EURASIP Journal on Advances in Signal Processing

RESEARCH

Open Access

CenterTransFuser: radar point cloud and visual information fusion for 3D object detection

Yan Li, Kai Zeng and Tao Shen*

*Correspondence: shentao@kmust.edu.cn

School of Information Engineering and Automation, Kunming University of Science and Technology, Kunming, China

Abstract

Sensor fusion is an important component of the perception system in autonomous driving, and the fusion of radar point cloud information and camera visual information can improve the perception capability of autonomous vehicles. However, most of the existing studies ignore the extraction of local neighborhood information and only consider shallow fusion between the two modalities based on the extracted global information, which cannot perform a deep fusion of cross-modal contextual information interaction. Meanwhile, in data preprocessing, the noise in radar data is usually only filtered by the depth information derived from image feature prediction, and such methods affect the accuracy of radar branching to generate regions of interest and cannot effectively filter out irrelevant information of radar points. This paper proposes the CenterTransFuser model that makes full use of millimeter-wave radar point cloud information and visual information to enable cross-modal fusion of the two heterogeneous information. Specifically, a new interaction called cross-transformer is explored, which cooperatively exploits cross-modal cross-multiple attention and joint cross-multiple attention to mine radar and image complementary information. Meanwhile, an adaptive depth thresholding filtering method is designed to reduce the noise of radar modality-independent information projected onto the image. The CenterTransFuser model is evaluated on the challenging nuScenes dataset, and it achieves excellent performance. Particularly, the detection accuracy is significantly improved for pedestrians, motorcycles, and bicycles, showing the superiority and effectiveness of the proposed model.

Keywords: Cross-transformer, Depth threshold filtering, 3D detection, Cross-modal fusion, Contextual interaction

1 Introduction

In recent years, autonomous driving technology has developed rapidly, and a single sensor with a small sensing range, little scene information, and a single processing method cannot adapt to the complex real-world environment [1, 2] and cannot effectively assist the car in environmental perception. Multi-sensor fusion enables the vehicle to perceive the surrounding environment and make decisions using 3D coordinates, depth, direction, speed, and other information about the perceived object. Vehicles are usually

Springer Open

© The Author(s) 2023. **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

equipped with multiple types of sensors, among which cameras have high resolution and can provide visual cues of the object to determine its location and category. However, cameras can only provide 2D information to the vehicle, and their 3D perception capability is limited [3]. Current studies such as [2, 4, 5] exploit LIDAR and visual information for 3D detection, but the detection accuracy is reduced because both sensors have poor stability in harsh weather environments and have difficulty in capturing long-range targets [6]. In contrast, millimeter-wave radar is less affected by extreme weather conditions [7, 8], has higher stability, and is relatively robust, inexpensive, and simple to maintain. Figure 1a indicates that the texture information of the camera is blurred and susceptible to the images in the night, rain, and other environments; Fig. 1b indicates that the radar points projected onto the images are largely unaffected by weather conditions and have better stability. However, millimeter-wave radar is not effective in sensing objects at high locations and objects such as pedestrians and bicycles due to the restricted vertical angle [9]. As a result, it is difficult to obtain context-aware information to directly detect the contours of objects. The detection capabilities of vision sensors and millimeter-wave radars can complement each other, and the detection algorithms based on millimeter-wave radar and vision fusion can significantly improve the perception capabilities of autonomous vehicles [10, 11], thus achieving environmental perception in complex scenarios.

Although 3D detection based on the fusion of millimeter-wave radar point cloud information and image vision information can improve vehicle perception of the environment, there are problems in data processing and fusion methods. Previous studies addressed the limited field-of-view problem of the data by stretching the radar points longitudinally, but they cannot capture lateral information. During the mapping process, the loss of radar information due to too dense instances and occlusion makes the radar information and the image object information inconsistent and leads to irrelevant information on the depth value mapping to the image [11–13]. Meanwhile, since radar features and visual features are not homogeneous, signal extraction from local domains

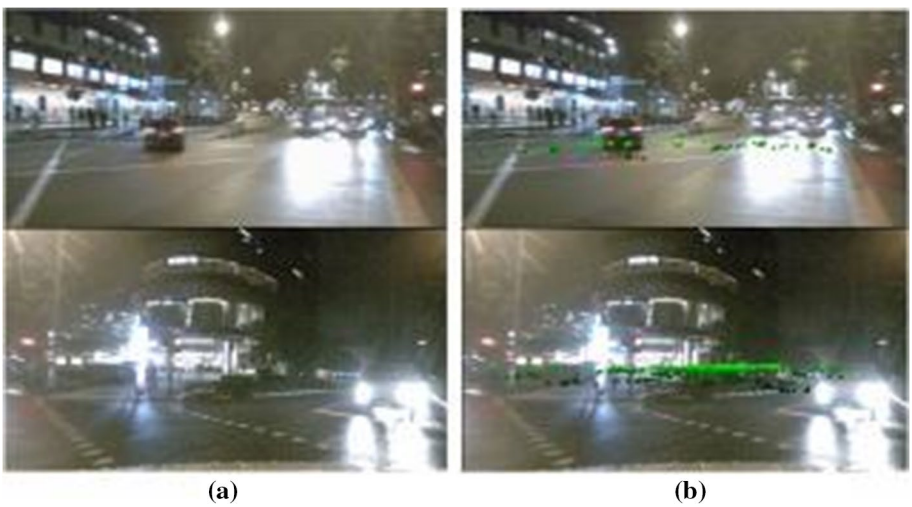

Fig. 1 Millimeter-wave radar and camera performance in diverse weather conditions, where **a** indicates the camera effect map at night and during rain; and **b** indicates the effect map of radar points onto the image

is ignored. For local information, generic object detectors usually ignore the contextual features of the local area. The study [14] shows that providing useful contextual information helps to extract more local information about the detected object. The experimental results of the study [15] demonstrate that for local regions, the detection accuracy is significantly improved by adding special contextual modules. The detection methods based on contextual information can extract more local information and improve the perception of small objects. Although the existing fusion methods [13, 16, 17] produce a joint representation of the information of two modalities, the representation cannot fully capture the complex connection between the two modalities. Also, it is difficult to exploit the heterogeneous and complementary information between the modal localities to organically combine the information of the two modalities for deep contextual interaction.

To solve the above problems, this paper proposes the CenterTransFuser model. First, a radar point spatial information enhancement method and a mask filtering method are proposed based on depth thresholding. The spatial enhancement method extends the radar point's height and width so that the radar point can contain all objects on the image. The depth thresholding filtering method uses adaptive depth thresholding to establish dynamic correlations between image information and radar information to filter out radar noise. Then, to reflect the importance of local information, the cross-transformer fusion model is designed. The model weighs different positions of radar point cloud features and visual information feature maps to learn contextually relevant representations and capture long-term dependencies in the input sequence. The cross-modal cross-multiple attention mechanism in this method enables radar and image information to guide and complement each other for cross-modal complementary information interaction. Meanwhile, the joint cross-multiheaded attention mechanism contextually interacts the original features with the features after cross-modal crossover, which utilizes both types of modal information while preserving their original information. Besides, the method adapts neural networks to generate weight matrices to fuse visual features, extract more local information, and fully reflect the importance of local neighborhoods, thus achieving better cross-modal fusion of radar point cloud information and visual information.

Overall, the main contributions of this paper are as follows.

- A cross-transformer method is proposed to capture the complementary information between the radar point cloud information and image information. It performs contextual interaction to make deep integration with the local information. Also, long and short jump connections are introduced to carry both shallow and deep information. Our model enables the point cloud features and image features to be fused more closely to reduce detection errors.
- The adaptive depth threshold is designed for the depth threshold mask filtering method to suppress irrelevant information in the radar point cloud. It provides different information for the radar according to different preliminary depths of the target in the image. Also, it reduces the loss of depth information caused by radar projection and a narrow field of view, which increases the correlation between the two modals.

- Our model is evaluated on the nuScenes dataset, and the experimental results show that our model achieves excellent performance. Particularly, the detection accuracy is significantly improved for pedestrians, motorcycles, and bicycles, showing the effectiveness and accuracy of the proposed model.

2 Related work

2.1 Monocular 3D object detection methods

3D object detection has received increasing attention in the field of autonomous driving. Monocular detection refers to the classification and localization of targets from the input signal with a single sensor. For monocular RGB images, Simonelli et al. proposed the MonoDIS model [18], which utilizes a novel deconvolution transform for 2D and 3D detection losses and untangles the dependencies of different parameters by isolating and processing parameter sets at the loss level. Zhou et al. proposed the CenterNet [19] model by using a keypoint detection network to find the image on the target centroids. Qi et al. used only the image features of the object to process point clouds and proposed Frustum point [20]. Meanwhile, they exploited the 2D box and depth information obtained from RGB target detectors to locate objects and group points and applied point networks on grouped points to extract target features for 3D bounding box prediction. Lang et al. proposed PointPillars [21] to directly use columns to process point cloud data to convert 3D space to 2D pseudo-image processing, which greatly improves the operation speed.

In addition, with the great popularity of Transformer [22] in image processing, researchers have applied Transformer to object detection and achieved good experimental results. Facebook first proposed the DETR [23] model for RGB images and applied it to 2D object detection. Then, an improved 3DETR [24] model was developed, and the Transformer architecture was applied to 3D object detection. The experimental results show that the model can improve detection accuracy and efficiency. For point clouds, Guo et al. proposed the PCT [25] model, which is well suited for unstructured, disordered point cloud data with irregular domains. Pan et al. proposed the Pointformer [26], which integrates high-resolution local and global features and captures the dependencies between multi-scale representations. At this stage, using a single sensor for object detection has achieved great progress, but a single sensor cannot meet the requirements of autonomous driving, and multiple sensors are required to work together to improve the safety of driverless cars.

2.2 Radar point cloud and image fusion object detection methods

For LiDAR and camera fusion, Chen et al. proposed MV3D [2] by using RGB images and LiDAR point cloud data as input to project 3D point clouds into aerial and foreground views. The integrated data are fused through a network to output classification results and bounding boxes. Li et al. proposed SPRCNN [4] by adding additional branches to predict sparse key points, view points, and object dimensions after a stereo region proposal network (RPN) and combining them with 2D left and right boxes to calculate a coarse 3D object bounding box.

For the fusion of millimeter-wave radar point cloud and vision information, Nobis et al. proposed CRFNet [13] to project radar detections onto the image plane to improve the current 2D object detection network by fusing camera data and projecting sparse radar data in the network layer. Nabati et al. [16] proposed generating 3D object proposals by using radar detection and then projecting them onto the image plane for joint 2D object detection. Simon et al. [27] proposed a spatial attention fusion (SAF) module based on millimeter-wave radar and vision sensors. The module can be embedded in the feature extraction stage and effectively exploits the features of millimeter-wave radar and vision sensors. Dong et al. [10] proposed to establish radar-camera correlations by deep representation learning to explore feature-level interactions and global inference. They adopted a novel sequential loss to enhance the critical association logic to improve the model performance. Nabati et al. proposed the CenterFusion model [11] by using a truncated cone-based association method to accurately associate radar detections with a target on the image.

Then, they exploited the initial detection results to generate a region of interest around the object in 3D space and used the fused features to accurately estimate the 3D properties of the object. In addition, the TransFuser model proposed by Prakash et al. [28] uses a self-attention mechanism to combine image and LiDAR representations to fuse the global context of the 3D scene into feature extraction layers of different modes, thus solving the problem of a high violation rate in complex driving scenes. These methods are different in terms of cascade fusion and element superposition fusion, and they use adaptive neural networks to generate attention weight matrices to fuse two modal features. However, fusion is only performed at a shallow information layer, which often results in wrong and missed detections and unsatisfactory detection results.

3 The proposed method

The architecture of our proposed sensor fusion network is shown in Fig. 2. The network takes two branches, i.e., radar point cloud and RGB image, as input and uses CenterTransFuser to fuse radar point cloud information and visual information. This enables deep cross-modal information interaction and contributes to superior performance. Firstly, the data of each sensor are processed independently and then fed together into the cross-transformer module for contextual information interaction.

3.1 Radar data preprocessing

Radar point cloud data are characterized by sparsity and a limited vertical field of view. In the road scenario, there are more than 3000 LiDAR points projected onto the camera, but less than 100 radar points after projection [29]. Meanwhile, due to the operational limitations of millimeter-wave radar sensors, radar measurements mainly focus on similar heights. Each radar point occupies only one pixel point in the image, which does not match the true height of the object.

To address the sparsity of radar data, the study [13] used 13 nearest time stamps (about 1 s) in the joint representation to increase the density of radar data, but the increase in radar data brings more noise. Our approach for noise handling is introduced in Sect. 3.2. For the missing radar height problem, this paper uses the same design as [13, 29] to extend the height of a given radar point to a range of 0.25 to 2.5 m, and the width is

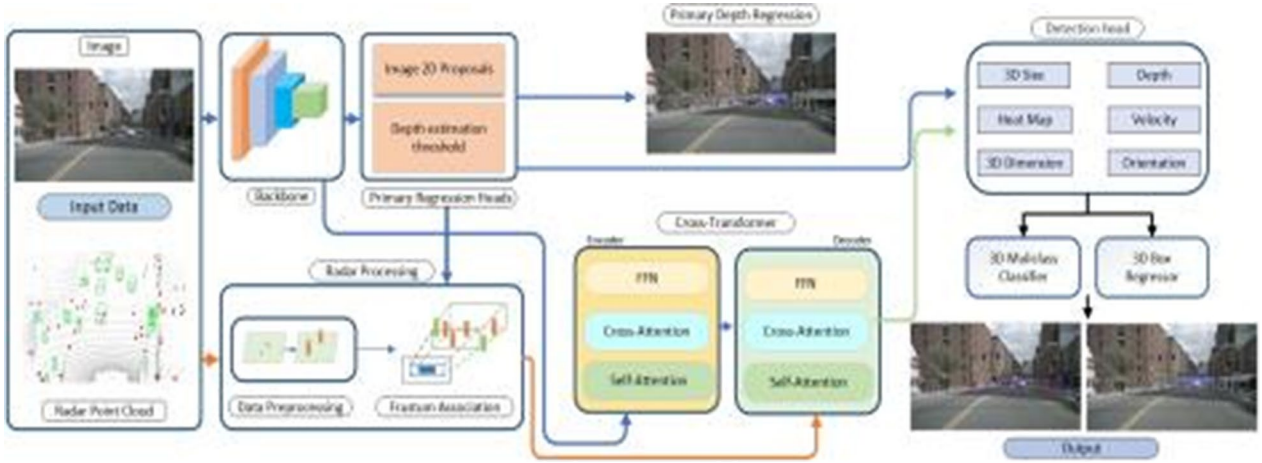

Fig. 2 CenterTransFuser model consists of two branches, namely radar and image. The image branch, after feature extraction by the backbone network, generates preliminary 3D information and 2D regions of interest for the radar branch. The radar branch is supplemented by data processing and frustum association with the information provided by the image. The information from both branches is fed into the cross-transformer module for contextual interaction and into the detection head for classification and regression

extended to a range of 0.05 to 0.25 m in the lateral direction so that the radar point has both vertical and lateral information. With this approach, the radar point can contain all objects on the image to mitigate the effect caused by the limited field of view of the radar. As shown in Fig. 3, the first column presents the radar point projected onto the image, and the second column adds height and width to the radar point for spatial information enhancement, which facilitates the extraction of object information.

Specifically, each radar detection is represented as a 3D point in the egocentric coordinate system and parameterized $P = (x, y, z, v_x, v_y)$, where (x, y, z) is the object position, (v_x, v_y) are the radial velocity of the object in the x and y directions and are compensated by the position of the millimeter-wave radar.

3.2 Dynamic depth thresholding and filtering method

In 3D detection, depth information is crucial for understanding the 3D structure of the scene from 2D images [30]. In 3D object detection with the fusion of point cloud information and visual information from millimeter-wave radar, the noise in the radar data is usually filtered with depth information. However, at this stage, the preliminary depth information of the object is only obtained through image feature prediction. As a result, the region of interest generated by the radar information using the preliminary information is not inaccurate enough to filter out irrelevant information in the radar points. Therefore, this paper designs an adaptive depth threshold function τ(d), and the radar branch information can be obtained by different depth information in the image to strengthen the correlation between the two modes and filter out the noise.

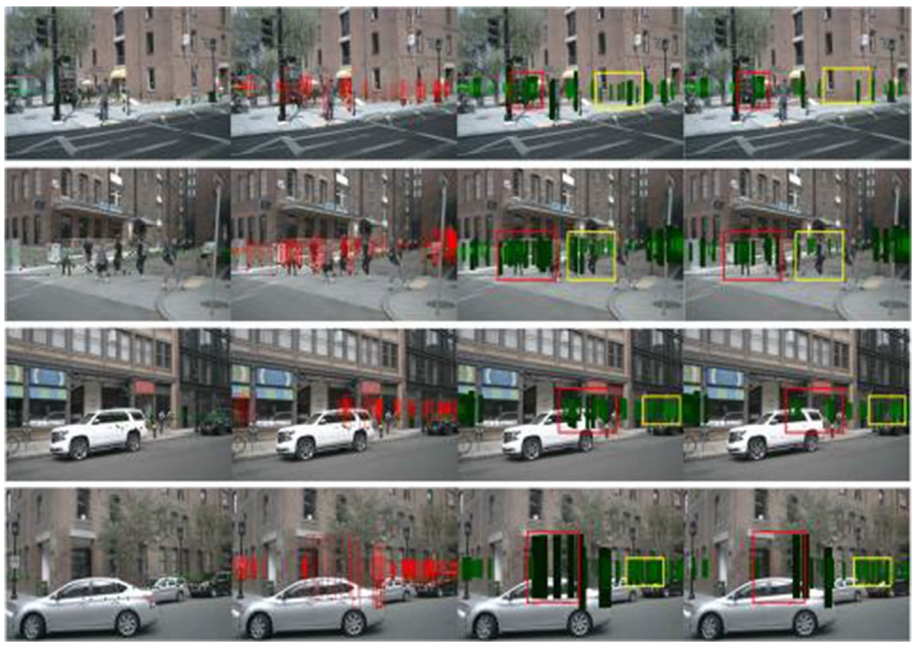

Fig. 3 Spatial information enhancement and the filtering effect. The first column presents the effect of mapping radar points to the image; the second column presents the spatial information enhancement of radar points; the third column presents the effect of mapping radar points to the image using the depth information after baseline spatial information enhancement; the fourth column presents the effect after depth filtering

For the input image xRGB and radar point xRadar, the radar point is projected onto the image. Specifically, the adaptive depth threshold τ(d) is given by Eq. 1:

$$P = f_{\text{stage}}(x_{\text{RGB}}, x_{\text{Radar}}) \tag{1}$$

where the nuScenes [31] dataset provides the calibration parameters needed to map the radar point cloud from the radar coordinate system to the auto-centricity and camera coordinate systems, f_{stage} indicates that the nuScenes dataset uses spatial calibration to project the radar points onto the image.

For the radar points projected onto the image, let the set of radar points $P \in \mathbb{R}^N$, where N is the number of radar points. The central scalar of each point $\hat{d}$ is predicted by the backbone network, and then the preliminary depth information based on the image features is derived through Eq. 2:

$$d = \frac{1}{\delta\left(\hat{d}\right)} - 1 \tag{2}$$

where δ is the sigmoid function. The uncertainty of depth prediction increases with the ground truth depth, but the preliminary depth of the target is only obtained based on the image feature prediction. In this case, the region of interest generated by the radar branch by using the preliminary information is not accurate enough.

Therefore, the depth thresholding method is designed as shown in Eq. 3:

$$\tau(d) = \exp\left(\frac{d * \log\left(\frac{\beta}{\alpha}\right)}{K} + \log(\alpha\beta)\right) \tag{3}$$

where d is the preliminary depth information obtained based on the image feature prediction; K is the number of subintervals discrete into in the (α, β) depth interval; α and β are the hyperparameters. In this paper, the experimental parameters are adjusted by taking $\alpha = 4, \beta = 5$, and the experiments show that the best results are obtained in this distinction.

For discrete intervals, this paper uses the SID strategy to perform discretization, as shown in Eq. 4:

$$\text{SID}: \ t_i = e^{\log(\alpha) + \frac{\log(\beta/\alpha) * i}{K}} \tag{4}$$

where $t_i = \{t_0, t_1, \ldots, t_K\}$ is the discretization threshold space.

By using depth thresholding, the training loss in areas with larger depth values can be reduced, and the image information can be more accurately predicted for relatively small and medium depths and reasonably estimated for larger depth values. Meanwhile, different information can be given to the radar based on the depth information of the target obtained from image prediction, and the loss of depth information due to the narrow field of view during radar projection can be alleviated.

For the filtering of radar noise, because there is no LiDAR ground truth to filter radar points in practical applications and there is no 3D structure information about the scene, it is difficult to eliminate radar point outliers. Previous methods [13] use an annotation

filter (AF) so that the filtered data contain only the objects detected by the radar at least once. However, the radar data output by this method partially filters out relevant object information, leading to serious under-detection. In 3D object detection with multimodal fusion, the anomalous noise generated by radar points projected onto the image is usually related to the depth value. So, this paper uses the above-mentioned depth threshold for filtering to reduce the noise of radar modality-independent information projected onto the image, as shown in Eq. 5:

$$P_d = f_{\text{stage}}(\tau(d), x_{\text{RGB}}, x_{\text{Radar}}). \tag{5}$$

According to Eq. 5, different information can be given to the radar according to the depth of the object predicted in the image. When the depth information of the target prediction $D \geq \tau(d)$, the radar information outside the range can be directly discarded to obtain $P_d \in \mathbb{R}^n$ as the number of effective radar points after filtering. The experimental result is illustrated in Fig. 3. In this figure, the third column presents the effect map of radar points mapped to images using the depth information after baseline spatial information enhancement, and it can be seen that much overlapping information contains noise; the fourth column presents the effect map after depth filtering, which is highlighted by red and yellow boxes, and it can be seen that radar noise is effectively suppressed.

3.3 Backbone network CenterNet

The CenterNet [19]network uses the keypoint detection network to find the target centroids on the image and regress to other object attributes. In this paper, the input image $I_m \in \mathbb{R}^{3\times H\times W}$ (H, W are the height and width of the image, respectively), $F_m \in \mathbb{R}^{C\times H\times W}$ is generated by a feature extractor (C is the number of image channels), and then a keypoint heat map is generated $\hat{Y} \in [0,1]^{C_l\times \frac{H}{R}\times \frac{W}{R}}$, R is the downsampling rate, and C_l is the number of target classes.

For the generated heat map, the focal loss function [19] is used, as shown in Eq. 6:

$$L_k = -\frac{1}{N}\sum_{xyc}\begin{cases}\left(1-\hat{Y}_{xyc}\right)^{\alpha}\log\left(\hat{Y}_{xyc}\right)\\ \left(1-Y_{xyc}\right)^{\beta}\left(\hat{Y}_{xyc}\right)^{\alpha}\log\left(1-\hat{Y}_{xyc}\right)\end{cases} \tag{6}$$

where N is the number of key points in the image; α, β are the hyperparameters, with $\alpha = 2, \beta = 4$ in this paper; $Y \in [0,1]^{C\times \frac{H}{R}\times \frac{W}{R}}$ is the ground truth heat map generated by the target.

In 3D detection using CenterNet, for each centroid, three additional attributes need to be regressed: depth, 3D dimension, and orientation. 3D dimensions are regressed with a separate head to their absolute values $\hat{\Gamma} \in [0,1]^{3\times \frac{H}{R}\times \frac{W}{R}}$, and $L1$ loss is used. For orientation, the specific dimension proposed by Mousavian et al. [32] is followed as $\text{Rot} = \mathbb{R}^{8\times \frac{H}{R}\times \frac{W}{R}}$. To recover the discrete error caused by the output step, for each point, a local offset $\hat{O} \in R^{2\times \frac{H}{R}\times \frac{W}{R}}$ is predicted, and these features are used with the $L1$ loss. For the processing of radar point cloud, radar data preprocessing is performed first, and preliminary radar features $F_r \in \mathbb{R}^{3\times H_0\times W_0}$ (3 is the number of channels) are obtained using the 2D box and the preliminary depth information and size provided by the image. Then,

the preliminary regions of interest are generated by frustum association [11]. Finally, the final output features of the radar image are input to the detection head, the properties of the object are regressed, and the classification and regression are output. This detection head is set up as that in [11], which consists of a convolution kernel and a convolution layer, and the output attribute information is used with L1 loss.

3.4 Cross-transformer

In this paper, the proposed cross-transformer can weigh different positions of radar features and image features, thus effectively utilizing the features of millimeter-wave radar and vision sensors to enable better interactions between the two modal contexts and cross-modal information. Specifically, the module has a radar branch and an image branch and uses the information from each branch as a query matrix to guide the other branch to extract target-related information. Then, it calculates the self-attention and joint cross-attention between all corresponding input information and performs deep fusion through multiple attention context interactions to extract features that are more relevant to the detected object.

Inspired by Transformer [22], this paper designs the cross-modal cross-multiple self-attention mechanism and the joint cross-multiple attention mechanism in Cross-Transformer. These two attention mechanisms work together to make the two modal information interact contextually to fuse deeply and extract more local information.

The cross-modal self-attention mechanism can guide the network to reinforce the target-related information in different modalities and perform contextual feature fusion on the information while keeping the query modal information stable, thus enhancing the robustness of the features. The joint cross-attention mechanism enables cross-modal information to interact with its modal information to facilitate the information extraction and focus on more contextual local information, thus obtaining information around the region of interest and improving object classification by learning the relationship between the object and the surrounding information. Since the inherent self-attention complexity of Transformer is $O(n^2)$, the computational overload and the high computational and storage requirements hinder its large-scale deployment on GPUs. The image branch is processed by CenterNet [19] as the backbone network to obtain a 448×800 RGB image x_{RGB}, which is further downsampled to $\cdot 112 \times 200$. The complexity becomes $O(kn^2)$, where k is the scaling ratio and set to $\frac{1}{16}$ in this paper. This reduces the high resource requirement for GPU and increases the scale of model deployment. Also, the convergence speed is improved, and the computation amount is reduced while ensuring accuracy. To alleviate the information loss caused by downsampling, a jump connection is introduced.

For the cross-modal self-attention mechanism and the joint cross-attention mechanism, the multiheaded attention mechanism is used to extract richer information. In this way, the model can learn relevant information in different representation subspaces, which enables better contextual interaction of image information and radar information and facilitates deep multimodal information fusion.

The multihead attention mechanism uses the scaled dot product between the query matrix (Q) and the key matrix (K) to calculate the similarity between the radar information and the image information. Then, it aggregates each query value (V) to reinforce the

complementary information of the radar image. Multihead is formed by stitching the self-attentive information of each head, as shown in Eq. 7:

$$\mathrm{ATT} = \mathrm{MultiHead}\ (Q, K, V) = \mathrm{cocat}(\mathrm{Att}_1, \ldots, \mathrm{Att}_h) W^O \tag{7}$$

where W^O is the weight of the self-attentive output, $(\mathrm{Att}_1, \ldots, \mathrm{Att}_h)$ is the number of attentions integrated by scaled dot product attention, h is the number of heads of attention, and the matrices Q, K, and V can be calculated using Eq. 8:

$$\begin{cases} Q = F_q W^q \\ K = F_k W^k \\ V = F_v W^v \end{cases} \tag{8}$$

where $W^q \in \mathbb{R}^{C \times D_{\mathrm{model}}}$, $W^k \in \mathbb{R}^{C \times D_{\mathrm{model}}}$, and $W^v \in \mathbb{R}^{C \times D_{\mathrm{model}}}$ are three different weight matrices. Each matrix can be mapped to the corresponding feature to obtain a matrix with different roles in the multiheaded attention. For each multiheaded attention module $\mathrm{MultiHead}_n(Q, K, V)$, the three weight matrices W_n^q, W_n^k and W_n^v map the corresponding features. In Eq. 5, F_q, F_k, and F_v can be equal or different. When $F_q \neq F_k$, it is possible to reinforce the relevant information in the model based on similarity. When $F_q = F_k$, the model becomes a multiheaded self-attention model. Additionally, the self-attention information of each head can be derived from Eq. 9.

$$\mathrm{Att}(Q, K, V) = \mathrm{softmax}\left(\frac{Q(K)^T}{\sqrt{D_{\mathrm{model}}}}\right) V \tag{9}$$

where $\frac{1}{\sqrt{D_{\mathrm{model}}}}$ is a scaling factor, D_{model} is the dimensionality of the query vector and the key vector, $\mathrm{softmax}(\cdot)$ is used to prevent the function from converging to an interval with a very small gradient when the size of the dot product becomes large.

Figure 4 shows the structure of the proposed cross-transformer, which consists of two parts, namely the image branch and the radar branch, with two inputs, radar features $F_r \in \mathbb{R}^{3 \times H_0 \times W_0}$ (3 is the number of channels) and image features $F_m \in \mathbb{R}^{C \times H_0 \times W_0}$. In the image branch, the radar information is used to guide the image information, and the weight relationship between the image information and the radar information is obtained. This paper maps the features F_r to the matrix Q_r, maps the feature F_m to the matrix K_m and matrix V_m to interact with the contextual information in the image. The cross-modal multiheaded self-attention mechanism reinforces the information related to the detected object in the two modal features, as shown in Fig. 5. The process is shown in Eq. 10.

$$\mathrm{att}_M = \mathrm{MultiHead}(Q_r, K_m, V_m) \tag{10}$$

In Eq. 6, the role of the reinforced radar information att_M in the information flow within the visual sensor is enhanced to better learn the relationship between the radar point cloud representation and the surrounding environment. To further fuse the radar information and image information, the radar information guided by the image information att_M is combined with the radar query matrix (Q_r) to obtain the joint radar–image cross-query matrix $(\underset{m}{\hat{Q}})$, as shown in Eq. 11.

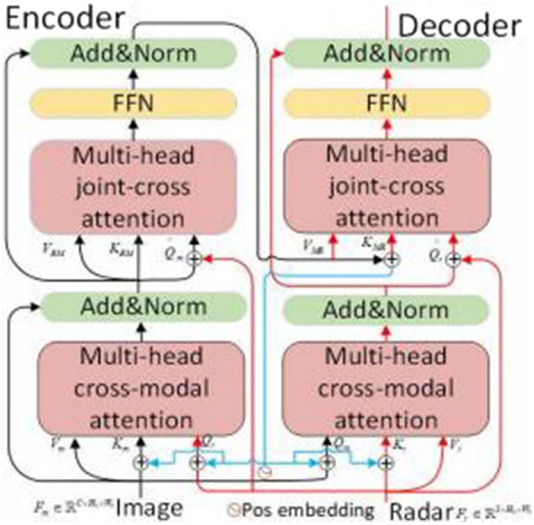

Fig. 4 Cross-transformer model mainly includes two parts, i.e., the encoder and the decoder. The query matrices of the radar branch and the image branch, respectively, guide image information and radar information into multihead cross-attention for cross-modal information interaction and then into multihead joint cross-attention for deep contextual interaction

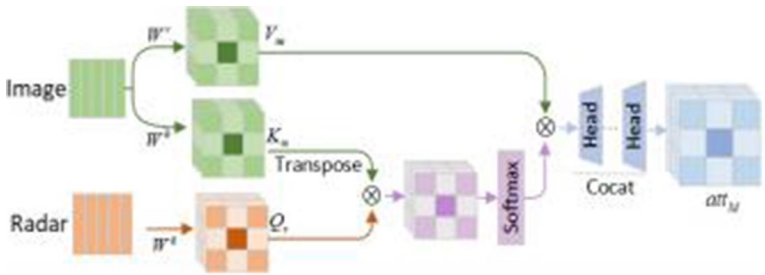

Fig. 5 The cross-modal attention model consisting of two branches, with the query matrix of the radar branch guiding the image for cross-modal interaction

$$\hat{Q}_m = \mathrm{att}_M + Q_r \tag{11}$$

Based on Eq. 12, the radar–image cross-joint query matrix $(\hat{Q}_m)$ is used to perform the similarity calculation with the cross-modal multiheaded self-attention result att_M to obtain the radar–image joint cross-headed multiheaded attention att_{RM} to realize deep interaction in the radar–image information context. K_{RM} is att_M the mapped key matrix, and V_{RM} is att_M of the mapped value matrix.

$$\mathrm{att}_{RM} = \mathrm{MultiHead}\left(\hat{Q}_m, K_{RM}, V_{RM}\right) \tag{12}$$

For the radar branch, the image information is used to guide the network to learn target-related information in the radar. The image features F_m are mapped as matrix Q_m, and the radar features F_r are mapped as matrices K_r and V_r. Meanwhile, the interactive contextual information in the radar cross-modal multihead attention is reinforced with the target-related information in the two modal features so that the image information

complements the radar information. Besides, the radar self-attention information att_R, which is weighted by the image query matrix Q_m, can be obtained through Eq. 13:

$$\text{att}_R = \text{MultiHead}\ (Q_m, K_r, V_r). \tag{13}$$

To achieve further cross-modal interaction between the image–radar interaction information and the radar information and not to lose the radar information, the query matrix of the radar information and the weighted result attRM of the joint radar–image cross-multiple attention are summed up, as shown in Eq. 14:

$$\hat{Q}_r = \text{att}_{RM} + Q_r \tag{14}$$

The cross-modal contextual interaction between image information and radar information is further enhanced to enable deep information fusion. The weighted result att_{RM} after the joint image–radar cross-multiple attention of the radar branch is obtained by using the interaction $\hat{Q}_r$ with the joint radar–image cross-multiple attention att_{RM}. K_{MR} is att_{RM} the mapped key matrix, and V_{MR} is att_{MR} the mapped value matrix, as shown in Eq. 15.

$$\text{att}_{MR} = \text{MultiHead}\left(\hat{Q}_r, K_{MR}, V_{MR}\right) \tag{15}$$

After updating the target features by the cross-modal cross-multiple attention module and joint cross-multiple attention module in the image branch and radar branch, the feed-forward network (FFN) is then applied to each target for further feature extraction. Then, the cross-transformer output is fed into the detection head with the same setting as that in [11] to obtain the desired output. This helps to learn higher-level features such as 3D size, heat map, orientation, rotation, velocity, etc. from the radar feature map.

4 Results and discussion

This section will analyze the data set, evaluation metrics, detail settings, and experimental content used for the experiments. The superiority of the model will be verified by comparing different modalities of current sensors in 3D target detection; the effectiveness of the model will be verified by analyzing the ablation experiments of each module.

4.1 Datasets and evaluation metrics

This paper uses a challenging public nuScenes [31] dataset. The dataset built by nuTonomy is the largest existing dataset for autonomous driving. This dataset not only provides camera and LIDAR data, but also contains millimeter-wave radar data. It includes six cameras, five radars, and one LIDAR. The dataset is organized by scenarios and the whole dataset contains 1000 scenarios that can be split into training and test sets, but only the annotations of the trainval split are publicly available (850 scenarios), of which 700 are training scenarios and 150 are validation scenarios. In each sample, it has 6 images and 5 radar scans in different directions. For the evaluation of the models, the same evaluation metrics as in the articles of [10, 18, 19] were used. Among them, mAP is a composite result that evaluates different classes of detection according to the average

accuracy metric (AP) to obtain the detection model, which can reflect the accuracy of the model detection.

In addition, in 3D target detection, the officials of nuScenes propose a new evaluation metric called nuScenes detection score (NDS), which is the combination of mAP with box position (ATE, average translation error), size (ASE, average scale error), orientation (AOE, average orientation error), attribute (AAE, average attribute error), and velocity (AVE, average velocity error) weighted average that captures all aspects of the nuScenes detection task. In particular, the larger the NDS, the smaller the error of these attribute metrics included, indicating better experimental results.

4.2 Experimental detail settings

This paper uses the DLA [19] backbone in the pre-trained CenterNet [19] network as the object detection network, where CenterNet is trained on the nuScenes dataset for 140 epochs and DLA uses an iterative depth aggregation layer to improve the resolution of the feature maps. The resolution of the camera data in the nuScenes dataset is pixels, where the input RGB images are adjusted by the camera parameters, which can speed up the training convergence. In addition, all models used in the experimental part are implemented in PyTorch [33], and our models are trained on two NVIDIA V100 GPUs for an additional 60 epochs (26 batch sizes) with a learning rate of 0.000025, with the same settings as [11] for the visual cone part of the radar processing to correlate the radar and image targets. Data augmentation was used during training to increase the generalization of the model using random bilateral left–right flips.

4.3 Comparison experiments of different modal 3D object detection methods

This paper proposed that CenterTransFuser model is compared with different modalities for 3D object detection methods on the nuScenes [31] dataset, as shown in Table 1, where CenterNet [19] and MonoDIS [18] are camera models; PointPillars [21], WYSIWYG [34], SARPNET [35], and InfoFocus [36] are LIDAR-based models; and SPRCNN [4] is a LIDAR and camera fusion model. CenterFusion [11] is a model for millimeter-wave radar and image fusion. Table 1 shows the performance comparison of the evaluation metrics for the nuScenes test set, and Table 2 shows the object detection results for the nuScenes dataset, where *L* represents LIDAR, *C* represents camera, and *R* represents millimeter-wave radar. For comparing the differences of algorithms in Table 2, we also

Table 1 Performance comparison results of 3D target detection models in the nuScenes test set

Methods	MDL	NDS (%)	mAP (%)	mATE (%)	mASE (%)	mAOE (%)	mAVE (%)	mAAE (%)
PointPillars [21]	*L*	0.453	0.305	0.517	0.290	0.500	0.316	0.319
SARPNET [35]	*L*	0.484	0.324	0.400	0.249	0.763	0.272	0.090
InfoFocus [36]	*L*	0.395	0.375	0.363	0.265	1.132	1.000	0.395
MonoDIS [18]	*C*	–	0.304	0.738	0.263	0.546	1.553	0.134
CenterNet [19]	*C*	0.400	0.338	0.658	0.255	0.629	1.629	0.142
SPRCNN [4]	*L*+*C*	–	0.361	0.751	0.231	0.571	1.672	0.112
CenterFusion [11]	*R*+*C*	0.449	0.326	0.631	0.261	0.516	0.614	0.115
Our model	*R*+*C*	0.471	0.347	0.628	0.252	0.523	0.527	0.135

MDL Modality

Table 2 Object classification detection results for the nuScenes dataset

Methods	Dataset	MDL	Car	Truck	Bus	Tralier	Const	Pedest	Motor	Bicycle	Traff	Barrier	mAP
PointPillar [21]	Test	*L*	0.684	0.230	0.282	0.234	0.041	0.597	0.274	0.011	0.308	0.389	0.305
	Val	*L*	0.705	0.250	0.334	0.167	0.045	0.599	0.200	0.016	0.296	0.332	0.295
WYSIWYG [34]	Test	*L*	0.791	0.304	0.466	0.401	0.071	0.650	0.182	0.001	0.401	0.347	0.350
	Val	*L*	0.800	0.358	0.541	0.285	0.075	0.669	0.185	0	0.279	0.345	0.354
SARPNET [35]	Test	*L*	0.599	0.187	0.194	0.180	0.116	0.694	0.298	0.142	0.446	0.383	0.324
	Val	*L*	–	–	–	–	–	–	–	–	–	–	–
InfoFocus [36]	Test	*L*	0.772	0.315	0.441	0.359	0.098	0.615	0.251	0.040	0.404	0.453	0.375
	Val	*L*	0.776	0.354	0.505	0.256	0.083	0.377	0.249	0.234	0.550	0.456	0.306
MonoDIS [18]	Test	*C*	0.478	0.220	0.188	0.176	0.074	0.370	0.290	0.207	0.583	0.533	0.338
	Val	*C*	–	–	–	–	–	–	–	–	–	–	–
CenterNet [19]	Test	*C*	0.536	0.270	0.248	0.251	0.086	0.375	0.291	0.207	0.583	0.533	0.338
	Val	*C*	0.484	0.231	0.340	0.131	0.035	0.389	0.305	0.229	0.563	0.470	0.332
CenterFusion [11]	Test	*R*+*C*	0.509	0.258	0.234	0.235	0.077	0.370	0.314	0.201	0.575	0.484	0.326
	Val	*R*+*C*	0.524	0.265	0.362	0.154	0.055	0.617	0.252	0.025	0.334	0.434	0.364
Our model	Test	*R*+*C*	0.535	0.278	0.328	0.243	0.065	0.422	0.387	0.217	0.601	0.476	0.347
	Val	*R*+*C*	0.549	0.289	0.348	0.265	0.063	0.412	0.379	0.215	0.589	0.479	0.379

do comparison experiments only in the test set because the authors of the cited MonoDIS and SARPNET models only give experimental data in the test set.

Analysis of Table 1 mainly illustrates the analytical comparison results of the existing 3D target detection models on the evaluation metrics NDS, mAP, and five error metrics officially given by nuScenes. Higher NDS and mAP indicate better performance, and smaller errors indicate better performance. The SARPNET [4] model is using LIDAR for 3D target detection, because LIDAR mainly emits a laser beam to detect the surrounding environment, the inherent advantage is its wider detection range and higher detection accuracy with higher precision. However, the performance in extreme weather such as rain, snow, and fog is poor, the amount of data collected is too large and very expensive, so millimeter-wave radar is usually chosen to be used. The analysis shows that, compared with the laser model, although our model cannot exceed all the evaluation indicators, it has obvious advantages for mASE and mAOE error indicators. In addition, especially in the subsequent comparison of the detection object classification results, for the detection of small targets, our model has a better detection effect.

It can be concluded that on the nuScenes split test set, the model in this paper improves NDS score and mAP by 1.8% and 4.2%, respectively, compared with the laser point cloud-based model PointPillars [21]; 7.1% and 0.9%, respectively, compared with the camera-based model CenterNet [19]; and with the LIDAR-based model SARPNET [35] and the camera-based model MonoDIS [18], the mAP of this paper is improved by 2.7% and 4.3%, respectively. The experiments show that the detection accuracy using a single sensor is generally low and the detection effect is unsatisfactory for autonomous vehicle applications, indicating that the use of a single sensor alone cannot meet the needs of autonomous driving, and the fusion of multiple sensors is usually required at this stage to improve the accuracy of autonomous vehicle detection. Therefore, compared with the current model CenterFusion [11], which fuses millimeter-wave radar point cloud and visual information for 3D target detection, the model in this paper improves 2.2% on NDS and 2.1% on mAP, which is a significant improvement in detection effectiveness. Comparing the error indicators of the fusion models of the two modalities, all the indicators decrease except mAAE which increases slightly, and the lowest indicators of all kinds are reached in the 3D target detection fusion algorithm, and the error indicators of the model in this paper are relatively reduced compared with the other ones, which reflects the superiority and effectiveness of the model in the field of 3D object detection.

Analyzing Table 2 in nuScenes dataset object classification detection results, for in the test set, it can be concluded that the classification results of our model are improved except for some large targets such as cars, trucks, buses, etc. in WYSIWYG [34] model. Compared with MonoDIS [18] and CenterFusion [11], the performance improvement is obvious. It can be concluded that compared with CenterNet [19] and CenterFusion [11], the improvement is 0.4%, 0.8%, and 8%; 2.6%, 2%, and 9.4% for car, truck, and bus detection, respectively, indicating that the model in this paper has a significant improvement. Because of the inherent advantages of LiDAR, our model has no advantage in the detection category of large targets compared with LiDAR-based models, but our model performs contextual information fusion to extract more local information, and for some small targets such as pedestrians, motorcycles,

bicycles, and traffic cones, our model shows significant advantages. However, LIDAR detection results are a combination of its hardware devices and algorithms, while our work is the design of millimeter-wave radar and image fusion algorithms, which are jointly presented here to compare the performance of 3D target detection methods in autonomous driving. In addition, our model essentially demonstrates the best performance for category classification detection in comparison with existing camera-based models, millimeter-wave radar point cloud-based and camera fusion models on 3D detection baseline.

Compared to MonoDIS [18], our model improves by 5.2%, 9.7%, 1%, and 1.8%, respectively; compared to CenterNet [19], 4.7%, 9.6%, 1%, and 1.8%, respectively; and compared to CenterFusion [11], 5.2%, 7.3%, 1.6%, and 2.6%, respectively. Compared to WYSIWYG [34], the improvements were 20.5%, 21.6%, and 20% on motorcycle, bicycle, and traffic cones, respectively. Similarly, analyzing the object classification results on the validation set, compared with other models, it can be obtained that our model has higher detection accuracy on some small targets and our model achieves excellent performance, reflecting the superiority of the model performance.

4.4 Comparison with existing millimeter-wave radar point cloud information and visual information fusion methods

In the field of autonomous driving, for the fusion of millimeter-wave radar and visual information for object detection, many approaches have been proposed by many scholars, which reflect excellent performance. Currently, the main models are RSF [16], RRPN [17], CRFNet [13], and CenterFusion [11].

In order to reflect the role of our model in this field, make a comparison between our model and these models, as shown in Table 3. The analysis shows that, compared to the RSF [16] model, for pedestrians and motorcycles, the improvement is 14.6% and 12.7%, respectively; for the other three models, for pedestrians, motorcycles, and bicycles, the improvement is 25.1%, 8.2%, and 0.3%; 7.5%, 17.7%, and 7.7%; and 5.2%, 7.3%, and 1.6%, respectively. For automobiles, our model improves by 1.2%, 11.7%, 4.4%, and 2.6%, respectively, compared to the four models. The most significant improvement is observed for pedestrians and motorcycles compared to all models, showing that the models in this paper can extract more information and improve the detection performance of these small targets. In addition, compared with the state-of-the-art model CenterFusion [11], which is based on the fusion of millimeter-wave

Table 3 Results of millimeter-wave radar and image fusion methods for object classification detection in the nuScenes test set

Methods	MDL	Car	Truck	Bus	Tralier	Const	Pedest	Motor	Bicycle	Traff	Barrier
RSF [16]	$R+C$	0.523	0.345	0.483	–	–	0.276	0.260	0.250	–	–
RRPN [17]	$R+C$	0.418	0.447	0.572	–	–	0.171	0.305	0.214	–	–
CRFNet [13]	$R+C$	0.491	0.267	0.431	–	–	0.347	0.210	0.140	–	–
CenterFusion [11]	$R+C$	0.509	0.258	0.234	0.235	0.077	0.370	0.314	0.201	0.575	0.484
Our model	$R+C$	0.535	0.278	0.328	0.243	0.065	0.422	0.387	0.217	0.601	0.476

MDL Modality

radar point cloud and visual information into 3D object detection, the models in this paper all improve, showing the effectiveness and superiority of the proposed model in the millimeter-wave radar point cloud and camera visual information fusion approach.

4.5 Ablation experiments

An ablation study is conducted in the nuScenes validation set to verify the rationality of each module of our model, in which the improved CenterNet [19] is used as a baseline to compare the experimental effects of each module after adding cross-transformer, depth thresholding, filtering, and double-test to reflect the effectiveness of our model after the radar point cloud data is processed by Frustum Association [11], respectively.

Analyzing Table 4 for the comparison of nuScenes performance metrics on the test set, it can be obtained that compared with baseline, the NDS score and mAP improve by 2.5% and 2.1%, respectively, after adding the cross-transformer module, indicating that the addition of the cross-transformer module is more low learning to the relevant information, making the image information and radar information to interact across modal contexts. After adding the depth thresholding module and the filtering module, the NDS score and mAP improve by 2.1% and 1.7%, respectively, indicating the effectiveness of our proposed adaptive thresholding filtering method.

All the errors are reduced except for the error mAAE. When the double-flip test is added, the data are enhanced and the performance of the model is improved. When all modules interact, our model achieves the best performance, reflecting the superiority and excellence of the model.

Analyzing Table 5 model in nuScenes test set object classification detection, it can be obtained that compared with baseline, the addition of cross-transformer module has the most obvious improvement for each category, for trailers, 10%; for some small targets such as pedestrians, motorcycles, bicycles, and traffic cones, 2.8%, 8.1%, 0.2%, and 4.6%, indicating that the cross-transformer module deepens the contextual interaction ability of local neighborhood information between radar images and visual images, and better extracts cross-modal information. After adding the depth thresholding module and the filtering module, for pedestrians, motorcycles, and traffic cones, the improvement is 2.8%, 6.4%, and 4.1%, respectively, indicating the

Table 4 Performance comparison results of CenterTransFuser model on nuScenes test set

Baseline	Cro-trans	DT	Filter	DP	NDS (%)	mAP (%)	mATE (%)	mASE (%)	mAOE (%)	mAVE (%)	mAAE (%)
✓					0.438	0.319	0.654	0.289	0.566	0.573	0.120
✓	✓				0.463	0.340	0.639	0.265	0.542	0.547	0.145
✓		✓			0.442	0.321	0.642	0.260	0.550	0.563	0.140
✓		✓	✓		0.459	0.336	0.637	0.258	0.542	0.559	0.139
✓	✓	✓	✓	✓	0.471	0.347	0.628	0.252	0.523	0.527	0.135

Cro-trans denotes cross-transformer

DT Depth threshold, *DP* Double-flip

Table 5 CenterTransFuser model in nuScenes test set object classification detection results

Baseline	Cro-trans	DT	Filter	DF	Car	Truck	Bus	Traier	Const	Pedest	Motor	Bicycle	Traff	Barrier
✓					0.508	0.232	0.310	0.132	0.046	0.373	0.273	0.213	0.535	0.459
✓	✓				0.523	0.257	0.319	0.232	0.055	0.401	0.354	0.215	0.581	0.470
✓		✓			0.496	0.241	0.296	0.159	0.032	0.389	0.295	0.201	0.545	0.457
✓		✓	✓		0.510	0.242	0.308	0.176	0.036	0.401	0.337	0.206	0.576	0.460
✓	✓	✓	✓	✓	0.535	0.278	0.328	0.243	0.065	0.422	0.387	0.217	0.601	0.476

effectiveness of our proposed adaptive thresholding filtering module for classification detection results. In addition, it can be obtained that our model achieves the best classification detection results when all modules interact with each other.

In addition, this paper visualized the final experimental detection effect plots to make a comparison between our model and the baseline detection model, as shown in Figs. 6 and 7. Among them, Fig. 6 shows the effect picture of missed detection, where the missed targets are marked by yellow rectangular boxes; Fig. 7 shows the effect picture of wrong detection, where the wrong targets are marked by yellow circular boxes. In the visualization images, the first column shows the estimated target depth value on the image, which can visualize the specific target on the image and facilitate the experimental comparison effect display; the second column shows the 3D block diagram of the baseline model detection target; the third column shows the specific classification and detection effect diagram of the baseline model detection target, which can directly reflect the target detection effect; the fourth column shows the 3D block diagram of our proposed The fourth column is the 3D block diagram of the target of our proposed CenterTransFuser model; the fifth column is the specific target classification and detection effect obtained by our model. From Fig. 6, it can be seen that for targets such as trucks, pedestrians, bicycles, traffic cones, etc., our model largely reduces the phenomenon of missed detection. Similarly, in Fig. 7, we can get that for the wrong detection phenomenon existing for targets such as motorcycles,

Fig. 6 Experimental effect of missed detection compared with the baseline model is plotted.The first column is the estimated depth value on the image, which can visualize the specific target on the image and facilitate the experimental comparison effect demonstration; the second column is the 3D block diagram of the target detected by the baseline model; the third column is the specific classification and detection effect diagram of the target detected by the baseline model, which can directly reflect the target detection effect; the fourth column is the 3D block diagram of the target of our proposed CenterTransFuser model; the fifth column is the specific target classification and detection effect diagram obtained by our model

Fig. 7 Experimental effect of error detection compared with the baseline model is plotted.The first column is the estimated depth value on the image, which can visualize the specific target on the image and facilitate the experimental comparison effect demonstration; the second column is the 3D block diagram of the target detected by the baseline model; the third column is the specific classification and detection effect diagram of the target detected by the baseline model, which can directly reflect the target detection effect; the fourth column is the 3D block diagram of the target of our proposed CenterTransFuser model; the fifth column is the specific target classification and detection effect diagram obtained by our model

pedestrians, and bicycles, our model has been effectively improved, showing the effectiveness and superiority of the model.

5 Conclusion

The fusion of radar point cloud and camera visual information is an important stage for object detection in autonomous driving. However, most of the existing studies ignore the extraction of local neighborhood information and only consider shallow fusion between the two modalities based on the extracted global information, which cannot perform a deep fusion of cross-modal contextual information interaction. Meanwhile, in data preprocessing, the noise in radar data is usually only filtered by the depth information derived from image feature prediction, and such methods affect the accuracy of radar branching to generate regions of interest and cannot effectively filter out irrelevant information of radar points. For dealing with these problems. This paper proposes the CenterTransFuser model that makes full use of millimeter-wave radar point cloud information and visual information to enable cross-modal fusion of the two heterogeneous information. The adaptive depth threshold is designed for depth threshold mask filtering method to suppress irrelevant information in the radar point cloud. It provides different information for the radar according to the different preliminary depths of the target in the image. The proposed model is evaluated on the nuScenes dataset, the experiments show it achieves excellent performance. In particular, the detection accuracy is

significantly improved for pedestrians, motorcycles, and bicycles, showing the effectiveness and accuracy of the model.

Acknowledgements
Not applicable.

Author contributions
YL and KZ contributed to conceptualization, methodology, and software; YL performed writing—original draft; YL and TS performed writing—review and editing. All authors have read and agreed to the published version of the manuscript. All authors read and approved the final manuscript.

Funding
The authors gratefully acknowledge support by the National Natural Science Foundation of China (No. 61971208), Yunnan Reserve Talents of Young and Middle-aged Academic and Technical Leaders (Shen Tao, 2018), Yunnan Young Top Talents of Ten Thousands Plan (Shen Tao, Zhu Yan, Yunren Social Development No. 2018 73), Major Science and Technology Projects in Yunnan Province (202002AB080001-8,202202AD080013), Development and Application of Blockchain Service Platform Supporting Regional Integrated Energy Transactions Project of China (No. SGIT0000XTJS1900433).

Availability of data and materials
Please contact the author for data requests.

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
The authors declare no conflict of interest.

Competing interests
The authors declare that they have no competing interests.

Received: 22 February 2022 Accepted: 2 November 2022
Published online: 11 January 2023

References

1. K. Ren, Q. Wang, C. Wang, Z. Qin, X. Lin, The security of autonomous driving: threats, defenses, and future directions. Proc. IEEE **108**(2), 357–372 (2019)
2. X. Chen, H. Ma, J. Wan, B. Li, T. Xia, Multi-view 3d object detection network for autonomous driving, in *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition* (2017), pp. 1907–1915
3. V. John, N.M. Karunakaran, C. Guo, K. Kidono, S. Mita, Free space, visible and missing lane marker estimation using the PsiNet and extra trees regression, in *2018 24th International Conference on Pattern Recognition (ICPR)*. (IEEE, 2018), pp. 189–194
4. P. Li, X. Chen, S. Shen, Stereo r-cnn based 3d object detection for autonomous driving, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2019), pp. 7644–7652
5. C. Michaelis, B. Mitzkus, R. Geirhos, E. Rusak, O. Bringmann, A.S. Ecker, M. Bethge, W. Brendel, Benchmarking robustness in object detection: autonomous driving when winter is coming (2019). arXiv preprint http://arxiv.org/abs/1907.07484
6. Z. Wei, F. Zhang, S. Chang, Y. Liu, H. Wu, Z. Feng, Mmwave radar and vision fusion for object detection in autonomous driving: a review (2021). arXiv preprint http://arxiv.org/abs/2108.03004
7. R. Zhang, S. Cao, Real-time human motion behavior detection via CNN using mmWave radar. IEEE Sens. Lett. **3**(2), 1–4 (2018)
8. K. Yoneda, N. Hashimoto, R. Yanase, M. Aldibaja, N. Suganuma, Vehicle localization using 76 GHz omnidirectional millimeter-wave radar for winter automated driving, in *2018 IEEE Intelligent Vehicles Symposium (IV)* (IEEE, 2018), pp. 971–977
9. Y. Kim, J.W. Choi, D. Kum, Grif net: gated region of interest fusion network for robust 3d object detection from radar point cloud and monocular image, in *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)* (IEEE, 2020), pp. 10857–10864
10. X. Dong, B. Zhuang, Y. Mao, L. Liu, Radar camera fusion via representation learning in autonomous driving, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2021), pp. 1672–1681
11. R. Nabati, H. Qi, Centerfusion: Center-based radar and camera fusion for 3d object detection, in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision* (2021), pp. 1527–1536
12. J.-T. Lin, D. Dai, L. Van Gool, Depth estimation from monocular images and sparse radar data, in *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)* (IEEE, 2020), pp. 10233–10240
13. F. Nobis, M. Geisslinger, M. Weber, J. Betz, M. Lienkamp, A deep learning-based radar and camera sensor fusion architecture for object detection, in *2019 Sensor Data Fusion: Trends, Solutions, Applications (SDF)* (IEEE, 2019), pp. 1–7
14. W. Xiang, D.-Q. Zhang, H. Yu, V. Athitsos, Context-aware single-shot detector, in *IEEE Winter Conference on Applications of Computer Vision (WACV)* (IEEE, 2018), pp. 1784–1793

15. F. Yu, V. Koltun, Multi-scale context aggregation by dilated convolutions (2015). arXiv preprint http://arxiv.org/abs/1511.07122
16. R. Nabati, H. Qi, Radar-camera sensor fusion for joint object detection and distance estimation in autonomous vehicles (2020). arXiv preprint http://arxiv.org/abs/2009.08428
17. R. Nabati, H. Qi, Rrpn: radar region proposal network for object detection in autonomous vehicles, in *2019 IEEE International Conference on Image Processing (ICIP)* (IEEE, 2019), pp. 3093–3097
18. A. Simonelli, S.R. Bulo, L. Porzi, M. Lopez-Antequera, P. Kontschieder, Disentangling monocular 3d object detection, in *Proceedings of the IEEE/CVF International Conference on Computer Vision* (2019), pp. 1991–1999
19. X. Zhou, D. Wang, P. Krähenbühl, Objects as points (2019). arXiv preprint http://arxiv.org/abs/1904.07850
20. C.R. Qi, W. Liu, C. Wu, H. Su, L.J. Guibas, Frustum pointnets for 3D object detection from Rgb-D data, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2018), pp. 918–927
21. A.H. Lang, S. Vora, H. Caesar, L. Zhou, J. Yang, O. Beijbom, Pointpillars: fast encoders for object detection from point clouds, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2019), pp. 12697–12705
22. A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, L. Kaiser, I. Polosukhin, Attention is all you need, in *Advances in Neural Information Processing Systems* (2017), pp. 5998–6008
23. N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, S. Zagoruyko, End-to-end object detection with transformers, in *European Conference on Computer Vision* (Springer, 2020), pp. 213–229
24. I. Misra, R. Girdhar, A. Joulin, An end-to-end transformer model for 3D object detection, in *Proceedings of the IEEE/CVF International Conference on Computer Vision* (2021), pp. 2906–2917
25. M.-H. Guo, J.-X. Cai, Z.-N. Liu, T.-J. Mu, R.R. Martin, S.-M. Hu, Pct: point cloud transformer. Comput. Vis. Med. **7**(2), 187–199 (2021)
26. X. Pan, Z. Xia, S. Song, L.E. Li, G. Huang, 3D object detection with pointformer, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2021), pp. 7463–7472
27. S. Chang, Y. Zhang, F. Zhang, X. Zhao, S. Huang, Z. Feng, Z. Wei, Spatial attention fusion for obstacle detection using mmWave radar and vision sensor. Sensors **20**(4), 956 (2020)
28. A. Prakash, K. Chitta, A. Geiger, Multi-modal fusion transformer for end-to-end autonomous driving, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2021), pp. 7077–7087
29. C.-C. Lo, P. Vandewalle, Depth estimation from monocular images and sparse radar using deep ordinal regression network, in *2021 IEEE International Conference on Image Processing (ICIP)* (IEEE, 2021), pp. 3343–3347
30. H. Fu, M. Gong, C. Wang, K. Batmanghelich, D. Tao, Deep ordinal regression network for monocular depth estimation, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2018), pp. 2002–2011
31. H. Caesar, V. Bankiti, A.H. Lang, S. Vora, V.E. Liong, Q. Xu, A. Krishnan, Y. Pan, G. Baldan, O. Beijbom, nuscenes: a multimodal dataset for autonomous driving, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2020), pp. 11621–11631
32. A. Mousavian, D. Anguelov, J. Flynn, J. Kosecka, 3D bounding box estimation using deep learning and geometry, in *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition* (2017), pp. 7074–7082
33. A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga et al., Pytorch: an imperative style, high-performance deep learning library. Adv. Neural. Inf. Process. Syst. **32**, 8026–8037 (2019)
34. P. Hu, J. Ziglar, D. Held, D. Ramanan, What you see is what you get: Exploiting visibility for 3d object detection, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2020), pp. 11001–11009
35. Y. Ye, H. Chen, C. Zhang, X. Hao, Z. Zhang, Sarpnet: shape attention regional proposal network for lidar-based 3D object detection. Neurocomputing **379**, 53–63 (2020)
36. J. Wang, S. Lan, M. Gao, L.S. Davis, Infofocus: 3D object detection for autonomous driving with dynamic information modeling, in *European Conference on Computer Vision* (Springer, 2020), pp. 405–420

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Submit your manuscript to a SpringerOpen journal and benefit from:

- ► Convenient online submission
- ► Rigorous peer review
- ► Open access: articles freely available online
- ► High visibility within the field
- ► Retaining the copyright to your article

Submit your next manuscript at ► springeropen.com