FairScene: Learning Class-Disentangled 2D/3D Representations for Semantic Scene Completion

Dian Jia[1] Pei Yu[2] Wei Tang[1]
[1]University of Illinois Chicago, Chicago, IL, USA
[2]Microsoft, Redmond, WA, USA
{djia7,tangw}@uic.edu, pei.yu@microsoft.com

Abstract

Semantic Scene Completion (SSC) aims to predict the semantic occupancy of each voxel within a 3D scene using sensor data, a critical task for autonomous driving and robotics. Despite recent progress, camera-based SSC remains challenging due to various difficulties, including voxel class imbalance, occlusion, and depth ambiguity. This paper introduces FairScene, a novel approach that learns class-disentangled 2D/3D representations to improve SSC. By ensuring balanced representations across classes, FairScene mitigates the dominance of majority classes and promotes fairer voxel categorization. Additionally, FairScene explicitly models spatial dependencies between different classes through a novel inter-class occupancy reasoning mechanism. Such explicit modeling helps alleviate occlusion and depth ambiguities in SSC. To address the scarcity of SSC training data, we propose OccMix, a novel augmentation strategy that generalizes MixUp from 2D to 2.5D and 3D metric spaces while maintaining geometric consistency. Extensive quantitative and qualitative experiments demonstrate that FairScene outperforms prior methods on both the SemanticKITTI and SSCBench-KITTI-360 benchmarks. The code is available at https://github.com/DianJJ/FairScene.

1. Introduction

The rapid development of autonomous driving poses new challenges for 3D scene understanding. To enhance safety in autonomous driving, Semantic Scene Completion (SSC) has been introduced to simultaneously predict the dense occupancy and semantics of a 3D scene using sensor data [9, 32, 34, 43], such as point clouds from LIDAR and RGB images from optical cameras [6, 14, 29, 49]. Compared to LIDAR-based solutions, camera-based SSC is more cost-effective and widely deployable, thereby drawing growing interest from both academia and industry.

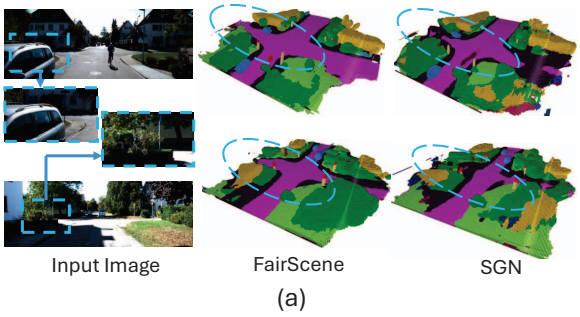


Figure 1. (a) FairScene effectively handles occluded scenes, accurately reconstructing road regions even when they are partially blocked. As shown in the image, FairScene infers the road behind occlusions (e.g., vehicles and vegetation) more accurately than its base model SGN [28]. (b) FairScene achieves a significant improvement on under-represented categories compared to its base model SGN. This result is reported on the SemanticKITTI hidden test set.

The typical pipeline for camera-based SSC includes image feature extraction, 2D-to-3D feature lifting, and voxel-wise semantic occupancy prediction. Many efforts have been made to enhance this pipeline by integrating geometric modeling with deep learning, *e.g.*, tri-perspective view encoding [14, 45], implicit stereo matching [16, 29], and normalized device coordinates [44]. Recent works have in-

troduced dedicated Transformer architectures [14, 15, 19, 22, 42] with sophisticated cross- and self-attention mechanisms. Despite notable progress, SSC remains a challenging task due to various difficulties, including voxel class imbalance, occlusion, and depth ambiguity.

This paper explores a fundamentally different approach to SSC, by learning class-disentangled representations from image space to the physical world. This approach offers two key benefits. First, it ensures balanced representations for each class, preventing majority classes from overshadowing minority ones in both 2D object recognition and 3D geometric processing, ultimately leading to fairer voxel categorization. Second, objects in the physical world are inherently related, for example, cars driving on the road, bicyclists riding bicycles, and traffic signs mounted on poles. Class-disentangled representations enable explicit modeling of these relational inductive biases within the deep network, which will help resolve ambiguities such as occlusion and depth uncertainty.

Our proposed approach, FairScene, introduces two new mechanisms. *(1) Class-Disentangled 2D-to-3D Representation Learning.* FairScene disentangles visual features extracted from the 2D backbone into class-specific representations and lifts them to 3D using depth estimates. Each 2D/3D class-specific representation encodes only the spatial and semantic information relevant to its corresponding class. This disentanglement is achieved through dense semantic guidance obtained by ray marching at target voxels. The disentangled representations promote class-balanced SSC and facilitate spatial relationship modeling between different classes. *(2) Inter-Class Occupancy Reasoning.* FairScene explicitly models interactions between 3D feature volumes of different classes by dynamically inferring each class's degree of dependency on others and using this information to guide inter-class spatial message passing. This mechanism helps resolve ambiguities in complex scenes while preserving class-balanced representations.

The limited size of SSC training data is another key challenge. Existing methods often use basic augmentations like cropping [10] or color jittering [41], or avoid augmentation due to difficulty in preserving 2D-3D geometric consistency. To address this, we propose OccMix, a generalization of MixUp [48] from 2D to 2.5D and 3D metric spaces. Unlike standard MixUp, which only mixes image pixels and labels, OccMix fuses samples across 2D pixels, 2.5D depth, and 3D occupancy while maintaining geometric consistency. This enables diverse, physically plausible augmentations that can be easily integrated into existing SSC frameworks.

The contributions of this paper are summarized below.

- We propose a novel approach that learns class-disentangled 2D/3D representations for SSC. It promotes balanced voxel categorization and facilitates spatial relationship modeling between different classes.
- We introduce a new mechanism that dynamically infers inter-class dependencies and explicitly models spatial interactions between different class-specific 3D feature volumes.
- We propose OccMix to generate physically plausible training data for SSC, by mixing data samples across 2D image space and 2.5D/3D metric spaces, while maintaining geometric consistency.
- Our approach outperforms previous state-of-the-art methods on SemanticKITTI and SSCBench-KITTI-360, with significant improvements on occluded areas and under-represented categories, as shown in Fig. 1. These results validate its effectiveness in learning balanced representations and alleviating geometric ambiguities in SSC.

2. Related Work

Camera-based Semantic Scene Completion. MonoScene [6] pioneered camera-based SSC with a feature lifting mechanism that samples RGB features along lines of sight. TPVFormer [14] encodes voxels via tri-perspective view planes, while VoxFormer [19] adopts a two-stage MAE-style framework. OccDepth [29] improves geometric projection using implicit stereo matching. Transformer-based methods like OccFormer [49] and NDC-Scene [44] introduce mask-wise prediction and normalized device coordinates to reduce geometric ambiguity. Recently, some methods have sought to address the under-represented categories. Symphonize [15] uses instance queries for better semantics and context modeling, while MonoOcc [50] introduces a privileged branch with knowledge distillation. In contrast, we propose a fundamentally different strategy: learning class-disentangled 2D-to-3D representations. While disentangled representations have been explored in 2D vision [24, 26, 35, 40], they are not applicable to 2D-to-3D lifting tasks.

Data Augmentation for Semantic Scene Completion. Data augmentation has proven crucial to improving model robustness and handling data imbalance in 3D perception tasks [7, 8, 21, 51]. Traditional approaches in SSC typically employ basic techniques such as random flipping and cropping. However, the unique challenges of SSC, particularly the strong projection constraints between 2D inputs and 3D labels, limit the applicability of conventional augmentation methods. Recent works in related 3D perception tasks have explored more sophisticated strategies: BEVDet [13] and BEVDepth [18] introduced specialized augmentation techniques for BEV representation, while works like [23] proposed geometry-aware augmentation methods for 3D object detection. However, developing effective augmentation strategies specifically for SSC remains challenging due to the need to maintain both geometric consistency and physical plausibility in the augmented data.

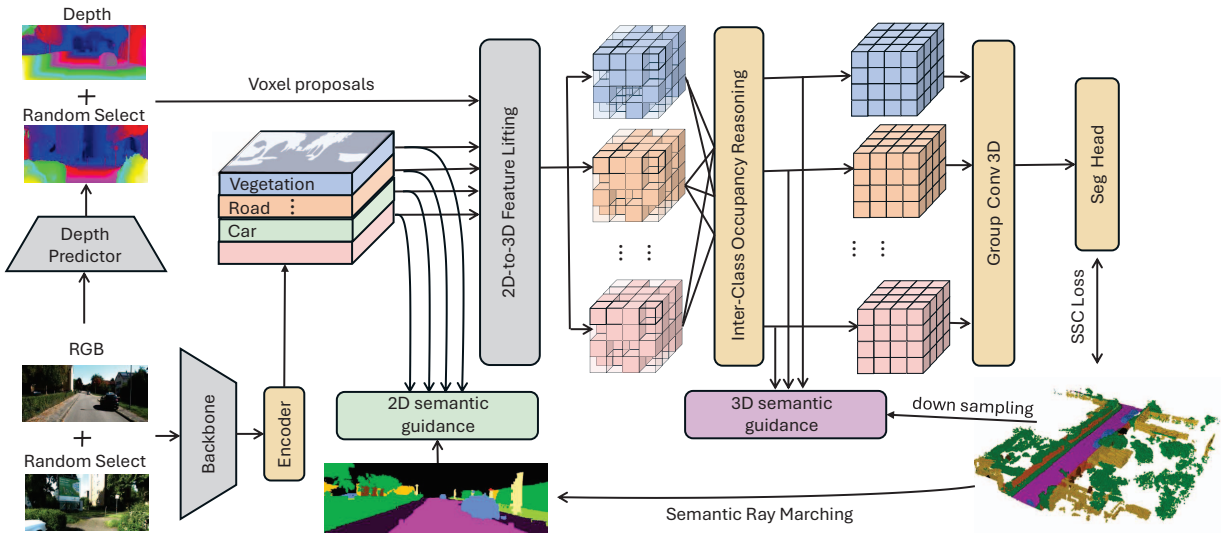


Figure 2. Overview of FairScene for Semantic Scene Completion. FairScene disentangles visual features extracted from the 2D backbone into class-specific representations and lifts them into 3D using depth estimates from an off-the-shelf stereo network. This disentanglement is guided by dense semantic guidance obtained by ray marching at target voxels. In 3D space, FairScene performs inter-class occupancy reasoning by dynamically inferring the degree of dependency between classes and leveraging this information to guide inter-class spatial message passing. Finally, a 3D group convolution block and a segmentation head generate voxel-wise classifications.

3. Method

We propose FairScene, a novel approach to semantic scene completion (SSC) by learning class-disentangled representations from image space to the physical world. As shown in Fig. 2, FairScene introduces two new mechanisms: *Class-Disentangled 2D-to-3D Representation Learning* and *Inter-Class Occupancy Reasoning*. Additionally, we propose *OccMix*, a geometry-preserving data augmentation technique tailored for SSC. In the rest of this section, we first present the two core components of FairScene in Sec. 3.1 and Sec. 3.2, respectively, and then elaborate OccMix in Sec. 3.3.

3.1. Class-Disentangled 2D-to-3D Representation Learning

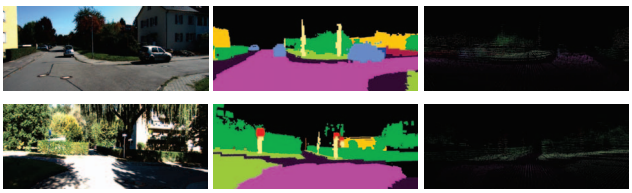
We disentangle visual features extracted from the 2D backbone into class-specific representations and lift them to 3D using depth estimates. Each 2D/3D class-specific representation encodes only the spatial and semantic information relevant to its corresponding class. This disentanglement is achieved through dense semantic guidance obtained by ray marching at target voxels. The disentangled representations promote class-balanced SSC and facilitate spatial relationship modeling between different classes.

Class-specific 2D Representations. A 2D backbone network first extracts a feature map from the input image. This feature map is then fed into two convolutional layers to produce a decoupled 2D representation $\boldsymbol{F}_k^{\text{2D}} \in \mathbb{R}^{Z\times H\times W}$ for each class k ($k \in 1,\dots,K$), where H and W are the height and width, Z is the number of channels, and K is the number of object classes. Let $\boldsymbol{F}^{\text{2D}} \in \mathbb{R}^{C\times H\times W}$ denote the collection of all class-specific representations, defined as $\boldsymbol{F}^{\text{2D}} = \texttt{Concat}(\boldsymbol{F}_1^{\text{2D}},\dots,\boldsymbol{F}_K^{\text{2D}})$, where $C = K \cdot Z$ and $\texttt{Concat}$ denotes concatenation along the channel dimension.

The decoupling of features is achieved through intermediate 2D semantic supervision, which transforms class-agnostic features into class-specific ones so that $\boldsymbol{F}_k^{\text{2D}}$ contains information specific to its corresponding class k. Concretely, we use each representation to predict a 2D segmentation mask for its respective class, which is then compared with 2D semantic guidance (described below) during training using a pixel-wise cross-entropy loss. Since $\boldsymbol{F}_k^{\text{2D}}$ is only responsible for predictions related to class k, it is guided to encode the spatial and semantic information of this class within the image.

Finally, we apply two *group* convolutional layers to $\boldsymbol{F}^{\text{2D}}$, each with K groups, to maintain and enhance class-specific features. For simplicity, we still use $\boldsymbol{F}^{\text{2D}} = \texttt{Concat}(\boldsymbol{F}_1^{\text{2D}},\dots,\boldsymbol{F}_K^{\text{2D}})$ to denote the enhanced class-specific representations.

Semantic Ray Marching. To obtain dense 2D semantic masks, a naive approach is to directly project ground truth voxel centers onto the image plane; however, this results in extremely sparse pixel annotations (see Fig. 3). Inspired by the neural radiation field (NeRF) [30, 38], we introduce Se-



(a) (b) (c)

Figure 3. (a) Input RGB image. (b) Segmentation mask obtained by Semantic Ray Marching. (c) Segmentation mask obtained by direct projection of voxel centers.

mantic Ray Marching (SRM). Instead of relying on sparse projections, SRM iteratively samples along each viewing ray, extracting semantic information from intersected voxels. This process generates continuous and robust 2D semantic guidance that effectively mitigates the discontinuity arising from sparse voxel projections.

For each pixel (u, v) in the image, we first compute the (unnormalized) ray direction using the camera's intrinsic matrix $\boldsymbol{E}$ and extrinsic matrix $\boldsymbol{R}$:

$$\boldsymbol{d}_{u,v} = \boldsymbol{R}^{-1}\boldsymbol{E}^{-1}\begin{bmatrix} u \\ v \\ 1 \end{bmatrix} \tag{1}$$

Then, the parametric equation for ray marching is given by:

$$\boldsymbol{r}_{u,v}(t) = \boldsymbol{c}^{\text{origin}} + t \cdot \frac{\boldsymbol{d}_{u,v}}{||\boldsymbol{d}_{u,v}||} \tag{2}$$

where $\boldsymbol{c}^{\text{origin}}$ is the camera origin, and t is the step parameter along the ray. At each step, we calculate the voxel indices as:

$$\boldsymbol{g}^{\text{index}} = \left\lfloor \frac{\boldsymbol{r}_{u,v}(t) - \boldsymbol{g}^{\text{origin}}}{\Delta g} \right\rfloor \tag{3}$$

where $\boldsymbol{g}^{\text{origin}}$ is the voxel grid origin, Δg is the voxel size.

The ray-marching process terminates upon encountering an occupied voxel, exiting the voxel grid, or reaching the step limit. If an occupied voxel is found, its semantic label is assigned to the corresponding pixel; otherwise, the pixel is labeled as empty. This process results in a dense 2D semantic segmentation, as shown in Fig. 3.

Class-Preserving 2D-to-3D Feature Lifting. We project 2D representations into 3D space in a class-preserving manner, resulting in a disentangled 3D representation for each class.

Each 2D feature map $\boldsymbol{F}_k^{\text{2D}} \in \mathbb{R}^{Z\times H\times W}$ ($k \in \{1, \ldots, K\}$) is first lifted to a 3D feature volume $\boldsymbol{F}_k^{\text{lift}} \in \mathbb{R}^{Z\times D'\times H'\times W'}$ via a common sparse-to-dense method [6], where pixels cast rays into 3D space, populating intersected voxels. D', H', and W' are the depth, height, and width of the target voxel grid, respectively. Following [19], voxel proposals are then generated by voxelizing depth-based pseudo LiDAR points [33] into an occupancy map $\mathbf{O} \in \{0, 1\}^{D'\times H'\times W'}$. The final class-disentangled 3D representation is obtained by selecting occupied voxels:

$$\boldsymbol{F}_k^{\text{3D}} = \{(\boldsymbol{f}_{k,\boldsymbol{p}}^{\text{lift}}, \boldsymbol{p}) : o_{\boldsymbol{p}} = 1\}, k = 1, \ldots, K \tag{4}$$

where $\boldsymbol{F}_k^{\text{3D}} \in \mathbb{R}^{Z\times D'\times H'\times W'}$ is a sparse 3D feature volume for class k, $\boldsymbol{p}$ is the 3D coordinate of a voxel, $o_{\boldsymbol{p}} \in \{0, 1\}$ and $\boldsymbol{f}_{k,\boldsymbol{p}}^{\text{lift}} \in \mathbb{R}^Z$ are the voxel's occupancy and 2D-to-3D lifted features from $\mathbf{O}$ and $\boldsymbol{F}_k^{\text{lift}}$ at position $\boldsymbol{p}$, respectively. We use a shared occupancy map across classes because it is both computationally efficient and preserves class-disentangled 2D/3D representations while allowing multiple semantic features to coexist at the same 3D location.

3.2. Inter-Class Occupancy Reasoning

Different classes in the physical world are inherently related, for example, `cars` driving on `roads`, `bicyclists` riding `bicycles`, and `traffic signs` mounted on `poles`. These relationships provide valuable priors that help resolve ambiguities in SSC, such as occlusion and depth uncertainty. Therefore, we propose a new inter-class occupancy reasoning mechanism that explicitly models spatial interactions between 3D feature volumes of different classes. It first dynamically infers each class's degree of dependency on others and then uses this information to guide inter-class spatial message passing.

Concretely, we formulate inter-class occupancy reasoning as follows:

$$\boldsymbol{h}_k = \texttt{AvePool3D}(\boldsymbol{F}_k^{\text{3D}}),\ k = 1, \ldots, K \tag{5}$$

$$\{a_{k,j}\}_{k,j=1}^{K} = \phi(\{\boldsymbol{h}_1, \ldots, \boldsymbol{h}_K\}) \tag{6}$$

$$\bar{\boldsymbol{F}}_k^{\text{3D}} = \boldsymbol{W}_k * \boldsymbol{F}_k^{\text{3D}} + \sum_{j=1}^{K} a_{k,j}\boldsymbol{W}_j * \boldsymbol{F}_j^{\text{3D}},\ k = 1, .., K \tag{7}$$

Eq. (5) uses 3D average pooling to extract a global representation $\boldsymbol{h}_k \in \mathbb{R}^Z$ from each class-specific 3D feature volume $\boldsymbol{F}_k^{\text{3D}}$. In Eq. (6), $a_{k,j} \in [0, 1]$ represents class k's dependency on class j, which is inferred from the global class representations. We model ϕ as a fully-connected layer followed by a softmax function. An alternative choice is the scaled dot-product attention [36], which we found to yield similar performance. In Eq. (7), $\bar{\boldsymbol{F}}_k^{\text{3D}}$ is the updated 3D feature volume for class k, and $*$ denotes (sparse) 3D convolution with $\boldsymbol{W}_k$ and $\boldsymbol{W}_j$ as the convolutional kernels.

Eq. (7) enables interactions between class-specific 3D volume representations based on their inter-class dependencies. If $a_{k,j}$ is high, which indicates class k strongly depends on class j, then class j's representation contributes more to the update of class k's representation. Interestingly, this operation could be viewed as a generalization

of graph convolution [2] from a 1D vector space to a high-dimensional tensor space (*e.g.*, 3D volume representations), enabling effective *spatial* message passing between different classes.

To preserve class-disentangled 3D representations, we apply 3D semantic guidance to the generated $\bar{\boldsymbol{F}}^{3D} = \texttt{Concat}(\bar{\boldsymbol{F}}_1^{3D}, \ldots, \bar{\boldsymbol{F}}_K^{3D})$. Concretely, we feed each $\bar{\boldsymbol{F}}_k^{3D}$ ($k \in \{1, \ldots, K\}$) to a $1 \times 1 \times 1$ convolutional layer for semantic occupancy prediction, which can be implemented easily using group convolution. The prediction is supervised by a voxel-wise classification loss.

Following [28], we construct a dense 3D volume representation by filling unoccupied voxels in the sparse volume $\bar{\boldsymbol{F}}^{3D}$ with a learnable vector $\boldsymbol{m} \in \mathbb{R}^C$:

$$\hat{\boldsymbol{f}}_{\boldsymbol{p}}^{3D} = \begin{cases} \bar{\boldsymbol{f}}_{\boldsymbol{p}}^{3D} & \text{if } o_{\boldsymbol{p}} = 1 \\ \boldsymbol{m} & \text{if } o_{\boldsymbol{p}} = 0 \end{cases} \tag{8}$$

where $\hat{\boldsymbol{f}}_{\boldsymbol{p}}^{3D}$ is the features of the dense 3D volume $\hat{\boldsymbol{F}}^{3D}$ at voxel position $\boldsymbol{p}$.

Finally, we obtain the SSC prediction through a 3D group convolution block, followed by a segmentation head:

$$\hat{\boldsymbol{Y}} = \texttt{SegHead}(\texttt{GroupConv3D}(\hat{\boldsymbol{F}}^{3D})) \tag{9}$$

where $\hat{\boldsymbol{Y}} \in [0, 1]^{N \times D' \times H' \times W'}$ is the voxel-wise classification output, and N is the number of categories.

Loss Functions. For the final prediction, we use the scene-class affinity loss and a voxel-wise cross-entropy loss following previous work [6]. For intermediate supervision on the class-specific representations, we take the pixel-wise cross-entropy loss as 2D semantic guidance and a combination of the voxel-wise cross-entropy loss and the lovasz loss [4] as the 3D semantic guidance. Detailed formulations of our loss functions are in the supplementary material.

3.3. OccMix

Traditional augmentation methods like random cropping and affine transforms are limited in SSC due to the strict geometric alignment required between images and voxels [8, 25]. Approaches such as CutMix [47] depend on accurate depth to insert objects consistently across scenes, which is often unavailable in camera-based SSC, leading to geometric inconsistencies. To address this, we propose OccMix—an extension of MixUp from 2D to 2.5D and 3D metric spaces. Unlike vanilla MixUp [37, 47], which breaks 3D projection rules when mixing images with different intrinsics, OccMix preserves geometric consistency across all dimensions. As illustrated in Fig. 4, it comprises three key components: *View-Occupancy Mixing*, *Voxel Proposal Fusion*, and *Mixed Feature Lifting*.

View-Occupancy Mixing. Images captured by different cameras have distinct geometric projection properties due to varying intrinsic parameters (focal length and principal point). To maintain geometric consistency and avoid depth ambiguity, we only mix images sharing identical intrinsic camera parameters.

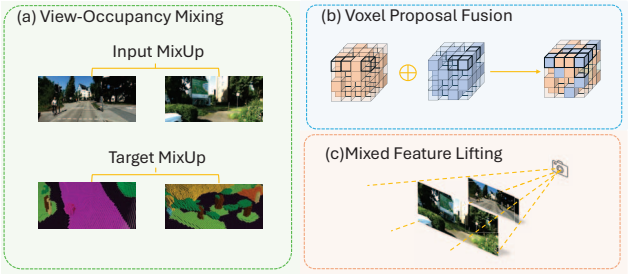


Figure 4. Illustration of the three components of OccMix. See Sec. 3.3 for explanations.

Let $(\boldsymbol{I}_i, \boldsymbol{Y}_i)$ denote a training sample, where $\boldsymbol{I}_i$ is the input image and $\boldsymbol{Y}_i$ is the target semantic occupancy map. We randomly choose another training sample $(\boldsymbol{I}_j, \boldsymbol{Y}_j)$ with the same intrinsic camera matrix. The View-Occupancy Mixing is then defined as:

$$\begin{cases} \tilde{\boldsymbol{I}}_i = \lambda \cdot \boldsymbol{I}_i + (1 - \lambda) \cdot \boldsymbol{I}_j \\ \tilde{\boldsymbol{Y}}_i = \lambda \cdot \boldsymbol{Y}_i + (1 - \lambda) \cdot \boldsymbol{Y}_j \end{cases} \tag{10}$$

where λ denotes the mix ratio, and $(\tilde{\boldsymbol{I}}_i, \tilde{\boldsymbol{Y}}_i)$ is the mixed training sample.

Voxel Proposal Fusion. Generate voxel proposals from a pre-estimated depth map is a common process in SSC [15, 19, 28], where each input image has its own set of voxel proposals to encode visible voxel features. We adopt a simple yet effective fusion strategy by taking the union of voxel proposals from two input samples to form the voxel proposals for the mixed sample. This ensures that informative features from both training samples are selected.

Mixed Feature Lifting. Directly lifting mixed 2D features to 3D is problematic due to differences in viewpoints. To address this, we perform feature lifting twice using each camera's extrinsic parameters individually, and then combine the resulting 3D features using the same mixing ratio as in the 2D mixing process. This approach ensures correct spatial mapping while preserving geometric relationships from both views.

4. Experiments

4.1. Dataset and Metrics

We evaluate FairScene on two standard outdoor SSC benchmarks, SemanticKITTI [3] and SSCBench-KITTI-360 [20], both containing real-world urban driving scenes with 19–20 semantic classes. Following the common practice in prior work, we voxelize a 51.2m $\times$ 51.2m $\times$ 6.4m volume into a $256 \times 256 \times 32$ grid and report Intersection-over-Union

Method	Input	Supervision	IoU	mIoU	road (15.30%)	sidewalk (11.13%)	parking (1.12%)	other-grnd. (0.56%)	building (14.1%)	car (3.92%)	truck (0.16%)	bicycle (0.03%)	motorcycle(0.03%)	other-veh. (0.20%)	vegetation (39.3%)	trunk (0.51%)	terrain (9.17%)	person (0.07%)	bicyclist (0.07%)	motorcyclist (0.05%)	fence (3.90%)	pole (0.29%)	traf.-sign (0.08%)
RenderOcc [31]	Multi	Voxel + Lidar	-	12.87	57.2	28.44	16.11	0.91	18.18	24.90	6.03	3.11	0.00	3.66	26.23	4.87	33.61	1.91	3.11	0.00	9.10	6.24	3.38
CGFormer [46]	Stereo	Voxel + Lidar	44.41	16.63	64.30	34.20	34.10	12.10	25.80	26.10	4.30	3.70	1.30	2.70	24.50	11.20	29.30	1.70	3.60	0.40	18.70	8.70	9.30
MonoScene [6]	Mono	Voxel	34.16	11.08	54.70	27.10	24.80	5.70	14.40	18.80	3.30	0.50	0.70	4.40	14.90	2.40	19.50	1.00	1.40	0.40	11.10	3.30	2.10
TPVFormer [14]	Mono	Voxel	34.25	11.26	55.10	27.20	27.40	6.50	14.80	19.20	3.70	1.00	0.50	2.30	13.90	2.60	20.40	1.10	2.40	0.30	11.00	2.90	1.50
SurroundOcc [42]	Mono	Voxel	34.72	11.86	56.90	28.30	30.20	6.80	15.20	20.60	1.40	1.60	1.20	4.40	14.90	3.40	19.30	1.40	2.00	0.10	11.30	3.90	2.40
OccFormer [49]	Mono	Voxel	34.53	12.32	55.90	30.30	**31.50**	6.50	15.70	21.60	1.20	1.50	1.70	3.20	16.80	3.90	21.30	2.20	1.10	0.20	11.90	3.80	3.70
VoxFormer [19]	Stereo	Voxel	42.95	12.20	53.90	25.30	21.10	5.60	19.80	20.80	3.50	1.00	0.70	3.70	22.40	7.50	21.30	1.40	2.60	0.20	11.10	5.10	4.90
SGN [28]	Stereo	Voxel	41.88	14.01	57.80	29.20	27.70	5.20	23.90	**24.90**	2.70	0.40	0.30	4.00	24.20	10.00	25.80	1.10	2.50	0.30	14.20	7.40	4.40
Symphonize [15]	Stereo	Voxel	42.19	15.04	**58.40**	29.30	26.90	**11.70**	24.70	23.60	3.20	**3.60**	2.60	5.60	24.20	10.00	23.10	**3.20**	1.90	**2.00**	16.10	7.70	8.00
HASSC [39]	Stereo	Voxel	**43.40**	13.34	54.60	27.70	23.80	6.20	21.10	22.80	4.70	1.60	1.00	3.90	23.80	8.50	23.30	1.60	4.00	0.30	13.10	5.80	5.50
MonoOcc [50]	Stereo	Voxel	-	13.80	55.20	27.80	25.10	9.70	21.40	23.20	5.20	2.20	1.50	5.40	24.00	8.70	23.00	1.70	2.00	0.20	13.40	5.80	6.40
FairScene (ours)	Stereo	Voxel	43.00	**15.76**	57.90	**30.60**	28.10	6.60	**24.80**	24.10	**7.60**	3.50	**2.70**	**6.10**	**24.70**	**12.00**	**26.00**	2.70	**4.40**	1.80	**16.30**	**9.90**	**9.60**

Table 1. Results evaluated on SemanticKITTI hidden **test** set. The method with the best performance is showcased in **bold** and the second best is showcased in underline.

Method	Input	Supervision	IoU	mIoU	car (2.85%)	bicycle (0.01%)	motorcycle (0.01%)	truck (0.16%)	other-veh. (5.75%)	person (0.02%)	road (14.98%)	parking (2.32%)	sidewalk (6.43%)	other-grnd. (2.05%)	building (15.67%)	fence (0.96%)	vegetation (41.99%)	terrain (7.10%)	pole (0.22%)	traf.-sign (0.06%)	other-struct. (4.33%)	other-obj. (0.28%)
LMSCNet [32]	Lidar	Voxel	47.35	13.65	20.91	0.00	0.00	0.26	0.58	0.00	62.95	13.51	33.51	0.20	43.67	0.33	40.01	26.80	0.00	0.00	3.63	0.00
SSCNet [34]	Lidar	Voxel	53.58	16.95	31.95	0.00	0.17	10.29	0.00	0.07	65.70	17.33	41.24	3.22	44.41	6.77	43.72	28.87	0.78	0.75	8.69	0.67
CGFormer [46]	Stereo	Voxel + Lidar	48.07	20.05	29.85	3.42	3.96	17.59	6.79	6.63	63.85	17.15	40.72	5.53	42.73	8.22	38.80	24.94	16.24	17.45	10.18	6.77
MonoScene [6]	Mono	Voxel	37.87	12.31	19.34	0.43	0.58	8.02	2.03	0.86	48.35	11.38	28.13	3.32	32.89	3.53	26.15	16.75	6.92	5.67	4.20	3.09
TPVFormer [14]	Mono	Voxel	40.22	13.64	21.56	1.09	1.37	8.06	2.57	2.38	52.99	11.99	31.07	3.78	34.83	4.80	30.08	17.52	7.46	5.86	5.48	2.70
OccFormer [49]	Mono	Voxel	40.27	13.81	22.58	0.66	0.26	9.89	3.82	2.77	54.30	13.44	31.53	3.55	36.42	4.80	31.00	19.51	7.77	8.51	6.95	4.60
VoxFormer [19]	Stereo	Voxel	38.76	11.91	17.84	1.16	0.89	4.56	2.06	1.63	47.01	9.67	27.21	2.89	31.18	4.97	28.99	14.69	6.51	6.92	3.79	2.43
SGN [28]	Stereo	Voxel	46.22	17.71	28.20	2.09	3.02	11.95	3.68	4.20	59.49	**14.50**	36.53	4.24	39.79	7.14	36.61	23.10	14.86	**16.14**	8.24	4.95
Symphonize [15]	Stereo	Voxel	44.12	18.58	**30.02**	1.85	5.90	**25.07**	**12.06**	**8.20**	54.94	13.83	32.76	**6.93**	35.11	**8.58**	**38.33**	11.52	14.01	9.57	**14.44**	**11.28**
FairScene (ours)	Stereo	Voxel	**47.79**	**19.57**	29.60	**3.14**	**7.13**	21.65	8.80	3.31	**61.13**	14.13	**38.37**	5.26	**41.9**	6.9	37.91	**23.71**	**16.07**	15.82	9.32	8.08

Table 2. Results evaluated on SSCBench-KITTI-360 **test** set. The method with the best performance is showcased in **bold** and the second best is showcased in underline.

(IoU) and mean IoU (mIoU) over all semantic classes except the 'unlabeled' category.

4.2. Implementation Details

We crop cam2 RGB images to 1220 × 370 for SemanticKITTI and cam1 images to 1408 × 376 for SSCBench-KITTI-360. The ResNet-50 [12] backbone and image encoder are initialized with MaskDINO [17] pretrained weights. Each group has 16 channels, yielding 320 and 304 total channels for the two datasets (20 and 19 categories, respectively). Despite exceeding the typical 128-channel design, the use of group convolution actually leads to a smaller number of parameters and flops [11]. FairScene is trained for 48 epochs on 4 L40S GPUs (batch size 4) using AdamW [27] with a learning rate of 2e-4 and weight decay of 1e-2. In OccMix, the mix ratio λ is sampled from $\mathrm{Beta}(\alpha, \alpha)$ with $\alpha = 0.75$. During training, OccMix is applied to 30% of samples to balance original and mixed data for stable convergence. We do not apply OccMix during inference.

4.3. Quantitative Results

We compare FairScene with state-of-the-art methods on SemanticKITTI and SSCBench-KITTI-360. On the SemanticKITTI hidden test set (Tab. 1), under stereo input and voxel supervision, FairScene achieves an mIoU of 15.76 and an IoU of 43.00, outperforming all other methods in mIoU and ranking second in IoU.

(Acc / Rec / F1)	SemanticKITTI		KITTI360	
	Symphonize	FairScene (Ours)	Symphonize	FairScene (Ours)
other vehicle	68.07/4.29/8.07	60.58/12.89/21.27	—	—
truck	95.14/0.81/1.61	89.01/4.39/8.37	92.78/4.94/9.38	91.70/6.00/11.26
traffic sign	36.56/25.20/29.84	37.46/30.70/33.75	46.73/35.37/40.26	45.33/41.23/43.18
person	56.74/2.10/4.05	41.15/26.75/32.44	67.58/12.79/21.51	55.23/23.37/32.84
bicyclist	78.87/0.72/1.43	50.94/26.01/34.42	—	—
motorcycle	82.44/1.11/2.19	80.74/3.90/7.44	86.37/4.27/8.14	79.36/13.04/22.40
motorcyclist	86.71/12.66/22.08	90.00/8.51/15.55	—	—
bicycle	57.33/2.04/3.96	47.70/15.01/22.83	74.36/12.82/21.87	71.75/14.34/23.90

Table 3. (Acc / Rec / F1) on SemanticKITTI and KITTI360.

Method	SemanticKITTI			KITTI360		
	F1 ↑	Acc ↑	Rec ↑	F1 ↑	Acc ↑	Rec ↑
SGN	12.16	11.28	17.79	12.63	9.93	25.36
Symphonize	13.26	12.58	17.05	14.28	11.43	26.94
FairScene (Ours)	**15.77**	**15.30**	**20.58**	**17.21**	**13.43**	**34.36**

Table 4. Occlusion Evaluation on SemanticKITTI and KITTI360.

Method	SemanticKITTI		KITTI360	
	Abs Rel ↓	RMSE ↓	Abs Rel ↓	RMSE ↓
SGN	0.3690	10.4465	0.4084	11.9624,
Symphonize	0.3457	9.1756	0.3852	10.3433,
FairScene (Ours)	**0.2263**	**8.7352**	**0.2695**	**9.2651**

Table 5. Depth Evaluation on SemanticKITTI and KITTI360.

FairScene also achieves overall better performance on under-represented categories (frequency $< 0.2\%$), including *other vehicle*, *truck*, *traffic sign*, *person*, *bicyclist*, *motorcycle*, *motorcyclist*, and *bicycle*. Although CGFormer [46] uses additional LiDAR supervision and 120% more parameters (122.4M vs 54.1M), FairScene outperforms CGFormer relatively by **42.4%** on under-represented categories. Compared to MonoOcc [50] and Symphonize [15], specifically designed to address under-represented categories, FairScene achieves **27.7%** and **56.1%** relative improvements in mIoU, respectively.

As shown in Tab. 2, FairScene also achieves strong results on SSCBench-KITTI-360 with 19.57 mIoU and 47.79 IoU. Remarkably, despite relying only on stereo input, it matches or exceeds several LiDAR-based methods in semantic accuracy, which typically benefit from precise depth.

Further analysis on under-represented categories. The mIoU metric often fluctuates on low-frequency classes due to its sensitivity to localization and depth errors—minor voxel shifts can cause large drops in IoU. To further assess model performance on under-represented categories, we adopt precision, recall, and F1 score. Tab. 3 compares FairScene with Symphonize [15] on classes with frequencies below 0.2%. While Symphonize achieve high precision, they suffer from very low recall (e.g., person: 2.10%), indicating poor coverage. In contrast, FairScene maintains similar precision while significantly improving recall and F1 score, demonstrating the effectiveness of our class-balanced 2D/3D representations in addressing class-imbalance challenges and achieving more robust performance.

Occluded regions. We also evaluate the performance of our method in occluded regions. We define occluded voxels as those that are not reachable by ray marching. Tab. 4 reports the class-wise mean F1 score, accuracy, and recall on these occluded voxels. The results show that FairScene consistently outperforms both SGN and Symphonize in all three metrics across both datasets, demonstrating its superior ability to recover occluded structures.

Depth ambiguities. We compute depth errors by tracing rays to the first occupied voxel and comparing its depth with LiDAR ground truth. Tab. 5 reports the results using two standard depth metrics, Abs Rel and RMSE [1, 5], showing that our method achieves lower depth errors than both Symphonize and our base model SGN.

4.4. Ablation Studies

We conduct ablation studies on the SemanticKITTI validation set to analyze the effectiveness of different components and each step in our proposed OccMix.

Ablation on architectural components. Tab. 6 shows the impact of key components in FairScene: CRL (Class-disentangled 2D-to-3D Representation Learning), IOR (Inter-Class Occupancy Reasoning), and SG (2D/3D Semantic Guidance). OccMix further boosts performance by enhancing sample diversity while maintaining geometric consistency.

CRL uses group convolution to learn class-specific features, but yields limited mIoU improvement alone. Introducing 2D SG provides stronger supervision and leads to a notable gain. IOR improves sparse 3D reasoning and enables inter-class interaction. Even without 3D SG, IOR increases mIoU by 0.97. Adding 3D SG brings an additional 0.42 improvement.

These results validate the complementary benefits of all components in FairScene.

Ablation on OccMix components. Tab. 7 compares the performance across OccMix components. Vanilla MixUp mixes an input image with a randomly selected sample and their corresponding targets, but uses only the original image's camera matrix for projection, ignoring 2D-to-3D geometry and thus degrading performance. We observe that selecting an image with the correct camera parameters and applying Mixed Feature Lifting with its corresponding camera matrix improves model performance. This improvement is due to View-Occupancy Mixing and Mixed Feature Lifting, which ensure that features from each viewpoint are accurately mapped to their respective 3D positions, maintaining spatial consistency and enhancing overall performance

Without Voxel Proposal Fusion, the model fails to incorporate voxel proposals from the secondary image, resulting in incorrect voxel proposals. Adding this component notably improves mIoU by better fusing proposals from both views, thus reducing geometric ambiguity.

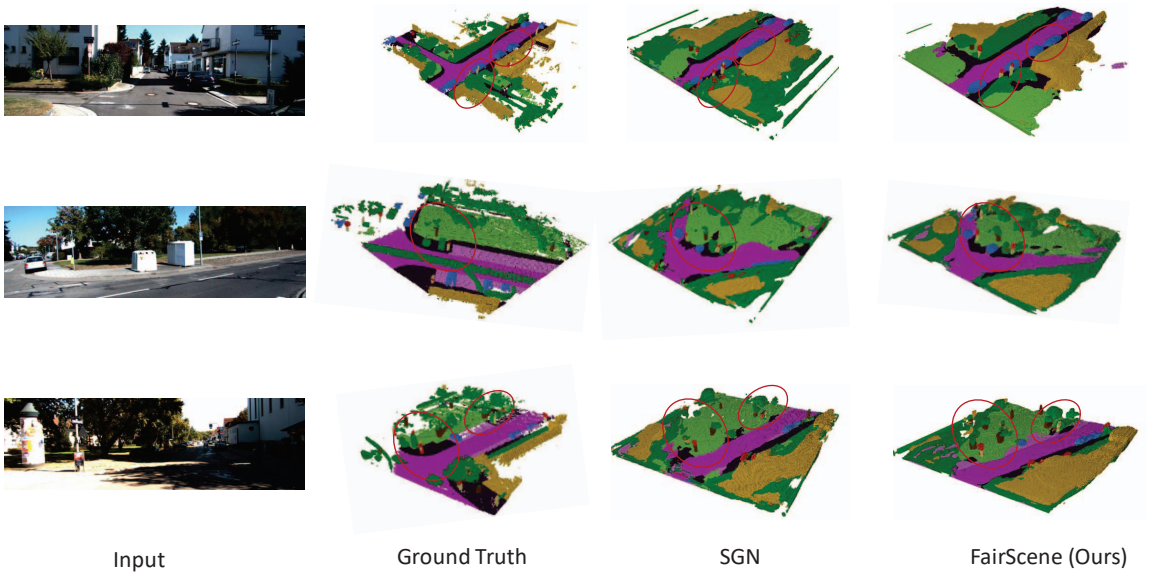


Figure 5. **Qualitative visualizations on SemanticKITTI val**. FairScene produces detailed predictions compared with its base model SGN. **Row 1**: The red circles indicate areas where FairScene better delineates traffic signs and vehicles, preserving clear boundaries between adjacent objects. **Row 2**: The red circles emphasize slender objects such as poles, where FairScene captures fine structural details that SGN tends to blur or merge with the background. **Row 3**: The red circles highlight tree trunks and poles, demonstrating FairScene's ability to produces detailed predictions, while maintaining coherent layouts.

OccMix	CRL	2D SG	IOR	3D SG	IoU	mIoU
					40.43	13.15
✓					41.40	13.82
✓	✓				41.60	14.01
✓	✓	✓			42.12	14.67
✓	✓	✓	✓		43.34	15.64
✓	✓	✓	✓	✓	**43.63**	**16.06**

Table 6. Ablation study on architectural components in FairScene. Results are reported on SemanticKITTI **val** set. **CRL**: Class-Disentangled 2D-to-3D Representatation Learning, **IOR**: Inter-Class Occupancy Reasoning, **SG**: Semantic Guidance.

Method	IoU	mIoU
w/o OccMix	42.79	15.66
w/ Vanilla MixUp	42.72	15.52
+ View-Occupancy Mixing	43.39	15.79
+ Mixed Feature Lifting	43.60	15.85
+ Voxel Proposal Fusion	**43.63**	**16.06**

Table 7. Performance comparison with different steps of OccMix. Results are reported on SemanticKITTI **val** set.

4.5. Qualitative Results

Fig. 5 showcases the visualizations on the SemanticKITTI validation set, comparing our method to SGN [28]. Our approach demonstrates significant improvements, particularly for the under-represented categories. These categories are often overlooked in traditional methods due to limited contextual cues and imbalanced representation in the dataset. Our method also performs very well in handling occluded regions. When dealing with roads occluded by vegetation, our model accurately predicts the extended areas. Additional qualitative comparisons with Symphonize [15] are included in the supplementary material.

5. Limitations

While FairScene has demonstrated strong overall performance in benchmark evaluations, it does not achieve the best results across all classes, suggesting that our proposed approach may complement existing methods. In future work, we aim to extend this research to a broader range of architectures to further enhance semantic scene completion.

6. Conclusion

This paper introduces a new approach, FairScene, for camera-based semantic scene completion. By learning class-disentangled 2D-to-3D representations, inter-class occupancy reasoning, and introducing a geometry-preserving data augmentation technique, our approach achieves substantial performance gains over existing methods on the SemanticKITTI and SSCBench-KITTI-360 datasets.

Acknowledgements. This work was supported in part by National Science Foundation (NSF) grants ECCS-2400900 and IIS-2442540, the National Artificial Intelligence Research Resource (NAIRR) Pilot, Amazon Web Services (AWS) provided through CloudBank, and NCSA Delta GPU resources provided through ACCESS.

References

[1] Vasileios Arampatzakis, George Pavlidis, Nikolaos Mitianoudis, and Nikos Papamarkos. Monocular depth estimation: A thorough review. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(4):2396–2414, 2023. 7

[2] Peter W Battaglia, Jessica B Hamrick, Victor Bapst, Alvaro Sanchez-Gonzalez, Vinicius Zambaldi, Mateusz Malinowski, Andrea Tacchetti, David Raposo, Adam Santoro, Ryan Faulkner, et al. Relational inductive biases, deep learning, and graph networks. *arXiv preprint arXiv:1806.01261*, 2018. 5

[3] Jens Behley, Martin Garbade, Andres Milioto, Jan Quenzel, Sven Behnke, Cyrill Stachniss, and Jurgen Gall. Semantickitti: A dataset for semantic scene understanding of lidar sequences. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019. 5

[4] 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. 5

[5] Aleksei Bochkovskii, AmaÃĢl Delaunoy, Hugo Germain, Marcel Santos, Yichao Zhou, Stephan R Richter, and Vladlen Koltun. Depth pro: Sharp monocular metric depth in less than a second. *arXiv preprint arXiv:2410.02073*, 2024. 7

[6] 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, 2, 4, 5, 6

[7] Hansheng Chen, Yuyao Huang, Wei Tian, Zhong Gao, and Lu Xiong. Monorun: Monocular 3d object detection by reconstruction and uncertainty propagation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10379–10388, 2021. 2

[8] Yongjian Chen, Lei Tai, Kai Sun, and Mingyang Li. Monopair: Monocular 3d object detection using pairwise spatial relationships. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12093–12102, 2020. 2, 5

[9] Ran Cheng, Christopher Agia, Yuan Ren, Xinhai Li, and Bingbing Liu. S3cnet: A sparse semantic scene completion network for lidar point clouds. *ArXiv*, abs/2012.09242, 2020. 1

[10] Ran Cheng, Christopher Agia, Yuan Ren, Xinhai Li, and Liu Bingbing. S3cnet: A sparse semantic scene completion network for lidar point clouds. In *Conference on Robot Learning*, pages 2148–2161. PMLR, 2021. 2

[11] Taco Cohen and Max Welling. Group equivariant convolutional networks. In *International conference on machine learning*, pages 2990–2999. PMLR, 2016. 6

[12] 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

[13] 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

[14] 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. 1, 2, 6

[15] Haoyi Jiang, Tianheng Cheng, Naiyu Gao, Haoyang Zhang, Tianwei Lin, Wenyu Liu, and Xinggang Wang. Symphonize 3d semantic scene completion with contextual instance queries. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20258–20267, 2024. 2, 5, 6, 7, 8

[16] Bohan Li, Yasheng Sun, Xin Jin, Wenjun Zeng, Zheng Zhu, Xiaoefeng Wang, Yunpeng Zhang, James Okae, Hang Xiao, and Dalong Du. Stereoscene: Bev-assisted stereo matching empowers 3d semantic scene completion. *arXiv preprint arXiv:2303.13959*, 1(3):6, 2023. 1

[17] Feng Li, Hao Zhang, Huaizhe Xu, Shilong Liu, Lei Zhang, Lionel M Ni, and Heung-Yeung Shum. Mask dino: Towards a unified transformer-based framework for object detection and segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3041–3050, 2023. 6

[18] 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

[19] 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. 2, 4, 5, 6

[20] Yiming Li, Sihang Li, Xinhao Liu, Moonjun Gong, Kenan Li, Nuo Chen, Zijun Wang, Zhiheng Li, Tao Jiang, Fisher Yu, Yue Wang, Hang Zhao, Zhiding Yu, and Chen Feng. Sscbench: A large-scale 3d semantic scene completion benchmark for autonomous driving, 2024. 5

[21] Zhuoling Li, Zhan Qu, Yang Zhou, Jianzhuang Liu, Haoqian Wang, and Lihui Jiang. Diversity matters: Fully exploiting depth clues for reliable monocular 3d object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2791–2800, 2022. 2

[22] 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. 2

[23] Zhenjia Li, Jinrang Jia, and Yifeng Shi. Monolss: Learnable sample selection for monocular 3d detection. In *2024 International Conference on 3D Vision (3DV)*, pages 1125–1135. IEEE, 2024. 2

[24] Mingyuan Liu, Dan Schonfeld, and Wei Tang. Exploit visual dependency relations for semantic segmentation. In *Proceed-*

ings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pages 9726–9735, 2021. 2

[25] Xianpeng Liu, Nan Xue, and Tianfu Wu. Learning auxiliary monocular contexts helps monocular 3d object detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 1810–1818, 2022. 5

[26] Francesco Locatello, Stefan Bauer, Mario Lucic, Gunnar Raetsch, Sylvain Gelly, Bernhard Schölkopf, and Olivier Bachem. Challenging common assumptions in the unsupervised learning of disentangled representations. In *Proceedings of the 36th International Conference on Machine Learning*, pages 4114–4124. PMLR, 2019. 2

[27] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 6

[28] Jianbiao Mei, Yu Yang, Mengmeng Wang, Junyu Zhu, Jongwon Ra, Yukai Ma, Laijian Li, and Yong Liu. Camera-based 3d semantic scene completion with sparse guidance network. *IEEE Transactions on Image Processing*, 2024. 1, 5, 6, 8

[29] Ruihang Miao, Weizhou Liu, Mingrui Chen, Zheng Gong, Weixin Xu, Chen Hu, and Shuchang Zhou. Occdepth: A depth-aware method for 3d semantic scene completion. *arXiv preprint arXiv:2302.13540*, 2023. 1, 2

[30] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1):99–106, 2021. 3

[31] 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. 6

[32] Luis Roldao, Raoul de Charette, and Anne Verroust-Blondet. Lmscnet: Lightweight multiscale 3d semantic completion. In *2020 International Conference on 3D Vision (3DV)*, pages 111–119. IEEE, 2020. 1, 6

[33] Faranak Shamsafar, Samuel Woerz, Rafia Rahim, and Andreas Zell. Mobilestereonet: Towards lightweight deep networks for stereo matching. In *Proceedings of the ieee/cvf winter conference on applications of computer vision*, pages 2417–2426, 2022. 4

[34] Shuran Song, Fisher Yu, Andy Zeng, Angel X. Chang, Manolis Savva, and Thomas Funkhouser. Semantic scene completion from a single depth image. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 190–198, 2017. 1, 6

[35] Luan Tran, Xi Yin, and Xiaoming Liu. Disentangled representation learning gan for pose-invariant face recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1415–1424, 2017. 2

[36] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 4

[37] Vikas Verma, Alex Lamb, Christopher Beckham, Amir Najafi, Ioannis Mitliagkas, David Lopez-Paz, and Yoshua Bengio. Manifold mixup: Better representations by interpolating hidden states. In *International conference on machine learning*, pages 6438–6447. PMLR, 2019. 5

[38] Peng Wang, Lingjie Liu, Yuan Liu, Christian Theobalt, Taku Komura, and Wenping Wang. Neus: Learning neural implicit surfaces by volume rendering for multi-view reconstruction. *arXiv preprint arXiv:2106.10689*, 2021. 3

[39] Song Wang, Jiawei Yu, Wentong Li, Wenyu Liu, Xiaolu Liu, Junbo Chen, and Jianke Zhu. Not All Voxels are Equal: Hardness-Aware Semantic Scene Completion with Self-Distillation . In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14792–14801, Los Alamitos, CA, USA, 2024. IEEE Computer Society. 6

[40] Xin Wang, Hong Chen, Zihao Wu, Wenwu Zhu, et al. Disentangled representation learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024. 2

[41] Yida Wang, David Joseph Tan, Nassir Navab, and Federico Tombari. Forknet: Multi-branch volumetric semantic completion from a single depth image. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 8607–8616, 2019. 2

[42] 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. 2, 6

[43] Xu Yan, Jiantao Gao, Jie Li, Ruimao Zhang, Zhen Li, Rui Huang, and Shuguang Cui. Sparse single sweep lidar point cloud segmentation via learning contextual shape priors from scene completion. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 3101–3109, 2021. 1

[44] Jiawei Yao, Chuming Li, Keqiang Sun, Yingjie Cai, Hao Li, Wanli Ouyang, and Hongsheng Li. Ndc-scene: Boost monocular 3d semantic scene completion in normalized device coordinates space. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9421–9431. IEEE Computer Society, 2023. 1, 2

[45] Zhu Yu, Runmin Zhang, Jiacheng Ying, Junchen Yu, Xiaohai Hu, Lun Luo, Si-Yuan Cao, and Hui-Liang Shen. Context and geometry aware voxel transformer for semantic scene completion. *arXiv preprint arXiv:2405.13675*, 2024. 1

[46] Zhu Yu, Runmin Zhang, Jiacheng Ying, Junchen Yu, Xiaohai Hu, Lun Luo, Si-Yuan Cao, and Hui-Liang Shen. Context and geometry aware voxel transformer for semantic scene completion. *Advances in Neural Information Processing Systems*, 37:1531–1555, 2025. 6, 7

[47] Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo. Cutmix: Regularization strategy to train strong classifiers with localizable features. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6023–6032, 2019. 5

[48] Hongyi Zhang, Moustapha Cissé, Yann Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. *ArXiv*, abs/1710.09412, 2017. 2

[49] Yunpeng Zhang, Zheng Zhu, and Dalong Du. Occformer: Dual-path transformer for vision-based 3d semantic occupancy prediction. In *Proceedings of the IEEE/CVF Inter-*

national Conference on Computer Vision, pages 9433–9443, 2023. 1, 2, 6

[50] Yupeng Zheng, Xiang Li, Pengfei Li, Yuhang Zheng, Bu Jin, Chengliang Zhong, Xiaoxiao Long, Hao Zhao, and Qichao Zhang. Monoocc: Digging into monocular semantic occupancy prediction. *arXiv preprint arXiv:2403.08766*, 2024. 2, 6, 7

[51] Zhikang Zou, Xiaoqing Ye, Liang Du, Xianhui Cheng, Xiao Tan, Li Zhang, Jianfeng Feng, Xiangyang Xue, and Errui Ding. The devil is in the task: Exploiting reciprocal appearance-localization features for monocular 3d object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2713–2722, 2021. 2