International Journal Of Engineering Research And Development
e- ISSN: 2278-067X, p-ISSN: 2278-800X, www.ijerd.com
Volume 21, Issue 5 (May 2025), PP 34-37

Efficient depth estimation method under the guidance of semantic segmentation

Jun Dai[1], Xu Guo[1], Zhongyu Cao[2], Hongyan Chen[3], Wanli Liu[4], Jie Wen[5], Yuegong Wang[5]

1.Henan Polytechnic University, Henan, China. 2. Zhengzhou University of Science and Technology, Henan, China. 3. Liaoning Technical University, Liaoning, China, 4. China University of Mining And Technology, Jiangsu, China, 5. Pingdingshan Pmj Coal Mine Machinery Eouipment Co..Ltd.

ABSTRACT: *Aiming at the problem of large computational complexity when mapping image data to three-dimensional space during the image-point cloud fusion process, this study proposes a depth estimation method using prior semantic information. When estimating the depth information of two-dimensional images, this method uses the semantic segmentation coefficients generated by the semantic segmentation network to guide the downsampling in depth estimation: dense sampling is conducted for the parts with higher semantic scores in the image (including vehicles, pedestrians, and cyclists), and sparse downsampling is carried out for the parts with lower semantic scores in the image (including road backgrounds, etc.). This method achieves the efficient operation of the depth estimation algorithm while maintaining the accuracy of multi-target recognition. Through the algorithm fusion strategy, the average recognition accuracy of vehicle categories has increased by 4.95%, and the average detection accuracy of pedestrian targets has increased by 5.27%.*

KEYWORDS: *Depth estimation; Semantic segmentation; Downsampling; Image point cloud fusion*

Date of Submission: 01-05-2025 Date of acceptance: 08-05-2025

I. INTRODUCTION

The image-point cloud BEV fusion method is a technology that jointly expresses and processes camera image information and lidar point cloud data on the bird 's-eye view (BEV) perspective. Its core significance lies in making full use of the complementary advantages of different sensors to achieve more accurate and robust environmental perception and target detection. This method is capable of comprehensively extracting the texture, color and semantic information of images, as well as the three-dimensional structure, distance and depth information of point clouds in complex scenes.

The key to BEV space fusion lies in how to conveniently and efficiently convert the features of two-dimensional images into the BEV space. The initial related algorithm was proposed by LSS (Lift-Splat-Shoot)[1]. It first predicts the grid-like depth distribution on two-dimensional features, and then "lifts" the two-dimensional features to the voxel space based on the depth. The core idea of this paradigm lies in achieving the mapping from two-dimensional images to three-dimensional space through geometric dimensionality escalation. BEVDepth[2] (ECCV 2022) introduces lidar point cloud as the depth supervision signal. By constructing an explicit depth loss function, the depth estimation error is reduced to 0.76m (nuScenes verification set), and the detection accuracy is improved to 48.1% mAP. In response to the time series modeling requirements of dynamic scenes, BEVFormer[3] (CVPR 2022) proposed the spatio-temporal Transformer architecture. Through the deformable attention mechanism, it aggregated the BEV features of multiple frames, designed the cross-attention module between surround-view cameras, and improved the ADE index of motion prediction by 16.2% on the Waymo dataset. Furthermore, STS[4](ICCV 2023) constructs a hierarchical temporal fusion network: pixel-level motion compensation is carried out at the bottom layer, and target-level trajectory association is performed at the top layer.

Fast-BEV[5] (RAL 2023) adopts a highly compressed depth encoder and replaces ResNet-101 with MobileNetV3, reducing the computational load by 78%. CaDDN[6] adopts a similar network to predict the classification depth distribution, compresses the voxel space features into the BEV space[7], and conducts three-dimensional detection at the end; OFT-Ne[8] fills the uniformly distributed 3D voxel feature grid by aggregating the image features of the corresponding projection regions, and then obtains the orthophoto BEV feature map through vertical summation; BEVFormer utilizes the cross-attention mechanism in the converter to enhance the modeling of 3D-2D view transformation.

In order to achieve efficient depth estimation in image point cloud fusion, this paper proposes a novel depth estimation method. This method uses the semantic segmentation results of image branches to guide the downsampling in the depth estimation process, effectively reducing the complexity of depth estimation and improving the operational efficiency of the overall system.

II. MATERIAL AND METHODS

In this study, the DeepLabV3+ semantic segmentation network is adopted. The Backbone of the original network is replaced by the lightweight MobileNetV4. Meanwhile, the attention mechanism and the bar pooling layer are added to obtain the lightweight IDV3+ network. After the original camera images pass through this network, semantic scores corresponding to different categories are obtained. During the process of depth estimation, it is only necessary to conduct dense sampling on the regions with higher semantic scores, namely pedestrians, cyclists and vehicles, and sparse sampling on the parts with lower semantic scores. The complete workflow diagram of this system is shown in Figure 1.

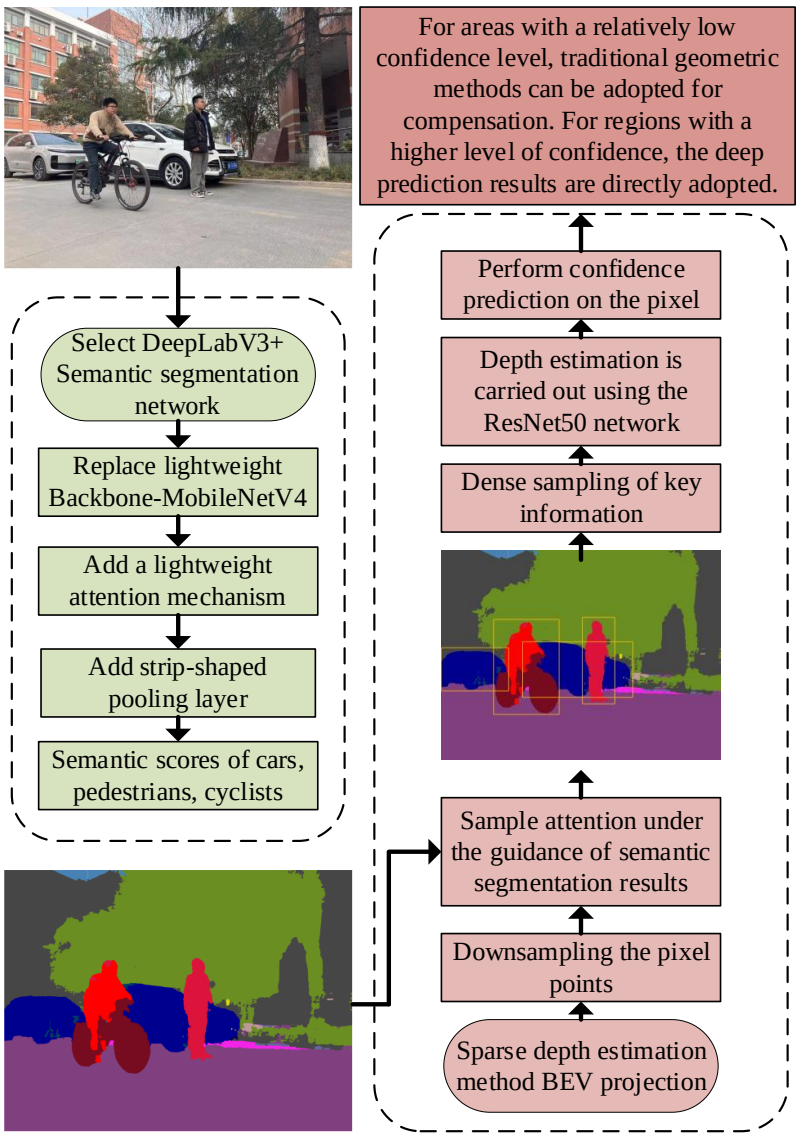

Figure1: System Operation Schematic Diagram

In response to the actual demand for the fusion of image and point cloud data in the BEV space, this study further explores how to effectively fuse the improved DeepLabv3+ semantic segmentation results in the image branch with the BEV features generated by lidar. To this end, we propose a fusion network design idea based on the sparse attention mechanism. This design takes into account the high efficiency of the traditional convolutional neural network (CNN) in local feature extraction and the advantages of the Transformer structure in capturing global information, thereby significantly improving the detection accuracy while maintaining real-time performance.

Firstly, in the design of the network structure, the initial features of the image and point cloud data are obtained respectively through their respective feature extraction networks: The image branch adopts the improved DeepLabv3+ network structure and introduces depth-separable convolution in the encoder and decoder structures, thereby improving the semantic segmentation accuracy while reducing the computational cost. Then, the semantic segmentation results are used to assist in depth estimation and map the two-dimensional image pixel data to the three-dimensional space. Further project the image features onto the BEV space; The point cloud branch utilizes sparse convolution and voxelization methods to map the original liDAR point cloud data to the BEV space, and achieves feature enhancement through height information encoding.

In order to achieve the effective fusion of the two, we have designed a dedicated fusion module in the network. After receiving the image and point cloud features in the BEV space, this module adopts a weighted fusion method based on the cross-modal attention mechanism to fuse the image features and lidar point cloud features in the BEV space. The specific process is as follows: Since both the image features and the point cloud features have been projected into the BEV space of the same size and resolution, that is, there are corresponding image features and point cloud features in each BEV grid cell, in order to ensure that the number of channels of the two different modal features remains consistent, the two features are first mapped to the same number of channels through a 1×1 convolution operation. Through experimental tests in this paper, Finally, the features of both different modes are mapped to 128 channels.

III. RESULTS AND DISCUSSION

To evaluate the running speed and detection accuracy of different downsampling strategies, five different downsampling strategy comparison experiments are set up in this chapter, as shown in Table 1: sparse downsampling with a fixed sampling rate, adaptive downsampling (including saliency guided sampling, multi-scale pyramid sampling, deformable sparse convolution sampling, etc.), and downsampling guided by semantic segmentation results. In terms of experimental configuration, the NVIDIA RTX 4090 24G graphics card is adopted, and the depth estimation encoder uniformly uses ResNet-50 for depth estimation.

Combining the experimental data obtained from Table 1 and conducting qualitative analysis based on the theoretical characteristics of each method, it can be known that coefficient subsampling with a fixed sampling rate performs the best in computational efficiency, but has deficiencies in the retention of detailed information, resulting in the lowest target detection rate and the largest root mean square error in depth estimation. Bilinear interpolation or other means need to be adopted to make up for the loss of detailed information. The method of adaptive sampling has achieved a good balance in aspects such as depth estimation error, detection rate and detection speed, but it requires the introduction of related additional processing networks, which indirectly increases the complexity of the model. On the contrary, thanks to the semantic segmentation coefficients already generated by the image processing branch in the third chapter, by adopting the sampling method guided by the semantic segmentation results, on the basis of the highest target detection rate, the smallest mean square error of depth estimation, and meeting the real-time requirements, no redundant network structure is added, reducing the complexity of the model.

Table 1: Comparative analysis of the Effects of Different downsampling Strategies for depth estimation

Sampling strategy	Root mean square error per meter	Target detection rate (%)	FPS
Sparse subsampling with a fixed sampling rate	3.8~4.5	60~70	**55**
Saliency guides multi-scale pyramid deformable sparse convolution	3.2~3.6	75~85	40
	2.9~3.3	80~88	30
	2.7~3.1	85~90	25
Sampling under the guidance of semantic segmentation results	**1.5~1.9**	**88~92**	22

To verify the effectiveness of the detection algorithm, this study conducted experimental evaluations under the KITTI dataset. Loss weights were set respectively, and different types of loss functions were specified, including binary cross-entropy loss and Smoot-L1 loss; Meanwhile, the Adam optimizer is selected. Set Batch_Size to 1, the learning rate to 0.003, and the Epoch to 80. A comparative experiment was conducted between the BEV object detection method based on image point cloud fusion and the object detection method without BEV fusion. The current object detection fusion algorithms with higher detection accuracy were also compared. The results are shown in Tables 5-3 and 5-4, respectively presenting the 3D experimental results of the 3D vehicle detection benchmark AP3D and the BEV experimental results of the top-down vehicle detection benchmark APBEV. The best results of each target detection are marked in the table in bold.

Table 2: Experimental Results of the 3D vehicle detection reference AP3D

Method	Data	Car（AP3D）/%			Pedestrian（AP3D）/%			Cyclist（AP3D）/%		
		Easy	Mid	Hard	Easy	Mid	Hard	Easy	Mid	Hard
VoxelNet	L	80.23	65.09	59.88	55.63	53.03	47.62	63.52	48.02	44.86
SECOND	L	81.72	70.56	64.36	50.23	40.66	35.86	68.38	51.47	44.79
PointPillars	L	83.74	73.31	67.11	52.02	47.48	42.87	67.24	50.53	46.73
CenterPoint	L	75.06	62.68	62.39	34.18	30.13	26.25	32.53	22.13	21.37
ACDet	L+C	86.41	75.39	70.92	52.83	43.17	40.33	83.77	64.81	58.37
F-PointNet	L+C	80.98	70.45	61.87	51.13	43.56	40.92	71.83	56.02	51.18
Fast-CLOCs	L+C	88.69	80.07	76.36	51.77	41.83	38.07	81.88	64.96	56.33
PointPainting	L+C	91.24	86.97	**85.78**	61.83	57.33	53.27	78.65	62.88	**59.76**
Ours	L+C	**91.69**	**87.86**	85.14	**65.89**	**60.57**	**55.46**	**78.93**	**72.36**	59.34

According to the experimental results of the table, in the Table 2, L indicates that the algorithm only uses lidar data, and L+C indicates that both lidar and image data are used simultaneously. Ours indicates that the algorithm we proposed includes the semantic segmentation module of image branches and BEV fusion 3D object detection. The comparison results of the two major categories of automobiles and pedestrians under the AP3D evaluation benchmark of the KITTI dataset show significant differences. Firstly, in terms of the detection of the automobile category, the average detection accuracies at the simple, medium and difficult levels are 91.69%, 87.86% and 85.14% respectively, and the overall effect is higher than that of the PointPainting algorithm. In terms of the detection of pedestrian categories, the Ours method has improved by 4.06%, 3.24% and 2.19% respectively at the simple, medium and difficult levels compared with the widely used PointPainting algorithm at present.

IV. CONCLUSION

In the image point cloud fusion module, this study proposes a two-dimensional depth information recovery method that uses the semantic segmentation results to guide the downsampling of depth estimation. Only dense depth estimation is performed on the prominent feature regions with high semantic segmentation coefficients in the image, and sparse depth estimation is adopted for unimportant regions such as the background. On the premise of ensuring feature retention, Reduced the computing resource occupation of the model; Finally, the lightweight CNN+Transformer neural network and the fusion strategy based on the cross-attention mechanism are adopted to fuse the image point cloud features in the BEV space. The experimental comparison through the KITTI dataset shows that compared with the method before fusion, The average detection accuracy rates for the categories of automobiles and pedestrians have increased by 4.95% and 5.27% respectively, verifying the effectiveness and accuracy of the algorithm proposed in this study.

ACKNOWLEDGEMENT

This work was sponsored by the National Key Technologies Research and Development Program of Henan Province in China (232102221028 and 242102 221039), Reform Project for Graduate Education at Henan Polytechnic University(2023YJ04)

REFERENCES

[1] Philion J, Fidler S. Lift, splat, shoot: Encoding iages from arbitrary camera rigs by implicitly unprojecting to 3d[C]//Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK,

[2] Li Y, Ge Z, Yu G, et al. Bevdepth: Acquisition of reliable depth for multi-view 3d object detection[C]//Proceedings of the AAAI conference on artificial intelligence. 2023, 37(2): 1477-1485.

[3] Li Z, Wang W, Li H, et al. Bevformer: learning bird's-eye-view representation from lidar-camera via spatiotemporal transformers[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2024.

[4] Sighencea B I, Stanciu I R, Căleanu C D. D-stgcn: Dynamic pedestrian trajectory prediction using spatio-temporal graph convolutional networks[J]. Electronics, 2023, 12(3): 611.

[5] Huang B, Li Y, Xie E, et al. Fast-BEV: Towards real-time on-vehicle bird's-eye view perception[J]. arXiv preprint arXiv:2301.07870, 2023.

[6] Guo X, Shi S, Wang X, et al. Liga-stereo: Learning lidar geometry aware representations for stereo-based 3d detector[C]//Proceedings of the IEEE/CVF international conference on computer vision. 2021: 3153-3163.

[7] Chen Y, Liu S, Shen X, et al. Dsgn: Deep stereo geometry network for 3d object detection[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2020: 12536-12545.

[8] Roddick T, Kendall A, Cipolla R. Orthographic feature transform for monocular 3d object detection[J]. arXiv preprint arXiv:1811.08188, 2018.