# Lightweight Spatial Embedding for Vision-based 3D Occupancy Prediction

Jinqing Zhang[1], Yanan Zhang[1], Qingjie Liu[1,2,3,*], Yunhong Wang[1,3]

[1]State Key Laboratory of Virtual Reality Technology and Systems, Beihang University, Beijing, China
[2]Zhongguancun Laboratory, Beijing, China
[3]Hangzhou Innovation Institute, Beihang University, Hangzhou, China

{zhangjinqing, zhangyanan, qingjie.liu, yhwang}@buaa.edu.cn

## Abstract

*Occupancy prediction has garnered increasing attention in recent years for its comprehensive fine-grained environmental representation and strong generalization to open-set objects. However, cumbersome voxel features and 3D convolution operations inevitably introduce large overheads in both memory and computation, obstructing the deployment of occupancy prediction approaches in real-time autonomous driving systems. Although some methods attempt to efficiently predict 3D occupancy from 2D Bird's-Eye-View (BEV) features through the Channel-to-Height mechanism, BEV features are insufficient to store all the height information of the scene, which limits performance. This paper proposes LightOcc, an innovative 3D occupancy prediction framework that leverages Lightweight Spatial Embedding to effectively supplement the height clues for the BEV-based representation while maintaining its deployability. Firstly, Global Spatial Sampling is used to obtain the Single-Channel Occupancy from multi-view depth distribution. Spatial-to-Channel mechanism then takes the arbitrary spatial dimension of Single-Channel Occupancy as the feature dimension and extracts Tri-Perspective Views (TPV) Embeddings by 2D convolution. Finally, TPV Embeddings will interact with each other by Lightweight TPV Interaction module to obtain the Spatial Embedding that is optimal supplementary to BEV features. Sufficient experimental results show that LightOcc significantly increases the prediction accuracy of the baseline and achieves state-of-the-art performance on the Occ3D-nuScenes benchmark.*

## 1. Introduction

With the gradual popularization of autonomous driving technology, the requirements for the perception ability of autonomous driving are getting higher. Unlike traditional

---

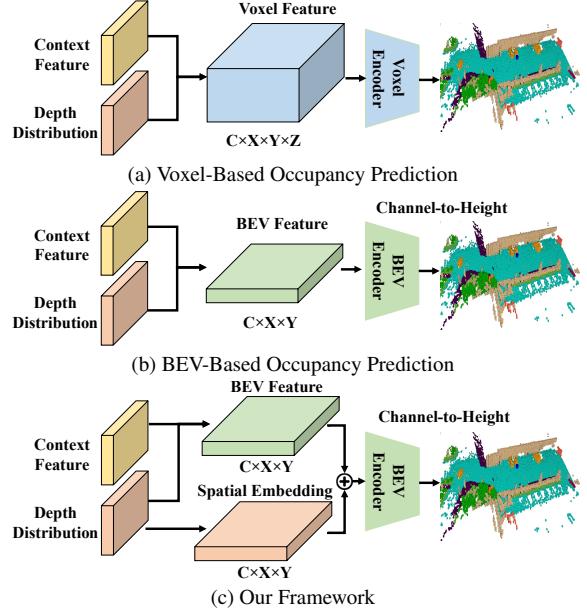*indicates the corresponding author.



Figure 1. Comparison of different occupancy prediction frameworks. BEV-based methods replace voxel features with BEV features for better deployability but lose a portion of height information. Our method utilizes a lightweight module to extract Spatial Embedding that effectively supplements height information to BEV features, enabling more accurate prediction.

close-set 3D object detection [23, 24, 27, 33] that only predicts the 3D bounding boxes of objects in limited categories, 3D occupancy prediction [28, 37, 43] partitions the 3D scene into grid cells and predicting semantic labels for each voxel. It enables a more fine-grained understanding of 3D scenes, stronger generalization to open-set objects, and greater robustness to occlusion and irregularly shaped objects, thereby enhancing high-level autonomous driving capabilities.

Since precise 3D occupancy prediction relies on understanding the overall 3D spatial information of driving scenes, current methods [3, 5, 8, 17, 22, 32, 34] typi-

cally adopt voxel features as the representation, as shown in Fig. 1(a). This poses a significant challenge for practical deployment, as 3D voxel features consume much more memory than 2D BEV features under the same spatial resolution. Additionally, processing voxel features requires 3D convolutions or transformers, which further increases latency. In contrast, several methods [36, 39] argue that the occupancies at different heights are tightly correlated in driving scenes. They apply Channel-to-Height mechanism to simultaneously predict the occupancy categories at different heights from the same BEV features, as shown in Fig. 1(b). Such approaches can achieve comparable performance to voxel-based methods while offering better deployability. However, unlike voxel-based methods, which store height information within the Z coordinates of arranged voxels, BEV-based methods accumulate features within the same pillar without preserving height values. As a result, BEV features are not capable of storing all the height information, limiting the models' ability to achieve optimal perception performance.

Therefore, as shown in Fig. 1(c), we propose Lightweight Spatial Embedding to represent the complete spatial information for the driving scene, including the height information lost by BEV features. The voxel-based methods generally transform multi-view image features to get the voxel features, which costs a large amount of memory. We assume that even the single-channel features can store most spatial information as long as there are voxel coordinates. We propose Global Spatial Sampling to transform the depth distribution predicted by multi-view image features into Single-Channel Occupancy. Although there is no category information, it inherits the spatial information estimated from image features. To extract the spatial information and supplement the BEV features, we apply Spatial-to-Channel Mechanism that takes each spatial dimension of Single-Channel Occupancy as the channel dimension. It allows the 2D convolutions to extract Tri-Perspective Views (TPV) Embeddings, as shown in Fig 3. BEV Embedding is obtained by taking Z dimension as the feature channel, while Front View (FV) Embedding and Side View (SV) Embedding take X and Y dimensions as the feature channel, respectively.

Among TPV Embeddings, BEV Embedding has the same spatial dimensions as the BEV features and can directly supplement the height information, but it only implicitly stores the height information in the channel dimension. On the contrary, FV Embedding and SV Embedding explicitly keep the Z dimension and can extract well-located height information. TPV-based methods [9, 30] rely on time-consuming cross-attention to implement the interaction between the TPV features. Here, we propose an alternative approach: multiplying any two TPV Embeddings to obtain new embeddings in the other view. After several in-

teraction operations, the information of TPV Embeddings is fused into the Spatial Embedding to further enhance the height information supplemented to BEV features. In addition, to augment data diversity, we further propose an innovative data augmentation strategy, termed BEV-CutMix. It enables the construction of new driving scenes through cutting and mixing BEV features of different scenes without introducing erroneous occlusion relationships.

Integrating the aforementioned components, we engineer a novel 3D occupancy prediction framework, which we denominate as LightOcc. We implement experiments on the commonly used Occ3D-nuScenes benchmark, and the results demonstrate that LightOcc can effectively improve the BEV-based 3D occupancy prediction baseline while maintaining model efficiency. The major contributions of this paper can be summarized as:

- We substitute the cumbersome voxel features with the Single-Channel Occupancy to equivalently store height information of driving scenes. The Spatial-to-Channel mechanism enables the lightweight 2D convolution to extract TPV Embeddings.
- We implement the Lightweight TPV Interaction between TPV Embeddings through the simple combination of matrix multiplication and convolution operations, effectively integrating both explicit and implicit height information.
- We propose BEV-CutMix to increase the data diversity while circumventing the occurrence of erroneous occlusion relationships.
- We conducted experiments on the Occ3D-nuScenes benchmark, validating that LightOcc significantly enhances prediction accuracy while preserving the computational efficiency of BEV-based models.

## 2. Related Work

### 2.1. BEV Representation

The BEV-based perception framework is competent for integrating multi-view image features into a uniform BEV representation, which improves performance in tasks like 3D object detection and BEV map segmentation. LSS [26] weights the image features at different depths to generate pseudo-points and accumulates the pseudo-points along the height dimension to obtain BEV representation. BEVDet [26] constructs the BEV-based 3D object detection framework and proposes corresponding data augmentation strategies. BEVDet4D [6] fuses the BEV features from past frames to help predict the velocity of the objects. BEVDepth [14] and BEVStereo [13] introduce the camera parameters and multi-view stereo to optimize the depth prediction. SA-BEV [40] further integrates the semantic information from perspective view into the BEV representation. BEVNext [18] utilizes modules with extended receptive fields to aggregate long-term temporal information.
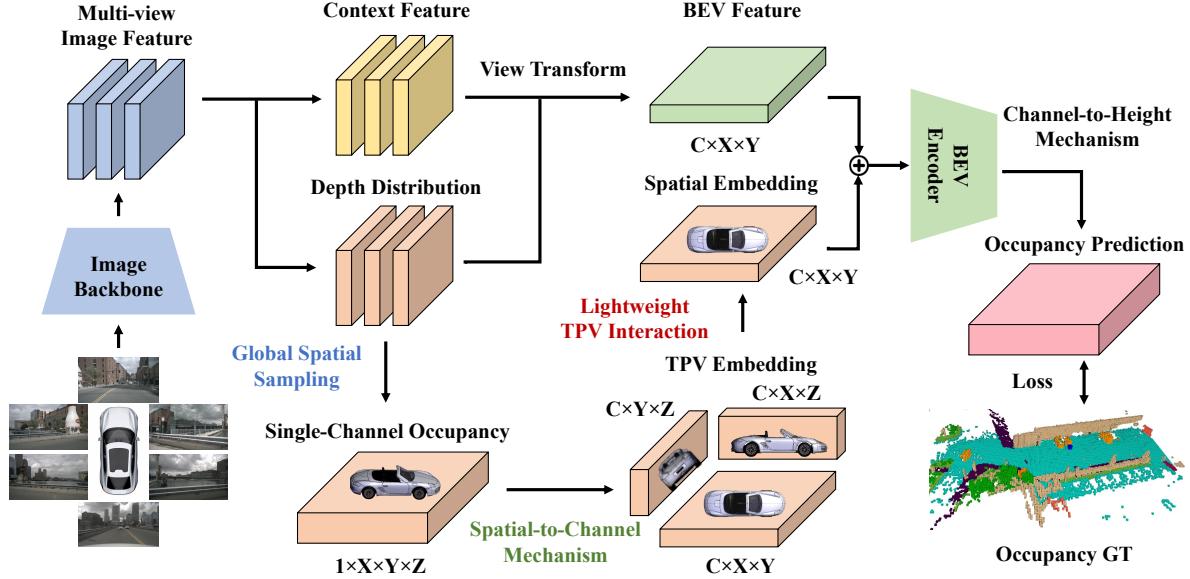
Figure 2. The overall architecture of LightOcc. Global Spatial Sampling gathers the predicted depth distribution into Single-Channel Occupancy, which is used to extract TPV Embeddings via Spatial-to-Channel mechanism. Lightweight TPV Interaction then fuses TPV Embeddings into Spatial Embedding and effectively supplements the lost height information to the BEV features.

GeoBEV [41] proposes RC-Sampling to efficiently generate BEV representation with high geometric quality.

Another branch of BEV-based perception methods uses Transformer to adaptively retrieve the corresponding image features to generate BEV representation. BEVFormer [16] applies cross-attention and self-attention to aggregate the image features and previous BEV features into the current BEV space. BEVFormerV2 [38] adds a 2D detection head and applies 2D detection supervision to optimize the image features. PolarFormer [11] transforms the BEV space into the polar space and designs a special detection head. DFA3D [12] inserts the explicit depth distribution into the BEV features when doing the cross-attention and simplifies the calculation of 3D Transformer.

## 2.2. Occupancy Prediction

Compared to 3D object detection, occupancy prediction provides a more detailed and comprehensive representation of the driving scene, aiding the vehicle in making more reasonable decisions. It also requires the ability of BEV framework to aggregate multi-view images, but generally represent the scene with 3D voxel features [3, 15, 42]. The classic 3D object detectors such as BEVDet and BEVFormer can be modified into the occupancy prediction method by transforming the image features into voxel features. SurroundOcc [35] expands the height dimension of BEV features by adopting spatial cross-attention and applies 3D convolution to upsample the volume features. FB-OCC [17] combines forward view transformation with the backward

approach to improve the BEV features, which are combined with voxel features for further process. COTR [22] reconstructs the compact 3D occupancy representation that has higher information density than voxel features. There are also some methods that pursue annotation-free training. RenderOcc [25] predicts the density and label for each voxel based on the voxel features and utilizes NeRF to render the depth map and semantic segmentation of the images. SelfOcc [10] employs a signed distance field to render 2D depth maps, along with raw color and semantic maps, under the supervision of the image sequence.

Unlike the approaches that utilize voxel features to represent the information of the driving scenes, several methods seek more efficient processing. TPVFormer [9] divides 3D space into three perspective views and utilizes attention modules to extract TPV features and apply cross-view interaction. S2TPVFormer [30] applies the temporal fusion based on TPVFormer, which proposes Temporal Cross-View Hybrid Attention to utilize the history TPV features as the latent spatial representation. FastOcc [5] replaces the voxel encoder with the BEV encoder to process BEV features and sample the image features to compensate for missing height information. FlashOcc [39] predicts 3D occupancy only based on the BEV features to achieve the best efficiency. It argues that occupancies at different heights are closely related and can be predicted utilizing the same features. DHD [36] follows the same conception and constructs three BEV features that represent the scene at different heights. After using extra modules to fuse these BEV

features, performance improvement is achieved. Different from these approaches that discard the height information or utilize heavy modules for compensation, we propose the Lightweight Spatial Embedding that preserves the comprehensive height information while having a trivial effect on the real-time performance.

## 3. Method

### 3.1. Overall Architecture

The overall architecture of our proposed LightOcc is shown in Fig. 2. Following the extraction of image features from multi-view images, the depth distribution is predicted for each view to facilitate the feature transformation from the perspective space to the BEV space. Subsequently, Global Spatial Sampling is employed to gather the depth distribution of each view and construct a Single-Channel Occupancy. Spatial-to-Channel mechanism enables the model to effortlessly extract embeddings from Tri-Perspective Views (TPV) by 2D convolutions. Lightweight TPV Interaction is then applied to fuse TPV Embeddings, producing the embedding that enriches the BEV features with the complete spatial information of the scene. In addition, a novel data augmentation strategy, BEV-CutMix, is introduced to increase data diversity during training.

### 3.2. Global Spatial Sampling

Recent improvements [7, 41] to LSS [26] have significantly boosted the efficiency of generating BEV features and circumvented the generation of huge 3D intermediate features, thereby enhancing the deployability of the model. However, the height information of the driving scenes, which is encoded in the height coordinates of 3D intermediate features, highly affects the accuracy of 3D occupancy prediction. Therefore, finding a way to preserve height information without resorting to heavy 3D intermediate features is crucial for enhancing BEV-based occupancy prediction approaches, such as FlashOcc [39].

Given that BEV features retain most information about driving scenes except for height details, we assume a 3D feature with few channels is sufficient to store the lost information. Consequently, we sample the multi-view depth distributions employed for view transformation, aiming to generate a Single-Channel Occupancy that represents comprehensive spatial information of the scene. Defining $D$ as the predicted depth channel and $H, W$ as the height and width of image features, the depth distribution can be represented as $\mathbf{D} \in \mathbb{R}^{D \times H \times W}$. We pre-define the $(x, y, z)$ as the 3D coordinates of Single-Channel Occupancy $\mathbf{O}_{SC} \in \mathbb{R}^{1 \times X \times Y \times Z}$, where $X, Y, Z$ are the length, width and height of the perception space, and project them into the image space as follows:

$$d(h, w, 1)^{\mathrm{T}} = \mathbf{K}^{-1}\left[\mathbf{R}(x, y, z)^{\mathrm{T}} + \mathbf{t}\right] \quad (1)$$
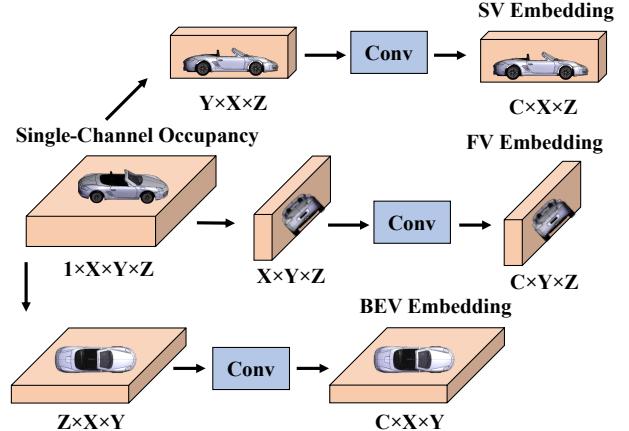


Figure 3. Spatial-to-Channel Mechanism

where $\mathbf{R}, \mathbf{t}, \mathbf{K}$ represent the rotation matrix, translation vector and intrinsic matrix of the camera. Regarding $(d, h, w)$ as the coordinates in $\mathbf{D}$, the Global Spatial Sampling can be formulated below:

$$\mathbf{O}_{SC}(x, y, z) = \sum_{i=1}^{N} \mathrm{Sampling}(\mathbf{D}_i, (d_i, h_i, w_i)) \quad (2)$$

where $N$ is the number of multi-view images and Sampling denotes the bilinear sampling operation.

Since the time cost by sampling operation is proportional to the number of feature channels, sampling $\mathbf{O}_{SC}$ is much more efficient than sampling 3D voxel features with the same spatial resolution. This ensures that Global Spatial Sampling does not affect the deployability of the BEV-based baseline.

### 3.3. Spatial-to-Channel Mechanism

Although $\mathbf{O}_{SC}$ offers an efficient and compact representation of height information, the challenge lies in effectively extracting the height information from $\mathbf{O}_{SC}$ and converting it into the Spatial Embedding that can be supplemented to BEV features. If we imitate the voxel-based method in how to process voxel features, a large amount of calculation will be required by multiple 3D convolution layers, which results in a significant drop in the efficiency of the model.

To this end, we propose the Spatial-to-Channel Mechanism, which can take arbitrary spatial dimension of $\mathbf{O}_{SC}$ as the feature channel. As shown in Fig. 3, taking the Z dimension as the feature channel, the BEV Embedding $\mathbf{E}_{BEV} \in \mathbb{R}^{C \times X \times Y}$ can be generated by:

$$\mathbf{E}_{BEV} = \mathrm{Conv}(\mathbf{O}_{SC} \to \mathbb{R}^{Z \times X \times Y}) \quad (3)$$

where $\to$ indicates the transpose operation of the matrix.

Similarly, the Front View (FV) Embedding $\mathbf{E}_{FV} \in \mathbb{R}^{C \times Y \times Z}$ and Side View (SV) Embedding $\mathbf{E}_{SV} \in \mathbb{R}^{C \times X \times Z}$
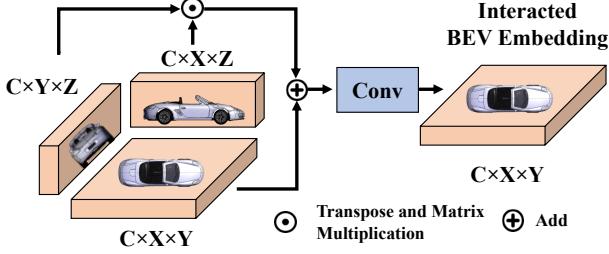
Figure 4. Lightweight TPV Interaction

can be obtained by:

$$\mathbf{E}_{FV} = \text{Conv}(\mathbf{O}_{SC} \rightarrow \mathbb{R}^{X \times Y \times Z}) \quad (4)$$

$$\mathbf{E}_{SV} = \text{Conv}(\mathbf{O}_{SC} \rightarrow \mathbb{R}^{Y \times X \times Z}). \quad (5)$$

Different from $\mathbf{E}_{BEV}$ that carries implicit height information in the feature channels, $\mathbf{E}_{FV}$ and $\mathbf{E}_{SV}$ keep the height dimension as the spatial dimension and thus can save explicit height information.

Unlike using multiple 3D convolution layers to gradually increase the receptive field until it covers all heights, Spatial-to-Channel mechanism largely improves the information density during processing $\mathbf{O}_{SC}$. Experiments show that using a single 2D convolution layer can effectively obtain the required spatial information.

### 3.4. Lightweight TPV Interaction

Among the TPV Embeddings extracted from $\mathbf{O}_{SC}$, $\mathbf{E}_{BEV}$ have the same spatial dimensions as the BEV features $\mathbf{F}_{BEV} \in \mathbb{R}^{C \times X \times Y}$, so they can be directly added to the BEV features and fed into the subsequent BEV encoder for further processing. However, $\mathbf{E}_{FV}$ and $\mathbf{E}_{SV}$ can not be directly fused with $\mathbf{F}_{BEV}$ because they have different spatial dimensions. Considering the importance of explicit height information carried $\mathbf{E}_{FV}$ and $\mathbf{E}_{SV}$, it is necessary to apply interaction between TPV Embeddings.

Methods such as TPVFormer [9] provide a direct TPV interaction strategy, which uses multi-layer Cross-View Hybrid-Attention to aggregate the TPV features. However, the heavy computation burden of Transformer runs counter to our goal of keeping the model efficient. As a result, we propose Lightweight TPV Interaction (LTI), which does matrix multiplication between the embeddings of any two views to obtain the embeddings with the same spatial dimension as the other view. As shown in Fig. 4, the interacted BEV Embedding $\mathbf{E}_{BEV}^I$ can be obtained by:

$$\mathbf{E}_{BEV}^M = (\mathbf{E}_{SV} \rightarrow \mathbb{R}^{C \times X \times Z}) \otimes (\mathbf{E}_{FV} \rightarrow \mathbb{R}^{C \times Z \times Y})$$
$$\mathbf{E}_{BEV}^I = \text{Conv}(\mathbf{E}_{BEV} + \mathbf{E}_{BEV}^M) \quad (6)$$

where $\otimes$ denotes matrix multiplication. The multiplied $\mathbf{E}_{BEV}^M \in \mathbb{R}^{C \times X \times Y}$ combines the information of $\mathbf{E}_{FV}$ and
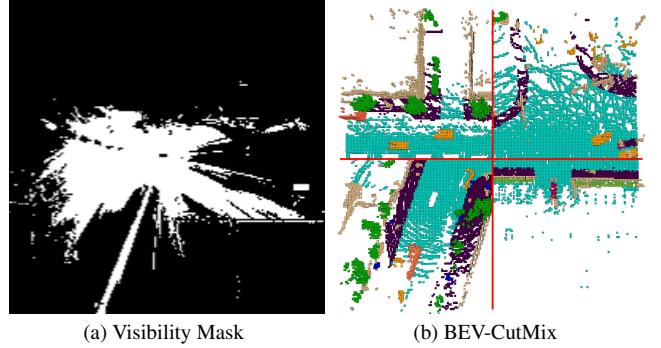


(a) Visibility Mask      (b) BEV-CutMix

Figure 5. Illustration of BEV-CutMix. Since the visibility mask is radially laid out from the center to the periphery. Cut the occupancy into 4 pieces from the center and randomly mix those pieces cross scenes do not cause wrong occlusion.

$\mathbf{E}_{SV}$ which are correctly mapped to the coordinates in BEV space. After adding $\mathbf{E}_{BEV}^M$ to $\mathbf{E}_{BEV}$, a single 2D convolution layer implements the interaction. Since $\mathbf{E}_{FV}$, $\mathbf{E}_{SV}$, $\mathbf{E}_{BEV}$ cover the $X, Y, Z$ dimensions respectively, LTI can quickly expand the receptive field to the whole space, enhancing the ability of information extraction.

After obtain $\mathbf{E}_{FV}^I$ and $\mathbf{E}_{SV}^I$ in the same way as follows:

$$\mathbf{E}_{FV}^M = (\mathbf{E}_{BEV} \rightarrow \mathbb{R}^{C \times Y \times X}) \otimes (\mathbf{E}_{SV} \rightarrow \mathbb{R}^{C \times X \times Z})$$
$$\mathbf{E}_{FV}^I = \text{Conv}(\mathbf{E}_{FV} + \mathbf{E}_{FV}^M) \quad (7)$$

$$\mathbf{E}_{SV}^M = (\mathbf{E}_{BEV} \rightarrow \mathbb{R}^{C \times X \times Y}) \otimes (\mathbf{E}_{FV} \rightarrow \mathbb{R}^{C \times Y \times Z})$$
$$\mathbf{E}_{SV}^I = \text{Conv}(\mathbf{E}_{SV} + \mathbf{E}_{SV}^M), \quad (8)$$

LTI is finally used to aggregate TPV Embeddings into the united Spatial Embedding $\mathbf{E}_S \in \mathbb{R}^{C \times X \times Y}$, which can be formulated by:

$$\mathbf{E}_S^M = (\mathbf{E}_{SV}^I \rightarrow \mathbb{R}^{C \times X \times Z}) \otimes (\mathbf{E}_{FV}^I \rightarrow \mathbb{R}^{C \times Z \times Y})$$
$$\mathbf{E}_S = \text{Conv}(\mathbf{E}_{BEV}^I + \mathbf{E}_S^M). \quad (9)$$

$\mathbf{E}_S$ carries the spatial information of the whole driving scene, including the implicit and explicit height clues, and can be directly added to $\mathbf{F}_{BEV}$ for subsequent processing and perception.

### 3.5. BEV-CutMix

Increasing diversity through data augmentation can effectively improve the generalization performance of the model. However, the augmentations on images have limited contribution to the diversity of data in BEV space. Some flexible augmentation strategies such as random scaling and random rotation are unsuitable for low-resolution and densely serialized occupancy annotations because they are likely to cause spatial misalignment. Inspired by BEV-Paste [40] that sum up the foreground-only BEV features of different scenes, we propose a novel data augmentation strategy

Table 1. 3D occupancy prediction performance on the Occ3D-nuScenes dataset. Both the small version and large version of LightOcc outperform the model that have similar settings.

| Method | History Frame | Resolution | Backbone | mIoU ↑ | others ↑ | barrier ↑ | bicycle ↑ | bus ↑ | car ↑ | cons. veh. ↑ | motorcycle ↑ | pedestrian ↑ | traffic cone ↑ | trailer ↑ | truck ↑ | drive. surf. ↑ | other flat ↑ | sidewalk ↑ | terrain ↑ | manmade ↑ | vegetation ↑ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| MonoScene [3] | ✗ | 928 × 1600 | R101 | 6.06 | 1.75 | 7.23 | 4.26 | 4.93 | 9.38 | 5.67 | 3.98 | 3.01 | 5.90 | 4.45 | 7.17 | 14.91 | 6.32 | 7.92 | 7.43 | 1.01 | 7.65 |
| CTF-Occ [32] | ✗ | 928 × 1600 | R101 | 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 |
| TPVFormer [9] | ✗ | 928 × 1600 | R101 | 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 [42] | ✗ | 256 × 704 | R50 | 20.40 | 6.62 | 32.57 | 13.13 | 20.37 | 37.12 | 5.04 | 14.02 | 21.01 | 16.96 | 9.34 | 20.64 | 40.89 | 27.02 | 27.43 | 18.65 | 18.78 | 16.90 |
| BEVDetOcc [8] | ✗ | 256 × 704 | R50 | 31.64 | 6.65 | 36.97 | 8.33 | 38.69 | 44.46 | 15.21 | 13.67 | 16.39 | 15.27 | 27.11 | 31.04 | 78.70 | 36.45 | 48.27 | 51.68 | 36.82 | 32.09 |
| FlashOcc (M1) [39] | ✗ | 256 × 704 | R50 | 32.08 | 6.74 | 37.65 | 10.26 | 39.55 | 44.36 | 14.88 | 13.4 | 15.79 | 15.38 | 27.44 | 31.73 | 78.82 | 37.98 | 48.7 | 52.5 | 37.89 | 32.24 |
| DHD-S [36] | ✗ | 256 × 704 | R50 | 36.50 | 10.59 | 43.21 | 23.02 | 40.61 | 47.31 | 21.68 | 23.25 | 23.85 | 23.40 | 31.75 | 34.15 | **80.16** | 41.30 | 49.95 | **54.07** | 38.73 | 33.51 |
| LightOcc-S | ✗ | 256 × 704 | R50 | **37.93** | **11.72** | **45.61** | **25.40** | **43.10** | **48.66** | 21.38 | **25.58** | **26.58** | **29.19** | **33.18** | **35.09** | 79.97 | **41.81** | **50.35** | 53.88 | **39.40** | **33.97** |
| BEVDetOcc-Stereo [8] | 1 | 512 × 1408 | SwinB | 42.02 | 12.15 | 49.63 | 25.10 | 52.02 | 54.46 | 27.87 | 27.99 | 28.94 | 27.23 | 36.43 | 42.22 | 82.31 | 43.29 | 54.62 | 57.90 | 48.61 | 43.55 |
| FlashOcc (M3) [39] | 1 | 512 × 1408 | SwinB | 43.52 | 13.42 | 51.07 | 27.68 | 51.57 | 56.22 | 27.27 | 29.98 | 29.93 | 29.80 | 37.77 | 43.52 | 83.81 | 46.55 | 56.15 | 59.56 | 50.84 | 44.67 |
| OSP [29] | 1 | 900 × 1600 | R101 | 39.41 | 11.20 | 47.25 | 27.06 | 47.57 | 53.66 | 23.21 | 29.37 | 29.68 | 28.41 | 32.39 | 39.94 | 79.35 | 41.36 | 50.31 | 53.23 | 40.52 | 35.39 |
| DHD-L [36] | 1 | 512 × 1408 | SwinB | 45.53 | 14.08 | 53.12 | 32.39 | 52.44 | 57.35 | 30.83 | 35.24 | 33.01 | 33.43 | 37.90 | 45.34 | **84.61** | 47.96 | 57.39 | 60.32 | 52.27 | 46.24 |
| FastOcc [5] | 16 | 640 × 1600 | R101 | 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 |
| PanoOcc [34] | 3 | 512 × 1408 | R101 | 42.13 | 11.67 | 50.48 | 29.64 | 49.44 | 55.52 | 23.29 | 33.26 | 30.55 | 30.99 | 34.43 | 42.57 | 83.31 | 44.23 | 54.40 | 56.04 | 45.94 | 40.40 |
| FB-Occ [17] | 4 | 512 × 1408 | R101 | 43.41 | 12.10 | 50.23 | 32.31 | 48.55 | 52.89 | 31.20 | 31.25 | 30.78 | 32.33 | 37.06 | 40.22 | 83.34 | **49.27** | 57.13 | 59.88 | 47.67 | 41.76 |
| OctreeOcc [21] | 4 | 512 × 1408 | R101 | 44.02 | 11.96 | 51.70 | 29.93 | 53.52 | 56.77 | 30.83 | 33.17 | 30.65 | 29.99 | 37.76 | 43.87 | 83.17 | 44.52 | 55.45 | 58.86 | 49.52 | 46.33 |
| COTR [22] | 8 | 512 × 1408 | SwinB | 46.20 | 14.85 | 53.25 | 35.19 | 50.83 | 57.25 | **35.36** | 34.06 | 33.54 | 37.14 | 38.99 | 44.97 | 84.46 | 48.73 | **57.60** | **61.08** | 51.61 | 46.72 |
| GEOcc [31] | 8 | 512 × 1408 | SwinB | 44.67 | 14.02 | 51.40 | 33.08 | 52.08 | 56.72 | 30.04 | 33.54 | 32.34 | 35.83 | 39.34 | 44.18 | 83.49 | 46.77 | 55.72 | 58.94 | 48.85 | 43.00 |
| LightOcc-L | 1 | 512 × 1408 | SwinB | 46.00 | 14.50 | 52.27 | 34.45 | **53.79** | 57.33 | 31.80 | 35.83 | 33.60 | 36.09 | 39.89 | 46.09 | 84.23 | 48.10 | 57.14 | 60.02 | 51.70 | 45.23 |
| LightOcc-L | 8 | 512 × 1408 | SwinB | **47.24** | **15.39** | **53.88** | **36.20** | 53.60 | **58.04** | 34.32 | **37.35** | **34.67** | **39.29** | **40.45** | **47.61** | 83.73 | 47.79 | 57.09 | 60.64 | **54.52** | **48.49** |

called BEV-CutMix, which cuts BEV features into several parts and mixes them into the BEV features of new scenes.

When applying BEV-CutMix, the difficulty lies in how to handle occlusion relationships for new scenes. In occupancy annotations, the occlusion relationship needs to be strictly obeyed, and the occluded regions are marked as invisible in the visibility mask. As a result, applying BEV-CutMix at random positions can introduce incorrect occlusion relationships, which hinders the model from learning accurate occupancy. Considering that the visibility mask radiates from the center to the edges as shown in Fig. 5(a), we divide the scene into four parts along the X and Y dimensions from the center, and if these four parts are arbitrarily combined as shown in Fig. 5(b), the introduction of false occlusion relationships can be avoided.

## 4. Experiments

### 4.1. Benchmark and Evaluation Metric

We conduct experiments on Occ3D-nuScenes [32] benchmark, which is built upon nuScenes dataset [2], the commonly used large-scale autonomous driving dataset. It contains 700 scenes for training and 150 scenes for validation and each scene has around 40 annotated samples. The occupancy annotations cover a spatial range from -40m to 40m along the X and Y dimensions, and -1m to 5.4m along the Z dimension. Each voxel in the occupancy annotations is a cube with a length of 0.4m, and the resolution of the occupancy is $200 \times 200 \times 16$. The voxels are annotated into 17 semantic categories and 1 free category, which means the voxels are not occupied by any objects. For evaluating the occupancy prediction performance, the mean intersection-over-union (mIoU) of all semantic categories is employed.

### 4.2. Implementation Details

We adopt the FlashOcc [39] as the baseline and apply our Lightweight Spatial Embedding to supplement the height information to BEV features. Following FlashOcc which constructs models with different scales to evaluate the best performance of accuracy and efficiency respectively, we implement two versions of LightOcc, namely LightOcc-S and LightOcc-L. LightOcc-S is basically consistent with the M1 version of FlashOcc, which utilizes the ResNet-50 [4] as the image backbone and inputs images in the size of $256 \times 704$. On the other hand, the model size of LightOcc-L is similar to the M3 version of FlashOcc, which utilizes Swin-Transformer-Base [19] as the image backbone and inputs the images in the size of $512 \times 1408$. The Multi-view Stereo module [13] used by FlashOcc (M3) is also employed. For training, we use the AdamW optimizer [20] with the learning rate $2 \times 10^{-4}$ to train the LightOcc for 48 epochs. All experiments are implemented by 8 RTX 3090 GPUs.

6

## 4.3. Main Results

We compare LightOcc with the state-of-the-art occupancy prediction method on Occ3D-nuScenes benchmark. The experiment results in Tab. 1 show that both LightOcc-S and LightOcc-L outperform other approaches that have similar model settings. LightOcc-S improves the performance of its baseline, i.e. FlashOcc (M1) model, by 5.85% mIoU and outperforms DHD-S [36] model by 1.43%. Light-L adopts the commonly used heavy configuration to pursue higher accuracy. When utilizing 1 history frame, Light-L improves the performance FlashOcc (M3) model by 2.48%. When utilizing 8 history frames, Light-L outperforms COTR [22] by 1.04% mIoU. These persuasive experiment results highlight the effectiveness of Lightweight Spatial Embedding in supplementing the height information to the BEV features.

## 4.4. Ablation Study

### 4.4.1. Effectiveness of Components

We evaluate the contributions of each component of LightOcc and show the results in Tab. 2. Based on the M0 version of FlashOcc, we construct a stronger baseline with several optimization measures, which include replacing view transformation operation with RC-Sampling [41], employing auxiliary loss such as Scene-Class Affinity Loss [3] and Lovász-Softmax Loss [1], and adjusting the BEV neck module. After optimization, the mIoU of the baseline is largely improved while the latency is decreased by 1.57 ms.

We gradually add our proposed component to the optimized strong baseline. After applying Spatial-to-Channel mechanism to the Single-Channel Occupancy $\mathbf{O}_{SC}$, BEV Embedding $\mathbf{E}_{BEV}$ is obtained and directly added to the BEV features $\mathbf{F}_{BEV}$. It brings a 1.3% mIoU improvement in accuracy, while the time cost is only slightly increased. After applying Lightweight TPV Interaction, TPV Embeddings efficiently interact with each other and fused into the Spatial Embedding $\mathbf{E}_S$, which replaces the $\mathbf{E}_{BEV}$ to provide more comprehensive spatial information of the driving scene to $\mathbf{F}_{BEV}$. It further increases the 0.56% mIoU. Another 0.41% mIoU is obtained by extending the training process from 24 epochs to 48 epochs. By joining BEV-CutMix to improve the data diversity, the mIoU achieves 37.93%, which is 5.85% higher than the FlashOcc, while just increasing 0.15 ms to the latency. Compared with the optimized baseline, the LightOcc still increases the performance by 2.93% mIoU, strongly demonstrating the effectiveness of our proposed components.

### 4.4.2. Lightweight Spatial Embedding

We conduct the ablation study of the Lightweight Spatial Embedding and show the results in Tab. 3. It can be found the activation function of depth distribution can largely affect the accuracy. This is because Softmax tends to centralize the weight into a single depth value, which does not

Table 2. Ablation study of our proposed component. Latencies are evaluated on a single RTX3090 GPU.

| Method | mIoU↑ | Latency↓ |
|---|---|---|
| FlashOcc (M1) [39] | 32.08 | 35.08 ms |
| + Baseline Optimzation | 35.00 | 33.51 ms |
| + Spatial-to-Channel | 36.30 | 34.81 ms |
| + Lightweight TPV Interaction | 36.86 | 35.23 ms |
| + Training for more epochs | 37.27 | 35.23 ms |
| + BEV-CutMix | 37.93 | 35.23 ms |

Table 3. Ablation study of Lightweight Spatial Embedding. "Depth Act" denotes the activation function used by depth distribution. "Mean" denotes to average the multiplied embeddings by the vanished dimensions. "Conv" denotes the number of convolution layers for each embedding extraction and interaction. All models are trained for 24 epochs without BEV-CutMix.

| Depth Act | Mean | Conv | mIoU↑ | Latency↓ |
|---|---|---|---|---|
| Softmax | ✗ | 1 | 36.11 | 35.23 ms |
| Sigmoid | ✗ | 1 | 36.71 | 35.23 ms |
| Sigmoid | ✓ | 1 | 36.86 | 35.23 ms |
| Sigmoid | ✓ | 2 | 36.92 | 36.40 ms |

Table 4. Ablation study of BEV-CutMix. "Cut X" and "Cut Y" denote cutting the occupancy from center along the X and Y dimensions respectively. "Mix Ratio" denotes the ratio of samples augmented by BEV-CutMix.

| Cut X | Cut Y | Mix Ratio | mIoU↑ |
|---|---|---|---|
| ✗ | ✗ | 0% | 37.27 |
| ✓ | ✗ | 50% | 37.90 |
| ✓ | ✗ | 100% | 37.93 |
| ✓ | ✓ | 100% | 37.72 |

match the representation required by occupancy prediction. On the contrary, Sigmoid maps each depth weight into [0,1], which can be regarded as the probability of being occupied. Besides, the multiplication between embeddings will vanish their common dimension and lead to the accumulation of values. Calculating the average of the values by the vanished dimension can balance the values of embeddings and bring 0.15% mIoU improvement. In addition, we find a single convolution layer is competent for embedding extraction and interaction. When utilizing two convolution layers, the mIoU is only improved by 0.06%, while the time cost is increased by 1.17 ms.

### 4.4.3. BEV-CutMix

We evaluate different settings of BEV-CutMix and show the results in Tab. 4. It can be found that cutting the BEV features from the center along the X dimension shows more importance than cutting along the Y dimension. The performance of the model is improved by 0.66% when cutting

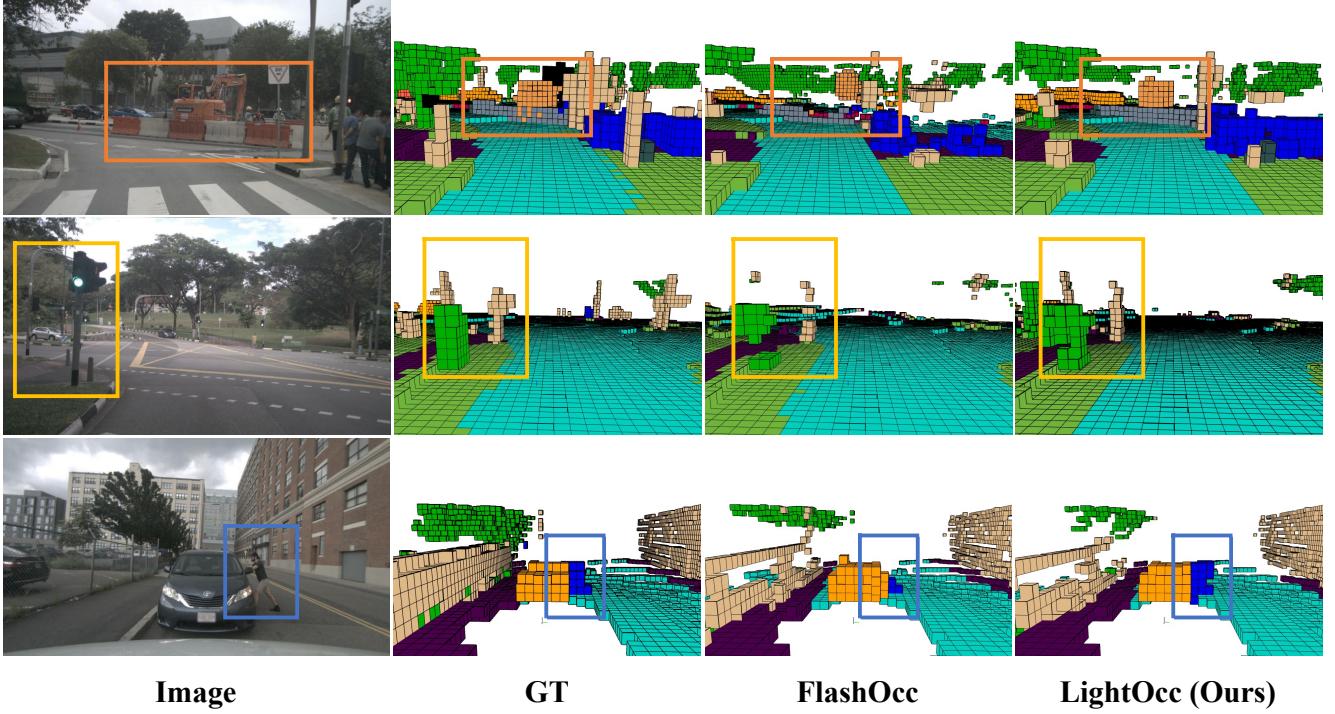|  |  |  |  |
|:---:|:---:|:---:|:---:|
| **Image** | **GT** | **FlashOcc** | **LightOcc (Ours)** |

Figure 6. Visualization results of LightOcc and FlashOcc [39]. The rectangles illustrate that Lightweight Spatial Embedding can effectively supplement the height clues to BEV features and significantly improve the accuracy of height in the occupancy prediction.

the occupancy into two parts along the X, but slightly declines when cutting the occupancy into four parts. This is probably because the X dimension generally coincides with the direction of the road and mixing such parts can generate more smooth and realistic scenes. We also adjust the ratio of the samples augmented by BEV-CutMix and find that it does not make much difference.

### 4.5. Visualization

In addition to quantitative analysis, we also qualitatively compare LightOcc with its baseline, i.e. FlashOcc, and show their occupancy predictions in Fig. 6. The obvious improvement in visualization is marked by rectangles. It can be found that the objects' height occupancy predicted by LightOcc is more precise than FlashOcc, such as the barriers in the first column and the pedestrian in the third column. Besides, the street lamp and traffic light in the second column have more complete structures in our prediction, while they are truncated in the prediction of FlashOcc. These qualitative results highlight the effectiveness of Lightweight Spatial Embedding in storing the height information of the driving scene and supplementing it to the BEV features for better prediction.

## 5. Conclusion

In this paper, we propose a novel vision-based occupancy prediction method, namely LightOcc. The main contribution of LightOcc lies in obtaining the Spatial Embedding of the whole driving scenes by lightweight module, which supplements the height clues lost by the BEV features. At first, we propose Global Spatial Sampling that accumulates the multi-view depth distribution into Single-Channel Occupancy. It can play the same role as voxel features, but requires much less time and memory. To quickly extract the spatial information, Spatial-to-Channel mechanism is proposed, which takes the arbitrary spatial dimension of Single-Channel Occupancy as the feature channel and extracts TPV Embeddings from different views by 2D convolution. Subsequently, Lightweight TPV Interaction module is proposed to efficiently combine the explicit and implicit height information of TPV Embeddings by matrix multiplication and 2D convolution. TPV Embeddings are finally fused into the unified Spatial Embedding, which can be directly added to the BEV features and boost the prediction performance. In addition, BEV-CutMix is proposed to increase the diversity of the data without introducing wrong occlusion in the generated scenes.

We conduct extensive experiments on Occ3D-nuScenes benchmark and the results indicate that LightOcc achieves

state-of-the-art performance, demonstrating the effectiveness of Lightweight Spatial Embedding in enhancing the height information carried by the BEV representation. Compared to its baseline, LightOcc significantly improves precision while keeping the high-efficiency characteristic, which means that LightOcc can be widely deployed in autonomous driving applications.

# References

[1] 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. 7, 1

[2] 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

[3] 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. 1, 3, 6, 7

[4] 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, 1

[5] Jiawei Hou, Xiaoyan Li, Wenhao Guan, Gang Zhang, Di Feng, Yuheng Du, Xiangyang Xue, and Jian Pu. Fastocc: Accelerating 3d occupancy prediction by fusing the 2d bird's-eye view and perspective view. *arXiv preprint arXiv:2403.02710*, 2024. 1, 3, 6

[6] Junjie Huang and Guan Huang. Bevdet4d: Exploit temporal cues in multi-camera 3d object detection. *arXiv preprint arXiv:2203.17054*, 2022. 2

[7] Junjie Huang and Guan Huang. Bevpoolv2: A cutting-edge implementation of bevdet toward deployment. *arXiv preprint arXiv:2211.17111*, 2022. 4, 1

[8] 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. 1, 6

[9] 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. 2, 3, 5, 6

[10] Yuanhui Huang, Wenzhao Zheng, Borui Zhang, Jie Zhou, and Jiwen Lu. Selfocc: Self-supervised vision-based 3d occupancy prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19946–19956, 2024. 3

[11] Yanqin Jiang, Li Zhang, Zhenwei Miao, Xiatian Zhu, Jin Gao, Weiming Hu, and Yu-Gang Jiang. Polarformer: Multi-camera 3d object detection with polar transformer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 1042–1050, 2023. 3

[12] Hongyang Li, Hao Zhang, Zhaoyang Zeng, Shilong Liu, Feng Li, Tianhe Ren, and Lei Zhang. Dfa3d: 3d deformable attention for 2d-to-3d feature lifting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6684–6693, 2023. 3

[13] Yinhao Li, Han Bao, Zheng Ge, Jinrong Yang, Jianjian Sun, and Zeming Li. Bevstereo: Enhancing depth estimation in multi-view 3d object detection with temporal stereo. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 1486–1494, 2023. 2, 6

[14] 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. 2

[15] 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. 3

[16] 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. 3

[17] 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, 3, 6

[18] Zhenxin Li, Shiyi Lan, Jose M Alvarez, and Zuxuan Wu. Bevnext: Reviving dense bev frameworks for 3d object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20113–20123, 2024. 2

[19] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021. 6

[20] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*. 6

[21] Yuhang Lu, Xinge Zhu, Tai Wang, and Yuexin Ma. Octreeocc: Efficient and multi-granularity occupancy prediction using octree queries. *arXiv preprint arXiv:2312.03774*, 2023. 6

[22] 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, 3, 6, 7

[23] Yuexin Ma, Tai Wang, Xuyang Bai, Huitong Yang, Yuenan Hou, Yaming Wang, Yu Qiao, Ruigang Yang, and Xinge

Zhu. Vision-centric bev perception: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024. 1

[24] Jiageng Mao, Shaoshuai Shi, Xiaogang Wang, and Hongsheng Li. 3d object detection for autonomous driving: A review and new outlooks. *arXiv preprint arXiv:2206.09474*, 1:1, 2022. 1

[25] Mingjie Pan, Jiaming Liu, Renrui Zhang, Peixiang Huang, Xiaoqi Li, Hongwei Xie, Bing Wang, Li Liu, and Shanghang Zhang. Renderocc: Vision-centric 3d occupancy prediction with 2d rendering supervision. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 12404–12411. IEEE, 2024. 3

[26] Jonah Philion and Sanja Fidler. Lift, splat, shoot: Encoding images from arbitrary camera rigs by implicitly unprojecting to 3d. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIV 16*, pages 194–210. Springer, 2020. 2, 4

[27] Rui Qian, Xin Lai, and Xirong Li. 3d object detection for autonomous driving: A survey. *Pattern Recognition*, 130: 108796, 2022. 1

[28] Yining Shi, Kun Jiang, Jiusi Li, Zelin Qian, Junze Wen, Mengmeng Yang, Ke Wang, and Diange Yang. Grid-centric traffic scenario perception for autonomous driving: A comprehensive review. *arXiv preprint arXiv:2303.01212*, 2023. 1

[29] Yiang Shi, Tianheng Cheng, Qian Zhang, Wenyu Liu, and Xinggang Wang. Occupancy as set of points. *arXiv preprint arXiv:2407.04049*, 2024. 6

[30] Sathira Silva, Savindu Bhashitha Wannigama, Roshan Ragel, and Gihan Jayatilaka. S2tpvformer: Spatio-temporal triperspective view for temporally coherent 3d semantic occupancy prediction. *arXiv preprint arXiv:2401.13785*, 2024. 2, 3

[31] Xin Tan, Wenbin Wu, Zhiwei Zhang, Chaojie Fan, Yong Peng, Zhizhong Zhang, Yuan Xie, and Lizhuang Ma. Geocc: Geometrically enhanced 3d occupancy network with implicit-explicit depth fusion and contextual self-supervision. *arXiv preprint arXiv:2405.10591*, 2024. 6

[32] 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, 6

[33] Li Wang, Xinyu Zhang, Ziying Song, Jiangfeng Bi, Guoxin Zhang, Haiyue Wei, Liyao Tang, Lei Yang, Jun Li, Caiyan Jia, et al. Multi-modal 3d object detection in autonomous driving: A survey and taxonomy. *IEEE Transactions on Intelligent Vehicles*, 8(7):3781–3798, 2023. 1

[34] 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. 1, 6

[35] 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. 3

[36] Yuan Wu, Zhiqiang Yan, Zhengxue Wang, Xiang Li, Le Hui, and Jian Yang. Deep height decoupling for precise vision-based 3d occupancy prediction. *arXiv preprint arXiv:2409.07972*, 2024. 2, 3, 6, 7

[37] Huaiyuan Xu, Junliang Chen, Shiyu Meng, Yi Wang, and Lap-Pui Chau. A survey on occupancy perception for autonomous driving: The information fusion perspective. *Information Fusion*, 114:102671, 2025. 1

[38] Chenyu Yang, Yuntao Chen, Hao Tian, Chenxin Tao, Xizhou Zhu, Zhaoxiang Zhang, Gao Huang, Hongyang Li, Yu Qiao, Lewei Lu, et al. Bevformer v2: Adapting modern image backbones to bird's-eye-view recognition via perspective supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17830–17839, 2023. 3

[39] 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. 2, 3, 4, 6, 7, 8, 1

[40] Jinqing Zhang, Yanan Zhang, Qingjie Liu, and Yunhong Wang. Sa-bev: Generating semantic-aware bird's-eye-view feature for multi-view 3d object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3348–3357, 2023. 2, 5

[41] Jinqing Zhang, Yanan Zhang, Yunlong Qi, Zehua Fu, Qingjie Liu, and Yunhong Wang. Geobev: Learning geometric bev representation for multi-view 3d object detection. *arXiv preprint arXiv:2409.01816*, 2024. 3, 4, 7, 1

[42] 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. 3, 6

[43] Yanan Zhang, Jinqing Zhang, Zengran Wang, Junhao Xu, and Di Huang. Vision-based 3d occupancy prediction in autonomous driving: a review and outlook. *arXiv preprint arXiv:2405.02595*, 2024. 1

# Lightweight Spatial Embedding for Vision-based 3D Occupancy Prediction

## Supplementary Material

This supplementary material provides more implementation details on LightOcc in Sec. A, more experiments results in Sec. B and visualization results in Sec. C.

## A. More Implementation Details

Following the general settings of the occupancy prediction methods [8], we apply data augmentation in both image space and BEV space. For image space, random scaling with a range of $[0.86, 1.25]$ and horizontal flipping with a probability of 0.5 are applied on the multi-view images. For BEV space, random scaling and random rotating are not employed because of the likely spatial misalignment on the low-resolution occupancy labels. In addition to BEV-CutMix, only random flipping with a probability of 0.5 is utilized. The size of BEV features obtained by view transformation matches the resolution of occupancy labels, which is $200 \times 200$. In LightOcc-S model, ResNet-18 [4] is employed as the BEV encoder to achieve the best efficiency, which is upgraded to ResNet-50 in LightOcc-L model to further improve the processing capability for BEV features. When utilizing history frames, the interval between each frame is 0.5s, which is consistent with previous methods.

## B. More Experiment Results

### B.1. Optimized Baseline

We optimize the M1 version of FlashOcc model [39] to obtain a stronger baseline. The effect of each optimization is shown in Tab. A. Replacing the original BEVPoolv2 [7] with RC-Sampling [41] can generate BEV with fine-grained geometric information and increases the mIoU by 0.62%. Besides, the former BEV FPN discards BEV features in $50 \times 50$ and utilizes too many convolution layers to fuse the $25 \times 25$ BEV features with $100 \times 100$ BEV features. On the contrary, we gradually fuse BEV features in different resolutions with fewer convolution layers, which reduces the latency by 1.21 ms and improves the accuracy. Furthermore, we employ auxiliary losses such as Scene-Class Affinity Loss [3] and Lovász-Softmax Loss [1] to assist the original Cross-Entropy Loss, which increase 1.17% mIoU without hindering the inference efficiency.

### B.2. BEV-CutMix

To avoid introducing wrong occlusion relationships in the generated scenes, BEV-CutMix cuts the BEV features from the center to keep the mixed visibility masks reasonable. As shown in Tab. B, cutting and mixing BEV features from random positions can increase the accuracy when compared

Table A. Details about the optimized baseline. "RC-Sampling" denotes using RC-Sampling for view transformation. "BEV FPN" denotes optimizing the structure of the BEV FPN. "Auxiliary Loss" denotes using other losses to assist Cross-Entropy Loss.

| RC-Sampling | BEV FPN | Auxiliary Loss | mIoU↑ | Latency |
|:---:|:---:|:---:|:---:|:---:|
| | | | 32.08 | 35.08 ms |
| ✓ | | | 32.70 | 34.72 ms |
| ✓ | ✓ | | 32.83 | 33.51 ms |
| ✓ | ✓ | ✓ | 35.00 | 33.51 ms |

Table B. Additional experiment about BEV-CutMix.

| Method | mIoU↑ |
|:---|:---:|
| w/o BEV-CutMix | 37.27 |
| Random Cutting and Mixing | 37.59 |
| Cutting and Mixing from Center | 37.93 |

with the model without BEV-CutMix. However, the wrong occlusion relationships damage the authenticity of the generated scenes and reduce the mIoU by 0.34%.

There is one thing to note when using BEV-CutMix in the model that fuses long-term temporal information. If cutting operation is applied to the BEV features of each frame, the mixed BEV features will have misalignments in content because of the movement of the ego car. As a result, BEV features of different frames are first transformed into the coordinate system of the current frame and concatenated into the final BEV features, which are then cut and mixed to create the representation of the new scene. The accuracy improvement of the LighOcc-L after inputting 8 history frames demonstrates the collaboration profit of BEV-CutMix and long-term temporal fusion.

## C. Visualization

More visualization comparison between the occupancy prediction of LightOcc and FlashOcc [39] is shown in Fig. A. In the first row, the cyclist in the scene is not represented in the occupancy predicted by FlashOcc, which poses a security risk. LightOcc, on the other hand, can perceive this crucial object precisely. Besides, the prediction of cross bar and pole in the second row indicates that LightOcc is good at predicting the structure of slender objects. In the night scene of the third row, FlashOcc fails to predict the road structure under the interference of the dazzling light, while LightOcc is unaffected. These visualization results demonstrate that LightOcc has higher perception ability and shows robustness in unconventional scenes.
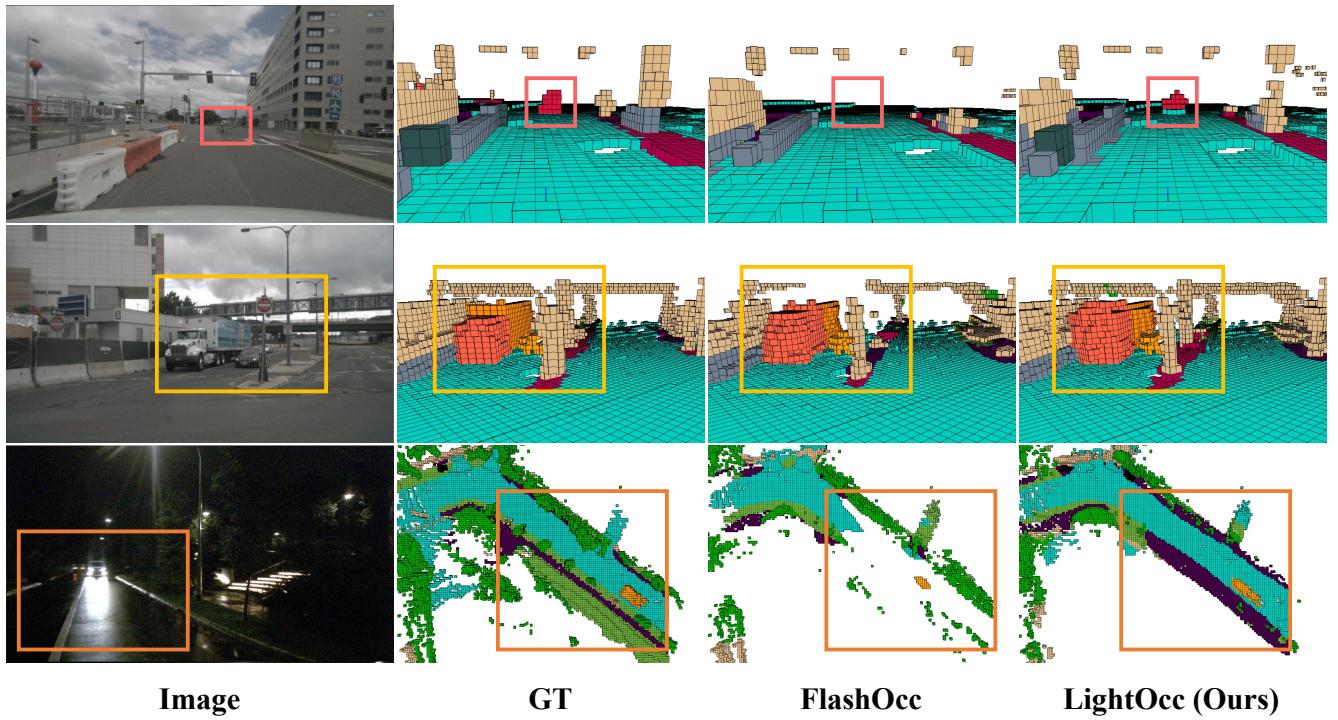
| **Image** | **GT** | **FlashOcc** | **LightOcc (Ours)** |

Figure A. More visualization results of LightOcc.