Journal of Electronic Research and Application, 2026, Volume 10, Issue 4
http://ojs.bbwpublisher.com/index.php/JERA
Online ISSN: 2208-3510
Print ISSN: 2208-3502

# SpecBEV-IR: Illumination-Robust Front-End Enhancement for Multi-View BEV 3D Object Detection

**Yu Lin, Shijie Jia***

Dalian Jiaotong University, Dalian 116028, Liaoning, China

***Corresponding author:** Shijie Jia, jsj@djtu.edu.cn



**Abstract:** Multi-view visual BEV 3D object detection projects image information from different camera views into a unified bird's-eye-view space and has become an important paradigm for autonomous driving perception due to its low cost, flexible deployment, and rich semantic information. However, under complex lighting conditions such as nighttime, backlighting, local overexposure, and uneven illumination, multi-view input images often suffer from degraded brightness distribution, local contrast, and structural details, which further affects image feature extraction, view transformation, and unified spatial modeling. To address this issue, this paper proposes SpecBEV-IR, an illumination-robust multi-view BEV 3D object detection method. Built upon the SpecBEV framework, the proposed method introduces an illumination-robust image front-end enhancement module, termed ICF, between the multi-view input images and the shared 2D encoder. The ICF module consists of an invariant cue extraction unit (ICE) and a fusion convolution unit (Fuse Conv). ICE extracts more stable illumination-invariant cues from raw images, while Fuse Conv integrates these cues with the original image content to generate enhanced input representations for subsequent feature encoding and view transformation. Different from conventional enhancement methods that mainly improve visual appearance, SpecBEV-IR emphasizes structural stability and cross-view consistency for downstream 3D detection. Experiments on the nuScenes dataset show that SpecBEV-IR achieves 0.4121 mAP and 0.5174 NDS on the validation set, while also obtaining better or more balanced performance on multiple error metrics, including mATE, mASE, mAOE, and mAAE. The results demonstrate that the proposed method effectively improves the overall robustness and detection performance of multi-view visual 3D object detection under complex lighting conditions.

**Keywords:** Multi-view BEV detection; Illumination-robust enhancement; Invariant cue extraction; 3D object detection

**Online publication:** May 21, 2026

## 1. Introduction

Multi-view camera-based bird's-eye-view (BEV) 3D object detection has become an important research

direction in autonomous driving perception [1,2]. By projecting image features from different camera views into a unified top-down spatial representation, BEV-based methods provide a structured description of surrounding traffic scenes and offer a convenient interface for downstream tasks such as object detection, scene understanding, motion planning, and decision making [3]. Compared with LiDAR-based solutions, camera-based perception is more attractive for large-scale deployment because of its lower hardware cost, richer semantic information, and greater flexibility in sensor configuration [4-6].

Despite the rapid progress of multi-view visual BEV perception, the performance of camera-based 3D detection remains highly sensitive to changes in imaging conditions [7]. In real driving environments, vehicles frequently encounter nighttime scenes, backlighting, local overexposure, uneven illumination, tunnels, and other challenging situations. Under such conditions, the input images may suffer from degraded brightness distribution, weakened local contrast, blurred structural details, and increased noise [8]. These degradations not only affect feature extraction within individual views, but may also propagate through subsequent view transformation and cross-view fusion, eventually reducing the quality of the unified BEV representation [7]. As a result, the detector may produce missed detections, false positives, and unstable localization.

Existing studies on illumination-robust visual perception mainly follow two technical routes. The first route enhances degraded images before they are fed into downstream models, as represented by methods such as Zero-DCE [9]. Such methods often improve image brightness and perceptual quality, but their optimization objectives are usually more aligned with image restoration itself than with the requirements of high-level perception tasks. In multi-view BEV detection, independently enhanced images may even introduce larger distribution discrepancies across different camera views, which can weaken the stability of shared feature encoding and subsequent BEV fusion. The second route attempts to learn task-oriented robust representations that are less sensitive to illumination changes [10,11]. Although this direction is more compatible with downstream perception, most existing studies are still developed for single-image tasks such as detection, segmentation, and recognition, and they do not sufficiently consider the coupling among input stability, cross-view consistency, and unified spatial modeling in multi-view BEV detection.

For multi-view visual BEV 3D object detection, illumination handling should therefore not be regarded as a simple pre-processing step. What the task requires is not merely visually enhanced input images, but more stable and more suitable input representations for shared encoding, view transformation, and BEV-space detection. In other words, a front-end module for this task should preserve scene semantics and structural cues under illumination variation, while avoiding unnecessary view-wise distribution shifts that may interfere with subsequent multi-view fusion [10]. This requirement makes the problem substantially different from conventional single-image enhancement [11].

Motivated by this observation, this paper proposes SpecBEV-IR, an illumination-robust front-end enhancement method for multi-view visual BEV 3D object detection under complex lighting conditions. Built upon the SpecBEV framework, the proposed method introduces an illumination-robust image front-end enhancement module, termed ICF, between the raw multi-view input images and the shared 2D encoder. The proposed ICF module consists of two components. First, an invariant cue extraction unit (ICE) is used to extract more stable illumination-invariant cues from the raw image. Second, a fusion convolution unit (Fuse Conv) integrates these cues with the original image content to generate an enhanced representation that is more suitable for subsequent feature encoding and unified spatial modeling. In addition, to better fit the multi-camera setting, a shared-across-views strategy is adopted so that all views follow a consistent enhancement mechanism while still preserving their own content-dependent responses.

Different from conventional enhancement methods that mainly aim to improve visual appearance, the proposed method is designed from the perspective of downstream BEV 3D detection. Its objective is to improve input robustness without disrupting the original scene content or the consistency required by multi-view fusion. By introducing task-oriented illumination-robust modeling at the image-domain front end, the proposed method establishes a more stable input foundation for subsequent shared feature extraction, view transformation, and BEV-space detection.

The main contributions of this paper are summarized as follows.

(1) An illumination-robust front-end enhancement framework, termed SpecBEV-IR, is proposed for multi-view visual BEV 3D object detection. The proposed method improves input stability under complex lighting conditions while preserving the original downstream detection pipeline;

(2) An ICF module is designed to construct task-oriented robust input representations. Specifically, ICE extracts illumination-invariant cues from raw images, and Fuse Conv integrates these cues with the original image content, so that the enhanced representation is more suitable for downstream shared encoding and BEV-space detection;

(3) A shared-across-views enhancement strategy is introduced for the multi-camera setting. This design helps reduce view-wise distribution discrepancy caused by inconsistent illumination across different camera views and improves the stability of subsequent cross-view fusion;

(4) Experiments on the nuScenes dataset demonstrate that the proposed method outperforms several representative front-end enhancement methods and achieves better overall detection performance under complex lighting conditions.

## 2. Literature review

### 2.1. Multi-view camera-based BEV 3D object detection

Camera-based 3D object detection is an important research direction in autonomous driving perception. Its main objective is to recover the spatial attributes of traffic participants, including location, size, and orientation, from image observations. Compared with LiDAR-based methods, visual approaches offer lower hardware cost, richer semantic information, and greater flexibility in large-scale deployment. These advantages make camera-based perception highly attractive for practical onboard systems. At the same time, image-based 3D perception remains inherently challenging, because a camera captures the projection of a three-dimensional scene onto a two-dimensional image plane, where explicit depth information is not directly available [12]. As a result, how to construct reliable spatial representations from image features has long been a central problem in visual 3D detection.

Early studies mainly focused on monocular or stereo 3D object detection. Monocular methods usually infer 3D attributes from a single RGB image by combining object appearance, geometric priors, and depth estimation. Although such methods are simple in sensor configuration, they often suffer from depth ambiguity and unstable spatial recovery, especially for distant or occluded targets. Stereo methods introduce disparity cues from paired images and provide stronger geometric constraints than monocular approaches. They can improve depth estimation to some extent, but their effectiveness often depends on fixed view overlap and carefully designed camera settings [1]. With the widespread adoption of surround-view camera systems in autonomous driving, research attention gradually shifted from local 3D recovery in monocular or stereo settings to unified spatial modeling under multi-view inputs [12].

Under this trend, bird's-eye-view representation has become a key paradigm for multi-view visual 3D perception. By projecting features from different camera views into a unified top-down space, BEV representation provides a more natural description of object distribution, relative position relationships, and road structure around the ego vehicle [3]. Compared with perspective-view features, BEV features are more consistent with physical scene geometry and are easier to integrate with downstream tasks such as detection, mapping, and planning. For this reason, BEV representation has become one of the most important intermediate forms in modern autonomous driving perception (**Figure 1**) [1,2].

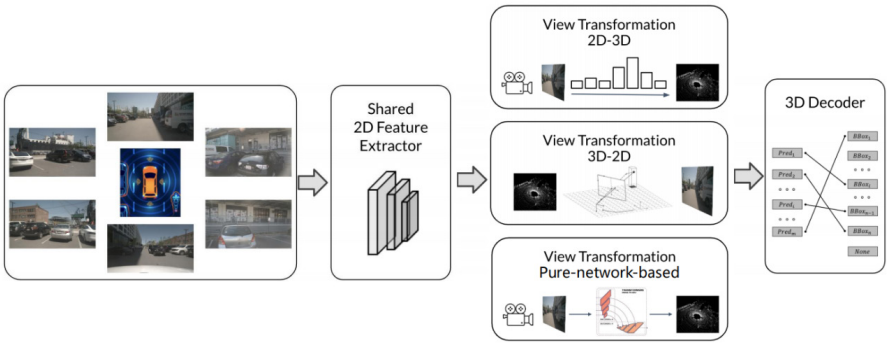


**Figure 1.** General pipeline of camera-based BEV 3D perception.

Existing multi-view BEV detection methods can generally be divided into two technical routes: explicit geometric transformation and implicit spatial modeling. Explicit geometric methods usually estimate depth distributions for image features and then lift the 2D features into 3D space according to camera parameters, followed by spatial aggregation in the BEV domain, as represented by methods such as Lift-Splat-Shoot, BEVDet, and BEVDepth [3–5]. This type of method has strong geometric interpretability and clearly reflects the transformation process from image space to unified spatial representation. It also forms the basis of many high-performance multi-camera BEV detection frameworks. In contrast, implicit spatial modeling methods do not rely entirely on explicit depth lifting. Instead, they directly learn the mapping from multi-view image features to BEV representations through learnable spatial queries, positional encoding, attention mechanisms, or transformer-based structures, as exemplified by BEVFormer, PETR, and DETR3D [6,14,15]. These methods are often more flexible in capturing global context and long-range dependencies, and they have further expanded the modeling capacity of visual BEV perception.

Although multi-view visual BEV 3D object detection has achieved significant progress, several difficulties remain unresolved as follows:

(1) Visual 3D perception is still sensitive to errors in depth recovery or spatial correspondence. Once the geometric relationship between image features and physical space is inaccurately modeled, the error may propagate to later stages and directly affect object localization and box regression;

(2) Feature fusion across different camera views is inherently challenging. Since the cameras differ in position, field of view, imaging angle, and visible content, the same object may appear with obvious variations across views [16]. After projection into the unified BEV space, these variations may still lead to semantic inconsistency or local geometric misalignment;

(3) Camera-based methods are strongly influenced by input image quality. When the scene involves

nighttime, backlighting, local overexposure, or uneven illumination, the degradation introduced at the image level may further affect feature extraction, view transformation, and BEV representation construction.

For this paper, high-quality multi-view BEV detection requires not only strong geometric consistency in the unified space, but also stable visual inputs before the shared image encoder and view transformation stages. Existing research has made substantial progress in depth modeling, BEV representation generation, and cross-view feature fusion. However, the effect of illumination-induced input degradation on the overall BEV detection pipeline has not been sufficiently addressed. This limitation motivates further investigation into illumination-robust enhancement for multi-view visual BEV 3D object detection.

## 2.2. Illumination-robust visual perception under complex lighting

Illumination-robust visual perception under complex lighting conditions mainly focuses on the performance degradation of visual models in scenarios such as nighttime, low-light environments, backlighting, local overexposure, uneven illumination, and rapid brightness transitions. Compared with normal imaging conditions, these scenarios often cause significant changes in brightness distribution, local contrast, texture details, and noise levels, which in turn affect the stability of subsequent feature extraction, object recognition, and spatial localization. For autonomous driving vision systems, the influence of complex lighting is not limited to images becoming simply darker or brighter. More importantly, the change in input distribution may disrupt the discriminative feature patterns learned during training and lead to noticeable performance fluctuations in high-level visual tasks.

Existing studies on this problem can generally be divided into two categories. The first category approaches the problem from the perspective of image enhancement. Representative methods in this line include SCI, NeRCo, and IAT [17–19]. These methods first restore degraded images and then feed the enhanced outputs into downstream models for detection, segmentation, or recognition. They usually aim to improve brightness, contrast, color fidelity, and fine details, expecting that visually clearer inputs will indirectly benefit downstream tasks. This line of research has developed into a relatively rich technical branch and often produces a direct improvement in image appearance. However, for high-level visual tasks, enhanced images are not always guaranteed to be beneficial. The main reason is that many enhancement methods are optimized for image restoration or perceptual quality rather than for preserving and strengthening task-relevant discriminative structures [20,21]. If the enhancement process introduces artificial textures, oversharpened edges, color shifts, or unbalanced brightness mapping, it may instead weaken the ability of downstream models to build stable representations of real target structures [22].

The second category places more emphasis on task-oriented robust representation learning. Instead of focusing only on improving image appearance, these methods attempt to learn features that are less sensitive to illumination variations and more suitable for high-level visual tasks [23]. Compared with conventional image enhancement, this line of research pays greater attention to whether the processed representation truly benefits downstream detection, classification, or segmentation. The focus gradually shifts from pixel-level restoration to feature stability and task consistency. This idea is closer to practical requirements in complex lighting scenarios, because real-world visual systems do not only face a single type of degraded input such as low-light images. They must remain stable under frequent transitions between normal lighting and challenging illumination conditions. For this reason, isolated processing designed only for degraded images is often insufficient to support robust perception across diverse real scenarios.

Although these studies have made meaningful progress, several limitations remain as follows:

(1) Many existing methods mainly deal with single-image inputs and concentrate either on image restoration itself or on conventional 2D tasks such as detection, classification, and segmentation. They rarely consider the consistency problem across different views in multi-camera systems;

(2) Illumination degradation not only affects local textures and edges, but also changes the statistical relationships among multi-view inputs. If different camera views are processed independently and produce larger feature distribution shifts, the stability of shared feature extraction may be further weakened, and additional disturbances may be introduced into later cross-view fusion;

(3) Many methods implicitly assume that illumination degradation has a relatively fixed pattern, whereas lighting changes in autonomous driving environments are much more dynamic. Variations in time, direction, scene structure, and environmental conditions can all make image degradation highly unpredictable, which increases the difficulty of robust modeling [8].

For multi-view visual BEV 3D object detection, the challenge posed by complex lighting conditions is more complicated than in single-image scenarios. This type of task requires not only sufficient discriminative information in each input image, but also stable and consistent feature representations across different camera views during shared encoding, view transformation, and unified spatial modeling. If complex illumination introduces large discrepancies among camera inputs, these discrepancies may be further amplified during BEV fusion and eventually affect the final detection results. Therefore, illumination handling for multi-view visual BEV detection should not remain at the level of single-view image enhancement. It needs to be considered together with the requirements of downstream detection, with joint attention to input stability, cross-view consistency, and unified spatial representation quality.

Motivated by this understanding, this paper focuses on the influence of illumination-induced multi-view input degradation on the subsequent BEV detection pipeline and studies the problem from the perspective of front-end robust enhancement in the image domain. The goal is not merely to improve visual appearance, but to provide more stable input representations for subsequent shared feature extraction, view transformation, and unified spatial modeling while preserving original scene content and view-specific information as much as possible. In this way, the overall performance of multi-view visual 3D object detection under complex lighting conditions can be improved.

## 2.3. Task-oriented robust representation learning

Traditional methods for handling complex illumination conditions mainly focus on restoring degraded input images. Their goal is to improve visual quality by enhancing brightness, increasing contrast, and compensating for missing details. Although such methods often produce obvious improvements at the image level, better perceptual appearance does not necessarily lead to better performance in high-level vision tasks [17,21,22]. The reason is that tasks such as object detection, segmentation, and scene understanding do not simply rely on images that look clearer or brighter. What they truly require are discriminative features that can stably reflect target structure, boundary information, and semantic differences. When image restoration places too much emphasis on visual quality itself, the structural cues needed by downstream models may instead be altered, weakened, or even obscured.

To address this issue, more recent studies have started to focus on task-oriented robust representation learning. These methods no longer treat illumination handling as an isolated image restoration problem.

Instead, they consider how to maintain stable feature representations under illumination changes from the perspective of downstream task requirements. Compared with conventional enhancement methods that mainly operate in the pixel domain, this line of research places greater emphasis on feature distribution consistency, structural information preservation, and continuity of task discriminability [23]. The objective is not merely to improve image quality, but to ensure that the model can still extract representations that remain truly effective for downstream tasks under different lighting conditions.

From the perspective of methodology, task-oriented illumination-robust representation learning usually follows two major directions. One direction attempts to directly model illumination-invariant representations in feature space. By constraining the relationship between samples captured under normal and challenging lighting conditions, the network is encouraged to learn features that are insensitive to illumination variation while remaining sensitive to semantic and structural information of targets. The other direction introduces a bridging mechanism between the input stage and the task network. In this case, the enhancement process is no longer regarded as an independent output, but is instead designed to directly serve the subsequent detection or recognition process. Under this idea, input enhancement, feature extraction, and high-level task modeling are no longer treated as separate stages, but are jointly designed around whether the processed representation is truly beneficial to downstream perception.

This research trend shifts the goal of illumination handling from image restoration to task consistency preservation. Such a shift is especially meaningful for real-world vision systems [11]. In practical applications, a model often needs to handle normal lighting, low-light scenes, backlighting, local overexposure, and rapid illumination transitions within the same operating environment. A method optimized only for one specific type of degradation may work well in a particular condition, but often struggles to maintain stable performance across diverse scenarios. Task-oriented robust representation learning pays more attention to unified modeling across different lighting conditions, aiming to preserve relatively stable feature extraction behavior when the input environment changes, so that fluctuations in downstream task performance can be reduced.

For multi-view visual BEV 3D object detection, this problem becomes even more complicated. In single-image tasks, illumination changes mainly affect local texture, edge quality, and contrast within the current image. Under multi-view settings, however, complex illumination also introduces statistical inconsistency across camera views, making the inputs from different cameras noticeably different even before they are fed into the shared encoder. If this discrepancy is not effectively reduced, the resulting error may further accumulate during later view transformation and unified spatial fusion. In this case, the challenge caused by complex lighting is no longer limited to degradation in an individual image, but becomes a joint problem involving unstable inputs, weakened cross-view consistency, and degraded unified spatial representation.

# 3. Method

## 3.1. Overall framework of SpecBEV-IR

To improve the input stability of multi-view visual BEV 3D object detection under complex lighting conditions, this paper proposes SpecBEV-IR based on the aforementioned multi-view BEV detection framework. Instead of changing the overall detection pipeline, the proposed method introduces an illumination-robust image front-end enhancement module, termed ICF, before the multi-view input images are fed into the shared 2D backbone. Through this design, the model can alleviate the adverse influence of

nighttime, backlighting, local overexposure, and uneven illumination on subsequent feature extraction and view transformation while preserving the original unified-space detection framework. In this way, a more stable input foundation is provided for 3D object detection in the BEV space (**Figure 2**).

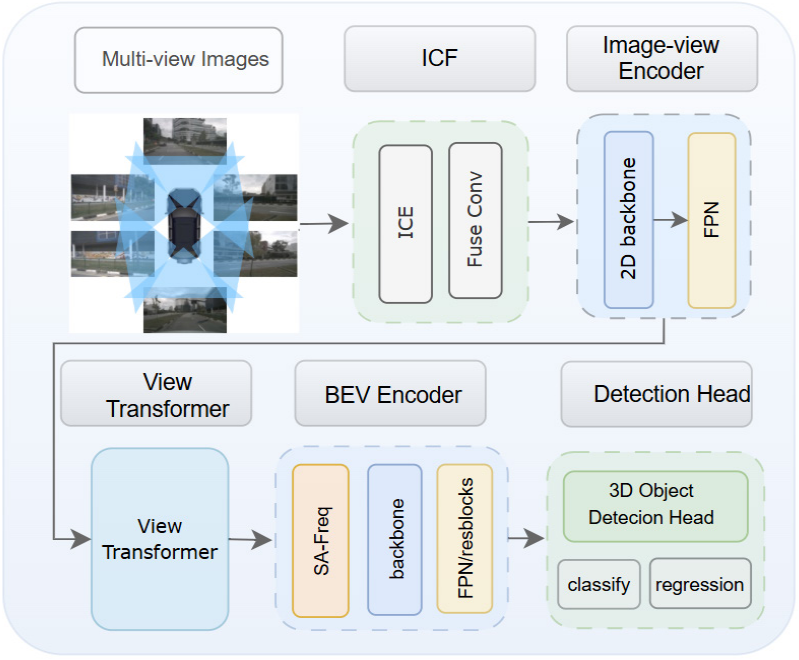


**Figure 2.** The illumination-robust BEV detection framework SpecBEV-IR.

From the perspective of the overall pipeline, the inputs of SpecBEV-IR are six surround-view camera images. Different from conventional multi-view visual detection methods that directly feed raw images into the shared image encoder, the proposed method first applies the ICF module to each view and generates enhanced inputs that are more suitable for subsequent visual representation learning. The enhanced multi-view images are then passed through the shared 2D backbone for feature extraction, followed by view transformation into a unified BEV space. On this basis, the following BEV-domain representation optimization and detection process are performed to accomplish 3D object detection. In other words, the main modification of SpecBEV-IR is introduced at the input stage. Its purpose is not simply to improve visual appearance, but to enhance the stability of shared encoding, view transformation, and unified spatial modeling through front-end robust enhancement.

Within this framework, image-domain front-end enhancement and unified-space representation optimization operate at different levels. The front-end ICF module is mainly responsible for alleviating appearance degradation in the input images under complex lighting conditions, so that the images entering the shared encoder become more stable in terms of brightness distribution, local structure, and discriminative visibility. In contrast, the subsequent representation optimization module in the unified BEV space mainly addresses redundant responses and geometric inconsistency after multi-view fusion. The former focuses on input quality and imaging stability across views, whereas the latter focuses on discriminative representation and geometric consistency in the unified space. These two parts are complementary in both processing level and functional objective. Through this hierarchical design that combines image-domain front-end

enhancement with unified-space representation optimization, the proposed framework can simultaneously improve input robustness under challenging illumination and the quality of BEV-space detection representations.

The ICF module consists of two main components, namely an illumination-invariant cue extraction branch and a feature fusion branch. The former is used to extract stable cues that are relatively insensitive to illumination variation from the raw image, while the latter jointly models these cues together with the original image content to generate an enhanced representation that preserves both scene information and illumination robustness. Since this module is placed before the shared 2D backbone, its output directly affects the quality of subsequent multi-view feature extraction and further influences the later view transformation and feature fusion process in the unified space. Different from traditional low-light enhancement methods that treat the enhanced image as an independent final output, the ICF module in this paper is designed to provide task-adaptive input representations for downstream detection, and its objective is consistently centered on improving BEV 3D detection performance.

For the multi-view input setting, this paper does not simply concatenate all camera views for joint processing. Instead, the ICF module is applied to each of the six views individually, while the same set of module parameters is shared across all views. The reason is that although different cameras may encounter different illumination conditions and scene contents, the front-end enhancement process should still follow a unified robust adjustment rule. If each view is processed by an independent enhancement network, the distribution gap among views may be further enlarged, which would negatively affect the stability of the shared encoder and the subsequent cross-view fusion process. In contrast, the parameter-sharing strategy allows each view to be enhanced according to its own image content while maintaining consistency in the enhancement mechanism, which is more suitable for the characteristics of multi-view visual BEV detection.

Based on the above design, SpecBEV-IR should not be regarded as a simple pre-processing step added before the original multi-view detection framework. Instead, it introduces a task-oriented input robustness modeling stage into the overall detection pipeline. This stage preserves useful scene information in the original multi-view images while supplementing and adjusting the inputs through illumination-invariant cues, thereby establishing a more stable input foundation for subsequent shared feature extraction, view transformation, and unified-space 3D object detection. The following subsections will further describe the detailed design and implementation of ICF, including the extraction of illumination-invariant cues, the fusion strategy, and the shared processing mechanism for multi-view inputs.

## 3.2. ICF: Illumination-robust front-end enhancement module

To reduce the instability of multi-view input images under complex lighting conditions, an illumination-robust image front-end enhancement module, termed ICF, is introduced between the raw multi-view images and the shared 2D backbone. The purpose of this module is not to improve perceptual image quality for visual inspection, but to provide more stable and task-adaptive inputs for subsequent feature extraction and view transformation.

As shown in **Figure 3**, the ICF module consists of two components: an illumination-invariant cue extraction branch and a feature fusion branch. The former extracts relatively stable structural information from the raw image under illumination variation, while the latter combines the extracted invariant cues with the original image content to generate an enhanced representation for the downstream detection pipeline. Under this design, the input image is not replaced by an independently restored result. Instead, the original

content is adjusted with additional stable cues that are useful for later representation learning.

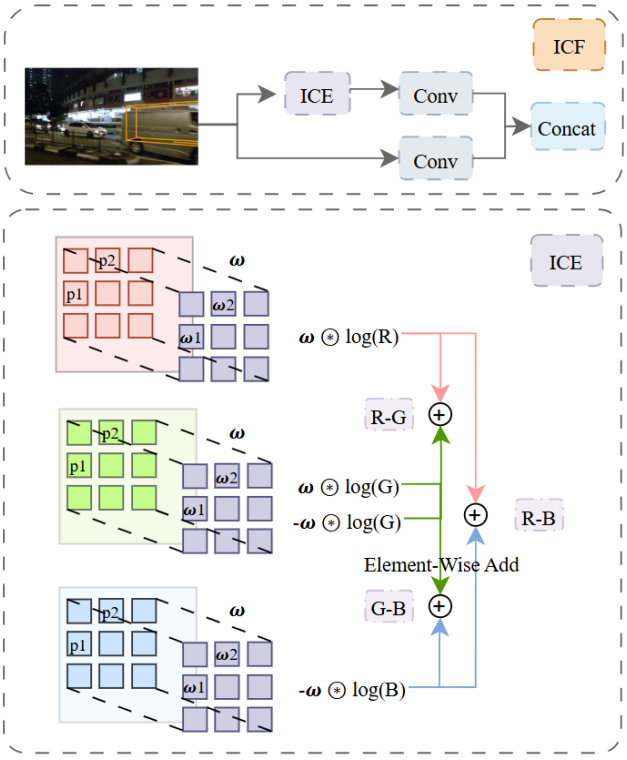


**Figure 3.** ICF module.

For the $i$-th camera view, let the raw input image be denoted as $I_i$. The ICF module first extracts an illumination-invariant representation $F_i^{inv}$ from $I_i$, and then fuses $F_i^{inv}$ with $I_i$ to produce the enhanced output $I_i^{enh}$. The process can be written as:

$$F_i^{inv} = ICE\left(I_i\right) \tag{1}$$

$$I_i^{enh} = Fuse\left(I_i, F_i^{inv}\right) \tag{2}$$

where $ICE(\cdot)$ denotes the invariant cue extraction operation and $Fuse(\cdot)$ denotes the fusion process. The enhanced image $I_i^{enh}$ is directly fed into the shared 2D backbone rather than being treated as an independent restoration output.

This two-stage design is motivated by the fact that degradation under complex lighting is not limited to reduced brightness. It may also involve local overexposure, uneven illumination, weakened edges, blurred structure, and amplified noise. Simple global adjustment of brightness or contrast is often insufficient for downstream detection, because it does not explicitly preserve the structural information required by the shared encoder and later BEV transformation. By first extracting illumination-invariant cues and then integrating them with the original image, the module can better preserve target contours, structural boundaries, and locally discriminative regions.

The ICF module is designed for downstream 3D detection rather than standalone image enhancement. Conventional low-light enhancement methods usually treat the enhanced image itself as the final output and

therefore focus on brightness recovery, color naturalness, or perceptual quality. In contrast, the output of ICF is only an intermediate representation for the following detection network. Its effectiveness is measured by whether it benefits shared feature extraction, view transformation, and BEV-space object detection.

For the multi-view setting, the ICF module is applied to each camera view individually, while all views share the same module parameters. This design keeps the enhancement process consistent across views and avoids introducing additional distribution gaps before shared encoding. At the same time, each view is still processed according to its own image content, which allows the module to adapt to local illumination differences without breaking the consistency required by multi-view fusion.

## 3.3. ICE: Invariant cue extraction

This paper introduces an illumination-robust image enhancement module, termed ICF (Invariant Cue Fusion), in the image-domain front-end to improve the structural stability and semantic discriminability of multi-view inputs before they are fed into the shared 2D encoder. The design of ICF follows a simple principle: illumination-invariant cues are first extracted and then fused with the original image information. The module consists of two parts, namely the illumination-invariant cue extraction unit, ICE (Invariant Cue Extraction), and the fusion convolution unit, Fuse Conv. Among them, ICE is responsible for extracting intermediate responses from raw RGB images that are less sensitive to illumination variation, while Fuse Conv jointly encodes the invariant-cue branch and the original-image branch to generate the final enhanced image representation. Different from conventional image enhancement methods that aim only at improving visual appearance, ICF does not pursue mere brightness enhancement. Instead, it places greater emphasis on the structural stability and discriminative representation required by downstream 3D detection, so that the enhanced result is more suitable as the input to the image feature encoder.

The core objective of ICE is to construct feature responses that are insensitive to illumination variation from the raw input image. The basic motivation is as follows: for RGB images, absolute intensity values may vary significantly with ambient illumination, shadow occlusion, and exposure conditions, whereas the relative relationships across different color channels and within local neighborhoods are usually more stable. Existing studies have shown that, under the Lambertian imaging assumption, pixel values can be decomposed into a position-related term, an illumination spectrum term, and an object reflectance term. If the first two factors can be eliminated through appropriate channel differencing and neighborhood differencing, an illumination-invariant representation closer to the intrinsic properties of the object can be obtained.

Based on this idea, this subsection first maps the input image into the logarithmic color domain and constructs three groups of cross-channel responses, corresponding to the local difference relationships between the red-green, red-blue, and green-blue channel pairs, respectively. Let the input image be denoted as $I$, and let its three color channels be denoted by $R$, $G$, and $B$. For an arbitrary learnable convolution kernel $W$, the following three groups of illumination-invariant responses are defined:

$$S^{RG} = W \circledast log(R) + (-W) \circledast log(G) \tag{3}$$

$$S^{RB} = W \circledast log(R) + (-W) \circledast log(B) \tag{4}$$

$$S^{GB} = W \circledast log(G) + (-W) \circledast log(B) \tag{5}$$

where $\circledast$ denotes the convolution operation.

The above formulation essentially exploits the cross-channel differencing form in the logarithmic color domain: the positive and negative responses of the same kernel on different color channels can suppress fluctuations related to the overall illumination intensity while highlighting more stable local structures and relative color relationships. Compared with directly using fixed difference templates, this convolutional formulation is more adaptive, since it can learn response patterns that better fit the downstream task through end-to-end training while satisfying the basic illumination-invariant constraints.

To further reduce the influence of residual illumination terms on the convolution responses, a zero-mean constraint is imposed on the convolution kernel. Let the kernel size be $k \times k$ . The constraint can be written as:

$$\bar{W} = \frac{1}{k^2} \sum_{i=1}^{k^2} w_i = 0 \tag{6}$$

where $w_i$ denotes the $i$ -th parameter in the convolution kernel.

The role of the zero-mean constraint is to make the kernel more inclined to represent local differences rather than overall brightness averages, thereby further weakening the direct influence of illumination variation on the responses. Existing related studies have shown that this constraint provides clear stabilization benefits in illumination-invariant feature learning.

In implementation, ICE organizes the three groups of cross-channel responses into a stacked invariant-cue representation, i.e.,

$$S = Concat\left(S^{RG}, S^{RB}, S^{GB}\right) \tag{7}$$

where $Concat(\cdot)$ denotes concatenation along the channel dimension.

The output $S$ of ICE is not the final enhanced image, but an intermediate invariant cue representation, which is essentially closer to an illumination-robust prompt for subsequent fusion. In this way, ICE can extract stable responses that are more valuable for object detection from the input without directly reconstructing the original image, thereby providing structural priors for the subsequent fusion process.

## 3.4. Fuse Conv: Fusion of invariant cues and original images

Although ICE can extract intermediate responses with strong illumination robustness from the input image, these responses still remain in the form of intermediate features and therefore cannot directly replace the original image as the input to the 2D backbone network. On the one hand, the invariant cues alone mainly reflect relative channel relationships and local structural variations, but lack complete texture and semantic appearance information. On the other hand, if they are directly fed into the 2D encoder, the resulting input distribution may become inconsistent with the pretrained feature distribution of the original visual backbone, thereby affecting subsequent feature extraction. Therefore, it is necessary to further jointly fuse the invariant cues produced by ICE with the original image information, so as to generate an enhanced image representation that both preserves the original scene content and exhibits stronger illumination robustness.

To this end, a Fuse Conv module is designed after ICE to perform dual-branch encoding and fusion on the invariant-cue branch and the original-image branch. Specifically, the input image $I$ is first sent to ICE to obtain the invariant-cue representation $S$ . Meanwhile, the original-image branch is retained, and a separate convolutional layer is used to extract shallow image representations. Subsequently, the two branches are respectively mapped by convolution, concatenated along the channel dimension, and further fused through convolution to form the final output. The process can be written as:

$$F_{ice} = \phi_1(S), F_{img} = \phi_2(I) \tag{8}$$

$$\hat{I} = \phi_f\left(Concat\left(F_{ice}, F_{img}\right)\right) \tag{9}$$

where $\phi_1(\cdot)$, $\phi_2(\cdot)$, and $\phi_f(\cdot)$ denote the convolutional mapping and fusion operations, respectively, and $\hat{I}$ denotes the output image enhanced by ICF.

Fuse Conv preserves the basic appearance information in the original-image branch, thereby avoiding the semantic deficiency caused by relying only on invariant responses. At the same time, it explicitly injects the more stable structural responses from the invariant-cue branch into the final input representation, thereby improving the ability of the subsequent 2D backbone to extract effective features under complex lighting conditions. This module is not a simple concatenation operation; rather, it serves the function of transforming invariant cues into an enhanced image representation that can be directly utilized by the subsequent visual encoder. According to **Figure 3**, the fusion strategy adopted by this module belongs to a lightweight front-end design. It does not modify the original multi-view BEV backbone structure, but only introduces a small amount of additional convolutional mapping and fusion overhead at the input stage, thereby providing good integration flexibility.

## 3.5. Shared enhancement strategy for multi-view inputs

This paper considers a six-camera input setting. Therefore, the processing strategy of ICF needs to account not only for illumination-robust enhancement of individual images, but also for consistency and scalability under multi-view inputs. To this end, a view-wise independent enhancement strategy with cross-view parameter sharing is adopted. Let the six camera-view inputs be denoted as $\{I_i\}_{i=1}^{6}$. Then, the same ICF module is applied to each view:

$$\hat{I}_i = ICF(I_i), i = 1, 2, \ldots, 6. \tag{10}$$

All views share the same set of ICE and Fuse Conv parameters, that is, the convolution kernels and fusion convolutions are completely shared across different camera views.

The motivation for parameter sharing mainly lies in two aspects as follows:

(1) Although different views vary in scene content and illumination conditions, the underlying mechanism by which illumination variation affects image formation is essentially consistent across views. Sharing parameters therefore helps the model learn a unified enhancement rule, while avoiding parameter redundancy and training instability that may arise if each view is learned independently;

(2) In multi-view BEV perception frameworks, the 2D backbone typically adopts shared weights to process images from different views. Applying shared parameters at the ICF stage as well maintains a consistent design logic between the input enhancement stage and the image encoding stage, which is beneficial for joint optimization of the overall network.

Parameter sharing does not mean that the outputs of all views are identical. Since each camera captures different image content, the invariant cues extracted by ICE and the fusion results generated by Fuse Conv still remain view-specific. As a result, the final output $\hat{I}_i$ is adaptively generated according to the content of each input image. What is shared is the enhancement mechanism rather than the enhancement result itself. Through this combination of a unified rule and view-wise responses, the model can preserve consistency in

enhancement strategy across all six views while retaining the actual differences among different camera inputs. This property is more suitable for multi-view BEV detection under complex and spatially uneven illumination environments.

The ICF module completes the full process from illumination-invariant cue extraction to enhanced image generation in the image-domain front end. Specifically, ICE is responsible for constructing cross-channel responses that are insensitive to illumination variation, Fuse Conv jointly integrates the invariant cues with the original image information, and the shared multi-view processing mechanism ensures that the module can be naturally embedded into a multi-camera BEV detection framework, thereby providing more stable inputs for subsequent 2D feature extraction, view transformation, and BEV representation optimization.

# 4. Experiment

## 4.1. Experimental environment and parameters

All experiments in this section are built upon the SpecBEV framework, and the overall model equipped with the illumination-robust image front-end enhancement module is denoted as SpecBEV-IR. Except for the image-domain front-end enhancement strategy, the image feature encoder, view transformation module, BEV encoder, detection head, and training strategy are kept the same as those of the baseline to ensure a fair comparison.

All experiments were conducted on a single NVIDIA GeForce RTX 4090D GPU with 24 GB of memory. The software environment was based on CUDA 12.4 and PyTorch 1.13. To ensure the stability and reproducibility of the experimental process, both training and evaluation were performed on a single GPU. For network implementation, ResNet-101-DCN was adopted as the 2D image feature extraction backbone, and the model was initialized with pretrained FCOS3D parameters to enhance feature extraction capability and accelerate convergence. The multi-view input consists of synchronized image sequences captured by six surround-view cameras. Images from all views are first processed by a shared-weight CNN backbone and then fed into a feature pyramid network to obtain multi-scale semantic representations. The resulting features are subsequently compressed to a fixed channel dimension to match the subsequent view transformation and BEV representation learning stages.

During training, the AdamW optimizer was employed, with the initial learning rate set to $2\times10^{-4}$. A cosine annealing strategy was adopted for learning rate decay. The model was trained for 200 epochs, and evaluation was conducted on the nuScenes validation set after each epoch. The total batch size for the six-camera input was set to 6. In addition, synchronized batch normalization was adopted to improve training stability.

To improve robustness and generalization, the conventional data augmentation strategies used in the baseline were retained during training, including random rotation, scaling, and flipping, together with slight view perturbation. For comparison, Zero-DCE, IAT, SCI, NeRCo, and SMG were all inserted into the same front-end position of the SpecBEV framework, that is, each of the six input views was first processed by the corresponding front-end method and then fed into the same multi-view visual BEV detection pipeline for feature extraction, view transformation, and 3D object detection. The proposed ICF module follows a view-wise processing strategy with cross-view parameter sharing, which preserves the consistency of the enhancement mechanism while maintaining adaptive responses to different camera inputs.

The experimental environment and parameter configuration of this paper are shown in **Table 1**.

**Table 1.** Experimental environment and parameter configuration

| Environment and Parameter | Name | Details |
|---|---|---|
| Hardware Environment | CPU | 18 vCPU AMD EPYC 9754 128-Core Processor |
| | GPU | RTX 4090D (24GB) |
| Software Environment | Programming Language | Python 3.10.18 |
| | Computing Platform | CUDA 12.4 |
| | Deep Learning Framework | PyTorch 1.13 |
| Training Setting | Batch Size | 6 |
| | Workers | 16 |
| | Initial Learning Rate | $2\times10^{-4}$ |
| | Learning Rate Schedule | Cosine annealing |
| | Epochs | 200 |
| | Optimizer | AdamW |
| Input Setting | Camera Views | 6 Surround-view Cameras |

## 4.2. Dataset introduction

The experiments are conducted on the nuScenes dataset using the official training and validation split [24]. nuScenes is a widely used large-scale multimodal dataset for autonomous driving, containing 1000 driving scenes in total, among which the training, validation, and test sets contain 700, 150, and 150 scenes, respectively. Each scene lasts approximately 20 s and is annotated at fixed key-frame intervals. In this work, synchronized RGB images from six surround-view cameras are used as inputs, and the multi-view image features are projected into a unified BEV space through the view transformation module to generate a bird's-eye-view representation with a resolution of 200 × 200.

Different from settings that focus only on unified-space representation optimization, this work further investigates the effect of image front-end enhancement on multi-view visual 3D object detection under complex lighting conditions. The nuScenes dataset covers diverse real-world driving scenarios, including daytime, nighttime, and rainy scenes, and can therefore effectively reflect the imaging variations and environmental disturbances encountered in practical multi-view visual perception. This makes it suitable for evaluating the robustness of the proposed method under challenging illumination conditions [24].

The complete scene distribution is used for unified training and evaluation, without any additional filtering of the training data according to illumination conditions. This setting is consistent with the objective of this work, namely, to improve the overall stability of multi-view BEV detection under diverse lighting variations rather than to specialize the model for one specific degradation scenario. All results are reported on the full nuScenes validation set to reflect the overall detection capability of different front-end enhancement strategies under realistic complex scene distributions (**Figure 4**).

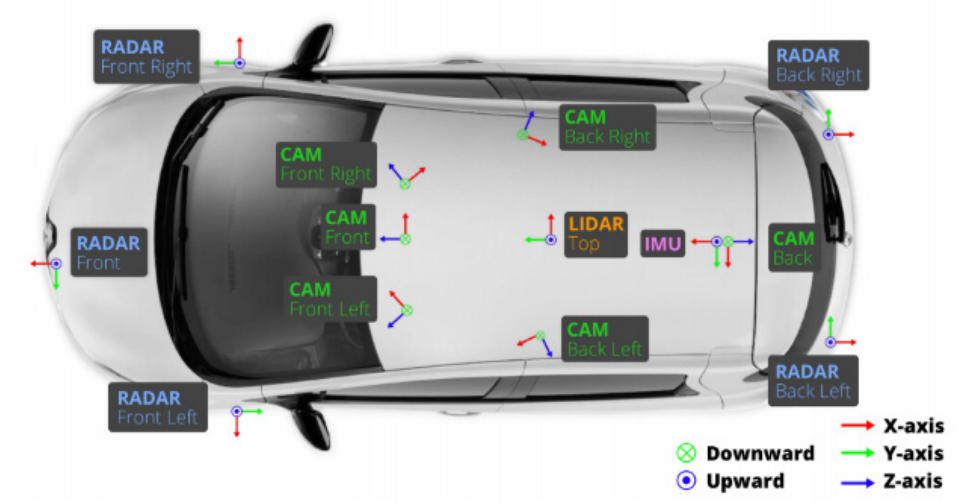


**Figure 4.** Sensor configuration of the data acquisition platform.[24]

## 4.3. Experimental evaluation metrics

The official nuScenes 3D object detection evaluation protocol is adopted for unified comparison among different methods. The main evaluation metrics include mean Average Precision (mAP) and nuScenes Detection Score (NDS). In addition, five true positive error metrics are also reported, namely mATE, mASE, mAOE, mAVE, and mAAE [24].

Among them, mAP is used to measure the overall detection accuracy under different object categories and distance thresholds. In the nuScenes evaluation protocol, mAP is computed based on the two-dimensional center distance between the predicted box and the ground-truth box, and the average precision of each category is calculated under multiple distance thresholds. It is defined as:

$$mAP = \frac{1}{|C|\ |D|} \sum_{c \in C} \sum_{d \in D} AP(c,d) \tag{11}$$

where $C$ denotes the set of object categories, corresponding to the 10 detection classes in nuScenes, and $D$ denotes the set of distance thresholds, defined as $\{0.5,1,2,4\}m$. Here, $AP(c,d)$ represents the average precision computed for category $c$ under distance threshold $d$.

Building upon this, NDS combines detection accuracy with several true positive error measures, enabling a more holistic assessment of the overall performance of 3D object detection models. It is defined as:

$$NDS = \frac{1}{10}\left[5 \cdot mAP + \sum_{mTP \in M} \left(1 - min\left(1, mTP\right)\right)\right] \tag{12}$$

where $M$ denotes the set of five true positive error metrics, including mATE, mASE, mAOE, mAVE, and mAAE. The term $min(1,mTP)$ is used to truncate the error values in order to avoid excessive influence from abnormal error cases on the overall score.

In addition to mAP and NDS, the five official nuScenes true positive error metrics are also reported to analyze the influence of different front-end enhancement methods on each subtask of 3D object detection. Specifically, mATE measures the translation error of the object center, mASE characterizes the scale discrepancy between the predicted box and the ground-truth box, mAOE reflects the orientation estimation error, mAVE measures the velocity prediction error, and mAAE evaluates the attribute prediction error. Lower values of these metrics indicate better performance on the corresponding subtasks. In the subsequent analysis, mAP and NDS are taken

as the primary criteria, while the five error metrics are used to further evaluate localization accuracy, scale estimation ability, orientation prediction ability, and dynamic attribute modeling capability.

## 4.4. Comparison of experimental results

To verify the effectiveness of the proposed SpecBEV-IR, quantitative comparisons were conducted on the nuScenes validation set against several representative image front-end enhancement methods, including IAT, Zero-DCE, SCI, NeRCo, and SMG [9,17–19,25]. All comparison methods were inserted into the same front-end position of the SpecBEV framework, while the image feature encoder, view transformation module, BEV representation learning pipeline, and detection head were kept unchanged to ensure a fair comparison. The quantitative results are reported in **Table 2**.

**Table 2.** Comparison with representative front-end enhancement methods on the nuScenes validation set

| Method | Year | mAP | NDS | mATE↓ | mASE↓ | mAOE↓ | mAVE↓ | mAAE↓ |
|---|---|---|---|---|---|---|---|---|
| IAT [19] | 2019 | 0.3424 | 0.4450 | 0.7041 | 0.2607 | 0.5485 | 0.5055 | 0.2431 |
| Zero-DCE [9] | 2020 | 0.3936 | 0.4968 | 0.5752 | 0.2417 | 0.5398 | 0.4185 | 0.2251 |
| SCI [17] | 2022 | 0.3836 | 0.4871 | 0.5841 | 0.2437 | 0.5599 | 0.4300 | 0.2294 |
| NeRCo [18] | 2023 | 0.3571 | 0.4609 | 0.6518 | 0.2616 | 0.5404 | 0.4877 | 0.2341 |
| SMG [25] | 2023 | 0.3584 | 0.4673 | 0.6383 | 0.2594 | 0.5426 | 0.4505 | 0.2278 |
| Ours | --- | 0.4121 | 0.5174 | 0.5637 | 0.2248 | 0.4632 | 0.4277 | 0.2074 |

From the overall results, the proposed method achieves the best performance in both mAP and NDS, reaching 0.4121 and 0.5174, respectively. Compared with the strongest competing method, Zero-DCE, the proposed method further improves mAP by 0.0185 and NDS by 0.0206 [9]. Compared with SCI, the improvements are 0.0285 in mAP and 0.0303 in NDS. These results indicate that the task-oriented enhanced representation generated by combining illumination-invariant cues with original image content is more suitable for the multi-view visual BEV detection pipeline than methods that mainly focus on improving the visual appearance of individual images.

With respect to the true positive error metrics, the proposed method achieves the best results in mATE, mASE, mAOE, and mAAE, with values of 0.5637, 0.2248, 0.4632, and 0.2074, respectively. The reduction in mATE indicates that the proposed method improves the accuracy of object center localization. The improvement in mASE shows that the enhanced inputs are more beneficial for object scale estimation. The lower mAOE suggests that the proposed method provides more stable geometric representations for orientation prediction, while the reduction in mAAE reflects improved attribute discrimination. The only metric on which the proposed method is not the best is mAVE. Zero-DCE obtains the lowest velocity error of 0.4185, whereas the proposed method achieves 0.4277 [9]. The gap, however, remains small, indicating that the proposed method still maintains competitive performance on velocity-related estimation while achieving the best overall detection results.

A closer examination of the compared methods shows that image-domain enhancement does not necessarily lead to monotonic improvements in 3D detection performance. IAT yields the weakest overall results, with mAP and NDS of only 0.3424 and 0.4450, respectively, and it also performs poorly across most error metrics [19]. Although NeRCo and SMG have shown effectiveness in certain image enhancement tasks, their performance in the current multi-view visual BEV detection framework remains limited, with mAP values of 0.3571 and 0.3584 and NDS values of 0.4609 and 0.4673, respectively [18,25]. This phenomenon suggests that, for

multi-view visual 3D detection, the effectiveness of a front-end enhancement method depends not only on whether brightness or contrast is improved, but also on whether the enhanced multi-view inputs can maintain stable and consistent representations throughout shared encoding, view transformation, and unified BEV fusion.

Zero-DCE and SCI achieve relatively strong overall results, indicating that moderate front-end adjustment of brightness and structural details can indeed improve input quality for downstream detection to some extent [9,17]. Among the comparison methods, Zero-DCE delivers the second-best results in mAP, NDS, mATE, mASE, mAOE, and mAAE, demonstrating strong overall competitiveness [9]. SCI performs slightly worse than Zero-DCE, but still substantially outperforms IAT, NeRCo, and SMG [9,17–19,25]. This observation further indicates that the requirement of the current task is not simply low-light restoration, but rather the preservation of original semantic information together with improved structural stability and reduced distribution shifts across different camera views.

The superior performance of the proposed method can be attributed to the fact that its front-end design is not limited to perceptual image enhancement, but instead performs input robustness modeling for downstream 3D detection. Specifically, ICE constructs intermediate responses that are less sensitive to illumination variation and thus provides more stable structural priors for the input. Fuse Conv further integrates these invariant cues with the original image content, so that the resulting enhanced representation preserves scene semantics while achieving stronger illumination robustness. In addition, the cross-view parameter-sharing strategy ensures consistency in the enhancement mechanism across all six camera views, thereby alleviating the negative influence of illumination-induced distribution shifts on the shared encoder and subsequent unified-space representation learning.

The results in **Table 2** demonstrate that front-end enhancement for multi-view visual BEV 3D object detection cannot be treated as a direct extension of single-image enhancement methods. Only when the enhancement process jointly considers input stability, structural preservation, and cross-view consistency can the front-end module truly improve downstream detection performance. The proposed SpecBEV-IR achieves the best overall results and the best performance on most true positive error metrics, which verifies its effectiveness in improving multi-view visual 3D object detection under complex lighting conditions.

## 4.5. Ablation experiment results

To further evaluate the effectiveness of the proposed illumination-robust front-end enhancement module, ablation experiments were conducted on the SpecBEV baseline. Except for the front-end design, all other settings, including the backbone, training schedule, and evaluation protocol, remained unchanged. The results are reported in **Table 3**.

**Table 3.** Ablation study of the proposed front-end enhancement module

| Model | ICE | Fuse Conv | Shared Across Views | mAP | NDS |
|---|---|---|---|---|---|
| SpecBEV baseline | | | | 0.3656 | 0.4711 |
| | √ | | | 0.3914 | 0.4854 |
| | | √ | | 0.3672 | 0.4609 |
| | √ | √ | | 0.4081 | 0.5058 |
| | √ | √ | √ | 0.4121 | 0.5174 |

Without any front-end enhancement, the baseline model achieves 0.3656 mAP and 0.4711 NDS on the

nuScenes validation set. This result indicates that under complex lighting conditions, instability in brightness distribution, local contrast, and structural details in the raw multi-view inputs still affects subsequent shared feature extraction and BEV-space detection.

When only the ICE branch is introduced, the performance increases to 0.3914 mAP and 0.4854 NDS, yielding gains of 0.0258 mAP and 0.0143 NDS over the baseline. This improvement shows that the illumination-invariant cues extracted by ICE can effectively alleviate the degradation caused by challenging lighting conditions and provide more stable structural information for subsequent feature encoding.

In contrast, using Fuse Conv alone leads to 0.3672 mAP and 0.4609 NDS. Although the mAP is slightly higher than that of the baseline, the NDS decreases. This suggests that fusion alone is insufficient to improve robustness when explicit illumination-invariant cues are absent. Without a stable cue extraction process, direct feature adjustment at the input stage may introduce additional disturbance to the original image distribution and weaken downstream spatial modeling.

After combining ICE and Fuse Conv into the complete ICF module, the performance further improves to 0.4081 mAP and 0.5058 NDS. Compared with the variant using ICE only, the complete module brings an additional gain of 0.0167 mAP and 0.0204 NDS. This result indicates that the two components play complementary roles. ICE is responsible for extracting relatively stable illumination-invariant information, while Fuse Conv integrates these cues with the original image content to construct a more suitable input representation for downstream BEV detection.

When the shared-across-views strategy is further introduced, the model reaches the best performance, with 0.4121 mAP and 0.5174 NDS. Compared with the non-shared ICF variant, this setting improves mAP by 0.0040 and NDS by 0.0116. The improvement suggests that, in the multi-view setting, parameter sharing provides a more consistent enhancement mechanism across different camera views, which is beneficial for reducing view-wise distribution discrepancies and improving the stability of subsequent shared encoding and BEV feature fusion.

These results show that the performance gain does not come from simply adding a front-end module. The main contribution lies in the explicit extraction of illumination-invariant cues and their effective fusion with the original image content. In addition, the shared processing strategy is beneficial for maintaining consistency across multi-view inputs, which is particularly important for multi-camera BEV 3D object detection under complex lighting conditions.

## 4.6. Visualization analysis

To further examine the influence of different front-end enhancement methods on downstream multi-view BEV 3D detection, qualitative results in a representative nighttime scene are presented in **Figure 5**. **Figure 5(a)**, **Figure 5(b)**, and **Figure 5(c)** correspond to the proposed method, Zero-DCE, and SMG, respectively [9,25]. In all cases, the red boxes denote the ground-truth annotations and the blue boxes denote the predicted results.


**Figure 5.** Qualitative comparison of detection results in a representative nighttime scene. In each subfigure, the left part shows six-camera-view detection results, where yellow boxes denote predictions and blue boxes denote ground truth. The right part shows the corresponding BEV visualization, where dark blue boxes denote predictions and red boxes denote ground truth.(a) Ours; (b) Zero-DCE; (c) SMG.

From the overall comparison, the proposed method produces prediction boxes that are more consistent with the ground-truth boxes. This advantage can be observed in both the image views and the BEV view. For several vehicle targets in the forward and right-side views, the predicted boxes of the proposed method

are closer to the ground-truth boxes in terms of position, scale, and orientation. This indicates that the proposed front-end enhancement module is more effective in preserving task-relevant structural information under complex nighttime illumination and in providing more stable inputs for the subsequent shared feature extraction, view transformation, and BEV-space detection process.

A more noticeable difference appears in the dense small-object region in the lower-left area of the BEV view. In **Figure 5(a)**, the proposed method yields more complete predictions for this region, and most predicted boxes maintain better spatial correspondence with the ground-truth annotations. By contrast, the predictions of Zero-DCE and SMG in the same area are less stable, with more obvious offsets for some objects and weaker separation among nearby targets [9,25]. This result is particularly meaningful for multi-view BEV detection, because dense small objects are usually more vulnerable to the combined influence of weak texture, insufficient illumination, and cross-view projection error. The improved performance in this region suggests that the proposed method is better at preserving fine target structure and maintaining detection reliability under challenging lighting conditions.

In the central and right regions of the BEV view, the predicted boxes generated by the proposed method also exhibit better alignment with the annotated objects. Some targets that are covered more accurately in **Figure 5(a)** show larger localization deviations or lower overlap quality in **Figure 5(b)** and **Figure 5(c)**. This observation suggests that the proposed method does not merely improve image appearance, but more importantly, it reduces the disturbance caused by illumination degradation to unified spatial representation learning, leading to more accurate object localization in BEV space.

Consistent behavior can also be observed across different camera views. In the front and front-right views, the proposed method provides more stable box placement for the main vehicle targets. In the rear-related views, although the scene remains dark and visually challenging, the proposed method still preserves relatively reliable responses for valid objects. This indicates that the shared-across-views enhancement strategy is beneficial for reducing view-wise distribution discrepancies during front-end processing, thereby improving the stability of shared encoding and subsequent cross-view fusion.

The qualitative results in **Figure 5** are consistent with the quantitative comparisons reported in **Table 2** and the ablation results in **Table 3**. The advantage of the proposed method lies not in stronger visual enhancement alone, but in its ability to construct a more suitable input representation for downstream multi-view BEV 3D detection. By extracting illumination-invariant cues and integrating them with the original image content, the proposed method improves detection stability, localization accuracy, and target completeness under complex nighttime lighting conditions.

## 5. Conclusion

This paper addresses the problem that multi-view visual BEV 3D object detection is highly vulnerable to input degradation under complex lighting conditions. In challenging scenarios such as nighttime, backlighting, local overexposure, and uneven illumination, the quality of raw multi-view images may deteriorate significantly, which further affects shared feature extraction, view transformation, and unified spatial representation learning. To alleviate this problem, this paper proposes SpecBEV-IR, an illumination-robust front-end enhancement method built upon the SpecBEV framework. By introducing task-oriented robustness modeling at the image-domain front end, the proposed method aims to provide more stable and more suitable input representations for downstream BEV 3D detection.

The proposed method introduces an illumination-robust image front-end enhancement module, termed ICF, between the multi-view input images and the shared 2D encoder. The ICF module consists of two components. ICE is designed to extract relatively stable illumination-invariant cues from raw images, while Fuse Conv integrates these cues with the original image content to generate enhanced representations that preserve scene semantics and structural information under illumination variation. In addition, a shared-across-views strategy is adopted so that all camera views follow a consistent enhancement mechanism, which helps reduce view-wise distribution discrepancy before shared encoding and subsequent BEV fusion. This design makes the proposed method different from conventional image enhancement approaches that mainly focus on improving visual appearance. Instead, the proposed front-end enhancement is explicitly designed to serve downstream multi-view BEV 3D detection.

Experimental results on the nuScenes validation set demonstrate the effectiveness of the proposed method. SpecBEV-IR achieves 0.4121 mAP and 0.5174 NDS, outperforming several representative front-end enhancement methods, including Zero-DCE, SCI, IAT, NeRCo, and SMG. The ablation study further verifies that ICE provides the main gain in illumination robustness, Fuse Conv improves the integration of invariant cues and original image content, and the shared-across-views strategy contributes to more stable multi-view enhancement. In addition, the qualitative results show that the proposed method produces more reliable predictions in nighttime scenes, especially in dense small-object regions and other locally challenging areas, where better target completeness and localization accuracy can be observed.

The results of this work indicate that illumination handling for multi-view visual BEV 3D object detection should be considered from the perspective of downstream detection rather than simple perceptual enhancement. A front-end module that explicitly takes input stability, structural preservation, and cross-view consistency into account is more beneficial to unified spatial representation learning under complex lighting conditions. This also suggests that robust input modeling is a meaningful direction for improving camera-based 3D perception in real-world driving environments.

Future work can be carried out in two directions. One is to further reduce the computational overhead of the front-end enhancement module and improve its suitability for real-time onboard deployment. The other is to extend the proposed method to more diverse degradation scenarios, such as rain, haze, glare, and stronger dynamic illumination changes, so as to further improve the robustness of multi-view visual 3D perception in complex real-world environments.

## Disclosure statement

The author declares no conflict of interest.

## References

[1] Li H, Sima C, Dai J, et al., 2023, Delving into the Devils of Bird's-Eye-View Perception: A Review, Evaluation and Recipe. IEEE Transactions on Pattern Analysis and Machine Intelligence.

[2] Ma Y, Wang T, Bai X, et al., 2024, Vision-Centric BEV Perception: A Survey. IEEE Transactions on Pattern Analysis and Machine Intelligence, 46(12): 10978–10997.

[3] Philion J, Fidler S, 2020, Lift, Splat, Shoot: Encoding Images from Arbitrary Camera Rigs by Implicitly Unprojecting to 3D, Computer Vision – ECCV 2020, 194–210.

[4] Huang J, Huang G, Zhu Z, et al., 2021, BEVDet: High-Performance Multi-Camera 3D Object Detection in Bird-Eye-View, arXiv preprint arXiv:2112.11790.

[5] Li Y, Ge Z, Yu G, et al., 2023, BEVDepth: Acquisition of Reliable Depth for Multi-View 3D Object Detection. Proceedings of the AAAI Conference on Artificial Intelligence, 37(2): 1477–1485.

[6] Li Z, Wang W, Li H, et al., 2022, BEVFormer: Learning Bird's-Eye-View Representation from Multi-Camera Images via Spatio Temporal Transformers, European Conference on Computer Vision, 1–18.

[7] Zhu Z, Zhang Y, Chen H, et al., 2023, Understanding the Robustness of 3D object Detection with Bird's-Eye-View Representations in Autonomous Driving, Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 21600–21610.

[8] Wang S, Zhao X, Xu H, et al., 2023, Towards Domain Generalization for Multi-View 3D Object Detection in Bird-Eye-View, Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 13333–13342.

[9] Guo C, Li C, Guo J, et al., 2020, Zero-Reference Deep Curve Estimation for Low-Light Image Enhancement, Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 1780–1789.

[10] Hashmi K, Suresh K, Stricker D, et al., 2025, TorchAdapt: Towards Light-Agnostic Real-Time Visual Perception, Proceedings of the IEEE/CVF International Conference on Computer Vision, 5645–5656.

[11] Wang S, Zeng S, Gu T, et al., 2025, From Enhancement to Understanding: Build a Generalized Bridge for Low-Light Vision via Semantically Consistent Unsupervised Fine-Tuning, Proceedings of the IEEE/CVF International Conference on Computer Vision, 13804–13814.

[12] Li H, Zhao Y, Zhong J, et al., 2025, Delving into the Secrets of BEV 3D Object Detection in Autonomous Driving: A Comprehensive Survey. IEEE Transactions on Intelligent Transportation Systems, 27(1): 119–144.

[13] Huang J, Huang G, 2022, BEVDet4D: Exploit Temporal Cues in Multi-Camera 3D Object Detection, arXiv preprint arXiv:2203.17054.

[14] Liu Y, Wang T, Zhang X, et al., 2022, PETR: Position Embedding Transformation for Multi-View 3D Object Detection, European Conference on Computer Vision, 531–548.

[15] Wang Y, Guizilini V, Zhang T, et al., 2022, DETR3D: 3D Object Detection from Multi-View Images via 3D-to-2D Queries, Conference on Robot Learning, 180–191.

[16] Li Y, Huang B, Chen Z, et al., 2024, Fast-BEV: A Fast and Strong Bird's-Eye-View Perception Baseline, IEEE Transactions on Pattern Analysis and Machine Intelligence, 2024.

[17] Ma L, Ma T, Liu R, et al., 2022, Toward Fast, Flexible, and Robust Low-Light Image Enhancement, Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 5637–5646.

[18] Yang S, Ding M, Wu Y, et al., 2023, Implicit Neural Representation for Cooperative Low-Light Image Enhancement, Proceedings of the IEEE/CVF International Conference on Computer Vision, 12918–12927.

[19] Zhang Y, Zhang J, Guo X, 2019, Kindling the Darkness: A Practical Low-Light Image Enhancer, Proceedings of the 27th ACM International Conference on Multimedia, 1632–1640.

[20] Wei C, Wang W, Yang W, et al., 2018, Deep Retinex Decomposition for Low-Light Enhancement, arXiv preprint arXiv:1808.04560.

[21] Liu R, Ma L, Zhang J, et al., 2021, Retinex-Inspired Unrolling with Cooperative Prior Architecture Search for Low-Light Image Enhancement, Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 10561–10570.

[22] Li C, Guo C, Loy C, 2021, Learning to Enhance Low-Light Image via Zero-Reference Deep Curve Estimation. IEEE Transactions on Pattern Analysis and Machine Intelligence, 44(8): 4225–4238.

[23] Hashmi K, Kallempudi G, Stricker D, et al., 2023, FeatEnhancer: Enhancing Hierarchical Features for Object Detection and Beyond under Low-Light Vision, Proceedings of the IEEE/CVF International Conference on Computer Vision, 6725–6735.

[24] Caesar H, Bankiti V, Lang A, et al., 2020, nuScenes: A Multimodal Dataset for Autonomous Driving, Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 11621–11631.

[25] Li J, Feng X, Hua Z, 2021, Low-Light Image Enhancement via Progressive-Recursive Network. IEEE Transactions on Circuits and Systems for Video Technology, 31(11): 4227–4240.