ECAI 2024
U. Endriss et al. (Eds.)

doi:10.3233/FAIA240480

TEOcc: Radar-Camera Multi-Modal Occupancy Prediction via Temporal Enhancement

Zhiwei Lin[a,1], **Hongbo Jin**[a,1,2], **Yongtao Wang**[a,*], **Yufei Wei**[b] **and Nan Dong**[b]

[a]Wangxuan Institute of Computer Technology, Peking University
[b]Chongqing Changan Automobile Co., Ltd.

Abstract. As a novel 3D scene representation, semantic occupancy has gained much attention in autonomous driving. However, existing occupancy prediction methods mainly focus on designing better occupancy representations, such as tri-perspective view or neural radiance fields, while ignoring the advantages of using long-temporal information. In this paper, we propose a radar-camera multi-modal temporal enhanced occupancy prediction network, dubbed TEOcc. Our method is inspired by the success of utilizing temporal information in 3D object detection. Specifically, we introduce a temporal enhancement branch to learn temporal occupancy prediction. In this branch, we randomly discard the $t-k$ input frame of the multi-view camera and predict its 3D occupancy by long-term and short-term temporal decoders separately with the information from other adjacent frames and multi-modal inputs. Besides, to reduce computational costs and incorporate multi-modal inputs, we specially designed 3D convolutional layers for long-term and short-term temporal decoders. Furthermore, since the lightweight occupancy prediction head is a dense classification head, we propose to use a shared occupancy prediction head for the temporal enhancement and main branches. It is worth noting that the temporal enhancement branch is only performed during training and is discarded during inference. Experiment results demonstrate that TEOcc achieves state-of-the-art occupancy prediction on nuScenes benchmarks. In addition, the proposed temporal enhancement branch is a plug-and-play module that can be easily integrated into existing occupancy prediction methods to improve the performance of occupancy prediction. The source code and models will be released at https://github.com/VDIGPKU/TEOcc.

1 Introduction

Three-dimensional occupancy prediction is a novel and important task for modern autonomous driving perception systems [27]. Compared to common 3D object detection, occupancy can represent objects in arbitrary shapes with continuous 3D grid cells and semantic labels. Thus, it can provide fine-grained geometry details, including the specific shape of the foreground object and the concrete geometry of the surrounding background in the whole scene for better perception [26]. Besides, it is not insurance enough for autonomous driving scenarios to recognize all predefined objects encountered during training [1]. Unseen objects may appear on the road and collide with the self-driving vehicle. In this situation, 3D occupancy can present novel objects with non-empty grid cells and the 'others' category, avoiding collision.

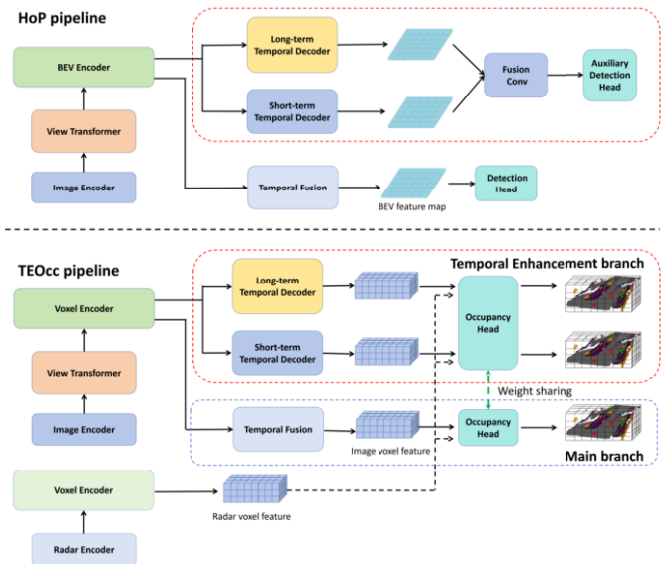

Figure 1. Differences between HoP and the proposed TEOcc. TEOcc uses independent long-term and short-term temporal decoders for 3D voxel feature generation and a shared head for occupancy prediction. Besides, TEOcc can incorporate radar-camera multi-modal inputs.

Current multi-view camera-based occupancy prediction methods mainly focus on how to represent occupancy, including voxels [31], bird's-eye-view (BEV) [7], tri-perspective view [12], and neural radiance fields (NeRF) [22]. Some of them use hierarchical representations to obtain fine-grained occupancy features from the coarse features [27]. Although numerous occupancy prediction methods [3, 27, 12, 17, 24] have been proposed, they have little exploration in long-term temporal modeling, which achieves great success in 3D object detection [8, 15, 19, 23, 29]. For instance, HoP [34] is an effective temporal modeling technique in 3D object detection. It generates a pseudo Bird's-Eye View feature of timestamp $t-k$ from its adjacent frames and utilizes this feature to predict the object set at timestamp $t-k$ in the training stage. However, HoP is only designed for BEV-based camera-only 3D object detection methods. It cannot be directly applied to the multi-modal occupancy prediction task because of different feature representations and complex multi-modal

* Corresponding Author. Email: wyt@pku.edu.cn
[1] Equal contribution.
[2] This work was done as an intern at PKU.

inputs.

To this end, we propose a radar-camera multi-modal occupancy prediction network with a temporal enhancement branch, named TEOcc, Specifically, in the temporal enhancement branch, we randomly mask one frame of the temporal multi-view camera inputs and generate its pseudo features with a long-term and short-term temporal decoder. Note that we only perform the temporal enhancement branch during training. Thus, no extra overheads are introduced during inference. Different from HoP, to reduce training costs and combine radar features, we design long-term and short-term temporal decoders with 3D convolutional layers. Besides, HoP concatenates features of long-term and short-term temporal decoders to produce one pseudo feature, while TEOcc uses independent temporal decoders to generate two pseudo features. Furthermore, we propose to use a shared 3D occupancy prediction head for two pseudo features and the main branch to predict corresponding occupancy results. The reason is that the occupancy prediction head is more like a dense classification head, which maps voxel features to semantic results. Thus, using a shared occupancy prediction head for all features can learn a better mapping.

The main contributions of this work are summarized as follows:

- We propose TEOcc, a radar-camera multi-modal temporal enhanced occupancy prediction network, which extends HoP to the multi-modal occupancy prediction task.
- Different from HoP, we use hand-craft long-term and short-term temporal decoders to independently predict occupancy with a shared occupancy prediction head.
- Building upon the current competitive 3D occupancy prediction method, TEOcc achieves state-of-the-art 3D occupancy estimation performance on the nuScenes dataset.

2 Related Works

2.1 3D Occupancy Prediction

The ability to forecast 3D occupancy, creating a comprehensive 3D voxel-based semantic depiction of a scene, is demanding and suited for autonomous driving systems. The emergence of Occ3D-nuScenes dataset [27] provided a fine-grained 3D occupancy ground truth and sparked more exploration in this field. Recently, several occupancy prediction methods have been proposed.

For point-based methods, S3CNet [4] formulates a sparse convolution-based neural network that deals with the sparsity of large-scale environments and predicts the semantically completed scene from the LiDAR point cloud. AIC-Net [14] presents an anisotropic convolutional network with adaptability to dimensional anisotropy and implicitly enables 3D kernels with varying sizes. Local-DIFs [25] produces a representation for 3D scenes by deep implicit functions with spatial support and generates point-like training targets from LiDAR data. More recently, PointOcc [35] introduces a cylindrical tri-perspective view to represent point clouds for occupancy prediction. OccWorld [32] proposes to learn the movement of the ego car and the evolution of the surrounding scenes simultaneously with a world model.

For camera-based methods, MonoScene [3] utilizes 3D Context Relation Prior and 2D-3D U-nets to enhance spatial-semantic awareness. SelfOcc [11] explores a self-supervised way to learn 3D occupancy using only video sequences. OccFormer [31] designs a dual-path transformer network to effectively process the 3D volume for semantic occupancy prediction. FB-OCC [17] proposes a bidirectional projection framework to utilize both forward and backward projection and avoid their limitations, obtaining better dense representation. TPVFormer [12] introduces a tri-perspective view as a new representation to depict the 3D scenes for predicting semantic occupancy. SurroundOcc [30] designs a pipeline to generate dense occupancy ground truth by fusing multi-frame LiDAR scans of dynamic objects and static scenes separately. RenderOcc [22] attempts to train multi-view 3D occupancy networks solely using 2D labels via volume rendering. Univision [6] simultaneously handles occupancy prediction and object detection utilizing a unified representation. FastOcc [7] accelerates the model while keeping its accuracy by replacing the time-consuming 3D convolution network with a novel residual architecture, where features are mainly extracted by a lightweight 2D BEV convolution network.

In contrast, our proposed TEOcc introduces temporal information in occupancy prediction via temporal enhancement and presents a radar-camera multi-modal occupancy prediction network.

2.2 Temporal Modeling in 3D Object Detection

Motion information has been extensively explored to improve performance with temporal cues in the 3D object detection task. BEVFormer [16] designs the temporal self-attention mechanism to dynamically fuse the previous BEV features by deformable attention [33] in an RNN manner. BEVDet4D [8] introduces the temporal modeling to lift BEVDet [10] to spatial-temporal 4D space. BEVStereo [15] proposes a dynamic temporal stereo technique for tackling the ill-posed issue of depth perception. STS [29] brings the temporal stereo technique in multi-view 3D object detection to facilitate accurate depth learning and 3D detection. SOLOFusion [23] utilizes a multi-view stereo (MVS) method to process high-resolution short-term images and then warp low-resolution long-term BEV features to produce a fused temporal BEV feature. PETRv2 [21] extends the position embedding transformation to temporal representation learning. HoP [34] encourages more accurate BEV feature learning via performing object detection in the historical frame. More recently, StreamPETR [28] proposes to utilize sparse object queries as intermediate representations to capture temporal information. SparseBEV [20] incorporates an adaptive spatio-temporal sampling module to perceive the BEV and temporal information dynamically.

These methods show great effectiveness for 3D object detection. However, few attempts have been made to explore temporal enhancement techniques for better perception performance in the 3D occupancy prediction task. Our proposed TEOcc tries to address this deficiency.

3 Method

3.1 Overall Architecture

Our overall architecture is shown in Figure 2. Specifically, we first extract multi-frame 3D volume features with the image encoder and view-transformation module. Meanwhile, we extract radar voxel features with a radar encoder. Then, in the temporal enhancement branch, we discard the image feature of the selected frame and reconstruct it using features from other frames and radar. The reconstructed pseudo feature and current 3D feature are sent to a shared occupancy head for final occupancy and semantic prediction.

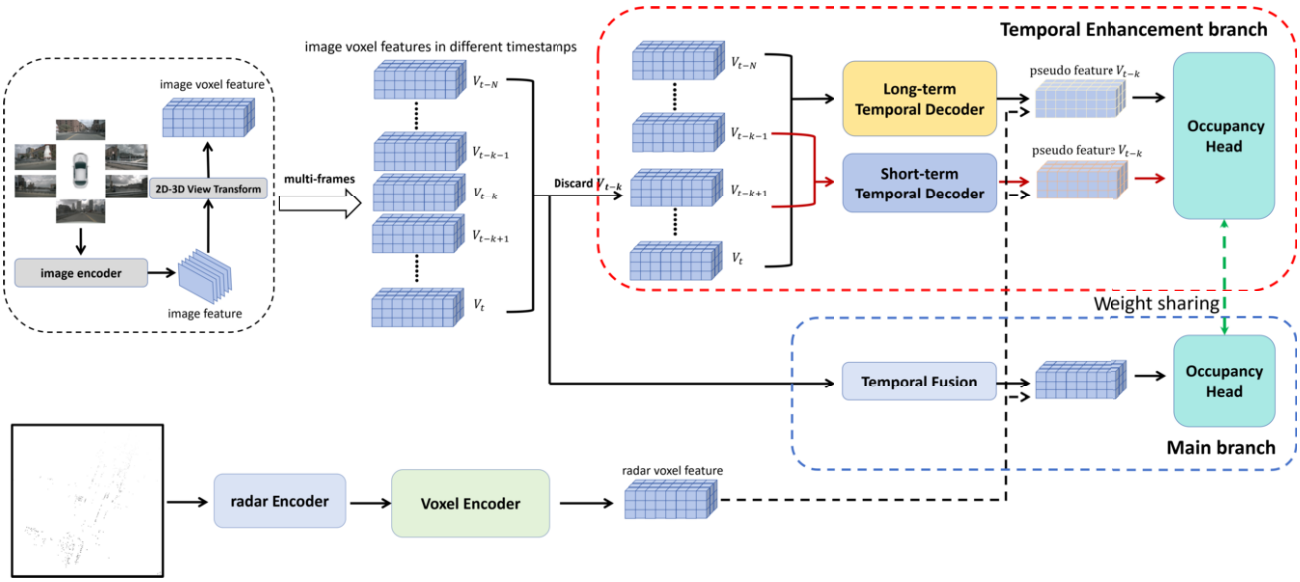

Figure 2. Overall pipeline of TEOcc. First, multi-frame multi-view camera features are extracted with an image encoder. The extracted 2D image features are transformed into 3D image voxel features with a 2D-3D view transformation module. Parallelly, we use a radar encoder and voxel encoder to extract radar voxel features. After that, in the main branch, all temporal image voxel features and radar voxel features are kept to predict final occupancy results. In the temporal enhancement branch, we discard one image voxel feature and use long-term and short-term decoders to generate corresponding pseudo features. Finally, a shared occupancy head is used to predict occupancy from generated pseudo voxel features.

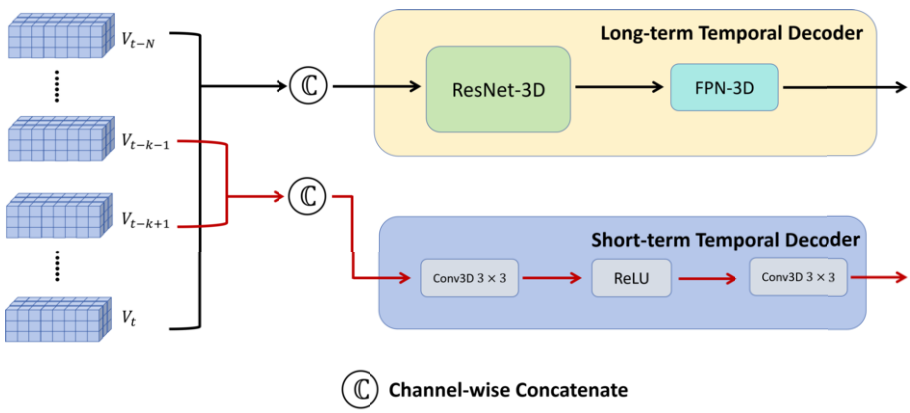

Figure 3. Architecture of the temporal enhancement module. The long-term temporal decoder consists of a ResNet-3D backbone and a FPN-3D neck to process multi-scale 3D voxel features. The short-term decoder is composed of two 3D convolution layers.

3.2 *Temporal Enhancement Branch*

As shown in Figure 3, our temporal enhancement branch comprises independent long-term and short-term temporal decoders. Specifically, given the 3D image voxel feature sequence $\{V_{t-N}, V_{t-N+1}, ..., V_t\}$ that consists of N historical image voxel features and the current image voxel feature, we randomly mask out the 3D image voxel feature V_{t-k}. Then, we send the remaining 3D image voxel feature sequence $\{V_{t-N}, ..., V_t\} - \{V_{t-k}\}$ and radar voxel feature to long-term temporal decoder. For short-term temporal decoder, we use adjacent features $\{V_{t-k-1}, V_{t-k+1}\}$ and radar voxel feature as the input. Two temporal decoders predict two pseudo-3D voxel features V_{t-k}. Finally, we use the occupancy head to predict the occupancy results for the $t-k$ frame.

Temporal Decoder. We designed two different temporal decoders based on the different temporal inputs following HoP. The short-term temporal decoder mainly focuses on the information from adjacent voxel features set $\{V_{t-k-1}, V_{t-k+1}\}$, while the long-term temporal decoder processes the whole temporal voxel feature set $\{V_{t-N}, ..., V_{t-k-1}, V_{t-k+1}, ..., V_{t-1}\}$. Due to the high temporal correlation between adjacent frames, the short-term temporal decoder can create a detailed spatial representation for V_{t-k}. In contrast, the long-term temporal decoder perceives the motion clues over long-term history, which improves the localization accuracy [23]. Therefore, these two branches are complementary to each other.

As depicted in HoP [34], Deformable Attention shows powerful historical object prediction ability because the coordinate offsets in Deformable Attention can match the movement of foreground objects and model temporal motion cues for the 3D object detection task. However, in the occupancy prediction task, every voxel needs to be classified rather than only predicting foreground objects. Therefore, Deformable Attention may not be appropriate for the 3D occupancy prediction task. Besides, full 3D Attention can capture all voxels in 3D space, but resulting in huge computational costs and time

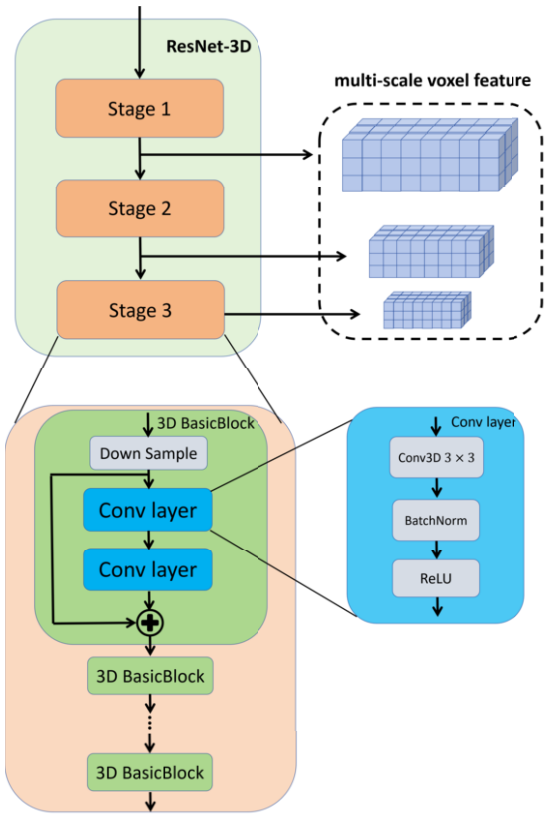

Figure 4. Architecture of ResNet-3D. ResNet-3D has three stages. Each stage consists of several 3D BasicBlocks.

overhead.

Fortunately, in our experiments, we find that simple 3D convolutions not only capture the movement of objects with temporal information, but also obtain a precise dense voxel representation for the 3D occupancy prediction task. Therefore, we replace Deformable Attention in HoP for temporal decoders with 3D convolution and specially design convolutional layers.

Specifically, as shown in Figure 3, we design ResNet-3D and FPN-3D as long-term temporal decoders. More concretely, ResNet-3D consists of three stages with the downsampling operation. Each stage is composed of several 3D BasicBlocks. Every 3D BasicBlocks has two 3D convolutional layers followed by a ReLU activation layer, as illustrated in Figure 4. In the three stages of ResNet-3D, we obtain three 3D voxel features in different scales. To further fuse these 3D voxel features with different resolutions, we send them to the FPN-3D. As shown in Figure 5, we first use trilinear interpolation to upsample these 3D voxel features to one resolution. Finally, we concatenate the upsampled 3D voxel features and send them to a convolutional layer followed by a norm and a ReLU activation layer. Using this multi-scale feature pyramid network to process image features, we can enhance spatial recognition ability and strong adaptability for the dense occupancy prediction task.

For short-term temporal decoders, since there is a high temporal correlation between voxel features from two adjacent frames, we find that two 3D convolutional layers with a ReLU activate layer can fuse the adjacent voxel features well.

Occupancy head. Different from HoP, we do not use an extra auxiliary occupancy head. In contrast, we use a shared occupancy head with multi-layer perceptions. The reason is that the occupancy head serves as the mapping function from voxel features to the occupancy category. Thus, it is not relevant to a specific frame, and the mapping can be learned by the same lightweight occupancy head used in the main branch.

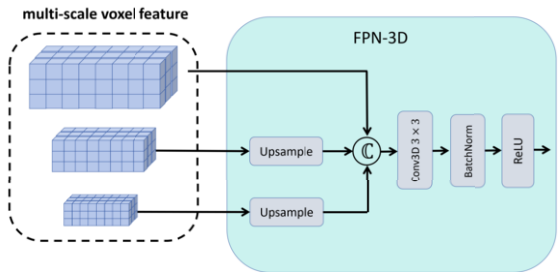

Figure 5. Architecture of FPN-3D. We upsample multi-scale voxel features into one scale and fuse them with a 3D convolution layer.

3.3 Radar-camera Fusion with Temporal Enhancement

In Sections 3.2, we discuss the temporal enhancement for multi-view camera-based 3D occupancy prediction methods. This section extends the temporal enhancement to radar-camera multi-modal fusion 3D occupancy prediction.

Specifically, we follow the pipeline of BEVFusion and replace the unified BEV space with the unified 3D voxel space. Specifically, as shown in Figure 2, we use a radar encoder and a voxel encoder to extract radar voxel features. Then, radar voxel features are sent to the main and temporal enhancement branches for radar-camera fusion. Besides, following BEVFusion, we concatenate the image voxel features and radar voxel features and use a 3D convolutional layer followed by a norm and a ReLU activation layer as the fusion layers. We utilize three independent fusion layers for the main and temporal enhancement branches. Finally, the fused multi-modal features are sent to the shared occupancy head to predict occupancy results.

3.4 Training and Inference

During the training stage, we keep the original occupancy loss of the main branch and add two additional occupancy losses from the temporal enhancement branch. The overall optimization objective is formulated as follows:

$$\mathcal{L} = \mathcal{L}_{Occ} + \mathcal{L}_{Occ\_long} + \mathcal{L}_{Occ\_short}, \tag{1}$$

where $\mathcal{L}_{Occ\_long}$ and $\mathcal{L}_{Occ\_short}$ denote the occupancy loss from long-term and short-term temporal decoders, respectively.

For inference, the temporal enhancement branch is removed. We only use the occupancy prediction from the main branch. Therefore, no extra inference cost is introduced.

4 Experiment

4.1 Dataset

We evaluate our method on Occ3D-nuScenes [27] benchmark. The Occ3D-nuScenes dataset is built upon the widely used large-scale autonomous driving dataset, nuScenes [2], which includes 1000 outdoor driving scenes with six surrounding-view cameras, LiDAR, and radar sensors. To provide high-quality occupancy labels, Occ3D-nuScenes first separates dynamic and static objects with LiDAR segmentation labels provided by nuScenes. Then, it aggregates multi-frame LiDAR points and utilizes the K-Nearest-Neighbor algorithm and mesh reconstruction to obtain a dense voxel with classification

Method	mIoU	Others	Barrier	Bicycle	Bus	Car	Cons.veh	Motorcycle	Pedestrian	Traffic cone	Trailer	Truck	Dri.sur	Other flat	Sidewalk	Terrain	Manmade	Vegetation
MonoScene	6.06	1.75	7.23	4.26	4.93	9.38	5.67	3.989	3.01	5.90	4.45	7.17	14.91	6.32	7.92	7.43	1.01	7.65
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
BEVStereo	24.50	15.73	38.41	7.88	38.70	41.20	17.56	17.33	14.69	10.31	16.84	29.62	54.08	28.92	32.68	26.54	18.74	17.49
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
RenderOcc	26.11	4.84	31.72	10.72	27.67	26.45	13.87	18.2	17.67	17.84	21.19	23.25	63.2	36.42	46.21	44.26	19.58	20.72
TPVFormer	27.83	7.2	38.9	13.7	40.8	45.9	17.2	20.0	18.8	14.3	26.7	34.2	55.6	35.5	37.6	30.7	19.4	16.78
SurroundOcc	37.18	8.97	46.33	17.08	46.54	52.01	20.05	21.47	23.52	18.67	31.51	37.56	81.91	41.64	50.76	53.93	42.91	37.16
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
FastOcc	39.21	12.06	43.53	28.04	44.80	52.16	22.96	29.14	29.68	26.98	30.81	38.44	82.04	41.93	51.92	53.71	41.04	35.49
TEOcc (Ours)	39.36	9.59	47.60	13.82	43.91	52.87	27.92	17.58	23.89	21.69	33.91	39.60	83.38	41.84	54.94	57.92	50.83	47.23
TEOcc-RC (Ours)	42.90	10.82	50.33	24.28	48.99	57.32	29.38	24.41	30.14	28.46	36.46	43.01	83.96	43.09	56.00	59.34	54.18	49.16

Table 1. Comparison of 3D occupancy prediction results on OCC3D-nuScenes val set. 'Cons.veh' and 'Dri.sur' are the shorts for construction vehicles and driveable surfaces. 'TEOcc-RC' means the radar-camera multi-modal version of TEOcc.

labels for occupancy. Occ3D-nuScenes provides 16 classes and a free class of 3D semantic labels for each scene. Each sample covers a range of [-40m, 40m], [-40m, 40m], and [-1m, 5.4m] with a voxel size of 0.4m for the x, y, and z-axis, respectively. The metric used in this benchmark is the mean Intersection over Union (mIoU) score. To obtain mIoU, we calculate the IoU value for each class and average the IoU value over 17 classes.

4.2 Implementation Details

We use a multi-view camera occupancy prediction method, BEVStereo [15], as the baseline for TEOcc. BEVStereo is composed of an image encoder, a 2D-3D view transformer, a BEV encoder, and an occupancy prediction head. We maintain the majority of the original structure of BEVStereo and add the proposed temporal enhancement module to construct the multi-view camera-based TEOcc. Then, we train radar-camera multi-modal TEOcc-RC based on the single-modal TEOcc.

For image feature extraction, we use ResNet50 [5] and FPN [18] as the image encoder. We employ 9 temporal frames. The image size is set to 256×704. For the radar encoder, we employ PointPillar with a voxel size of [0.4, 0.4, 0.4] for the x, y, and z axes. PointPillar first divides radar points into several pillars according to the voxel size. Then, it uses a simplified version of PointNet to extract features of radar points in each pillar. Finally, the pillar features are scattered to create a 2D Bird's-Eye View feature. We use Adam [13] as the optimizer with a batch size of 4. We train the network 24 epochs with a learning rate of 1e-4. For data augmentation, we use the same image augmentation with BEVPoolv2 [9], *i.e.*, image rotation and flip. Besides, we use horizontal flips as the augmentation in voxel space.

4.3 Main Results

We compare the proposed TEOcc with previous state-of-the-art 3D occupancy prediction methods on the Occ3D-NuScenes validation set in Table 1. TEOcc shows competitive 3D occupancy prediction performance. Specifically, TEOcc achieves a mIoU of 39.36, outperforming all previous multi-view camera-based occupancy prediction methods, including TPVFormer, SurroundOcc, and FastOcc. Besides, the challenge of dynamic object recognition described in RenderOCC is alleviated by our method. In particular, for dynamic objects like cars, buses, trailers, and trucks, TEOcc significantly improves the occupancy performance compared with RenderOcc. In addition, for static objects, TEOcc is still ahead of RenderOcc, demonstrating the effectiveness of temporal enhancement for the comprehensive understanding of 3D spatial relationships. Furthermore, when combined with radar inputs, TEOcc-RC improves TEOcc by 3.54 mIoU, surpassing the previous state-of-the-art 3D occupancy prediction method FastOcc by 3.69 mIoU.

long-term	short-term	random	mIoU
			21.02
✓			24.33
✓	✓		26.12
✓	✓	✓	26.95

Table 2. Ablation of main components. Each component improves the 3D occupancy performance consistently.

long-term	short term	fusion	mIoU
✓			24.33
✓	✓		26.12
✓	✓	✓	24.65

Table 3. Ablation of independent temporal decoders. Using independent temporal decoders achieves better results than fusing 3D voxel features from two decoders.

In summary, the results indicate that TEOcc with temporal enhancement can improve the construction and perception of occupancy representations for existing frameworks and enhance the overall understanding of 3D scenes.

4.4 Ablation Study

To validate the effectiveness of the proposed module, we conduct extensive ablation studies on the Occ3D-NuScenes dataset. In this section, to reduce the training costs, we train TEOcc with 0.5× schedule, *i.e.*, 12 epochs.

Main Components. We conduct comprehensive experiments to verify the effectiveness of each module. The main results of the experiment are shown in Table 2. Long-term enhancement networks provide the most significant performance improvement, with 3.31 mIoU. The short-term temporal decoder helps to improve the final prediction results with 1.79 mIoU. It is worth noting that we use a random selection strategy to select which frame is masked in the temporal enhancement branch, while HoP only chooses the $t-1$ frame for masking. The results show that using the random selection strategy results in a noticeable performance boost. We speculate that the random selection strategy makes the model pay more attention to the temporal changes of surrounding scenes rather than mechanically memorizing the positional relationship between the past and the current frame.

Independent Temporal Decoders. Different from HoP [34], we obtain two independent 3D voxel features with fine-grained granularity from the two temporal decoders. As shown in Table 3, using independent temporal decoders shows better occupancy prediction performance compared with fusing the 3D voxel features from two tem-

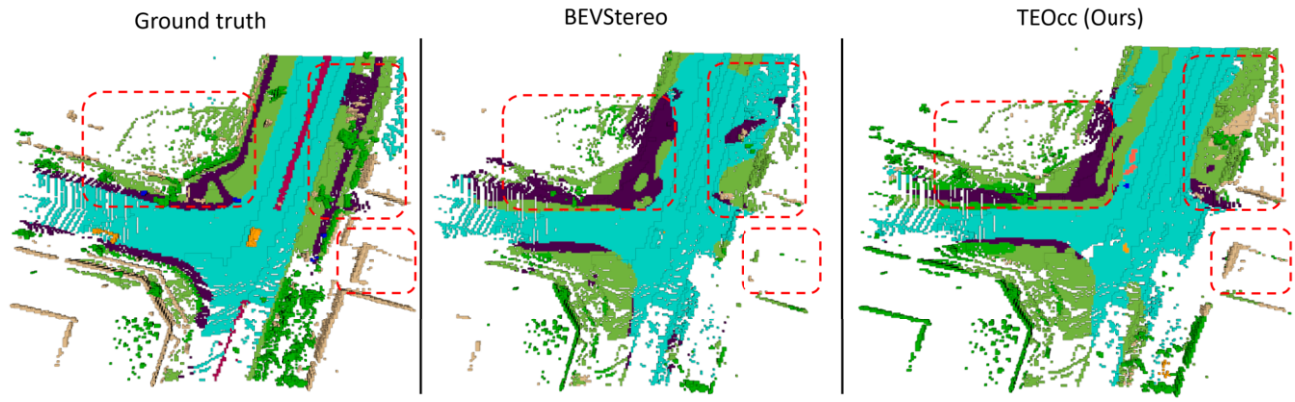

Figure 6. Occupancy visualization in global view. From left to right, we show the ground truth, BEVStereo (baseline), and TEOcc prediction results. We can see that TEOcc is more accurate in perceiving global details, especially distant information and occluded parts.

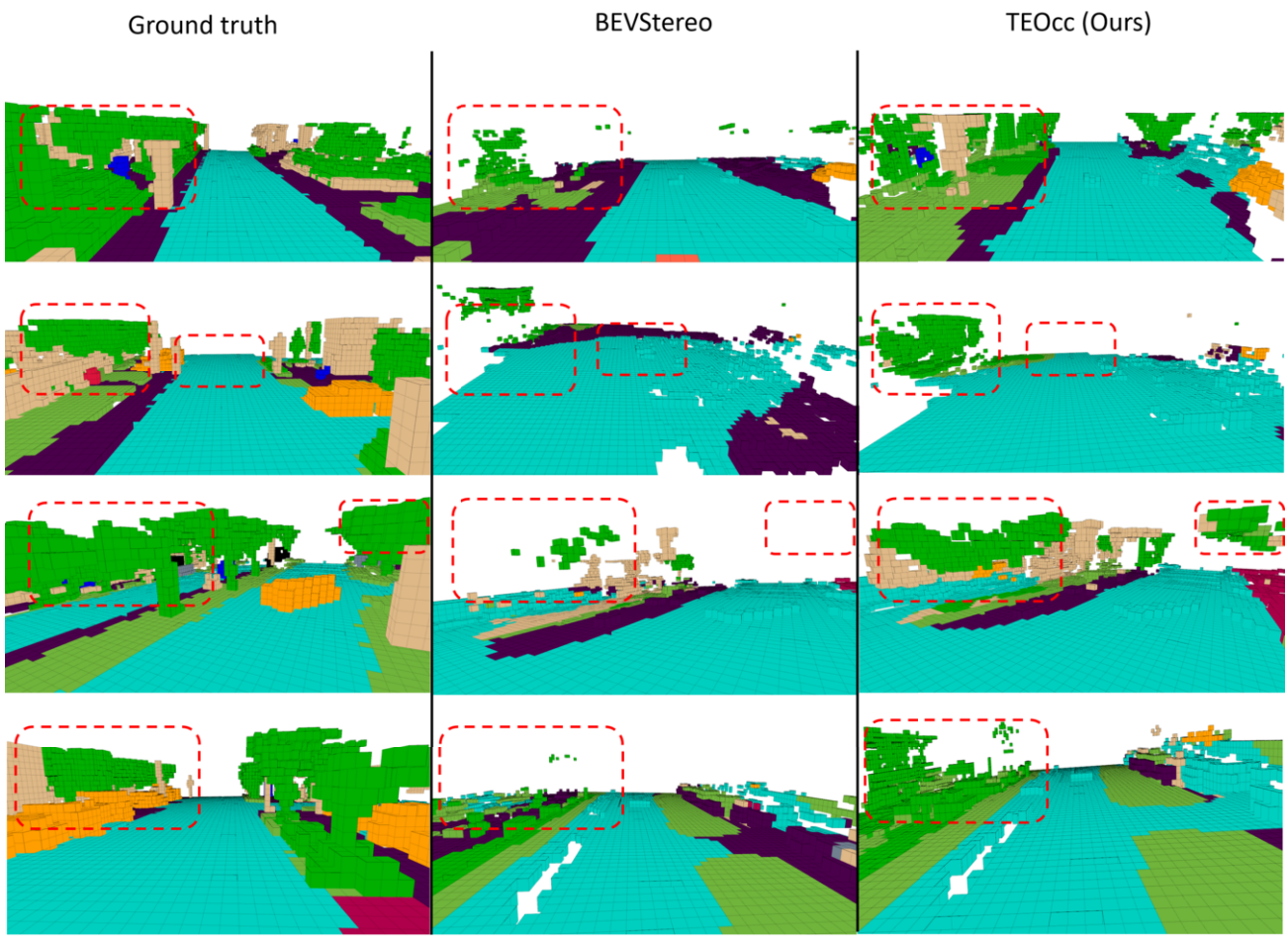

Figure 7. Occupancy visualization in local view. From left to right, we show ground truth, BEVStereo (baseline), and TEOcc prediction results. It can be observed that TEOcc is better at predicting object shape.

flip augmentation	mIoU
	31.16
✓	32.05

Table 4. Ablation of voxel data augmentation. Filp data augmentation in voxel space improves occupancy performance.

shared occupancy head	mIoU
	24.63
✓	26.12

Table 5. Ablation of occupancy head. Sharing an occupancy head obtains better occupancy results.

Method	GPU Memory	Training Time
BEVStereo	1×	1×
TEOcc	1.8×	1.1×

Table 6. Ablation of training cost. TEOcc brings marginal training time and GPU memory increasing.

poral decoders. We speculate that the long-term and short-term temporal encoders learn different 3D voxel features due to the different temporal lengths. Fusing them into one 3D voxel feature may lead to feature conflicts.

Data Augmentation. In addition to image augmentation, we follow the BEV data augmentation in BEVPoolv2 [9] to add voxel data augmentation. As shown in Table 4, we find horizontal flip data augmentation in voxel space can improve the occupancy prediction performance from 31.16 mIoU to 32.05 mIoU.

Shared Occupancy Head. As shown in Table 5, we compare the results of using the additional auxiliary occupancy head and shared occupancy head for the final prediction. The results show that employing a sharing occupancy head obtains 1.51 mIoU improvement compared with the additional auxiliary head. The reason is that the occupancy head is a dense classification head, which maps 3D voxel features to the occupancy category. Thus, the shared occupancy head forces the generated pseudo 3D voxel features to share the same feature space with the main branch, allowing the temporal enhancement branch to learn a more unified temporal representation with smaller training costs.

Training Costs. Because we add the temporal enhancement branch during the training stage, the training cost increases. To evaluate the efficiency of TEOcc in training, we compare its training time and GPU memory consumption with the BEVStereo baseline. This training time is evaluated with 8 NVIDIA A800 GPUs. As shown in Table 6, the additional training time brought by TEOcc is negligible. However, due to the utilization of 3D convolution in temporal decoders, the memory consumption increases to 1.8×.

4.5 Visualization

We provide the occupancy visualization of the global view and local view in Figure 6 and 7, respectively. The figures show that, with the proposed temporal enhancement module, TEOcc can capture more accurate long-distance view information and predict detailed scenes locally.

5 Conclusion

In this paper, we propose a radar-camera multi-modal temporal enhanced occupancy prediction network, named TEOcc. Specifically, we generate a pseudo voxel feature of timestamp $t-k$ from its adjacent frames and utilize this feature to predict the occupancy results at timestamp $t-k$. Besides, we design 3D convolutional-based short-term and long-term temporal decoders to predict the pseudo voxel features. Furthermore, we propose to use independent temporal decoders and a shared occupancy prediction head in TEOcc. As a plug-and-play method, the temporal enhancement module can be easily incorporated into existing occupancy prediction methods with additional 0.1× training time costs. Extensive experiments on the Occ-3D nuScenes validation dataset show the effectiveness of the proposed TEOcc. Specifically, TEOcc-RC achieves 42.90 mIoU, outperforming all the previous occupancy networks and achieving new state-of-the-art occupancy prediction results on the leaderboard.

Acknowledgements

This work was supported by National Key R&D Program of China (Grant No. 2022ZD0160305).

References

[1] S. Boeder, F. Gigengack, and B. Risse. Occflownet: Towards self-supervised occupancy estimation via differentiable rendering and occupancy flow. *arXiv preprint arXiv:2402.12792*, 2024.
[2] H. Caesar, V. Bankiti, A. H. Lang, S. Vora, V. E. Liong, Q. Xu, A. Krishnan, Y. Pan, G. Baldan, and O. Beijbom. nuscenes: A multimodal dataset for autonomous driving. In *CVPR*, pages 11621–11631, 2020.
[3] A.-Q. Cao and R. De Charette. Monoscene: Monocular 3d semantic scene completion. In *CVPR*, pages 3991–4001, 2022.
[4] R. Cheng, C. Agia, Y. Ren, X. Li, and L. Bingbing. S3cnet: A sparse semantic scene completion network for lidar point clouds. In *Conference on Robot Learning*, pages 2148–2161, 2021.
[5] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016.
[6] Y. Hong, Q. Liu, H. Cheng, D. Ma, H. Dai, Y. Wang, G. Cao, and Y. Ding. Univision: A unified framework for vision-centric 3d perception. *arXiv preprint arXiv:2401.06994*, 2024.
[7] J. Hou, X. Li, W. Guan, G. Zhang, D. Feng, Y. Du, X. Xue, and J. Pu. Fastocc: Accelerating 3d occupancy prediction by fusing the 2d bird's-eye view and perspective view. *arXiv preprint arXiv:2403.02710*, 2024.
[8] J. Huang and G. Huang. Bevdet4d: Exploit temporal cues in multi-camera 3d object detection. *arXiv preprint arXiv:2203.17054*, 2022.
[9] J. Huang and G. Huang. Bevpoolv2: A cutting-edge implementation of bevdet toward deployment. *arXiv preprint arXiv:2211.17111*, 2022.
[10] J. Huang, G. Huang, Z. Zhu, Y. Ye, and D. Du. Bevdet: High-performance multi-camera 3d object detection in bird-eye-view. *arXiv preprint arXiv:2112.11790*, 2021.
[11] Y. Huang, W. Zheng, B. Zhang, J. Zhou, and J. Lu. Selfocc: Self-supervised vision-based 3d occupancy prediction. *arXiv preprint arXiv:2311.12754*, 2023.
[12] Y. Huang, W. Zheng, Y. Zhang, J. Zhou, and J. Lu. Tri-perspective view for vision-based 3d semantic occupancy prediction. In *CVPR*, pages 9223–9232, 2023.
[13] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015.
[14] J. Li, K. Han, P. Wang, Y. Liu, and X. Yuan. Anisotropic convolutional networks for 3d semantic scene completion. In *CVPR*, pages 3351–3359, 2020.
[15] Y. Li, H. Bao, Z. Ge, J. Yang, J. Sun, and Z. Li. Bevstereo: Enhancing depth estimation in multi-view 3d object detection with temporal stereo. In *AAAI*, volume 37, pages 1486–1494, 2023.
[16] Z. Li, W. Wang, H. Li, E. Xie, C. Sima, T. Lu, Y. Qiao, and J. Dai. Bevformer: Learning bird's-eye-view representation from multi-camera images via spatiotemporal transformers. In *ECCV*, pages 1–18, 2022.
[17] Z. Li, Z. Yu, D. Austin, M. Fang, S. Lan, J. Kautz, and J. M. Alvarez. Fb-occ: 3d occupancy prediction based on forward-backward view transformation. *arXiv preprint arXiv:2307.01492*, 2023.
[18] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie. Feature pyramid networks for object detection. In *CVPR*, pages 2117–2125, 2017.
[19] X. Lin, T. Lin, Z. Pei, L. Huang, and Z. Su. Sparse4d: Multi-view 3d object detection with sparse spatial-temporal fusion. *arXiv preprint arXiv:2211.10581*, 2022.
[20] H. Liu, Y. Teng, T. Lu, H. Wang, and L. Wang. Sparsebev: High-performance sparse 3d object detection from multi-camera videos. In *ICCV*, 2023.
[21] Y. Liu, J. Yan, F. Jia, S. Li, A. Gao, T. Wang, and X. Zhang. Petrv2: A unified framework for 3d perception from multi-camera images. In *ICCV*, pages 3262–3272, 2023.
[22] M. Pan, J. Liu, R. Zhang, P. Huang, X. Li, H. Xie, B. Wang, L. Liu, and S. Zhang. Renderocc: Vision-centric 3d occupancy prediction with 2d rendering supervision. In *ICRA*, pages 12404–12411, 2024.
[23] J. Park, C. Xu, S. Yang, K. Keutzer, K. M. Kitani, M. Tomizuka, and W. Zhan. Time will tell: New outlooks and A baseline for temporal multi-view 3d object detection. In *ICLR*, 2023.
[24] L. Peng, J. Xu, H. Cheng, Z. Yang, X. Wu, W. Qian, W. Wang, B. Wu, and D. Cai. Learning occupancy for monocular 3d object detection. *arXiv preprint arXiv:2305.15694*, 2023.
[25] C. B. Rist, D. Emmerichs, M. Enzweiler, and D. M. Gavrila. Semantic scene completion using local deep implicit functions on lidar data. *IEEE TPAMI*, 44(10):7205–7218, 2021.

[26] Y. Shi, K. Jiang, J. Li, J. Wen, Z. Qian, M. Yang, K. Wang, and D. Yang. Grid-centric traffic scenario perception for autonomous driving: A comprehensive review. *arXiv preprint arXiv:2303.01212*, 2023.

[27] X. Tian, T. Jiang, L. Yun, Y. Mao, H. Yang, Y. Wang, Y. Wang, and H. Zhao. Occ3d: A large-scale 3d occupancy prediction benchmark for autonomous driving. *NeurIPS*, 36, 2024.

[28] S. Wang, Y. Liu, T. Wang, Y. Li, and X. Zhang. Exploring object-centric temporal modeling for efficient multi-view 3d object detection. In *ICCV*, 2023.

[29] Z. Wang, C. Min, Z. Ge, Y. Li, Z. Li, H. Yang, and D. Huang. Sts: Surround-view temporal stereo for multi-view 3d detection. *arXiv preprint arXiv:2208.10145*, 2022.

[30] Y. Wei, L. Zhao, W. Zheng, Z. Zhu, J. Zhou, and J. Lu. Surroundocc: Multi-camera 3d occupancy prediction for autonomous driving. In *ICCV*, pages 21729–21740, 2023.

[31] Y. Zhang, Z. Zhu, and D. Du. Occformer: Dual-path transformer for vision-based 3d semantic occupancy prediction. In *ICCV*, pages 9433–9443, 2023.

[32] W. Zheng, W. Chen, Y. Huang, B. Zhang, Y. Duan, and J. Lu. Occworld: Learning a 3d occupancy world model for autonomous driving. *arXiv preprint arXiv:2311.16038*, 2023.

[33] X. Zhu, W. Su, L. Lu, B. Li, X. Wang, and J. Dai. Deformable DETR: deformable transformers for end-to-end object detection. In *ICLR*, 2021.

[34] Z. Zong, D. Jiang, G. Song, Z. Xue, J. Su, H. Li, and Y. Liu. Temporal enhanced training of multi-view 3d object detector via historical object prediction. In *ICCV*, pages 3781–3790, 2023.

[35] S. Zuo, W. Zheng, Y. Huang, J. Zhou, and J. Lu. Pointocc: Cylindrical tri-perspective view for point-based 3d semantic occupancy prediction. *arXiv preprint arXiv:2308.16896*, 2023.