion from a single depth image. In: CVPR (2017)
40. Straub, J., Whelan, T., Ma, L., Chen, Y., Wijmans, E., Green, S., Engel, J.J., Mur-Artal, R., Ren, C., Verma, S., et al.: The replica dataset: A digital replica of indoor spaces. arXiv preprint arXiv:1906.05797 (2019)
41. Tan, M., Le, Q.: Efficientnet: Rethinking model scaling for convolutional neural networks. In: ICML (2019)
42. Wang, F., Galliani, S., Vogel, C., Speciale, P., Pollefeys, M.: Patchmatchnet: Learned multi-view patchmatch stereo. In: CVPR (2021)
43. Wang, X., Zhu, Z., Xu, W., Zhang, Y., Wei, Y., Chi, X., Ye, Y., Du, D., Lu, J., Wang, X.: Openoccupancy: A large scale benchmark for surrounding semantic occupancy perception. ICCV (2023)

44. Wang, Y., Wang, P., Yang, Z., Luo, C., Yang, Y., Xu, W.: Unos: Unified unsupervised optical-flow and stereo-depth estimation by watching videos. In: CVPR (2019)
45. Watson, J., Mac Aodha, O., Prisacariu, V., Brostow, G., Firman, M.: The temporal opportunist: Self-supervised multi-frame monocular depth. In: CVPR (2021)
46. Wei, Y., Zhao, L., Zheng, W., Zhu, Z., Zhou, J., Lu, J.: Surroundocc: Multi-camera 3d occupancy prediction for autonomous driving. In: ICCV (2023)
47. Wu, S.C., Tateno, K., Navab, N., Tombari, F.: Scfusion: Real-time incremental scene reconstruction with semantic completion. In: 3DV (2020)
48. Xia, P., Zhang, L., Li, F.: Learning similarity with cosine similarity ensemble. Information sciences **307**, 39–52 (2015)
49. Yan, X., Gao, J., Li, J., Zhang, R., Li, Z., Huang, R., Cui, S.: Sparse single sweep lidar point cloud segmentation via learning contextual shape priors from scene completion. In: AAAI (2021)
50. Zhang, H., Shen, C., Li, Y., Cao, Y., Liu, Y., Yan, Y.: Exploiting temporal consistency for real-time video depth estimation. In: ICCV (2019)
51. Zhang, J., Zhao, H., Yao, A., Chen, Y., Zhang, L., Liao, H.: Efficient semantic scene completion network with spatial group convolution. In: ECCV (2018)
52. Zhang, Y., Zhu, Z., Du, D.: Occformer: Dual-path transformer for vision-based 3d semantic occupancy prediction. ICCV (2023)