Enhancing Generalizability via Utilization of Unlabeled Data for Occupancy Perception

Ruihang Li, Tao Li, Shanding Ye, Kaikai Xiao, Zheng Huangnan, Zhe Yin, Zhijie Pan

College of Computer Science and Technology, Zhejiang University
Hangzhou, 310027 China
12221089@zju.edu.cn

Abstract

3D occupancy perception accurately estimates the volumetric status and semantic labels of a scene, attracting significant attention in the field of autonomous driving. However, enhancing the model's ability to generalize across different driving scenarios or sensing systems, often requires redesigning the model or extra-expensive annotations. To this end, following a comprehensive analysis of the occupancy model architecture, we proposed the UGOCC method that utilizes domain adaptation to efficiently harness unlabeled autonomous driving data, thereby enhancing the model's generalizability. Specifically, we design the depth fusion module by employing self-supervised depth estimation, and propose a strategy based on semantic attention and domain adversarial learning to improve the generalizability of the learnable fusion module. Additionally, we propose an OCC-specific pseudo-label selection tailored for semi-supervised learning, which optimizes the overall network's generalizability. Our experiment results on two challenging datasets nuScenes and Waymo, demonstrate that our method not only achieves state-of-the-art generalizability but also enhances the model's perceptual capabilities within the source domain by utilizing unlabeled data.

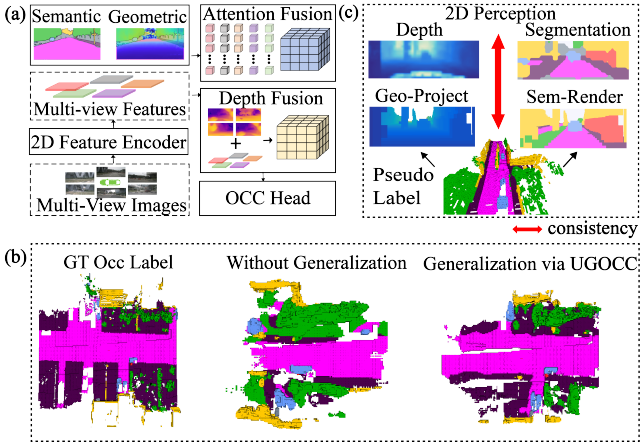

Figure 1: (a) Contemporary occupancy perception pipeline (b) Enhancement of 3D perception generalization in terms of geometry and semantics via 2D information through pseudo-labels. (c) Visualization of UGOCC generalization strategy.

Introduction

Vision-centric 3D Occupancy Perception (Pan et al. 2024a) benefits from the implicit feature representations in multi-camera systems. It voxelizes the environment around autonomous vehicles and assigns semantic information, playing an important role in modern autonomous driving and robotics. However, it typically demands supervised learning using occupancy labels obtained from multi-frame semantic LiDAR reconstructions, which incurs substantial cost overhead. Additionally, autonomous vehicles often operate in environments characterized by substantial feature variation and are equipped with diverse sensor suites (Ayala and Mohd 2021), including varying numbers of cameras and different fields of views (FOV). This leads to significant performance decline in vision-based OCC perception models when transitioning from the source to the target domain, as depicted in Fig. 1-(b).

Currently, the state-of-the-art occupancy perception models are divided into three main parts (Busch et al. 2024; Schramm et al. 2024; Li et al. 2024b; Harley et al. 2023): **feature extraction**, **feature fusion**, and **occupancy decoding head**, as shown in Fig. 1-(a). In the feature extraction module (2D Feature Encoder), which shares weights across views, the generalization decline from source to target domains—similar to 2D perception tasks—can be effectively migrated using semi-supervised and self-supervised learning, as established in 2D generalization techniques (Schwonberg et al. 2023). The feature fusion principal employs two strategies: explicit fusion based on depth information and implicit fusion utilizing attention-based learning. The generalization performance of these strategies is influenced by various sources of generalization bias, including depth estimation performance, network sensitivity to the intrinsic and extrinsic parameters of the sensor suite, and discrepancies in scene characteristics. As for occupancy decoding head, during the domain transfer process, the primary factors affecting performance are the decline in the generalization ability of the unified feature map and overfitting caused by the geometric distribution and characteristics of semantic categories.

Enhancing the domain generalization of the feature ex-

traction network through 2D perception tasks, and guiding the fusion and decoding modules with the 2D network's capabilities, is a natural strategy to improve the generalization of feature fusion and 3D occupancy perception. Such as, utilizing 2D-to-3D rendering consistency and projection consistency to transfer the semantic and depth generalization capabilities of 2D networks to 3D tasks, respectively, as shown in Fig. 1-(c). Consequently, a critical challenge emerges: **how to effectively utilize unlabeled data in the target domain to enhance generalization and improve model performance and adaptability?**

To this end, we introduce **UGOCC**, an occupancy perception strategy that leverages unlabeled data to enhance target domain generalization and improve source domain performance. Our model integrates three key modules: **Semantic Query Adversarial Fusion**, **Self-supervised Depth Augmentation**, and **Semantic and Geometric Awared Pesudo-label Selection**, creating a unique pipeline for effective module utilization. The Semantic Query Adversarial Module boosts generalization through direct interaction in both 2D and 3D spaces by incorporating semantic attention as a learnable query in the fusion process. This facilitates active alignment of BEV features across domains, enhancing semantic feature generalization. The Self-supervised Depth Augmentation module uses unlabeled temporal data to improve 2D depth estimation generalization across different viewpoints. Additionally, our semi-supervised OCC perception approach employs thresholding and a 2D-to-3D pseudo-label guidance strategy, significantly boosting model performance. As shown in Fig. 1-(b), **UGOCC** effectively reduces inter-domain discrepancies, with noticeable generalization improvements in the target domain.

In summary, the contributions of this paper are:

- Our approach, **UGOCC**, pioneers the study of occupancy perception by efficiently leveraging unlabeled data to enhance model generalization and robustness, integrating pseudo-label-based learning into this field.
- We propose the **Semantic Query Adversarial Fusion** and **Self-supervised Depth Augmentation** modules, achieving significant enhancements in the generalization capabilities of the fusion module in the OCC perception network and its ability to learn geometric and semantic information.
- We introduce a novel **pseudo-label selection strategy** for **semi-supervised** learning in OCC perception, efficiently transferring the generalization capabilities of 2D networks to 3D OCC perception and enhancing the overall network's generalizability.
- Our method demonstrates excellent performance in transferring from source to target domains. Additionally, we have open-sourced our code, contributing to the community.

Related Work

Voxel-level 3D Occupancy Perception

3D occupancy perception evolved from Occupancy Grid Maps (Milstein 2008) (OGM) and used deep learning to acquire semantic occupancy voxel information from images, guiding subsequent perception, planning, and navigation tasks. Existing studies utilize three types of supervision: sparse, dense, and 2D supervision. Sparse supervision uses semantic lidar point clouds for training and testing, as seen in early methods like Monoscene (Cao and De Charette 2022) and TPVFormer (Huang et al. 2023). Dense supervision focuses on creating occupancy labels tailored for visual perception, SurroundOcc introduces a SSC method enhancing label quality effectively (Wei et al. 2023b). Occ3D introduces the Occ-nuScenes and Occ-Waymo datasets with a coarse-to-fine voxel encoder for occupancy representation (Tian et al. 2024). RenderOcc, OccNerf, and SelfOcc (Pan et al. 2024b; Zhang et al. 2023; Huang et al. 2024) employ 2D labels for occupancy learning, proposing a new pipeline. However, these methods do not address occupancy networks' generalization in cross-domain applications or transfer learning for unlabeled models, prompting us to propose the UGOCC strategy.

Domain Adaptation

Domain generalization focuses on improving model performance on unlabeled target domains. Various approaches have been applied to 2D detection, such as aligning feature distributions and using pseudo-labels. These techniques mainly address domain shifts due to environmental changes like rain or low lighting. Recent studies such as STMono3D (Yang et al. 2024) and M2ATS (Guo et al. 2023) have explored these challenges in 3D vision. In the field of autonomous driving perception, DA-Bev (Jiang et al. 2024) enhances cross-domain capabilities of 3D object detection through active learning and self-training, improving model adaptability in extreme conditions like night and rain. PD-BEV (Lu et al. 2023) attributes cross-domain loss primarily to differences in sensor suite parameters and proposes solutions. We conducted cross-domain analysis for the occupancy task and performed decoupling analysis of network modules, proposing targeted generalization strategies.

Method

Preliminary

Our approach focuses on enhancing the performance of camera-only occupancy perception tasks within the source domain and improving generalization in the target domain. It involves a labelled source domain $D_S = \{I_s^n, K_s^n, O_s, D_s^n\}$, and unlabelled target target domain $D_t = \{I_s^n, K_s^n\}$, where I represents input images, K denotes camera parameters, O stands for occupancy labels, D indicates depth maps, the superscript n represents the number of views, and the subscripts s and t refer to the source and target domains, respectively. Our goal is to achieve accurate occupancy perception results in the source domain and highly generalized results in the target domain.

Model Architecture

Our approach consists of two main components: the inference network of occupancy perception, as shown in Fig. 2-

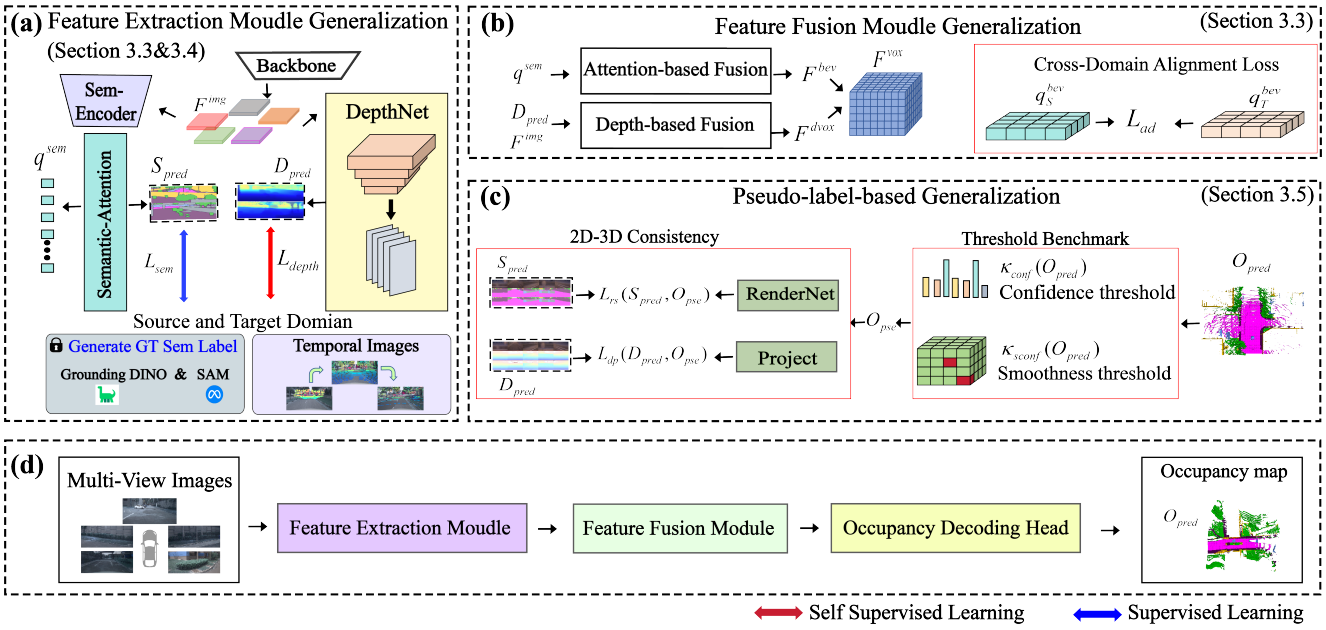

Figure 2: (a) Enhance feature extraction module's generalization via Semantic Attention and Self-supervised Depth Augmentation. (b) Enhance feature fusion module's generalization, by cross-domain alignment. (c) Occ-specific Pseudo-learning, incorporated to enhance overall metwork's performance. (d) Our network comprises three main modules.

(a)(b)(c), and the generalization strategy from the source domain to the target domain, as illustrated in Fig. 2-(d).

Feature Extraction Module The image feature extraction module takes multi-view images as input $I^{N_{cam}} \in \mathbb{R}^{n\times 3\times H\times W}$, where N_{cam} represents the number of views, and H and W denote the height and width of the original images, respectively. We employ ResNet (He et al. 2016) as a multi-scale feature extractor to obtain image features at multi-scales. Then we deploy a Feature Pyramid Network to aggregate these image features. The final output of the feature extraction module is denoted as F^{img}.

Feature Fusion Module In the Feature Fusion Module, image features from multiple views are lifted to 3D and fused into a unified volumetric feature map $F^{vox} \in C_v \times \frac{X}{ds} \times \frac{Y}{ds} \times \frac{Z}{ds}$, where C_v is the embedding dim, ds is the downsampling ratio of each dimension of the volumetric feature map F^{vox}.

As dispaly in Fig. 2-(b), We initially decode the semantic map S_{pred} and the depth map D_{pred} from F^{img} though the SemanticNet $\mathcal{S}$ and DepthNet $\mathcal{D}$. For the explicit depth fusion, we adopt LSS (Philion and Fidler 2020) to fuse image features from various views using depth estimates of each pixel that account for uncertainty, resulting in a depth-based feature map $F^{dvox} = LSS(F^{img}, D, K)$. For the attention-based fusion, our approach innovatively proposes an attention fusion strategy named **Semantic Query Adversarial Fusion** (Section 3.3), which incorporates semantic supervision priors in the fusion process. This approach enhances the overall network's utilization ability of 2D semantic information, resulting in the fused BEV feature with semantic priors F^{bev}. Finally, we unsqueeze F^{bev} and add them with F^{dvox} them in order to obtain the unify volumetirc representation F^{vox}.

Occupancy Perception Head As depicted in Fig. 2-(b), the purpose of this module is to leverage 3D volumetric representations F^{vox} to regress the occupancy prediction results. We use a simple multi-layer convolutional network for regression, similar to the approach in BEVDet (Huang et al. 2021). This ultimately results in the occupancy map $O_{pred} \in cls \times X \times Y \times Z$, where cls denotes the semantic category, (X, Y, Z) correspond to the dimensions of the occupancy space.

Semantic Query Adversarial Fusion

Depth features of various perspectives are explicitly utilized in the process of occupancy perception, whereas semantic features are not directly involved. Specifically, the depth map D serves as the input to the depth fusion module, explicitly guiding the populating of image features into volume representational space. This enhancement of depth estimation accuracy directly improves the precision of projections and reduces the domain discrepancies. Meanwhile, semantic features embedded within the image features implicitly contribute to feature fusion, intensifying domain discrepancies, which cannot be explicitly reinforced.

To mitigate this discrepancies, we introduced a module called Semantic Query Adversarial Fusion, as shown in Fig. 3. It utilizes a cross-domain supervision-capable module to reduce the cross-domain loss associated with the extraction of 2D semantic features, and employs a learnable approach to incorporate this guidance into the fusion module. As depicted in Fig. 3-(a), we employ Semantic-

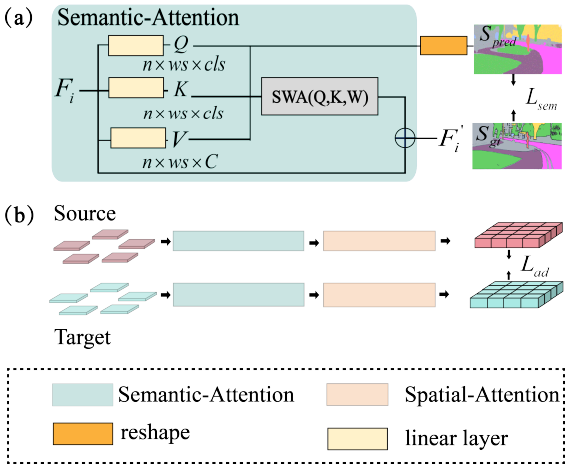

Figure 3: Semantic attention query Q with feature dimension cls matches semantic segmentation. Adversarial learning loss $\mathcal{L}_{ad}$ in cross-domain alignment uses fused BEV features from source and target domains.

Attention as the final layer of SemanticNet. It divides image features F^{img} into three entities: $Q \in \{n \times ws \times cls\}$, $K \in \{n \times ws \times cls\}$, $V \in \{n \times ws \times C\}$, where Q, K are projected into the semantic space with feature channels aligned to the number of categories in semantic segmentation and V is projected into the embedding feature space, with ws denotes the window size (Liu et al. 2021; Chen et al. 2023; Liu et al. 2022). Subsequently, Q transforms into the semantic map S_{pred}, which will be supervised with SAM-generated (Osco et al. 2023) ground truth S_{gt} to compute the semantic segmentation loss $\mathcal{L}_{sam}$. Leveraging SAM for ground truth labels, we establish low-cost 2D semantic labels in both domains, enhancing domain alignment with the semantic attention mechanism.

The semantic-aware query is computed as shown in Eq.1, and we adopt a fusion method similar to Spatial-Attention to derive bev query q^{bev}, where q^{dbev} is the query transformed from F^{dvox} compressed along the z-axis, SPA denotes Spatial-attention (Li et al. 2022).

$$\begin{aligned} q^{bev} &= SPA(q^{dbev}, q^{sem}, q^{sem}) \\ q^{sem} &= \text{SoftMax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \end{aligned}, \tag{1}$$

To reduce the cross-domain discrepancies, we employ a domain classification loss $\mathcal{L}_{ad}$ for cross-domain alignment between source domain bev query q_S^{bev} and target domain bev query q_T^{bev}. The overall semantic cross-domain alignment loss can be defined as:

$$\mathcal{L}_{sal} = \lambda_{sem} \cdot \mathcal{L}_{sem}(S_{gt}, S_{pred}) + \lambda_{ad} \cdot \mathcal{L}_{ad}(q_S^{bev}, q_T^{bev}) \tag{2}$$

This strategy employs multi-level alignment across image and bird's-eye view feature space, positively impacts the generalization of semantic features. Furthermore, the derived bev features are utilized as input image features in a deep fusion module, facilitating adaptive alignment during the training process.

Self-supervised Depth Augmentation

Depth estimation performance directly determines the accuracy of image feature projection in depth-based fusion (Philion and Fidler 2020), which is crucial for the generalization of the fusion module. Consequently, we propose self-supervised depth estimation to enhance multi-view depth estimation by effectively utilizing image data from both source and target domains, thereby improving the generalizability of the depth fusion module.

Specifically, we utilize the geometric consistency of temporal data to establish reprojection loss $\mathcal{L}rep$ and loss of adjacent views $\mathcal{L}adj$, as denoted in Eq. 3. This approach is frequently employed in multi-view self-supervised depth estimation (Li et al. 2024a; Wei et al. 2023a), as denotes in Eq.3,

$$\begin{aligned} \mathcal{L}_{rep} &= (1-\alpha)\|I_t - \tilde{I}_t\|_1 + \alpha \frac{\left(1 - \text{SSIM}(I_t, \tilde{I}_t)\right)}{2} \\ \mathcal{L}_{adj} &= \frac{1}{K}\sum_{i=1}^{K}\left[(u_i - \tilde{u}_i)^2 + (v_i - \tilde{v}_i)^2\right]^{\frac{1}{2}}, \end{aligned} \tag{3}$$

where I_t and $\tilde{I}_t$ denote the original and projected images, respectively. SSIM (Palubinskas 2017) represents the Structural Similarity Index. Coordinates (u, v) and $(\tilde{u}, \tilde{v})$ indicate the original and projected positions of adjacent keypoints in the image coordinate system, and α is an adjustable weight.

Due to the employment of camera's extrinsic parameters in the viewpoint transformations to supervise between adjacent views. Depth maps achieve the ability of represent metric distances rather than non-metric depth disparities. During training in the source domain, we continue to use supervised depth loss $\mathcal{L}_{sp}$. The overall loss function can be expressed as follows:

$$\mathcal{L}_{depth} = \lambda_s \mathcal{L}_{sp} + \lambda_{rep}\mathcal{L}_{rep} + \lambda_{adj}\mathcal{L}_{adj}, \tag{4}$$

where λ_s is used only in the source domain.

Pesudo-label Selection

Our approach develops strategies for extracting geometric and semantic features in 2D spaces and enhancing fusion module generalizability. We employ pseudo-labeling to efficiently use unlabeled data in the target domain and enhance cross-domain adaptability. The selection of pseudo-labels, focused on efficiency and accuracy, critically affects model performance. Utilizing semantic maps (S_{pred}) and depth maps (D_{pred}) aids in effective pseudo-label generation and selection. Leveraging a 2D feature generalization strategy, our network minimizes domain discrepancies, making 2D-based pseudo-label selection advantageous. We categorize our strategies into two: a **2D-to-3D pseudo-label guidance strategy** for semantic and geometric consistency, and a **pseudo-label selection strategy** that includes confidence thresholds ($\mathcal{K}_{\text{conf}}$) and spatial smoothness thresholds ($\mathcal{K}_{\text{sconf}}$).

For model perception in the target domain, represented as $O_T \in \{\text{cls} \times X \times Y \times Z\}$, pseudo-labels are initially selected based on the $\mathcal{K}_{\text{conf}}$ and $\mathcal{K}_{\text{sconf}}$ thresholds.

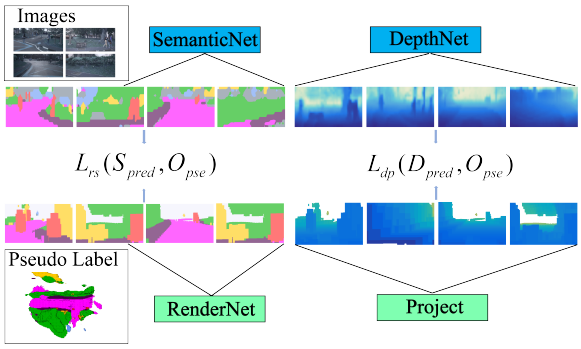

Figure 4: Consistency between pseudo occupancy label O_{pseudo} and 2d perception results D_{pred}, S_{pred}.

$$O_{pse} = \begin{cases} \arg\sigma(O_{pred}) & \text{if } C \geq \mathcal{K}_{conf} \text{ and } S \geq \mathcal{K}_{sconf} \\ \text{null} & \text{otherwise} \end{cases}, \tag{5}$$

where C is the confidence, S is the spatial smoothness, O_{pse} is the pseudo-label, σ is the softmax function, arg denotes the class with the maximum probability.

Spatial smoothness mitigates the occurrence of discrete points in occupancy pseudo-labels and minimizes positional inaccuracies arising from the network's limited geometric adaptability, as denoted in Eq. 6. Here, 'div' denotes the divergence in three-dimensional space, and the size of the occupancy grid signifies the infinitesimal elements along the three axes.

$$S(x, y, z) = \frac{1}{div(O_{pse}(x, y, z))} \tag{6}$$

Subsequently, we propose two consistency measures to filter out noise in pseudo-labels and enhance their quality. For semantic consistency, as depicted in Fig. 4, It can be observed that cross-domain decline in occupancy manifest as inaccuracies in scale and localization information, which can be effectively corrected in the 2D space via S_{pred}. We use RenderNet to render the pseudo-labels O_{pse} from various perspectives, forming S_{pseudo}. Then, we apply a cross-entropy loss to calculate the confidence, excluding the *free* category, to further enhance the quality of the pseudo-labels, we employ a 2D-3D semantic consistency loss, as defined in Eq.7..

$$\mathcal{L}_{\text{rs}} = CE(S_{pred}, S_{pseudo}), \tag{7}$$

Multi-view depth map can be regarded as labels within the occupancy space $O_D = project(D, K)$, containing only *free* and *non-free* categories which can be integrated with the confidence and mask of pseudo-labels. As depicted in Fig. 4, We propose a geometry consistency loss:

$$\mathcal{L}_{\text{pd}} = BCE(O_D, O_T), \tag{8}$$

The essence of this approach is to reintroduce previously low-confidence *free* labels into the pseudo-label set, without impacting the segments that maintain high confidence. The overall pseudo-label loss $\mathcal{L}_p$ can be expressed as:

$$\mathcal{L}_p = \lambda_{rs}\mathcal{L}_{rs} + \lambda_{pd}\mathcal{L}_{pd} + \lambda_{pl}CE(O_{pred}, O_{pse}), \tag{9}$$

nuScenes to Waymo										
	BEVDet*		BEVF*		FBOCC*		Flash*		UGOCC*	
	S	T	S	T	S	T	S	T	S	T
Dri-S.	58.9	18.0	58.3	21.4	61.8	6.8	59.1	4.1	62.7	49.2
Veh.	35.7	5.6	34.0	4.8	38.1	4.8	35.6	4.7	38.4	23.4
Ped.	22.9	1.3	21.9	1.2	24.9	1.2	23.2	0.9	24.4	14.7
Bui.	18.8	3.1	15.9	2.9	22.2	3.2	18.6	2.5	23.2	13.8
Veg.	23.8	4.5	21.7	3.7	26.6	3.5	23.8	3.7	26.7	13.1
Sid.	25.4	2.1	33.3	3.8	38.7	1.1	35.3	1.1	39.7	22.7
Others	28.8	0.3	27.0	0.1	29.5	0.2	29.2	0.1	28.1	4.2
Tra.	23.9	0.0	23.7	0.0	25.2	0.0	25.4	0.0	21.9	1.8
2W.	23.7	0.4	22.9	0.5	24.8	0.4	23.8	0.3	23.2	8.1
mIoU	30.2	3.9	28.8	4.3	**32.5**	**2.3**	30.5	1.9	**32.0**	**16.7**

Table 1: **Evaluation on nuScenes to Waymo Transition:** S tests in the source domain, T in the target, with $*$ for our implementation.

Waymo to nuScenes										
	BEVDet*		BEVF*		FBOCC*		Flash*		UGOCC*	
	S	T	S	T	S	T	S	T	S	T
Dri-S.	63.1	15.9	66.6	22.9	63.4	11.1	48.1	0.7	60.6	48.7
Veh.	24.7	2.4	23.0	1.0	27.2	1.7	24.6	0.1	25.2	21.6
Ped.	17.9	1.0	17.8	0.3	21.3	0.1	18.6	0.2	17.4	11.5
Bui.	14.6	2.2	14.0	1.5	16.3	1.6	15.4	1.2	18.4	12.2
Veg.	13.9	4.5	13.4	4.6	15.0	3.7	14.1	4.7	16.9	12.4
Sid.	39.7	6.8	37.3	6.8	40.3	4.6	32.8	0.3	38.4	22.8
Others	9.6	0.0	7.9	0.1	10.2	0.0	8.3	0.0	10.2	8.9
Tra.	8.5	0.0	7.2	0.0	7.0	0.0	7.1	0.0	11.1	4.3
2W.	10.9	0.9	9.1	0.9	18.4	0.0	12.4	0.4	11.6	10.6
mIoU	22.5	3.8	21.1	4.2	**24.4**	**2.6**	20.2	0.8	**23.3**	**16.8**

Table 2: **Evaluation on Waymo to nuScenes Transition:** S tests in the source domain, T in the target, with $*$ for our implementation.

where the weight λ_{pd} is positively correlated with the confidence of the pseudo-labels.

Overall Supervision

During the overall training process, our loss function is defined as:

$$\mathcal{L} = \mathcal{L}_{sal} + \lambda_{spv}\mathcal{L}_{occ}(O_{pred}, O_{gt}) + \lambda_{pse}\mathcal{L}_p + \lambda_d\mathcal{L}_{depth}, \tag{10}$$

where λ_{spv} is used only in the source domain, and λ_{pse} is used only in the target domain, alternative weight settings can be found in the corresponding module.

Experiments

Datasets

Our experiments use the nuScenes and Waymo datasets (Caesar et al. 2020; Sun et al. 2020) with

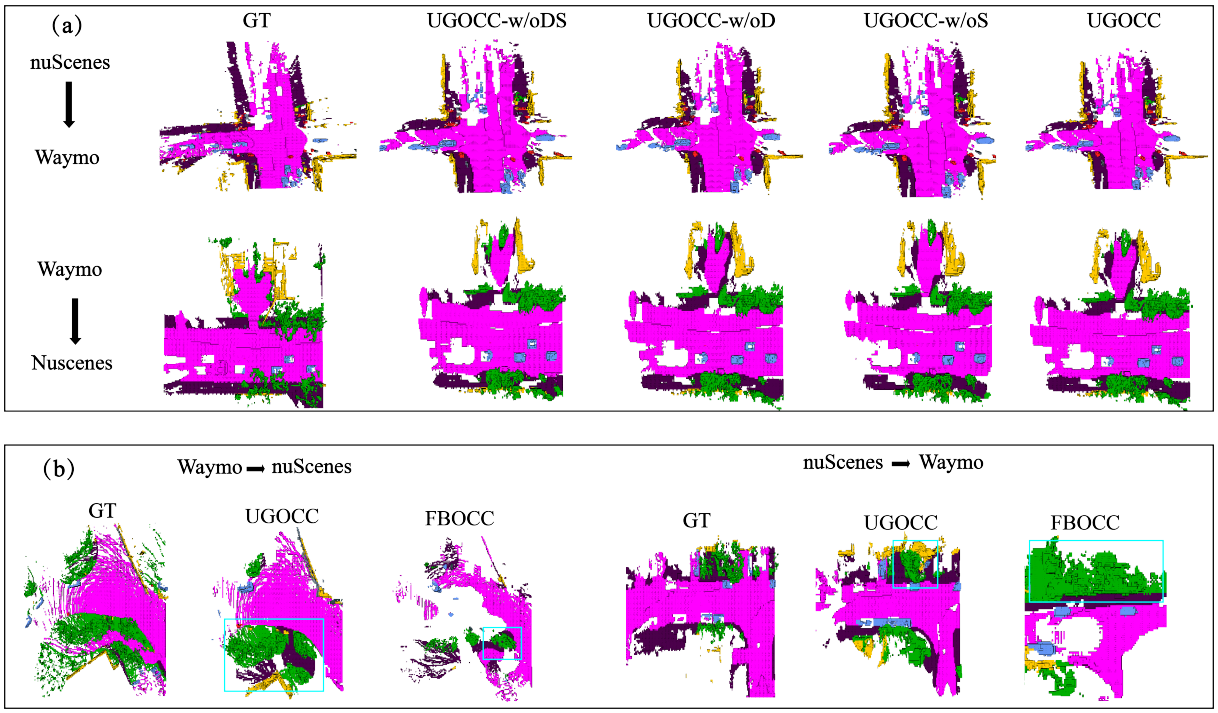

Figure 5: **Visualization of experimental results:** (a) Visualizations of ablation studies under the nuScenes-to-Waymo and Waymo-to-nuScenes benchmarks, and (b) Visualizations of the overall optimization of our method on occupancy perception.

OCC labels from OCC-nuScenes and OCC-Waymo. The nuScenes dataset obtains a six-camera surround-view system, while Waymo utilize a five-camera wide-angle system. OCC labels are constrained within -40m to 40m on the X and Y axes, and -1m to 5.4m on the Z axis, with a voxel resolution of 0.4m. We construct semantic maps of key frames using Grounded-SAM for automated annotation.

In order to enhance the benchmarking of the network's perceptual capabilities across different datasets, we standardized the labels by mapping the OCC classes in both OCC-nuScenes and OCC-Waymo to ten categories. Detailed information regarding label quality, predicted semantic maps, and the exact mapping relationships for OCC label transformation is provided in the appendix.

Implementation Details

We selected four baselines as benchmark approaches, each employing distinct strategies in their multi-view feature fusion modules, BEVDetOcc (Depth Fusion) (Huang and Huang 2022), BEVFormer (Attention Fusion) (Li et al. 2022), FBOCC (both Depth fusion and Attention Fusion) (Li et al. 2023) and FlashOcc (2d representation via Depth Fusion) (Yu et al. 2023).

To ensure fairness, all camera images are resized to 256×704. Each method uses ResNet50 with ImageNet-pretrained weights as the 2D backbone for feature extraction. The Feature Fusion Module utilizes 3D volumetric (100, 100, 8) or 2D bird-eye-view (200, 200) representations. Each sample includes five images for perspective consistency across datasets. Methods are trained for 24 epochs in the source domain using the AdamW optimizer with a learning rate of 7×10^{-5}. For cross-domain learning, we pretrain self-supervised depth estimation and Semantic Query Adversarial Fusion, setting their backpropagation ratios to 0.1. Confidence and smoothness thresholds are 0.85 and 0.8, respectively. Experiments run on six 4090 devices, with a learning rate of 3.5×10^{-5}, continuing until the pseudo-label selection ratio exceeds 35%. All the experiments are process without occupancy mask.

Quantitative Experiments and Ablation Studies

Our experiments were conducted under two benchmarks: using nuScenes as the source domain and Waymo as the target domain, and using Waymo as the source domain and nuScenes as the target domain. We identified several baselines with typical features in multi-view feature strategies for occupancy perception experiments, including the following: **BEVDetOCC:** Utilizing the depth fusion module (LSS) to obtain a volumetric representation space. **BEVFormer:** Utilizing the transformer-based learnable fusion approach to obtain a volumetric representation space. **FBOCC:** Utilizing both a depth estimation-based fusion approach and a learnable BEV fusion approach to produce a unified feature map. **FlashOCC:** this method constructs a highly concentrated BEV feature map with a resolution of (200×200), and uses a specialized head for decoding. For our **UGOCC** method, we conducted ablation studys with four different configurations:

- **UGOCC-w/oD**: Without enhancement from the depth fusion module based on self-supervised depth estimation

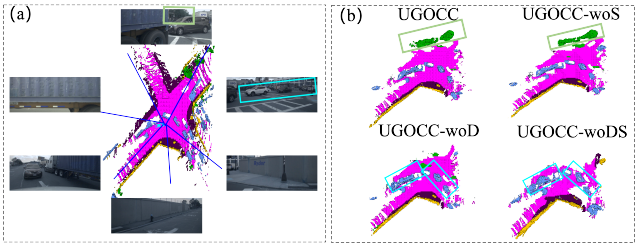

Figure 6: **Analysis of Enhanced Generalization Capabilities in Fusion:** (a) Ground truth occupancy label and surround-views images. (b) Occupancy prediction results of UGOCC, UGOCC-w/oD, UGOCC-w/oS, and UGOCC-w/oDS.

		nuScenes to Waymo				Waymo to nuScenes			
		Dri-S.	Veh.	Ped.	mIoU	Dri-S.	Veh.	Ped.	mIoU
w/o DS	S	59.8	37.9	23.6	30.8	59.3	23.5	11.5	18.8
	T	37.3	16.4	10.8	**11.8**	42.7	19.2	9.4	**13.1**
w/o S	S	62.9	38.5	24.6	32.3	59.5	25.4	18.0	22.9
	T	46.4	20.6	13.2	**15.2**	46.1	19.0	10.6	**14.9**
w/o D	S	61.6	37.8	22.7	30.9	60.8	22.5	10.2	19.1
	T	38.9	17.7	10.8	**13.1**	48.1	23.0	9.0	**15.3**
UGOCC	S	62.7	38.4	24.4	32.0	60.6	25.2	17.4	23.3
	T	49.2	23.4	14.7	**16.7**↑	48.7	21.6	11.5	**16.8**↑

Table 3: **Ablation Studie:** w/oDS denotes UGOCC-w/oD , w/oS denotes UGOCC-w/oD, w/oDS denotes UGOCC-w/oDS, as depict in section 4.2.

pre-training.

- **UGOCC-w/oS**: Without using the Semantic Query Adversarial Fusion module to enhance the learnable semantic module.
- **UGOCC-w/oDS**: Without both two modules.
- **UGOCC**: The complete method, as shown in Fig.2.

We adjusted the occupancy labels in both nuScenes and Waymo, necessitating the retraining of all baseline methods. As shown in Tab.1 and Tab.2, every baseline method experiences a significant performance drop in the target domain (mIoU$<$5) across both benchmarks. Fig. 5 illustrates that during the transfer process, models retain some semantic recognition abilities, with generally accurate category identification, as indicated by the cyan boxes. However, they exhibit significant deficiencies in geometric scale, distance, and spatial semantic distribution.

Our method effectively mitigates these issues, significantly enhancing cross-domain generalization. Incorporating unlabeled data from the target domain through semi-supervised learning substantially improves the model's performance. The complete UGOCC approach achieves a score of 16.8 in the target domain, reaching 400% of the baseline, demonstrating its effectiveness. A detailed analysis of each module is provided below.

Semantic and Geometric Awared Pesudo-label Selection: In the nuScenes to Waymo benchmark, the UGOCC-woDS mIoU reached 11.8 and 13.1 in the Waymo and nuScenes target domains, respectively. While this marks a significant improvement over baseline methods, it indicates that relying solely on unsupervised generalization during the loss phase is insufficient for achieving high-level generalization in the fusion module.

Self-supervised Depth Augmentation: UGOCC-woS enhances the depth estimation module through semi-supervised learning and self-supervised depth estimation, improving performance across target domain viewpoints. As shown by the green boxes in Fig. 6-(b), this enhancement significantly improves the network's ability to perceive occluded objects at greater distances, due to better integration of features into the unified volumetric feature map. These objects occupy a smaller proportion in the original view, but the opposite is true in occupancy space. Precise depth projection overcomes the difficulty in establishing connections with the 2D semantic map during rendering.

Semantic Query Adversarial Fusion: Semantic Query Adversarial Fusion enhances the fusion module's generalization, while rendering consistency for pseudo-labels supervises the model's adaptation to semantic associations and sensor suite parameters in the source domain. As shown by the cyan boxes in Fig. 6, the introduction of this module has a significant positive impact on the geometric scale of nearby cross-view objects and their semantic associations in occupancy space.

For the complete method, the model's overall generalization performance further improves. This demonstrates that the synergy between depth fusion and semantic generalization strategies enhances overall model generalization. This aligns with the distinction between geometric and semantic information in occupancy perception results.

Discuss

Quantitative experiments reveal that the predicted mIoU for occupancy in the Waymo dataset is significantly lower than that in nuScenes. We attribute this primarily to the camera field of view and the evaluation mask, which are discussed in detail in the supplementary material. We conducted quantitative analyses to evaluate the sensitivity and importance of three specific modules in cross-domain transfer, which support our motivation. Detailed results are presented in the appendix.

Conclusion

We introduce UGOCC, a novel strategy to enhance the generalization of occupancy perception via unlabeled data. By analyzing the impact of domain differences across the network components, we developed specific strategies: Semantic Query Adversarial Fusion, Self-supervised Depth Augmentation, and Semantic and Geometric Aware Pseudo-label Selection. Our experiments on the large-scale autonomous driving datasets, nuScenes and Waymo, demonstrate that UGOCC achieve state-of-the-art performance in the generalization of occupancy prediction models, thereby significantly advancing their practical deployment.

Acknowledgments

This work was supported by the Key Research and Development Program of Zhejiang Province in China (No. 2023C01237).

References

Ayala, R.; and Mohd, T. K. 2021. Sensors in autonomous vehicles: A survey. *Journal of Autonomous Vehicles and Systems*, 1(3): 031003.

Busch, D.; Freeman, I.; Meyes, R.; and Meisen, T. 2024. Improved Single Camera BEV Perception Using Multi-Camera Training. *arXiv preprint arXiv:2409.02676*.

Caesar, H.; Bankiti, V.; Lang, A. H.; Vora, S.; Liong, V. E.; Xu, Q.; Krishnan, A.; Pan, Y.; Baldan, G.; and Beijbom, O. 2020. nuscenes: A multimodal dataset for autonomous driving. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 11621–11631.

Cao, A.-Q.; and De Charette, R. 2022. Monoscene: Monocular 3d semantic scene completion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3991–4001.

Chen, K.; Liu, S.; Zhu, T.; Qiao, J.; Su, Y.; et al. 2023. Improving expressivity of gnns with subgraph-specific factor embedded normalization. In *ACM SIGKDD Conference on Knowledge Discovery and Data Mining*.

Guo, D.; Lin, Y.; You, X.; Yang, Z.; Zhou, J.; Yang, B.; Zhang, J.; Shi, H.; Hu, S.; and Zhang, Z. 2023. M2ATS: A Real-world Multimodal Air Traffic Situation Benchmark Dataset and Beyond. In *Proceedings of the 31st ACM International Conference on Multimedia*, 213–221.

Harley, A. W.; Fang, Z.; Li, J.; Ambrus, R.; and Fragkiadaki, K. 2023. Simple-bev: What really matters for multi-sensor bev perception? In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, 2759–2765. IEEE.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

Huang, J.; and Huang, G. 2022. Bevdet4d: Exploit temporal cues in multi-camera 3d object detection. *arXiv preprint arXiv:2203.17054*.

Huang, J.; Huang, G.; Zhu, Z.; Ye, Y.; and Du, D. 2021. Bevdet: High-performance multi-camera 3d object detection in bird-eye-view. *arXiv preprint arXiv:2112.11790*.

Huang, Y.; Zheng, W.; Zhang, B.; Zhou, J.; and Lu, J. 2024. Selfocc: Self-supervised vision-based 3d occupancy prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 19946–19956.

Huang, Y.; Zheng, W.; Zhang, Y.; Zhou, J.; and Lu, J. 2023. Tri-perspective view for vision-based 3d semantic occupancy prediction. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 9223–9232.

Jiang, K.; Huang, J.; Xie, W.; Li, Y.; Shao, L.; and Lu, S. 2024. DA-BEV: Unsupervised Domain Adaptation for Bird's Eye View Perception. *arXiv preprint arXiv:2401.08687*.

Li, R.; Ye, S.; Yin, Z.; Li, T.; Zhang, Z.; Xiao, K.; and Pan, Z. 2024a. M2Depth: A Novel Self-Supervised Multi-Camera Depth Estimation with Multi-Level Supervision. In *2024 IEEE International Conference on Multimedia and Expo (ICME)*, 1–6. IEEE.

Li, Z.; Lan, S.; Alvarez, J. M.; and Wu, Z. 2024b. BEVNeXt: Reviving Dense BEV Frameworks for 3D Object Detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 20113–20123.

Li, Z.; Wang, W.; Li, H.; Xie, E.; Sima, C.; Lu, T.; Qiao, Y.; and Dai, J. 2022. Bevformer: Learning bird's-eye-view representation from multi-camera images via spatiotemporal transformers. In *European conference on computer vision*, 1–18. Springer.

Li, Z.; Yu, Z.; Wang, W.; Anandkumar, A.; Lu, T.; and Alvarez, J. M. 2023. Fb-bev: Bev representation from forward-backward view transformations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 6919–6928.

Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; and Guo, B. 2021. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, 10012–10022.

Liu, Z.; Ning, J.; Cao, Y.; Wei, Y.; Zhang, Z.; Lin, S.; and Hu, H. 2022. Video swin transformer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 3202–3211.

Lu, H.; Zhang, Y.; Lian, Q.; Du, D.; and Chen, Y. 2023. Towards generalizable multi-camera 3d object detection via perspective debiasing. *arXiv preprint arXiv:2310.11346*.

Milstein, A. 2008. Occupancy grid maps for localization and mapping. *Motion planning*, 381–408.

Osco, L. P.; Wu, Q.; de Lemos, E. L.; Gonçalves, W. N.; Ramos, A. P. M.; Li, J.; and Junior, J. M. 2023. The segment anything model (sam) for remote sensing applications: From zero to one shot. *International Journal of Applied Earth Observation and Geoinformation*, 124: 103540.

Palubinskas, G. 2017. Image similarity/distance measures: what is really behind MSE and SSIM? *International Journal of Image and Data Fusion*, 8(1): 32–53.

Pan, M.; Liu, J.; Zhang, R.; Huang, P.; Li, X.; Xie, H.; Wang, B.; Liu, L.; and Zhang, S. 2024a. Renderocc: Vision-centric 3d occupancy prediction with 2d rendering supervision. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, 12404–12411. IEEE.

Pan, M.; Liu, J.; Zhang, R.; Huang, P.; Li, X.; Xie, H.; Wang, B.; Liu, L.; and Zhang, S. 2024b. Renderocc: Vision-centric 3d occupancy prediction with 2d rendering supervision. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, 12404–12411. IEEE.

Philion, J.; and Fidler, S. 2020. 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*, 194–210. Springer.

Schramm, J.; Vödisch, N.; Petek, K.; Kiran, B. R.; Yogamani, S.; Burgard, W.; and Valada, A. 2024. BEVCar: Camera-Radar Fusion for BEV Map and Object Segmentation. *arXiv preprint arXiv:2403.11761*.

Schwonberg, M.; Niemeijer, J.; Termöhlen, J.-A.; Schmidt, N. M.; Gottschalk, H.; Fingscheidt, T.; et al. 2023. Survey on unsupervised domain adaptation for semantic segmentation for visual perception in automated driving. *IEEE Access*, 11: 54296–54336.

Sun, P.; Kretzschmar, H.; Dotiwalla, X.; Chouard, A.; Patnaik, V.; Tsui, P.; Guo, J.; Zhou, Y.; Chai, Y.; Caine, B.; et al. 2020. Scalability in perception for autonomous driving: Waymo open dataset. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2446–2454.

Tian, X.; Jiang, T.; Yun, L.; Mao, Y.; Yang, H.; Wang, Y.; Wang, Y.; and Zhao, H. 2024. Occ3d: A large-scale 3d occupancy prediction benchmark for autonomous driving. *Advances in Neural Information Processing Systems*, 36.

Wei, Y.; Zhao, L.; Zheng, W.; Zhu, Z.; Rao, Y.; Huang, G.; Lu, J.; and Zhou, J. 2023a. Surrounddepth: Entangling surrounding views for self-supervised multi-camera depth estimation. In *Conference on robot learning*, 539–549. PMLR.

Wei, Y.; Zhao, L.; Zheng, W.; Zhu, Z.; Zhou, J.; and Lu, J. 2023b. Surroundocc: Multi-camera 3d occupancy prediction for autonomous driving. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 21729–21740.

Yang, F.; Chen, H.; He, Y.; Zhao, S.; Zhang, C.; Ni, K.; and Ding, G. 2024. Geometry-Guided Domain Generalization for Monocular 3D Object Detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 6467–6476.

Yu, Z.; Shu, C.; Deng, J.; Lu, K.; Liu, Z.; Yu, J.; Yang, D.; Li, H.; and Chen, Y. 2023. Flashocc: Fast and memory-efficient occupancy prediction via channel-to-height plugin. *arXiv preprint arXiv:2311.12058*.

Zhang, C.; Yan, J.; Wei, Y.; Li, J.; Liu, L.; Tang, Y.; Duan, Y.; and Lu, J. 2023. Occnerf: Self-supervised multi-camera occupancy prediction with neural radiance fields. *arXiv preprint arXiv:2312.09243*.