International Journal for Housing Science and Its Applications
Publish August 6, 2025. Volume 46, Issue 3 Pages 6293-6304

https://doi.org/10.70517/ijhsa463540

Research on the application of end-to-end deep learning model-based multimodal target detection technique in point cloud data

Niya Dong[1] and Yi Lin[1,*]
[1] College of Communication and Information Engineering, Chongqing College of Mobile Communication, Chongqing, 401520, China
Corresponding authors: (e-mail: cqcyit@163.com).

Abstract Computer vision is an important field in the digital era, and target detection technology plays a key role in it. Traditional methods have accuracy and robustness limitations in complex environments, and point cloud data has gradually become a research hotspot due to its advantage of 3D spatial information. Multimodal deep learning effectively solves the limitation problem of single modality by fusing different data sources, and significantly improves the performance of target detection. In this paper, an end-to-end deep learning model (MANet) based on mutual attention mechanism is proposed to realize the effective fusion of point cloud data and RGB image features for 3D target detection. The point cloud data is first preprocessed with statistical filtering and RANSAC ground segmentation, and then an end-to-end deep learning network composed of four modules: point cloud feature learning, image feature learning, mutual attention feature fusion, and target detection is designed. Through the mutual attention mechanism, the alignment and fusion of point cloud and image features are realized, and the 3D target detection performance is improved. Experiments on the KITTI dataset show that the proposed MANet algorithm achieves 86.13% accuracy on the Car AP 3D metric with medium difficulty, which is a 5.66% improvement over MAFF-Net, and 92.27% on the Car AP BEV metric.Ablation experiments on the Waymo Open dataset demonstrate the effectiveness of the mutual-attention feature fusion to make the 3D mAP of LEVEL_1 to improve from 84.57% to 85.84%. The experimental results show that the proposed multimodal fusion method can effectively improve the accuracy and robustness of 3D target detection, which is of great application value in the fields of autonomous driving and smart city.

Index Terms multimodal target detection, deep learning, point cloud data, mutual attention mechanism, feature fusion, 3D target detection

I. Introduction

With the advent of the digital era, computer vision has become an important field, and many people are researching how to make computers able to perceive, understand, analyze, and process images and video data like human beings [1]-[3]. Among them, target detection technology is one of the most basic technologies in computer vision because it can help computers automatically identify, localize, and track target objects in images, which can not only improve people's quality of life, but also be widely used in industry, medicine, security, and other fields [4]-[7]. At present, end-to-end deep learning model of multimodal target detection technology has become mainstream, which realizes the detection of targets in images by training neural networks, while multimodal deep learning adds more modal information on the basis of retaining the traditional visual information, making the detection effect more accurate and stable, and it has important applications in point cloud data [8]-[11].

Point cloud data is a collection of a large number of points in three-dimensional space, which is used to represent the shape of an object or scene [12]. It is presented with the coordinate information of the points, which can intuitively reflect the geometric characteristics of the spatial target [13]. Point cloud data are often acquired with the help of LiDAR, which scans the collection of points in the surrounding environment, and photogrammetry can also acquire point clouds, and the 3D coordinates of points are calculated from multi-angle images [14]-[16]. 3D scanner is one of the commonly used devices to acquire high precision point cloud data [17]. And the end-to-end deep learning model of multimodal target detection technique can largely improve the accuracy and robustness of 3D environment perception by fusing point cloud and other modal data [18]-[20].

Computer vision has become an important field in the digital era, and many people are researching how to make computers like human beings to be able to perceive, understand, analyze and process images and video data. Target detection technology is one of the most basic technologies in computer vision because it can help computers

automatically identify, localize and track target objects in images, which can not only improve people's quality of life, but also be widely used in industry, medicine, security and other fields. At present, end-to-end deep learning model of multimodal target detection technology has become mainstream, which realizes the detection of targets in images by training neural networks, while multimodal deep learning adds more modal information on the basis of retaining the traditional visual information, which makes the detection effect more accurate and stable, and it has important applications in point cloud data. Point cloud data is a collection of a large number of points in three-dimensional space, which is used to represent the shape of an object or scene. It is presented with the coordinate information of the points, which can intuitively reflect the geometric characteristics of the spatial target. Point cloud data is often obtained with the help of LIDAR, which scans the collection of points in the surrounding environment, and photogrammetry can also be used to obtain the point cloud, and the 3D coordinates of the points are calculated through multi-angle images. 3D scanners are one of the common devices used to acquire high-precision point cloud data. The end-to-end deep learning model for multimodal target detection technology can largely improve the accuracy and robustness of 3D environment sensing by fusing point cloud and other modal data. However, the disorder, rotational invariance, and density inhomogeneity of point cloud data bring challenges to the design and implementation of algorithms. Existing single-modal methods are difficult to fully utilize the geometric properties of point cloud data, resulting in limited detection accuracy and robustness. In addition, how to effectively fuse the complementary information of different modal data is also a problem to be solved. Therefore, the development of a 3D target detection method that can fully utilize multi-modal information has important theoretical value and application significance. This study will focus on solving the following problems: first, how to effectively preprocess the point cloud data to remove noise and irrelevant information; second, how to design a suitable feature extraction network to obtain the deep semantic features of the point cloud and the image; and third, how to realize the effective fusion of multimodal features to make full use of the complementary nature of different modal information. To this end, this paper proposes a multimodal target detection network MANet based on mutual attention, which preprocesses the point cloud data by statistical filtering and RANSAC ground segmentation, extracts the point cloud features using VoxelNet, extracts the image features based on ResNet, and designs the mutual attention module to realize the feature fusion, and finally achieves target classification and 3D box estimation to improve the accuracy and robustness of 3D target detection.

II. 3D point cloud target detection model based on multimodal feature fusion

In order to improve the 3D target detection accuracy by utilizing the image information to assist the point cloud data, this chapter proposes an end-to-end deep learning network for 3D target detection with adaptive fusion of multimodal features based on the preprocessing of point cloud data.

II. A.Point cloud data and its characteristics

II. A. 1) Point cloud data

Point cloud data is a three-dimensional data representation that is typically acquired through sensors such as LIDAR. Based on the measured distance and angle information, the radar detection results can be converted into point cloud data. The coordinates of each point in the point cloud are determined by its distance, bearing, and pitch angle, so the three-dimensional shape of the target can be constructed from the point cloud data.

The point cloud data can be expressed in the form of equation (1):

$$P = \{x_i, y_i, z_i, t_i \mid i = 1, 2, \ldots, N\} \tag{1}$$

The x_i, y_i, z_i in the formula represent the 3D coordinate information of the i th point, and t_i represents the possible other attributes of the i th point, such as the color information, the intensity information, etc., which makes the point cloud data have many advantages over the 2D image data.

II. A. 2) Characterization of Point Cloud Data

3D point cloud data has disorder, rotational invariance and density inhomogeneity.

(1) Point cloud disorder

The disorder of point cloud data refers to the fact that there is no clear order relationship between the points in the point cloud, unlike the pixels in the image data which are arranged in an orderly manner according to a two-dimensional matrix. The acquisition results of point cloud data may be affected by the data acquisition equipment, scanning method, or other factors. As a result, the order of point cloud data is usually random or irregular. Points in a point cloud are discrete, and each point has its own 3D coordinates and incidental other attributes, but the order between these points does not affect the point cloud representation.

(2) Rotational invariance of point clouds

The rotation operation of point cloud data in 3D space does not change the geometric structure and features described by the point cloud data. In other words, if any rotation operation is performed on the point cloud data, its overall shape and structure remain unchanged. For some specific application scenarios, such as point cloud alignment and alignment, it is sometimes necessary to consider the rotation of point cloud data. In these cases, rotational invariance can be used to simplify the processing of the problem and improve the efficiency and robustness of the algorithm.

(3) Density inhomogeneity of point cloud

Unlike image data where pixels are uniformly arranged according to a two-dimensional matrix, the distribution of points in each region of point cloud data is not uniform, i.e., there is a phenomenon that some regions may be dense while others may be sparse. This non-uniformity can be due to a variety of reasons, including the resolution of the acquisition device, the sampling method, the complexity of the object surface, and occlusion.

II. B.Point cloud data preprocessing

II. B. 1) Point Cloud Statistical Filtering Processing

Point cloud data are usually collected by LIDAR, and due to the error of the equipment itself and the influence of environmental factors, the collected data may contain some noise and outliers. These noise and outliers will affect the subsequent point cloud processing and analysis, resulting in false detection of the target, so it is necessary to remove these noise. The noise distribution of outlier points in point cloud data is relatively sparse, and the statistical filter is a filtering method based on the principle of statistics, which can determine whether a point is an outlier by calculating the statistical characteristics of the neighboring points around each point in the point cloud, and remove it from the point cloud.

For each point in the point cloud, statistical filtering will look for k points around it and compute the Euclidean distance between that point and other points. The calculation formula is shown in equation (2):

$$d_{ij} = \frac{1}{n}\sum_{j=1}^{n}\sqrt{(x_i - x_j)^2 + (y_i - y_j)^2 + (z_i - z_j)^2} \tag{2}$$

The obtained inter-point distances are Gaussian distributed, and then the mean μ and the standard deviation σ are obtained, and the calculation of the mean and the standard deviation are shown in Eq. (3) and Eq. (4):

$$\mu = \frac{1}{nk}\sum_{i=1}^{n}\sum_{j=1}^{k} d_{ij} \tag{3}$$

$$\sigma = \sqrt{\frac{1}{(n-1)k}\sum_{i=1}^{n}\sum_{j=1}^{k}(d_{ij} - \mu)^2} \tag{4}$$

Based on the neighborhood statistics information, a distance threshold d_m is calculated to determine whether a point is is a noise point, and if the distance between the point and the key point is greater than this threshold, it is an outlier point that needs to be removed. The threshold is calculated by the formula:

$$d_m = \mu + \alpha\sigma \tag{5}$$

where α is a constant, which can be expressed as a standard deviation multiplier, the size of the value affects the denoising effect. For each point, according to its neighborhood statistical information and threshold value, determine whether it is noise or outliers. If it meets the filtering conditions, it is retained, otherwise it is regarded as a noise or outlier and removed or replaced with one of the values in the neighborhood statistical information. The principle of statistical filtering is shown in Fig. 1, the radius of the center of the circle is the calculated distance threshold, there is one point in the neighborhood of the pl point which is judged to be an outlier, and there are two points in the neighborhood of the p2 point which are judged to be outliers, and these outliers will be removed in the filtering process.

II. B. 2) RANSAC ground point cloud data segmentation

Removing the ground point cloud and then detecting the obstacles can avoid the problem of ground irrelevant point cloud interfering with the detection. Therefore the segmentation of ground point cloud is very important in 3D detection. In this paper, we use the Randomized Sample Consistent (RANSAC) algorithm [21] to remove the ground point cloud.

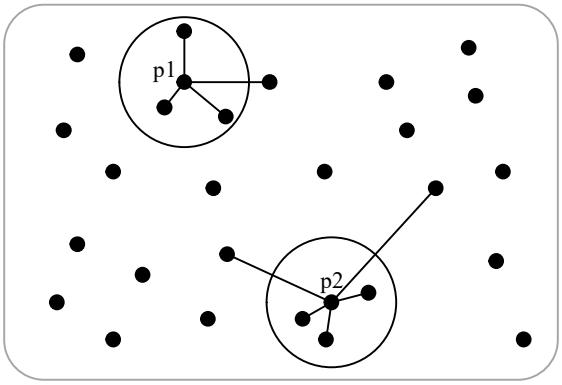

Figure 1: Schematic diagram of Statistical filtering

The RANSAC algorithm divides the data into interior and exterior points, the interior points are the data points that conform to the fitted model, i.e., those that are close to the results of the model fitting. Outpoints are data points that do not conform to the results of the fitted model, usually noise or anomalies. In the RANSAC algorithm, outlier points are excluded from the fitting process to minimize their impact on the model fit. The advantage is the ability to identify results quickly and accurately and with some robustness in samples that contain erroneous data.

The steps for ground segmentation of point cloud data using RANSAC algorithm are as follows:

(1) For the three points $p_1=(x_1,y_1,z_1)$, $p_2=(x_2,y_2,z_2)$, $p_3=(x_3,y_3,z_3)$ form these 3 points into a plane $Ax+By+Cz+D=0$ and use Eq. (6) to compute the normal vector $\vec{n}$ of the plane:

$$\vec{n}=(p_2-p_1)\times(p_3-p_1) \tag{6}$$

(2) Calculate the distance from a point to a plane in a point cloud:

$$distance_i=\frac{\vec{n}^T(p_i-p_1)}{\|\vec{n}\|} \tag{7}$$

(3) Set the threshold value τ and compare it with the $distance$ of each point calculated in Equation (7), if $distance>\tau$, the point is an outer point, and the points with $distance<\tau$ are recognized as inner points, and then calculate the ratio of the resulting number of inner points to the whole point cloud data ω, as shown in Equation (8), n_{in} represents the number of inner points and n_{out} represents the number of outer points:

$$\omega=\frac{n_{in}}{n_{in}+n_{out}} \tag{8}$$

(4) Repeat the above operation to reach the iteration threshold T through continuous iteration, and then find out the parametric model with the highest number of interior points. Denote by p the probability that a point randomly selected from the data set belongs to the inner point. Assuming that the model selects a total of n points, the probability that all n points belong to the inner point is ω^n, and the probability that at least one of them belongs to the outer point is $1-\omega^n$, which can be shown that $(1-\omega^n)T$ denotes the probability of never picking n points that are all inside points, and is equal to $1-p$, so that equation (9) can be obtained:

$$(1-\omega^n)T=1-p \tag{9}$$

Taking the logarithm of Eq. (9) yields Eq. (10), obtaining the iteration threshold T:

$$T=\frac{\log(1-p)}{\log(1-\omega^n)} \tag{10}$$

The RANSAC ground filtering algorithm mainly contains two core parameters: the distance threshold τ and the number of iterations T. Among them, the distance threshold is especially critical, which directly determines the accuracy of extracting valid information from the original data set. If it is set too strictly, points that originally belong to the set of interior points may be wrongly eliminated. If it is set too loosely, the outer points may be misjudged as inner points, thus affecting the accuracy of the algorithm. The number of iterations T is mainly dependent on the amount of data in the original 3D point cloud and is closely related to the operation efficiency of the algorithm.

Theoretically, the size of T is proportional to the final result of extracting the plane, but too large a value of T will lead to a decrease in the running speed of the algorithm and affect the real-time segmentation.

II. C.3D target detection model construction with multimodal fusion

In this paper, we propose a deep neural network MANet based on mutual attention, which aims to align and fuse the features of 2D RGB images and 3D point clouds in the feature learning stage for more effective 3D target detection.

II. C. 1) Deep Neural Network for 3D Target Detection

The 3D target fusion detection deep learning network is shown in Fig. 2, which mainly includes point cloud feature learning, image feature learning, mutual attention feature fusion, and target detection modules. In the feature learning stage, point cloud features are extracted based on VoxelNet [22] and image features are extracted based on ResNet [23]. In the multimodal feature fusion stage, mutual attention is calculated based on feature correlation, and the effect of fusing multimodal features is achieved through feature correction. Finally, target classification and 3D box estimation are realized based on Regional Proposal Network (RPN) [24] and Classification Regression Multi-Task Learning Network.

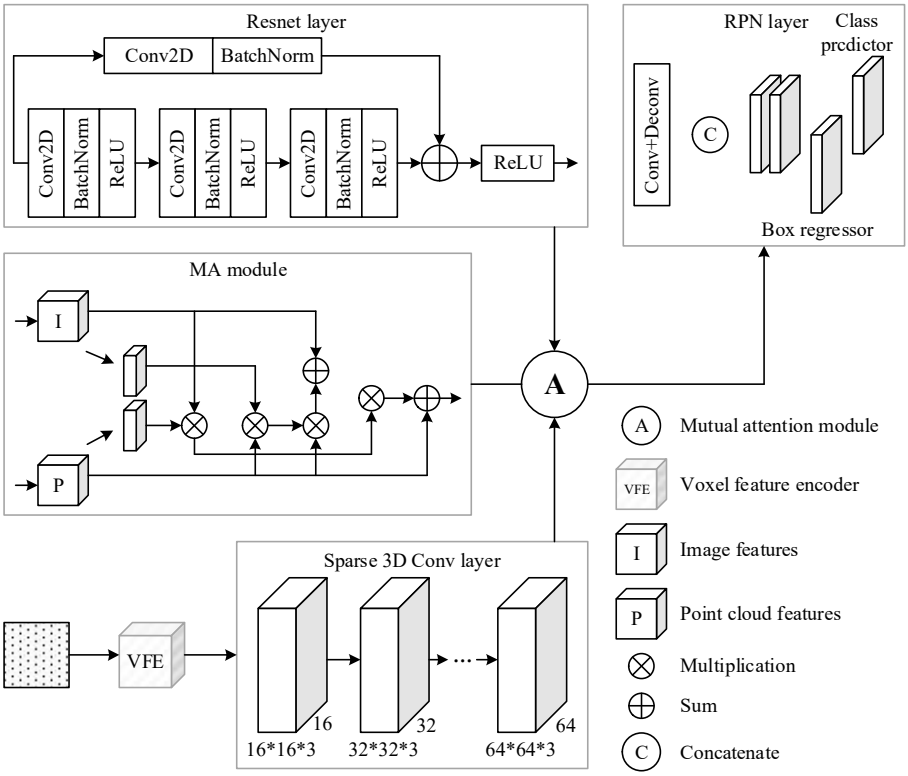

Figure 2: The structure of 3D object detection neural network

II. C. 2) Feature learning

Feature learning consists of two branches: the Lidar branch and the camera branch, which extract point cloud features and image features, respectively. For the Lidar branch, it is assumed that a 3D object contains N points, and the point cloud data of the object is represented as $\{x_i, y_i, z_i, r_i\}$, where x_i, y_i, z_i is the spatial coordinate of the i rd point and r_i is the size of the corresponding reflection value of the point. Similar to VoxelNet, the point cloud data sampled by LIDAR is first converted to a 3D voxel grid, where a certain number of point clouds are selected by random sampling, and the point cloud features are learned by a forward neural network to obtain the voxel features based on the maximum pooling operation, and then the features corresponding to the point cloud data are learned by the 3D sparse convolutional neural network. For the camera branch, image features are extracted using Resnet2D convolutional neural network.

II. C. 3) Mutual Attention Feature Fusion Module

Inspired by the concept of mutual correlation in signal processing and the attention mechanism in the field of machine learning, the mutual attention module is designed for the information fusion of features between different modalities.

The point cloud features and image features are denoted as G_p and G_i, respectively, and the steps for the fusion of the two features are as follows:

(1) Calculate the correlation degree value between G_p and G_i:

$$R_i = \tanh(G_p \cdot (W_i G_i + b_i)) \tag{11}$$

$$R_p = \tanh(G_i \cdot (W_p G_p + b_p)) \tag{12}$$

In Eqs. (11) and (12), in order to realize the alignment of the two modal features at the eigenspace level, the eigenspace transformations of G_p and G_i are carried out by multiplying them by the matrices W_p and W_i, respectively, plus the bias vectors b_i and b_p The feature space translation is achieved so that the point cloud feature space and the image feature space are aligned by learning to optimize the values of W_p, W_i, b_i and b_p. The two correlation values are then transformed using the tanh function to obtain the values of the cross-correlation function between G_p and G_i, R_i and R_p, respectively.

(2) Use the obtained correlation function values to compute the attention scores for each component in the point cloud features and each component in the image features:

$$A_i = softmax(R_i) \tag{13}$$

$$A_p = softmax(R_p) \tag{14}$$

A_i and A_p are the mutual attention scores between point cloud features and image features. Numerical transformation of the correlation function values using $softmax$ allows normalization on the one hand, transforming the original correlation scores into a probability distribution in which the sum of the weights of all elements is 1. On the other hand, through the function transformation mechanism inherent in $softmax$ to highlight more important relevant attention scores, amplifying the saliency feature weights will be more conducive to the extraction of deep semantic features.

(3) Multiply the attention weights and the feature vector matrix to get the image correction matrix after updating by the attention mechanism:

$$C_i = G_p \cdot A_i \tag{15}$$

The image features corrected by the point cloud information are:

$$G_i = G_i + C_i \tag{16}$$

(4) Use the corrected image features G_i and the attention matrix to obtain the correction matrix for the point cloud features:

$$C_p = G_i \cdot A_p \tag{17}$$

The point cloud features corrected by the image features based on the attention mechanism are:

$$G_p = G_p + C_p \tag{18}$$

G_p is the feature that fuses the image information and the point cloud information to provide the feature base for the subsequent 3D target detection. In the above feature fusion process, $\cdot$ denotes the matrix dot product, and W_p, W_i, b_i and b_p are the parameters that need to be optimized, which is achieved by the optimization of the loss function in the subsequent target detection network to optimize W_p, W_i, b_i and b_p parameter optimization.

II. C. 4) Classification and regression multitasking networks

In this step, an RPN architecture is constructed for 3Dbox estimation. The architecture of this RPN consists of three stages, each combining a convolutional layer, batch regularization and a linear rectifier unit layer (ReLU). The fused features are fed into this network and the output of each stage of this part is upsampled to a feature map with the same dimensions. Next, the feature maps of these three stages are stitched together into a single feature map. Finally, a convolutional layer using three 1×1 sized convolutional kernels predicts the category, offset, and

orientation. In the last layer of this network, a non-great suppression layer is added to generate a 3D outsourced rectangular box of the final detected object.

II. C. 5) Loss function

In order to perform the 3D target classification and 3Dbox localization tasks simultaneously, the loss function is defined as the sum of the classification and regression loss functions, i.e., the multi-task loss function, as shown in equation (19) below:

$$L_{total} = L_{cls} + L_{reg} \tag{19}$$

where L_{ck} and L_{reg} represent classification loss and regression loss, respectively, and L_{cls} is used with the FocalLoss function shown in equation (20):

$$L_{cls} = -\alpha_t (1 - p_t)^{\gamma} \log(p_t) \tag{20}$$

where p_t is the class probability estimated by the model. α_t and γ are the parameters of FocalLoss, and α_t and γ are set to 0.3 and 3, respectively, during training.

For 3D target detection, L_{reg} consists of 7 parameters, and the real 3Dbox is represented as $(x_c^g, y_c^g, z_c^g, l^g, w^g, h^g, \theta^g)$, where (x_c^g, y_c^g, z_c^g) denotes the center point of the 3Dbox, (l^g, w^g, h^g) denotes the length, width, and height of the 3Dbox, and θ^g is the orientation angle of the 3D object. For the anchor box of positive samples, this paper parameterizes it as $(x_c^a, y_c^a, z_c^a, l^a, w^{\alpha}, h^{\alpha}, \theta^{\alpha})$. Define the residual vector $\tau^* \in R^7$, the vector τ^* contains the seven target parameters to be regressed, which are denoted as the positional residuals $(\Delta x, \Delta y, \Delta z)$, the residuals of the length, width and height in the three dimensions $(\Delta l, \Delta w, \Delta h)$, and orientation angle residuals $\Delta\theta$, which are calculated by the following equation:

$$\begin{aligned} \Delta x &= \frac{x_c^g - x_c^a}{d^a} \\ \Delta y &= \frac{y_c^g - y_c^a}{d^a} \\ \Delta z &= \frac{z_c^g - z_c^a}{h^a} \end{aligned} \tag{21}$$

$$\begin{aligned} \Delta l &= \log\left(\frac{l^g}{l^a}\right) \\ \Delta w &= \log\left(\frac{w^g}{w^a}\right) \\ \Delta h &= \log\left(\frac{h^g}{h^a}\right) \end{aligned} \tag{22}$$

$$\Delta\theta = \theta^g - \theta^a \tag{23}$$

where $d^a = \sqrt{(l^a)^2 + (w^a)^2}$ is the length of the diagonal in the horizontal plane of any ANCHOR. In order to correctly retrieve the parameters of the real 3Dbox from the positive samples to be matched, the regression loss is calculated using the following equation:

$$L_{reg} = \frac{1}{N_{pos}} \sum_i SmoothL1(\tau_i^*) \tag{24}$$

In Eq. (24), τ_i^* denotes the vector of residuals between the i th real 3Dbox and the 3Dbox predicted by the model. Where the $SmoothL1$ function is calculated as follows:

$$SmoothL1(\tau_i^*) = \begin{cases} 0.5(\tau_i^*)^2, if \mid \tau_i^* \mid < 1 \\ \tau_i^* - 0.5, otherwise \end{cases} \tag{25}$$

III. Multimodal target detection experiment based on point cloud data

In order to verify the effectiveness and correctness of the proposed multimodal target detection method, this paper conducts a large number of experiments on the KITTI dataset, compares and analyzes it with a variety of more popular algorithms, and explores the validity of the modules in the model through ablation experiments.

III. A. Data sets

In this paper, the algorithm performance is evaluated using the KITTI dataset, which is the largest dataset in the world for evaluating algorithms in autonomous driving scenarios, including point clouds and images of three categories: cars, pedestrians and cyclists. For each category, the detection results are evaluated according to three difficulty levels: easy, medium, and difficult, which are determined based on target size, occlusion status, and truncation level, respectively. The algorithm is comprehensively evaluated and the training data is subdivided into training set and validation set to obtain 3824 data samples for training and 3875 data samples for validation. After segmentation, samples of the same sequence are not included in both training and validation sets.

III. B. Evaluation indicators

In this paper, based on the official evaluation protocol provided by KITTI, the algorithm of this paper and various existing algorithms are compared and experimented on the validation set. The protocol requires that target detection needs to complete the work of target category judgment and location selection at the same time.

Among them, the judgment of target localization is based on judging whether the degree of overlap between the prediction frame and the truth frame, i.e., the intersection and merger ratio (IoU), reaches a certain threshold, if the intersection and merger ratio is greater than the threshold, it is considered that the localization is accurate, and vice versa, it is considered that the localization is incorrect. The judgment of target classification is based on whether the classification confidence reaches a certain determined threshold, if it is greater than the threshold, the classification is considered accurate, and vice versa, the classification is considered incorrect.

The final target detection correctness is determined by combining the judgments of target localization and classification correctness, converting the problem of detecting multi-category targets into a binary classification problem so that a confusion matrix can be constructed, and evaluating the model accuracy using a series of metrics for target classification. Experiments were conducted to compare the algorithms using the average precision (AP) metric, which is the mean value of precision at different recall rates, and the mean average precision (mAP).

III. C. Experiments on point cloud random sampling threshold analysis

In order to analyze the impact of point cloud random sampling thresholds on the performance of the algorithm, this paper carries out controlled experiments on this paper's algorithm applying different sampling thresholds (T) at three difficulty levels of the automotive category of the KITTI validation set, and the experiments set up five kinds of sampling thresholds: 15, 30, 40, 50, and 60, and use the average accuracy index to measure the algorithm's accuracy, and record the average time the algorithm spends on each detection. The experimental results for different point cloud sampling thresholds are shown in Table 1.

It can be seen that the time overhead of the algorithm rises as the random sampling threshold increases, because a higher sampling threshold means that the network needs to process more points. As the random sampling threshold in the algorithm increases from 15 to 40, the average accuracy of target detection is significantly improved. When the threshold is raised from 40 to 50, the detection accuracy shows a decrease in easy difficulty and only a small increase in medium and hard difficulty. And when the threshold was raised from 50 to 60, the detection accuracy showed a decrease. The experimental results show that too low a sampling threshold can lead to a point cloud that is too sparse, missing spatial information and affecting the detection accuracy. Too high a sampling threshold will cause the algorithm to be affected by the uneven density of the point cloud data, and will lead to higher computational overhead. Considering the detection speed and accuracy, this paper determines the random sampling threshold of the point cloud as 40 in the subsequent experiments.

Table 1: Experimental results of cloud sampling thresholds at different points

Point cloud sampling threshold	Time cost /ms	Detection accuracy /%		
		Simple	Medium	Difficult
15	118	64.87	59.26	53.64
30	185	79.43	64.51	60.07
40	232	83.86	68.75	65.41
50	269	82.05	69.42	66.73
60	284	81.32	67.31	64.58

III. D. Experiments on density analysis of voxel divisions

In order to analyze the effect of voxel division density on the detection effect of the algorithm, this paper sets up three different voxel division setting schemes:

(1) D'=10, W'=250, H'=100.

(2) D'=20, W'=500, H'=200.

(3) D'=30, W'=100, H'=400.

where D', W', and H' are the number of voxels contained in the Z, Y, and X-axis direction dimensions of the point cloud after division, respectively.

Comparison experiments of the algorithms using the three grouping methods were conducted on all difficulty levels of the three categories on the KITTI dataset, and the results of the experiments at different voxel dimensions are shown in Table 2.

Analyzing the data shows that the second division achieved the best detection accuracy on all difficulty levels of the three categories. The experimental results show that too sparse division will ignore the local information of the input data, while too dense division will ignore the connection between features due to too much focus on localization. In the subsequent experiments, this algorithm will adopt the second division method.

Table 2: Experimental results at different voxel sizes

Voxel division	Detection accuracy /%								
	Automobile Category			Pedestrian category			Bicycle category		
D'×W'×H'	Simple	Medium	Difficult	Simple	Medium	Difficult	Simple	Medium	Difficult
10*250*100	80.51	65.36	63.21	57.52	53.74	49.20	65.57	47.35	46.28
20*500*200	84.83	69.59	66.43	60.84	57.62	52.28	70.54	50.46	48.19
30*1000*400	80.72	66.14	63.45	57.93	53.48	59.56	66.31	48.12	46.74

III. E. Experiments for quantitative evaluation of model performance

This section aims to evaluate the performance and effectiveness of the proposed MANet-based multimodal target detection algorithm through quantitative analysis.

In the experiments based on the KITTI dataset, the dataset is first divided into dataset and validation set on a 1:1 basis for validation experiments. The performance comparison of this paper's algorithm with other algorithms performed on the KITTI validation set is shown in Table 3.

From the analysis, it can be concluded that the performance of the proposed MANet algorithm on the KITTI validation set is improved compared to the performance of the representative algorithms. In the Car AP 3D medium standard evaluation metric, the MANet algorithm achieves 86.13% accuracy, which is an improvement of 5.66% compared to the advanced MAFF-Net algorithm. The MANet algorithm also achieved a very good performance in the Car AP BEV metric, with an accuracy of 96.16%.

Table 3: Performance evaluation of representative algorithms and the algorithm in this paper

Algorithm	Car AP 3D /%			Car AP BEV /%		
	Simple	Medium	Difficult	Simple	Medium	Difficult
VoxelNet	82.25	66.43	62.75	-	-	-
Frustum PointNets	84.84	72.92	64.39	-	-	-
SECOND	88.07	77.21	70.05	-	-	-
Point-GNN	89.04	78.37	78.59	91.27	88.2	88.83
PointRCNN	89.64	79.91	78.88	-	-	-
Frustum ConvNet	89.23	79.47	78.54	91.75	89.52	87.68
MAFF-Net	90.57	80.47	76.31	94.82	90.98	87.38
MANet (Ours)	93.69	86.13	83.54	96.16	92.27	90.11

Meanwhile, the dataset is divided into dataset and validation set to train the model in the ratio of 8:2, and the model is submitted to the KITTI official website to test the detection performance of the algorithm on the KITTI training set. In this section of the algorithm, the detection is mainly performed from three categories: cars, pedestrians and bicycles, and there are four detection criteria for each category, which are: 2D detection accuracy, orientation accuracy, 3D detection accuracy and BEV detection accuracy. The performance evaluation results of

the algorithm on the KITTI training set are shown in Table 4. It can be seen that the algorithm in this paper achieves superior detection results.

Table 4: Performance evaluation of KITTI training set for MANet algorithm

Testing standard	Simple	Medium	Difficult
Car (Detection)	98.45%	95.51%	93.72%
Car (Orientation)	98.47%	96.05%	93.57%
Car (3D Detection)	91.74%	83.35%	78.04%
Car (Bird's Eye View)	94.77%	91.68%	87.19%
Pedestrian (Detection)	60.64%	54.70%	52.76%
Pedestrian (Orientation)	58.08%	50.19%	48.90%
Pedestrian (3D Detection)	44.05%	38.40%	38.22%
Pedestrian (Bird's Eye View)	49.50%	44.33%	42.59%
Cyclist (Detection)	80.22%	71.84%	67.42%
Cyclist (Orientation)	79.76%	73.02%	66.52%
Cyclist (3D Detection)	74.73%	60.87%	57.17%
Cyclist (Bird's Eye View)	77.67%	64.75%	59.95%

In order to validate the effectiveness of the algorithm, in addition to evaluating it using the detection criteria described above, the category Car results on the training set of the algorithm were compared with other state-of-the-art algorithms. This allows for a more comprehensive assessment of the algorithm's performance in the target detection task. The results of the algorithm performance comparison are shown in Table 5.

Based on the comparison results, it can be concluded that the multimodal target detection algorithm MANet in this paper has a significant improvement in AP 3D and AP BEV metrics compared to the state-of-the-art algorithms, which proves the advantage of the algorithm. Compared with the state-of-the-art fusion algorithm EPNet, the Car AP 3D metrics are improved by 0.63%, 2.38%, and 2.56% at three different detection difficulties: simple, medium, and difficult, respectively, which proves that the algorithm's target detection accuracy is significantly improved at different difficulty levels. On the Car AP BEV metric, the MANet algorithm improves 0.27%, 1.94%, and 2.19% relative to EPNet under three different detection difficulties: simple, medium, and difficult, respectively, demonstrating that the algorithm also significantly improves the detection accuracy under the BEV perspective for the automobile category.

Table 5: Comparison results of algorithm performance

Algorithm	Data mode	AP 3D /%			AP BEV /%		
		Simple	Medium	Difficult	Simple	Medium	Difficult
MV3D	L&R	75.84	64.87	54.56	87.54	79.91	70.56
FrustumPointNets	L&R	82.12	71.41	63.42	89.96	85.12	76.52
AVOD	L&R	82.91	72.47	67.26	89.33	84.83	78.69
MMF	L&R	87.82	77.83	69.47	-	-	-
MVAF-Net	L&R	89.22	79.66	76.42	92.95	88.97	85.73
CLOCs	L&R	89.68	81.94	78.04	94.08	90.74	87.92
Fast-CLOCs	L&R	89.72	81.54	78.19	93.85	90.71	87.51
3D-CVF	L&R	90.02	81.14	73.86	94.95	90.38	83.52
EPNet	L&R	90.69	80.25	75.79	95.04	89.17	84.67
MANet (Ours)	L&R	91.32	82.63	78.35	95.31	91.11	86.86

III. F. Model ablation experiments

To further demonstrate the effectiveness of the proposed multimodal target detection model, this section conducts ablation experiments on the Waymo Open dataset, which is one of the largest and most comprehensive open datasets for autonomous driving. The dataset is collected from Waymo's self-driving fleet while driving on a variety of city streets, suburban roads, and highways, and has a total of about 12 million 3D labels and 10 million 2D labels for four types of targets: vehicles, pedestrians, signs, and bicycles.

In the ablation experiments, this paper explores the effects of the point cloud feature learning and image feature learning modules (Module 1) and the mutual attention feature fusion module (Module 2) on the performance of 3D

target detection, and a comparison of the results of the ablation experiments is shown in Fig. 3. The metrics of the ablation experiment results are shown in Table 6.

The experimental results show that the 3D mAP is 84.57% and 75.64% for LEVEL_1 and LEVEL_2, while the 3D mAPH is 82.72% and 74.57% for LEVEL_1 and LEVEL_2, respectively, in the case of modeling with only point cloud features and image features. Even without the mutual attention feature fusion module, the model can effectively improve the accuracy of target detection by deeply learning and fusing different point cloud features and image features, but it is not yet the best performance. When modeled with only inter-attentive features, the 3D mAP of LEVEL_1 and LEVEL_2 decreases to 83.49% and 74.35%, respectively, and the 3D mAPH decreases to 81.16% and 72.83% accordingly. In comparison, the performance decreases but still demonstrates the effectiveness of the mutual attention feature fusion module in capturing the interrelationships between points and the spatial structure information in point cloud data. The model performs best when both the point cloud feature learning and image feature learning modules and the mutual attention feature fusion module are used, with the 3D mAP improving to 85.84% and 76.41% for LEVEL_1 and LEVEL_2, respectively, while the 3D mAPH improves to 84.28% and 76.35%, respectively. The results clearly demonstrate the significant impact of the two modules on improving the overall performance of the model when they are combined, indicating that they complement each other in improving the target detection accuracy, and effectively enhance the recognition capability of the model through deep feature fusion and relational modeling.

Table 6: Ablation experiment results of the Waymo dataset

Module		LEVEL_1		LEVEL_2	
Module 1	Module 2	3D mAP	3D mAPH	3D mAP	3D mAPH
√	×	84.57	82.72	75.64	74.57
×	√	83.49	81.16	74.35	72.83
√	√	85.84	84.28	76.41	76.35

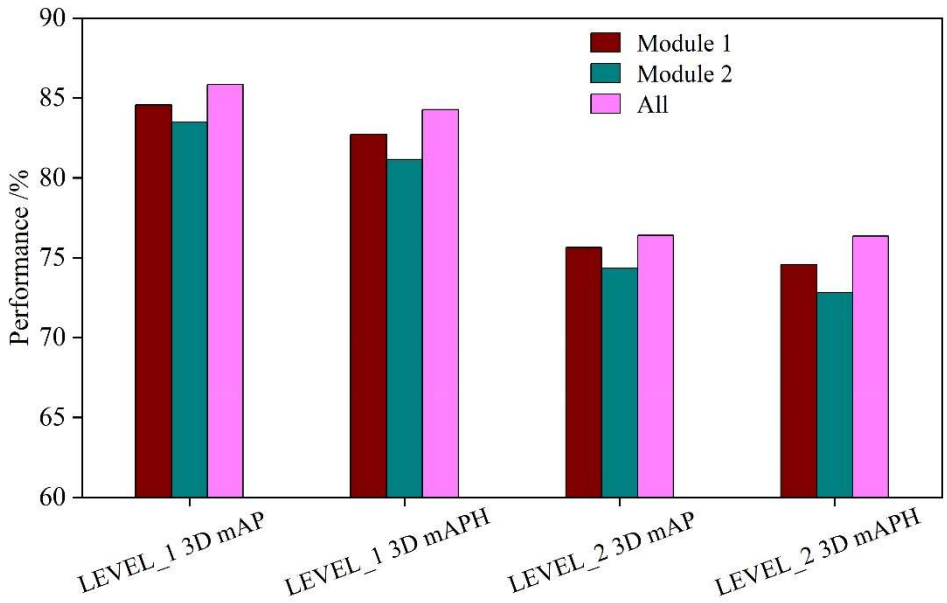

Figure 3: Comparison of ablation experiments

IV. Conclusion

The multimodal target detection technique effectively solves the limitations of single-modal methods in complex scenes by fusing the complementary information provided by different data sources. In this paper, MANet, a multimodal target detection network based on the mutual attention mechanism, is proposed to provide a new idea for 3D target detection in point cloud data. MANet realizes the effective fusion of point cloud features and image features by designing the mutual attention module, which makes full use of the advantages of the two modalities. Experiments on the KITTI validation set show that MANet achieves 86.13% detection accuracy on the Car AP 3D metrics with moderate difficulty, which is 5.66% better than the advanced MAFF-Net algorithm; and achieves 96.16% accuracy on the Car AP BEV metrics, which demonstrates superior performance. Compared with the current state-

of-the-art fusion algorithm EPNet, MANet improves the Car AP 3D metrics on the KITTI training set by 0.63%, 2.38%, and 2.56% at three different difficulties, namely easy, medium, and hard, respectively. Ablation experiments on the Waymo Open dataset further demonstrate the effectiveness of the inter-attentive feature fusion module, which improves the 3D mAP of LEVEL_1 from 84.57% to 85.84% when both point cloud feature learning and inter-attentive feature fusion module are used. The experimental analysis of point cloud random sampling threshold and voxel division density shows that reasonable parameter settings have a significant impact on model performance. Future work will further optimize the algorithm structure to improve the computational efficiency and explore more flexible and efficient multimodal fusion strategies to adapt to more complex and changing application scenarios.

References

[1] Paneru, S., & Jeelani, I. (2021). Computer vision applications in construction: Current state, opportunities & challenges. Automation in Construction, 132, 103940.

[2] Gao, J., Yang, Y., Lin, P., & Park, D. S. (2018). Computer vision in healthcare applications. Journal of healthcare engineering, 2018, 5157020.

[3] Khan, A. I., & Al-Habsi, S. (2020). Machine learning in computer vision. Procedia Computer Science, 167, 1444-1451.

[4] Liang, F., Zhou, Y., Chen, X., Liu, F., Zhang, C., & Wu, X. (2021, January). Review of target detection technology based on deep learning. In Proceedings of the 5th international conference on control engineering and artificial intelligence (pp. 132-135).

[5] Yue, X., Wang, Q., He, L., Li, Y., & Tang, D. (2022). Research on tiny target detection technology of fabric defects based on improved YOLO. Applied Sciences, 12(13), 6823.

[6] Rawat, S. S., Verma, S. K., & Kumar, Y. (2020). Review on recent development in infrared small target detection algorithms. Procedia Computer Science, 167, 2496-2505.

[7] Jiang, W., Ren, Y., Liu, Y., & Leng, J. (2022). Artificial neural networks and deep learning techniques applied to radar target detection: A review. Electronics, 11(1), 156.

[8] Vink, J. P., & de Haan, G. (2015). Comparison of machine learning techniques for target detection. Artificial Intelligence Review, 43, 125-139.

[9] Iftikhar, S., Asim, M., Zhang, Z., Muthanna, A., Chen, J., El-Affendi, M., ... & Abd El-Latif, A. A. (2023). Target detection and recognition for traffic congestion in smart cities using deep learning-enabled UAVs: A review and analysis. Applied sciences, 13(6), 3995.

[10] Wang, J., Zhang, T., & Cheng, Y. (2021). Deep Learning for Object Detection: A Survey. Computer Systems Science & Engineering, 38(2).

[11] Tahir, A., Munawar, H. S., Akram, J., Adil, M., Ali, S., Kouzani, A. Z., & Mahmud, M. P. (2022). Automatic target detection from satellite imagery using machine learning. Sensors, 22(3), 1147.

[12] Woo, H., Kang, E., Wang, S., & Lee, K. H. (2002). A new segmentation method for point cloud data. International Journal of Machine Tools and Manufacture, 42(2), 167-178.

[13] Wang, Q., Tan, Y., & Mei, Z. (2020). Computational methods of acquisition and processing of 3D point cloud data for construction applications. Archives of computational methods in engineering, 27(2), 479-499.

[14] Goodbody, T. R., Coops, N. C., Tompalski, P., Crawford, P., & Day, K. J. (2017). Updating residual stem volume estimates using ALS-and UAV-acquired stereo-photogrammetric point clouds. International journal of remote sensing, 38(8-10), 2938-2953.

[15] Carlsson, G. (2014). Topological pattern recognition for point cloud data. Acta Numerica, 23, 289-368. Moyano, J., Nieto-Julián, J. E., Lenin, L. M., & Bruno, S. (2022). Operability of point cloud data in an architectural heritage information model. International Journal of Architectural Heritage, 16(10), 1588-1607.

[16] Cura, R., Perret, J., & Paparoditis, N. (2017). A scalable and multi-purpose point cloud server (PCS) for easier and faster point cloud data management and processing. ISPRS Journal of Photogrammetry and Remote Sensing, 127, 39-56.

[17] Wang, J., Zhang, J., & Xu, Q. (2014, July). Research on 3D laser scanning technology based on point cloud data acquisition. In 2014 International Conference on Audio, Language and Image Processing (pp. 631-634). IEEE.

[18] Mirzaei, K., Arashpour, M., Asadi, E., Masoumi, H., Bai, Y., & Behnood, A. (2022). 3D point cloud data processing with machine learning for construction and infrastructure applications: A comprehensive review. Advanced Engineering Informatics, 51, 101501.

[19] Huang, S., Liu, L., Fu, X., Dong, J., Huang, F., & Lang, P. (2022). Overview of LiDAR point cloud target detection methods based on deep learning. Sensor Review, 42(5), 485-502.

[20] Bello, S. A., Yu, S., Wang, C., Adam, J. M., & Li, J. (2020). Deep learning on 3D point clouds. Remote Sensing, 12(11), 1729.

[21] Qian Sun, Ming Diao, Yibing Li & Ya Zhang. (2017). An improved binocular visual odometry algorithm based on the Random Sample Consensus in visual navigation systems. The Industrial Robot,44(4),542-551.

[22] Vishwanath A. Sindagi, Yin Zhou & Oncel Tuzel. (2019). MVX-Net: Multimodal VoxelNet for 3D Object Detection. CoRR,abs/1904.01649.

[23] Yan Li, Chunping Li, Tingting Zhu, Shurong Zhang, Li Liu & Zhanpeng Guan. (2025). A recognition model for winter peach fruits based on improved ResNet and multi-scale feature fusion. Frontiers in Plant Science,16,1545216-1545216.

[24] Li Yuezun, Chang Ming-Ching, Sun Pu, Qi Honggang, Dong Junyu & Lyu Siwei. (2021). TransRPN: Towards the Transferable Adversarial Perturbations using Region Proposal Networks and Beyond. Computer Vision and Image Understanding,213.