Occupancy Learning with Spatiotemporal Memory

Ziyang Leng[1] Jiawei Yang[2] Wenlong Yi[1] Bolei Zhou[1]
[1]University of California, Los Angeles [2]University of Southern California

Abstract

3D occupancy becomes a promising perception representation for autonomous driving to model the surrounding environment at a fine-grained scale. However, it remains challenging to efficiently aggregate 3D occupancy over time across multiple input frames due to the high processing cost and the uncertainty and dynamics of voxels. To address this issue, we propose ST-Occ, a scene-level occupancy representation learning framework that effectively learns the spatiotemporal feature with temporal consistency. ST-Occ consists of two core designs: a spatiotemporal memory that captures comprehensive historical information and stores it efficiently through a scene-level representation and a memory attention that conditions the current occupancy representation on the spatiotemporal memory with a model of uncertainty and dynamic awareness. Our method significantly enhances the spatiotemporal representation learned for 3D occupancy prediction tasks by exploiting the temporal dependency between multi-frame inputs. Experiments show that our approach outperforms the state-of-the-art methods by a margin of 3 mIoU and reduces the temporal inconsistency by 29%. The code and model are available at https://github.com/matthew-leng/ST-Occ.

1. Introduction

In recent years, vision-centric 3D occupancy representation has gained significant interest in autonomous driving [5, 14, 16, 20, 27, 28, 33]. Closely related to Bird's Eye View (BEV) representations, many prior efforts have sought to extend common BEV perception pipelines and techniques [20]—such as view transformation, decoder designs, and temporal fusion—to obtain high-quality 3D occupancy representations.

Recent works leverage temporal information to improve the robustness of occupancy prediction [16, 21, 27, 28, 32, 33]. To achieve this, historical features are typically stored on a frame-wise basis, aligned with the current frame, and processed in a recurrent [15] or stacked manner [8]. However, with the extended height dimension in 3D occupancy representation, the memory and computation overheads of the temporal fusion process become a critical issue [20]. Moreover, occupancy prediction tasks require voxel-level detail, which demands higher granularity than the BEV representations for 3D detection tasks. Consequently, existing temporal fusion paradigms are often inefficient and insufficient in exploiting spatiotemporal information for 3D occupancy learning.

Much less has been explored on utilizing spatiotemporal information for occupancy representation learning. While high granularity in occupancy representation aids the prediction task [28], the performance gains from temporal information integration remain limited [16]. We attribute this to three main challenges: 1) Efficiency. The voxel-wise detail in occupancy representation makes it large and dense, so storing and processing multiple frames of historical features is resource-intensive, limiting the number of frames that can be used in temporal fusion. 2) Uncertainty. Due to factors like occlusion and varying lighting conditions [11], voxel-level uncertainties arise across frames, potentially accumulating noise and error during temporal fusion and negatively impacting prediction robustness and accuracy [5]. 3) Dynamics. Dynamic instances in the scene introduce voxel shifts, resulting in misaligned historical features if not accurately modeled, which can hinder performance on dynamic instances.

To address the aforementioned problems, we propose constructing a spatiotemporal memory under scene-centered coordinates instead of ego vehicle-centered coordinates, demonstrated in Fig. 1. In this way, we can not only store and process the historical feature efficiently in a recurrent way but also incorporate temporal clues that mitigate the uncertainty and compensate for the dynamics in occupancy representation. Therefore, we introduce a new paradigm: unified temporal modeling, along with ST-Occ, a scene-level Spatiotemporal (ST) Occupancy representation learning framework designed to efficiently exploit the spatiotemporal information for the 3D occupancy prediction task. ST-Occ can perform 3D occupancy prediction in a streaming video approach [24] with a variable number of frames for temporal modeling, which makes it effective for constructing a complete and holistic representation of a large-scale scene with strong temporal consistency and ro-

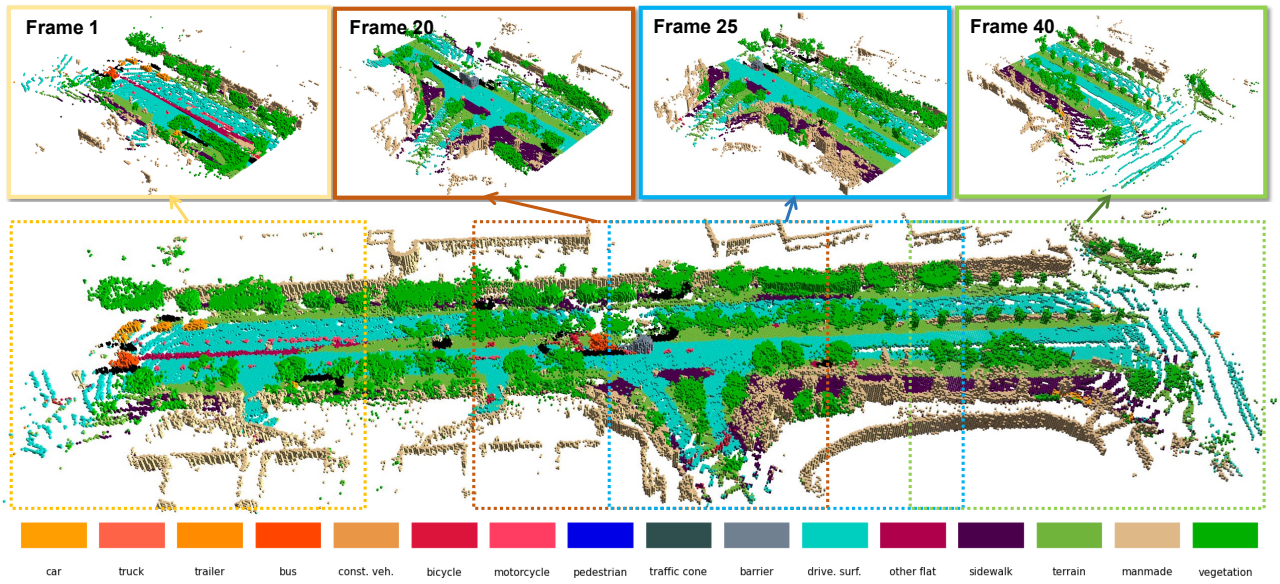


Figure 1. Occupancy prediction for a large-scale scene using our proposed ST-Occ. The first row shows ego-vehicle-centered predictions at different time steps and locations. The second row presents the scene-level occupancy prediction derived from our spatiotemporal memory, aggregating all 40 frames in the scene. The dashed rectangles are colored to correspond with their respective source frames above.

bustness. Figure 1 shows one occupancy prediction result by ST-Occ.

ST-Occ consists of two core modules: a spatiotemporal memory bank and a memory attention. The spatiotemporal memory bank is constructed under scene-centered coordinates, which aim to capture comprehensive historical information, including historical representation, that assists the temporal modeling. The memory attention conditioned the current frame occupancy representation on the historical information from the spatiotemporal memory. This incorporates spatiotemporal information with uncertainty and dynamic awareness, which benefits the occupancy prediction. These two modules together compose the unified temporal modeling for the framework.

Compared with the existing occupancy representation learning methods, the experiment shows the proposed ST-Occ surpasses the state-of-the-art method by a margin of 3 mIoU on the Occ3D benchmark [27] while maintaining computational and memory efficiency. Besides, our unified temporal modeling is 2.8× effective in utilizing temporal information. To further evaluate the temporal consistency of our framework's prediction, we design an evaluation metric to measure the occupancy prediction inconsistency between frames, and our method results in a 29% decrease in temporal inconsistency. We summarize our contributions as follows:

- We design a new unified temporal modeling paradigm to achieve memory and computation-efficient temporal fusion.
- We propose a scene-level occupancy representation learning framework that implements the unified temporal modeling. It exploits the spatiotemporal information with uncertainty and dynamic awareness.
- We conduct extensive experiments on the occupancy prediction task and our proposed temporal consistency evaluation metric, and our method outperforms state-of-the-art methods by substantial margins.

2. Related Work

Camera-based 3D Occupancy Prediction. 3D occupancy prediction aims to predict whether a voxel in the 3D space is occupied or not and its semantic class if occupied [27, 28, 35]. The use of occupancy maps can be traced back to robotics mapping and planning tasks [6, 25]. The occupancy network introduced by Tesla [1] brings the occupancy to autonomous driving perception. Recently, different camera-based 3D occupancy prediction works [4, 16, 27, 28, 30] have been developed by extending vision-centric BEV perception pipelines [9, 15, 17]. For instance, OccNet [28] utilizes a cascade voxel decoder derived from the BEV decoder to reconstruct 3D occupancy. FB-OCC [16] constructs the 3D features via forward-backward transformations used in FB-BEV [17]. Another line of research aims to improve efficiency. For instance, FlashOcc [33] and COTR [21] compress the intermediate representation to BEV or a smaller-scale representation. OctOcc [23] and SparseOcc [20] design intermediate representations with varied levels of granularities to reduce the computational and memory cost, and they refine occupancy representations in a coarse-

to-fine manner. These works contribute to obtaining a fine-grained representation with efficiency.

Orthogonal to prediction precision, PasCo [5] highlights the uncertainty awareness in occupancy prediction, which enables the model deployment in real-world safety-critical scenarios with noisy and ambiguous data. However, efforts to model the uncertainty and dynamics of occupancy from the perspective of temporal modeling remain less explored. If not modeled appropriately, these factors would lead to temporal inconsistency in the occupancy representation, resulting in decreased prediction performance with less robustness. Our work incorporates these factors into the temporal modeling process with efficiency, thus obtaining a fine-grained and robust representation for better prediction.

Temporal Modeling. Vision-centric perception significantly benefits from temporal modeling, which leverages cross-time information to obtain a better representation [8, 15, 19, 24, 29, 31]. The historical information used in the temporal fusion helps the perception in the scenarios of occlusion, distortion, and lighting changes, among many others [11]. Previous efforts mainly focus on temporal modeling for BEV perception. BEVFormer [15] introduces temporal self-attention, which recurrently attends to previous ego-aligned BEV features. BEVDet and its follow-up [8, 9] differ in how to fuse past information: they align and concatenate historical features with present ones. Current occupancy prediction frameworks mostly extend these techniques from 2D BEV features to 3D occupancy features [16, 26–28, 33]. However, the memory and computational costs brought by the dense occupancy representation are non-trivial and limit the scope of fused historical frames. Our work proposes a unified temporal modeling, which is more memory and computation-efficient and effective in utilizing temporal information.

3. Preliminaries

We introduce the preliminaries and limitations of 3D occupancy representation and temporal fusion. We then discuss our unified temporal modeling method to address these limitations at the end.

3D Occupancy Representation. 3D occupancy representation is a world representation that offers a holistic view of the ego vehicle and its surroundings. Denoted as $\mathbf{V} \in \mathbb{R}^{H \times W \times Z \times C}$, an occupancy representation is a discretized 3D volumetric feature volume with a spatial shape of $H \times W \times Z$ and an additional feature dimension C, where each voxel contains scene features at its physical position [10]. With the additional height dimension compared with BEV representations, 3D occupancy representation usually contains more fine-grained geometric and semantic details

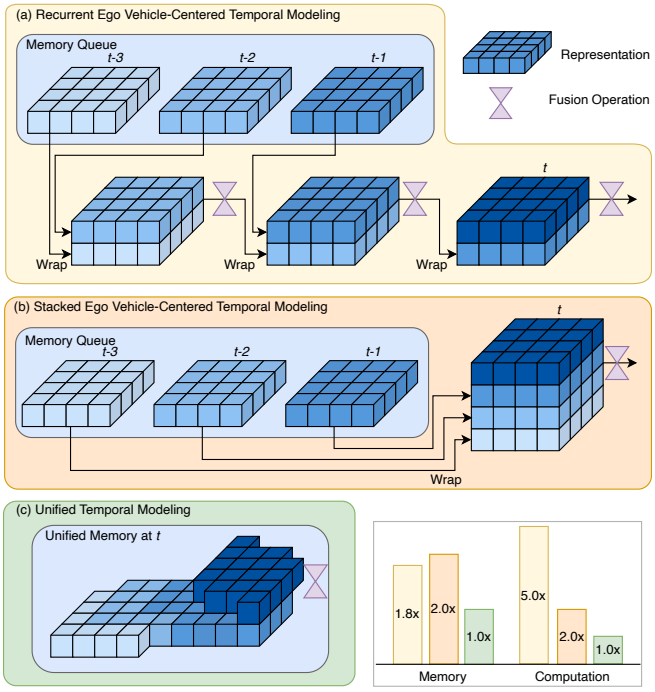


Figure 2. Comparison of different temporal fusion paradigms, including (a) recurrent-based, (b) stacking-based ego vehicle-centered modeling, and (c) our proposed unified temporal modeling. Compared to previous approaches, our method requires significantly less memory and computation (multipliers are scaled relative to our approach).

of the scene and instances [20]. Despite its success, predicting occupancy frame by frame ignores the mutual information across time and is more vulnerable to sensor noise. Incorporating temporal information is thus an important component for robust occupancy learning.

Queue-based Temporal Fusion. Temporal fusion aims to use historical features to facilitate the perception in occluded or uncertain regions and dynamic instances. Given a set of historical features $F = \{F^{t-k}, F^{t-k+1}, \dots, F^{t}\}$ captured at different instantaneous timesteps from $t-k$ to t, we denote the temporal fusion process as

$$\tilde{F}_{\text{out}} = \psi(F), \tag{1}$$

where ψ is the temporal fusion function, k is number of historical frames used, $\tilde{F}_{\text{out}}$ is the output feature that incorporates temporal information.

Due to the ego vehicle motion, pose information T_t is used to align the historical feature through the transformation matrix T_{t-1}^{t}, which is calculated as

$$T_{t-1}^{t} = T_{t}^{inv} \cdot T_{t-1}. \tag{2}$$

Typical temporal fusion paradigms store historical frame features in a queue, align with the current frame, and pro-

cess in a recurrent [15] or stacked manner [8, 13], shown in Fig. 2 (a) and (b), which can be expressed as

$$\tilde{F}^t_{\text{recurrent}} = \psi(F^t, T^t_{t-1}\psi(F^{t-1}, T^{t-1}_{t-2}\psi(F^{t-2}, \dots))), \quad (3)$$

$$\tilde{F}^t_{\text{stack}} = \psi(F^t, T^t_{t-1} \cdot F^{t-1}, T^t_{t-2} \cdot F^{t-2}, \dots). \quad (4)$$

The above approaches are widely adopted and effective in methods involving BEV representation. However, the extended height dimension considerably increases the storage and processing costs regarding occupancy representation.

Unified Temporal Modeling Unified temporal modeling replaces the memory-heavy queue with a unified memory M under scene-centered coordinates in Fig. 2 (c). The ego pose T_t determines the region of interest (RoI) of timestamp t in the unified memory. The process of incorporating temporal information is defined as:

$$\tilde{F}^t_{\text{unified}} = \psi(F^t \mid \chi[M_t, T_t]), \quad (5)$$

where $\chi[\cdot]$ is the feature sampling operation and M_t denotes the unified memory at t timestamp.

The RoI in the unified memory is updated reversely using

$$\chi[M_{t+1}, T_t] = \tilde{F}^t_{\text{unified}}. \quad (6)$$

4. Method

ST-Occ employs a new unified temporal modeling paradigm to exploit the spatiotemporal information in occupancy representation learning with uncertainty and dynamic awareness. We first introduce the pipeline of our proposed framework ST-Occ in (§4.1). We then talk about the two core modules (§4.2) and (§4.3) that realize the unified temporal modeling. Lastly, we describe the temporal consistency evaluation (§4.4) and the loss functions (§4.5).

4.1. ST-Occ

Our method aims to exploit the spatiotemporal information to learn the occupancy representation with strong temporal consistency and performance on occupancy prediction. As depicted in Fig. 3, ST-Occ contains two components for unified temporal modeling: a spatiotemporal memory that preserves the temporal clues of historical input frames and a memory attention that conditions current frame occupancy representation on the spatiotemporal memory to incorporate historical information.

Fig. 3 illustrates our pipeline. Specifically, given multi-view images input I_t captured at timestamp t, the occupancy encoder extracts their ego vehicle-centered occupancy representations $\mathbf{V}_t$. Then, the memory attention conditions $\mathbf{V}_t$ on the historical information $\mathbf{H}_t$ to obtain the fused occupancy representation $\tilde{\mathbf{V}}_t$, where $\mathbf{H}_t$ is extracted from the corresponding region of interest (RoI) in the spatiotemporal memory. Lastly, we update this part of memory using the fused occupancy representation. Next, we will introduce the details of each component.

4.2. Spatiotemporal Memory

We design the spatiotemporal memory to efficiently store comprehensive historical information for temporal modeling of the occupancy representation learning. The spatiotemporal memory is constructed as a representation $\mathbf{M} \in \mathbb{R}^{H_G \times W_G \times Z_G \times C_G}$ at the beginning of each scene sequence. While the spatiotemporal memory representation is slightly larger than the ego vehicle-centered representation $\mathbf{V}$ due to ego motion, it is much more memory efficient when multiple temporal frames are used for temporal fusion.

When a total of k temporal frames are used in the temporal modeling, the typical paradigms retain k total representations while our unified temporal modeling only requires one representation, thus more memory efficient.

We introduce not only the historical representation $\mathbf{V}$ but also other useful temporal attributes μ to preserve comprehensive information that would facilitate the temporal fusion process to the spatiotemporal memory.

We define the temporal attributes of a voxel at position $\mathbf{p}$ as $\mu_{\mathbf{p}} \equiv \{\mathbf{c}_{\mathbf{p}}, \delta_{\mathbf{p}}, \mathbf{f}_{\mathbf{p}}\}$. For simplicity, we omit the subscript $\mathbf{p}$ from now on. Among these attributes, $\mathbf{c} \in \mathbb{R}^N$ is the historical class activation vector after the softmax operation, where N is the number of classes and $\sum_{i=0}^{N-1} c_i = 1$; $\delta \in \mathbb{R}$ is the average log variance $\mathbf{s}$ over classes; and $\mathbf{f} \in \mathbb{R}^2$ is the occupancy flow vector in a top-down view.

The historical class activation of temporal attributes is updated using the class activation $\mathbf{c}_t$ from occupancy head, with an exponential decay of α follows

$$\chi[\mathbf{M}_{t+1}\langle \mathbf{c} \rangle, T_t] = \text{softmax}\left(\alpha \mathbf{c}_t + (1-\alpha)\chi[\mathbf{M}_t\langle \mathbf{c} \rangle, T_t]\right), \quad (7)$$

where the $\langle \cdot \rangle$ is the extraction operation from spatiotemporal memory. We also add two additional networks in the occupancy head to predict log variance $\mathbf{s}$ and occupancy flow $\mathbf{f}$. They are updated to the spatiotemporal memory by

$$\chi[\mathbf{M}_{t+1}\langle \delta \rangle, T_t] = \delta_t, \quad \chi[\mathbf{M}_{t+1}\langle \mathbf{f} \rangle, T_t] = \mathbf{f}_t. \quad (8)$$

Finally, we update the historical representation using the memory attention conditioning

$$\chi[\mathbf{M}_{t+1}\langle \mathbf{V} \rangle, T_t] = \tilde{\mathbf{V}}_t, \quad (9)$$

which is introduced as follows.

4.3. Memory Attention

We design the memory attention to condition the initial occupancy representation $\mathbf{V}_t$ on the historical information $\mathbf{H}_t$

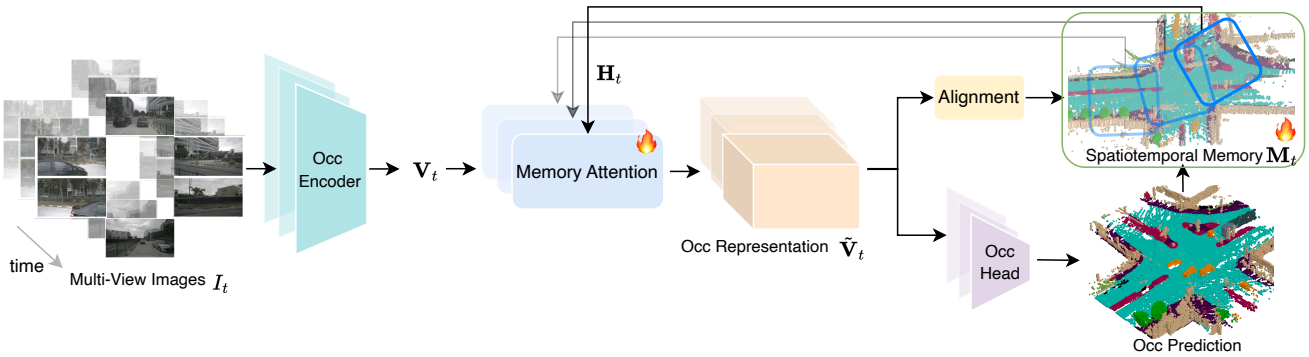


Figure 3. Overview of our ST-Occ. ST-Occ implements the unified temporal modeling using a spatiotemporal memory and a memory attention. The spatiotemporal memory captures comprehensive historical information in a scene-centered coordinate system, and the memory attention conditions the current occupancy representation on the spatiotemporal memory with uncertainty and dynamic awareness.

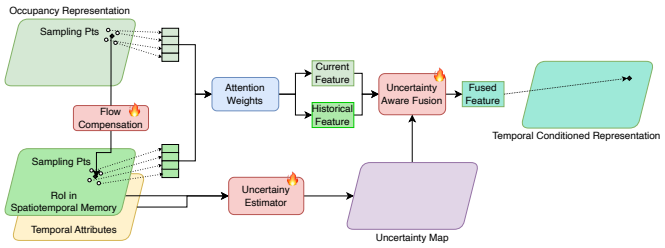


Figure 4. Illustration of memory attention with uncertainty and dynamic awareness.

at timestamp t. The temporal information are incorporated into $\tilde{\mathbf{V}}_t$ follows

$$\tilde{\mathbf{V}}_t = \psi(\mathbf{V}_t \mid \mathbf{H}_t) = \psi(\mathbf{V}_t \mid \chi[\mathbf{M}_t, T_t]), \tag{10}$$

The historical information $\mathbf{H}_t$ consists of historical representation and temporal attributes. They are retrieved from the spatiotemporal memory using grid sampling.

To incorporate uncertainty and dynamic awareness in the memory attention, we use the uncertainty u to balance the fused feature between historical and current representations, and we use the occupancy flow $\mathbf{f}$ to compensate the motion for voxels corresponding to dynamic instances.

To predict the uncertainty u, we use an MLP to encode temporal attributes as

$$u = \text{MLP}(\mathbf{c}, \delta, \varepsilon), \tag{11}$$

where ε is the cosine similarity between current and historical representation at the same physical position.

The temporal conditioning process in our framework is built on the temporal self-attention (TSA) layer [15], with the encoded attributes incorporated to enable uncertainty and dynamic awareness. We use the initial occupancy representation as the *query*, historical representation in RoI of spatiotemporal memory as *key*, and *value*. The process depicted in Fig. 4 uses deformable attention (DA) [36] which can be formulated as

$$(1-u)\,\text{DA}(V_{t_p}, p+f, V_t) + u\,\text{DA}(V_{t_p}, p+f, \chi[\mathbf{M}_t, T_t]), \tag{12}$$

where V_{t_p} denotes the initial occupancy representation V_t located at $p = (x, y, z)$. This design avoids feature misalignment and noise aggregation that causes temporal inconsistency while remaining entirely learnable.

4.4. Measuring Temporal Consistency

To assess the temporal consistency of occupancy predictions across frames in a sequence, we propose a new evaluation metric: mean Spatiotemporal Classification Variability (mSTCV). This metric quantifies the classification variability of voxels representing the same real-world location over time, thereby measuring how stable occupancy predictions are across frames.

To construct the correspondence of voxel between frames, we utilize our spatiotemporal memory to additionally store the historical occupancy prediction results $\mathbf{P}$ and occupancy ground truth $\mathbf{G}$ under scene-centered coordinates and update as follows

$$\chi[\mathbf{M}_{t+1}\langle\mathbf{P}\rangle, T_t] = \mathbf{P}_t, \quad \chi[\mathbf{M}_{t+1}\langle\mathbf{G}\rangle, T_t] = \mathbf{G}_t. \tag{13}$$

The STCV for timestamp t is defined as

$$\frac{\sum \mathbb{1}[(\chi[\mathbf{M}_t\langle\mathbf{P}\rangle, T_t] \neq \mathbf{P}_t) \wedge (\chi[\mathbf{M}_t\langle\mathbf{P}\rangle, T_t] \neq \text{Free})]}{\sum \mathbb{1}[\mathbf{P}_t \neq \text{Free}]}, \tag{14}$$

which calculates the percentage of classification changes in non-free voxels over the total number of non-free voxels. The mSTCV is computed by averaging STCV across all frames:

$$\text{mSTCV} = \frac{1}{T}\sum_{t}^{T} \text{STCV}_t \tag{15}$$

4.5. Loss

Our final loss function comprises three parts. The occupancy prediction loss $\mathcal{L}_{occ}$ follows the formulation in FB-

Method	Backbone	mIoU	others	barrier	bicycle	bus	car	const. veh.	motorcycle	pedestrian	traffic cone	trailer	truck	driv. surf.	other flat	sidewalk	terrain	manmade	vegetation
TPVFormer	Res101	27.83	7.22	38.90	13.67	40.78	45.90	17.23	19.99	18.85	14.30	26.69	34.17	55.65	35.47	37.55	30.70	19.40	16.78
OccFormer		21.93	5.94	30.29	12.32	34.40	39.17	14.44	16.45	17.22	9.27	13.90	26.36	50.99	30.96	34.66	22.73	6.76	6.97
BEVFormer		26.88	5.85	37.83	17.87	40.44	42.43	7.36	23.88	21.81	20.98	22.38	30.70	55.35	28.36	36.00	28.06	20.04	17.69
CTF-Occ		28.53	8.09	39.33	20.56	38.29	42.24	16.93	24.52	22.72	21.05	22.98	31.11	53.33	33.84	37.98	33.23	20.79	18.00
FB-OCC†	Res50	37.39	12.17	44.83	25.73	42.61	47.97	23.16	25.17	25.77	26.72	31.31	34.89	78.83	41.42	49.06	52.22	39.07	34.61
FB-OCC		39.11	13.57	44.74	27.01	45.41	49.10	25.15	26.33	27.86	27.79	32.28	36.75	80.07	42.76	51.18	55.13	42.19	37.53
ST-Occ (ours)		**42.13**	**14.36**	**49.62**	**27.77**	**46.28**	**52.55**	**26.87**	**29.79**	**29.83**	**31.39**	**35.40**	**39.03**	**84.26**	**47.72**	**56.09**	**59.85**	**45.27**	**40.11**
ViewFormer†*	Res50	37.80	9.90	44.89	22.67	42.84	48.90	21.39	24.52	25.22	24.93	29.18	34.56	81.93	44.07	53.72	55.50	42.18	36.29
ViewFormer*		41.44	11.63	50.16	26.49	44.39	**53.36**	22.85	27.80	27.74	29.95	33.04	39.39	84.67	48.08	57.43	59.64	47.57	40.38
ViewFormer† + ST-Occ (ours)		**42.30**	**12.00**	**50.61**	**27.93**	**45.81**	53.24	**24.79**	**29.18**	**28.63**	**30.93**	**33.53**	**39.58**	**85.28**	**49.42**	**58.39**	**60.39**	**48.02**	**41.42**

Table 1. 3D occupancy prediction results on Occ3D benchmark. † without temporal information. * reproduced using its official code. The *direct* baselines of our method are colored by steelblue (Best viewed in color).

OCC as

$$\mathcal{L}_{occ} = \mathcal{L}_{fl} + \mathcal{L}_{ls} + \mathcal{L}^{geo}_{scal} + \mathcal{L}^{sem}_{scal} + \mathcal{L}_{d}, \tag{16}$$

which contains the Focal loss [18], affinity loss following MonoScene [4], Lovasz softmax loss [2], and depth loss following FB-OCC [16]. The log variance prediction uses the Gaussian negative log-likelihood loss $\mathcal{L}_{nll}$ following

$$\mathcal{L}_{nll} = \sum_{i} \frac{1}{2} \left(\exp(-s_i) \|y_i - \hat{y}_i\|^2 + s_i \right). \tag{17}$$

Besides, an L1 loss $\mathcal{L}_{of}$ is used for the occupancy flow prediction. The training loss thus becomes

$$\mathcal{L} = \mathcal{L}_{occ} + \mathcal{L}_{nll} + \mathcal{L}_{of}. \tag{18}$$

5. Experiments

5.1. Dataset and Metrics

We evaluate our method on the 3D occupancy benchmark, Occ3D [27], constructed using the nuScenes dataset [3]. The benchmark consists of 1,000 driving sequences, each lasting 20 seconds, with RGB images captured from six cameras providing a 360-degree view. These sequences are divided into 700 training scenes, 150 validation scenes, and 150 test scenes. For each frame, the dataset includes 3D occupancy annotations within a range of [-40m, -40m, -1m, 40m, 40m, 5.4m] in ego-vehicle coordinates, with a voxel size of 0.4 meters. There are 18 voxel classes in total, one of which represents an unoccupied, free region. Additionally, the Occ3D benchmark provides per-frame visibility masks, indicating whether each voxel is visible in the current camera view, to support both training and evaluation processes.

To assess the performance of our 3D occupancy prediction model, we use the mean Intersection-over-Union (mIoU) metric. Furthermore, we use our defined metric, mSTCV, designed to evaluate the temporal consistency of occupancy predictions across consecutive frames.

5.2. Implementation Details

Network. Our framework builds on the recent FB-OCC method [16] and follows its experimental setup. We employ a ResNet50 [7] backbone to extract perspective-view features from images of size 256×704. For more architectural details about the baseline, we refer readers to [16]. The output occupancy representation $\mathbf{V}$ is with dimensions $H = 100, W = 100, Z = 8$, and $C = 80$. For fair comparison, we build on FB-OCC without its temporal fusion module to demonstrate the effectiveness of our ST-Occ.

Our memory attention includes three temporal self-attention layers for temporal conditioning, along with a four-layer MLP to encode temporal attributes. The occupancy head includes three parallel three-layer convolutional networks to predict class activation $\mathbf{c}$, log variance $\mathbf{s}$, and occupancy flow $\mathbf{f}$. The decay factor of historical class activation (*i.e.*, α in Eq. (7)) is set to 0.5.

Training. We train our ST-Occ with a learning rate of 2×10^{-4} for 26 epochs. Temporal modeling is excluded for the first three epochs to stabilize training. The ground truth for occupancy flow used in our flow prediction training is derived from nuScenes annotations in real-time by computing the instance bounding box offsets across time. Lastly, we utilize grid sampling function with bilinear interpolation to update our spatiotemporal memory.

5.3. Results

We conduct experiments to compare our proposed ST-Occ with previous state-of-the-art 3D occupancy prediction models, including TPVFormer [10], OccFormer [34], BEVFormer [15], CTF-Occ [27], FB-OCC [16], and ViewFormer [12]. We replace the temporal fusion modules of FB-OCC and ViewFormer with our framework while keeping other settings identical. Tab. 1 presents the 3D occupancy prediction results of all methods on the Occ3D benchmark [27]. Compared with the previous state-of-the-art FB-OCC, our method achieves a substantial improvement of 3 mIoU, with consistent performance gains

Method	mSTCV (%) ↓	mSTCV † (%) ↓
FB-OCC	12.18	8.57
Mem. Attn.	9.25	6.64
Mem. Attn. + Uncertainty	8.85	6.53
Mem. Attn. + Uncertainty + Dynamics	**8.68**	**6.48**

Table 2. Temporal consistency evaluation results on Occ3D datasets of FB-OCC and ST-Occ under various settings. *Mem. Attn.* denotes the memory attention. † without applying voxel visibility mask. Our default setting is colored in grey.

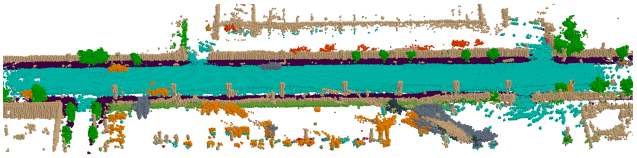
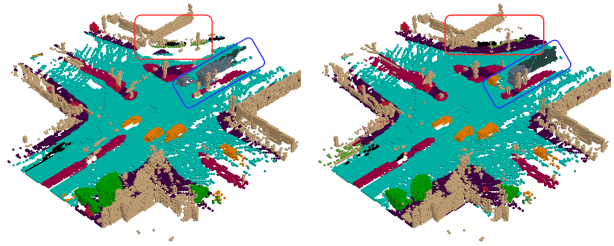


Figure 5. An example visualization of our aggregated spatiotemporal memory.



(a) Prediction from FB-OCC **(b)** Prediction from ST-Occ

Figure 6. A qualitative comparison on the occupancy prediction results between FB-OCC and ST-Occ.

across all classes. In examining the impact of temporal modeling, FB-OCC's improvement with temporal fusion is 1.72 mIoU, while our approach achieves an increase of 4.74 mIoU—approximately 2.8× more effective than FB-OCC. Our framework also achieves 4.5 mIoU improvement on ViewFormer without temporal modeling, which is 25% more effective in temporal modeling and surpasses the baseline by 0.9 mIoU. These results demonstrate the superior performance of our method and its effectiveness in temporal modeling.

We also compute our proposed temporal consistency metric, mSTCV, for both our method and the baselines, as shown in Tab. 2. Compared with FB-OCC, our method reduces the temporal inconsistency by more than 25%, whether applying the voxel visibility mask or not. These demonstrate the robustness and enhanced temporal consistency achieved by our method in occupancy prediction.

Qualitative Results Fig. 1 presents an example of occupancy prediction results from our method. Our method produces high-quality occupancy predictions centered around the ego vehicle at various timestamps, displayed in the first row. The spatiotemporal memory retains the entire historical spatiotemporal representation, preserving a holistic view of the scene. The second row shows occupancy predictions derived from our spatiotemporal memory. Fig. 5 demonstrate our ST-Occ on another large-scale scene.

Fig. 6 shows a qualitative comparison. Compared to FB-OCC, our approach more accurately predicts occupancy in some occluded regions, provides better instance classification and reduces noise in the predictions. These results demonstrate our method's capability to model long-term temporal dependencies with both dynamic and uncertainty awareness, resulting in a more refined spatiotemporal representation and more robust occupancy prediction.

Temporal Modeling	Training Mem. (GB) ↓	Fusion Time (ms) ↓	Inference Mem. (GB) ↓	FPS↑
Recurrent	12.89	705	10.08	5.95
Stacked	19.02	84	11.29	5.42
Unified (ours)	**10.90**	**24**	**5.57**	**8.65**

Table 3. Training and inference efficiency comparison between three temporal modeling paradigms. Our approach achieves the best efficiency across training and inference.

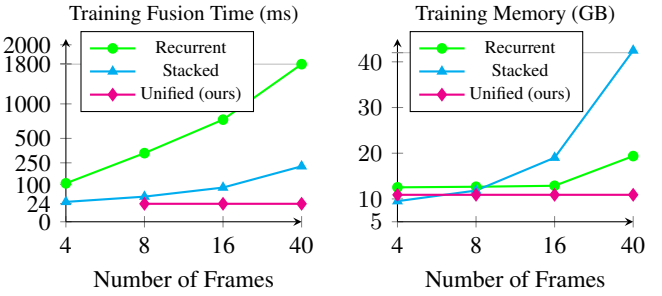


Figure 7. Effect of the number of temporal frames incorporated on the training efficiency of three temporal fusion paradigms. Our temporal modeling yields significantly less computational and memory footprint across number of frames.

5.4. Ablation Study

Temporal Modeling Efficiency We compare the efficiency of our temporal modeling against a recurrent-based approach (VoxFormer) [14] and a stacking-based approach (FB-OCC) [16] in Tab. 3. We use the same fusion operation, number of historical frames, and occupancy size for all methods to ensure fair comparison. These results indicate that our approach is the most efficient across metrics. Fig. 7 further demonstrates that our approach remains cost-efficient when incorporating more frames.

Uncertainty & Dynamic Awareness To understand our ST-Occ, we ablate different design choices in Tab. 4. The *No Temporal* setting is identical to FB-OCC but without temporal fusion. The *Mem. Attn.* corresponds to the memory attention in Sec. 4.3, with *Dynamics* and *Uncertainty* awareness optionally incorporated. Our method ST-Occ incorporates both uncertainty and dynamic awareness.

Settings	mIoU	others	barrier	bicycle	bus	car	const. veh.	motorcycle	pedestrian	traffic cone	trailer	truck	driv. surf.	other flat	sidewalk	terrain	manmade	vegetation
No Temporal	37.39	12.17	44.83	25.73	42.61	47.97	23.16	25.17	25.77	26.72	31.31	34.89	78.83	41.42	49.06	52.22	39.07	34.61
Mem. Attn.	41.17	13.96	48.62	27.81	44.83	51.01	27.07	28.86	29.22	29.28	33.96	38.41	83.62	46.36	54.94	58.76	43.93	39.15
Mem. Attn. + Dynamics	41.73	14.03	48.46	**28.02**	**47.23**	52.10	**27.72**	28.81	29.43	29.97	**36.28**	38.81	84.00	45.66	55.45	59.24	44.43	39.75
Mem. Attn. + Uncertainty	41.85	14.35	**50.31**	27.48	46.26	51.93	27.62	28.68	29.28	29.82	34.60	38.92	84.09	47.32	**56.23**	59.54	45.00	39.98
ST-Occ	**42.13**	**14.36**	49.62	27.77	46.28	**52.55**	26.87	**29.79**	**29.83**	**31.39**	35.40	**39.03**	**84.26**	**47.72**	56.09	**59.85**	**45.27**	**40.11**

Table 4. 3D occupancy prediction results of ST-Occ with different design settings on Occ3D benchmark.

Compared with the model without temporal information, our vanilla memory attention achieves a performance increase of 3.78 mIoU, validating the effectiveness of our temporal fusion paradigm in exploiting historical information for occupancy prediction. Additionally, when augmented with our proposed dynamic awareness, our memory attention achieves around 1 IoU increase across classes covering dynamic instances. These improvements demonstrate our method's ability to capture instance dynamics. Further, incorporating uncertainty awareness yields a complementary performance boost, particularly in static classes, with approximately a 1 IoU increase. This result indicates that our approach can leverage uncertainty-aware attention to further mitigate inter-frame noise for better prediction in static regions. Lastly, equipping our memory attention with both uncertainty and dynamic awareness combines their strength and leads to the best performance across static and dynamic regions, increasing performance for 1 mIoU and consistently enhancing performance for all classes.

c	ε	δ	**f**	mIoU
				41.17
✓				41.45
✓	✓			41.73
✓	✓	✓		41.85
			✓	41.73
✓	✓	✓	✓	42.13

Table 5. Ablation study on different sub-components of ST-Occ on Occ3D benchmark. **c**, ε, δ, and **f** correspond to the historical class activation, feature similarity, averaged log variance, and occupancy flow.

Sub-components of ST-Occ Tab. 5 presents an ablation study on the sub-components of our ST-Occ, including 1) the use of historical class activation **c**, feature similarity ε, and averaged log variance δ in uncertainty estimation. 2) occupancy flow **f**, used to compensate for the movement of dynamic voxels across time. In line with our motivation, integrating more historical information into uncertainty estimation leads to more gains in performance, as the model can recurrently refine its predictions of uncertain regions using cross-time information.

Number of Frames for Temporal Fusion In Fig. 8, we evaluate the impact of varying temporal fusion lengths on our method's performance. Results show that our method benefits from long-term temporal fusion, achieving higher mIoU and reduced temporal inconsistency as the number of frames increases. Notably, compared to FB-OCC which uses 16 temporal frames, our method attains similar performance while only using 8 temporal frames, reducing temporal inconsistency by 20% with an equal number of temporal frames during inference.

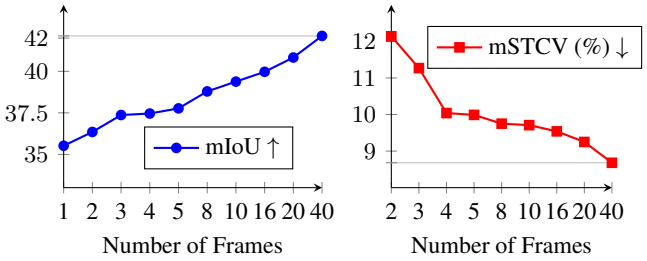


Figure 8. Effect of number of frames incorporated during inference on the precision and robustness of occupancy prediction. We observe a strong correlation between the number of incorporated frames and the studied metrics.

6. Conclusion

We propose ST-Occ, a scene-level occupancy representation learning framework. Our approach introduces a new temporal fusion paradigm, unified temporal modeling, designed to capture long-term temporal dependencies in occupancy prediction. Our framework leverages spatiotemporal memory and memory attention, incorporating both uncertainty and dynamic awareness. Our ST-Occ achieves significant improvement over prior occupancy prediction methods. Consistent improvements on prediction precision (mIoU) and robustness (mSTCV) from extensive experiments demonstrate the effectiveness of our ST-Occ.

There are limitations in our current approach that suggest potential directions for future work. Our method models dynamic voxels using an explicit estimator trained on existing nuScenes annotations. Future research could integrate this dynamic modeling directly into the temporal fusion process by deriving occupancy flow from temporal information, thereby reducing reliance on annotations. Another promising direction is to extend the unified temporal modeling to sparse query-based perception methods.

Acknowledgements

This work was supported by the NSF under Grants IIS-2339769 and CNS-2235012, and the Sony Focused Research Award.

References

[1] Tesla AI Day. https://www.youtube.com/watch?v=j0z4FweCy4M, 2021. 2

[2] Maxim Berman, Amal Rannen Triki, and Matthew B Blaschko. 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*, pages 4413–4421, 2018. 6

[3] Holger Caesar, Varun Bankiti, Alex H Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuscenes: A multimodal dataset for autonomous driving. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11621–11631, 2020. 6

[4] Anh-Quan Cao and Raoul De Charette. Monoscene: Monocular 3d semantic scene completion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3991–4001, 2022. 2, 6

[5] Anh-Quan Cao, Angela Dai, and Raoul de Charette. Pasco: Urban 3d panoptic scene completion with uncertainty awareness. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14554–14564, 2024. 1, 3

[6] Jean Dezert, Julien Moras, and Benjamin Pannetier. Environment perception using grid occupancy estimation with belief functions. In *2015 18th international conference on information fusion (Fusion)*, pages 1070–1077. IEEE, 2015. 2

[7] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 6

[8] Junjie Huang and Guan Huang. Bevdet4d: Exploit temporal cues in multi-camera 3d object detection. *arXiv preprint arXiv:2203.17054*, 2022. 1, 3, 4

[9] Junjie Huang, Guan Huang, Zheng Zhu, Yun Ye, and Dalong Du. Bevdet: High-performance multi-camera 3d object detection in bird-eye-view. *arXiv preprint arXiv:2112.11790*, 2021. 2, 3

[10] Yuanhui Huang, Wenzhao Zheng, Yunpeng Zhang, Jie Zhou, and Jiwen Lu. Tri-perspective view for vision-based 3d semantic occupancy prediction. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9223–9232, 2023. 3, 6

[11] Heng Li, Yuenan Hou, Xiaohan Xing, Xiao Sun, and Yanyong Zhang. Occmamba: Semantic occupancy prediction with state space models. *arXiv preprint arXiv:2408.09859*, 2024. 1, 3

[12] Jinke Li, Xiao He, Chonghua Zhou, Xiaoqiang Cheng, Yang Wen, and Dan Zhang. Viewformer: Exploring spatiotemporal modeling for multi-view 3d occupancy perception via view-guided transformers. In *European Conference on Computer Vision*, pages 90–106. Springer, 2025. 6

[13] Yinhao Li, Zheng Ge, Guanyi Yu, Jinrong Yang, Zengran Wang, Yukang Shi, Jianjian Sun, and Zeming Li. Bevdepth: Acquisition of reliable depth for multi-view 3d object detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 1477–1485, 2023. 4

[14] Yiming Li, Zhiding Yu, Christopher Choy, Chaowei Xiao, Jose M Alvarez, Sanja Fidler, Chen Feng, and Anima Anandkumar. Voxformer: Sparse voxel transformer for camera-based 3d semantic scene completion. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9087–9098, 2023. 1, 7

[15] Zhiqi Li, Wenhai Wang, Hongyang Li, Enze Xie, Chonghao Sima, Tong Lu, Yu Qiao, and Jifeng Dai. Bevformer: Learning bird's-eye-view representation from multi-camera images via spatiotemporal transformers. In *European conference on computer vision*, pages 1–18. Springer, 2022. 1, 2, 3, 4, 5, 6

[16] Zhiqi Li, Zhiding Yu, David Austin, Mingsheng Fang, Shiyi Lan, Jan Kautz, and Jose M Alvarez. Fb-occ: 3d occupancy prediction based on forward-backward view transformation. *arXiv preprint arXiv:2307.01492*, 2023. 1, 2, 3, 6, 7

[17] Zhiqi Li, Zhiding Yu, Wenhai Wang, Anima Anandkumar, Tong Lu, and Jose M Alvarez. Fb-bev: Bev representation from forward-backward view transformations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6919–6928, 2023. 2

[18] T Lin. Focal loss for dense object detection. *arXiv preprint arXiv:1708.02002*, 2017. 6

[19] Xuewu Lin, Tianwei Lin, Zixiang Pei, Lichao Huang, and Zhizhong Su. Sparse4d: Multi-view 3d object detection with sparse spatial-temporal fusion. *arXiv preprint arXiv:2211.10581*, 2022. 3

[20] Haisong Liu, Yang Chen, Haiguang Wang, Zetong Yang, Tianyu Li, Jia Zeng, Li Chen, Hongyang Li, and Limin Wang. Fully sparse 3d occupancy prediction. In *European Conference on Computer Vision*, pages 54–71. Springer, 2024. 1, 2, 3

[21] Qihang Ma, Xin Tan, Yanyun Qu, Lizhuang Ma, Zhizhong Zhang, and Yuan Xie. Cotr: Compact occupancy transformer for vision-based 3d occupancy prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19936–19945, 2024. 1, 2

[22] OpenDriveLab. OpenDriveLab Challenge 2024: Occupancy and Flow Track, 2024. 1

[23] Wenzhe Ouyang, Xiaolin Song, Bailan Feng, and Zenglin Xu. Octocc: High-resolution 3d occupancy prediction with octree. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 4369–4377, 2024. 2

[24] Jinhyung Park, Chenfeng Xu, Shijia Yang, Kurt Keutzer, Kris Kitani, Masayoshi Tomizuka, and Wei Zhan. Time will tell: New outlooks and a baseline for temporal multi-view 3d object detection. *arXiv preprint arXiv:2210.02443*, 2022. 1, 3

[25] Marcel Schreiber, Vasileios Belagiannis, Claudius Gläser, and Klaus Dietmayer. Dynamic occupancy grid mapping

with recurrent neural networks. In *2021 IEEE International Conference on Robotics and Automation (ICRA)*, pages 6717–6724. IEEE, 2021. 2

[26] Hao Shi, Song Wang, Jiaming Zhang, Xiaoting Yin, Zhongdao Wang, Zhijian Zhao, Guangming Wang, Jianke Zhu, Kailun Yang, and Kaiwei Wang. Occfiner: Offboard occupancy refinement with hybrid propagation. *arXiv preprint arXiv:2403.08504*, 2024. 3

[27] Xiaoyu Tian, Tao Jiang, Longfei Yun, Yucheng Mao, Huitong Yang, Yue Wang, Yilun Wang, and Hang Zhao. Occ3d: A large-scale 3d occupancy prediction benchmark for autonomous driving. *Advances in Neural Information Processing Systems*, 36, 2024. 1, 2, 6

[28] Wenwen Tong, Chonghao Sima, Tai Wang, Li Chen, Silei Wu, Hanming Deng, Yi Gu, Lewei Lu, Ping Luo, Dahua Lin, et al. Scene as occupancy. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8406–8415, 2023. 1, 2, 3

[29] Shihao Wang, Yingfei Liu, Tiancai Wang, Ying Li, and Xiangyu Zhang. Exploring object-centric temporal modeling for efficient multi-view 3d object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3621–3631, 2023. 3

[30] Xiaofeng Wang, Zheng Zhu, Wenbo Xu, Yunpeng Zhang, Yi Wei, Xu Chi, Yun Ye, Dalong Du, Jiwen Lu, and Xingang Wang. Openoccupancy: A large scale benchmark for surrounding semantic occupancy perception. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 17850–17859, 2023. 2

[31] Yuqi Wang, Yuntao Chen, Xingyu Liao, Lue Fan, and Zhaoxiang Zhang. Panoocc: Unified occupancy representation for camera-based 3d panoptic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 17158–17168, 2024. 3

[32] Yi Wei, Linqing Zhao, Wenzhao Zheng, Zheng Zhu, Jie Zhou, and Jiwen Lu. Surroundocc: Multi-camera 3d occupancy prediction for autonomous driving. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 21729–21740, 2023. 1

[33] Zichen Yu, Changyong Shu, Jiajun Deng, Kangjie Lu, Zongdai Liu, Jiangyong Yu, Dawei Yang, Hui Li, and Yan Chen. Flashocc: Fast and memory-efficient occupancy prediction via channel-to-height plugin. *arXiv preprint arXiv:2311.12058*, 2023. 1, 2, 3

[34] Yunpeng Zhang, Zheng Zhu, and Dalong Du. Occformer: Dual-path transformer for vision-based 3d semantic occupancy prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9433–9443, 2023. 6

[35] Wenzhao Zheng, Weiliang Chen, Yuanhui Huang, Borui Zhang, Yueqi Duan, and Jiwen Lu. Occworld: Learning a 3d occupancy world model for autonomous driving. In *European Conference on Computer Vision*, pages 55–72. Springer, 2025. 2

[36] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection. *arXiv preprint arXiv:2010.04159*, 2020. 5