A Spatiotemporal Approach to Tri-Perspective Representation for 3D Semantic Occupancy Prediction

Sathira Silva[*1,2], **Savindu Wannigama**[*1], **Gihan Jayatilaka**[3], **Muhammad Haris Khan**[2], **Roshan Ragel**[1]

[1]University of Peradeniya, Peradeniya 20400, Sri Lanka
{e17369,roshanr}@eng.pdn.ac.lk
[2]Mohamed bin Zayed University of AI, Abu Dhabi, UAE
{sathira.silva,muhammad.haris}@mbzuai.ac.ae
[3]University of Maryland, College Park, MD 20742, USA
gihan@cs.umd.edu

Abstract

Holistic understanding and reasoning in 3D scenes are crucial for the success of autonomous driving systems. The evolution of 3D semantic occupancy prediction as a pretraining task for autonomous driving and robotic applications captures finer 3D details compared to traditional 3D detection methods. Vision-based 3D semantic occupancy prediction is increasingly overlooked in favor of LiDAR-based approaches, which have shown superior performance in recent years. However, we present compelling evidence that there is still potential for enhancing vision-based methods. Existing approaches predominantly focus on spatial cues such as tri-perspective view (TPV) embeddings, often overlooking temporal cues. This study introduces S2TPVFormer, a spatiotemporal transformer architecture designed to predict temporally coherent 3D semantic occupancy. By introducing temporal cues through a novel Temporal Cross-View Hybrid Attention mechanism (TCVHA), we generate Spatiotemporal TPV (S2TPV) embeddings that enhance the prior process. Experimental evaluations on the nuScenes dataset demonstrate a significant $+4.1\%$ of absolute gain in mean Intersection over Union (mIoU) for 3D semantic occupancy compared to baseline TPVFormer, validating the effectiveness of S2TPVFormer in advancing 3D scene perception.

Introduction

Accurate and comprehensive 3D scene understanding and reasoning are crucial for the advancement of robotic and autonomous driving systems (Li et al. 2022b; Huang et al. 2023; Wei et al. 2023; Cao and de Charette 2022). This reasoning encompasses two essential dimensions: spatial reasoning and temporal reasoning. Vision-based approaches to 3D perception (Lang et al. 2019; Zhu et al. 2021; Roldão, de Charette, and Verroust-Blondet 2020; Shi, Wang, and Li 2018) present distinct advantages over LiDAR-based methods that rely on explicit depth measurements. Notably, vision-centric methods excel in identifying road elements, such as traffic lights and road signs, a task that proves challenging for LiDAR-based approaches.

[*]These authors contributed equally.

Workshop on Machine Learning for Autonomous Driving at AAAI 2025 (ML4AD2025), Philadelphia, PA USA. https://ml4ad.github.io

For an extended period, one of the most prominent 3D perception tasks has been 3D object detection (Simonelli et al. 2019; Wang et al. 2021b; Li et al. 2022b; Huang et al. 2021), which are constrained by the limited expressiveness of their 3D bounding box outputs.

This limitation was overcome by generalizing the expression of one cuboid into a collection of smaller cubes (voxels) that can collectively approximate arbitrary shapes by the introduction of a vision-centric 3D semantic occupancy prediction (SOP) task (Cao and de Charette 2022). 3D SOP aims to capture the intricate details of the surrounding scene, leveraging information derived from surrounding multi-camera images captured from different perspective views. TPVFormer (Huang et al. 2023) introduces a Tri-Perspective View (TPV) representation and Cross-View Hybrid Attention (CVHA) as a self-attention mechanism over the three planes, for compute-efficient 3D semantic occupancy prediction.

Previous works (Li et al. 2022b; Huang and Huang 2022; Li et al. 2021) have emphasized the importance of temporal fusion in 3D object detection. However, earlier approaches to 3D SOP (Huang et al. 2023; Wei et al. 2023; Tian et al. 2023; Zhang, Zhu, and Du 2023) frequently overlooked the benefits of leveraging temporal information. This is evidenced by TPVFormer relying solely on the spatially fused features of the current scene for semantic predictions. Spatial fusion is the process of fusing 2D-to-3D lifted features from multi-camera views into a unified spatial representation. Building on this foundation, we propose using Cross-View Hybrid Attention (CVHA) to exchange spatiotemporal information across tri-perspective views. This exchange can be achieved through temporal feature fusion using one of the following approaches:

(1) Fusion of historical data only through the Bird's Eye View (BEV) plane.

(2) Fusion of historical data through all tri-perspective views.

Approach (1) has been explored in the literature using BEV warping (Li et al. 2022b; Huang and Huang 2022; Sima et al. 2023; Zhang et al. 2022). To consider the possibility of implementing approach (2), several important details must be taken into account. The pitch and roll of the

ego vehicle are often ignored due to their insignificance. The more significant yaw axis is aligned parallel to the Front and Side planes in the TPV representation. Changes in the yaw angle cause shifts in the position of these planes resulting in occupying different slices of the ego-space at different timestamps. Therefore, warping is feasible only on the BEV plane, complicating the implementation of approach (2). Additionally, BEV warping can lead to information loss. UniFusion (Qin et al. 2022), a spatiotemporal transformer method for map segmentation, addresses this issue by introducing *virtual views* for parallel and adaptive spatiotemporal fusion across all camera views and time steps.

To bridge the gaps identified in the 3D SOP literature, we propose **S2TPVFormer**; a unified spatiotemporal TPV encoder. We adopt TPV as the latent ego-space representation, harnessing the strengths of BEV and Voxel representations while maintaining computational efficiency. Our spatiotemporal transformer encoder produces temporally rich S2TPV embeddings, enabling the prediction of dense and temporally coherent 3D semantic occupancy through a lightweight MLP decoder. For the spatiotemporal fusion of multi-camera views into the TPV representation, we first transform historical camera views to the current time step using Virtual View Transformation (VVT) and then fuse the multi-camera features into the TPV representation for each time step. To facilitate the effective interaction of features across all time steps and TPV planes, we propose *Temporal Cross-View Hybrid Attention (TCVHA)*. This mechanism allows features to interact not only within the same time step but also across different time steps, enhancing spatiotemporal context awareness and resulting in a unified spatiotemporal representation.

A summary of our main contributions is as follows:

- We introduce S2TPVFormer, featuring a novel temporal fusion workflow for TPV representation, and demonstrate how CVHA facilitates the sharing of spatiotemporal information across the three planes.
- S2TPVFormer achieves significant improvements in the 3D SOP task on the nuScenes validation set, with a **+4.1%** mIOU gain over the baseline TPVFormer, highlighting that vision-based 3D SOP still has considerable potential for improvement.

Related Work

Latent 3D Scene Representations: The effectiveness of 3D scene understanding heavily relies on the representation of the 3D environment as illustrated in figure 1. Traditional approaches (Wang et al. 2020; Rukhovich, Vorontsova, and Konushin 2021) involve dividing the 3D space into voxels and assigning each voxel a vector to denote its status, which is computationally expensive. Alternatively, BEV-based methods (Li et al. 2022b; Huang et al. 2021; Li et al. 2022a; Min et al. 2023; Philion and Fidler 2020) perform remarkably well in tasks such as 3D object detection and map segmentation where height information is not significant. Some 3D SOP methods (Sima et al. 2023) use BEV as the latent 3D scene embedding, but have to employ complex decoders to reconstruct the lost height information from BEV. TPVFormer (Huang et al. 2023) introduces a Tri-Perspective View (TPV) representation generalizing the BEV representation by incorporating two additional orthogonal planes.

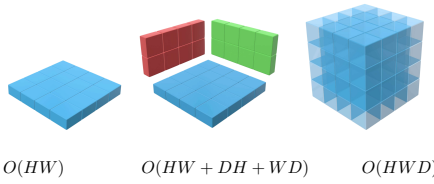


Figure 1: **Comparison of BEV, TPV, and Voxel latent vector fields used to represent 3D scenes.**

2D-3D View Transformation: Transforming 2D perspective observations into 3D space latent embeddings can be considered an ill-posed problem due to the lack of depth information in 2D input images, but can be made feasible by incorporating a strong *geometric prior*. Monocular single-camera approaches address this challenge by predicting explicit depth maps (Li et al. 2023; Philion and Fidler 2020). For example, LSS (Philion and Fidler 2020) "lifts" each perspective view image individually into a frustum of features, then "splats" all frustums into a rasterized BEV grid. In contrast to LSS-based methods (Li et al. 2022a; Huang et al. 2021), *spatial fusion* is an alternative approach (Huang et al. 2023; Li et al. 2022b; Wei et al. 2023; Qin et al. 2022) which uses a spatial query-based transformer approach while leveraging camera parameters as a geometric prior to fuse spatial information from 2D perspective views into a unified latent representation for the ego-space. We adapt *spatial fusion* since LSS-based view transformations tend to generate relatively sparse 3D representations.

Temporal Reasoning: Temporal reasoning holds equal importance to spatial reasoning in a cognitive perception system for identifying occluded objects and determining the motion state of entities. Spatial fusion provides a basis for temporal fusion. BEVFormer(Li et al. 2022b) recurrently fuses BEV features where the history features are warped to align with the ego-space of the current frame. The problem here is that since the warping occurs in ego BEV space which is pre-defined with bounded limits, some warped points are mapped outside the bounds of the original ego BEV space, leading to information loss. UniFusion (Qin et al. 2022) employs vanilla attention for attending to spatially mapped BEV features across all camera views and all time steps. In this work we adapt this method of parallel spatiotemporal fusion.

Vision-centric 3D Semantic Occupancy Prediction: The objective of *3D SOP* is to intricately reconstruct the 3D environment surrounding an entity by incorporating detailed geometric information and semantic understanding. In the context of autonomous driving, 3D SOP serves as the academic alternative to occupancy networks (Mescheder et al. 2019).

MonoScene (Cao and de Charette 2022) is a pioneering work in vision-based 3D SOP, specifically focusing

on Semantic Scene Completion (SSC). It introduces the first single-camera framework for SSC, enabling the reconstruction of outdoor scenes using RGB inputs alone. Building upon the foundation of MonoScene, TPVFormer (Huang et al. 2023), the first multi-camera method for 3D SOP, introduces a tri-perspective view representation with a transformer-based TPV encoder. A more recent line of research (Tian et al. 2023; Sima et al. 2023; Wei et al. 2023) suggests that dense semantic occupancy predictions require dense labels and proposes pipelines for generating densified ground-truth voxel semantics. With our method, we demonstrate that leveraging temporal information provides an effective alternative to densifying supervision for achieving accurate SOP.

Methodology

Overall Architecture

Here we discuss our S2TPVFormer pipeline, which consists of four major modules, as illustrated in figure 2. The 2D image backbone is detailed in the following section, and the spatiotemporal 2D-3D encoder is covered in the section after that. The remaining modules include a simple feature aggregator to generate voxel semantic occupancy features and a lightweight MLP head for predicting the semantic labels of individual voxels (Huang et al. 2023).

Image Backbone

The image backbone consists of two networks; a feature extractor network and a neck module, which extracts multi-scale features for enhanced granularity. The image backbone network extracts multi-scale features from all the input surrounding multi-camera images simultaneously at a given timestep, providing the foundation for the S2TPV encoder. We employ a ResNet (He et al. 2016) as the image feature extractor and an FPN (Lin et al. 2017) to produce multi-scale features. Given the N_{cam} surround multi-camera images $\mathbf{I}_t$ at time step t, the image backbone is used to extract multi-level 2D perspective view features for each camera view. We denote these as $\mathbf{F}_t = \{\{F_t^{ij}\}_{j=1}^{N_{scale}}\}_{i=1}^{N_{cam}}$. Since the proposed pipeline is not limited to a specific image backbone, it can be replaced with any other feature extractor network such as ViT (Dosovitskiy et al. 2021), or SwinTransformer (Liu et al. 2021) along with any FPN variant such as BiFPN (Tan, Pang, and Le 2020) or NAS-FPN (Ghiasi et al. 2019).

S2TPV Encoder

During inference, S2TPVFormer caches $\mathbf{F}_t$ in a queue for each time step. These history feature maps, along with the current feature maps $\{\mathbf{F}_{t-k}^{ij}\}_{k=0}^{M}$, where M is the total number of temporal fusion steps, are fed to the Unified Spatiotemporal Fusion module to fuse features across all camera views and time steps onto the S2TPV queries. Essentially, this module does the following, (1) Virtual View Transformation (VVT) to view camera features as if they were present in the current time step, followed by Spatial Cross Attention (SCA) to fuse virtual camera view features onto S2TPV queries for each time step, and (2) Fuse the virtual spatial TPV features across all time steps via TCVHA. The Temporal Cross-View Hybrid Attention (TCVHA), that we introduce extending CVHA, is realized via concatenating previous S2TPV features with current spatial TPV features as shown below the TCVHA module in figure 2. A separate CVHA module is used to self-attend to S2TPV features to refine the queries and produce temporally coherent semantic occupancy embeddings. The S2TPV occupancy embeddings are finally aggregated and fed through a lightweight MLP head.

Unified Spatiotemporal Fusion: Since the spatial and temporal fusion in S2TPVFormer is parallel, history frames across the camera views has to be aligned with the current ego space. Given a past time frame, we use the VVT, as expressed in equations (1) and (2). In these equations, $R_i^{v,p}$ and $t_i^{v,p}$ represent the rotation and translation of the virtual view transformation for the p^{th} time step. R_i and t_i denote the rotation and translation from the camera sensor to ego-space for the i^{th} camera. R_c and t_c are the transformations from ego-space to global coordinates for the current time step, while R_p and t_p correspond to the ego-space to global coordinate transformation for the past time step. Together, $R_i^{v,p}$ and $t_i^{v,p}$ transform an ego-space point from a past time step to a virtual point in the current time step, as viewed from the perspective of the i^{th} camera sensor.

$$R_i^{v,p} = R_i^{-1} R_p^{-1} R_c \tag{1}$$

$$t_i^{v,p} = R_i^{-1} R_p^{-1} t_c - R_i^{-1} R_p^{-1} t_p - R_i^{-1} t_i \tag{2}$$

We implement spatial fusion using 3D Deformable Attention (Li et al. 2022b) to reduce the computational burden of using vanilla attention. After the VVT, the virtual views are passed through the Spatial Cross-Attention (SCA) module to project them into the current ego TPV space. Taking advantage of the deformable attention mechanism, we implement VVT by employing the reference points defined in equation (4). For a S2TPV query $q_{h,w} \in Q^{HW}$ located at (h, w), we uniformly sample N_{ref}^{HW} reference points along the orthogonal direction of the plane as described in equation (3) (Lang et al. 2019). These points are then transformed using the VVT and camera intrinsic parameters, which provide the geometry prior for the 2D-3D lifting, resulting in the final virtual image reference points. For the attention-based fusion, only the views where the projected reference point, $\mathbf{Ref}_{h,w}^{v,i}$, falls within the image bounds are considered. The SCA function that performs spatial fusion is described in equation (5).

$$\mathbf{Ref}_{h,w}^{ego} = \{(h, w, d_k)\}_{k=1}^{N_{ref}^{HW}} \tag{3}$$

$$\mathbf{Ref}_{h,w}^{v,i} = K_i R_i^{v,p} \mathbf{Ref}_{h,w}^{ego} + t_i^{v,p} \tag{4}$$

$$\text{SCA}(q_{h,w}, \mathbf{F}_t) = \frac{1}{|V_{hit}|} \sum_{i \in V_{hit}} \text{3DDeformAttn}(q_{h,w}, \mathbf{Ref}_{h,w}^{v,i}, \mathbf{F}_t^i) \tag{5}$$

In equations (3) and (4), $\mathbf{Ref}_{h,w}^{ego}$ represents the reference points generated in the ego-TPV space for each TPV plane,

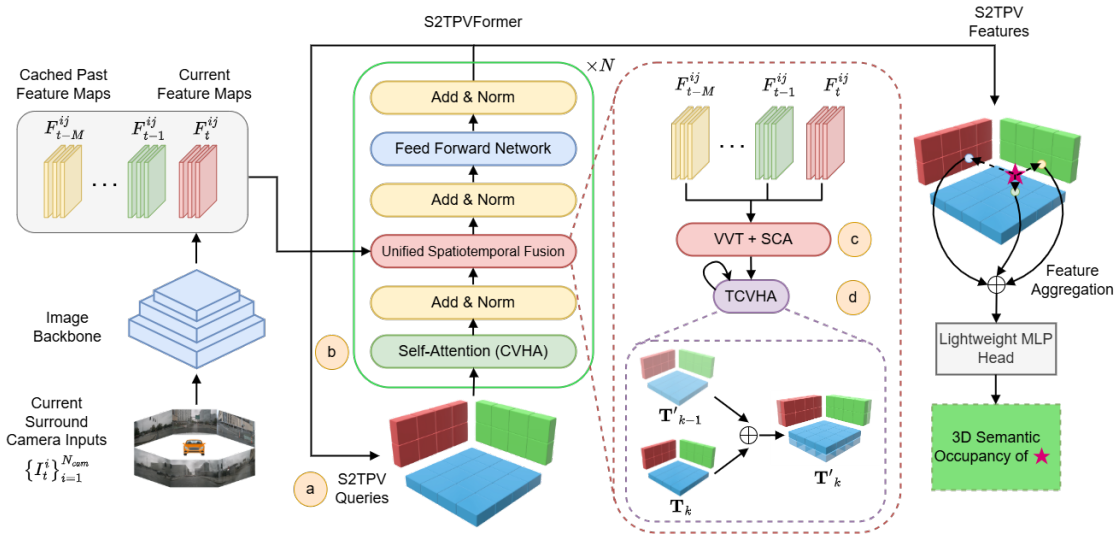


Figure 2: **The 3D SOP pipeline for the proposed S2TPVFormer architecture.** The S2TPVFormer encoder layers consist of four main components: (a) Three learnable grid-shaped parameters to learn spatiotemporal queries, (b) Self-Attention module, (c) Spatial Fusion (VVT + SCA) Module, and (d) Temporal Cross-View Hybrid Attention (TCVHA) Module. Both (c) and (d) are encapsulated as the *Unified Spatiotemporal Fusion* Module in the block diagram.

$\mathbf{Ref}^{v,i}_{h,w}$ denotes the i^{th} virtual camera view reference points for 3D deformable attention, and K_i is the camera intrinsic matrix of the i^{th} camera sensor. In equation (5), V_{hit} denotes the set of hit camera views. Note that the above formulations consider only the Q^{HW} query plane. The computations for the other two planes will follow the same approach. After spatial cross-attention, the resulting features are;

$$\mathbf{T}_t = T_t^{HW} \cup T_t^{DH} \cup T_t^{WD} \tag{6}$$

Temporal Cross-View Hybrid Attention: Realizing the capability of CVHA in self-attending to S2TPV representation, we construct TCVHA, essentially for the queries to interact with history features. As depicted in figure 2, for a given BEV query feature q at a point $p = (h, w)$, it interacts with four types of feature points: (1) history points (temporal fusion), (2) self points, (3) front viewpoints, and (4) side viewpoints. Note that this diagram only illustrates interactions with BEV queries and does not show interactions with previous front and side view features. Given the spatially fused TPV features $\mathbf{T}_t$ for all the time steps in the history queue, the queries for the TCVHA are created iteratively, as described in equations (7), (8), and (9). Here, $q'_{k,h,w} \in \mathbf{Q}'_k$ represents the queries for the TCVHA operation at the k^{th} iteration, and $\{\cdot\}$ denotes the concatenation operation. For the first iteration, the spatially fused TPV features from the last temporal fusion step, $\mathbf{T}_{t-M}$, are concatenated with themselves. The cross-view reference points, $\mathbf{Ref}^{cross}_{h,w}$, are generated in the same way as in (Huang et al. 2023). Using these intermediate features as queries, TCVHA computes the temporally fused intermediate S2TPV features at k^{th} iteration as expressed in equation (9). This is recursively repeated for M number of temporal fusion steps until we get the final unified spatiotemporal features $\mathbf{T}'_t$. Figure 2 illustrates the recurrent-style temporal fusion of TCVHA, shown below the TCVHA block.

$$\mathbf{Q}'_k = \{T'^{HW}_{k-1}, T^{HW}_k\} \cup \{T'^{DH}_{k-1}, T^{DH}_k\} \cup \{T'^{WD}_{k-1}, T^{WD}_k\} \tag{7}$$

$$\text{TCVHA}(q'_{k,h,w}) = \text{DeformAttn}(q'_{k,h,w}, \mathbf{Ref}^{cross}_{h,w}, \mathbf{T'}_k) \tag{8}$$

$$\mathbf{T'}_k = \text{TCVHA}(q'_{k,h,w}) \tag{9}$$

Experimental Setup and Implementation

Datasets and Evaluation Metrics

We use nuScenes (Fong et al. 2021), which is a large-scale dataset for autonomous driving research, providing 1000 urban driving scenes with annotations for object detection. The dataset contains 28,130 training and 6,018 validation keyframes, captured at 20Hz. It incorporates data from monocular cameras, LiDAR, RADAR, and GPS, with 1.1 billion LiDAR points manually annotated for 32 classes.

For 3D perception tasks such as *LiDAR Segmentation* and *3D SOP*, mean Intersection over Union (mIoU) stands as a significant validation metric to measure model accuracy. A higher IoU signifies better overlap between predicted and ground truth bounding boxes or segmented regions, offering a precise assessment of localization accuracy. Complementing this, mIoU calculates the average IoU across multiple classes, providing an overall performance indicator.

Training for 3D Semantic Occupancy Prediction and LiDAR Segmentation

Ground Truth Labels for Supervision: Models for both 3D SOP and LiDAR segmentation are trained with supervision from the training set of the nuScenes dataset. The

Model Config	Embedding Dimensionality	Backbone
S2TPVFormer (base)	256	ResNet101
S2TPVFormer (small)	128	ResNet50

Table 1: **Model configurations used to run experiments**

nuScenes dataset comprises images, sparse LiDAR points from sweeps across the scene, and annotations of 16 semantic classes for each point. To train our models we divide the 3D space into a voxel grid and assign the semantic label of the LiDAR points that fall into a voxel as the semantic label of the voxel itself. A new semantic class, 'empty', is assigned to all voxels without any LiDAR points falling onto them for training the 3D SOP model. This approach is in line with the standard practices proposed in various studies (Huang et al. 2023; Wei et al. 2023; Zhang, Zhu, and Du 2023; Tian et al. 2023). It is important to mention that we do not generate super-resolution voxel semantic labels, as has been explored in some related work (Tian et al. 2023; Wei et al. 2023).

Loss Functions & Training: We use two loss functions during training: (a) a Cross-entropy loss to improve voxel classification accuracy and (b) a Lovasz-softmax loss (Berman, Triki, and Blaschko 2018) to maximize the IoU score across classes. When training for **3D SOP**, we supervise voxel predictions with the Lovasz loss and LiDAR point predictions with Cross-Entropy loss. Conversely, when training for **LiDAR Segmentation**, the supervision is reversed, as suggested by the ablation study results of TPVFormer (Huang et al. 2023). For both tasks we apply an equal-weighted summation to get the total loss. This approach helps the latent representation learn the discretization strategy inherent to the voxel space. We have also used several data augmentation techniques, including image scaling, color distortion, and Gridmask (Chen et al. 2020).

Implementation Details

To highlight the encoder's efficiency, we use a lightweight MLP decoder composed of two linear layers with a Softplus activation layer in between. For different configurations in table 1, S2TPVFormer (base) employs a ResNet101-DCN (Dai et al. 2017) initialized from an FCOS3D (Wang et al. 2021a) checkpoint, while S2TPVFormer (small) employs a ResNet50 (He et al. 2016) pre-trained on the ImageNet dataset (Deng et al. 2009). For both configurations, we set the input image resolution to 1600x900, TPV resolution to 100x100x8, and the number of transformer encoder layers to $N = 3$ for all experiments, unless stated otherwise.

Results and Analysis

Analysis of 3D Semantic Occupancy Prediction Results

Quantitative Analysis: The experimental results demonstrate that S2TPVFormer outperforms the TPVFormer baseline in 3D Semantic Occupancy Prediction (SOP). As shown in table 2, we achieve a 4.1% improvement over TPVFormer for SOP. This highlights the contribution of our temporal attention mechanism. It is also noteworthy that the IoU increases for fourteen out of the sixteen classes, demonstrating the robustness of the proposed methodology.

Qualitative Analysis: Figure 3 demonstrates the model's capability to predict 3D semantic occupancy around the ego vehicle. This figure presents six input camera images fed into the model, alongside eight representations of the semantic occupancy predictions made by the model for the same frame from the nuScene validation set. Through a comparative study with TPVFormer, we highlight the enhancements achieved through our novel temporal attention module.

Our analysis particularly focuses on two critical objects identified in the camera images for this frame; (a) a truck passing closely by the ego vehicle on its left, highlighted with a blue circle. TPVFormer misclassifies this truck as a car. We believe this is due to the truck's proximity to the ego vehicle, which causes only the top half of the truck to be visible in the camera image, making it resemble a car. Conversely, S2TPVFormer accurately identifies it as a truck. We argue that this accuracy stems from the model's ability to integrate information from preceding frames, where the truck is captured in full from a distance. This allows the temporal fusion capability of S2TPVFormer to effectively utilize past frame data for accurate prediction. (b) A construction vehicle (more specifically a craine), highlighted with a red circle, visible in the distance in the front-left camera image. We contend that the model's access to temporally enriched image features enables S2TPVFormer to identify distant objects such as this.

Another notable observation from our analysis is that the predictions generated by S2TPVFormer are significantly denser than those of TPVFormer, even though both models are trained on the same sparse ground truth from the nuScenes dataset.

Analysis of LiDAR Segmentation Results

We test the performance of S2TPVFormer (base) for LiDAR segmentation to assess the generalization capabilities of our model, with a particular focus on the novel temporal attention module. We report the results of LiDAR segmentation on the nuScenes test and validation sets in tables 3 and 4, respectively. In table 3, we present results for some of the best-performing methods in the nuScenes LiDAR segmentation challenge, including models that use both cameras and LiDAR as input modalities. Our model achieves promising results that are comparable with state-of-the-art methods in the literature.

Ablation Study

We present two main ablation studies to investigate: (a) the range of temporal attention during inference, and (b) the dimensionality of the S2TPV embedding, in the context of 3D SOP.

Range of Temporal Attention: As discussed in section , the training of our S2TPVFormer model is conducted using a single previous time frame for temporal attention. This

Method	mIoU (%)	barrier	bicycle	bus	car	const. veh.	motorcycle	pedestrian	traffic cone	trailer	truck	drive. surf.	other flat	sidewalk	terrain	manmade	vegetation
TPVFormer	52.0	59.6	**26.3**	77.6	74.1	30.9	47.5	**41.8**	20.2	44.9	67.8	86.3	54.5	55.5	54.6	47.5	44.0
S2TPVFormer (Base)	**56.1**	**60.1**	16.5	**85.9**	**74.3**	**42.2**	**51.5**	37.0	**21.2**	**49.4**	**74.2**	**86.4**	**56.3**	**57.9**	**55.0**	**65.4**	**65.0**
S2TPVFormer (Small)	43.4	54.3	17.2	66.0	69.5	28.2	22.8	32.1	15.1	31.7	59.6	82.4	49.9	47.8	47.4	34.9	36.0

Table 2: **3D Semantic Occupancy Prediction results on the nuScenes validation set.** It is fair to compare the results of TPVFormer and S2TPVFormer (Base) as our Base configuration is the same as the configuration TPVFormer has used for 3D SOP.

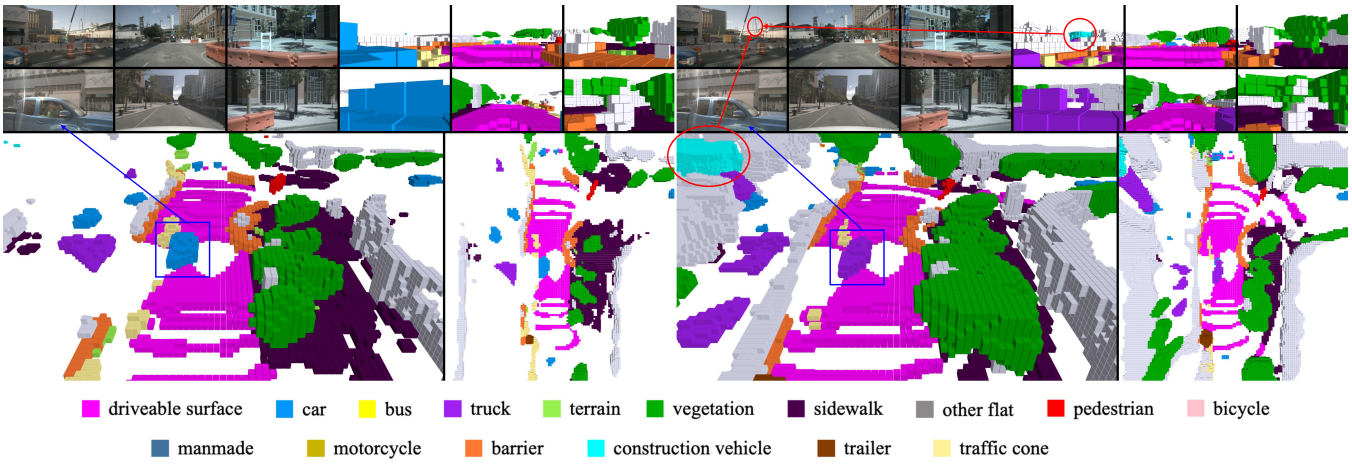


Figure 3: **Qualitative results on nuScenes validation set.** TPVFormer's (Huang et al. 2023) predictions are visualized on the left side, and S2TPVFormer's predictions are on the right side.

Method	Input Modality	mIoU (%)	barrier	bicycle	bus	car	const. veh.	motorcycle	pedestrian	traffic cone	trailer	truck	drive. surf.	other flat	sidewalk	terrain	manmade	vegetation
MINet	LiDAR	56.3	54.6	8.2	62.1	76.6	23.0	58.7	37.6	34.9	61.5	46.9	93.3	56.4	63.8	64.8	79.3	78.3
LidarMultiNet	LiDAR	81.4	80.4	48.4	94.3	90.0	71.5	87.2	85.2	80.4	86.9	74.8	97.8	67.3	80.7	76.5	92.1	89.6
UniVision	LiDAR	72.3	72.1	34.0	85.5	89.5	59.3	75.5	69.3	65.8	84.2	71.4	96.1	67.4	71.9	65	77.9	71.7
PanoOcc	LiDAR	71.4	82.5	32.3	88.1	83.7	46.1	76.5	67.6	53.6	82.9	69.5	96.0	66.3	72.3	66.3	80.5	77.3
OccFormer	LiDAR	70.8	72.8	29.9	87.9	85.6	57.1	74.9	63.2	53.5	83	67.6	94.8	61.9	70.0	66.0	84.0	80.5
TPVFormer-Small†	Camera	59.2	65.6	15.7	75.1	80.0	45.8	43.1	44.3	26.8	72.8	55.9	92.3	53.7	61.0	59.2	79.7	75.6
TPVFormer-Base†	Camera	69.4	74.0	27.5	86.3	85.5	60.7	68.0	62.1	49.1	81.9	68.4	94.1	59.5	66.5	63.5	83.8	79.9
S2TPVFormer (Base)	Camera	60.4	61.2	18.2	80.6	78.1	55.2	57.6	41.5	26.4	76.1	61.3	89.8	49.4	56.6	58.0	79.3	76.4

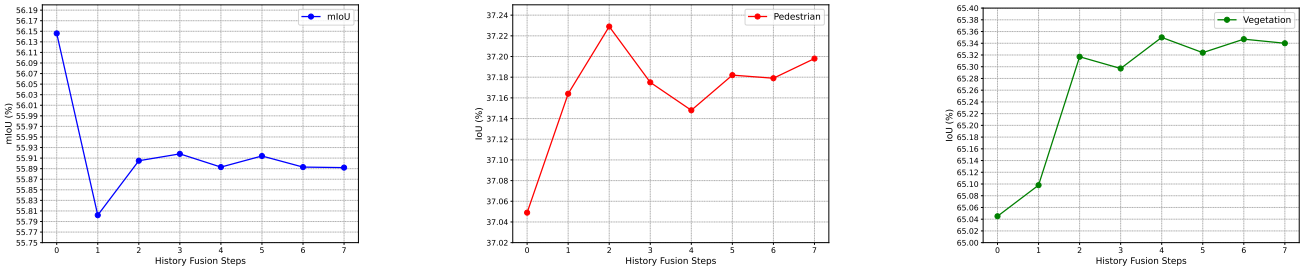
Table 3: **LiDAR Segmentation performance on the nuScenes test set.** † represents that TPVFormer-Small and TPVFormer-Base are different from S2TPVFormer (small) and S2TPVFormer (base)

study aims to examine the variation in performance of the model for 3D SOP as a function of varying extents of temporal attention. It is important to note that we change the history fusion steps only for inference.

As depicted in figure 4, we present an analysis where the number of temporal history fusion steps is varied across eight different values, examining their impact on the IoU across two semantic classes as well as on the mean IoU. It is observed that the optimal number of history fusion steps necessary to achieve the most favorable outcomes differs

Method	Input Modality	mIoU	barrier	bicycle	bus	car	const. veh.	motorcycle	pedestrian	traffic cone	trailer	truck	drive. surf.	other flat	sidewalk	terrain	manmade	vegetation
BEVFormer	Camera	56.2	54.0	22.8	76.7	74.0	45.8	53.1	44.5	24.7	54.7	65.5	88.5	58.1	50.5	52.8	71.0	63.0
TPVFormer-Base†	Camera	68.9	70.0	40.9	93.7	85.6	49.8	68.4	59.7	38.2	65.3	83.0	93.3	64.4	64.3	64.5	81.6	79.3
TPVFormer-Small†	Camera	59.3	64.9	27.0	83.0	82.8	38.3	27.4	44.9	24.0	55.4	73.6	91.7	60.7	59.8	61.1	78.2	76.5
S2TPVFormer (base)	Camera	61.6	62.9	25.5	87.4	81.3	51.6	64.2	45.7	22.0	57.4	77.5	89.3	50.4	56.5	58.9	78.7	76.4

Table 4: **LiDAR Segmentation results on nuScenes validation set.** * represents the produced upon completion of training over four epochs. † represents that TPVFormer-Small and TPVFormer-Base are different from S2TPVFormer (small) and S2TPVFormer (base)



(a) Mean IoU (b) Pedestrian (c) Vegetation

Figure 4: Potential of long-range temporal fusion.

among the semantic classes. This observation underscores the inherent potential for improving temporal fusion within our model, although it remains underexploited at the current juncture.

Ablation	mIoU (%)
TPVFormer-Small*	**44.4**
S2TPVFormer (Small)	43.4
TPVFormer	52.0
S2TPVFormer (Base)	**55.0**

Table 5: **Summary of the ablation study on embedding dimensionality.** * represents the reproduced results using our implementation of the TPVFormer's architecture.

S2TPV Embedding Dimensionality: For this study, we train S2TPVFormer and TPVFormer using the S2TPVFormer (small) configuration outlined in table 1. From the mIoU scores in table 5, we draw two important observations: **(a)** the mIoU scores of both TPVFormer and S2TPVFormer increase with the enhancement of embedding dimensionality, and **(b)** TPVFormer attains a higher mIoU than S2TPVFormer in the small configuration, even though the opposite is true for the base configuration. These observations lead us to conclude that **(a)** a higher embedding dimensionality is required to facilitate the TPV representation to learn and retain the additional information it receives via temporal attention, and **(b)** our model reveals promising scalability compared to TPVFormer.

Conclusion

Overshadowed by the increased performance of its LiDAR-based counterpart, the task of vision-based 3D Semantic Occupancy Prediction (3D SOP) has gradually lost traction within the academic community in the recent years. However, the vision-based approach still holds untapped potential for improvement. In this paper, we show one such improvement by introducing the novel approach of leveraging spatiotemporal information in the TPV representation to enhance the temporal coherence of 3D SOP. Our method specifically utilizes temporal attention to enhance the model's ability to comprehend and predict the 3D scene over time.

As the first to incorporate this method into the TPV representation, we demonstrate significant improvements in the accuracy of vision-based 3D SOP, reiterating its relevance despite the prominence of LiDAR methods. Our results show that incorporating temporal information can bridge some performance gaps between vision-based systems and their LiDAR counterparts. However, the full potential of long-range temporal information in these domains remains untapped. Future research should focus on further exploring our methodology, possibly focusing on the integration of dense semantic labels, to explore the complete capability of temporal attention in improving 3D scene understanding.

Acknowledgements

We sincerely thank Honglu Zhou for providing Lambda credits, which enabled the execution of our experiments.

References

Berman, M.; Triki, A. R.; and Blaschko, M. B. 2018. The lovász-softmax loss: A tractable surrogate for the optimization of the intersection-over-union measure in neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 4413–4421.

Cao, A.-Q.; and de Charette, R. 2022. MonoScene: Monocular 3D Semantic Scene Completion. In *CVPR*.

Chen, P.; Liu, S.; Zhao, H.; Wang, X.; and Jia, J. 2020. GridMask Data Augmentation.

Dai, J.; Qi, H.; Xiong, Y.; Li, Y.; Zhang, G.; Hu, H.; and Wei, Y. 2017. Deformable Convolutional Networks. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*.

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, 248–255. Ieee.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; Uszkoreit, J.; and Houlsby, N. 2021. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. arXiv:2010.11929.

Fong, W. K.; Mohan, R.; Hurtado, J. V.; Zhou, L.; Caesar, H.; Beijbom, O.; and Valada, A. 2021. Panoptic nuScenes: A Large-Scale Benchmark for LiDAR Panoptic Segmentation and Tracking. *arXiv preprint arXiv:2109.03805*.

Ghiasi, G.; Lin, T.-Y.; Pang, R.; and Le, Q. V. 2019. NAS-FPN: Learning Scalable Feature Pyramid Architecture for Object Detection. arXiv:1904.07392.

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.

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.

Huang, Y.; Zheng, W.; Zhang, Y.; Zhou, J.; and Lu, J. 2023. Tri-Perspective View for Vision-Based 3D Semantic Occupancy Prediction.

Lang, A. H.; Vora, S.; Caesar, H.; Zhou, L.; Yang, J.; and Beijbom, O. 2019. PointPillars: Fast Encoders for Object Detection from Point Clouds.

Li, Q.; Wang, Y.; Wang, Y.; and Zhao, H. 2021. HDMapNet: An Online HD Map Construction and Evaluation Framework.

Li, Y.; Ge, Z.; Yu, G.; Yang, J.; Wang, Z.; Shi, Y.; Sun, J.; and Li, Z. 2022a. BEVDepth: Acquisition of reliable depth for multi-view 3D object detection.

Li, Y.; Yu, Z.; Choy, C.; Xiao, C.; Alvarez, J. M.; Fidler, S.; Feng, C.; and Anandkumar, A. 2023. VoxFormer: Sparse Voxel Transformer for Camera-based 3D Semantic Scene Completion.

Li, Z.; Wang, W.; Li, H.; Xie, E.; Sima, C.; Lu, T.; Yu, Q.; and Dai, J. 2022b. BEVFormer: Learning bird's-eye-view representation from multi-camera images via spatiotemporal transformers.

Lin, T.-Y.; Dollár, P.; Girshick, R.; He, K.; Hariharan, B.; and Belongie, S. 2017. Feature Pyramid Networks for Object Detection. arXiv:1612.03144.

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.

Mescheder, L.; Oechsle, M.; Niemeyer, M.; Nowozin, S.; and Geiger, A. 2019. Occupancy Networks: Learning 3D Reconstruction in Function Space.

Min, C.; Xu, X.; Li, F.; Si, S.; Xue, H.; Jiang, W.; Zhang, Z.; Li, J.; Zhao, D.; Xiao, L.; Xu, J.; Nie, Y.; and Dai, B. 2023. Occ-BEV: Multi-Camera Unified Pre-training via 3D Scene Reconstruction.

Philion, J.; and Fidler, S. 2020. Lift, splat, shoot: Encoding images from arbitrary camera rigs by implicitly unprojecting to 3D.

Qin, Z.; Chen, J.; Chen, C.; Chen, X.; and Li, X. 2022. UniFusion: Unified multi-view fusion transformer for spatial-temporal representation in bird's-eye-view.

Roldão, L.; de Charette, R.; and Verroust-Blondet, A. 2020. LMSCNet: Lightweight Multiscale 3D Semantic Completion.

Rukhovich, D.; Vorontsova, A.; and Konushin, A. 2021. ImVoxelNet: Image to Voxels Projection for Monocular and Multi-View General-Purpose 3D Object Detection.

Shi, S.; Wang, X.; and Li, H. 2018. PointRCNN: 3D object proposal generation and detection from point cloud.

Sima, C.; Tong, W.; Wang, T.; Chen, L.; Wu, S.; Deng, H.; Gu, Y.; Lu, L.; Luo, P.; Lin, D.; and Li, H. 2023. Scene as Occupancy.

Simonelli, A.; Bulò, S. R. R.; Porzi, L.; López-Antequera, M.; and Kontschieder, P. 2019. Disentangling monocular 3D object detection.

Tan, M.; Pang, R.; and Le, Q. V. 2020. EfficientDet: Scalable and Efficient Object Detection. arXiv:1911.09070.

Tian, X.; Jiang, T.; Yun, L.; Wang, Y.; Wang, Y.; and Zhao, H. 2023. Occ3D: A Large-Scale 3D Occupancy Prediction Benchmark for Autonomous Driving. *arXiv preprint arXiv:2304.14365*.

Wang, T.; Zhu, X.; Pang, J.; and Lin, D. 2021a. FCOS3D: Fully convolutional one-stage monocular 3D object detection.

Wang, Y.; Chao, W.-L.; Garg, D.; Hariharan, B.; Campbell, M.; and Weinberger, K. Q. 2020. Pseudo-LiDAR from Visual Depth Estimation: Bridging the Gap in 3D Object Detection for Autonomous Driving.

Wang, Y.; Guizilini, V.; Zhang, T.; Wang, Y.; Zhao, H.; and Solomon, J. 2021b. DETR3D: 3D Object Detection from Multi-view Images via 3D-to-2D Queries.

Wei, Y.; Zhao, L.; Zheng, W.; Zhu, Z.; Zhou, J.; and Lu, J. 2023. SurroundOcc: Multi-Camera 3D Occupancy Prediction for Autonomous Driving.

Zhang, Y.; Zhu, Z.; and Du, D. 2023. OccFormer: Dual-path transformer for vision-based 3D semantic occupancy prediction.

Zhang, Y.; Zhu, Z.; Zheng, W.; Huang, J.; Huang, G.; Zhou, J.; and Lu, J. 2022. BEVerse: Unified Perception and Prediction in Birds-Eye-View for Vision-Centric Autonomous Driving.

Zhu, X.; Zhou, H.; Wang, T.; Hong, F.; Li, W.; Ma, Y.; Li, H.; Yang, R.; and Lin, D. 2021. Cylindrical and asymmetrical 3D convolution networks for LiDAR-based perception.