IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 23 August 2024, accepted 4 September 2024, date of publication 11 September 2024, date of current version 30 September 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3458412

RESEARCH ARTICLE

MonoMPV: Monocular 3D Object Detection With Multiple Projection Views on Edge Devices

ZHAOXUE DENG[1,2], BINGSEN HAO[1], GUOFANG LIU[3], XINGQUAN LI[2,4], (Fellow, IEEE), HANBING WEI[1], FEI HUANG[5], AND SHENGSHU LIU[5]

[1]School of Mechatronics and Vehicle Engineering, Chongqing Jiaotong University, Chongqing 400074, China
[2]Research and Development Department, Chongqing Changan Automobile Company Limited., Chongqing 400023, China
[3]China Society of Automotive Engineers, Beijing 100176, China
[4]Vehicle Engineering Institute, Chongqing University of Technology, Banan District, Chongqing 400054, China
[5]China Road and Bridge Corporation, Beijing 100010, China

Corresponding author: Bingsen Hao (bingsen_hao@163.com)

This work was supported in part by the Science and Technology Innovation Key R&D Program of Chongqing under Grant CSTB2022TIAD-STX0003, in part by the National Natural Science Foundation of China under Grant 52172381, and in part by the Research and Innovation Program for Graduate Students in Chongqing Jiaotong University under Grant YYK202405.

ABSTRACT In the field of autonomous driving, monocular 3D object detection is focused on the task of representing 3D scenes using a single camera image and conducting 3D object detection. While Bird's-Eye View (BEV) method effectively decreases the computational burden associated with 3D scene representation, its limitation in considering height information can lead to a less accurate depiction of complex 3D structures. This study introduces an innovative monocular 3D object detection framework called MonoMPV. This framework represents a complete 3D scene by mapping spatial objects onto Multi-Projection Views (MPV) without the need for voxelization, thus simplifying the process. Notably, MPV systems consist of Feature Cross-Attention (FCA) and Projection Cross-Attention (PCA) components. FCA aims to enhance image features to MPV level, while PCA enables direct information interaction among the views within MPV. Furthermore, Triplet Loss for Top Feature (TLTF) was employed in conjunction with FCA and PCA to distinguish effectively between top-plane and background features. By engaging in this practice, it is possible to develop more complex 3D structural models and establish precise optimization objectives through TLTF. Consequently, this approach enhances the effective utilization of data by the model. Experimental results on the nuScenes dataset illustrate that this approach surpasses existing monocular 3D object detection methods. To implement algorithms on on-board edge computing devices, the monocular 3D object detection task has been executed on the edge device Jetson Orin NX, ensuring high precision.

INDEX TERMS Autonomous driving, 3D object detection, multi-projection views, edge computing.

I. INTRODUCTION

Accurately perceiving 3D environment is crucial in autonomous driving systems. Advanced perception techniques employ a variety of sensors such as LiDAR [1], [2] radar [3], stereo vision [4], [5] or their combinations to achieve accurate depth perception. Despite the absence of direct depth perception, visual-based models have demonstrated significant advancements in various tasks, including depth estimation [6], semantic map construction [7], and 3D object detection, through the utilization of surrounding cameras. While these systems demonstrate exceptional performance, they present challenges due to their complexity, high expenses, and maintenance complexities. In contrast, monocular 3D object detection, which involves using 2D monocular images for perceiving 3D objects, has been increasingly recognized for its cost-effectiveness.

In recent years, significant developments have been achieved in the effective depiction of 3D scenes for monocular 3D object detection. Two predominant frameworks utilized for this purpose are voxel and Bird's-Eye View (BEV) representations. Although voxel models incorporate

The associate editor coordinating the review of this manuscript and approving it for publication was Domenico Rosaci.

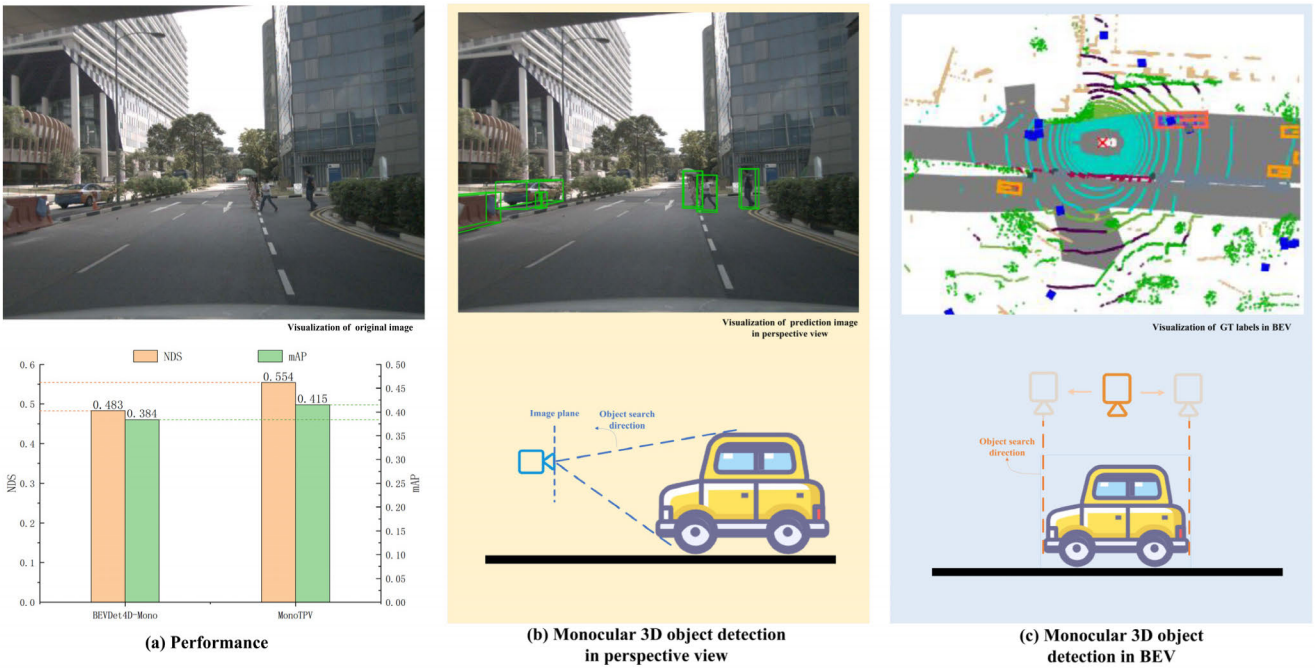


FIGURE 1. Motivation and comparative analysis of detection methods: (a) MPV-based method surpasses existing BEV-based monocular 3D object detection methods in accuracy; (b) For monocular 3D object detection, the search direction is aligned horizontally with the road, leading to potential depth estimation errors that can impair 3D bounding box localization and reduce overall detection accuracy; (c) In BEV-based object detection, the search direction is typically vertical relative to the road. However, BEV generation process can result in information loss, diminishing the accuracy of object detection.

techniques such as sparse convolution [8], they still pose significant computational costs. Similarly, BEV models [9], [10] primarily concentrate on the plane exhibiting the most significant information change, offering a cohesive viewpoint for autonomous driving. They illustrate the location and scale of objects, rendering them appropriate for tasks related to perception and planning. BEV grid vectors implicitly represent 3D information of each object. Despite their computational efficiency, methods based on BEV do not exhibit a significant performance advantage over other approaches in monocular 3D detection. The primary factor is that 3D detection relies on robust BEV features to accurately predict 3D bounding boxes; however, the process of BEV generation inherently results in certain information loss. Therefore, a profound comprehension of 3D environments is crucial for the development of more secure and precise visual-based autonomous driving systems. Meanwhile, it is essential to take into account the constraints posed by limited computational resources and latency when implementing a panoramic driving perception system on edge devices that are frequently utilized in autonomous vehicles. Exploring strategies to enhance the promotion of BEVs through the utilization of refined 3D structural models, while concurrently upholding performance and efficiency, represents a valuable research avenue.

This paper aims to investigate the challenge of absent height information in BEV representation within monocular 3D object detection. This paper elaborates on a more sophisticated 3D spatial environment. Several mutually perpendicular cross-sections are generated by combining BEV with two vertical planes. Multi-Projection Views (MPV) approach enables the description of 3D scenes at various resolutions and the generation of unique attributes for points within 3D space. The research proposes a monocular 3D object detection framework, MonoMPV, which effectively extracts features of MPV from 2D images. The initial step involves the mapping of 2D image features onto MPV spatial grid using image cross-view attention, which serves to enhance the information dimensionality. Subsequently, cross-view mixture attention is conducted among the features of MPV to enhance the interaction across the three planes. Furthermore, to improve the emphasis on top-plane characteristics, Transfer Learning is utilized with TLTF between Feature Cross-Attention (FCA) and Projection Cross-Attention (PCA) to distinctly separate top-plane attributes from background features.

The contributions of this study are summarized as follows:

- A novel monocular 3D object detection framework (MonoMPV) is proposed. The framework addresses the limitations of BEV by advancing to MPV for monocular 3D object detection. This paper provides a comprehensive analysis and detailed explanation of the underlying mechanisms driving this advancement. We have deployed monocular 3D object detection tasks on the Jetson Orin NX embedded platform, enabling the network for real-world scenarios.

- A tailored loss function for top-plane features is introduced, measuring the similarity between top-plane and background features using Triplet Loss. This approach effectively enhances the focus on top-plane features, improving detection accuracy.
- MonoMPV framework demonstrates superior performance on the nuScenes dataset. Experimental results show that MonoMPV achieves improvements of 3.9% in NuScenes Detection Score (NDS) and 3.1% in mean Average Precision (mAP) metrics compared to existing monocular 3D object detection methods.

II. RELATED WORK

A. VOXEL-BASED ENVIRONMENTAL REPRESENTION

This study involves the discretization of 3D space into voxels and the assignment of a vector to each voxel, thus enhancing the representation of 3D structures. This voxel representation technique demonstrates superior performance in tasks such as LiDAR segmentation and 3D scene reconstruction. However, it exhibits lower performance for 3D object detection when compared to methods utilizing BEV [11], [12]. Furthermore, although voxel representations have demonstrated significant advancements in LiDAR perception systems, their utilization in visual-based autonomous driving systems is limited [13]. MonoScene was a pioneer in creating initial voxel representations through the projection of image features along reverse rays into different potential locations in 3D space [14]. Subsequently, these representations were processed using a 3D UNet structure. However, the substantial computational overhead associated with voxel representations poses a significant challenge in expanding multi-view image perception into 3D space. Therefore, the objective is to discover more effective and expressive techniques for representing the complex structure of 3D scenes.

B. BEV-BASED ENVIRONMENTAL REPRESENTION

Research in the field of perspective transformation can be classified into two primary categories: geometric-based and transformer-based approaches. For example, BEVFormer facilitates perspective transformation by generating a BEV grid query set and subsequently engaging in cross-attention with image features using these queries. The Transformer-based Detr3D [15] model is designed to predict 3D object bounding boxes directly from multi-view image data. This model facilitates the transformation from a 2D image representation to a 3D object view. Geometric techniques such as Lift-Splat-Shoot (LSS) [16] incorporate perspective transformation by associating image characteristics with a feature cone determined by depth [17] or height [18], and subsequently projecting the cone features onto a grid known as BEV. BEVDet performs direct image-to-BEV space mapping for the purpose of 3D object detection. Subsequent research endeavors have sought to incorporate LiDAR sensor depth supervision [19] or multi-view stereo technology [20] to improve the accuracy of depth prediction and attain optimal performance. However, the application of BEV for monocular 3D object detection presents significant challenges, necessitating a more detailed delineation of 3D spatial environment.

C. IMPLICIT ENVIRONMENTAL REPRESENTATION

Recent studies have introduced a novel implicit representation technique for scenes that leverages continuous functions. The inputs consist of the 3D coordinates of points, which are subsequently utilized to model the surface of scenes or objects through continuous mathematical functions. Implicit representations, unlike explicit representations such as voxels and BEV, have the capability to model various resolutions while maintaining high computational efficiency [21]. These advantages empower individuals to manage larger and more complex scenarios, thereby enhancing their ability to offer more comprehensive descriptions. Some scholars are specifically motivated by recent developments in hybrid implicit and explicit representations, which represent the pioneering efforts in employing implicit representations for simulating 3D perception in outdoor autonomous driving environments [22]. For monocular 3D object detection, MonoMPV enhances detection performance and efficiency by expanding BEV to a model that captures complex 3D structures.

III. METHOD

This paper expands BEV method to include detailed 3D structure modeling. It introduces MonoMPV framework, which aims to represent the complex 3D structure of a scene on a singular planar surface. In Part 1, a comprehensive explanation has been provided on the transition process from BEV to the generation of MPV. This involves enhancing image features from the initial plane to MPV plane. In Part 2, a monocular 3D object detection framework (MonoMPV) is proposed, building upon MPV to efficiently extract features from a single 2D image. In Part 3, the objective function is explicitly optimized by employing Transfer Learning with a TLTF to effectively distinguish top-plane features from the background.

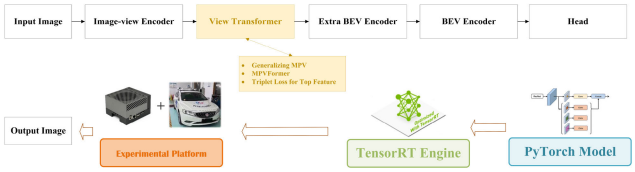


FIGURE 2. The block diagram of our MonoMPV method. (Note: explain our methodology using a block diagram and have implemented model deployment on the embedded device Jetson Orin NX.)

A. GENERALIZING MPV

1) MOTIVATION

Autonomous driving perception commonly necessitates the effective representation of complex 3D scenes, with voxel and BEV representations being widely embraced frameworks.

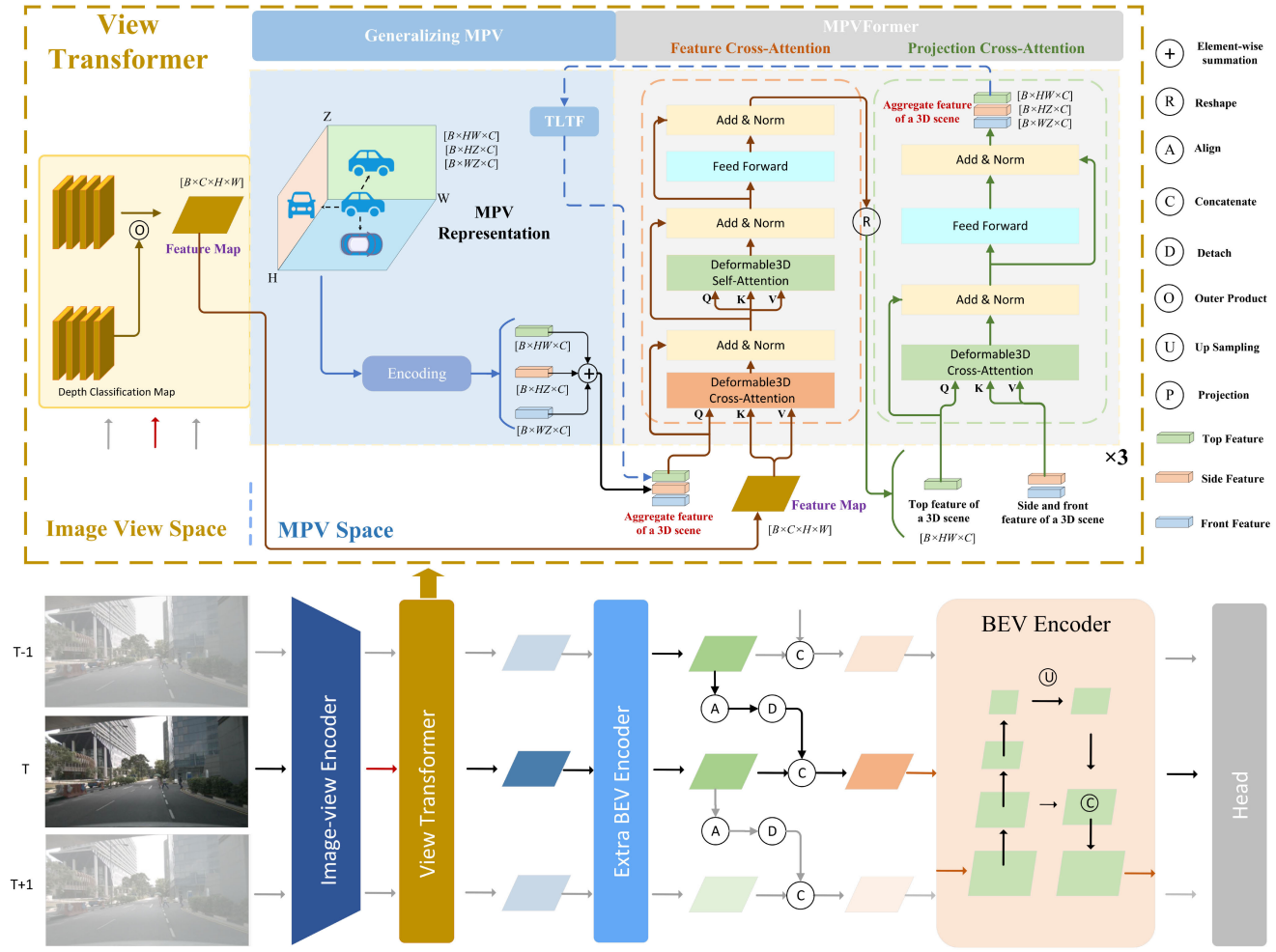


FIGURE 3. Architecture of MonoMPV method. (Note:MonoMPV retains intermediate features from the previous frame and fuses them with features generated from the current frame to serve as input for subsequent processing. Image-view Encoder generates feature maps using ResNet and FPN structure. View Transformer converts features from the image view to MPV, with FCA and PCA implementing attention on MPV features and feature maps, respectively. Candidate features are adjusted using an Extra BEV Encoder prior to temporal fusion. BEV Encoder performs further encoding on BEV features. Finally, a task-specific head is built based on BEV features to predict the target values of 3D objects.)

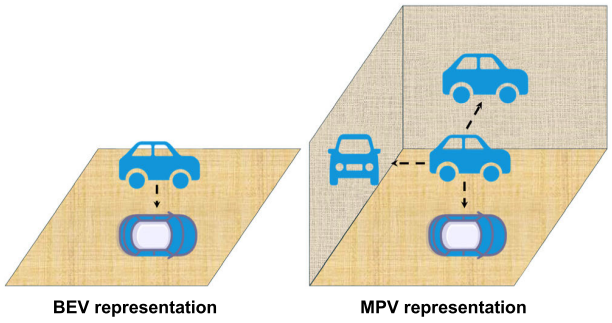


FIGURE 4. Comparisons of proposed MPV, voxel, and BEV representations. (Note: The proposed MPV representation is compared with voxel and BEV representations. While BEV is more efficient than voxel representation, it discards height information and cannot comprehensively describe a 3D scene.)

Voxel features pose computational intensity and present challenges for real-time onboard applications. While BEV decreases computational burden, the complete exclusion of height can negatively affect its level of expressiveness. To address this issue, it is suggested to employ an MPV representation that can comprehensively depict 3D space without diminishing any dimension and circumventing cubical intricacy (Figure 4).

2) OVERALL STRUCTURE

The initial step involves projecting the features onto three planes to derive the characteristics of a point within a 3D space. Subsequently, the features of each projection point are acquired through bilinear interpolation. The comprehensive features of 3D point are derived by amalgamating the three projection features. Therefore, MPV representation illustrates the different characteristics of points within 3D space.

MPV plane is formally examined, which is comprised of three projection planes that are mutually perpendicular as

follows:

$$F = [F^{\mathrm{HW}}, F^{\mathrm{DH}}, F^{\mathrm{WD}}], \tag{1}$$

where $F^{\mathrm{HW}} \in R^{H\times W\times C}$, $F^{\mathrm{DH}} \in R^{D\times H\times C}$,and $F^{\mathrm{WD}} \in R^{W\times D\times C}$ represent the top, side, and front projection views of a 3D scene, respectively; C represents the dimension of features; while H, W, and D represent the resolution of the three planes. The intuitive understanding of these angles can offer supplementary insights into the scene, facilitating the comprehension of a complex scenario when observed from various viewpoints.

3) MPV FEATURE FORMULATION

MPV representation integrates characteristics from its projections in the top, side, and front perspectives to provide a comprehensive description of a point (x, y, z) in the physical world. Specifically, the initial step involves projecting the point onto MPV plane to derive coordinates $[(h,w),(d,h),(w,h)]$. Subsequently, sampling is conducted from these locations on MPV plane to acquire corresponding features $[f_{h,w}, f_{d,h}, f_{w,d}]$. Finally, the aggregation of these three features results in the generation of $f_{x,y,z}$ as follows:

$$\begin{aligned} f_{h,w} &= \mathrm{Sam}(F^{\mathrm{HW}}, (h, w)) = \mathrm{Sam}(F^{\mathrm{HW}}, T_{\mathrm{hw}}(x, y)), \\ f_{d,h} &= \mathrm{Sam}(F^{\mathrm{DH}}, (d, h)) = \mathrm{Sam}(F^{\mathrm{DH}}, T_{\mathrm{dh}}(z, x)), \\ f_{w,d} &= \mathrm{Sam}(F^{\mathrm{WD}}, (w, d)) = \mathrm{Sam}(F^{\mathrm{WD}}, T_{\mathrm{wd}}(y, z)), \quad (2) \\ f_{x,y,z} &= J(f_{h,w}, f_{d,h}, f_{w,d}), \quad (3) \end{aligned}$$

where the aggregation function J and the sampling function Sam are achieved through summation and bilinear interpolation, respectively. Due to the alignment of MPV plane with the real-world plane, each projection function T performs simple scaling on the real-world plane.

MPV plane is unfolded along its orthogonal directions and combined to create a complete 3D feature space that resembles a voxel feature space. The computational complexity of MPV representation is O(HW+DH+WD), which is one order of magnitude lower compared to voxel feature representation. In summary, the representations of MPV can range from a single BEV to complementary MPV, offering a more detailed comprehension of 3D environment, all the while upholding efficiency.

B. MPVFORMER

1) MOTIVATION

The significant information loss that occurs during BEV generation process presents challenges for the utilization of BEV in 3D space for detailed representation. Consequently, BEV representation is expanded to MPV representation of complex 3D structures. Given that attention mechanisms demonstrate proficiency in managing lengthy sequences and complex interdependencies, this study exploits the robust aggregation capacity of the Transformer model to acquire sophisticated MPV features. To achieve this objective, a Transformer-based encoder (MPVFormer) is introduced, which leverages attention mechanisms to enhance the detailed representation of objects in 3D space.

2) OVERALL STRUCTURE

In Figure 3, MPVFormer model comprises two main components: FCA and PCA. Specifically, MPV queries pertain to groups of feature vectors as defined by Equation (2). Each MPV query represents the features of a grid cell in one of the three planes. These queries are utilized to encode precise view information associated with the pillar area. Subsequently, FCA leverages deformable attention [23] to consolidate visual information from the image features. Given the significance of top-plane features in detection, PCA mechanism is employed to extend attention to MPV features in orthogonal directions, thereby enhancing the collection of contextual information. Finally, MPVFormer model is constructed by sequentially stacking three FCA blocks and three PCA blocks. In the subsequent sections, a detailed description of each module in the proposed MPVFormer will be provided.

3) MPV QUERIES

While MPV queries and MPV planes both pertain to the identical collection of 2D features as outlined in (1), they serve different functions for attention mechanisms and 3D presentation scenarios. Each MPV query is associated with a 2D grid region of size $b \times bm^2$ in the designated viewpoint, which then corresponds to a 3D pillar region extending from the viewpoint vertically. During the proposed operational process, MPV queries serve to initially enhance the original visual information of image features through FCA block. Subsequently, the contextual information of other queries obtained from PCA block is employed for optimization. In practice, this study designates MPV queries as learnable parameters and pre-insert 3D position embeddings in the initial encoder layer.

4) FCA MODULE

The initial segment of MPVFormer employs FCA module to efficiently extract visual information from image features (Figure 3). Specifically, FCA is employed to enhance the multi-scale image feature maps to MPV level. Thanks to the high-resolution characteristics of MPV queries and image feature mapping, the computation of full-sized cross-attention is infeasible. Consequently, an effective Deformable3D Cross-Attention is proposed for FCA.

The process of querying all projection views in MPV exhibits similarity. The top view is considered as an illustrative example. Upon sampling the reference point (h, w), the coordinates (x, y) is initially determined in the real-world top view by applying the inverse projection function $\mathrm{Proj}_{\mathrm{top}}^{-1}$. Subsequently, $N_{\mathrm{ref}}^{\mathrm{top}}$ reference points are uniformly sampled along the top view direction for querying $f_{h,w}$, which can be

expressed as follows:

$$(x, y) = \text{Proj}_{\text{hw}}^{-1}(h, w) = (b \times (h - \frac{H}{2}), b \times (w - \frac{W}{2})), \tag{4}$$

$$\text{Refer}_{\text{top}}^{\text{world}} = (\text{Proj}_{\text{top}}^{-1}(h, w), Z) = \{(x, y, z_i)\}_{\text{i=1}}^{N_{\text{ref}}^{\text{top}}}, \tag{5}$$

where $\text{Refer}_{\text{top}}^{\text{world}}$ denotes the set of reference points used to query point $f_{h,w}$ in the world coordinate. It is important to note that the quantity of reference points N_{ref} may vary as a result of different axis ranges. After obtaining the reference point of $f_{h,w}$, it is necessary to project it onto pixel coordinates as follows:

$$\text{Refer}_{\text{top}}^{\text{pix}} = \text{Proj}_{\text{pix}}(\text{Ref}_{\text{top}}^{\text{world}}) = \text{Proj}_{\text{pix}}(\{(x, y, \boldsymbol{z_i})\}), \tag{6}$$

where $\text{Refer}_{\text{top}}^{\text{pix}}$ represents the set of reference points used to query point $f_{h,w}$ in the pix coordinate; and Proj_{pix} represents the projection function determined by the exterior and interior of camera.

Subsequently, a functional architecture is utilized, consisting of Deformable3D Cross-Attention, Deformable3D Self-Attention, and a Feed-Forward Network (FFN). Residual connections are employed around each of the three sub-layers, followed by normalization. The output of each sub-layer is $\text{LN}(x + \text{Sublayer}(x))$, where $\text{Sublayer}(x)$ represents the function implemented by the respective sub-layer. FCA is formally implemented as follows:

$$\text{FCA}(f_{h,w}, I) = \text{LN}(\text{FFN}(\text{Phase}_2) + \text{Phase}_2), \tag{7}$$

where LN represents LayerNorm function; $\text{FCA}(f_{h,w}, I)$ represents the feature through FCA; and Phase_2 represents the intermediate features from the second stage of FCA, which can be determined as follows:

$$\text{Phase}_2 = \text{LN}(\text{SelfDA3D}(\text{Phase}_1) + \text{Phase}_1), \tag{8}$$

where SelfDA3D represents Deformable3D Self-Attention; and Phase_1 represents the intermediate features from the first stage of FCA, which can be expressed as follows:

$$\text{Phase}_1 = \text{LN}(\text{CrossDA3D}(f_{h,w}, \text{Refer}_{\text{top}}^{\text{pix,j}}, I) + I), \tag{9}$$

where $f_{h,w}$ represents the top view feature from MPV sample and CrossDA3D represents Deformable3D Cross-Attention.

5) PCA MODULE

PCA block utilizes 3D deformable attention to capture contextual information following FCA block operation. Specifically, MPVFormer employs FCA to extract image features; however, it does not incorporate direct interactions among MPV. Therefore, PCA is selected, which enables queries to share information across various perspectives, thereby aiding in context extraction. Meanwhile, deformable attention is employed to mitigate computational burden. Specifically, given that critical features are top-view directional features in MPV, top-view directional features are designated as queries, while the other two directions are designated as key and value. Taking MPV query located at (h,w) as an example, the reference points are divided into three separate subsets as follows:

$$\text{Ref}_{h,w} = \text{Ref}_{h,w}^{\text{top}} \cup \text{Ref}_{h,w}^{\text{side}} \cup \text{Ref}_{h,w}^{\text{front}}, \tag{10}$$

where $\text{Ref}_{h,w}^{\text{top}}$, $\text{Ref}_{h,w}^{\text{side}}$, and $\text{Ref}_{h,w}^{\text{front}}$ represent the reference points for the top, side, and front planes, respectively. Several points near $f_{h,w}$ were randomly selected to serve as reference points for the top plane. For the side and front planes, the first step involves uniformly sampling 3D points in the direction of the top view. Subsequently, these points are projected onto the side and front planes:

$$\text{Ref}_{h,w}^{\text{side}} = \{(d_i, h)\}_i, \text{Ref}_{h,w}^{\text{front}} = \{(w, d_i)\}_i \tag{11}$$

Subsequently, a functional architecture is utilized, consisting of Deformable3D Cross-Attention and a FFN. Residual connections are applied around each of the two sub-layers, followed by normalization. The output of each sub-layer is $\text{LN}(x + \text{Sublayer}(x))$, where $\text{Sublayer}(x)$ represents the function implemented by the respective sub-layer. Specifically, PCA is implemented as follows:

$$\text{PCA}(f_{h,w}, T) = \text{LN}(\text{FFN}(\text{Phase}_{\text{pca}}) + \text{Phase}_{\text{pca}}), \tag{12}$$

where $\text{PCA}(f_{h,w}, T)$ represents a feature through FCA; $\text{FCA}(f_{h,w}, I)$ is reformulated into $f_{h,w}$ and T, where T is composed of the side and front planes; and $\text{Phase}_{\text{pca}}$ represents the intermediate feature of PCA, which can be determined as follows:

$$\text{Phase}_{\text{pca}} = \text{LN}(\text{CrossDA3D}(f_{h,w}) + f_{h,w}), \tag{13}$$

where $\text{CrossDA3D}(f_{h,w})$ represents Deformable3D Cross-Attention, which is expressed as follows:

$$\text{CrossDA3D}(f_{h,w}) = \text{CrossDA3D}(f_{h,w}, \text{Ref}_{h,w}, T), \tag{14}$$

Finally, MPVFormer was assembled by vertically stacking three HCAB blocks and three HAB blocks. Through this process, MPVFormer facilitates the exchange of information between feature maps and MPV, as well as within MPV itself.

C. TRIPLET LOSS FOR TOP FEATURE (TLTF)

1) MOTIVATION

The traditional method for assessing loss relies on the similarity of probabilistic distributions [24], [25]. The cross-entropy loss between two vectors is computed and iteratively minimized during training to increase their similarity. This study aims to emphasize the central region of the feature map by consolidating data from various locations according to the prominent top-plane features. To enhance the prominence of the top-plane features in MPVFormer, a clear differentiation is made between two attention graphs (i.e., G_n and G_m). This differentiation serves to amplify the similarity or discrepancy between the two vectors. Therefore, the conventional cross-entropy loss is unsuitable for the specific task. Due to the significant role that top-plane features play in the detection process, TLTF is employed to assess the dissimilarity

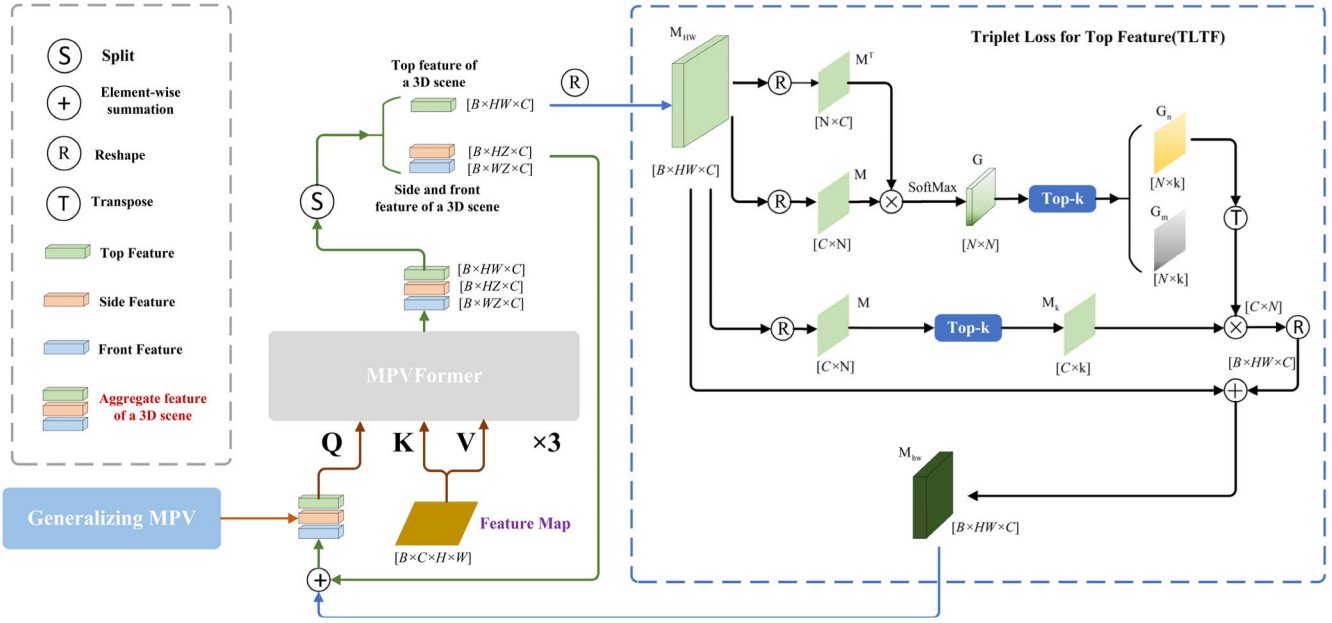


FIGURE 5. Architecture of proposed TLTF. (Note: Given the crucial role of top-plane features in the detection process, TLTF differentiates these features into two attention graphs, focusing on top-plane features and distinctly separating them from the background.)

between vectors of top-plane features. By employing this method, this study emphasized the top-plane features and created a difference between the top-plane feature and the background.

2) TOP-PLANE FEATURES EXTRACTION

The study aims to enhance feature representations by conducting contextual correlation analysis, strengthening essential components of the feature map through attention aggregation facilitated by graph-based information. Firstly, the top-plane features $M_{\mathrm{HW}} \in R^{B\times \mathrm{HW}\times C}$ of MPVFormer was sued as input for TLTF. Secondly, M_{HW} was reformulated into $\widetilde{M} \in R^{B\times HW\times C}$ to obtain the attention map, where $N = B \times HW$ represents the total number of pixels of top-plane features. The spatial attention map $G \in R^{N\times N}$ formally indicates generated by matrix multiplication and softmax process of $\widetilde{M}$ and $\widetilde{M^T}$:

$$G = \mathrm{SoftMax}(\widetilde{M^T}\widetilde{M}), \tag{15}$$

where SoftMax represents the row-wise softmax function.

Subsequently, the top-k ranking function [26] was utilized to extract two attention maps $G_n, G_m \in R^{N\times k}$ from the spatial attention graph G, where k represents the half number of pixels ($k = N/2$). Since Convolutional Neural Networks (CNNs) are designed to activate or highlight necessary features, the more precise components of the top-plane features align with G_n. In contrast to this, the background associated with G_m remains unaffected by the top-plane features.

The operations described are conducted on set G_n to emphasize the top-plane characteristics and highlight more prominent objects. Firstly, the feature map $\widetilde{M_k} \in R^{C\times k}$ was developed by applying the top-k function to $\widetilde{M}$, which has similar attributes to G_n. Secondly, $\widetilde{M_k}$ and G_n^T were multiplied to highlight the characteristics of G_n for the consolidation of the condensed data from $\widetilde{M_k}$ and G_n. Secondly, $\widetilde{M_k}G_n^T$ was normalized using the appropriate scale parameter t, and the results were reformulated to match the size of $R^{C\times H\times W}$. The input features of TLTF was then added to M_{HW} to maintain its initial behavior. Formally, the generation process of the feature map $M_{\mathrm{hw}} \in R^{C\times H\times W}$ is defined as follows:

$$M_{\mathrm{hw}} = \mathrm{Reshape}(t\widetilde{M_k}G_n^T) + M_{\mathrm{HW}}, \tag{16}$$

where Reshape[·] represents the reshape operation of $R^{C\times H\times W}$; and t represents the parameter initialized to 0, which is learned. Therefore, M_{hw} is the version that only enhances the core part of M_{HW}, thereby directing the network's attention towards the aggregation of top-plane features.

3) APPLYING TRIPLET LOSS

Triplet Loss is utilized to quantify the dissimilarity between feature vectors in the top-plane. This method not only prioritizes the top-plane features but also efficiently distinguishes them from the background. Firstly, an auxiliary anchor feature map G_a is defined as a reference point, which acts as the foundation for the two outputs generated by the top-k sorting function. Here, the value of G_a is calculated by subtracting the average of each element in G_n from G_m. Conceptually, G_a encompasses normalized characteristics akin to those in G_n, while also integrating the contextual features of G_m. G_a was utilized to differentiate between G_n and G_m in the discrimination learning process. Formally, a triplet loss function is formulated with respect to three entities denoted as G_a, G_n,

and G_m.

$$f(G_a, G_n, G_m) = [||G_a - G_n||^2 - a]_+ + [a - ||G_a - G_m||^2]_+, \quad (17)$$

where $[\cdot]_+$ performs the same function as $\max\{0, \cdot\}$; the marginal δ is set to 1; G_a attracts G_n, positioning it near the δ edge of G_n. Conversely, the incorporation of the image background in G_m is likely to cause a shift away from G_a.

Finally, the study introduces a module called TLTF, designed to facilitate discriminative learning through the utilization of triplet loss. The training procedure involves the automatic generation of three attention graphs (G_a, G_n, and G_m) for each iteration. Formally, the overall loss function of the network is expressed as follows:

$$L_{\text{total}} = L + \eta L_{\text{triplet}}, \quad (18)$$

where L represents the loss function at the baseline; η represents the regularization coefficient, set to 0.5; L_{triplet} represents a triplet loss for similarity learning; and L_{triplet} is added as a regularization term to the total loss L_{total}. The definition of L_{triplet} is formally determined as follows:

$$L_{\text{triplet}} = \frac{1}{N_G} \sum_{i}^{N_G} f(G_{\text{ai}}, G_{\text{ni}}, G_{\text{mi}}), \quad (19)$$

where L_{triplet} represents the average of $f(G_a, G_n, G_m)$ based on G; N_G represents the number of all possible triplet tuples that can be obtained from G_a, G_n, and G_m. The regularization factor L_{triplet} is valuable due to its ability to stimulate the attention map within TLTF, facilitating discriminative learning through similarity learning.

The combination of MPVFormer and TLTF has been shown to improve the performance and generalization capacity of the model, thereby aiding in the development of more precise and effective models. MPVFormer exhibits significant benefits in representing complex 3D structural models, whereas TLTF offers explicit optimization objectives for the model. By engaging in this process, the model can enhance its data utilization efficiency, continuously optimizing its performance to attain specific objectives.

IV. EXPERIMENTS

This study conducts a comprehensive evaluation of the performance of the proposed models and validate their superiority in practical applications. A comparison is conducted between MonoMPV and established monocular 3D object detection models. Ablation and comparison studies have been conducted to examine the effect of different parameters.

A. EXPERIMENTS SETTINGS

1) DATASET

In this study, the efficacy of the proposed framework was evaluated using the extensive dataset nuScenes [27], which is widely regarded as a benchmark in the field of 3D object detection. This dataset encompasses multimodal data gathered from 1,000 scenes, encompassing RGB images captured by six surround-view cameras, as well as point cloud data obtained from one LiDAR sensor and five Radars. The dataset contains annotations for ten distinct categories of 3D bounding boxes, comprising over 1.4 million instances for detection. The dataset is partitioned into training, validation, and testing sets with a split ratio of 700/150/150, respectively. This partitioning strategy ensures equitable comparisons of various models and approaches. Due to its rich scene diversity and comprehensive categorical information, the nuScenes dataset is increasingly being adopted as a leading standard for 3D object detection. Consequently, it serves as the primary benchmark for assessing the performance of the proposed method.

2) IMPLEMENTATION DETAILS

The experiments were conducted on a hardware setup featuring an Intel I9-13900k CPU, GeForce RTX 4090 GPU, 32GB of RAM, and running the Ubuntu 20.04 operating system. The model framework was implemented using PyTorch 1.10.0 with CUDA 11.3. The feature dimensions for the channel C, FFN-based, and Mlp-based detection headers were uniformly set to 256. A batch size of 16 was used across all experiments. The dimensions of the MPV spaces (mpv_h, mpv_w, and mpv_z) were established as 16, 44, and 8, respectively. The MonoMPV model was trained over 40 epochs with a learning rate of 8e-4. The optimizer employed in this study was AdamW, with a weight decay of 1e-4. The learning rate was adjusted using a OneCycle learning rate scheduler with a linear decay strategy. For deployment on an edge device, the operating environment consisted of Ubuntu 20.04, with CUDA 11.4, TensorRT 8.5, and cuDNN 8.6 ensuring compatibility and performance optimization.

FIGURE 6. Jetson Orin NX edge device.

TABLE 1. Deployment environment.

Category	Details
Operating System	Jetpack 5.1.1(ubuntu 20.04))
Memory	Jetson AGX Xavier(32G), Jetson Orin NX(16G)
GPU Library	CUDA==11.4, TensorRT==8.5.2.2, cuDNN==8.6.0.166
Opencv-python	Opencv-C++==4.5.4

TABLE 2. Results on the nuScenes dataset. (Note: The best results are emphasized using bold numbers, while the second-best results are represented with blue. The table presents the results of BEVDet4D-Mono, which were replicated on a single GeForce RTX 4090 GPU using only the front camera.)

Methods	Modality	NDS	mAP	mATE	mASE	mAOE	mAVE	mAAE
CenterFusion	Camera & Radar	0.453	0.332	0.649	0.263	0.535	0.54	0.142
CenterPoint v2	Camera & LiDAR & Radar	0.714	0.671	0.249	0.236	0.350	0.250	0.136
PointPillars (Light)	LiDAR	0.453	0.305	0.517	0.290	0.500	0.316	0.368
LRM0	Camera	0.371	0.294	0.752	0.265	0.603	1.582	0.14
MonoDIS [28]	Camera	0.384	0.304	0.738	0.263	0.546	1.553	0.134
CenterNet [29]	Camera	0.4	0.338	0.658	0.255	0.629	1.629	0.142
Noah CV Lab	Camera	0.418	0.331	0.66	0.262	0.354	1.663	0.198
FCOS3D [30]	Camera	0.428	0.358	0.69	0.249	0.452	1.434	0.124
PGD [31]	Camera	0.448	0.386	0.626	0.245	0.451	1.509	0.127
BEVDet4D-Mono	Camera	0.483	0.384	0.514	0.173	0.302	1.482	0.105
MonoMPV	Camera	**0.522**	**0.415**	0.429	0.144	0.217	0.962	0.108

3) EVALUATION METRICS

This study presented standardized metrics for evaluating 3D object detection, including mAP, Average Velocity Error (AVE), Average Scale Error (ASE), NDS, Average Translation Error (ATE), Average Attribute Error (AAE), and Average Orientation Error (AOE). The nuScenes metrics encompass five True Positive (TP) metrics specifically designed to assess errors in translation, scale, orientation, velocity, and attributes. Specifically, this study focuses on two of the most commonly employed metrics for quantitative assessment:

- mAP: $\text{mAP} = \frac{1}{|C||D|} \sum_{c \in C} \sum_{d \in D} \text{AP}_{c,d}$.
  The mAP metric in 3D object detection is analogous to that used in 2D object detection, providing a measure of precision and recall. The mAP is calculated based on the center distance on the ground plane rather than the 3D Intersection over Union (IoU). This approach ensures that the prediction results are aligned more accurately with the ground truth. Here, c represents the number of categories and d denotes 2D center distance on the ground plane.
- NDS: $\text{NDS} = \frac{1}{10}[5\text{mAP} + \sum_{\text{mTP} \in \text{TP}}(1 - \min(1, \text{mTP}))]$.
  The NDS offers a comprehensive assessment by integrating various detection quality indicators. The NDS metric is segmented into two components: one focusing on assessesing detection performance, specifically mAP, and another that evaluates detection quality based on metrics related to location, size, orientation, properties, and speed measurements (ATE, ASE, AOE, AVE, and Average Acceleration Error (AAE)). Given that mAVE, mAOE, and mATE have the potential to exceed 1, these metrics are normalized to a range of 0 to 1 to maintain consistency in evaluation.

B. PERFORMANCE COMPARISON

This study presents both quantitative and qualitative results and conduct a comprehensive ablation study on the crucial factors that contribute to the state-of-the-art performance of the proposed method.

Table 2 presents results from the nusense dataset. A quantitative analysis was performed on the authoritative benchmark dataset for 3D object detection, nuScenes. The proposed method demonstrates superior performance by achieving state-of-the-art results on both NDS and mAP evaluation metrics.

To enhance the intuitive validation of this method's effectiveness, qualitative results on the nuScenes dataset are presented in Figure 7.

C. DEPLOYMENT IN EDGE DEVICE

Existing monocular 3D object detection algorithms are exclusively trained and validated on training devices. However, these training devices cannot be implemented in real vehicles because of their high-power consumption and bulky size. Regrettably, when compared to training devices, the edge devices commonly utilized in vehicles exhibit lower power consumption and limited computational capabilities. Therefore, when implementing these devices on embedded systems frequently utilized in autonomous vehicles, it is essential to consider the constraints of computational resources and latency. The proposed method was optimized and implemented utilizing the mmdeploy framework on Jetson Orin NX embedded device. As a result, Jetson Orin NX was able to perform inference in 183ms, demonstrating the practical applicability of the proposed network.

During deployment, the ONNX model is imported into TensorRT for simplification and FP16 acceleration. Network simplification merges Conv, BN, and ReLU layers into CBR layers. As shown in Figure 9, for Inception subnetwork, the

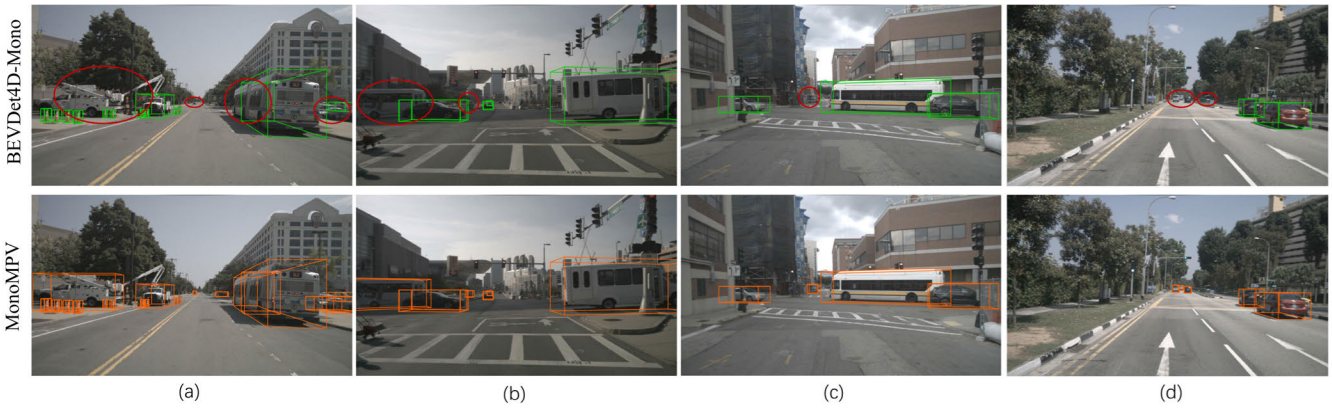


FIGURE 7. Visualization of 3D object detection results. (Note: Monocular 3D object detection based on BEV is represented by the green rectangle (top), while MonoMPV is represented by the orange rectangle (bottom). Detection differences are emphasized with red circular boxes.)

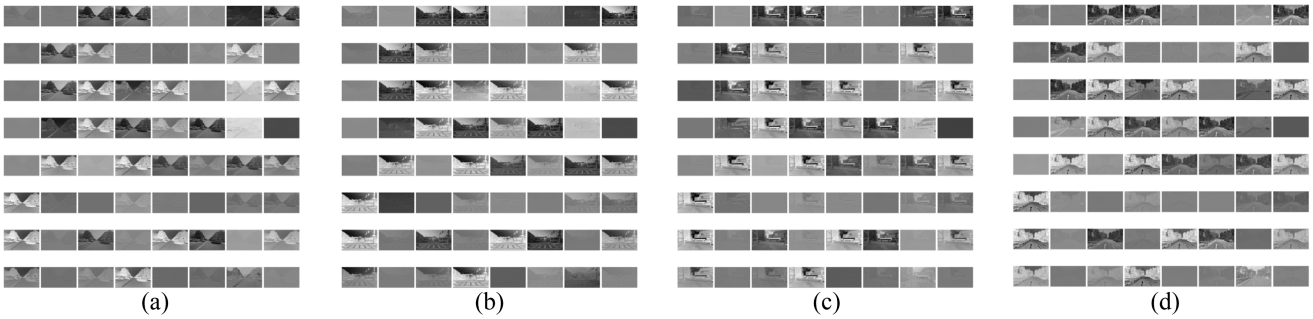


FIGURE 8. Visualization of the grayscale map of model weights. During inference, we selected the first layer of a 64-channel feature extraction network for visualizing grayscale images across the scenes in Figure 7. By analyzing the grayscale images of features learned by the model through specific channels, we can effectively analyze the model's learning of abstract features (e.g., edges, textures) at different channel levels.

original architecture is transformed into Figure 9(b) via TensorRT simplification. For the horizontal combination portion, further optimization fuses input layers with the same input. The specific process from Figure 9(a) to 8(b) to 8(c) is: first, TensorRT simplifies the computational process by integrating the network's Conv, BN, and ReLU layers into a single CBR layer. Then, the system directly concatenates inputs to subsequent operations, skipping the additional concatenation step, which reduces system throughput and enhances processing efficiency. With TensorRT's support, FP16 acceleration technology reduces data precision from 32-bit to 16-bit floating-point numbers, improving calculation and computation efficiency.

To elucidate the effect of deploying edge devices in real vehicles more intuitively, the qualitative results is illustrated in the application setting in Figure 10. The proposed approach involves utilizing the nuScenes dataset for training and integrating the manually adjusted camera intrinsics. Subsequently, the trained weights are implemented in real-world scenarios. In Figure 10, the detection accuracy of targets in close proximity is high. However, there is still potential for enhancing the predictive accuracy of small targets located at a distant range. It is important to note that motorcycles and tricycles, as two types of vehicles, were not part of the nuScenes dataset gathered from overseas. Consequently, they were not accurately identified in Figure 10 (c) depicting the real-life scenario. However, as a result of variations in the training and testing conditions, this approach demonstrates competitive results in the surrounding area, underscoring its capacity for generalization within a specific setting.

V. DISCUSSION

In monocular 3D object detection, a significant challenge associated with BEV representations is the absence of height information, which limits the accurate representation of 3D environment. Consequently, this problem hinders the overall enhancement of detection performance. Building upon current monocular 3D object detection algorithms that utilize BEV representations, the application of BEV is expanded to provide a more detailed depiction of 3D structures, thereby improving the precision of object detection. Furthermore, the introduction of TLTF enhances the model's ability to optimize data utilization by offering a different optimization objective.

In Table 2, a compilation of early monocular methods that depend on supplementary data is presented, alongside image-only methods that have recently demonstrated significant results on the nuScenes dataset. Upon validation,

TABLE 3. Average precision of each class on nuScenes benchmark. (Note: CV and TC represent Construction Vehicle and Traffic Cone, respectively, which are abbreviations in the table.)

Methods	car	truck	bus	trailer	CV	ped	motor	bicycle	TC	barrier	mAP
LRM0	0.467	0.21	0.17	0.149	0.061	0.359	0.287	0.246	0.476	0.512	0.294
MonoDIS	0.478	0.22	0.188	0.176	0.074	0.37	0.29	0.245	0.487	0.511	0.304
CenterNet	0.536	0.27	0.248	0.251	0.086	0.375	0.291	0.207	0.583	0.533	0.338
Noah CV Lab	0.515	0.278	0.249	0.213	0.066	0.404	0.338	0.237	0.522	0.49	0.331
FCOS3D	0.524	0.27	0.277	0.255	0.117	0.397	0.345	0.298	0.557	0.538	0.358
PGD	0.561	0.299	0.285	0.266	0.134	0.441	0.397	0.314	0.605	0.561	0.386
MonoMPV(Ours)	0.617	0.311	0.319	0.278	0.182	0.474	0.416	0.308	0.653	0.592	0.415

TABLE 4. Ablation study on nuScenes validation 3D detection dataset: (a) Baseline model (Add only 3D object detection branch); (b) MPVFormer module (Add only MPV); (c) TLTF module (Add only TLTF); (d) Add both MPV module and TLTF module.

Ablation	NDS	mAP	mATE	mASE	mAOE	mAVE	mAAE
a	0.483	0.384	0.514	0.173	0.302	1.482	0.105
b	0.504	0.411	0.469	0.182	0.265	1.435	0.098
c	0.492	0.396	0.483	0.190	0.282	1.241	0.103
d	0.522	0.415	0.429	0.144	0.217	0.962	0.108

this study initially compared all methods that utilize RGB images as input data. The results indicated that the proposed MonoMPV method demonstrated superior performance compared to the other methods. The proposed approach yielded a mAP of 41.5% and a NDS of 52.2%. It is noteworthy that the proposed model demonstrated superior performance compared to the best previous methods by 2.9% in the mAP measurement. The study also included the results of benchmark testing conducted with alternative data models. The tests were carried out utilizing CenterFusion algorithm based on RGB images and radar data, the real-time LiDAR sensors' PointPillars approach, and the CenterPoint method that amalgamates data from all sensors. Significantly, MonoMPV outperforms both PointPillars and CenterFusion in both mAP and NDS. This suggests that when provided with an adequate amount of data, MonoMPV effectively addresses the depth ambiguity problem associated with utilizing only a single RGB image. Unfortunately, a disparity exists between the proposed approach and the High-Performance CenterPoint. This disparity can be ascribed to the inherent limitations of the dataset, as the proposed method relies only on a single RGB image, making it challenging to rival methodologies that leverage Camera, LiDAR, and Radar simultaneously. Meanwhile, employing alternative modal data techniques generally leads to enhanced NDS performance. The rationale behind this phenomenon is rooted in the utilization of methods that forecast the velocity of object movement by analyzing continuous multi-frame point clouds or radar velocity measurements, consequently resulting in a reduced value for mAVE. In contrast to the proposed model, MonoMPV employs a single frame of RGB images. However, MonoMPV has shown competitive performance in situations where RGB images are the only input data. This emphasizes the potential efficacy in specific circumstances.

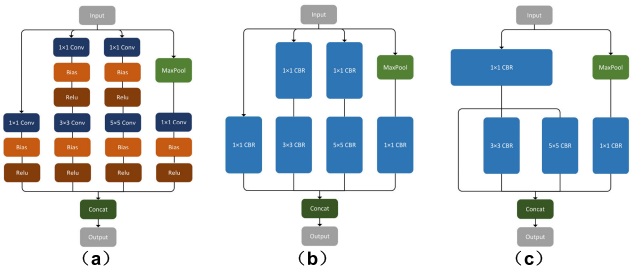


FIGURE 9. Principle of model simplification in TensorRT.

To evaluate the efficacy of different proposed approaches, an ablation study was conducted in Table 4. The experimental results suggest that the lack of these modules results in a decline in performance when gradually integrating MPVFormer and TLTF into the baseline model. This statement validates the significance of their role in improving overall performance. Retention of these modules, in comparison to their removal, can substantially improve performance, thereby reinforcing the efficacy of the proposed approach. The ablation experiment conducted on MPVFormer module is presented in Table 4(b). The ablation experiment conducted on MPVFormer in comparison to the baseline model resulted in the following results: an increase of 2.1% in NDS and 2.7% in mAP. MonoMPV surpassed the baseline in both NDS and mAP metrics, showcasing its capacity to illustrate more complex 3D structural representations, consequently improving the precision of object detection. Table 4(c) demonstrates that the model incorporating only TLTF module exhibits superior performance compared to the baseline model. Thanks to the clarity established by TLTF module as the optimization target for the model, the model can enhance its data utilization efficiency. This validates the capability of the introduced TLTF module to maintain optimization for achieving the anticipated

FIGURE 10. Visualization of 3D object detection on real vehicles with edge device. (Note: The orange box represents the results of 3D object detection.

objective, thereby enhancing the detection of monocular 3D targets. While optimizing the loss function in TLTF module seems to have minimal effect on enhancing the accuracy of object detection, the inclusion of both modules substantially improves the overall accuracy. MPVFormer demonstrates significant advantages in representing complex 3D structural models. TLTF module focuses on highlighting the significant regions of feature maps by consolidating diverse location information on the top plane, thereby offering a precise target for model optimization.

Figure 7 illustrates the qualitative results of proposed method on the nuScenes dataset. Current monocular 3D object detection techniques relying on BEV face challenges in effectively identifying distant small objects and objects obstructed by other entities. Furthermore, these methods are susceptible to misclassification errors caused by environmental interferences. In these challenging scenarios, MonoMPV demonstrates the capability to accurately predict distant small objects and occluded objects, suggesting that the proposed approach is proficient in capturing detailed 3D structures and exhibits robust overall perception. Regrettably, the long-distance trucks (Figure 7(d)) were not identified. This variation can be ascribed to the disparity in the number of training samples between trucks and cars in the nuScenes dataset. The restricted amount of training data presents challenges to the network's capacity to precisely recognize these categories, resulting in performance fluctuations, a prevalent issue in many monocular 3D object detection systems. Therefore, a crucial research focus involves exploring the utilization of data augmentation methods (e.g., diffusion models) to enhance the volume of data in 3D object detection datasets with limited samples.

VI. CONCLUSION

This paper introduces a monocular 3D object detection framework (MonoMPV) designed to accurately map spatial objects onto MPV. In this regard, cross-attention is utilized to map image features onto MPV and to enable information exchange among these views. Therefore, the monocular 3D object detection is framed as the task of detailed 3D scene description. Furthermore, Triplet Loss is utilized to optimize the model's target, guiding the model towards the direction of optimization. By engaging in this process, this study develops a model that can effectively depict the complex details of 3D environments, thereby enhancing the precision of monocular 3D object detection. The efficacy of the proposed method was assessed using the nuScenes benchmark dataset, showcasing enhanced performance in contrast to established vehicle detection techniques. The network is capable of performing inference on Jetson Orin NX edge device with a latency of 183ms, thereby guaranteeing its applicability in real-world scenarios. This study did not investigate the utilization of diffusion models in the limited class samples of an augmented training dataset. Future research endeavors will concentrate on enhancing performance in this area.

ACKNOWLEDGMENT

The authors would like to thank the reviewers for their corrections and helpful suggestions.

REFERENCES

[1] X. Zhu, H. Zhou, T. Wang, F. Hong, Y. Ma, W. Li, H. Li, and D. Lin, "Cylindrical and asymmetrical 3D convolution networks for LiDAR segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 9934–9943.

[2] Z. Wang, Z. Huang, Y. Gao, N. Wang, and S. Liu, "MV2DFusion: Leveraging modality-specific object semantics for multi-modal 3D detection," 2024, *arXiv:2408.05945*.

[3] G. Zhang, L. Fan, C. He, Z. Lei, Z. Zhang, and L. Zhang, "Voxel Mamba: Group-free state space models for point cloud based 3D object detection," 2024, *arXiv:2406.10700*.

[4] P. Li, X. Chen, and S. Shen, "Stereo R-CNN based 3D object detection for autonomous driving," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 7636–7644.

[5] W. Peng, H. Pan, H. Liu, and Y. Sun, "IDA-3D: Instance-depth-aware 3D object detection from stereo vision for autonomous driving," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 13012–13021.

[6] Y. Wei, L. Zhao, W. Zheng, Z. Zhu, Y. Rao, G. Huang, J. Lu, and J. Zhou, "SurroundDepth: Entangling surrounding views for self-supervised multi-camera depth estimation," in *Proc. Conf. Robot Learn.*, Mar. 2023, pp. 539–549.

[7] Y. Zhang, Z. Zhu, W. Zheng, J. Huang, G. Huang, J. Zhou, and J. Lu, "BEVerse: Unified perception and prediction in birds-eye-view for vision-centric autonomous driving," 2022, *arXiv:2205.09743*.

[8] C. Choy, J. Gwak, and S. Savarese, "4D spatio-temporal ConvNets: Minkowski convolutional neural networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 3070–3079.
[9] T. Liang, H. Xie, K. Yu, Z. Xia, Z. Lin, Y. Wang, T. Tang, B. Wang, and Z. Tang, "BEVFusion: A simple and robust LiDAR-camera fusion framework," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 35, 2022, pp. 10421–10434.
[10] Z. Li, W. Wang, H. Li, E. Xie, C. Sima, T. Lu, Y. Qiao, and J. Dai, "BEVFormer: Learning bird's-eye-view representation from multi-camera images via spatiotemporal transformers," in *Proc. Eur. Conf. Comput. Vis.*, Oct. 2022, pp. 1–18.
[11] J. Huang and G. Huang, "BEVDet4D: Exploit temporal cues in multi-camera 3D object detection," 2022, *arXiv:2203.17054*.
[12] L. Chambon, E. Zablocki, M. Chen, F. Bartoccioni, P. Pérez, and M. Cord, "PointBeV: A sparse approach for BeV predictions," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2024, pp. 15195–15204.
[13] Y. Li, Y. Chen, X. Qi, Z. Li, J. Sun, and J. Jia, "Unifying voxel-based representation with transformer for 3D object detection," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 35, 2022, pp. 18442–18455.
[14] Y. B. Can, A. Liniger, D. P. Paudel, and L. Van Gool, "Structured bird's-eye-view traffic scene understanding from onboard images," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 15641–15650.
[15] Y. Wang, V. C. Guizilini, T. Zhang, Y. Wang, H. Zhao, and J. Solomon, "DETR3D: 3D object detection from multi-view images via 3D-to-2D queries," in *Proc. Conf. Robot Learn.*, Jan. 2022, pp. 180–191.
[16] J. Philion and S. Fidler, "Lift, splat, shoot: Encoding images from arbitrary camera rigs by implicitly unprojecting to 3D," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 194–210.
[17] L. Piccinelli, Y.-H. Yang, C. Sakaridis, M. Segu, S. Li, L. Van Gool, and F. Yu, "UniDepth: Universal monocular metric depth estimation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2024, pp. 10106–10116.
[18] L. Yang, K. Yu, T. Tang, J. Li, K. Yuan, L. Wang, X. Zhang, and P. Chen, "BEVHeight: A robust framework for vision-based roadside 3D object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 21611–21620.
[19] Y. Li, Z. Ge, G. Yu, J. Yang, Z. Wang, Y. Shi, J. Sun, and Z. Li, "BEVDepth: Acquisition of reliable depth for multi-view 3D object detection," in *Proc. AAAI Conf. Artif. Intell.*, Jun. 2023, vol. 37, no. 2, pp. 1477–1485.
[20] Y. Li, H. Bao, Z. Ge, J. Yang, J. Sun, and Z. Li, "BEVStereo: Enhancing depth estimation in multi-view 3D object detection with temporal stereo," in *Proc. AAAI Conf. Artif. Intell.*, Jun. 2023, vol. 37, no. 2, pp. 1486–1494.
[21] B. Fei, J. Xu, R. Zhang, Q. Zhou, W. Yang, and Y. He, "3D Gaussian splatting as new era: A survey," 2024, *arXiv:2402.07181*.
[22] Y. Huang, W. Zheng, Y. Zhang, J. Zhou, and J. Lu, "Tri-perspective view for vision-based 3D semantic occupancy prediction," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 9223–9232.
[23] S. Wu, T. Jakab, C. Rupprecht, and A. Vedaldi, "DOVE: Learning deformable 3D objects by watching videos," *Int. J. Comput. Vis.*, vol. 131, no. 10, pp. 2623–2634, Oct. 2023.
[24] R. Nabati and H. Qi, "CenterFusion: Center-based radar and camera fusion for 3D object detection," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2021, pp. 1526–1535.
[25] T. Yin, X. Zhou, and P. Krähenbühl, "Center-based 3D object detection and tracking," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 11779–11788.
[26] D. H. Kim, F. Anvarov, J. M. Lee, and B. C. Song, "Metric-based attention feature learning for video action recognition," *IEEE Access*, vol. 9, pp. 39218–39228, 2021.
[27] A. H. Lang, S. Vora, H. Caesar, L. Zhou, J. Yang, and O. Beijbom, "PointPillars: Fast encoders for object detection from point clouds," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 12689–12697.
[28] A. Simonelli, S. R. Bulo, L. Porzi, M. Lopez-Antequera, and P. Kontschieder, "Disentangling monocular 3D object detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 1991–1999.
[29] X. Zhou, D. Wang, and P. Krähenbühl, "Objects as points," 2019, *arXiv:1904.07850*.
[30] T. Wang, X. Zhu, J. Pang, and D. Lin, "FCOS3D: Fully convolutional one-stage monocular 3D object detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshops (ICCVW)*, Oct. 2021, pp. 913–922.
[31] T. Wang, Z. H. U. Xinge, J. Pang, and D. Lin, "Probabilistic and geometric depth: Detecting objects in perspective," in *Proc. Conf. Robot Learn.*, 2022, pp. 1475–1485.

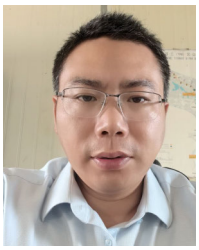
ZHAOXUE DENG was born in Shandong, China, in 1985. He received the Ph.D. degree in automotive engineering from Chongqing University, Chongqing, China, in 2015. He is currently an Associate Professor with the Automotive Engineering Department, Chongqing Jiaotong University, Chongqing. His research interests include semi-active/active suspension systems and smart vehicle dynamic control.

BINGSEN HAO was born in Hebei, China, in 1999. He received the B.S. degree from Lanzhou University of Technology, in 2022. He is currently pursuing the master's degree with Chongqing Jiaotong University. His research interests include autonomous driving 3D object detection and the deployment of deep learning.

GUOFANG LIU was born in Shanxi, China, in 1984. She received the master's degree in automotive engineering from Beihang University, Beijing, China, in 2011. She is currently the Vice Director of the Automotive Electrification Research Center, China Society of Automotive Engineers. Her research interests include new energy automobile industry and technology route research.

XINGQUAN LI (Fellow, IEEE) was born in Guizhou, China, in 1981. He received the B.S. degree in vehicle engineering from Chongqing Jiaotong University, in 2004, and the Ph.D. degree in vehicle engineering from Chongqing University, in 2012. He joined Chongqing Changan Automobile Company Ltd., in 2016. His main research interests include vehicle NVH control, intelligent vehicle automatic driving behavior, and cooperative control.

HANBING WEI was born in China, in 1979. He is currently a Professor and a Ph.D. Supervisor with Chongqing Jiaotong University, the Director of the Engineering Center, Chongqing Jiaotong University, and a Visiting Scholar with The University of Queensland, Australia, and Kettering University, USA. His research interests include smart vehicle perception, decision, and control.



FEI HUANG received the bachelor's degree from Changsha University of Technology, China, in 2003. From 2006 to 2013, he participated in the construction of bridge and highway projects in China. He is currently a Senior Engineer in civil engineering and also the Project Manager of the Dakar BRT Project, China Road and Bridge Corporation. His research interests include road engineering and smart transportation.

SHENGSHU LIU received the bachelor's degree in civil engineering from Wuhan University of Technology and the master's degree in structural engineering from Tongji University. He is currently the Deputy Project Manager of the Dakar Rapid Transit Project, China Road and Bridge Corporation. He is mainly engaged in project lifecycle construction and management, BIM research, practical application of special structural engineering, civil engineering design, and other related work.