DOI: 10.23977/jeis.2023.080102
ISSN 2371-9524 Vol. 8 Num. 1

DETR 3D Object Detection Method Based on Fusion of Depth and Salient Information

Yonggui Wang[1,*], Jian Li[1], Zaicheng Zhang[1], Bin He[2]

[1]School of Electronic Information and Artificial Intelligence, Shaanxi University of Science and Technology, Xi'an, Shaanxi, 710016, China
[2]School of Electronics and Information Engineering, Tongji University, Shanghai, 200000, China
1403439107@qq.com
[*]Corresponding author

Keywords: Monocular 3D Object Detection, Depth Module, Global Relationship of Features, Transformer, Saliency Network

Abstract: Most of the existing monocular 3D object detection algorithms combine geometric relationships and convolutional neural networks to predict the 3D attributes of the object, lacking depth feature information and global relationship of features. To solve these problems, a DETR monocular 3D object detection algorithm combining depth and salient information is proposed. A lightweight unsupervised depth module is constructed to extract object depth feature information, and Transformer model is introduced to obtain the global relationship of features. In addition, aiming at the high computational cost of Transformer model in the algorithm, a remarkable network is designed to reduce the computational load of Transformer encoder. The experimental results in KITTI official dataset show that the proposed algorithm achieves the optimal detection accuracy in multiple indicators compared with other current advanced detection algorithms, and the effectiveness of each module in the algorithm is proved through ablation experiments.

1. Introduction

As one of the key tasks of computer vision, 3D object detection aims at identifying and locating corresponding objects in images. It has been widely used in automatic driving and indoor object detection. At present, the algorithm with the optimal detection accuracy mainly relies on LiDAR laser point cloud data to provide accurate target depth information. However, the detection algorithm based on LiDAR has some problems, such as high cost of equipment, high requirement of environment in the process of data acquisition and high cost of calculation. Compared with laser LiDAR equipment, monocular camera has the characteristics of low equipment cost, easy portability and simple data acquisition, so many scholars at home and abroad have paid attention to it, and many excellent monocular 3D target detection algorithms have been proposed.

Most of the existing monocular 3D detection algorithms use convolutional neural networks to restore the 3D bounding frame of the target, which is mainly divided into three steps: (1) predict the location of the 2D center point of the target; (2) Using geometric algorithm or projection algorithm to locate the 3D center of the target and predict the target depth information; (3) Aggregate target visual feature information, so as to restore the 3D surrounding box of the target. However, the convolutional neural network mainly extracts the local relation between features, and lacks the global relation of features. Moreover, the depth information predicted by geometric algorithm or projection algorithm has serious position deviation from the real label, which leads to the decline of detection accuracy.

In order to solve the above problems, a DETR3D object detection algorithm integrating depth and salient information is proposed in this paper. The algorithm includes backbone network, depth module, salient network and DETR model integrating depth and salient information. Transformer[1] encoder and decoder make up the DETR model.

The contributions of this paper are summarized as follows:

(1) A monocular 3D target detection algorithm of DETR fusion depth and salient information is proposed. The algorithm fuses the depth feature and token salient information on the basis of the visual feature information to improve the detection accuracy and efficiency of the algorithm. Experimental results of KITTI dataset prove that compared with other monocular 3D target detection algorithms, the algorithm in this paper achieves the optimum in multiple detection indexes.

(2) Aiming at the lack of depth information in the monocular 3D object detection algorithm, an unsupervised depth information module is proposed to extract the depth feature information in the image.

(3) In view of the lack of global relationship of feature map in monocular 3D target detection algorithm, combining CNN and Transformer model to enrich local and global relationship of features, and in view of the high calculation cost of Transformer model in the algorithm, A significant network is proposed to reduce the computing cost of Transformer encoder.

2. Related Work

This section introduces related work in two aspects: one is the monocular 3D object detection algorithm based on convolutional neural network, and the other is the object detection algorithm based on Transformer.

Convolutional neural network-based method: The 3D detection algorithm based on convolutional neural network mainly relies on geometric algorithm or projection algorithm to predict the 3D information of the target. For example, Li et al. [2] proposed GS3D, which uses the ratio of the height of the 2D bounding box of the target to the true height to predict the depth information of the object, so as to restore the 3D bounding box of the target. Brazil et al. [3] put forward M3D-RPN, established a Region Proposal for 2D detection and 3D detection, and also used 2D bounding box information to restore the target 3D bounding box. Liu et al. [4] proposed AutoShape, established the two-dimensional and three-dimensional geometric constraints between each target, and fitted the 3D key points of the target by fitting the three-dimensional object model of deformation and the automatic model of object mask. Park et al. [5] proposed DD3D, which uses pseudo-lidar for pre-training, so as to provide target depth information labels for monocular 3D target detection. Li et al. [6] proposed DCD, which uses a dense projection constraint algorithm from multiple directional edges to restore the target 3D enclosing box through multiple projection constraints and output more candidate depths.

Method based on Transformer: In recent years, many researchers have introduced Transformer model in the field of object detection to solve the problem of lack of global relationship in feature graph. For example, Carion et al. [1] proposed DETR and built the first end-to-end 2D target detection algorithm based on Transformer model. Zhu et al. [7] proposed Deformable DETR, on which multi-scale feature information was introduced to improve the convergence of the model.

Wang et al. [8] proposed DETR3D and used Transformer model to extract multi-view feature information to restore 3D enclosing box of the target. Zhang et al. [9] proposed MonoDETR and designed a deep regression decoder to predict 3D position information of the target. Huang et al. [10] proposed MonoDTR and used Transformer model to extract depth and visual feature information in depth image and RGB image respectively to predict 3D target position information, and introduced depth position coding to improve 3D detection accuracy. Although the detection algorithm of convolutional neural network above has made continuous progress in detection accuracy, it still lacks depth feature information and global relationship of feature. While the detection method introduced in Transformer model solves the problem of lack of global relationship in feature graph, it also greatly increases the calculation cost of algorithm. Therefore, this paper proposes a DETR3D target detection algorithm that integrates depth and significance information. The algorithm not only contains depth feature information and global relationship of feature, but also reduces calculation cost of Transformer encoder in the algorithm by using token significance information.

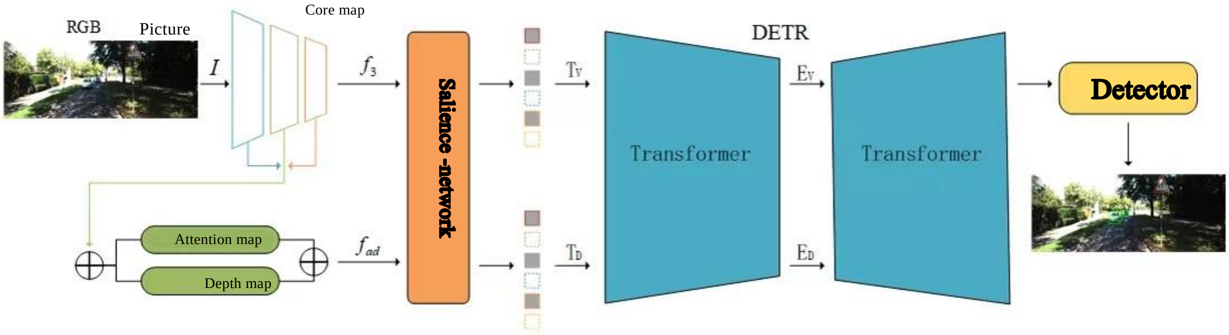

Figure 1: Overall architecture of algorithm

3. Algorithm

The overall framework of the algorithm in this paper is shown in Figure 1, which mainly includes five components: backbone network, depth module, salient network, DETR module integrating depth and salient information, and detection header. The backbone network is ResNet-50[11], which is used to learn the visual feature information of the target. The input is image I, and the output is three feature maps of the subsampling scale of 1/8, 1/16 and 1/32. The depth module is used to learn the target depth feature information (Section 3.1). It uses the attention mechanism to reduce the redundant information in the feature map. The whole training process does not require additional depth information. The significance network is used to predict the significance of tokens (Section 3.2), and the tokens k% before significance are screened out as inputs to the DETR module to reduce computation in Transformer encoders. The DETR module includes a Transformer encoder and decoder (Section 3.3) where the encoder encodes the visual and depth information separately and the decoder aggregates the visual and depth characteristics. The detection header consists of a multitask loss function (Section 3.4, p.4), which is used to restore the target 3D enclosing box.

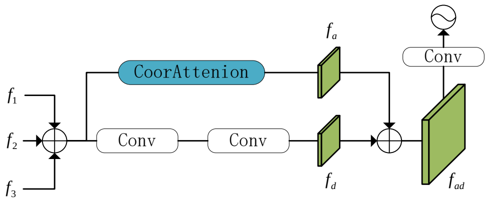

Figure 2: Deep module

3.1 Depth Module

Inspired by S2R-DepthNet[12], this paper designed an unsupervised dual-threaded lightweight depth module composed of attention mechanism and convolutional neural network, as shown in Figure 2.

(1) The feature diagram of the last three layers of the module is ResNet-50$f_1 \in \mathbb{R}^{\frac{H}{8}\times\frac{H}{8}\times C}$, $f_2 \in \mathbb{R}^{\frac{H}{16}\times\frac{H}{16}\times C}$ and $f_3 \in \mathbb{R}^{\frac{H}{32}\times\frac{H}{32}\times C}$The multi-scale semantic information is generated by combining the down-sampling algorithm with element addition$f_4 \in \mathbb{R}^{\frac{H}{16}\times\frac{H}{16}\times C}$

(2) Secondly, f_4 The depth feature map of the imageisgenerated after two convolution$f_{\mathrm{d}} \in \mathbb{R}^{\frac{H}{16}\times\frac{H}{16}\times C}$,simultaneously f_4 Generate the attention feature map through location attention $f_{\mathrm{a}} \in \mathbb{R}^{\frac{H}{16}\times\frac{H}{16}\times C}$,then f_d with f_a Element addition is combined to generate $f_{\mathrm{ad}} \in \mathbb{R}^{\frac{H}{16}\times\frac{H}{16}\times C}$.

(3) Follow LID[9] to generate pseudo-depth labels by convolution block pairsf_{ad}The depth position is coded, and the depth prediction of each pixel is supervised using focus loss and pseudo-depth tags.Be denoted asf_{dmap}.

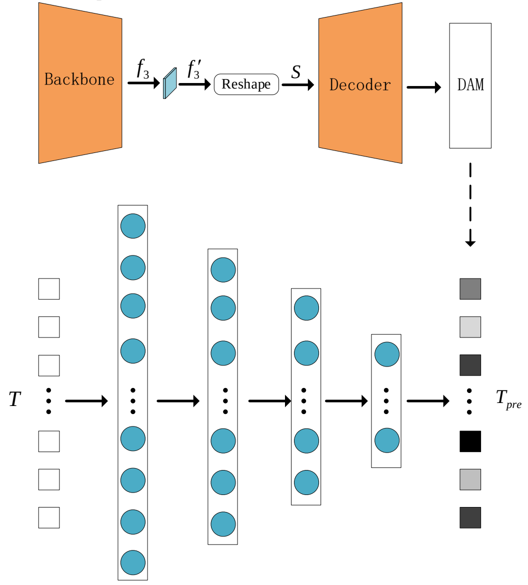

Figure 3: Significant network

3.2 Salient Network

In the field of computer vision, the monocular 3D target detection task requires higher dimensional feature information than other target detection tasks, so the network model usually has a higher computational amount. The introduction of Transformer model into the 3D target detection network model solves the problem of lack of global relationship between feature graphs, but further increases the calculation cost of the algorithm, thus affecting the portability and embeddability of the algorithm. In order to reduce the computational load of Transformer encoder in 3D target detection algorithm, a significant network prediction token significance is proposed in this paper, and the token with high significance is reserved to reduce the computational load of Transformer model in monocular 3D target detection. As shown in Figure 3, this module generates significance pseudo-tags mainly through the backbone network, lightweight decoder and binarization DAM[13], and uses the pseudo-tags and linear layer to supervise and train to predict the significance of tokens, specifically as follows:

(1) The input image passes through the backbone network to get the I feature map of the last layer f_3(Section 3.1), f_3 After the convolution layer of location mapping and reduced dimension generated $f_3' \in \mathbb{R}^{h\times w\times d}(d=256)$,By combining features f_3' the h with w and $\mathrm{S} \in \mathbb{R}^{h\times w\times d}$. Cross-attention mapping of decoders can be used to measure significance [13].Therefore S A binary DAM is generated by a decoder which is composed of a self-attention layer and a deformable attention layer as the pseudo tag of the significance token.

(2) Enter the token $\mathrm{T} \in \mathbb{R}^{h\times w\times d}$after four layers of full connection, the length of the first three layers is 256, 128 and 64, respectively. The last layer uses the cross entropy loss function to predict the token significance, as calculated by formula (1):

$$\mathrm{L_{sig}} = \mathrm{BCE}\left(\mathrm{T_{pe}^{i}}, \mathrm{S^{i}}\right) \tag{1}$$

Among them T_{pre}^{i} To predict the i significance of the fourth token, S^{i} Is the dummy tag corresponding to the i fourth token.

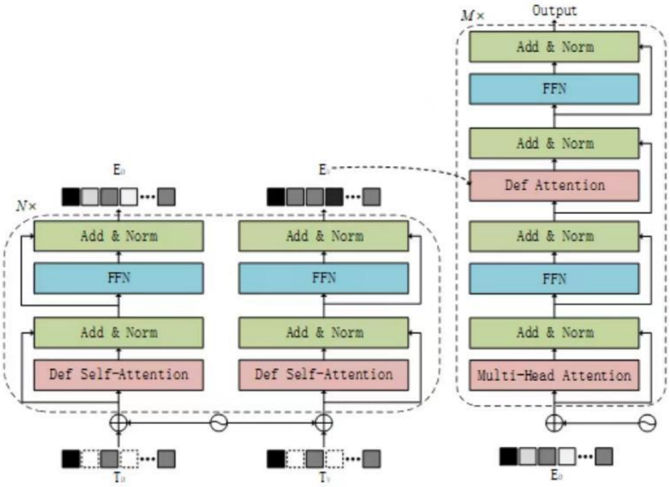

Figure 4: DETR module

3.3 DETR Module

The DETR module consists of a Transformer encoder, which encodes the visual and depth salient tokens separately, and a decoder, which aggregates the visual and depth feature information. The input of Transformer encoder is a significant token, which is generated as follows:(1) Visual feature map f_3 And depth feature map f_{ad} Serialization (Section 3.2, p. 42) generates the corresponding token;(2) Before selecting through the salient network $K\%$ Token of significance, Generate visually significant tokens T_V Significant token with depth T_D .

Encoder: Transformer encoder receives visually significant tokens T_V Significant token with depth T_D ,The aggregation depth position code is used as an encoder input.As shown in FIG. 4(left), each encoder block is successively composed of a Self-Attention layer composed of deformable attention and a feedforward neural layer (FFN).Separate pairTo encode, T_V and T_D ,Generate visual coding tokens E_V and Deep coded token E_D ,The whole encoder contains 6 encoder blocks.

Decoder: The decoder follows the standard architecture of the deep decoder [9], as shown in Figure 4(right)The decoder receives the output visual coding token from the encoder E_V With depth coded tokens E_D ,The initial object query is the object query that can be learned $q \in \mathbb{R}^{N\times C}$, The value of N is 50. Each decoding block is successively composed of multiple attention layer, feedforward nerve layer, deformable attention layer and feedforward nerve layer. The decoder contains 6 decoding blocks. First of all E_D By aggregating depth features through multiple attention layers and then conducting feature interactions between queries through self-attention layers,Finally introduce E_V Visual features are aggregated through deformable layers of attention.

3.4 Detection head

The detection head conducts supervision training through the multi-task loss function, as shown in Formula 2:

$$L = L_{class} + L_{size} + L_{orien} + L_{depth} + L_{sig} + L_{dmap} \qquad (2)$$

Among them, L_{class} , L_{size} , L_{orien} , L_{depth} , L_{sig} (Section 3.2) and L_{dmap} (Section 3.1) respectively represents the classification loss, three-dimensional dimension loss, orientation loss, depth information loss, token significance loss and depth map loss of the target.

(1) Classification loss: It consists of focus lossandisdefinedasFormula3:

$$L_{class} = \begin{cases} -a(1-y')'\log y' & y=1 \\ -(1-a)y'^{\gamma}\log(1-y') & y=0 \end{cases} \qquad (3)$$

Among them, y' After the predictive output of the sigmoid activation function (values between 0 and 1), a =0.25, γ =2.

Three-dimensional dimensional loss: it is composed of IoU optimization loss [14], which is defined as Formula 4:

$$L_{size} = \left\|\frac{(s-s^*)}{s}\right\|_1 \qquad (4)$$

Among them, ||·||1It's the L1 norm, $s^* = [h^*, w^*, l^*]_{3D}$ The true length, width and height of the target, $s = [h, w, l]_{3D}$ Represents the length, width and height of the predicted target.

(3) Orientation Loss: composed of Multi-Bin loss, it is defined as formula 5:

$$L_o = -\frac{1}{n_{\theta^*}}\sum\cos(\theta^* - c_i - \Delta\theta_i) \qquad (5)$$

Among them, n_{θ^*} Is the number of 3D enclosing boxes in the image, θ^* Is the real Angle value of the target, c_i As the goal i the3DCenter Angle of enclosing frame, $\Delta\theta_i$ Is the change in the center Angle.

(4) Depth prediction: It is composed of depth loss [9], which is defined as formula 6:

$$L_{depth} = \mathrm{Depth}\left(d_{gt} - d_{pred}\right) \qquad (6)$$

Among them, d_{gt} Is the target's ground depth tag, d_{pred} Predict the depth value for the target.

4. Experiment

4.1 Data Set

This paper conducted experiments on KITTI monocular 3D target detection data set [15], which contained 7,481 images, including 3,712 training images and 3,769 verification images, all of which had a high resolution of 384*1280 and were divided into three categories, namely, automobile, pedestrian and bicycle. Reported in section 4.3 simple, moderate difficulty, three levels of test results, and in three dimensional space boundary box and bird 's-eye view of the average precision (AP | R40) evaluation of performance, were recorded as AP3D | R40 [15] and APBEV | R40 [15], is located in 40 recalled.

4.2 Experimental Setting

The mono 3D object detection algorithm proposed in this paper is implemented based on Pytorch framework. ResNet-50 is used as the backbone network and Tesla V100 GPU with 16GB video memory is used for training and testing in Ubuntu16 environment. In the training process, the input image Size was 384*1280, the Batch Size was set to 8, the Adam optimizer was used for optimization, the initial learning rate was set to 2x10-4, the training was 200 rounds, and the learning rate was reduced by 0.1 times at the 125th and 165 epoch. In Transformer module,6 layers of coding blocks and decoding blocks are set, the number of queries is 50, and the depth range is set as [0m,60m] according to MonoDLE[14].

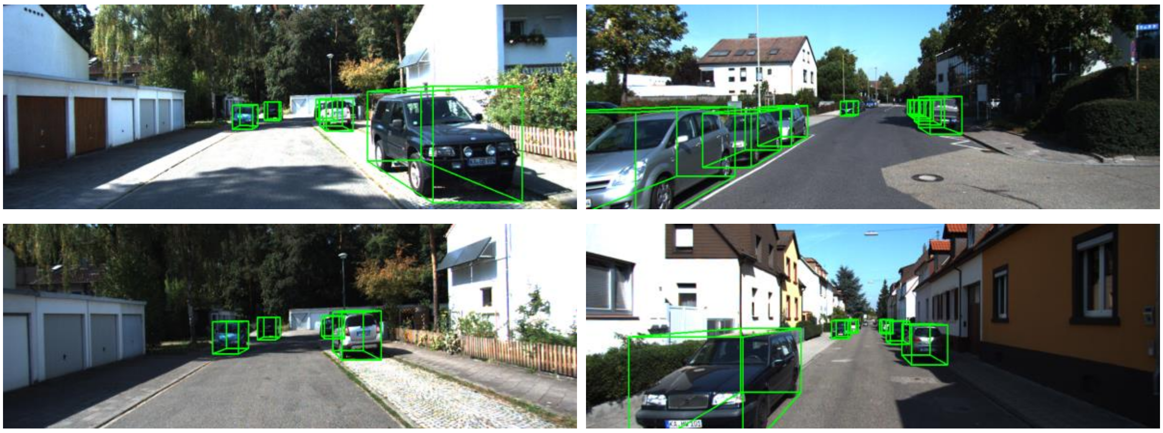

Figure 5: Visualization results

4.3 Experimental Results and Analysis

The experiment in this section compares and analyzes the detection accuracy of the current mainstream monocular 3D target detection algorithm and the model proposed in this paper on KITTI verification set. In addition to the monocular 3D target detection algorithm based on convolutional neural network, the comparison experiment also includes the 3D target detection algorithm based on Transformer model. In order to guarantee the reliability of the contrast experiment results and fairness, adopted in the experimental contrast KITTI official evaluation index, respectively 40 recall the location of the 3 d detection average precision (AP3D | R40) with aerial view perspective 3 d detection average precision (APBEV | R40), including the IoU for occurring simultaneously, The accuracy of three detection levels of easy, medium and difficult is given respectively under each evaluation index. All the detection accuracy given by the algorithm in this paper is that the top 50% tokens are retained as encoder input. Visualization is performed according to the detection results. Figure 5 shows the effect of the target 3D surrounding box projected on the RGB image.

Table 1: Comparative experiment of IoU ≥ 0.7

algorithm	model	AP3D|R40[%](IoU≥0.7)			APBEV|R40[%](IoU≥0.7)		
		Easy	Mod	Hard	Easy	Mod	Hard
M3D-RPN[3]	CNN	14.53	11.07	8.65	26.86	21.15	18.36
RTM3D[16]		19.17	14.20	11.99	24.74	22.03	18.05
MonoFlex[17]		23.64	17.51	14.83_x0007_-	-	-	
MonoDLE[14]		17.45	13.66	11.68	24.97	19.33	17.01
GrooMeD-NMS[18]		19.67	14.32	11.27	27.38	19.75	15.92
Ground-Aware[19]		23.63	16.16	12.06	-	-	-
AutoShape[4]		20.09	14.65	12.07	-	-	-
GUPNet[20]		21.10	15.48	12.88	28.58	20.92	17.83
DEVIANT[21]		24.63	16.54	14.52	32.60	23.04	19.99
MonoDETR[9]	CNN Transformer	22.54	15.86	12.93	33.53	22.37	19.12
Algorithm of this paper	CNN Transformer	24.38	17.78	14.69	36.02	25.68	21.75
promotion		-0.25	+1.24	-0.14	+2.49	+2.64	+1.76

Table 1 and Table 2 respectively show the average detection accuracy of experimental results of the proposed algorithm and other monocular 3D target detection algorithms on KITTI data set when IoU≥0.7 and IoU≥0.5, in which the underline accuracy is the optimal result. As shown in table 1, this algorithm under the medium level of detection of AP3D | R40 and all testing level APBEV | R40 achieve optimal, compared with suboptimal detection accuracy improved 1.24% and 2.49%, respectively, 2.64% and 1.76%. While this article algorithm under the simple and difficult detection level of AP3D | R40 for subprime, but compared with the detection precision of the optimal fell by 0.25% and 0.14%.

Table 2: Comparative experiment of IoU ≥ 0.5

algorithm	model	AP3D|R40[%](IoU≥0.5)			APBEV|R40[%](IoU≥0.5)		
		Easy	Mod	Hard	Easy	Mod	Hard
M3D-RPN[3]	CNN	49.89	36.14	28.08	55.87	41.36	34.08
RTM3D[16]		52.59	40.96	34.95	56.90	44.69	41.75
MonoFlex[17]		-	-	-	-	-	-
MonoDLE[14]		55.41	43.42	37.81	60.73	46.87	53.22
GrooMeD-NMS[18]		55.62	41.07	32.89	61.83	44.98	36.29
Ground-Aware[19]		58.95	43.99	38.07	64.60	47.76	42.97
AutoShape[4]		-	-	-	-	-	-
GUPNet[20]		58.95	43.99	38.07	64.60	47.76	42.97
DEVIANT[21]		61.00	46.00	40.18	65.28	49.63	43.50
MonoDETR[9]	CNN Transformer	60.56	43.73	37.28	67.10	47.85	42.18
Algorithm of this paper	CNN Transformer	66.62	48.02	42.56	71.03	52.78	46.04
promotion		+5.62	+2.02	+2.38	+3.93	+3.15	+2.54

As shown in table 2, the algorithm under all testing level AP3D | R40 and APBEV | R40 achieves optimal, compared with the detection precision of subprime respectively increased by 5.62%, 2.02%, 2.38%, 3.93%, 3.15% and 2.54%.

4.4 Ablation Experiment

In order to further illustrate the depth of the modules is given in this paper, significant network, and combines the depth information and significant information of DETR module in monocular effectiveness of 3 d target detection algorithm, this article on the ablation experiments KITTI validation set, evaluation index for IoU acuity 0.7 when three testing level under AP3D | R40. For the convenience of comparison, ablation experiments will be conducted separately for each module in this section. The baseline algorithm includes depth module, significant network and DETR module that integrates depth information and significant information, in which the top 50% encoder token of significance is reserved as input for DETR module.

Table 3: Deep module ablation experiment

name	AP3D|R40[%](IoU≥0.7)		
	Easy	Mod	Hard
Attention diagram	22.29	16.57	14.19
Depth module	21.17	16.10	13.94
reference	24.38	17.78	14.69

4.4.1 Depth Module Ablation Experiment

Table 3 shows the ablation results of the depth module. When the attention diagram in the depth module is missing, the detection accuracy decreases by 2.09%, 1.21% and 0.5% respectively. When the whole depth module is missing, the depth feature and depth position coding information are missing in the whole detection algorithm, so the detection accuracy is reduced by 3.21%, 1.68% and 0.75% respectively. Table 4 Significant network ablation experiments

Table 4: Significant network ablation experiment

Token number	Amount of computation	AP3D|R40[%](IoU≥0.7)		
		Easy	Mod	Hard
10%	0.86G	23.80	17.11	14.49
20%	1.52G	23.93	16.59	14.56
30%	2.10G	24.02	17.46	14.62
50%	2.99G	24.38	17.78	14.69
100%	5.10G	23.55	17.77	14.72

4.4.2 Significant Network Ablation Experiment

In the ablation experiment in this section, the top 10%, 20%, 30%, 50% and 100% toke ns with high significance are retained as input of Transformer encoder after the token passe s through the significant network. Table 4 shows the experimental results of corresponding detection accuracy and calculation amount of Transformer encoder respectively. The calculati on quantity follows DETR[1] calculation criteria. While keeping the top 100% in difficult t o detect when the token level under AP3D | R40 achieve optimal, its corresponding encode r computation is 5.10 G. Keep top 50% tokens in simple and moderate detection under the difficulty of AP3D | R40 achieve optimal, difficult to detect the difficulty of AP3D | R40 subprime, only was reduced by 0.03% compared with the optimal results, but fell by 41.4% in the amount of calculation. According to the comprehensive analysis of detection accurac y and computation amount, the performance is optimal when the top 50% of significant tok ens are retained. Therefore, 50% tokens are selected as the optimal result of the experiment in this paper.

4.4.3 DETR Module Ablation Experiment

In the ablation experiment in this section, the DETR module that integrates depth information and significant information in this algorithm is replaced by the DETR[1] module based on visual features, so as to prove the effectiveness of this module in monocular 3D target detection.

Table 5: 3DETR module ablation experiment

name	AP3D|R40[%](IoU≥0.7)		
	Easy	Mod	Hard
Visual features DETR	19.63	15.36	13.93
reference	24.38	17.78	14.69

Table 5 shows the ablation experiment results of DETR module integrating depth information and significant information. Because the DETR module based on visual features lacks encoders and decoders related to depth feature information, the detection accuracy decreases by 4.75%, 2.42% and 0.76%, respectively.

5. Summary

In order to solve the problems of lack of depth feature information, missing feature global relationship and high computing cost in monocular 3D target detection task in automatic driving scene, a DETR monocular 3D target detection algorithm integrating depth and salient information was proposed, which mainly includes trunk network, depth module, salient network and DETR model integrating depth and salient information. The main network extracted the visual feature information, and the depth module extracted the depth feature information. The saliency network predicts the saliency of each token and retains the first K% of tokens as input to Transformer encoders in the DETR module to reduce computational costs; The DETR model aggregates visual and depth feature information and provides feature global relationships. Experiments on KITTI data set show that the proposed algorithm outperforms the existing monocular 3D target detection algorithm in several detection indexes, and the effectiveness of each module of the algorithm is verified by ablation experiments. In the follow-up work, we will continue to improve the network structure, further improve the detection accuracy and expand in more

References

[1] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In ECCV, 2020.
[2] Li B, Ouyang W, Sheng L, et al. GS3D: An Efficient 3D Object Detection Framework for Autonomous Driving[J]. 2019.
[3] Brazil G., Liu X: M3D-RPN: Monocular 3D region proposal network for object detection. In: ICCV, 2019.
[4] Liu Z, Zhou D, Lu F, et al. Autoshape: Real-time shape-aware monocular 3d object detection[C]. Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021: 15641-15650.
[5] Park D, Ambrus R, Guizilini V, et al. Is pseudo-lidar needed for monocular 3d object detection?[C]. Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021: 3142-3152.
[6] Li Y, Chen Y, He J, et al. Densely Constrained Depth Estimator for Monocular 3D Object Detection[C]. European Conference on Computer Vision. Springer, Cham, 2022: 718-734.

[7] Zhu X, Su W, Lu L, et al. Deformable detr: Deformable transformers for end-to-end object detection[J]. arXiv preprint arXiv:2010.04159, 2020.

[8] Wang Y, Guizilini V C, Zhang T, et al. Detr3d: 3d object detection from multi-view images via 3d-to-2d queries[C]. Conference on Robot Learning. PMLR, 2022: 180-191.

[9] Zhang R, Qiu H, Wang T, et al. Monodetr: Depth-aware transformer for monocular 3d object detection[J]. arXiv preprint arXiv:2203.13310, 2022.

[10] Huang K C, Wu T H, Su H T, et al. MonoDTR: Monocular 3D Object Detection with Depth-Aware Transformer[C]. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022: 4012-4021.

[11] He K, Zhang X, Ren S, et al. Deep residual learning for image recognition[C]. Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 770-778.

[12] Chen X, Wang Y, Chen X, et al. S2r-depthnet: Learning a generalizable depth-specific structural representation[C]. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021: 3034-3043.

[13] Roh B, Shin J W, Shin W, et al. Sparse detr: Efficient end-to-end object detection with learnable sparsity[J]. arXiv preprint arXiv:2111.14330, 2021.

[14] Ma X, Zhang Y, Xu D, et al. Delving into localization errors for monocular 3d object detection[C]. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021: 4721-4730.

[15] Geiger A, Lenz P, Urtasun R. Are we ready for autonomous driving? the kitti vision benchmark suite[C]. 2012 IEEE conference on computer vision and pattern recognition. IEEE, 2012: 3354-3361.

[16] Li P, Zhao H, Liu P, et al. Rtm3d: Real-time monocular 3d detection from object keypoints for autonomous driving[C]. European Conference on Computer Vision. Springer, Cham, 2020: 644-660.

[17] Zhang Y, Lu J, Zhou J. Objects are different: Flexible monocular 3d object detection[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021: 3289-3298.

[18] Kumar A, Brazil G, Liu X. Groomed-nms: Grouped mathematically differentiable nms for monocular 3d object detection[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021: 8973-8983.

[19] Liu Y, Yixuan Y, Liu M. Ground-aware monocular 3d object detection for autonomous driving[J]. IEEE Robotics and Automation Letters, 2021, 6(2): 919-926.

[20] Lu Y, Ma X, Yang L, et al. Geometry uncertainty projection network for monocular 3d object detection[C]. Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021: 3111-3121.

[21] Kumar A, Brazil G, Corona E, et al. Deviant: Depth equivariant network for monocular 3d object detection[C]. European Conference on Computer Vision. Springer, Cham, 2022: 664-683. 1989-07-26 (in Chinese).