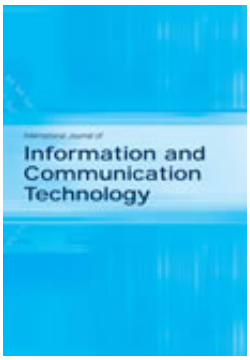
INDERSCIENCE PUBLISHERS
Linking academia, business and industry through research



International Journal of Information and Communication Technology

ISSN online: 1741-8070 - ISSN print: 1466-6642
https://www.inderscience.com/ijict

3D image reconstruction using an improved BEV model and global convolutional attention fusion

HuaShun Yan, XiaoJie Li, ZeLin Mou

Article History:

Received:	03 December 2024
Last revised:	08 January 2025
Accepted:	08 January 2025
Published online:	31 March 2025

3D image reconstruction using an improved BEV model and global convolutional attention fusion

HuaShun Yan*, XiaoJie Li and ZeLin Mou

Computer Science and Technology,
Chengdu University of Information Technology,
Chengdu, Sichuan, 610225, China
Email: 18788936063@163.com
Email: lixj@cuit.edu.cn
Email: 1084433641@qq.com
*Corresponding author

Abstract: In autonomous driving and computer vision, 3D object detection plays a critical role but faces challenges related to the effective extraction and integration of multi-view features. The existing BEVFormer model, which uses CNNs to convert images into a bird's-eye view (BEV), shows potential but struggles to capture fine-grained details and multi-scale information, especially in high-resolution, complex scenes. To address these limitations, we propose the MultiCAN-DEBEV model, which integrates the MSF-DySample, GCAF, and MSDE modules. These modules improve the handling of multi-scale features, enhance feature expressiveness, and strengthen detail representation. Experiments on the nuScenes dataset show significant performance improvements, and the modular design ensures broad adaptability to other 3D detection models.

Keywords: computer vision; 3D object detection; BEV object detection; autonomous driving.

Reference to this paper should be made as follows: Yan, H., Li, X. and Mou, Z. (2025) '3D image reconstruction using an improved BEV model and global convolutional attention fusion', *Int. J. Information and Communication Technology*, Vol. 26, No. 6, pp.98–116.

Biographical notes: HuaShun Yan received his BS degree in IoT from Chengdu University of Information Technology, City, China, in 2023. He is currently working toward his Master degree in Computer Technology with the School of Chengdu University of Information Technology, City, China. His research interests include computer vision, 3D object detection; and autonomous driving.

XiaoJie Li earned her doctorate in Computer Science and Technology from Sichuan University in June 2015. She joined the School of Computer Science at Chengdu University of Information Technology in July 2015. In 2017, she was a visiting scholar at the Image Science Institute of Vanderbilt University (VUIIS) in the United States. Her main research interests include machine learning and image processing.

ZeLin Mou received his BS degree in Computer Science and Technology from Aba Teachers University, Sichuan, China, in 2022. He is currently working toward his Master degree in Computer Technology with the School of Chengdu University of Information Technology, City, China. His research interests include object detection and neural networks.

1 Introduction

3D object detection is a pivotal technology in the realms of autonomous driving and computer vision. It is instrumental in recognising and localising diverse objects, including vehicles and pedestrians, within intricate traffic environments, thereby enhancing the safety and efficiency of autonomous driving systems (Wu et al., 2021). Accurate detection in complex urban settings necessitates the extraction of robust features from multi-angle images and their integration to produce precise 3D positional and configurational information (Chen, 2024).

Investigations into multi-view 3D object detection predominantly rely on deep learning technologies, particularly the utilisation of convolutional neural networks (CNNs) and attention mechanisms The BEVFormer model is a notable technique widely used for generating target detection from a bird's-eye view (BEV), enabling accurate predictions (Li et al., 2022). This model integrates data from diverse sensors, including cameras positioned in multiple directions and LiDAR, effectively addressing projection issues from 2D to 3D while enhancing detection accuracy and efficiency (Li et al., 2023).

Despite the advancements achieved by the BEVFormer in feature extraction and integration, challenges persist in complex scenarios, especially when addressing small or distant targets. Issues such as the loss of vertical structural details and information regarding occluded objects may still arise. Specifically, these models exhibit limitations in integrating multi-scale feature information, capturing contextual information, and managing detailed features. Optimising these issues, especially in leveraging detailed information from high-resolution images, is crucial for detecting small and distant targets in autonomous driving scenarios within complex environments (Chen et al., 2022).

To tackle these challenges, this study introduces an enhanced BEVFormer model, designated as multi-scale convolutional attention network with detail enhancement for BEV (MultiCAN-DEBEV). This model introduces a multi-scale feature fusion dynamic upsampling (MSF-DySample) module that adaptively adjusts the upsampling strategy based on varying input features, thereby enhancing the model's capability to manage inputs of different scales and improving its handling of multi-scale features. The global convolutional attention fusion (GCAF) module enhances the capture of global information, thereby improving the model's ability to infer the potential locations and shapes of occluded objects through enhanced contextual information acquisition. The multi-scale detail enhancement (MSDE) module effectively extracts and enhances detailed features in images by applying deformable and dilated convolutions. The integration of these modules not only strengthens the model's capacity to address the challenges of lost vertical structural details and occluded object information in complex scenarios but also enhances the performance of downstream tasks such as 3D object detection. Furthermore, due to its modular nature, these components can be easily applied to other similar tasks. Experimental results on the nuScenes dataset indicate that the improved MultiCAN-DEBEV model outperforms the baseline model.

In conclusion, the proposed MultiCAN-DEBEV model not only demonstrates theoretical innovation but also exhibits improved performance in practical applications, thereby offering substantial technical support and a theoretical foundation for the future advancement of autonomous driving technology and associated computer vision tasks.

2 Related work

3D object detection holds a pivotal role in domains such as autonomous driving and robotic vision. Relevant studies can be primarily classified into three categories: traditional methods based on LiDAR, multi-view imaging methods based on cameras, and advanced technologies that integrate CNNs with attention mechanisms (Yin et al., 2021; Liu et al., 2021).

- *Challenges in multi-view image fusion:* Multi-view image fusion and dynamic upsampling play a crucial role in autonomous driving applications. The BEVFormer model enhances the performance of three-dimensional object detection by converting multi-view images into a BEV. Nonetheless, existing methodologies often face substantial computational costs and restricted feature expression capabilities when tackling multi-view image fusion and dynamic upsampling. In recent years, dynamic upsampling technologies have attracted significant attention. For example, CARAFE facilitates efficient upsampling by learning to generate dynamic convolutional kernels, whereas FADE and SAPA integrate attention mechanisms to further improve the performance of dynamic upsampling (Wang et al., 2019; Zhang et al., 2020; Liu et al., 2020). Although these methods have advanced in achieving efficient upsampling, they still incur substantial computational costs. To tackle this issue, we propose the MSF-DySample module, which integrates multi-scale feature fusion and dynamic upsampling techniques, utilising a dual-branch architecture to concurrently capture high-frequency and low-frequency information, thereby enhancing feature expression capabilities.
- *Enhancing feature fusion with CNNs and attention:* The integration of CNNs and attention mechanisms within the BEVFormer model has garnered significant attention (Zhong and Hu, 2022). While CNNs have demonstrated considerable success in various tasks, including image classification, object detection, and image segmentation, they are inherently limited to local receptive fields, which hampers their ability to effectively capture global contextual information and long-range dependencies. To mitigate this challenge, researchers have increasingly adopted attention mechanisms that dynamically adjust feature weights to better capture long-range dependencies. The BEVFormer model incorporates the Transformer architecture, utilising self-attention mechanisms to enhance the feature fusion capabilities of multi-view images (Li et al., 2022). However, despite these advancements, BEVFormer still exhibits limitations in capturing detailed features and multi-scale information. To address these shortcomings, we propose the GCAF module, which integrates convolutional and attention mechanisms to improve feature expression, thereby facilitating the concurrent capture of both local and global features.
- *Limitations of traditional LiDAR methods:* Traditional 3D object detection methods predominantly rely on LiDAR data, such as PointNet++ and VoxelNet (Qi et al., 2017; Zhou and Tuzel, 2018). These methods facilitate 3D object detection through the processing of point cloud data. However, the acquisition cost of LiDAR data is substantial, and its performance is suboptimal under adverse weather conditions. With technological advancements, camera-based 3D object detection methods have increasingly garnered attention. These methods utilise multi-view images to generate

BEVs, exemplified by the BEVFormer model, which transforms multi-view images into BEVs and integrates CNNs for feature extraction and fusion, achieving notable results. Although BEVFormer excels in 3D object detection tasks, it still exhibits limitations in capturing detailed features and multi-scale information. Therefore, this paper proposes an enhanced method by introducing a MSDE to address these challenges.

- *Limitations of traditional LiDAR methods:* 3D object detection methods predominantly rely on LiDAR data, exemplified by techniques such as PointNet++ and VoxelNet (Qi et al., 2017; Zhou and Tuzel, 2018). These methods facilitate 3D object detection through the processing of point cloud data. However, the acquisition costs associated with LiDAR data are substantial, and its performance can be suboptimal under adverse weather conditions. With advancements in technology, camera-based 3D object detection methods have garnered increasing attention. These methods utilise multi-view images to generate BEVs, as demonstrated by the BEVFormer model, which transforms multi-view images into BEVs and integrates CNNs for feature extraction and fusion, achieving notable results. Although the BEVFormer model excels in 3D object detection tasks, it still exhibits limitations in capturing detailed features and multi-scale information. To address these challenges, this paper proposes an enhanced approach by introducing a MSDE.

In summary, while existing 3D object detection methods have achieved significant advancements across various dimensions, they still demonstrate limitations in capturing detailed features and processing multi-scale information. As the demands for detection accuracy and real-time performance in practical applications, such as autonomous driving and intelligent monitoring, continue to escalate, there is a pressing need for further optimisation and enhancement of these methods. Consequently, the proposed MultiCAN-DEBEV model aims to enhance the performance of 3D object detection by integrating several advanced modules designed to meet the requirements of practical applications.

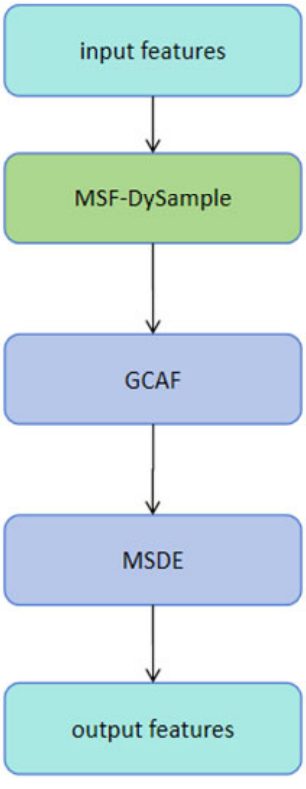
3 MultiCAN-DEBEV

3.1 Model design overview

This paper presents the MultiCAN-DEBEV model, enhancing the accuracy of BEV representation and improving the performance of 3D object detection by incorporating a MSF-DySample module, a GCAF module, and a MSDE based on the BEVFormer model.

First, the input multimodal data undergoes upsampling through the MSF-DySample module to retain more detailed information. Next, the enhanced feature maps are further refined by the GCAF module and aggregated through the MSDE module. This process ultimately generates feature maps for 3D object detection.

Figure 1 Framework diagram of the MultiCAN-DEBEV model (see online version for colours)

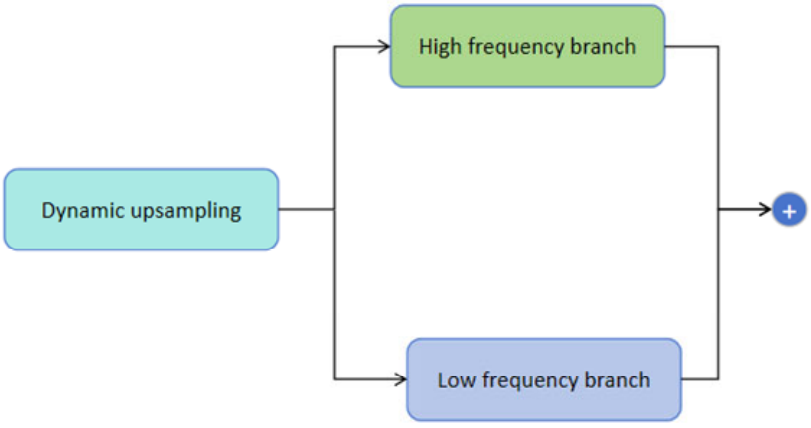


Notes: This model enhances the BEVFormer by incorporating the MSF-DySample module, the GCAF module, and the multi-scale detail enhancement (MSDE) module, aiming to improve the accuracy of BEV representation and the performance of 3D object detection.

3.2 MSF-DySample module

The MSF-DySample module introduces a convolution operator termed AttnConv, which employs an attention-based approach and capitalises on the benefits of shared weights and context-aware weights for local perception. Specifically, the MSF-DySample module is characterised by a dual-branch architecture, wherein one branch utilises AttnConv to capture high-frequency information, while the other branch employs vanilla attention and downsampling techniques to capture low-frequency information (Chen et al., 2021; Vaswani et al., 2017). This innovative dual-branch design enables the MSF-DySample module to concurrently capture both high-frequency and low-frequency information within a unified framework, thereby facilitating the effective fusion of multi-scale features. Consequently, this enhancement significantly improves the feature representation capability of the model and bolsters its performance in complex scenarios.

Figure 2 Overall framework diagram of the MSF-DySample module (see online version for colours)



Notes: This module (MSF-DySample) aims to enhance feature representation by combining dynamic upsampling techniques with multi-scale feature fusion.

3.2.1 Implementation of dynamic sampling

To dynamically adjust the position of each pixel based on the content of the input feature map, an initial convolution layer is used to generate offsets.

$$(O):[O=\mathrm{Conv}(X)] \tag{1}$$

where $(O\in R^{B\times 2GS^2\times H\times W})$.

Subsequently, to provide each pixel with an initial offset, ensuring that each pixel has a reasonable starting point during the upsampling process, the initial position matrix is computed.

$$\begin{aligned}&(I):\left[Ih=\left[\frac{-S+1}{2},\frac{-S+3}{2},\ldots,\frac{S-2}{2}\right]\Big/S\right]\left[I_w=\left[\frac{-S+1}{2},\frac{-S+3}{2},\ldots,\frac{S-1}{2}\right]\Big/S\right]\\&[I=Ih\times Iw][I=I\times G][I=I\cdot\mathrm{reshape}(1,-1,1,1)]\end{aligned} \tag{2}$$

where Ih and Iw are the vertical (height) and horizontal (width) offsets, respectively.

Simultaneously, to ensure that each pixel's position is mapped within the range [–1, 1], normalised coordinates are calculated. The pixels are also rearranged for interpolation using the GridSample function:

$$\begin{aligned}&(N):[N_h=[0.5,1.5,\ldots,H-0.5]][N_w=[0.5,1.5,\ldots,W-0.5]]\\&[N=N_h\times N_w]\\&[N=N\cdot\mathrm{unsqueeze}(1)\cdot\mathrm{unsqueeze}(0)\cdot\mathrm{type}(X\cdot\mathrm{dtype})\cdot\mathrm{to}(X\cdot\mathrm{device})]\\&[N=2\cdot(N+O')/T-1]\end{aligned} \tag{3}$$

$$\begin{aligned}&[N'=N\cdot\mathrm{view}(B,2,-1,H,W)]\\&[N'=N'\cdot\mathrm{permute}(0,2,3,4,1)\cdot\mathrm{contiguou}()\cdot\mathrm{flatten}(0,1)]\end{aligned} \tag{4}$$

where N_h = [0.5, 1.5, …, H – 0.5] and N_w = [0.5, 1.5, …, W – 0.5]represent the position of each pixel in the feature map. T = [W, H] used to normalise coordinates into the range [–1, 1].

Finally, the GridSample function is used to perform bilinear interpolation on the input feature map based on the calculated coordinates to retain more detail features.

$$\begin{aligned}&\left[\mathrm{Y}=\mathrm{GridSample}\left(\mathrm{X}\cdot\mathrm{reshape}(\mathrm{B}\times\mathrm{G},-1,\mathrm{H},\mathrm{W}),\mathrm{N}',\right.\right.\\&\left.\left.\mathrm{mode}=' \mathrm{bilinear}',\mathrm{align\_corners}=\mathrm{False},\mathrm{padding\_mode}=\text{"border"}\right)\right]\\&\left[\mathrm{Y}=\mathrm{Y}\cdot\mathrm{view}(\mathrm{B},-1,\mathrm{SH},\mathrm{SW})\right]\end{aligned}\tag{5}$$

3.2.2 Implementation of dual-branch high and low frequency feature fusion

For the high-frequency branch, the model is designed with small convolutional kernels and attention mechanisms to capture detail-rich high-frequency features, enhancing the model's spatial resolution understanding. The process is as follows:

Initially, feature mappings are generated:Fqkv:

$$\left[\mathrm{Fqkv}=\mathrm{Conv}3\times3^{\mathrm{m}\times\mathrm{d}}(\mathrm{X})\mathrm{with}\ \mathrm{m}=\frac{\mathrm{num\_heads}}{\sum\mathrm{i}^{\mathrm{n}}\mathrm{g}_{\mathrm{i}}},\mathrm{d}=\frac{\mathrm{dim}}{\mathrm{num\_heads}}\right]\tag{6}$$

where (X) is the input feature map, m is the number of heads within each attention group, d is the dimension of each head, num_heads is the total number of heads, dim is the channel number of the input features, ($\mathrm{Conv}_{3\times3}$) represents convolutional operations using 3 × 3 kernels, and g_i indicates the number of heads in the i group.

Subsequently, the feature's expressiveness is enhanced using the AttnMap activation function.

$$\left[\mathrm{Q},\mathrm{K},\mathrm{V}=\mathrm{Split}\left(\mathrm{AttnMap}\left(\mathrm{F}_{\mathrm{qkv}}\right)\right)\right]\tag{7}$$

where Split function decomposes Q, K, V into query: Q, key: K and value: V, then, scaled attention is calculated:

$$\left[\mathrm{Attn}=\tanh\left(\mathrm{scalor}\times(\mathrm{Q}\cdot\mathrm{K})\right)\right]\tag{8}$$

where scalor is the scaling factor, typically set $\frac{1}{\sqrt{\mathrm{d}}}$ to stabilise gradients.

Finally, attention and value are merged:

$$\left[\mathrm{Z}_{\mathrm{high}}=\mathrm{Attn}\odot\mathrm{V}\right]\tag{9}$$

where $\odot$ represents element-wise multiplication, Z_{high} is the high frequency feature after fusion.

For the low-frequency information branch, the model focuses on overall context information through average pooling and global query operations, complementing macro-structural features overlooked by the high-frequency branch. The process is as follows:

Initially, a global query is generated:

$$\left[\mathrm{Q}_{\mathrm{global}} = \mathrm{Conv}_{1\times1}^{g\times d}(\mathrm{X})\right] \tag{10}$$

where g is the number of heads in the low-frequency branch, which is used to capture global contextual information of the image.

Then, low-frequency keys and values are generated through global average pooling:

$$\left[\mathrm{KVlow} = \mathrm{AvgPoolws}\left(\mathrm{Conv}_{1\times1}^{2g\times d}(\mathrm{X})\right)\right] \tag{11}$$

where AvgPoolws represents average pooling operations with a (ws × ws) window size, which is used for downsampling to extract low-frequency features.

Subsequently, low-frequency features are obtained through global context queries and local aggregation.

$$\begin{aligned}&\left[\mathrm{Attnlow} = \mathrm{softmax}\left(\mathrm{scalor}\times\left(\mathrm{Qglobal}\cdot\mathrm{KVlow}^{\mathrm{T}}\right)\right)\right]\\&\left[\mathrm{Zlow} = \mathrm{Attnlow}\cdot\mathrm{KVlow}\right]\end{aligned} \tag{12}$$

3.3 GCAF module

This paper introduces the GCAF module, which aims to significantly enhance the capability of feature representation in capturing contextual information by integrating convolutional and attention mechanisms. Specifically, the GCAF module is structured into two branches: a local branch and a global branch. The local branch employs convolution and channel shuffling techniques for local feature extraction, thereby ensuring the precise capture of detailed information. In contrast, the global branch models long-range dependencies through attention mechanisms (Cao et al., 2019; Zhang et al., 2018), effectively capturing global contextual information. Furthermore, the global branch leverages global information from the feature maps across spatial dimensions (height and width) to generate attention maps. These attention maps are subsequently utilised to weight the input feature maps, significantly enhancing the contextual information within the feature representation (Jiang et al., 2021). In this manner, the GCAF module is capable of concurrently capturing both local and global features, thereby substantially improving the overall feature representation capability and bolstering the model's performance in complex scenarios.

3.3.1 Local feature extraction branch

First, for a given input feature map, local feature maps are generated through a convolution layer:

$$\mathrm{Local}(\mathrm{x}) = \mathrm{Conv}(\mathrm{x}) \tag{13}$$

where Conv represents the standard convolution operation, x is the input feature map.

Subsequently, channel shuffling is used. By rearranging multiple groups of channels of the feature maps, the information can be better integrated across different channels. Specifically, the channel shuffle operation can be expressed as:

$$\mathrm{Shuffle}(\mathrm{x}) = \mathrm{ChannelShuffle}\left(\mathrm{Local}(\mathrm{x})\right) \tag{14}$$

In particular, channel shuffling is achieved by dividing the channels of the feature map into multiple groups, and then rearranging the channels within each group. The mathematical expression is:

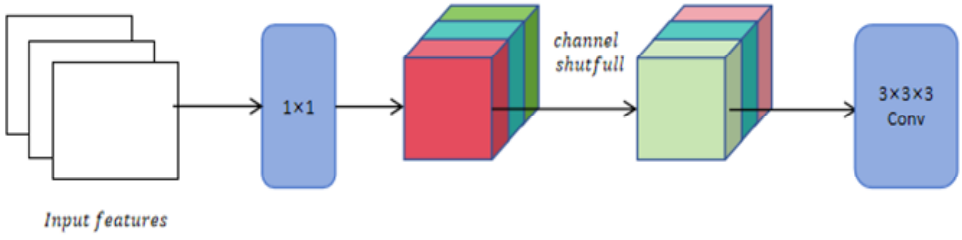
$$\text{Shuffle}(x) = \text{Reshape}\left(x, (B, G, C/G, H, W)\right) \tag{15}$$

$$\text{Shuffle}(x) = \text{Transpose}\left(x, (0, 2, 1, 3, 4)\right) \tag{16}$$

$$\text{Shuffle}(x) = \text{Reshape}\left(x, (B, C, H, W)\right) \tag{17}$$

where C is the number of channels of the input feature map, G is the number of groups, B is the batch size, H and W are the feature map respectively; C/G is the number of channels per group.

Figure 3 Logic diagram of the local feature extraction branch (see online version for colours)



Notes: The local feature extraction branch employs convolution operations and channel shuffling techniques to efficiently extract local features.

3.3.2 Global feature extraction branch

First, compute the dot product of Q and K, and obtain the attention weights through the softmax function:

$$\text{Attention}(Q, K) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right) \tag{18}$$

where d_k is the dimension of the keys, QK^T representing the dot product of queries and keys. Specifically, the dot product operation can be represented as:

$$QK^T = \sum_{i=1}^{d_k} Q_i K_i^T \tag{19}$$

Subsequently, apply the attention weights to the values, obtaining the weighted feature map that effectively captures global contextual information and long-range dependencies:

$$\text{Global}(x) = \text{Attention}(Q, K)V \tag{20}$$

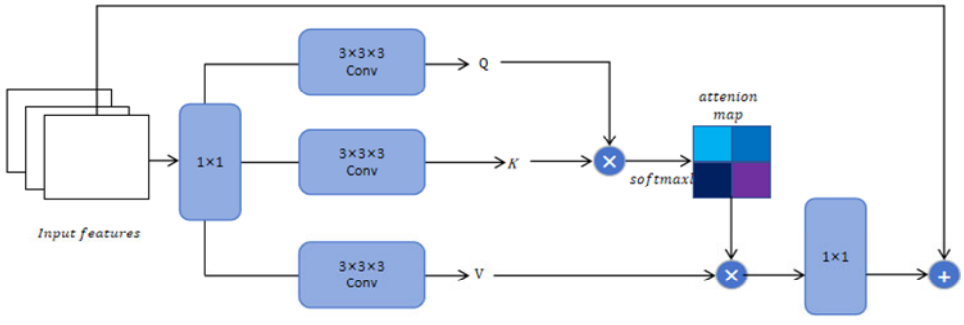
Next, the feature maps $X \in R^{B\times C\times H\times W}$ are divided into G groups, with each group containing C/G channels. The grouped feature maps are represented as:

$$X^{B\in G\times\frac{C}{G}\times H\times W} \tag{21}$$

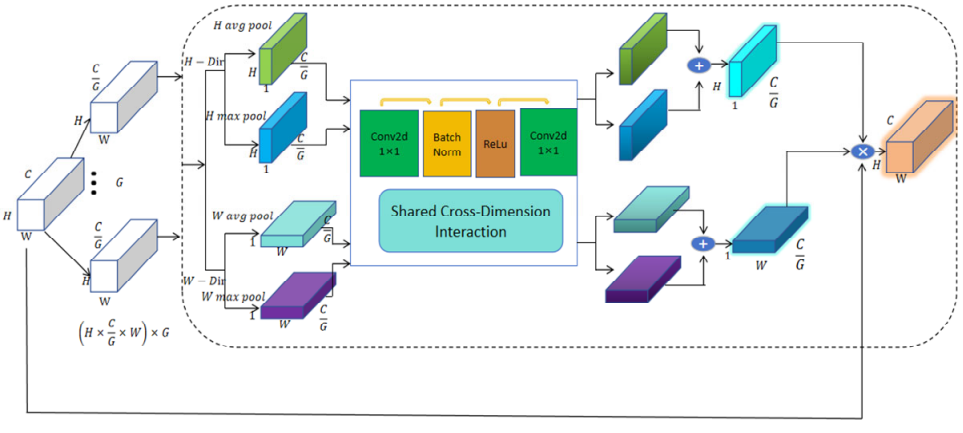
Subsequently, global average pooling and global max pooling operations are performed on the grouped feature maps in both the height and width dimensions:

$$\begin{aligned} X_{h,avg} &= AvgPool(X) \in R^{B\times G\times \frac{C}{G}\times H\times 1} \\ X_{h,max} &= MaxPool(X) \in R^{B\times G\times \frac{C}{G}\times H\times 1} \\ X_{w,avg} &= AvgPool(X) \in R^{B\times G\times \frac{C}{G}\times 1\times W} \\ X_{w,max} &= MaxPool(X) \in R^{B\times G\times \frac{C}{G}\times 1\times W} \end{aligned} \tag{22}$$

Figure 4 Logic diagram of the global feature extraction branch, (a) the global feature extraction branch uses attention mechanisms to model long-range dependencies, thereby capturing global features (b) it generates attention maps using global information from the feature maps in the spatial dimensions (height and width) and weights the input feature maps with these attention maps to enhance feature representation capability (see online version for colours)



(a)



(b)

After the operations are completed, a shared convolution layer is applied for feature processing on each grouped feature map. This shared convolution layer consists of two

1 × 1 convolution layers, a batch normalisation layer, and a ReLu activation function, which are used to reduce and restore the channel dimensions:

$$\begin{aligned} &Y_{h,avg} = Conv\left(X_{h,avg}\right),\ Y_{h,max} = Conv\left(X_{h,max}\right) \\ &Y_{w,avg} = Conv\left(X_{w,avg}\right),\ Y_{w,max} = Conv\left(X_{w,max}\right) \end{aligned} \tag{23}$$

Then, by summing the outputs of the convolution layers and applying a Sigmoid activation function, attention weights for the height and width dimensions are generated:

$$\begin{aligned} &A_h = \sigma\left(Y_{h,avg} + Y_{h,max}\right) \in R^{B\times G\times \frac{C}{G}\times H\times 1} \\ &A_w = \sigma\left(Y_{w,avg} + Y_{w,max}\right) \in R^{B\times G\times \frac{C}{G}\times 1\times W} \end{aligned} \tag{24}$$

where σ is a Sigmoid activation function.

Finally, the input feature maps are weighted according to the attention weights to obtain the final global feature map:

$$O = X \times A_h \times A_w \in R^{B\times C\times H\times W} \tag{25}$$

3.4 MSDE module

The MSDE module proposed in this paper significantly enhances the detailed features in images by integrating various convolutional operations (Zhang et al., 2018). Specifically, the MSDE module combines the weights and biases of different convolutional operations to construct a comprehensive convolutional kernel, thereby markedly improving the feature representation capability without incurring additional computational costs (Yu and Koltun, 2016). This module employs multi-scale atrous convolutions to extract features from diverse receptive fields, ensuring a thorough capture of detailed information, and further enhances feature expressiveness by incorporating channel and spatial attention mechanisms. The multi-scale atrous convolutions consist of five branches, each utilising distinct dilation rates for convolution and global average pooling to extract rich global features (Hu et al., 2018; Woo et al., 2018). Ultimately, the outputs of all branches are concatenated and fused through channel and spatial attention mechanisms, ensuring that detailed features are adequately extracted and enhanced, thereby improving the model's performance in complex scenarios.

3.4.1 Dynamic convolution branch

First, dynamic convolution kernel weights are generated through a convolution layer:

$$\text{Weights}(x) = \text{Conv}(x) \tag{26}$$

where Conv represents the standard convolution operation; x is the input feature map.

Then, the feature map is convoluted using the dynamic convolution kernel weights:

$$\text{DEConv}(x) = \text{DynamicConv}\left(x,\ \text{Weights}(x)\right) \tag{27}$$

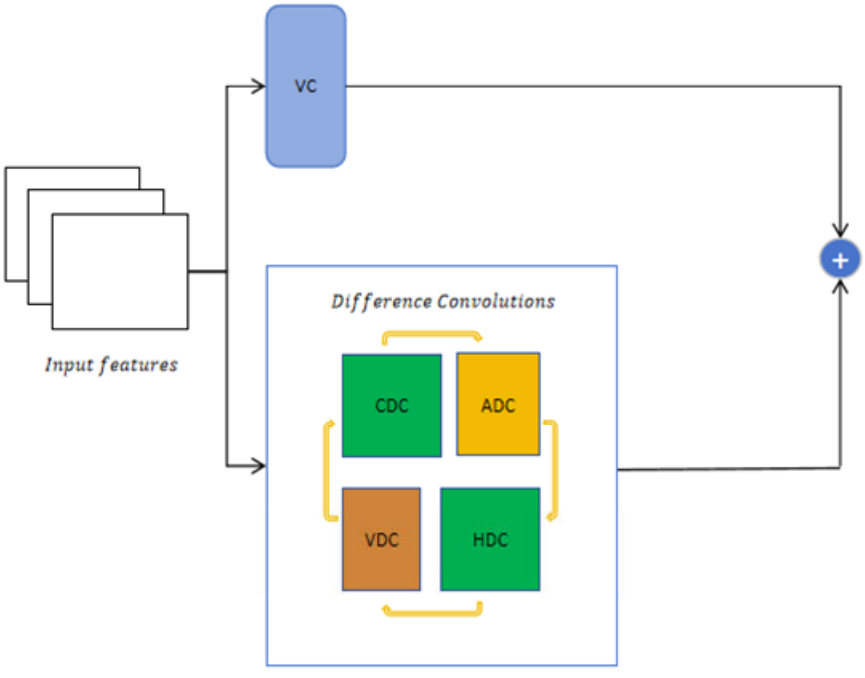
where DynamicConv represents the dynamic convolution operation.

The implementation of dynamic convolution can be expressed as:

$$\text{DynamicConv}\left(x, \text{Weights}(x)\right) = \sum_{i=1}^{k} \sum_{j=1}^{k} \text{Weights}_{ij}(x) \cdot x_{ij} \tag{28}$$

where k is the size of the convolution kernel; Weights_{ij} are the dynamically generated convolution kernel weights; x_{ij} is a local region of the input feature map.

Figure 5 Dynamic convolution branch logic diagram (see online version for colours)



Notes: By merging multiple convolution operations, a composite convolution kernel is formed.

3.4.2 Multi-scale dilated convolution

First, for a given input feature map x, multi-scale feature maps are generated using several different dilation rates in dilated convolution layers:

$$\text{AtrousConv}_i(x) = \text{Conv}_{d_i}(x) \tag{29}$$

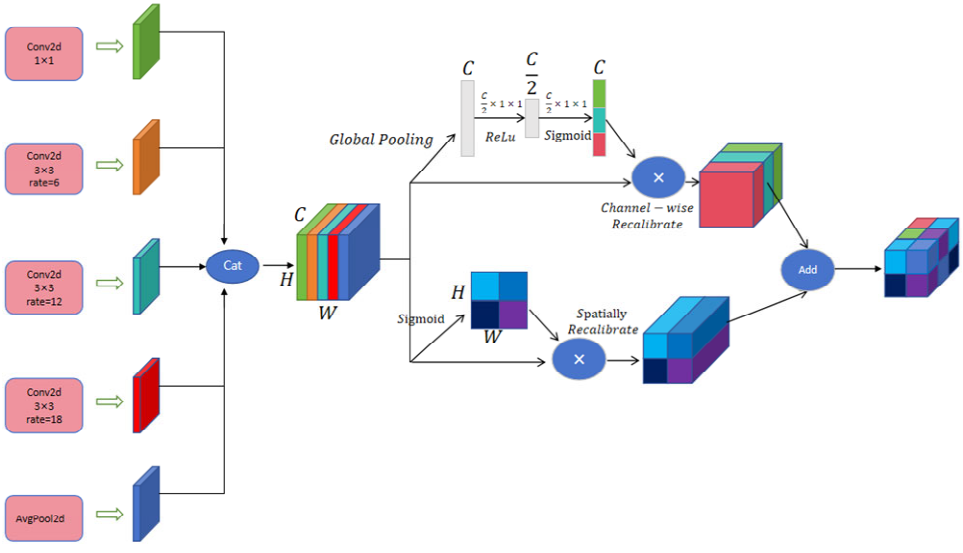
where Conv_{d_i} represents the dilated convolution operation with dilation rate d_i.

Then, the dilated convolution feature maps with different dilation rates are concatenated to obtain:AtrousConv(x):

$$\begin{aligned} &\text{AtrousConv}(x) \\ &= \text{concat}\left(\text{AtrousConv}_1(x), \text{AtrousConv}_2(x), \ldots, \text{AtrousConv}_n(x)\right) \end{aligned} \tag{30}$$

where concat represents the concatenation operation, assuming each dilated convolution feature map has dimensions B × C × H × W, and the dimension of the concatenated feature map is B × (n × C) × H × W, where n is the number of dilated convolutions.

Figure 6 Multi-scale dilated convolution logic diagram (see online version for colours)



Notes: Enhanced feature maps are obtained by concatenating different dilated convolutions.

4 Experiments

This section provides a comprehensive overview of the experimental setup employed to evaluate the MultiCAN-DEBEV model, encompassing the dataset, experimental configuration, 3D object detection results, ablation studies, and visualisation outcomes. The primary objective of these experiments is to assess the model's effectiveness and its adaptability in complex environments.

4.1 Dataset

This study utilises the nuScenes dataset to evaluate the performance of the enhanced BEVFormer model in 3D object detection tasks. The nuScenes dataset is a large-scale dataset for autonomous driving, designed to advance the development of autonomous driving technologies. It integrates data from multiple sensors, including multi-view images, LiDAR point clouds, and GPS/IMU information, thereby providing rich multimodal data suitable for various computer vision applications (Caesar et al., 2020).

The nuScenes dataset consists of 850 scenes spanning urban, suburban, and highway environments, with significant diversity in terms of geographical locations, weather conditions, and lighting. This variety makes the dataset highly representative of real world conditions. Each scene is meticulously annotated with multiple object types such as vehicles, pedestrians, and cyclists, with high annotation accuracy, providing a solid foundation for the training and evaluation of 3D object detection models. To ensure fairness and reproducibility in the experiments, the dataset is divided into a training set (700 scenes) and a validation set (150 scenes). The training set is primarily used for model training and parameter tuning, while the validation set is used to assess the

model's performance on unseen data. This division method effectively prevents overfitting and ensures the model's applicability in the real world.

4.2 Experimental setup

Experiments were conducted on a system with an NVIDIA RTX A5000 GPU to optimise training and inference efficiency. The model is implemented using the PyTorch framework, with AdamW as the optimiser, and R101 as the Backbone. The batch size is set to 24 to balance training speed and memory usage effectively. To evaluate model performance, the BEVFormer-s model is used as the baseline, with normalised detection score (NDS) and mean average precision (mAP) as evaluation metrics. NDS considers the accuracy, recall, and classification performance of detection, while mAP focuses on the precision of the detection results. These settings provide a solid foundation for subsequent experiments (Liu et al., 2022; Everingham et al., 2010).

4.3 3D object detection results

First, compare the MultiCAN-DEBEV model with the baseline model BEVFormer-S and other mainstream 3D object detection models (such as BEVDet-pure, SRCN3D, and PiCasso_OD_v1.0) on the nuScenes dataset.

Table 1 3D detection results of different models on nuScenes

Model	*NDS* ↑	*mAP* ↑
HENet_Sp	0.707	0.645
HaomoAI perception model	0.624	0.624
BEVDet-pure	0.463	0.398
SRCN3D	0.463	0.396
PiCasso_OD_v1.0	0.369	0.307
BEVFormer-s	0.462	0.409
MultiCAN-DEBEV	0.469	0.418

Notes: Compare the 3D detection performance of models contemporaneous with BEVFormer-S and those current at this time.

The enhanced MultiCAN-DEBEV model demonstrates superior performance compared to the BEVFormer-S model and several contemporaneous models on the nuScenes dataset, as evidenced by improvements in both the NDS and mAP metrics. Specifically, the NDS increased by 0.7% and the mAP by 0.9% relative to the BEVFormer-S model, thereby validating the effectiveness of the MultiCAN-DEBEV model. Notably, the MultiCAN-DEBEV model excels in complex scenarios, particularly in pedestrian and vehicle detection tasks, where it effectively captures detailed features, thereby enhancing detection accuracy. However, given that the BEVFormer-S was introduced in 2022 and that the MultiCAN-DEBEV is an optimisation of the BEVFormer framework, a significant performance gap remains when compared to various new models introduced after 2023. Nevertheless, the MSF-DySample, GCAF, and MSDE modules incorporated into the MultiCAN-DEBEV model are modular in nature, endowing them with strong universality. Furthermore, with the anticipated introduction of BEVFormer V2 and

CLIP-BevFormer, which are based on the BEVFormer model, it is expected that these modules will be applicable to these new models in future experiments, thereby facilitating improvements and optimisations that enhance overall performance.

4.4 Ablation study

To analyse the contribution of each module to the model's performance, ablation experiments were conducted by systematically removing the MSF-DySample, GCAF, and MSDE modules, and recording the changes in model performance.

Table 2 Detection results of the MultiCAN-DEBEV model on the nuScenes dataset for each module

Model	*NDS* ↑	*mAP* ↑
BEVFormer-s	0.462	0.409
+MSF-DySample ↑	0.463	0.413
+GCAF ↑	0.462	0.410
+MSDE ↓	0.461	0.400
+MSF-DySample + GCAF ↑	0.464	0.411
+MSF-DySample + MSDE ↑	0.485	0.464
+GCAF +MSDE ↑	0.464	0.412
MultiCAN-DEBEV ↑	0.469	0.418

Note: Each module was tested in combination on the nuScenes dataset.

In comparison to the baseline model BEVFormer-S, the incorporation of the MSF-DySample and GCAF modules results in a modest increase in both NDS and mAP metrics. Conversely, the exclusive insertion of the MSDE module leads to a slight decline in NDS and mAP. Nevertheless, given the MSDE module's enhanced capability for extracting detailed features in complex scenarios, subsequent integration experiments were conducted. From the analysis of individual module insertions, it can be concluded that the inclusion of the MSF-DySample and GCAF modules facilitates dynamic upsampling and the capture of long-range dependencies, thereby improving model performance and optimisation. While the introduction of the MSDE module results in a minor reduction in performance, it remains justifiable due to its enhancement of the model's ability to recognise detailed features through dilated fusion.

Subsequent experiments were performed with combinations of the MSF-DySample + GCAF, MSF-DySample + MSDE, and GCAF + MSDE modules. The results indicate that these pairwise combinations yield improvements in both NDS and mAP metrics. Notably, the combination of the MSF-DySample and MSDE modules produced the most significant performance enhancement, with an increase of 2.3% in NDS and 5.5% in mAP compared to the baseline model BEVFormer-S. The experimental findings suggest that capturing high and low-frequency features from the original feature maps, followed by dilated fusion to extract detailed features, can substantially enhance model performance and target recognition. Additionally, other pairwise combinations also demonstrated slight improvements, further validating the efficacy of simultaneous module utilisation.

Ultimately, the simultaneous incorporation of the MSF-DySample, GCAF, and MSDE modules into the baseline model resulted in a modest performance improvement

and an enhancement in three-dimensional recognition capabilities. The experimental results illustrate that by enhancing the extraction of high and low-frequency features with the MSF-DySample module, capturing both global and local features with the GCAF module, and employing the MSDE for multi-scale feature fusion to refine detailed features, the original model's feature maps can be significantly enhanced, leading to an overall improvement in model performance and an increased rate of three-dimensional object recognition.

4.5 Visualisation results

To further validate the efficacy of the MultiCAN-DEBEV model in three-dimensional object detection, a selection of detection results was subjected to visualisation analysis.

Figure 7 Visualisation of the MultiCAN-DEBEV model in complex scenarios (see online version for colours)

Notes: The model performs object detection and labelling for pedestrians and vehicles under challenging conditions.

The images presented illustrate the detection results of the MultiCAN-DEBEV model within complex scenes, encompassing bounding boxes for both pedestrians and vehicles. It is apparent that the MultiCAN-DEBEV model excels in accurately identifying and localising various categories of targets, even in instances of occlusion. Furthermore, the model demonstrates a high degree of accuracy for distant targets. This effectively underscores the contributions of the MSF-DySample and GCAF modules in enhancing multi-scale features and long-range dependencies for three-dimensional detection within this framework.

Figure 8 Visualisation of the MultiCAN-DEBEV model in a parking lot scenario with multiple vehicles and long-distance detection (see online version for colours)

Notes: In this complex situation, where multiple vehicles and distant cars are present, the model performs object detection and labelling.

The images illustrate that, through detail enhancement, the model effectively recognises vehicle information at extended distances. Furthermore, by employing multi-scale convolution, it accurately identifies and precisely localises various types of vehicles. This

robust performance underscores the significant role of the MSDE module in enhancing detailed features and utilising multi-scale dilated convolution for three-dimensional detection within this framework.

5 Discussion and conclusions

This study presents the MultiCAN-DEBEV model, an advanced iteration of the BEVFormer-S algorithm designed to enhance BEV representation and 3D object detection. The model consists of three main modules: MSF-DySample, GCAF, and MSDE. MSF-DySample employs a dual-branch framework for dynamic upsampling, effectively capturing both high-frequency and low-frequency information. GCAF integrates local and global features to enhance contextual acquisition and feature robustness, while MSDE refines feature representation through the application of deformable and dilated convolutions. Collectively, these three modules address challenges such as the loss of vertical structural details and information regarding occluded objects that may arise in complex scenarios.

Experimental results indicate that MultiCAN-DEBEV outperforms BEVFormer-S on the nuScenes dataset, excelling in detail acquisition in complex scenarios and significantly reducing the rates of missed detections and false positives. Its modular design facilitates seamless integration with other models. This research advances the theory of 3D object detection and its applications in autonomous driving and intelligent monitoring; however, the integration of these modules introduces operations such as dynamic upsampling, global convolution, and multi-scale feature fusion, which increases computational complexity. This leads to prolonged training times and a reduction in real-time performance. Therefore, future research will focus on refining the model modules and employing optimisation techniques, such as pruning, quantisation, and knowledge distillation, to further improve the model's real-time performance while maintaining detection accuracy. Additionally, since the model has only been validated on the nuScenes dataset, there may be a risk of overfitting. Therefore, in subsequent research, data augmentation techniques will be employed, and validation will be conducted on other datasets, such as KITTI, to ensure the model's generalisation capability. Finally, errors in feature extraction or fusion could introduce artefacts in the reconstructed 3D images. To mitigate this, future research will apply post-processing techniques to the reconstructed 3D images, aiming to reduce the impact of such artefacts as much as possible.

Declaration of interests

All authors declare that they have no conflicts of interest.

References

Caesar, H., Yang, X. and Beutler, B. (2020) 'nuScenes: a large-scale dataset for autonomous driving', *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp.11621–11631.

Cao, Y., Shen, Z. and Zhang, L. (2019) *Global Context Attention for Semantic Segmentation*, arXiv preprint arXiv:1904.03140.

Chen, L., Wang, Z. and Huang, G. (2022) 'Multi-scale feature fusion for 3D object detection in autonomous driving', *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp.2672–2681.

Chen, X., Li, Y. and Liu, W. (2021) 'Attentive convolutions for accurate image classification', *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp.3680–3689.

Chen, Y. (2024) *Research on Key Technologies for 3D Object Detection in Front of the Vehicle Based on Deep Learning*, Jimei University, DOI: 10.27720/d.cnki.gjmdx.2024.000085.

Everingham, M., Van Gool, L., Williams, C.K., Winn, J. and Zisserman, A. (2010) 'The Pascal visual object classes (VOC) challenge', *International Journal of Computer Vision*, Vol. 88, pp.303–338, https://doi.org/10.1007/s11263-009-0275-4.

Hu, J., Shen, L. and Sun, G. (2018) 'Squeeze-and-excitation networks', *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp.7132–7141.

Jiang, Y., Lu, L. and Xu, J. (2021) 'Enhanced view-independent representation method for skeleton-based human action recognition', *International Journal of Information and Communication Technology*, Vol. 19, No. 2, pp.201–218.

Li, Z., Wang, W., Li, H., Xie, E., Sima, C., Lu, T., Qiao, Y. and Dai, J. (2022) *BEVFormer: Learning Bird's-Eye-View Representation from Multi-Camera Images via Spatiotemporal Transformers*, arXiv preprint arXiv:2203.17270.

Li, Z., Yan, F. and Zhang, Y. (2023) 'BEVFusion: bridging LiDAR-camera features for multi-view 3D object detection', *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 45, No. 3, pp.890–903.

Liu, S., Qi, L., Qin, H., Shi, J. and Jia, J. (2020) 'Simple and effective adaptive point-wise attention for scene parsing', *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp.7692–7701.

Liu, W., Liao, S. and Lin, T.Y. (2022) *NDS: Normalized Detection Score for Object Detection*, arXiv preprint arXiv:2202.01512.

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S. and Guo, B. (2021) 'Swin transformer: hierarchical vision transformer using shifted windows', in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp.10012–10022.

Qi, C.R., Yi, L., Su, H. and Guibas, L.J. (2017) 'PointNet++: deep hierarchical feature learning on point sets in a metric space', *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp.5105–5114.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.S., Kaiser, Ł. and Polosukhin, I. (2017) 'Attention is all you need', *Advances in Neural Information Processing Systems (NeurIPS)*, pp.5998–6008.

Wang, J., Chen, K., Xu, R., Liu, Z., Loy, C.C. and Lin, D. (2019) 'Carafe: content-aware reassembly of features', in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp.3007–3016, arXiv preprint arXiv:1905.02188.

Woo, S., Park, J., Lee, J.Y. and Kweon, I.S. (2018) 'CBAM: convolutional block attention module', *Proceedings of the European Conference on Computer Vision (ECCV)*, pp.3–19.

Wu, H., Li, Y., Zhang, J. and Kuang, Y. (2021) 'Fast road scenarios recognition of intelligent vehicles by image processing', *International Journal of Information and Communication Technology*, Vol. 18, No. 1, pp.1–15.

Yin, T., Zhou, X. and Krähenbühl, P. (2021) *CenterPoint: A Two-Stage Object Detector for 3D Point Cloud*, arXiv preprint arXiv:2012.11490.

Yu, F. and Koltun, V. (2016) *Multi-Scale Context Aggregation by Dilated Convolutions*, arXiv preprint arXiv:1511.07122.

Zhang, X., Zhang, T., Qi, X., Li, H. and Huang, X. (2020) 'FADE: fused attention in multi-scale decoder for instance segmentation', *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp.16282–16291.

Zhang, X., Zhou, X., Lin, M. and Sun, J. (2018) 'ShuffleNet: an extremely efficient convolutional neural network for mobile devices', *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp.6848–6856.

Zhong, S. and Hu, T. (2022) 'A multi-attribute recognition method of vehicle's line-pressing in parking lot based on multi-task convolution neural network', *International Journal of Information and Communication Technology*, Vol. 20, No. 3, pp.308–324.

Zhou, Y. and Tuzel, O. (2018) 'VoxelNet: end-to-end learning for point cloud based 3D object detection', *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp.4490–4499.