EA-BEV: Attention-enhanced Multimodal Fusion Method for 3D Object Detection

Yehui Ding[1*], Yuntao Shi[1]

[1] North China University of Technology, Beijing, 100144, China

dingyehui@mail.ncut.edu.cn

https://doi.org/10.70695/AA1202502A18

Abstract

To address the issue of low accuracy in object detection for autonomous driving, we propose an attention-enhanced multi-modal fusion three-dimensional object detection method (EA-BEV). This method incorporates a self-attention mechanism in the image processing network, which effectively extracts deep features and reduces the problem of insufficient image feature extraction caused by semantic information blurriness. In the point cloud processing network, we designed a high-order convolutional spatial attention mechanism that significantly enhances the network's ability to model and express non-linear deep features of point clouds, thereby improving the global descriptive capability of point cloud information. We conducted comparative experiments on the nuScenes dataset, and the results show that the mAP metric is 76.2% and the NDS metric is 74.4%. The EA-BEV method demonstrates a clear advantage in the accuracy of 3D object detection, providing a new approach for environmental perception in autonomous driving.

Keywords BEV; 3D Object Detection; Multimodal Fusion

1 Introduction

With advancements in autonomous driving and smart transportation systems, autonomous driving technology for electric vehicles has also advanced rapidly, attracting widespread attention from both academia and industry [1]. The goal of autonomous driving is to enable vehicles to intelligently perceive their surrounding environment. Environmental perception is one of the core technologies of autonomous driving, and the accuracy of perception directly affects the stability and safety of the autonomous driving system. Currently, 3D object detection methods for autonomous driving can be categorized into three types based on the type of sensors used: camera-based 3D object detection methods, LiDAR-based 3D object detection methods, and camera-LiDAR fusion-based 3D object detection methods [2]. In camera-based 3D object detection methods, a surround-view camera (multi-camera) strategy is commonly adopted. Multiple cameras can comprehensively capture environmental information around the autonomous vehicle (e.g., BEVFormer [3]). However, since cameras capture 2D image data lacking depth information and are susceptible to lighting conditions, detection accuracy is relatively low, and false detections are common, which limits their application in autonomous driving scenarios.

In contrast, LiDAR-based 3D object detection methods use multi-beam LiDAR to collect information around the vehicle (e.g., PointPillars [4]). Although the captured point clouds contain depth information and are unaffected by lighting, they are sparse and lack texture details, which can also lead to false detections. Camera-LiDAR fusion-based 3D object detection methods (e.g., TransFusion [5]) can leverage the rich semantic information in images and provide highly accurate depth information, enabling better performance and broader application scenarios. Therefore, exploring 3D object detection methods based on image and point cloud fusion offers a valuable approach for the industry.

The 3D object detection method based on image and point cloud fusion can overcome the limitations caused by the sparsity of point clouds and the loss of semantic information in single sensors. Depending on the fusion strategy, it can be divided into three types: early fusion (data-level fusion), deep fusion (feature-level fusion), and late fusion (object-level fusion) [6]. Classic camera and LiDAR fusion methods focus on feature-level fusion (i.e., deep fusion), using heuristic algorithms to separately process the data features from cameras and LiDAR. A fusion module then generates BEV (Bird's Eye View) features to achieve 3D object detection. The BEVFusion [7] algorithm is a pioneering algorithm for scene-level BEV fusion. BEVFusion focuses on how to perform the fusion and designs modules for the

camera data stream, LiDAR data stream, and fusion module, unifying the camera and LiDAR features into the BEV space. The MetaBEV [8] algorithm introduces cross-modal attention integration through a Mixture of Experts (MoE), further enhancing the BEV features. The Is-Fusion [9] algorithm proposes an instance scene feature fusion method, strengthening the capability of query features. The BEVDiffuser [10] algorithm introduces a diffusion model to further improve the fusion ability of network features. However, these algorithms mainly enhance the BEV feature queries on the BEV feature maps, while overlooking the insufficient semantic information in the image and point cloud features obtained from various sensors, which ultimately limits the detection performance after fusion. Most existing fusion algorithms are unable to directly process both image and point cloud data. For image data, this often results in inadequate extraction of texture information and blurred semantic representation. For point cloud data, the use of a single CNN-based convolution model relies only on first-order statistics, which fails to capture global information and ignores the ability to model the nonlinear deep features inherent in spatial point cloud data.

To address the above issues, we propose an attention-enhanced multimodal fusion 3D object detection method, which we call EV-BEV. This method fully leverages the two modalities of multi-camera and LiDAR data to generate high-quality fused BEV features. Our contributions are as follows:

A high-order convolutional spatial attention mechanism is designed in the LiDAR data stream, which computes the covariance matrix of the mapped point cloud features to obtain high-order features of the point cloud, thereby enhancing the network's capability to model spatial correlations in point cloud data.

A self-attention mechanism is designed in the image data stream to enhance the extraction of image texture information and capture deep features with rich semantic content.

Experiment on nuScenes dataset, and the results demonstrate that our EV-BEV method achieves superior performance in multimodal fusion 3D object detection.

2 Implementation Method

In this section, we first introduce the overall structure of the EA-BEV algorithm that we designed; next, we analyze the theories of high-order convolutional spatial attention and self-attention mechanisms.

2.1 EM-BEV Overall Architecture

The overall architecture of the EM-BEV we designed is shown in Figure 1. EM-BEV consists of a multi-camera processing module (Camera Stream), a point cloud processing module (LiDAR Stream), and a BEV feature fusion module. In the Camera Stream, we design a self-attention mechanism module to improve the network's feature extraction capability from multi-camera views, enhancing the network's ability to extract image texture information and obtain rich semantic deep features. In the LiDAR Stream, we design a high-order convolution module to enhance the network's ability to capture the nonlinear (second-order) deep features of point cloud data and improve its capacity to describe global information. The fusion module employs a fusion mechanism similar to BEVFusion [7]. First, the BEV features of point clouds and images are concatenated along the channel dimension, and the concatenated features are then extracted through a convolutional network. Next, global average pooling and convolution prediction are used to adaptively select the concatenated features. Finally, the fused BEV features are output and fed into the detection head for detection to obtain the final detection results.

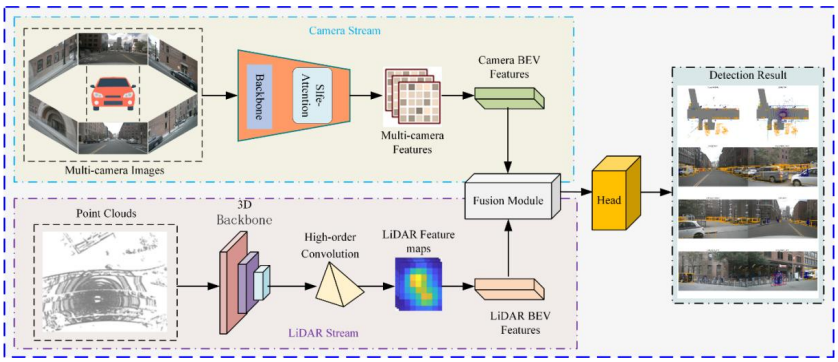

Fig. 1. Overall architecture of EM-BEV

2.2 High-order Convolution Mechanism

The feature map output from the convolution processing of point cloud data in the 3D Backbone is denoted as ***F***, which serves as the input for higher-order convolution. Here $F \in$ $^{H\times W\times C}$, where H represents the height of the feature map, W represents the width of the feature map, and C represents the number of channels in the feature map. The feature map $F_i\left(i=1,\ldots,C\right)$ is processed into a column vector f_i with a channel count of 1, and then the correlation covariance of each column vector (feature channel) is calculated as:

$$P=\begin{bmatrix} \operatorname{cov}(f_1,f_1) & \operatorname{cov}(f_1,f_2) & \cdots & \operatorname{cov}(f_1,f_c) \\ \operatorname{cov}(f_2,f_1) & \operatorname{cov}(f_2,f_2) & \cdots & \operatorname{cov}(f_2,f_c) \\ \vdots & \vdots & \vdots & \vdots \\ \operatorname{cov}(f_c,f_1) & \operatorname{cov}(f_c,f_2) & & \operatorname{cov}(f_c,f_c) \end{bmatrix} \tag{1}$$

For the sake of convenience in calculations, we compose the feature matrix $A \in$ $^{C\times M}$ from the feature channel vector f_i ,where $M=H\times W$,with M representing the M-dimensional local features in a channel space. The covariance matrix P is obtained through the potential interactions of the convolution features, and the expression for P is as follows:

$$P=A\chi A^T \tag{2}$$

where χ is defined as:

$$\chi=\frac{1}{M}\left(I-\frac{1}{M}NN^T\right) \tag{3}$$

where I is the identity matrix of size $M\times M$,and $N=[1,1,\ldots,1]$ is an n-dimensional vector where all elements are 1. The covariance matrix A uses a global CP method based on iterative matrix square root normalization instead of ordinary average pooling. Any symmetric positive definite matrix has a unique square root that can be accurately computed through EIG (Eigenvalue Decomposition) or SVG (Singular Value Decomposition). Therefore, the matrix A is decomposed into its EIG as follows:

$$EIG(A)=B\Lambda \mathrm{B}^T \tag{4}$$

where B is an orthogonal matrix, and $\Lambda=diag(\lambda_1,\ldots,\lambda_c)$ is a diagonal matrix with the eigenvalues λ_i of A on its diagonal. Therefore, A has a square root η ,

$$\eta_l=Bdiag(\lambda_i^{\frac{1}{2}})B^T \tag{5}$$

Among them, $\eta^2=A$, but EIG and SVD are not supported on graphic processors. Higham et al. [11] studied a class of iterative algorithms for computing the square root of a matrix. Subsequently, Li et al. [12] proposed an iterative matrix square root normalization for CP using the Newton-Schultz iteration based on this method. This process consists of three distinct stages: pre-normalization, Newton-Schultz iteration, and post-compensation.

First, this iterative method converges within a local boundary [13], and the pre-normalization is achieved by computing the trajectory of P, defined as follows:

$$A=\frac{1}{tr(P)}P \tag{6}$$

Next, the Newton-Schultz iteration is performed, where this stage is repeated k times to produce the approximate matrix square root η_l $(l=1,\ldots,c)$. Based on the given initial matrices $\eta_0=A$ and $\mu_0=I$, the computation structure for the square root Y of A is as follows:

$$\eta_l = \frac{1}{2}\eta_{l-1}(3I - \mu_{l-1}\eta_{l-1}), \mu_l = \frac{1}{2}(3I - \mu_{l-1}\eta_{l-1})\mu_{l-1} \tag{7}$$

The post-compensation η_l is defined as follows:

$$cp = \sqrt{tr(P)}\eta_l \tag{8}$$

The final square root *cp* obtained is the result after the CP operation, which is then input into the fully connected layer for downstream tasks.

To optimize the efficiency of calculating the covariance matrix, *H* and *W* are simplified to *H′* and *W′*. The feature $F' \in \mathbb{R}^{H' \times W'}$ is interpreted as having *c*-dimensional local features at each spatial location, forming a two-dimensional space with dimensions $H' \times W'$. The spatial covariance matrix of size $M \times M$ after second-order pooling is as follows:

$$P' = A\chi A^T \tag{9}$$

The *i*-th row of P', where $i = (1, \ldots, N)$ represents the statistical dependence relationship between the spatial feature *i* and all features in P'. The spatial dimensions are scaled from *H′* to *H* and from *W′* to *W*.

The specific operations are as follows: perform row-wise convolution on the covariance matrix, with the number of input channels being *M* and the number of output channels being *4M*, resulting in a tensor of size *4M* ×*1*. Next, the tensor is adaptively downsampled and then convolved again, with the number of channels being *M*. After applying the sigmoid activation function, the tensor is upsampled to a resolution of *H* ×*W*, generating a weight tensor $V \in \mathbb{R}^{H \times W}$. As shown in Figure 2.

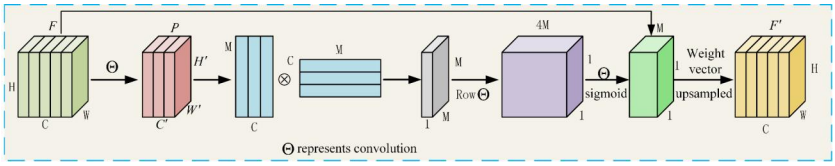

Fig. 2. Flowchart of high-order convolution

2.3 Self-attention Mechanism

To address the issues of insufficient texture information extraction and vague semantic information during feature extraction in image data, a self-attention mechanism has been adopted. The self-attention mechanism plays a key role in image feature extraction. By calculating the weighted relationships between input features, it effectively captures long-range dependencies and important features within the image. This mechanism enables the model to adaptively focus on different regions when processing features, thereby enhancing the expression of important information and suppressing redundant or irrelevant features, significantly improving the effectiveness of object detection. The structural diagram of the self-attention mechanism is shown in Figure 3.

Self-attention mechanism [14] Map the input sequence $X = (x_1, x_2, \ldots, x_n)$ linear transformation into query vector $Q = (q_1, q_2, \ldots, q_n)$, bond vector $K = (k_1, k_2, \ldots, k_n)$ and value vector $V = (v_1, v_2, \ldots, v_n)$ .By calculating the similarity between vector *Q* and vector *K* $S_{i,j} = Q_i^T K_j$,Then convert it into attention distribution through *softmax* function *SAtt*. And weighted sum of the value vector *V* with this distribution to obtain the upper and lower vector *C*. *SAtt* is defined as:

$$SAtt = \exp(S_{i,j}) \cdot (\sum_k^n \exp(S_{i,k}))^{-1} \tag{10}$$

$$C_i = \sum_{j=1}^n A_{i,j} v_{i,j} \tag{11}$$

The *SAtt* calculation formula can be written as:

$$SAtt(Q,K,V) = soft\max\left(\frac{QK^T}{\sqrt{d_k}}\right) \tag{12}$$

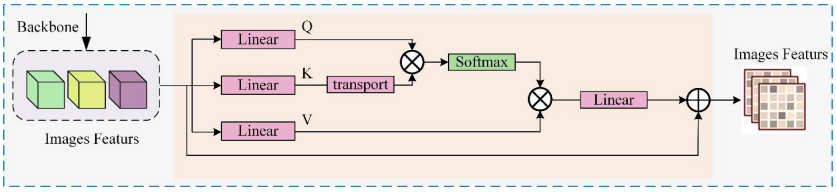

Fig.3. Self-Attention mechanism structural diagram

3 Experimental Results and Analysis

In this section, we first introduce the dataset and evaluation metrics used by the EA-BEV algorithm; next, we present the settings of the parameters and the operating environment during algorithm training; finally, we analyze the experimental results and the ablation experiments.

3.1 Dataset and Evaluation Metrics

Dataset

The nuScenes [15] dataset is a large-scale dataset for 3D perception, containing over 40,000 annotated scenes. Each sample in the dataset is equipped with inputs from lidar and surrounding cameras. This dataset not only provides multi-view image data but also includes rich environmental information, such as road signs, pedestrians, and vehicles. This makes nuScenes an important resource for research in autonomous driving and intelligent transportation systems, allowing researchers to utilize its multimodal data for model training and validation. The specific number of scenes used in this experiment is as follows: the training set consists of 700 scenes, the validation set has 150 scenes, and the test set contains 150 scenes (excluding annotated data).

Evaluation Metrics

The official nuScenes dataset provides seven evaluation metrics, which are mAP, NDS, mATE, mASE, mAOE, mAVE, and mAAE.

3.2 Experimental Parameters

Experimental Setup

The training process for the MLDF-BEV model employed the AdamW optimizer [16], featuring a weight decay coefficient of 0.01 and an initial learning rate of 2.0×10^{-4}, which was gradually decreased using a cosine annealing schedule [17]. In line with the CenterPoint method [18], the ground truth for 3D spatial instances was randomly rotated within a range of ±22.5°.

EA-BEV Operating Environment

The entire training process was conducted on four RTX 4090 GPUs over 24 epochs, with a batch size of 8. During model inference, no test time augmentation (TTA) techniques were applied. The software and hardware configurations used in this experiment, as well as the version details, are as follows: CPU is Gold 5218R×2, RAM is 224GB, GPU is RTX 4090×4, the operating system is Ubuntu 18.04, CUDA version is 11.1, and PyTorch version is 1.9.1.

3.3 Experimental Results

Our EA-BEV model is first evaluated on the nuScenes validation set and compared with current state-of-the-art methods, with the detailed comparison results shown in Table 2. To ensure the objectivity of the comparison data and eliminate other interfering factors, we trained all the algorithms in Table 2 on our server for 24 epochs. The experimental parameters were set the same as those of our EA-BEV, and the results after training are shown in Table 1. Our EA-BEV has improved the mAP value by 0.8% compared to V-Fusion[19]. Our EA-BEV algorithm has improved the NDS value by 1.2% compared to Is-Fusion[9]. Similarly, we conducted inference tests on the nuScenes test set, and the experimental

results are shown in Table 2. Our EA-BEV has improved the mAP value by 0.6% compared to V-Fusion[19]. Our EA-BEV algorithm has improved the NDS value by 0.8% compared to BEVDiffuser [9].

The reasons for the superior performance of our algorithm can be attributed to two main factors. First, during the feature extraction process, we introduced an image self-attention mechanism. This mechanism effectively enhances the ability to extract semantic information and deep features, reducing semantic ambiguity between features and ensuring that the model can accurately understand and process complex images. This means that the model can more precisely capture important information within the images, thereby improving overall performance. Second, we adopted a higher-order spatial attention mechanism based on higher-order convolutions. This mechanism leverages the covariance matrix of the point cloud feature maps to obtain higher-order features of the point cloud. This approach not only allows the model to better understand spatial relationships within the point cloud data but also enhances the network's ability to model the spatial correlations of the point cloud data. This improved modeling capability enables our algorithm to exhibit greater robustness and accuracy when handling point cloud data in practical applications. Overall, these two factors collectively drive the superior performance of our algorithm, allowing it to achieve favorable results across various tasks. The visualization results are shown in Figure 4.

Table 1. Comparison of 3D object detection performance based on nuScenes image val set. The best and second-best performances are marked in red and blue, respectively.

Methods	Backbone	mAP↑	NDS↑	mATE↓	mASE↓	mAOE↓	mAVE↓	mAAE↓
TransFusion[5]	ResNet	0.675	0.713	0.283	0.278	0.285	0.301	0.246
BEVFusion[7]	ResNet	0.685	0.714	0.271	0.276	0.269	0.286	0.238
MetaBEV[8]	ResNet	0.680	0.715	0.268	0.265	0.262	0.272	0.228
Is-Fusion[9]	ResNet	0.720	0.732	0.273	0.254	0.287	0.260	0.219
BEVDiffuser[10]	ResNet	0.719	0.692	0.276	0.252	0.294	0.266	0.184
V-Fusion[19]	ResNet	0.727	0.730	0.251	0.243	0.258	0.259	0.199
SparseFusion[20]	ResNet	0.710	0.731	0.262	0.250	0.264	0.256	0.201
EA-BEV	ResNet+EA	0.735	0.744	0.253	0.241	0.255	0.259	0.198

Table 2. Comparison of 3D object detection performance based on nuScenes image test set. The best and second-best performances are marked in red and blue, respectively.

Methods	Backbone	mAP↑	NDS↑	mATE↓	mASE↓	mAOE↓	mAVE↓	mAAE↓
TransFusion[5]	ResNet	0.716	0.689	0.285	0.277	0.283	0.303	0.248
BEVFusion[7]	ResNet	0.729	0.702	0.269	0.273	0.267	0.286	0.237
MetaBEV[8]	ResNet	0.745	0.714	0.269	0.268	0.267	0.278	0.229
Is-Fusion[9]	ResNet	0.752	0.730	0.277	0.258	0.289	0.253	0.221
BEVDiffuser[10]	ResNet	0.753	0.733	0.277	0.252	0.296	0.265	0.188
V-Fusion[19]	ResNet	0.756	0.732	0.253	0.246	0.254	0.257	0.201
SparseFusion[20]	ResNet	0.738	0.720	0.264	0.255	0.267	0.259	0.204
EA-BEV	ResNet+EA	0.762	0.741	0.255	0.244	0.257	0.261	0.200

Fig.4. The visualization results of EA-BEV

3.4 Ablation Study

To verify the effectiveness of the component modules designed in EA-BEV for 3D object detection in autonomous driving, we will integrate the designed high-order convolutional layer and self-attention mechanism into the ResNet network for performance comparison. The setup of the ablation experiments and the experimental results are shown in Table 3. In the feature extraction network, when only the high-order convolution module is used, the mAP metric improves by 2.6%, and the NDS metric improves by 1.6%. When only the self-attention module is used, the mAP metric improves by 2.2%, and the NDS metric improves by 1.1%. When both modules are used simultaneously, the mAP metric improves by 3.7%, and the NDS metric improves by 2.8%. The high-order convolutional spatial attention module can capture higher-order features of point cloud information, thereby enhancing the network's capability to model global information representation. The self-attention module improves the ability to extract texture semantic information during image feature extraction and enhances the effectiveness of information capture in deep image features.

Table 3. The ablation studies of the EA-BEV module.

High-order Convolution	Self-attention	mAP	NDS
-	-	0.698	0.716
√	-	0.724	0.732
-	√	0.720	0.727
√	√	0.735	0.744

4 Conclusion

In the EA-BEV model we proposed, different feature extraction enhancement methods were adopted for image data processing and point cloud data processing, respectively. The image data processing utilized a self-attention mechanism, which effectively reduces the issues of insufficient image feature extraction caused by the ambiguity of semantic information. In the point cloud processing network, we designed a high-order convolutional spatial attention mechanism, which can effectively enhance the network's global descriptive ability for point cloud information. This improved the EA-BEV model's object detection capability, while also increasing the model's mAP and NDS values. The EA-BEV method proposed in this paper has achieved significant progress in the multimodal fusion technology for autonomous driving. However, this method also has certain limitations. The use of attention based on the Transformer mechanism in the model results in a longer training time for the detector. To address this issue, further optimization of the network model structure can be implemented to enhance the training speed.

Acknowledgement

This work was supported without any funding.

Conflicts of Interest

The authors declare no conflicts of interest.

References

1. Guo, J., Kurup, U., & Shah, M. (2019). Is it safe to drive? An overview of factors, metrics, and datasets for driveability assessment in autonomous driving. IEEE Transactions on Intelligent Transportation Systems, 21(8), 3135-3151.
2. Ma, Y., Wang, T., Bai, X., Yang, H., Hou, Y., Wang, Y., ... & Zhu, X. (2024). Vision-centric bev perception: A survey. IEEE Transactions on Pattern Analysis and Machine Intelligence.
3. Li, Z., Wang, W., Li, H., Xie, E., Sima, C., Lu, T., ... & Dai, J. (2025). BEVFormer: Learning Bird's-Eye-View Representation From LiDAR-Camera via Spatiotemporal Transformers. IEEE Transactions on Pattern Analysis and Machine Intelligence, 47(03), 2020-2036.
4. Lang, A. H., Vora, S., Caesar, H., Zhou, L., Yang, J., & Beijbom, O. (2019). Pointpillars: Fast encoders for object detection from point clouds. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 12697-12705).
5. Bai, X., Hu, Z., Zhu, X., Huang, Q., Chen, Y., Fu, H., & Tai, C. L. (2022). Transfusion: Robust lidar-camera fusion for 3d object detection with transformers. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 1090-1099).
6. Huang, K., Shi, B., Li, X., Li, X., Huang, S., & Li, Y. Multi-modal sensor fusion for auto driving perception: A survey. arXiv 2022. arXiv preprint arXiv:2202.02703.
7. Liang, T., Xie, H., Yu, K., Xia, Z., Lin, Z., Wang, Y., ... & Tang, Z. (2022). Bevfusion: A simple and robust lidar-camera fusion framework. Advances in Neural Information Processing Systems, 35, 10421-10434.
8. Ge, C., Chen, J., Xie, E., Wang, Z., Hong, L., Lu, H., ... & Luo, P. (2023). Metabev: Solving sensor failures for 3d detection and map segmentation. In Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 8721-8731).
9. Yin, J., Shen, J., Chen, R., Li, W., Yang, R., Frossard, P., & Wang, W. (2024). Is-fusion: Instance-scene collaborative fusion for multimodal 3d object detection. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 14905-14915).
10. Ye, X., Yaman, B., Cheng, S., Tao, F., Mallik, A., & Ren, L. (2025). BEVDiffuser: Plug-and-Play Diffusion Model for BEV Denoising with Ground-Truth Guidance. In Proceedings of the Computer Vision and Pattern Recognition Conference (pp. 1495-1504).
11. Higham, N. J. (2008). Functions of Matrices: Theory and Computation. Springer-Verlag.
12. Li, P., Xie, J., Wang, Q., & Gao, Z. (2018). Towards faster training of global covariance pooling networks by iterative matrix square root normalization. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 947-955).
13. Chakravarthy, S. S., Bharanidharan, N., & Rajaguru, H. (2023). Deep learning-based metaheuristic weighted k-nearest neighbor algorithm for the severity classification of breast cancer. IRBM, 44(3), 100749.

14. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. Advances in neural information processing systems, 30.
15. Caesar, H., Bankiti, V., Lang, A. H., Vora, S., Liong, V. E., Xu, Q., ... & Beijbom, O. (2020). nuscenes: A multimodal dataset for autonomous driving. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 11621-11631).
16. Loshchilov, I., & Hutter, F. (2017). Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101.
17. Loshchilov, I., & Hutter, F. (2016). Sgdr: Stochastic gradient descent with warm restarts. arXiv preprint arXiv:1608.03983.
18. Yin, T., Zhou, X., & Krahenbuhl, P. (2021). Center-based 3d object detection and tracking. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 11784-11793).
19. Li, Z., Zhao, X., Bian, J., Liu, B., Li, W., & Zhang, L. (2025, April). V-Fusion: 2D Detection-enhanced Multimodal 3D BEV Object Detection. In ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 1-5). IEEE.
20. Zhou, Z., & Tulsiani, S. (2023). Sparsefusion: Distilling view-conditioned diffusion for 3d reconstruction. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 12588-12597).

Biographies

1. **Yehui Ding** currently pursuing the phd degree in Control Science and Engineering at the School of Electrical and Control Engineering, North China University of Technology, Beijing, China. His research interests focus on environmental perception for autonomous driving, including computer vision, object detection, and multimodal perception.
2. **Yuntao Shi** professor, master's student advisor, and doctoral student advisor at the School of Electrical and Control Engineering, North China University of Technology, Beijing, China. His research interests include cloud computing, the industrial internet, and environmental perception for autonomous driving.

一種用於三維目標檢測的註意力增強多模態融合方法

丁葉輝[1]，史運濤[1]

[1] 北方工業大學，北京，中國，100144

摘要：為了解決自動駕駛中目標檢測精度低的問題，我們提出了一種註意力增強的多模態融合三維目標檢測方法（EA-BEV）。該方法在圖像處理網絡中引入了自註意力機製，能夠有效提取深層特征，減少因語義信息模糊而導致的圖像特征提取不足的問題。在點雲處理網絡中，我們設計了一種高階卷積空間註意力機製，顯著增強了網絡對點雲非線性深層特征的建模和表達能力，從而提高了點雲信息的全局描述能力。我們在nuScenes數據集上進行了對比實驗，結果顯示mAP指標為 76.2%，NDS 指標為 74.4%。EA-BEV方法在三維目標檢測的精度方面表現出明顯的優勢，為自動駕駛的環境感知提供了一種新方法。

關鍵詞：BEV；三維目標檢測；多模態融合

1. 丁葉輝，目前在北方工業大學電氣與控製工程學院攻讀控製科學與工程專業的博士學位。他的研究興趣集中在自動駕駛的環境感知領域，包括計算機視覺、目標檢測以及多模態感知；
2. 史運濤，北方工業大學電氣與控製工程學院的教授，同時擔任碩士生導師和博士生導師。他的研究興趣包括雲計算、工業互聯網以及自動駕駛的環境感知。