

---

# See through the Dark: Learning Illumination-affined Representations for Nighttime Occupancy Prediction

---

Yuan Wu<sup>1\*</sup> Zhiqiang Yan<sup>2\*</sup> Yigong Zhang<sup>3†</sup> Xiang Li<sup>3</sup> Jian Yang<sup>1†</sup>

<sup>1</sup>PCA Lab<sup>‡</sup>, Nanjing University of Science and Technology

<sup>2</sup>National University of Singapore <sup>3</sup>Nankai University

## Abstract

Occupancy prediction aims to estimate the 3D spatial distribution of occupied regions along with their corresponding semantic labels. Existing vision-based methods perform well on daytime benchmarks but struggle in nighttime scenarios due to limited visibility and challenging lighting conditions. To address these challenges, we propose **LIAR**, a novel framework that learns illumination-affined representations. LIAR first introduces Selective Low-light Image Enhancement (SLLIE), which leverages the illumination priors from daytime scenes to adaptively determine whether a nighttime image is genuinely dark or sufficiently well-lit, enabling more targeted global enhancement. Building on the illumination maps generated by SLLIE, LIAR further incorporates two illumination-aware components: 2D Illumination-guided Sampling (2D-IGS) and 3D Illumination-driven Projection (3D-IDP), to respectively tackle local underexposure and overexposure. Specifically, 2D-IGS modulates feature sampling positions according to illumination maps, assigning larger offsets to darker regions and smaller ones to brighter regions, thereby alleviating feature degradation in underexposed areas. Subsequently, 3D-IDP enhances semantic understanding in overexposed regions by constructing illumination intensity fields and supplying refined residual queries to the BEV context refinement process. Extensive experiments on both real and synthetic datasets demonstrate the superior performance of LIAR under challenging nighttime scenarios. The source code and pretrained models are available here.

## 1 Introduction

Understanding the 3D structure of the environment [15, 16, 52–55] is a core task in autonomous driving, as it allows vehicles to perceive their surroundings and make informed decisions. Recently, vision-based occupancy prediction [12, 18, 26, 31] has attracted growing interest due to its ability to estimate the spatial layout of occupied regions together with their semantic labels. Although existing methods perform well under daytime conditions, their performance drops significantly in nighttime scenes. The need for reliable perception in such low-light environments underscores the importance of advancing nighttime occupancy prediction. However, this task remains particularly challenging, as limited visibility and complex lighting conditions can severely degrade the quality of visual inputs.

As illustrated in Fig. 1(a), nighttime images not only exhibit low visibility but also contain both underexposed and overexposed regions [23, 39, 57]. Underexposure caused by insufficient illumination severely degrades visual features, whereas overexposure from artificial light sources, such as vehicle headlights and streetlamps, results in significant semantic deficiency in the saturated regions.

---

\*Equal contribution.

†Corresponding authors.

<sup>‡</sup>PCA Lab, Key Lab of Intelligent Perception and Systems for High-Dimensional Information of Ministry of Education, School of Computer Science and Engineering, Nanjing University of Science and Technology.

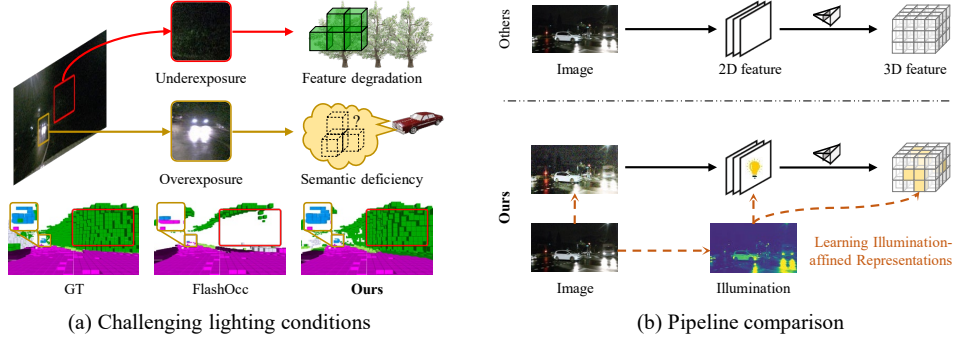


Figure 1: (a) Nighttime images often suffer from both underexposure and overexposure, resulting in feature degradation and semantic deficiency. (b) Our approach learns illumination-aware representations to enhance both the fundamental 2D and 3D stages.

To address these challenges, we propose LIAR, a novel framework that learns illumination-affined representations to enhance and leverage nighttime visual information. LIAR introduces a Selective Low-light Image Enhancement (SLLIE) module, which first evaluates the global illumination of a nighttime image using priors derived from daytime scenes. This selective strategy ensures that only genuinely low-light images are enhanced, enabling the model to focus on challenging inputs while avoiding over-amplification of well-lit images that could otherwise lead to contrast degradation or overexposure. As shown in Fig. 1(b), building on the illumination maps produced by SLLIE, LIAR further incorporates two illumination-aware components, 2D Illumination-guided Sampling (2D-IGS) and 3D Illumination-driven Projection (3D-IDP), to respectively address underexposure and overexposure. In the 2D occupancy stage, 2D-IGS leverages illumination maps to adaptively generate feature sampling points, assigning larger offsets to darker regions and smaller offsets to brighter regions. This approach enables the model to effectively compensate for underexposed regions by aggregating semantic cues from adequately-exposed regions, thereby mitigating feature degradation and improving feature quality. In the 3D occupancy stage, 3D-IDP constructs 3D illumination intensity fields to refine the projection process by placing greater emphasis on overexposed regions. This targeted strategy mitigates the adverse effects of semantic deficiency caused by overexposure. Concurrently, these three modules enable LIAR to effectively address both global and local illumination challenges, improving occupancy prediction in complex nighttime environments.

In summary, our contributions are summarized as follows:

- To the best of our knowledge, we are the first to propose LIAR, a novel illumination-driven occupancy prediction framework tailored to the challenges of nighttime environments.
- We introduce three key components: SLLIE, 2D-IGS, and 3D-IDP. SLLIE enables global yet targeted image enhancement, while 2D-IGS and 3D-IDP respectively mitigate the negative effects of underexposure and overexposure in local regions.
- Extensive experiments on both real and synthetic datasets demonstrate the superiority of our approach, achieving up to a 7.9-point improvement over the second-best method. Source code and pretrained models are released for peer research.

## 2 Related Work

**Vision-based Occupancy Prediction.** Recently, vision-based 3D occupancy prediction has attracted growing interest, with studies exploring both supervised and unsupervised learning paradigms [12, 31, 32, 49, 64, 68]. Among supervised approaches, MonoScene [3] is a pioneering method for monocular occupancy prediction. BEVDet [11] adopts the Lift-Splat-Shoot (LSS) [34] for view transformation. BEVDet4D [9] further explores the temporal fusion strategy by fusing history frames. Building upon the BEVDet [11, 20, 21] series, FlashOcc [63] introduces a channel-to-height mechanism for memory-efficient occupancy prediction. In addition, several transformer-based methods have been proposed. For instance, BEVFormer [25] utilizes spatiotemporal transformers to construct BEV features. SparseOcc [27] takes the first step to explore the fully sparse architecture. For unsupervised methods [13, 66], SelfOcc [13] and OccNeRF [66] are two representative methods that employ volume

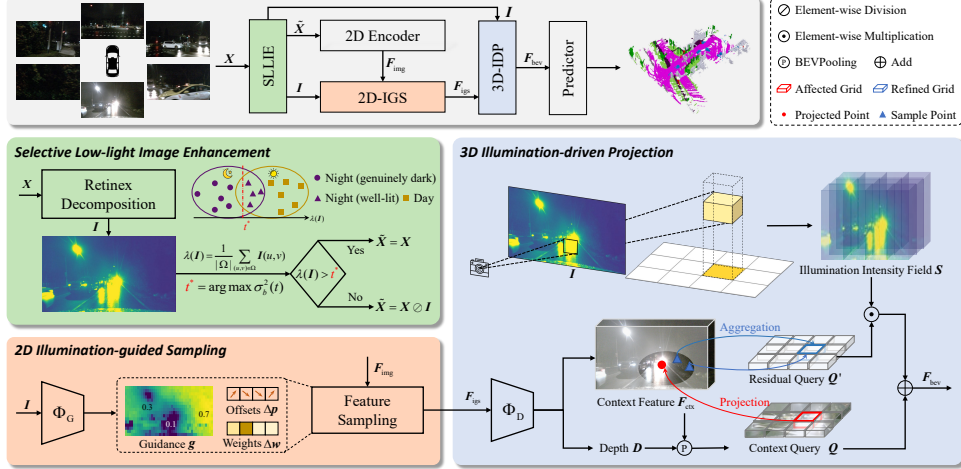


Figure 2: Pipeline of LIAR. The input  $X$  is initially processed by SLLIE to generate an illumination map  $I$  and an enhanced image  $\tilde{X}$ . Using  $I$ , the 2D-IGS module modulates feature sampling to enhance degraded features in underexposed regions, while the 3D-IDP module constructs 3D illumination intensity fields to compensate for deficient semantics in overexposed regions. Finally, a prediction head is deployed to output the occupancy.

rendering to generate self-supervised signals. Recently, Gaussian-based representations [6, 17, 42] have emerged as a powerful approach for 3D scene modeling. For example, GaussianTR [17] aligns Gaussian features with foundation models, achieving state-of-the-art zero-shot performance.

**Low-light Image Enhancement.** Low-light image enhancement [8, 23, 36, 41, 48, 57] aims to improve the brightness and contrast. Histogram Equalization [35], along with its variant CLHE [35], is a classical method for enhancing the global contrast of images. However, their effectiveness diminishes in the presence of significant image noise or non-uniform illumination. Retinex-based methods [36] offer an alternative strategy. Recently, network-based Retinex methods [2, 28, 36, 46] combine convolutional neural networks with Retinex theory to achieve improved accuracy. Nevertheless, in real-world nighttime driving scenarios, artificial light sources often cause spatially uneven illumination [23, 39]. To address this issue, LCDPNet [39] leverages local color distributions to correct illumination imbalance, while CSEC [23] models color distribution shifts to enhance image quality. Different from the aforementioned methods, we do not uniformly treat all nighttime images as requiring enhancement. Instead, we selectively enhance images that suffer from poor illumination, as indiscriminate enhancement may degrade the visual quality of already well-lit ones.

**View Transformation.** Transforming 2D image features into 3D space is a crucial step in 3D perception tasks [58–62]. Bottom-up methods [20, 21, 34] rely on depth estimation [43–45, 56] to project multi-view features into the BEV space, a paradigm first introduced by LSS [34]. However, the resulting BEV features are typically sparse, posing challenges for tasks that require dense spatial representations. In contrast, top-down transformer-based methods [22, 25, 47] generate dense BEV features, but their lack of explicit geometric constraints makes them vulnerable to occlusions and depth inconsistencies. Recent studies [24, 31, 37, 65] attempt to integrate the strengths of both paradigms. For instance, FB-OCC [24] leverages LSS-generated features to initialize transformer queries, while COTR [31] extends this strategy by employing compact occupancy representations to preserve rich geometric information. Nevertheless, these methods primarily address view transformation from a projection-based perspective, neglecting the influence of environmental factors such as illumination.

### 3 Method

#### 3.1 Overview

As illustrated in Fig. 2, our LIAR comprises a SLLIE, 2D encoder, 2D-IGS, 3D-IDP and predictor. The pipeline begins with SLLIE, which takes a nighttime image  $X$  as input and outputs an illumination map  $I$  along with an enhanced image  $\tilde{X}$ . The 2D encoder then extracts image feature  $F_{img}$  from  $\tilde{X}$ .

Guided by the spatial distribution of  $\mathbf{I}$ , 2D-IGS performs adaptive feature sampling on  $\mathbf{F}_{\text{img}}$ , resulting in illumination-aware feature  $\mathbf{F}_{\text{igs}}$ . Subsequently, 3D-IDP integrates  $\mathbf{I}$  and  $\mathbf{F}_{\text{igs}}$  to project features into 3D space. Specifically, DepthNet  $\Phi_D$  is employed to generate context feature  $\mathbf{F}_{\text{ctx}}$  and depth prediction  $\mathbf{D}$ , which are processed through BEVPooling to produce context query  $\mathbf{Q}$ . Meanwhile, 3D illumination intensity field  $\mathbf{S}$  is derived from  $\mathbf{I}$  to modulate deformable querying between  $\mathbf{Q}$  and  $\mathbf{F}_{\text{ctx}}$ . Finally, the resulting BEV feature  $\mathbf{F}_{\text{bev}}$  is passed to the predictor for occupancy prediction.

### 3.2 Selective Low-light Image Enhancement

**Retinex Decomposition.** Based on the Retinex theory [19], given a nighttime image  $\mathbf{X} \in \mathbb{R}^{3 \times H \times W}$ , the Retinex-enhanced image  $\mathbf{X}_{\text{enh}}$  is formulated as:

$$\mathbf{X}_{\text{enh}} = \mathbf{X} \oslash \mathbf{I}, \quad \mathbf{I} \in (0, 1], \quad (1)$$

where  $\mathbf{I} \in \mathbb{R}^{1 \times H \times W}$  represents the illumination map and  $\oslash$  denotes element-wise division. To accurately estimate  $\mathbf{I}$  from the RGB input, we adapt a stage-wise estimation strategy using a self-calibrated module from SCI [30]. Specifically,  $\mathbf{I}$  is derived from the input  $\mathbf{X}$  via a cascaded learning process in a self-supervised manner. To reduce training overhead, we pretrain this module on the nighttime subset of the nuScenes dataset [1] and freeze its weights within the LIAR framework.

**Selective Mechanism.** Leveraging the estimated illumination map  $\mathbf{I}$ , we define an illumination factor  $\lambda(\mathbf{I})$  to quantify the overall image brightness:

$$\lambda(\mathbf{I}) = \frac{1}{|\Omega|} \sum_{(u,v) \in \Omega} \mathbf{I}(u,v), \quad (2)$$

where  $\Omega$  denotes the set of all pixel coordinates in the image. As illustrated in Fig. 3, we assume that daytime images generally exhibit appropriate illumination and can thus serve as a natural reference for well-exposed scenes. Notably, certain nighttime images yield  $\lambda(\mathbf{I})$  values comparable to those of daytime images, indicating that not all nighttime images suffer from insufficient illumination. Based on this observation, nighttime images can be categorized into low-light and well-lit subsets. To avoid unnecessary enhancement, only low-light images should be processed, while well-lit ones remain unchanged. To this end, we aim to determine an illumination threshold  $t^*$  that effectively separates low-light ( $\lambda(\mathbf{I}) \leq t^*$ ) from well-lit ones ( $\lambda(\mathbf{I}) > t^*$ ). Following the principle of maximum inter-class variance [33], we formulate the threshold selection as an optimization problem. For a candidate threshold  $t$ , let  $\omega_0(t)$  and  $\omega_1(t) = 1 - \omega_0(t)$  denote the proportions of images whose illumination factor satisfies  $\lambda(\mathbf{I}) \leq t$  and  $\lambda(\mathbf{I}) > t$ , respectively. Let  $\mu_0(t)$  and  $\mu_1(t)$  denote the corresponding mean illumination values for the two groups. The global mean illumination is computed as:

$$\mu_T = \omega_0(t)\mu_0(t) + \omega_1(t)\mu_1(t), \quad (3)$$

and the inter-class variance is given by:

$$\sigma_b^2(t) = \omega_0(t) (\mu_0(t) - \mu_T)^2 + \omega_1(t) (\mu_1(t) - \mu_T)^2. \quad (4)$$

The optimal threshold  $t^*$  is obtained by maximizing the inter-class variance:

$$t^* = \arg \max_t \sigma_b^2(t). \quad (5)$$

This threshold separates low-light nighttime images from well-lit ones, enabling selective enhancement while preventing over-processing. The enhanced output  $\tilde{\mathbf{X}}$  is given by:

$$\tilde{\mathbf{X}} = \begin{cases} \mathbf{X}_{\text{enh}}, & \text{if } \lambda(\mathbf{I}) \leq t^*, \\ \mathbf{X}, & \text{otherwise.} \end{cases} \quad (6)$$

### 3.3 2D Illumination-guided Sampling

As illustrated in Fig. 2, the 2D-IGS module enhances 2D image feature representations by learning adaptive sampling points from the illumination maps. Specifically, given an illumination map  $\mathbf{I}$ , we first apply a lightweight guidance network  $\Phi_G$  to downsample its resolution, producing  $\mathbf{I}'$  with the



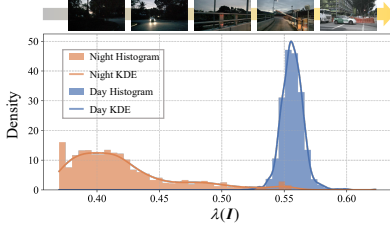


Figure 3: Distributions of the illumination factor in night and day scenes.

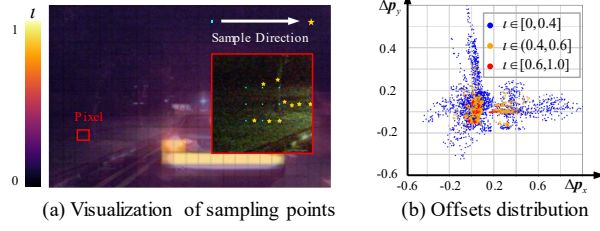


Figure 4: (a) Visualization of sampling points in low-light regions. (b) Statistical analysis of the offset distribution.

same resolution as the image feature  $F_{\text{img}}$ . Subsequently, we compute the illumination guidance map  $g$  by applying min-max normalization to the inverse of  $I'$ :

$$g = \frac{I'^{-1} - \min(I'^{-1})}{\max(I'^{-1}) - \min(I'^{-1})}. \quad (7)$$

Next, we map  $I'$  to offset  $\Delta p = (\Delta p_x, \Delta p_y)$  and weight  $\Delta w$  via one convolution layer, allowing illumination-aware control over the receptive field. Then, we modulate  $\Delta p$  using  $g$ :

$$\Delta \tilde{p} = \Delta p \odot g, \quad (8)$$

where  $\odot$  denotes element-wise multiplication. Finally, these illumination-aware parameters are then utilized in a deformable convolution [5] operation  $\mathcal{F}_{\text{dcn}}(\cdot)$ , which performs differentiable warping on the image feature  $F_{\text{img}}$ . The illumination-aware feature  $F_{\text{igs}}$  is computed as:

$$F_{\text{igs}} = \mathcal{F}_{\text{dcn}}(F_{\text{img}}, p + \Delta \tilde{p}) \odot \Delta w + F_{\text{img}}, \quad (9)$$

where  $p$  denotes the regular grid of pixel coordinates. Fig.4(a) provides an intuitive visualization of this process, where the sampled points in underexposed regions generate offsets toward adequately-exposed regions. Furthermore, we perform a statistical analysis to investigate the relationship between the offset magnitude at each pixel and its corresponding illumination value, where  $\iota$  denotes the pixel value in  $I$ . As shown in Fig.4(b), pixels in adequately-exposed regions (red and yellow dots) exhibit minor offsets, while those in underexposed regions (blue dots) tend to have significantly larger offsets. These results indicate that 2D-IGS efficiently mitigates feature degradation in 2D representations.

### 3.4 3D Illumination-driven Projection

**Illumination Modeling.** For each BEV grid centered at  $(x, y)$ , we first uniformly sample  $N_z$  points along the vertical axis, yielding heights  $\{z_j\}_{j=1}^{N_z}$  within a predefined range. Each 3D point  $(x, y, z_j)$  is projected onto the 2D image plane using the camera projection matrix  $M \in \mathbb{R}^{3 \times 4}$ :

$$d_j [u_j \ v_j \ 1]^\top = M [x \ y \ z_j \ 1]^\top, \quad \forall j \in \{1, \dots, N_z\}, \quad (10)$$

where  $(u_j, v_j)$  are the projected pixel coordinates, and  $d_j$  represents the corresponding depth. The illumination intensity at each height is then sampled from the illumination map  $I$  at position  $(u_j, v_j)$ . Finally, the aggregated illumination value at the BEV location  $(x, y)$  is given by:

$$S(x, y) = \frac{1}{N_z} \sum_{j=1}^{N_z} I(\lfloor v_j \rfloor, \lfloor u_j \rfloor), \quad (11)$$

where  $\lfloor \cdot \rfloor$  denotes the floor operation to discretize continuous coordinates into pixel indices. This aggregated  $S$  provides an informative estimation of 3D illumination distribution, enabling the model to localize overexposed regions and emphasize them during subsequent feature reasoning.

**BEV Context Refinement.** As shown in Fig. 2, given  $F_{\text{igs}}$  as input, the DepthNet  $\Phi_D$  is first employed to predict context-aware feature  $F_{\text{ctx}}$  and depth  $D$ . Subsequently, BEVPooling [10] projects  $F_{\text{ctx}}$  into BEV space based on  $D$ , producing context query  $Q$ . To refine the BEV context features, we propose constructing illumination intensity fields to emphasize overexposed regions. Specifically, each BEV query position  $(x, y)$  is first lifted into a vertical set of  $N_z$  points  $\{z_j\}_{j=1}^{N_z}$

Table 1: Quantitative comparison on the Occ3D-nuScenes dataset. "1f" denotes single-frame method. "2f", "4f" and "8f" denotes methods fusing temporal information from 2, 4 and 8 frames.

Method	mIoU	others	barrier	bicycle	car	motorcycle	pedestrian	traffic cone	truck	drive. surf.	other flat	sidewalk	terrain	manmade	vegetation	Venue
<i>Train: night    Test: night</i>																
BEVDetOcc (1f) [11]	13.22	0.0	0.0	0.0	34.5	0.0	0.0	0.0	0.1	<b>58.7</b>	1.0	<u>24.1</u>	<u>26.0</u>	<u>19.7</u>	<u>21.1</u>	arXiv22
FlashOcc (1f) [63]	<u>13.42</u>	0.0	0.0	0.0	35.3	0.0	0.6	0.0	3.6	<u>57.2</u>	2.3	21.8	25.9	19.6	<b>21.7</b>	arXiv23
SparseOcc (1f) [27]	13.32	<b>10.4</b>	<u>0.5</u>	<u>6.4</u>	<u>37.3</u>	<u>10.9</u>	<u>8.2</u>	<b>1.4</b>	<b>13.1</b>	46.4	<u>4.4</u>	19.7	9.6	8.8	9.3	ECCV24
OPUS (1f) [40]	11.32	0.2	0.0	0.0	33.2	4.6	0.1	0.0	<u>13.0</u>	50.9	1.0	15.0	13.5	11.4	15.8	NIPS24
LIAR (1f)	<b>19.27</b>	<u>9.7</u>	<b>11.2</b>	<b>8.5</b>	<b>37.7</b>	<b>14.8</b>	<b>12.2</b>	<u>1.1</u>	12.5	<b>58.7</b>	<b>8.5</b>	<b>27.2</b>	<b>27.2</b>	<b>20.1</b>	20.6	NIPS25
BEVDetOcc (2f) [11]	15.86	0.6	0.0	0.0	40.3	0.0	6.0	0.0	4.5	62.0	3.0	28.3	28.8	23.7	<u>24.8</u>	arXiv21
BEVFormer (2f) [25]	16.57	3.6	0.0	0.0	40.3	16.1	9.9	0.0	10.1	62.1	4.8	19.9	24.4	18.8	22.2	ECCV22
FlashOcc (2f) [63]	18.15	5.4	0.0	0.0	41.5	9.5	10.5	0.0	<u>17.7</u>	<u>63.5</u>	1.7	26.8	28.0	25.0	24.6	arXiv23
FBOcc (2f) [24]	19.79	9.3	<b>16.3</b>	5.4	40.0	13.6	<u>12.5</u>	0.1	17.6	59.4	<u>6.8</u>	24.5	<u>29.7</u>	20.0	21.9	CVPR23
OPUS (2f) [40]	12.77	1.5	0.0	0.0	35.4	8.5	1.1	0.0	17.0	52.9	2.2	16.6	14.6	13.0	16.1	NIPS24
SparseOcc (2f) [27]	14.29	12.0	1.9	<u>7.9</u>	37.9	9.8	9.3	<u>0.4</u>	16.6	46.9	4.8	21.8	11.4	9.7	9.9	ECCV24
COTR (2f) [31]	<u>20.01</u>	<b>15.3</b>	0.3	1.5	<b>44.0</b>	<u>18.1</u>	10.1	0.0	7.7	63.2	<b>7.0</b>	<b>30.3</b>	<b>31.1</b>	<u>25.1</u>	<b>26.6</b>	CVPR24
LIAR (2f)	<b>22.09</b>	<u>13.0</u>	<u>5.3</u>	<b>13.5</b>	<u>42.8</u>	<b>19.3</b>	<b>18.6</b>	<b>1.1</b>	<b>20.2</b>	<b>64.3</b>	2.8	<u>29.0</u>	29.3	<b>25.4</b>	24.7	NIPS25
<i>Train: day &amp; night    Test: night</i>																
BEVDetOcc (1f) [11]	<u>18.96</u>	<u>4.7</u>	22.5	<u>2.6</u>	38.5	6.6	<u>6.5</u>	<u>0.0</u>	12.5	<b>63.6</b>	<u>5.7</u>	<b>29.0</b>	28.8	21.3	<b>23.1</b>	arXiv22
FlashOcc (1f) [63]	18.93	4.3	<u>1.0</u>	0.0	<u>38.7</u>	<u>7.7</u>	5.8	<u>0.0</u>	<u>13.6</u>	<u>60.6</u>	2.1	27.1	<u>30.0</u>	<b>21.7</b>	<u>22.5</u>	arXiv23
LIAR (1f)	<b>23.67</b>	<b>12.0</b>	<b>36.1</b>	<b>14.5</b>	<b>40.7</b>	<b>19.2</b>	<b>14.0</b>	<b>1.3</b>	<b>19.3</b>	60.5	<b>11.1</b>	<u>28.9</u>	<b>30.5</b>	<u>21.4</u>	21.8	NIPS25
BEVDetOcc (2f) [11]	21.98	9.1	16.5	2.4	44.4	8.2	11.4	0.0	28.9	64.6	8.6	29.9	31.4	26.1	<b>26.3</b>	arXiv21
BEVFormer (4f) [25]	13.77	3.1	19.8	0.7	44.1	16.6	<u>14.5</u>	0.0	22.3	35.8	5.6	13.9	8.4	3.1	5.0	ECCV22
FlashOcc (2f) [63]	23.40	13.4	18.5	2.3	<b>46.5</b>	10.7	14.3	0.0	<u>30.0</u>	<u>66.5</u>	5.9	<b>32.5</b>	<u>32.9</u>	<b>28.1</b>	<u>26.1</u>	arXiv23
OPUS (8f) [40]	20.28	14.0	10.4	12.1	40.4	15.2	13.0	0.0	27.9	62.9	5.0	25.5	19.8	17.1	20.6	NIPS24
SparseOcc (8f) [27]	22.79	15.8	<u>40.0</u>	<b>23.0</b>	43.5	15.3	13.6	<u>0.4</u>	28.9	58.8	<u>10.2</u>	26.2	16.7	13.2	13.5	ECCV24
COTR (2f) [31]	<u>25.17</u>	<b>16.0</b>	<b>41.5</b>	8.9	42.1	<u>17.4</u>	12.5	0.0	26.6	65.2	<b>11.2</b>	27.0	<b>33.7</b>	24.8	25.6	CVPR24
LIAR (2f)	<b>27.33</b>	<u>15.9</u>	37.7	<u>19.0</u>	<u>45.4</u>	<b>19.1</b>	<b>17.8</b>	<b>1.9</b>	<b>33.6</b>	<b>67.1</b>	8.1	<u>31.2</u>	<b>33.7</b>	<u>27.4</u>	24.7	NIPS25

along the height dimension. For each point  $(x, y, z_j)$ , its corresponding 2D coordinates are computed using the projection function  $\mathcal{P}$  defined in Eq. 10. The residual query  $\mathbf{Q}'$  at  $(x, y)$  is defined as:

$$\mathbf{Q}'(x, y) = \sum_{j=1}^{N_z} \mathcal{F}_{\text{dca}}(\mathbf{Q}(x, y), \mathcal{P}(x, y, z_j), \mathbf{F}_{\text{ctx}}). \quad (12)$$

In this formulation,  $\mathcal{F}_{\text{dca}}$  denotes the deformable cross-attention module [67], which leverages the query  $\mathbf{Q}(x, y)$  to sample relevant context feature  $\mathbf{F}_{\text{ctx}}$  around the projected coordinates. Finally, the refined BEV feature  $\mathbf{F}_{\text{bev}}$  is given by:

$$\mathbf{F}_{\text{bev}} = \mathbf{Q} + \mathbf{Q}' \odot \mathbf{S}, \quad (13)$$

where the illumination intensity field  $\mathbf{S}$  acts as a spatial weighting factor.

### 3.5 Training Loss

First, we employ a weighted cross-entropy loss  $\mathcal{L}_{\text{ce}}$  to supervise the learning of the predictor:

$$\mathcal{L}_{\text{ce}} = - \sum_{v=1}^{N_{\text{vox}}} \sum_{m=1}^{N_{\text{cla}}} c \hat{r}_{v,m} \log \left( \frac{e^{r_{v,m}}}{\sum_m e^{r_{v,m}}} \right), \quad (14)$$

where  $N_{\text{vox}}$  and  $N_{\text{cla}}$  denote the total number of voxels and classes. Here,  $r_{v,m}$  is the prediction logit for  $v$ -th voxel belonging to class  $m$ ,  $\hat{r}_{v,m}$  is the corresponding label. The weight  $c$  is a class-wise balancing factor computed as the inverse of the class frequency. In addition, inspired by MonoScene [3], we incorporate two auxiliary losses:  $\mathcal{L}_{\text{scal}}^{\text{sem}}$  and  $\mathcal{L}_{\text{scal}}^{\text{geo}}$ , which respectively regularize the semantic and geometric consistency of the predictions. Finally, the total training loss  $\mathcal{L}_{\text{t}}$  is formulated as:

$$\mathcal{L}_{\text{t}} = \alpha \mathcal{L}_{\text{ce}} + \beta \mathcal{L}_{\text{scal}}^{\text{sem}} + \gamma \mathcal{L}_{\text{scal}}^{\text{geo}}, \quad (15)$$

where  $\alpha$ ,  $\beta$  and  $\gamma$  are hyper-parameters and we empirically set to 10, 0.2, and 0.2, respectively.

Table 2: Quantitative comparison on the nuScenes-C dataset under three severity levels.

Method	mIoU	others	barrier	bicycle	bus	car	const. veh.	motorcycle	pedestrian	traffic cone	trailer	truck	drive. surf.	other flat	sidewalk	terrain	manmade	vegetation
<i>Severity: easy</i>																		
BEVDetOcc (1f) [11]	<u>15.08</u>	<u>0.5</u>	<u>12.4</u>	1.3	14.1	<u>24.9</u>	<u>7.7</u>	<u>3.8</u>	<u>7.8</u>	<u>4.5</u>	2.1	<u>10.1</u>	<u>62.5</u>	<u>12.0</u>	<u>25.5</u>	<u>32.0</u>	<u>17.1</u>	<u>18.2</u>
FlashOcc (1f) [63]	12.47	0.2	7.5	<u>1.9</u>	<u>16.1</u>	19.1	6.6	3.0	5.6	3.4	<u>4.6</u>	9.1	49.7	9.9	19.7	25.8	14.9	14.7
LIAR (1f)	<b>22.52</b>	<b>3.7</b>	<b>19.4</b>	<b>12.1</b>	<b>24.4</b>	<b>31.8</b>	<b>11.0</b>	<b>14.8</b>	<b>16.3</b>	<b>14.4</b>	<b>11.6</b>	<b>17.9</b>	<b>67.9</b>	<b>21.5</b>	<b>34.1</b>	<b>38.2</b>	<b>21.8</b>	<b>22.2</b>
BEVDetOcc (2f) [11]	20.22	0.9	18.2	4.3	20.4	35.6	13.2	8.3	12.4	9.3	3.4	18.3	67.2	14.4	29.2	35.6	28.0	24.9
BEVFormer (4f) [25]	14.13	0.7	20.4	3.8	28.9	33.6	4.4	8.3	11.3	7.8	10.2	18.0	34.4	14.0	17.4	11.2	7.2	8.6
FlashOcc (2f) [63]	21.83	1.7	19.5	7.7	24.0	35.0	14.6	11.6	14.3	14.3	4.4	19.0	63.3	21.3	33.2	33.3	28.9	<u>25.1</u>
OPUS (8f) [40]	<u>23.63</u>	<u>4.0</u>	22.2	14.0	<u>31.7</u>	<u>37.5</u>	<u>16.3</u>	15.5	13.3	13.8	<u>12.0</u>	<u>24.0</u>	65.6	<u>23.1</u>	<u>34.4</u>	29.4	21.2	23.7
SparseOcc (8f) [27]	21.98	3.9	23.0	<u>16.6</u>	28.7	37.5	13.2	<u>18.6</u>	<u>18.9</u>	<b>24.7</b>	7.4	21.8	58.0	18.9	26.9	23.3	16.5	15.8
COTR (2f) [31]	21.36	1.3	<u>23.6</u>	10.2	16.0	31.4	8.2	11.1	17.1	18.5	3.3	15.0	<u>67.6</u>	19.0	33.0	<u>35.9</u>	<u>29.4</u>	22.5
LIAR (2f)	<b>30.66</b>	<b>5.1</b>	<b>31.6</b>	<b>17.2</b>	<b>32.2</b>	<b>43.8</b>	<b>16.7</b>	<b>19.8</b>	<b>21.5</b>	<u>23.9</u>	<b>20.0</b>	<b>27.6</b>	<b>75.5</b>	<b>33.1</b>	<b>42.5</b>	<b>45.0</b>	<b>35.7</b>	<b>30.1</b>
<i>Severity: moderate</i>																		
BEVDetOcc (1f) [11]	<u>11.50</u>	<u>0.3</u>	<u>6.6</u>	<u>0.8</u>	10.0	<u>20.7</u>	<u>6.1</u>	<u>2.8</u>	<u>6.0</u>	<u>2.2</u>	1.0	<u>5.5</u>	<u>56.5</u>	<u>5.7</u>	<u>18.2</u>	<u>24.8</u>	<u>13.6</u>	<u>14.8</u>
FlashOcc (1f) [63]	9.05	0.0	3.1	0.7	<u>13.4</u>	15.4	5.9	1.7	4.3	1.7	<u>1.5</u>	<u>5.5</u>	42.0	3.3	13.2	18.4	11.9	11.9
LIAR (1f)	<b>18.44</b>	<b>2.2</b>	<b>11.1</b>	<b>11.1</b>	<b>23.0</b>	<b>28.0</b>	<b>8.5</b>	<b>13.8</b>	<b>14.0</b>	<b>12.5</b>	<b>5.0</b>	<b>13.7</b>	<b>62.8</b>	<b>12.0</b>	<b>27.5</b>	<b>31.4</b>	<b>17.8</b>	<b>19.0</b>
BEVDetOcc (2f) [11]	15.95	0.4	10.3	3.2	14.9	31.8	10.1	5.7	9.7	4.9	1.7	12.6	62.4	8.5	22.2	28.8	23.6	<u>20.5</u>
BEVFormer (4f) [25]	9.98	0.4	11.6	2.9	25.2	29.5	1.5	5.4	8.6	2.7	<u>3.8</u>	11.9	27.3	8.1	11.9	7.5	4.8	6.8
FlashOcc (2f) [63]	17.34	0.9	12.6	5.4	17.6	31.4	10.7	8.6	11.8	9.2	2.2	12.7	60.6	<u>14.0</u>	26.0	28.1	23.2	19.9
OPUS (8f) [40]	17.50	<u>2.0</u>	11.6	9.6	<u>25.9</u>	32.9	<b>13.4</b>	12.1	10.7	8.2	3.3	<u>17.3</u>	59.4	13.6	24.8	18.4	15.4	18.9
SparseOcc (8f) [27]	16.81	1.9	13.6	<u>11.7</u>	22.8	<u>33.5</u>	9.2	15.5	<u>16.0</u>	<u>19.2</u>	2.0	16.2	53.2	8.3	19.3	16.3	13.9	13.1
COTR (2f) [31]	<u>17.77</u>	0.8	<u>18.2</u>	8.6	11.4	28.4	6.8	9.7	14.1	15.0	1.3	11.3	<u>62.9</u>	13.2	<u>27.1</u>	<u>29.9</u>	<u>25.1</u>	18.3
LIAR (2f)	<b>25.67</b>	<b>2.7</b>	<b>20.0</b>	<b>13.2</b>	<b>28.7</b>	<b>39.8</b>	<u>12.8</u>	<b>17.6</b>	<b>18.5</b>	<b>20.2</b>	<b>11.9</b>	<b>22.8</b>	<b>72.0</b>	<b>25.6</b>	<b>37.0</b>	<b>39.6</b>	<b>29.6</b>	<b>24.7</b>
<i>Severity: hard</i>																		
BEVDetOcc (1f) [11]	<u>7.81</u>	<u>0.0</u>	<u>2.1</u>	<u>0.3</u>	5.2	<u>14.6</u>	<u>3.1</u>	<u>1.4</u>	<u>3.2</u>	<u>0.8</u>	0.3	2.1	<u>49.4</u>	<u>1.4</u>	<u>11.2</u>	<u>16.0</u>	<u>10.2</u>	<u>11.6</u>
FlashOcc (1f) [63]	5.82	<u>0.0</u>	0.7	0.1	<u>7.9</u>	11.0	1.3	0.5	2.4	0.4	<u>0.5</u>	<u>2.2</u>	34.3	0.2	7.3	11.0	9.2	9.9
LIAR (1f)	<b>12.21</b>	<b>0.8</b>	<b>4.7</b>	<b>7.4</b>	<b>15.4</b>	<b>20.9</b>	<b>3.7</b>	<b>9.1</b>	<b>8.8</b>	<b>7.5</b>	<b>2.2</b>	<b>6.8</b>	<b>53.3</b>	<b>2.2</b>	<b>16.1</b>	<b>21.9</b>	<b>12.3</b>	<b>14.5</b>
BEVDetOcc (2f) [11]	10.46	0.1	4.2	1.6	7.1	23.9	3.8	2.5	5.5	1.4	0.6	5.1	53.1	2.8	14.1	19.5	17.5	<u>14.9</u>
BEVFormer (4f) [25]	5.71	0.2	<u>5.1</u>	0.9	13.1	22.3	0.2	2.4	4.8	0.6	0.5	5.7	19.4	2.8	6.3	5.1	3.0	4.7
FlashOcc (2f) [63]	11.76	0.2	4.8	2.0	9.7	25.0	5.0	5.2	7.2	3.9	<u>0.8</u>	5.3	<u>56.3</u>	4.9	16.5	<u>22.3</u>	16.4	14.3
OPUS (8f) [40]	10.80	<u>0.5</u>	3.1	5.8	<b>14.7</b>	25.0	<b>10.6</b>	5.3	6.4	2.9	0.5	<u>8.6</u>	49.7	3.9	13.6	10.3	9.0	13.7
SparseOcc (8f) [27]	10.77	<b>0.8</b>	<u>5.1</u>	<u>7.9</u>	10.8	<u>25.9</u>	4.2	<u>8.2</u>	<u>11.3</u>	<u>11.0</u>	0.3	8.3	45.2	1.0	11.6	10.6	10.9	10.0
COTR (2f) [31]	<u>12.01</u>	0.4	<b>9.1</b>	6.0	3.9	23.4	1.7	6.4	8.3	9.2	0.4	5.6	53.9	<u>5.3</u>	<u>18.6</u>	21.0	<u>18.2</u>	13.0
LIAR (2f)	<b>17.31</b>	<b>0.8</b>	<b>9.1</b>	<b>8.3</b>	<u>14.2</u>	<b>31.5</b>	<u>9.0</u>	<b>10.2</b>	<b>12.3</b>	<b>13.2</b>	<b>2.3</b>	<b>13.1</b>	<b>63.0</b>	<b>14.4</b>	<b>26.0</b>	<b>30.1</b>	<b>19.5</b>	<b>17.6</b>

## 4 Experiments

**Dataset.** We evaluate our method on both real and synthetic nighttime scenarios. (1) **Occ3D-nuScenes** [38] includes 700 scenes for training and 150 for validation, with annotations spanning a spatial range of -40m to 40m along both the X and Y axes, and -1m to 5.4m along the Z axis. The occupancy labels are defined using voxels with a size of  $0.4\text{m} \times 0.4\text{m} \times 0.4\text{m}$ , containing 17 categories. Notably, the training and validation sets include 84 and 15 real-world nighttime scenes, respectively. (2) **nuScenes-C** [51] is a synthetic benchmark that introduces eight types of data corruptions, each applied at three intensity levels to the validation set of nuScenes. Among them, the *Dark* corruption simulates nighttime by reducing brightness and contrast and adding random noise.

**Implementation Details.** We present two variants of LIAR: a non-temporal version and a temporal version that incorporates one historical frame. Both of them are built upon the FlashOcc [63] series and utilize ResNet-50 [7] as the 2D encoder. During training, we employ the AdamW optimizer [29] with a learning rate of  $2 \times 10^{-4}$ , training each model for 24 epochs. Our implementation is based on MMDetection3D [4], and experiments are conducted on 4 NVIDIA GeForce RTX 4090 GPUs.

### 4.1 Comparison with the State-of-the-Art

To comprehensively evaluate the performance under nighttime conditions, we design two experimental settings. (1) We train the models on the Occ3d-nuScenes dataset using either only nighttime data or the full dataset (daytime + nighttime), and evaluate both models on the nighttime subset (see Tab. 1). (2) We take the model trained on the full dataset and evaluate it on the nuScenes-C (see Tab. 2).

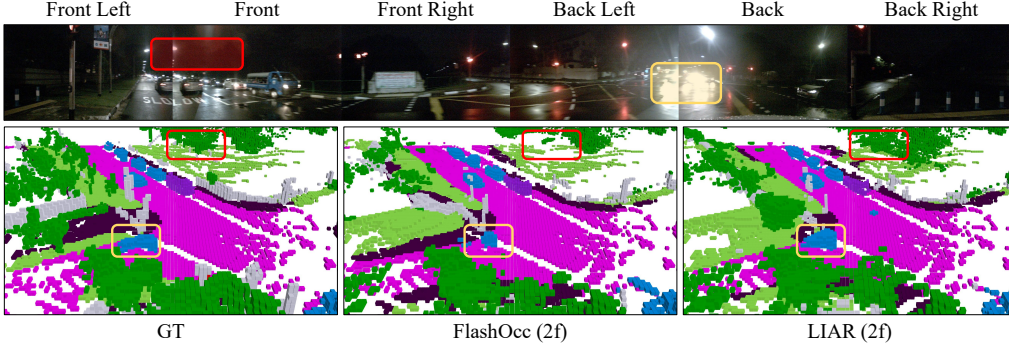


Figure 5: Visual results with underexposure and overexposure on the Occ3D-nuScenes dataset.

Table 3: Ablation study of LIAR on the Occ3D-nuScenes.

LIAR	SLLIE	2D Feature Enh.			View Transformation		mIoU
		Add	Concat	2D-IGS	BEVPooling	3D-IDP	
baseline					✓		13.42
i	✓				✓		14.39
ii		✓			✓		13.86
iii			✓		✓		13.81
iv				✓	✓		14.11
v						✓	14.88
vi	✓			✓		✓	15.31



Figure 6: Adverse effect of indiscriminate enhancement.

**Comparison on Real-world Dataset.** We first evaluate our LIAR on the Occ3d-nuScenes dataset. As shown in Tab. 1, LIAR consistently achieves superior performance across both training settings. Notably, when trained solely on nighttime data, our model achieves the best performance, outperforming the second-best approach by 5.85 mIoU. Additionally, when a single history frame is introduced, LIAR (2f) surpasses the suboptimal method (COTR (2f) [31]) by 2.08 in mIoU. Furthermore, when training on the full dataset, LIAR continues to demonstrate its superiority. As illustrated in Fig. 5, LIAR (2f) delivers superior results, particularly in challenging regions such as *car* and *vegetation*, where overexposure and poor visibility degrade the performance of FlashOcc (2f) [63].

**Comparison on Synthetic Dataset.** Given the limited availability of real-world nighttime data, we further evaluate our LIAR on the nuScenes-C dataset. As shown in Tab. 2, LIAR outperforms all competing methods across all three severity levels. Under the *easy* setting, LIAR achieves the best performance in both the single-frame and temporal fusion configurations, surpassing the second-best models by 7.44 and 7.03, respectively. Notably, under the *moderate* and *hard* settings, our single-frame model even outperforms methods that incorporate temporal fusion. For instance, under the *hard* setting, LIAR (1f) outperforms BEVFormer (4f) and SparseOcc (8f) by 6.50 and 1.44 mIoU, respectively. Fig. 7 presents visual comparisons across all three severity levels, illustrating that our model maintains strong performance even under extremely low-light conditions.

## 4.2 Ablation Study

All experiments in this section are conducted on the nighttime split of the Occ3D-nuScenes dataset, with all models are trained exclusively using the cross-entropy loss.

**LIAR Designs.** Tab. 3 presents the ablation results of LIAR. The baseline model excludes SLLIE, 2D-IGS, and 3D-IDP, adopting BEVPooling for view transformation. Building on this, LIAR-i introduces SLLIE for low-light image enhancement, resulting in a 0.97 mIoU improvement. To investigate illumination integration for 2D feature enhancement, LIAR-ii and LIAR-iii apply naive fusion methods by directly adding or concatenating illumination maps with extracted 2D image features, yielding marginal gains of 0.44 and 0.39 mIoU, respectively. In contrast, LIAR-iv replaces these naive approaches with the proposed 2D-IGS, boosting performance to 14.11 mIoU and underscoring the benefits of guided sampling. Furthermore, LIAR-v incorporates 3D-IDP to address overexposure in the 3D space, achieving an additional 1.46 mIoU improvement over the baseline. Finally, LIAR-vi

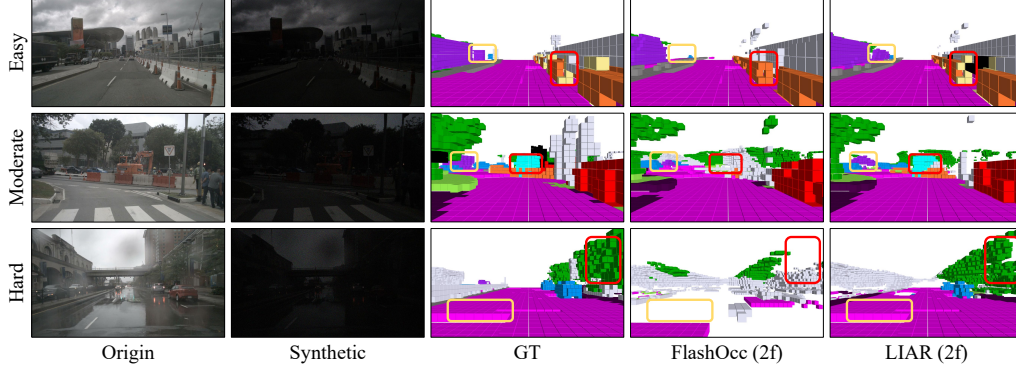


Figure 7: Visual comparisons across the three severity levels on the nuScenes-C dataset.

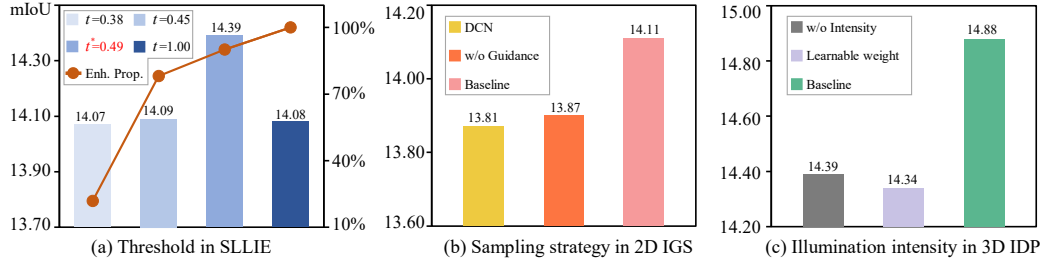


Figure 8: Ablation study on the impact of illumination in our SLLIE, 2D-IGS, and 3D-IDP.

integrates SLLIE, 2D-IGS, and 3D-IDP simultaneously, attaining the best overall performance with a 1.89 mIoU gain over the baseline. Overall, each component positively impacts the baseline.

**SLLIE.** Fig. 8(a) presents the ablation study of illumination threshold in SLLIE. The baseline is LIAR-i in Tab. 3. We manually select several illumination thresholds  $t$  (0.38, 0.45, 1.00) to compare against the statistically derived optimal  $t^*$ . The corresponding line graph illustrates how the proportion of enhanced nighttime images varies with different thresholds. As the proportion increases, mIoU also improves, peaking at  $t = t^*$ . However, when  $t = 1.00$ , meaning all nighttime images are enhanced indiscriminately, the mIoU decreases from 14.39 to 14.08. This performance drop occurs due to over-brightening images that already possess adequate illumination, which introduces artifacts, amplifies noise, and ultimately deteriorates the overall visual quality, as illustrated in Fig. 6.

**2D-IGS.** Fig. 8(b) shows the ablation study in 2D-IGS. The baseline (pink bar) is LIAR-iv in Tab. 3. When the sampling module is replaced with a standard DCN [5], the mIoU decreases from 14.11 to 13.81, confirming the effectiveness of learning sampling positions according to illumination maps. Additionally, removing only the illumination guidance from the baseline model results in a 0.24 decrease in mIoU, indicating that the guidance map provides essential spatial priors that modulate the sampling offsets. These results validate the rational behind illumination-guided sampling.

**3D-IDP.** Fig. 8(c) illustrates the ablation of illumination intensity field in 3D-IDP. The baseline (green bar) is LIAR-v in Tab. 3. First, removing the illumination intensity field leads to a performance drop from 14.88 to 14.39 mIoU. To further demonstrate the effectiveness of modeling 3D illumination, we design a network that generates BEV weights as a substitute, which results in a decrease of 0.54 mIoU. These results underscore the positive impact of explicitly modeling illumination in 3D space.

## 5 Conclusion

We propose LIAR in this work. Our LIAR is the first occupancy framework that addresses the challenges of nighttime scenes by harnessing illumination-affined representations. LIAR first introduces SLLIE to adaptively enhance limited visibility. Subsequently, 2D-IGS and 3D-IDP are designed to weaken the adverse effects of underexposure and overexposure, respectively. Extensive experiments validate the effectiveness of our LIAR and its superiority in challenging nighttime environments.



**Limitation.** We evaluate our method on both real and synthetic nighttime scenarios using Occ3D-nuScenes [38] and nuScenes-C [50], respectively. For Occ3D-nuScenes, although it serves as the most widely adopted benchmark for occupancy prediction, its nighttime subset is relatively small, and several categories are absent. These limitations may reduce the diversity of evaluated scenarios. As for nuScenes-C, it is a synthetic nighttime dataset generated by darkening daytime images and injecting noise. Due to its synthetic nature, nuScenes-C inevitably differs from real-world data. This domain gap can be broadly categorized into two aspects: First, it retains the semantics and motion patterns of daytime scenes, which differ from actual nighttime driving scenarios. Second, it fails to replicate sensor-induced degradations, such as thermal noise, motion blur, and compression artifacts.

**Broader Impact.** With the increasing development of vision-centric autonomous driving systems, addressing safety concerns under adverse environmental conditions has become critical for ensuring the reliability of autonomous vehicles. LIAR represents a significant step toward robust 3D occupancy prediction in challenging nighttime scenarios. Given the broad applicability of this task, our approach has the potential to benefit a wide range of autonomous driving applications.

## 6 Acknowledgment

This work was supported by the National Science Fund of China under Grant Nos. U24A20330, 62361166670, and 62306155, and by the Technology Major Project No. 2022ZD0116305.

## References

- [1] Holger Caesar, Varun Bankiti, Alex H Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nusenes: A multimodal dataset for autonomous driving. In *CVPR*, pages 11621–11631, 2020.
- [2] Yuanhao Cai, Hao Bian, Jing Lin, Haoqian Wang, Radu Timofte, and Yulun Zhang. Retinex-former: One-stage retinex-based transformer for low-light image enhancement. In *ICCV*, pages 12504–12513, 2023.
- [3] Anh-Quan Cao and Raoul De Charette. Monoscene: Monocular 3d semantic scene completion. In *CVPR*, pages 3991–4001, 2022.
- [4] MMDetection3D Contributors. MMDetection3D: OpenMMLab next-generation platform for general 3D object detection. <https://github.com/open-mmlab/mmdetection3d>, 2020.
- [5] Jifeng Dai, Haozhi Qi, Yuwen Xiong, Yi Li, Guodong Zhang, Han Hu, and Yichen Wei. Deformable convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, pages 764–773, 2017.
- [6] Wanshui Gan, Fang Liu, Hongbin Xu, Ningkai Mo, and Naoto Yokoya. Gaussianocc: Fully self-supervised and efficient 3d occupancy estimation with gaussian splatting. *arXiv preprint arXiv:2408.11447*, 2024.
- [7] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016.
- [8] Mingbo Hong, Shen Cheng, Haibin Huang, Haoqiang Fan, and Shuaicheng Liu. You only look around: Learning illumination invariant feature for low-light object detection. *arXiv preprint arXiv:2410.18398*, 2024.
- [9] Junjie Huang and Guan Huang. Bevdet4d: Exploit temporal cues in multi-camera 3d object detection. *arXiv preprint arXiv:2203.17054*, 2022.
- [10] Junjie Huang and Guan Huang. Bevpoolv2: A cutting-edge implementation of bevdet toward deployment. *arXiv preprint arXiv:2211.17111*, 2022.
- [11] 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.

- [12] Yuanhui Huang, Wenzhao Zheng, Yunpeng Zhang, Jie Zhou, and Jiwen Lu. Tri-perspective view for vision-based 3d semantic occupancy prediction. In *CVPR*, pages 9223–9232, 2023.
- [13] Yuanhui Huang, Wenzhao Zheng, Borui Zhang, Jie Zhou, and Jiwen Lu. Selfocc: Self-supervised vision-based 3d occupancy prediction. In *CVPR*, pages 19946–19956, 2024.
- [14] Bernd Jähne. *Digital image processing*. Springer Science & Business Media, 2005.
- [15] Mingqian Ji, Jian Yang, and Shanshan Zhang. Enhancing pseudo-boxes via data-level lidar-camera fusion for unsupervised 3d object detection. *arXiv preprint arXiv:2508.20530*, 2025.
- [16] Mingqian Ji, Jian Yang, and Shanshan Zhang. Ocrfdet: Object-centric radiance fields for multi-view 3d object detection in autonomous driving. *arXiv preprint arXiv:2506.23565*, 2025.
- [17] Haoyi Jiang, Liu Liu, Tianheng Cheng, Xinjie Wang, Tianwei Lin, Zhizhong Su, Wenyu Liu, and Xinggang Wang. Gausstr: Foundation model-aligned gaussian transformer for self-supervised 3d spatial understanding. *arXiv preprint arXiv:2412.13193*, 2024.
- [18] Jungho Kim, Changwon Kang, Dongyoung Lee, Sehwan Choi, and Jun Won Choi. Protoocc: Accurate, efficient 3d occupancy prediction using dual branch encoder-prototype query decoder. In *AAAI*, pages 4284–4292, 2025.
- [19] Edwin H Land. The retinex theory of color vision. *Scientific american*, 237(6):108–129, 1977.
- [20] Yinhao Li, Han Bao, Zheng Ge, Jinrong Yang, Jianjian Sun, and Zeming Li. Bevstereo: Enhancing depth estimation in multi-view 3d object detection with temporal stereo. In *AAAI*, pages 1486–1494, 2023.
- [21] 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 *AAAI*, pages 1477–1485, 2023.
- [22] 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 *CVPR*, pages 9087–9098, 2023.
- [23] Yiyu Li, Ke Xu, Gerhard Petrus Hancke, and Rynson WH Lau. Color shift estimation-and-correction for image enhancement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 25389–25398, 2024.
- [24] Zhiqi Li, Zhiding Yu, Wenhai Wang, Anima Anandkumar, Tong Lu, and Jose M Alvarez. Fb-bev: Bev representation from forward-backward view transformations. In *ICCV*, pages 6919–6928, 2023.
- [25] Zhiqi Li, Wenhai Wang, Hongyang Li, Enze Xie, Chonghao Sima, Tong Lu, Qiao Yu, and Jifeng Dai. Bevformer: learning bird’s-eye-view representation from lidar-camera via spatiotemporal transformers. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.
- [26] Zhimin Liao, Ping Wei, Shuaijia Chen, Haoxuan Wang, and Ziyang Ren. Stcocc: Sparse spatial-temporal cascade renovation for 3d occupancy and scene flow prediction. *arXiv preprint arXiv:2504.19749*, 2025.
- [27] Haisong Liu, Yang Chen, Haiguang Wang, Zetong Yang, Tianyu Li, Jia Zeng, Li Chen, Hongyang Li, and Limin Wang. Fully sparse 3d occupancy prediction. In *ECCV*, pages 54–71. Springer, 2024.
- [28] Risheng Liu, Long Ma, Jiaao Zhang, Xin Fan, and Zhongxuan Luo. Retinex-inspired unrolling with cooperative prior architecture search for low-light image enhancement. In *CVPR*, pages 10561–10570, 2021.
- [29] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.
- [30] Long Ma, Tengyu Ma, Risheng Liu, Xin Fan, and Zhongxuan Luo. Toward fast, flexible, and robust low-light image enhancement. In *CVPR*, pages 5637–5646, 2022.



- [31] Qihang Ma, Xin Tan, Yanyun Qu, Lizhuang Ma, Zhizhong Zhang, and Yuan Xie. Cotr: Compact occupancy transformer for vision-based 3d occupancy prediction. In *CVPR*, pages 19936–19945, 2024.
- [32] Gyeongrok Oh, Sungjune Kim, Heeju Ko, Hyung-gun Chi, Jinkyu Kim, Dongwook Lee, Daehyun Ji, Sungjoon Choi, Sujin Jang, and Sangpil Kim. 3d occupancy prediction with low-resolution queries via prototype-aware view transformation. *arXiv preprint arXiv:2503.15185*, 2025.
- [33] Nobuyuki Otsu et al. A threshold selection method from gray-level histograms. *Automatica*, 11 (285-296):23–27, 1975.
- [34] Jonah Philion and Sanja Fidler. Lift, splat, shoot: Encoding images from arbitrary camera rigs by implicitly unprojecting to 3d. In *ECCV*, pages 194–210. Springer, 2020.
- [35] Stephen M Pizer, E Philip Amburn, John D Austin, Robert Cromartie, Ari Geselowitz, Trey Greer, Bart ter Haar Romeny, John B Zimmerman, and Karel Zuiderveld. Adaptive histogram equalization and its variations. *Computer vision, graphics, and image processing*, 39(3): 355–368, 1987.
- [36] Shangquan Sun, Wenqi Ren, Jingyang Peng, Fenglong Song, and Xiaochun Cao. Di-retinex: Digital-imaging retinex theory for low-light image enhancement. *arXiv preprint arXiv:2404.03327*, 2024.
- [37] Xin Tan, Wenbin Wu, Zhiwei Zhang, Chaojie Fan, Yong Peng, Zhizhong Zhang, Yuan Xie, and Lizhuang Ma. Geocc: Geometrically enhanced 3d occupancy network with implicit-explicit depth fusion and contextual self-supervision. *IEEE Transactions on Intelligent Transportation Systems*, 2025.
- [38] Xiaoyu Tian, Tao Jiang, Longfei Yun, Yucheng Mao, Huitong Yang, Yue Wang, Yilun Wang, and Hang Zhao. Occ3d: A large-scale 3d occupancy prediction benchmark for autonomous driving. *NeurIPS*, 36:64318–64330, 2023.
- [39] Haoyuan Wang, Ke Xu, and Rynson WH Lau. Local color distributions prior for image enhancement. In *European conference on computer vision*, pages 343–359. Springer, 2022.
- [40] Jiabao Wang, Zhaojiang Liu, Qiang Meng, Liujiang Yan, Ke Wang, Jie Yang, Wei Liu, Qibin Hou, and Ming-Ming Cheng. Opus: occupancy prediction using a sparse set. *arXiv preprint arXiv:2409.09350*, 2024.
- [41] Kun Wang, Zhenyu Zhang, Zhiqiang Yan, Xiang Li, Baobei Xu, Jun Li, and Jian Yang. Regularizing nighttime weirdness: Efficient self-supervised monocular depth estimation in the dark. In *ICCV*, pages 16055–16064, 2021.
- [42] Letian Wang, Seung Wook Kim, Jiawei Yang, Cunjun Yu, Boris Ivanovic, Steven Waslander, Yue Wang, Sanja Fidler, Marco Pavone, and Peter Karkus. Distillnerf: Perceiving 3d scenes from single-glance images by distilling neural fields and foundation model features. *NeurIPS*, 37:62334–62361, 2024.
- [43] Zhengxue Wang, Zhiqiang Yan, and Jian Yang. Sgnet: Structure guided network via gradient-frequency awareness for depth map super-resolution. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 5823–5831, 2024.
- [44] Zhengxue Wang, Zhiqiang Yan, Ming-Hsuan Yang, Jinshan Pan, Jian Yang, Ying Tai, and Guangwei Gao. Scene prior filtering for depth map super-resolution. *CoRR*, 2024.
- [45] Zhengxue Wang, Zhiqiang Yan, Jinshan Pan, Guangwei Gao, Kai Zhang, and Jian Yang. Dornet: A degradation oriented and regularized network for blind depth super-resolution. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 15813–15822, 2025.
- [46] Chen Wei, Wenjing Wang, Wenhan Yang, and Jiaying Liu. Deep retinex decomposition for low-light enhancement. *arXiv preprint arXiv:1808.04560*, 2018.

- [47] Yi Wei, Linqing Zhao, Wenzhao Zheng, Zheng Zhu, Jie Zhou, and Jiwen Lu. Surroundocc: Multi-camera 3d occupancy prediction for autonomous driving. In *ICCV*, pages 21729–21740, 2023.
- [48] Jiangwei Weng, Zhiqiang Yan, Ying Tai, Jianjun Qian, Jian Yang, and Jun Li. Mamballie: Implicit retinex-aware low light enhancement with global-then-local state space. *Advances in Neural Information Processing Systems*, 37:27440–27462, 2024.
- [49] Yuan Wu, Zhiqiang Yan, Zhengxue Wang, Xiang Li, Le Hui, and Jian Yang. Deep height decoupling for precise vision-based 3d occupancy prediction. *arXiv preprint arXiv:2409.07972*, 2024.
- [50] Shaoyuan Xie, Lingdong Kong, Wenwei Zhang, Jiawei Ren, Liang Pan, Kai Chen, and Ziwei Liu. Robobev: Towards robust bird’s eye view perception under corruptions. *arXiv preprint arXiv:2304.06719*, 2023.
- [51] Shaoyuan Xie, Lingdong Kong, Wenwei Zhang, Jiawei Ren, Liang Pan, Kai Chen, and Ziwei Liu. Benchmarking and improving bird’s eye view perception robustness in autonomous driving. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025.
- [52] Huaiyuan Xu, Junliang Chen, Shiyu Meng, Yi Wang, and Lap-Pui Chau. A survey on occupancy perception for autonomous driving: The information fusion perspective. *Information Fusion*, 114:102671, 2025.
- [53] Zhiqiang Yan, Xiang Li, Kun Wang, Zhenyu Zhang, Jun Li, and Jian Yang. Multi-modal masked pre-training for monocular panoramic depth completion. In *European Conference on Computer Vision*, pages 378–395. Springer, 2022.
- [54] Zhiqiang Yan, Kun Wang, Xiang Li, Zhenyu Zhang, Jun Li, and Jian Yang. Rignet: Repetitive image guided network for depth completion. In *European Conference on Computer Vision*, pages 214–230. Springer, 2022.
- [55] Zhiqiang Yan, Xiang Li, Kun Wang, Shuo Chen, Jun Li, and Jian Yang. Distortion and uncertainty aware loss for panoramic depth completion. In *International Conference on Machine Learning*, pages 39099–39109. PMLR, 2023.
- [56] Zhiqiang Yan, Kun Wang, Xiang Li, Zhenyu Zhang, Jun Li, and Jian Yang. Desnet: Decomposed scale-consistent network for unsupervised depth completion. In *Proceedings of the AAAI conference on artificial intelligence*, pages 3109–3117, 2023.
- [57] Zhiqiang Yan, Yupeng Zheng, Deng-Ping Fan, Xiang Li, Jun Li, and Jian Yang. Learnable differencing center for nighttime depth perception. *Visual Intelligence*, 2(1):15, 2024.
- [58] Zhiqiang Yan, Jianhao Jiao, Zhengxue Wang, and Gim Hee Lee. Event-driven dynamic scene depth completion. *arXiv preprint arXiv:2505.13279*, 2025.
- [59] Zhiqiang Yan, Xiang Li, Le Hui, Zhenyu Zhang, Jun Li, and Jian Yang. Rignet++: Semantic assisted repetitive image guided network for depth completion: Z. yan et al. *International Journal of Computer Vision*, pages 1–23, 2025.
- [60] Zhiqiang Yan, Kun Wang, Xiang Li, Guangwei Gao, Jun Li, and Jian Yang. Tri-perspective view decomposition for geometry aware depth completion and super-resolution. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025.
- [61] Zhiqiang Yan, Zhengxue Wang, Haoye Dong, Jun Li, Jian Yang, and Gim Hee Lee. Ducos: Duality constrained depth super-resolution via foundation model. *arXiv preprint arXiv:2503.04171*, 2025.
- [62] Zhiqiang Yan, Zhengxue Wang, Kun Wang, Jun Li, and Jian Yang. Completion as enhancement: A degradation-aware selective image guided network for depth completion. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 26943–26953, 2025.
- [63] Zichen Yu, Changyong Shu, Jiajun Deng, Kangjie Lu, Zongdai Liu, Jiangyong Yu, Dawei Yang, Hui Li, and Yan Chen. Flashocc: Fast and memory-efficient occupancy prediction via channel-to-height plugin. *arXiv preprint arXiv:2311.12058*, 2023.

- [64] Zhu Yu, Bowen Pang, Lizhe Liu, Runmin Zhang, Qiang Li, Si-Yuan Cao, Maochun Luo, Mingxia Chen, Sheng Yang, and Hui-Liang Shen. Language driven occupancy prediction. *arXiv preprint arXiv:2411.16072*, 2024.
- [65] 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.
- [66] Chubin Zhang, Juncheng Yan, Yi Wei, Jiaxin Li, Li Liu, Yansong Tang, Yueqi Duan, and Jiwen Lu. Occnerf: Advancing 3d occupancy prediction in lidar-free environments. *arXiv preprint arXiv:2312.09243*, 2023.
- [67] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection. *arXiv preprint arXiv:2010.04159*, 2020.
- [68] Ziyue Zhu, Shenlong Wang, Jin Xie, Jiang-jiang Liu, Jingdong Wang, and Jian Yang. Voxel-splat: Dynamic gaussian splatting as an effective loss for occupancy and flow prediction. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 6761–6771, 2025.

## NeurIPS Paper Checklist

### 1. Claims

Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

Answer: [\[Yes\]](#)

Justification: The abstract and introduction clearly state our contributions and the claims made well match the experimental results.

Guidelines:

- The answer NA means that the abstract and introduction do not include the claims made in the paper.
- The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
- The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
- It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

### 2. Limitations

Question: Does the paper discuss the limitations of the work performed by the authors?

Answer: [\[Yes\]](#)

Justification: The limitations of our work are discussed in Sec. 5.

Guidelines:

- The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

### 3. Theory assumptions and proofs

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [\[NA\]](#)

Justification: Our paper does not include theoretical results.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

#### 4. Experimental result reproducibility

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [\[Yes\]](#)

Justification: We provide the detailed implementation in the Sec. 4.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general, releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  - (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  - (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  - (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  - (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

#### 5. Open access to data and code

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: The datasets used in our work are all publicly available, and we have properly cited these datasets. Source codes are available here

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (<https://nips.cc/public/guides/CodeSubmissionPolicy>) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so “No” is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (<https://nips.cc/public/guides/CodeSubmissionPolicy>) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

## 6. Experimental setting/details

Question: Does the paper specify all the training and test details (e.g., data splits, hyper-parameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [Yes]

Justification: We have provided the experimental details in Sec. 4.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

## 7. Experiment statistical significance

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [No]

Justification: The datasets used in our experiments are large, and it is very costly to train our model for many times. In summary, we don't report the error bar.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).

- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

#### 8. Experiments compute resources

Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [Yes]

Justification: We have reported the information of compute resources used in our experiments.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

#### 9. Code of ethics

Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics <https://neurips.cc/public/EthicsGuidelines>?

Answer: [Yes]

Justification: Our research conforms with the NeurIPS Code of Ethics.

Guidelines:

- The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

#### 10. Broader impacts

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [Yes]

Justification: We have discussed the broader impacts of our work in Sec. 5.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.



- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

## 11. Safeguards

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [NA]

Justification: Our work poses no such risks.

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

## 12. Licenses for existing assets

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

All assets used in our work are properly cited.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, [paperswithcode.com/datasets](https://paperswithcode.com/datasets) has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.

- If this information is not available online, the authors are encouraged to reach out to the asset’s creators.

### 13. **New assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [NA]

Justification: This paper does not release new assets.

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

### 14. **Crowdsourcing and research with human subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [NA]

Justification: Our paper does not involve crowdsourcing nor research with human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

### 15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [NA]

Justification: Our paper does not involve crowdsourcing nor research with human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

### 16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

Answer: [NA]

Justification: LLMs were not involved in any methodological component of the research.

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (<https://neurips.cc/Conferences/2025/LLM>) for what should or should not be described.

## A Technical Appendices and Supplementary Material

Table 4: Comparison of computational cost. All metrics were measured on the NVIDIA 4090 GPU.

Method	Flops (G)	Memory (GB)	Params (M)	FPS (img/s)	mIoU
BEVDetOcc [9]	541.21	4.71	34.97	1.1	21.98
FlashOcc [63]	439.71	2.76	58.67	7.5	23.40
OPUS [40]	215.54	2.00	73.17	22.4	20.28
COTR [31]	740.89	12.12	37.71	0.5	25.17
<b>Our LIAR (2f)</b>	690.40	5.05	64.39	4.7	27.33

Table 5: Quantitative comparison of SLLIE and other low-light image enhancement methods. Here,  $\gamma$  denotes the gamma correction factor, and  $c, T$  are the clip limit and tile grid size used in CLAHE.

Method	Description	mIoU
Gamma Correction [14]	$\gamma = 2.2$	13.77
Histogram Equalization [35]	equalize each channel	13.36
CLAHE [35]	$c = 2.0, T = (8, 8)$	13.84
<b>Our SLLIE</b>	selective enhancement	14.39

### A.1 Metrics

Following prior works [24, 31, 49, 63], we report the mean Intersection-over-Union (mIoU) as the evaluation metric for voxel-wise occupancy predictions. Notably, in the Occ3D-nuScenes dataset [38], the classes *bus*, *construction vehicle*, and *trailer* are absent from the nighttime validation split. Therefore, the mIoU is calculated over the remaining 14 semantic categories.

### A.2 Computational Cost

Tab. 4 presents the comparison of computational cost on the Occ3D-nuScenes dataset. LIAR achieves the highest mIoU, demonstrating superior performance. However, this improvement comes at the cost of increased computational complexity, primarily due to the Retinex decomposition network used to generate illumination maps. Therefore, future research could focus on more efficient Retinex decomposition techniques that can be seamlessly integrated into the occupancy framework.

### A.3 SLLIE Pretraining

The SLLIE module is pretrained on the nighttime split of the nuScenes training set in a self-supervised manner. Following SCI [30], we use the fidelity loss as follows:

$$\mathcal{L}_f = \sum_{t=1}^T \|x^t - (y + s^{t-1})\|^2, \quad (16)$$

where  $y$  is the input low-light image,  $x^t$  is the illumination estimate at stage  $t$  and  $s^{t-1}$  is the output of the self-calibrated module from the previous stage. When the input image  $y$  is already well-exposed (e.g., daytime image), the pseudo-target  $y + s^{t-1}$  becomes over-bright, thereby misleading the network to over-enhance the illumination  $x^t$ . Consequently,  $\mathcal{L}_f$  generates erroneous gradients that hinder the model’s ability to generalize to genuine low-light scenarios. Given that the full training set comprises a large proportion of daytime images (approximately 88%), jointly training the SLLIE module with the rest of the occupancy model results in optimization collapse. Therefore, we pretrain the SLLIE module on nighttime data and freeze its weights during training to ensure stability.

### A.4 Comparison of SLLIE and Other Methods

We compare the low-light enhancement performance of SLLIE with Gamma Correction [14], Histogram Equalization [35], and CLAHE [35]. All models are trained on the nighttime split of the Occ3D-nuScenes dataset [38], using cross-entropy loss exclusively. As listed in Tab. 5, SLLIE



Figure 9: Qualitative comparison of SLLIE and other low-light image enhancement methods.

Table 6: Comparison with low-light enhancement methods on the Occ3D-nuScenes.

Method	mIoU
BEVDetOcc [9]	15.86
BEVDetOcc + LCDPNet [39]	16.18 (+0.32)
BEVDetOcc + CSEC [23]	16.23 (+0.37)
FlashOcc [63]	18.15
FlashOcc + LCDPNet	18.29 (+0.14)
FlashOcc + CSEC	18.23 (+0.08)
COTR [31]	20.01
COTR + LCDPNet	20.22 (+0.21)
COTR + CSEC	20.17 (+0.16)
<b>Our LIAR</b>	<b>22.09</b>

Table 7: Performance on the Occ3D-nuScenes.

Method	mIoU		
	all	day	night
BEVDetOcc (2f) [9]	36.10	36.85	21.98
BEVFormer (4f) [25]	23.67	24.17	13.77
FlashOcc (2f) [63]	37.84	38.90	23.40
OPUS (8f) [40]	33.20	33.96	20.28
SparseOcc (8f) [27]	31.08	31.59	22.79
COTR (2f) [31]	38.70	39.47	25.17
<b>Our LIAR (2f)</b>	<b>39.57</b>	<b>40.42</b>	<b>27.33</b>

achieves the best performance. For instance, it surpasses the second best CLAHE [35] by 0.55 mIoU. Fig. 9 presents the qualitative comparisons in challenging lighting conditions. It is evident that other methods tend to amplify overexposed regions and introduce excessive noise in low-light regions, whereas SLLIE effectively avoids over-enhancement and improves overall visual quality.

Moreover, we conduct experiments where we apply state-of-the-art low-light enhancement methods as a preprocessing step before standard occupancy models. Specifically, we use LCDPNet [39] and CSEC [23] as low-light enhancement modules. As shown in Tab. 6, while applying enhancement methods leads to marginal improvements, our LIAR consistently outperforms existing methods. For example, our LIAR outperforms COTR + LCDPNet and COTR + CSEC by 1.87 and 1.92 mIoU, respectively. This highlights that the improvements are not solely attributable to low-light preprocessing, but also stem from our illumination-aware designs.

### A.5 Performance on Daytime Scenes

While our method is primarily designed for nighttime scenarios, we also evaluate our method on daytime scenes. Tab. 7 shows that our LIAR consistently outperforms existing methods on both the daytime and full datasets, surpassing the second-best COTR [31] by 0.87 and 0.95 mIoU, respectively. These results highlight the generalizability of our approach beyond nighttime conditions.

### A.6 Additional Visual Results

Fig. 10 illustrates results on real-world nighttime data from the Occ3D-nuScenes, while Fig. 11 shows visual comparisons on the same scene under all three severity levels on the nuScenes-C.

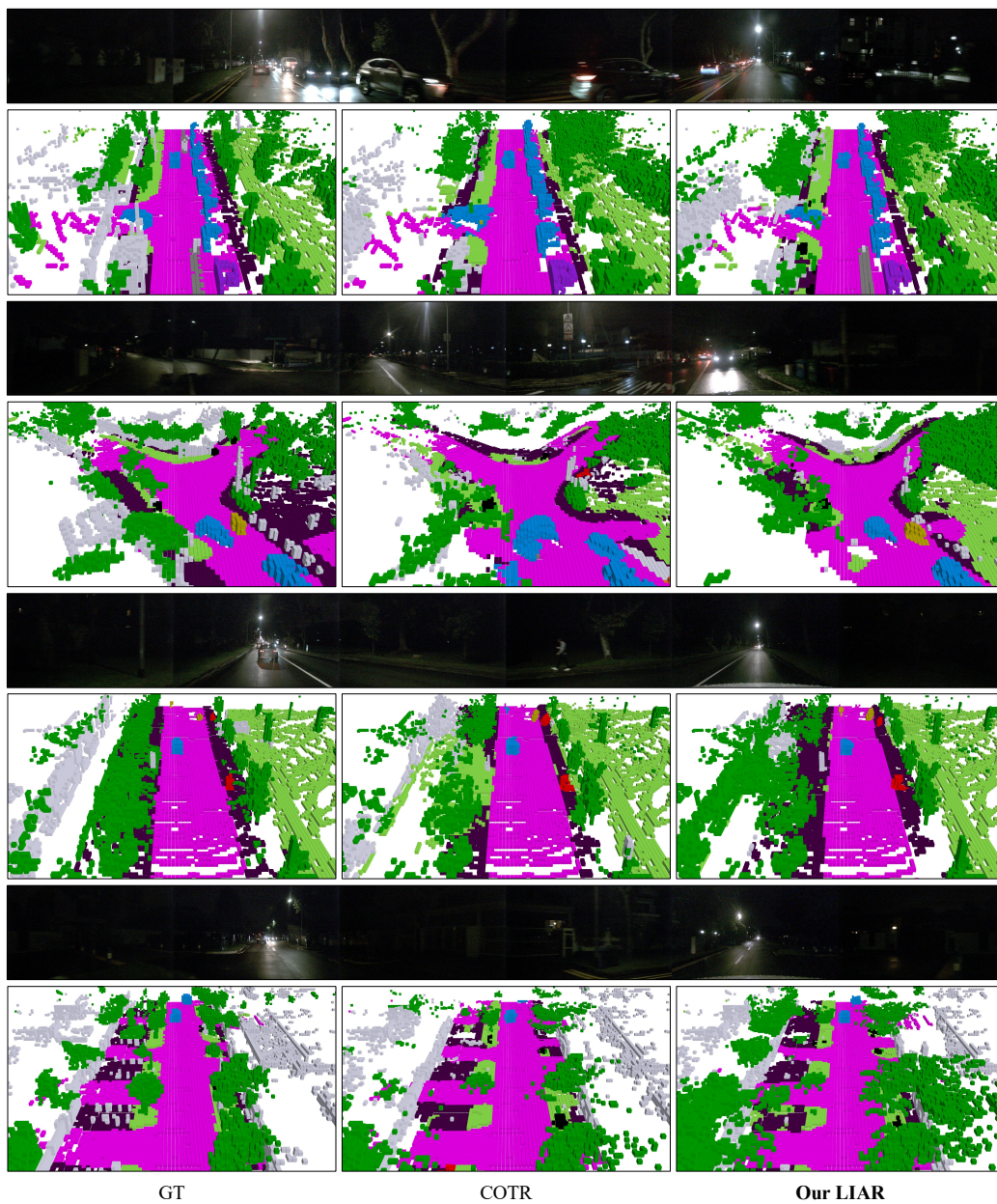


Figure 10: Qualitative comparison results on the Occ3D-nuScenes dataset.



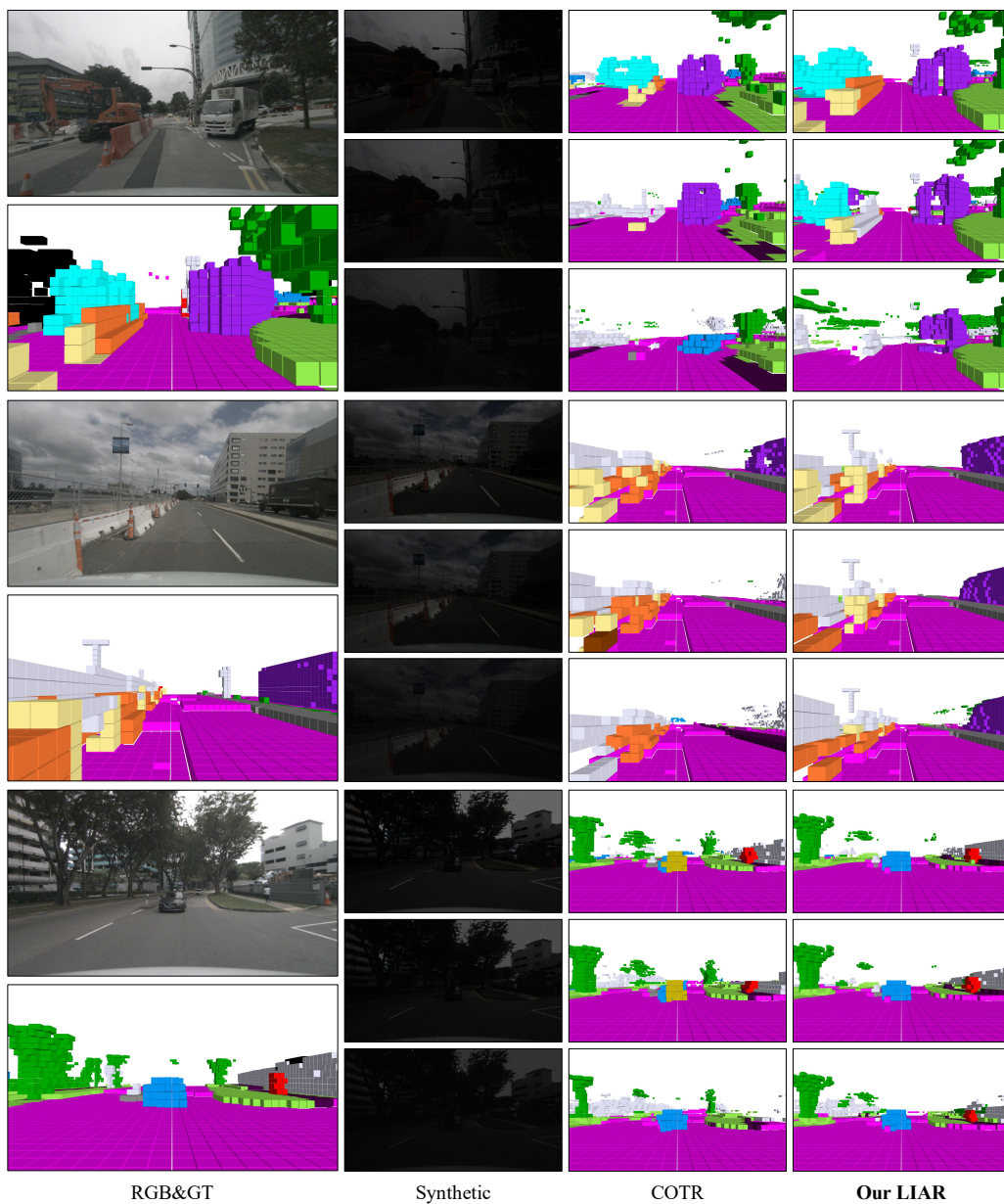


Figure 11: Qualitative comparison results on the nuScenes-C dataset across three severity levels.