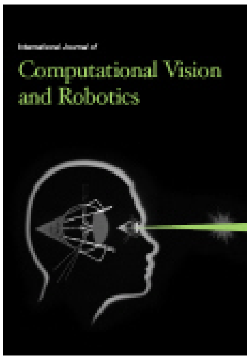


International Journal of Computational Vision and Robotics

ISSN online: 1752-914X - ISSN print: 1752-9131
https://www.inderscience.com/ijcvr

ResNet-based surface normal estimator with multilevel fusion approach with adaptive median filter region growth algorithm for road scene segmentation

Yachao Zhang, Yuxia Yuan

DOI: 10.1504/IJCVR.2022.10049783

Article History:

Received:	12 May 2022
Accepted:	05 July 2022
Published online:	01 December 2023

ResNet-based surface normal estimator with multilevel fusion approach with adaptive median filter region growth algorithm for road scene segmentation

Yachao Zhang and Yuxia Yuan*

School of Electronic and Electrical Engineering,
Zhengzhou University of Science and Technology, China
Email: 316168430@qq.com
Email: byoungholee@qq.com
*Corresponding author

Abstract: As an integral part of information processing, road information has important application value in map drawing, post-disaster rescue and military application. In this paper, convolutional neural network is used to fuse lidar point cloud and image data to achieve road segmentation in traffic scenes. We first use adaptive median filter region growth algorithm to preprocess the input image. The semantic segmentation convolutional neural network with encoding and decoding structure of ResNet is used as the basic network to cross and fuse the point cloud surface normal features and RGB image features at different levels. After fusion, the data is restored into the decoder. Finally, the detection result is obtained by activation function. The KITTI data set is used for evaluation. Experimental results show that the proposed fusion scheme has the best segmentation performance. Compared with other road detection methods, the results show that the proposed method can achieve better overall performance. In terms of AP, the value of proposed method exceeds 95% for UM, UMM scene.

Keywords: road segmentation; adaptive median filter region growth; data fusion; point cloud surface normal feature; encoding and decoding structure.

Reference to this paper should be made as follows: Zhang, Y. and Yuan, Y. (2024) 'ResNet-based surface normal estimator with multilevel fusion approach with adaptive median filter region growth algorithm for road scene segmentation', *Int. J. Computational Vision and Robotics*, Vol. 14, No. 1, pp.99–117.

Biographical notes: Yachao Zhang is with the School of Electronic and Electrical Engineering, Zhengzhou University of Science and Technology, China. His major is road segmentation and image processing.

Yuxia Yuan is with the School of Electronic and Electrical Engineering, Zhengzhou University of Science and Technology, China. Her major is image processing, and signal processing.

1 Introduction

Road detection is an important part of environment identification in automatic driving, and it is the premise of automatic driving. At present, most autonomous vehicles use multi-sensor data fusion to realise road detection (Kim et al., 2021; Gao et al., 2021). The most common one is the fusion of lidar data and RGB image data. Existing studies show that the fusion of these two sensors can improve the accuracy of road detection. The latest fusion method uses convolutional neural network (CNN) as a fusion tool to fuse the data of two modes (Yin et al., 2019), and uses semantic segmentation to detect the road. However, how to better fuse the data of the two sensors is still an urgent problem to be solved in this research field.

In view of the above problems, this paper proposes a variety of pixel-level, feature-level and decision-level fusion schemes. In particular, four cross-fusion schemes are designed in feature-level fusion, and the best fusion scheme is obtained through comparative study of various schemes. In terms of network architecture, semantic segmentation CNN with encoding and decoding structure is adopted as the basic network. The point cloud depth map is represented by an ordinary graph (Xu et al., 2021). The normal graph features and RGB image features are cross-fused at different levels. This method can better learn the correlation between lidar point cloud information and camera image information, cross-supplement point cloud and image information, reduce the loss of feature information.

The main contributions of this paper are as follows:

1 The pixel-level, feature-level and decision-level fusion schemes of point cloud and image data fusion based on CNN are proposed to realise road detection in traffic scenes. In particular, four kinds of cross-fusion schemes are designed in feature-level fusion, and the best fusion scheme is obtained through comparative study of various schemes.

2 KITTI data set is used for experimental evaluation, and the experimental results of various fusion methods are compared and analysed. Experimental results show that the optimal fusion method proposed in this paper can significantly improve the segmentation effect of roads.

The structure of this paper is as follows. In Section 2, we give the related works. Section 3 detailed introduces the proposed road segmentation method. Experiments and analysis are shown in Section 4. There is a conclusion in Section 5.

2 Related works

The traditional road detection method is to distinguish the road from the vertical object according to the geometric properties of the scene to achieve the purpose of road detection. In recent years, CNN has become the mainstream way for road segmentation due to its strong feature extraction and characterisation ability. Road segmentation methods based on deep learning can be divided into semantic segmentation based on image and lidar image fusion.

2.1 Semantic segmentation method based on image.

Image-based semantic segmentation considers road detection as a semantic segmentation task. Semantic segmentation networks mostly adopt encoder-decoder structure. The encoder extracts the effective features, the decoder restores the features, and then it realises the road segmentation by integrating all features and optimisation functions through the full connection layer. U-net is a common segmentation model in encoder-decoder structure. At present, there are many new CNNs based on U-Net structure (Shi et al., 2022). U-net ++ (Zhou et al., 2018) improves the connection mode of the decoder in U-Net by adding the dense connection mechanism similar to DenseNet (Huang et al., 2017), which contributes to the improvement of accuracy. Theoretically, with the increase of network depth, more complex feature extraction can be carried out, and the segmentation performance will become better. However, the deepening of the network often brings about the problem of degradation, and there will be over-fitting phenomenon. Res-UNet (Xiao et al., 2018) is inspired by residual network (ResNet) principle and adds residual unit through short-circuit mechanism, which greatly eliminates the problem of degraded over-fitting caused by deep neural network. Chen et al. (Chen et al., 2018) used DeepLabv3 as an encoder module and a simple decoder module to refine segmentation results, and applied deep separable volume product to ASPP module and decoder module to obtain a faster and stronger encoder-decoder network for semantic segmentation. SegNet used the maximum pooled pixel index in the encoder to de-pool in the decoder, thus eliminating the need to learn up-sampling and saving computing time. Softmax classification was used to output the probability of a category for each pixel.

OFANet (Zhang et al., 2019) used a ‘1-N substitution’ strategy for training, discussed the mutual enhancement effect between detection task and semantic segmentation, and greatly solved a series of problems caused by too few data sets. MultiNet (Teichmann et al., 2018) proposed an approach that combined classification, detection, and semantic segmentation, where the encoder stages of the three tasks were shared, using deep CNN to produce rich shared features that could be used across all tasks. These features were then used by three task-oriented decoders, which produced results in real time. Shared computing reduced the time that it took to perform all tasks, and performance remained to be improved. RBNet (Chen and Chen, 2017) conducted road detection and road boundary detection at the same time, and studied the contextual relationship structure and boundary arrangement between roads. Then, the probability of the pixels in the image was estimated by the Bayesian model belonging to the road and the road boundary, eliminating the potential misjudgment outside the boundary. Multi-task CNN (Oeljeklaus et al., 2018) proposed a compact multi-task CNN architecture to effectively detect and estimate the dryness terrain of objects and basic automotive environment models under the computational resource constraints of embedded systems. It introduced a simple extended 3D boundary box estimation scheme based on detection decoder and analysis geometry.

2.2 Based on lidar and image fusion method

Multi-sensor fusion is to process multi-source information data using certain methods and criteria to achieve the required estimation decision. In the field of autonomous driving, data information such as lidar sensors and cameras are mostly fused to sense the surrounding environment. Schlosser et al. (2016) preprocessed 3D point cloud data of lidar into HHA (horizontal parallax, height of ground, angle) data, and input them together with RGB images. The fusion method of pixel addition was adopted in different specific layers of CNN network, proving that the strongest effect would be obtained in the middle layer of the network. LidCamNet (Caltagirone et al., 2019) adopted the method of feature fusion and trainable linear superposition to compare the experimental results with the fusion results in the early and late periods. Trainable parameters had certain flexibility in data fusion and good segmentation results further verified the feasibility of this method in semantic segmentation. Chen et al. (2019) adopted the progressive lidar adaptive cascade fusion structure. It used lidar data to assist image data for road segmentation. Using trainable parameters, the lidar features and RGB features were adaptively processed. In strong light or strong shadow conditions, it could achieve better fusion effect. Van Gansbeke et al. (2019) proposed a fusion method to correct the prediction of point cloud information by taking RGB image as the guidance and using its target information to reduce the misjudgment probability of point cloud. Wang et al. (2019) used lidar sensors and stereoscopic binocular cameras to estimate depth with the stereoscopic matching network of the two enhancement technologies, instead of direct fusion, which improved detection accuracy to a certain extent.

Zhang and Funkhouser (2018) adopted the RGB-D depth complemented method based on the deep learning. It input a RGB-D graph to predict surface normals and object edge occlusion for all planes in the RGB graph. The depth map was used as regularisation to solve the global linear optimisation problem, and finally the completed depth map was obtained, which provided better data information for the automatic driving environment perception. In order to simultaneously extract RGB image and depth map features, the two approaches were fused in Wang et al. (2021), and the fused image was transformed into HHG image. Jaritz et al. (2018) proposed a 3D object pose estimation based on dual-sensor information fusion (visual cone PointNet target positioning algorithm), which further proved the feasibility of multi-data fusion. SNE-roadSeg (Fan et al., 2020) adopted encoder-decoder structure to perform feature fusion on data input of dual sensors in the encoder to achieve accurate free space detection. The method of transforming point cloud depth graph into normal feature graph was proposed, and the surface normal estimation problem was transformed into the least square plane fitting estimation problem. The difficulty of estimating the normal of every point on the three-dimensional surface was that the three-dimensional points on roads and pavements had very similar surface normals.

3 Proposed road segmentation method

The proposed network infrastructure is shown in Figure 1, which consists of adaptive median filter region growth algorithm, an encoder with ResNet, a decoder with skip connection and dense connection (as shown in Figure 2), and a surface normal-estimator. The input image is RGB-D, and the lidar depth map is processed by surface normal estimator (SNE) into normal map. Features of two input signals are extracted by two encoders, and then they are restored by decoder. Finally, the sigmoid activation function is used to generate road segmentation results.

Normals are used to enrich feature information and correct shadows and other visual effects produced by light sources. The depth map has only a small amount of depth feature information in a single layer. The processed normal map can better distinguish road surface from non-road surface according to the principle that each point is located in different planes and the normal direction of the surface is also different.

Figure 1 Proposed network structure

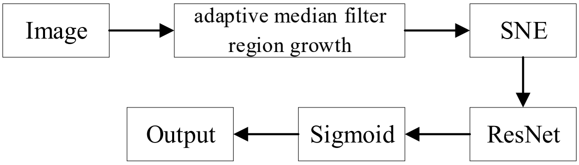
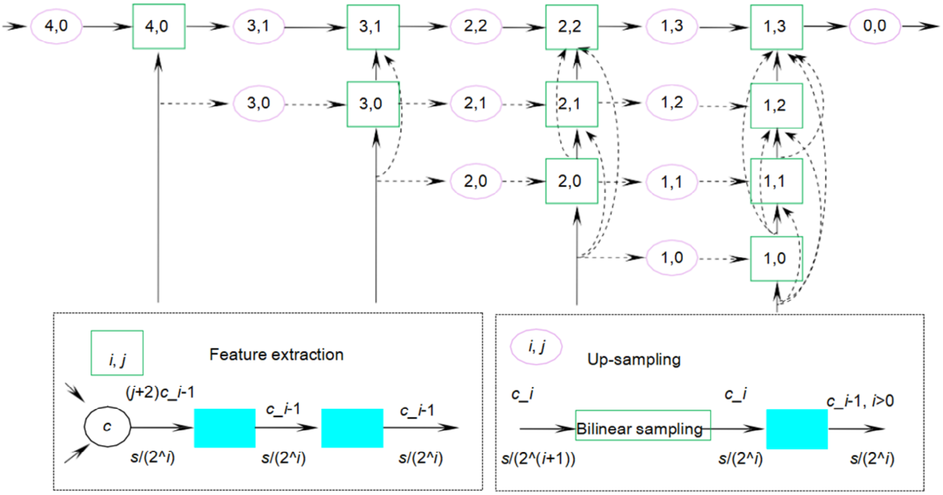


Figure 2 Decoder structure diagram (see online version for colours)



RGB encoders and surface normal encoders adopt a ResNet structure as their backbone, which is identical to each other. As shown in Figure 1, input data first passes through an initial block (consisting of convolution kernel 7 × 7, convolution layer with step 2, batch normalisation layer (BN) and ReLU activation layer). Then, a maximum pooling layer and four Res-layers are successively used to gradually reduce the resolution and increase the number of feature map channels. Each of the four Res-layers consists of n bottleneck blocks. The bottleneck block consists of three convolution layers with convolution kernels of 1 × 1, 3 × 3 and 1 × 1, respectively. ResNet has a variety of architectures, this paper uses ResNet-152. The number of feature mapping channels from c_0 to c_4 is 64, 256, 512, 1024, and 2048 respectively. The number of bottleneck blocks of the four Res-layers is 3, 8, 36, and 3 respectively. s represents the resolution of the input image.

The decoder (decoder square block in Figure 1), as shown in Figure 2, consists of two different types of modules (feature extractor and up-sampling layer), which decodes the encoded feature map to restore the resolution of feature mapping. In each layer of the decoder, feature layers generated in the corresponding coding stage are introduced respectively, and they are closely connected to achieve flexible feature fusion. The curved arrow represents the skip connection, and the bottom-up straight arrow represents the feature graph generated during the introduction of coding. Feature extractor is used to extract features and ensure that the resolution of feature image is unchanged. The upper sampling layer is used to improve the resolution and reduce the number of channels of feature image. The rectangular frames shared by feature extractor and up-sampling layer are composed of convolution layer, BN layer and ReLU layer with convolution kernel of 3 × 3 and step size 1 and Padding1.

3.1 Adaptive median filter region growth algorithm

In image analysis, the image quality has a great influence on the recognition effect. Therefore, preprocessing is necessary in image analysis. The main purpose of image preprocessing is to eliminate useless information and improve the reliability of feature extraction, image segmentation, matching and recognition. The road scene image has the characteristics of uneven grey distribution, the overall grey value is low, and the noise is much, so the operation of image segmentation may be affected. Therefore, before carrying out segmentation, we should first carry out roughly preliminary processing. Histogram equalisation is used in this paper. Histogram equalisation is a way to change the contrast quality of an image by adjusting it. In the program through the function histeq() to grey image histogram equalisation processing. It makes images clearer and more detailed.

Median filtering is a nonlinear image processing. It determines the grey scale of the centre pixel by taking the middle of the pixels in the neighbourhood from small to large. It has a better effect on pulse noise filtering, can ensure that the edge of the signal is not lost, the details of the image has a better protection, so the use of median filter is wide. The adaptive median filtering algorithm is as follows:

1 The first noise detection. Let the matrix $[x_{i,j}]$ be a digital noise image to be detected (it represents the positions of point *i* and *j*). Firstly, a noise identification matrix $[f_{i,j}]$ of the same dimension as $[x_{i,j}]$ is defined, which represents the noise points in the original image and initialises $[f_{i,j}]$ into a matrix of all zeros. If $[f_{i,j}] = 1$ exists in the identification matrix, $x_{i,j}$ is the pixel point of impulse noise or noise pollution. If there is $f_{i,j} = 0$ in the identity matrix $[f_{i,j}]$, $x_{i,j}$ indicates that the point is not polluted by noise.

According to the median idea, image pixels are classified according to the 3 × 3 template:

$$x_{i,j} = \begin{cases} N, x_{i,j} = \min\left(W\left[x_{i,j}\right]\right), \max\left(W\left[x_{i,j}\right]\right) \\ S, \min\left(W\left[x_{i,j}\right]\right) < x_{i,j} < \max\left(W\left[x_{i,j}\right]\right) \end{cases} \tag{1}$$

where *N* is the signal point. *S* is the noise point. The minimum value of $W[x_{i,j}]$ in the neighbourhood of a pixel point is represented by $\min(W[x_{i,j}])$, and the maximum value is represented by $\max(W[x_{i,j}])$.

2 Second noise detection.

$$S_{i,j} = \left\{x_{i+k,j+r} \mid k, r = 0, \pm 1, \cdots, \pm n\right\} \tag{2}$$

$$Average\left(S_{i,j}\right) = \frac{1}{(2n+1)^2} \sum_{k=-n}^{n} \sum_{r=-n}^{n} x_{i+k,j+r} \tag{3}$$

$$\left|x_{i,j} - Average\left(S_{i,j}\right)\right| > g_{i,j} \tag{4}$$

$$\left|x_{i,j} - Average\left(S_{i,j}\right)\right| \le g_{i,j} \tag{5}$$

$S_{i,j}$ is the grey value set of all elements in the filtering window. $Average(S_{i,j})$ is the average grey value of all pixels in the filtering window. Where $g_{i,j}$ is noise sensitivity coefficient, which is defined as:

$$g_{i,j} = \frac{1}{3}\sqrt{\sum_{k=-n}^{n}\sum_{k=-n}^{n}\left[x_{i+k,j+r} - Average\left(S_{i,j}\right)\right]^2} \tag{6}$$

The second noise detection method is used to judge all the pixels identified as noise points for the first time again. For the pixels satisfying formula (1)~(4), their identification $f_{i,j}$ does not change. For the pixel points satisfying formula (1)~(5), the identifier is changed to $f_{i,j} = 0$.

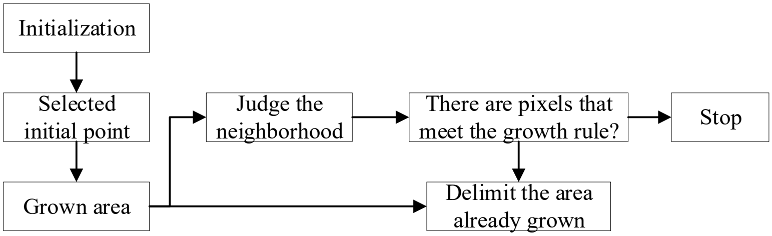
3 Image filtering. For $x_{i,j}$, which is determined as a noise point in the noise detection stage, the noise pollution degree ρ in the filtering window with $x_{i,j}$ as the centre point is calculated. ρ is the ratio of the total number of noise points in the filtering window to the total number of pixels.

$$G_{med}(i, j) = med\left\{A(i, j) \mid P(i, j) = 0\right\} \tag{7}$$

where $G_{med}(i, j)$ is the median of signal pixels in window A. A stands for filtering window. $P(i, j)$ is the element value at the corresponding position in the identification matrix. $A(i, j)$ is the grey value of pixels in window A.

The region growth method is to gather together the pixels or sub-regions that meet the conditions according to a certain rule. The process starts from a group of growing points, and the collection of adjacent pixels or regions with similar properties to the growing point becomes a new starting point. The process is repeated until the condition is not met. As for the judgment basis can generally be considered grey value, image texture. The steps are shown in Figure 3.

Figure 3 Flow chart of region growth method


Initialization
Selected initial point
Judge the neighborhood
There are pixels that meet the growth rule?
Stop
Grown area
Delimit the area already grown

Morphological operation is a method based on image shape. Two parameters need to be entered, raw image and structured elements, with corrosion and bloat being the most common. The corrosion and expansion operations are complementary. The process of corrosion followed by expansion is called open operation, which is mainly used to remove bright areas and separate objects at fine points, mainly to smooth the boundary.

And expansion before corrosion is closed operation which is to fill the area of the smaller black hole in the connection, iterative processing, it can also play a part of the smooth. For the sake of image effect, this paper adopts the method of open operation and close operation to carry on morphological processing to the image after region growth.

3.2 Surface normal estimator

Surface normal is an important attribute of geometric surface. It refers to the straight line (vector) that passes through a point on the surface and is perpendicular to the tangent plane of that point. Surface normals are widely used in 3D modelling to correct shadows and other visual effects caused by light sources. By processing the depth map into a normal map, objects of different planes and heights can be better distinguished.

Surface normals can be calculated by performing three filtering operations on inverse depth images or parallax images, namely two image gradient filters (one in horizontal and one in vertical directions) and an average/median filter. The SNE is shown in Figure 3, which is developed from the 3F2N method. Many experiments in Yu et al. (2020) have proved that better segmentation results can be obtained by using this deep data processing method. The estimation of surface normals can be transformed into the least square plane fitting estimation problem, which estimates the plane normals tangent to the surface at each point on the three-dimensional surface.

Firstly, each point $p = [u, v]^T$ on the depth map is connected with the point $P = [X, Y, X]^T$ on the space through the coordinate transformation equation (8), and then a local plane is fitted [as shown in equation (9)]. By convolving the anti-depth image Z with the horizontal image gradient filter and the vertical image gradient filter respectively, n_x, n_y are obtained, and substituted into the plane formula, so equation (10) is obtained:

$$K\begin{bmatrix} u \\ v \\ 1 \end{bmatrix} = \begin{bmatrix} f_x & 0 & u_0 \\ 0 & f_y & v_0 \\ 0 & 0 & 1 \end{bmatrix}\begin{bmatrix} X \\ Y \\ Z \end{bmatrix} \tag{8}$$

$$n_x X + n_y Y + n_z Z + d = 0 \tag{9}$$

$$n_x = -df_x g_x,\ n_y = -df_y g_y,\ n_z = d\frac{f_x \Delta X_i g_x + f_y \Delta Y_i g_y}{\Delta Z_i}(i = 1, \cdots, 8) \tag{10}$$

where K is the camera internal parameter matrix. $p_0 = [u_0, v_0]^T$ is the centre of the image. f_x and f_y are the focal length of the camera in pixels. $n = [n_x, n_y, n_z]^T$ is the surface normal. d is constant. g_x and g_y are the x and y derivatives of the inverse depth image Z.

Surface normal at estimation point P requires surrounding points information (also known as k neighbourhood). $N_P = [Q_1, \cdots, Q_k]^T$ is k nearest neighbour of P. Given any $Q_i \in N_P$, $Q_i - P = [\Delta X_i, \Delta Y_i, \Delta Z_i]^T$. can generate $k = 1, 2, \ldots, 8$ normalised surface normals $\overline{n}, \cdots, \overline{n}_k$, where:

$$\overline{n}_i = \frac{n_i}{\| n_i \|_2} = \left[\overline{n}_{xi}, \overline{n}_{yi}, \overline{n}_{zi}\right]^T \tag{11}$$

Since any normalised surface normal can be projected onto a sphere with radius 1 and centre (0, 0, 0). Therefore, we assume that the optimal surface normal $\hat{n}$ of *P* is also projected at some positions on the same sphere. Therefore, *k* normalised surface normals of the same point are normalised, and $\hat{n}$ is expressed in spherical coordinate system [equation (12)], and the optimal surface normals are obtained. Where $\theta \in [0, \pi]$ represents inclination angle and $\varphi \in [0, 2\pi]$ represents azimuth angle.

$$\hat{n} = [\sin\theta\cos\varphi, \sin\theta\sin\varphi, \cos\theta]^T \tag{12}$$

$$\varphi = \arctan\left(\frac{f_y g_y}{f_x g_x}\right) \tag{13}$$

$$\theta = \arctan\left(\frac{\sum_{i=1}^{k} \overline{n}_{xi}\cos\varphi + \sum_{i=1}^{k} \overline{n}_{yi}\sin\varphi}{\sum_{i=1}^{k} \overline{n}_{zi}}\right) \tag{14}$$

Assuming that the angle between any pair of normalised surface normals is less than $\pi / 2$, therefore, $\hat{n}$ can be obtained by minimising .

3.3 Fusion method

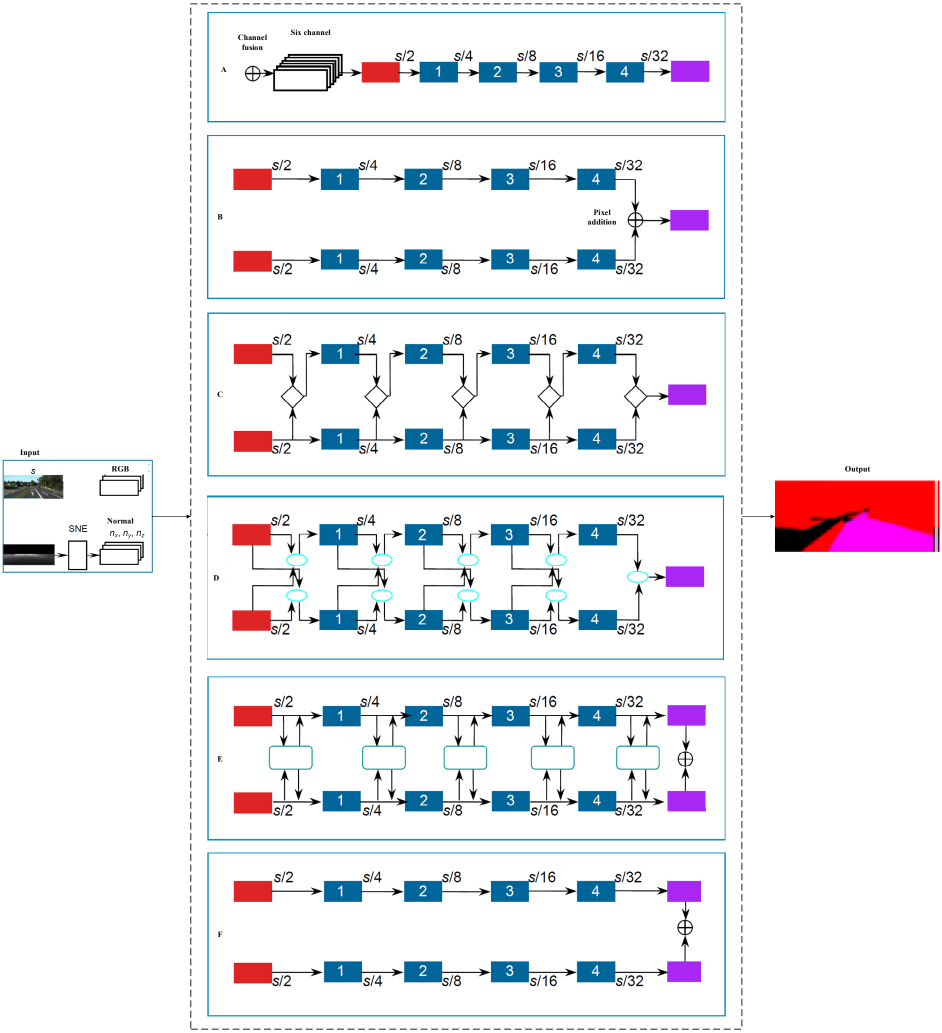
Multi-sensor information fusion can be divided into three levels: pixel level fusion, feature level fusion and decision level fusion according to the degree of abstraction of information processing. Aiming at the problem of how to adopt and at what stage fusion can achieve better results, this paper designs and tests a variety of fusion strategies (as shown in Figure 4). Pixel-level fusion belongs to the underlying data fusion method (such as fusion method A). The original observation information of the two sensors is directly fused after data preprocessing, and the six-channel observation data is entered into the encoder-decoder structure to extract features and conduct judgment and recognition.

Feature-level fusion belongs to the middle level and secondary fusion (such as fusion B, C, D and E). Representative features are extracted from the original observation information of the two sensors and appropriate features are selected for cross-fusion:

- Fusion B: The original data sets are respectively entered into the encoder structure to extract features, and then the two channels of feature data after encoding are fused. The fused data is sent to the decoder to obtain segmentation results.
- Fusion C: The original data sets are respectively entered into the encoder network structure, and cross method 1 (diamond box in Figure 4) is adopted in the five stages of the encoder. As shown in Figure 5(a), information supplement is made for RGB feature graph.
- Fusion D: The original data sets are respectively entered into the encoder network structure, and crossover method 2 (elliptic box in Figure 4) is adopted in the five stages of the encoder. As shown in Figure 5(b), information supplement is made for RGB feature graph.
- Fusion E: The original data sets are respectively entered into the encoder network structure, and cross method 3 (rounded rectangular box in Figure 4) is adopted in the five stages of the encoder, as shown in Figure 5(c). The fusion method is the

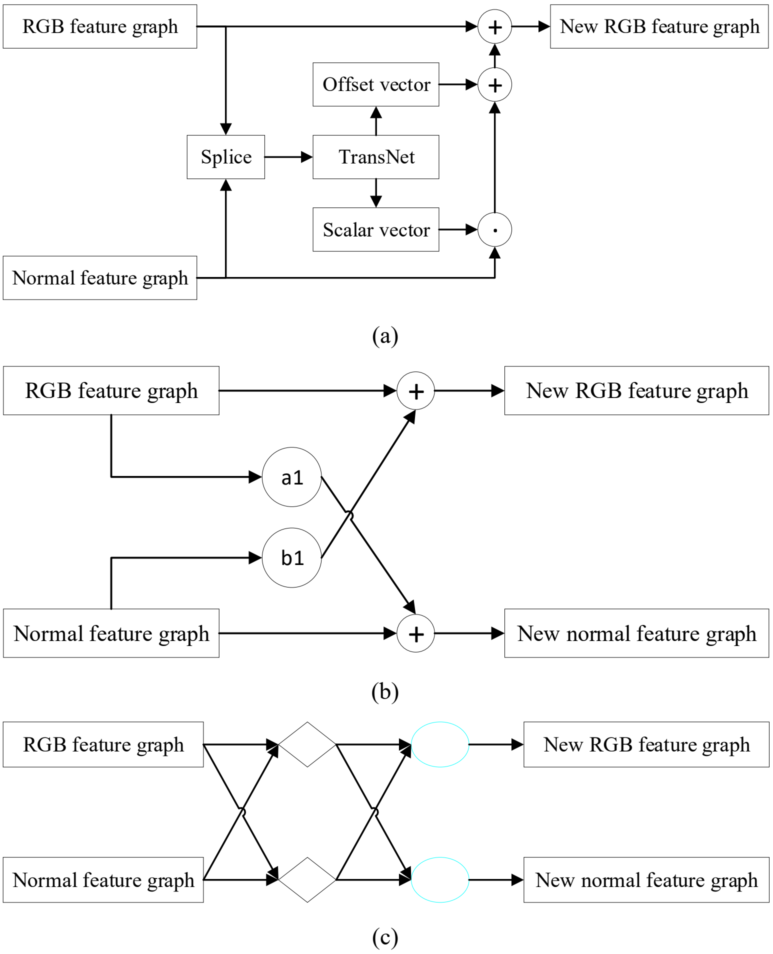
synthesis of schemes C and D. The normal feature is spliced with RGB feature channel, and two parameters and are obtained through training and learning. According to these two parameters, the transformed normal data feature map is obtained. The transformed RGB feature map is superimposed with RGB feature map. Similarly, the transformed normal feature map is obtained. Then the transformed normal feature graph is multiplied by the trainable parameter B again, and finally it superimposes with the transformed RGB feature graph to obtain the new RGB feature graph. In the other way, the new normal feature map after fusion can be obtained similarly. Then, the two channels of fusion data are sent to the decoder structure for restoration. Finally, the fusion is performed again in Sigmoid layer.

Figure 4 Network structure with different fusion strategies (see online version for colours)



Decision level fusion belongs to high-level and secondary fusion (such as fusion F), and the output is a joint decision result. Theoretically, this kind of joint decision is better than the decision based on single sensor. The two sensor data sets are entered into the encoder network respectively, spliced after decoding, and then fused in the Sigmoid layer to obtain the segmentation result.

Figure 5 Cross fusion method, (a) cross method 1 (b) cross method 2 (c) cross method 3 (see online version for colours)



4 Experiments and analysis

Experimental data is KITTI road data set, which consists of three subsets: training set (289 images), verification set (32 images), and testing set (290 images). A validation set is a set of images used for model validation in the training set. KITTI provides truth values for adjusting the model hyperparameters and evaluating the model capabilities. The test set is only used to evaluate the performance of the final model. KITTI does not provide the truth value and requires the researcher to provide the test results. The test results are compared with the truth value by KITTI, which ensures the fairness of the comparison between different methods. The KITTI image consists of three scenarios: unmarked urban (UU), urban mark (UM), and urban multi-marked motorway (UMM). There are five evaluation indexes: accuracy (ACC), precision (P), recall (R), F1 (F1-score) and PR curve (AP).

$$ACC = \frac{n_{tp} + n_{tn}}{n_{tp} + n_{tn} + n_{fp} + n_{fn}} \tag{15}$$

$$P = \frac{n_{tp}}{n_{tp} + n_{tn}} \tag{16}$$

$$R = \frac{n_{tp}}{n_{tp} + n_{fn}} \tag{17}$$

$$F1 = 2 \cdot \frac{P \cdot R}{P + R} \tag{18}$$

$$IoU = \frac{n_{tp}}{n_{tp} + n_{fp} + n_{fn}} \tag{19}$$

$$AP = \int_0^1 P(R)dR \tag{20}$$

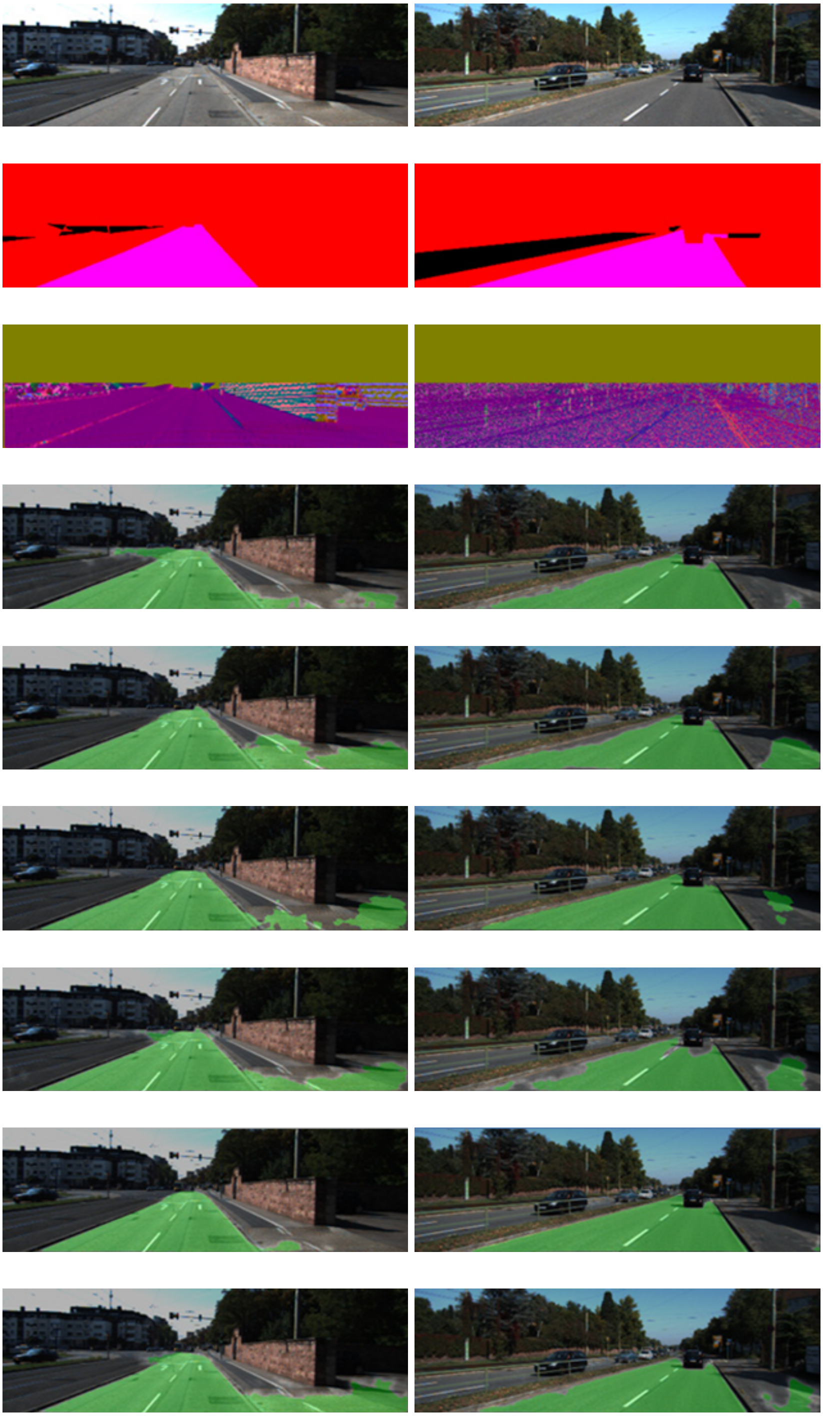
where n_{tp}, n_{tn}, n_{fp} and n_{fn} are the true positive pixel numbers, true negative pixel numbers, false positive pixel numbers and false negative pixel numbers in all images. AP (average accuracy) is the area under the PR curve (with recall as the horizontal axis and precision as the vertical axis). P(R) is the accuracy corresponding to different recall rates.

In addition, stochastic gradient descent with momentum (SGDM) optimiser is used to minimise the loss function, and the initial learning rate is set to 0.1. An early stop mechanism is used on the validation set to avoid over-fitting, and then the performance is quantified using the test set. The experiment is divided into two parts: first, the segmentation results of different fusion methods are compared on the same basic network structure to determine the best fusion method. Secondly, the segmentation effect of the proposed method is compared with other road segmentation methods to verify the road segmentation performance of the proposed method.

4.1 Comparison of various fusion schemes

The comparison of various fusion schemes is carried out on the verification set image, and each index is obtained by comparing obtained values with the truth values. The input data of the network are RGB images collected by camera and depth images obtained by lidar. The surface normal estimation of depth data is realised in the data preprocessing. Different fusion methods are adopted to supplement the feature information. Encoder-decoder structure is used to extract features and perform road segmentation.

Figure 6 Examples of experimental results of different fusion methods (see online version for colours)



Note: From first row-ninth row: raw image, truth, normal image, fusion A~fusion F.

Table 1 shows the performance indexes and Loss values of experimental results obtained on the verification set using different fusion methods. Compared with pixel-level fusion (fusion A) and decision level fusion (fusion F), accuracy, precision, F1-score and IoU of fusion A are 0.2%, 2.5%, 0.6% and 1% higher than those of fusion F, respectively, and only recall is 1.4% lower than it. In all feature-level fusion methods, fusion E has excellent performance in all aspects of performance indicators, Loss is only 0.022, accuracy is improved to 99.6%, Precision is improved to 98.1%, Recall increases by 1.9%, F1 increases by 2.8%, and IoU increases to 97.0%.

Table 1 Performance comparison between different methods

	Loss	*ACC*	*P*	*R*	*F1*	*IoU*
Fusion A	0.049	0.986	0.961	0.948	0.956	0.913
Fusion B	0.065	0.984	0.943	0.951	0.948	0.899
Fusion C	0.050	0.987	0.945	0.969	0.958	0.917
Fusion D	0.047	0.985	0.953	0.946	0.950	0.904
Fusion E	*0.022*	*0.996*	*0.981*	*0.988*	*0.986*	0.970
Fusion F	0.058	0.984	0.936	0.962	0.950	0.903

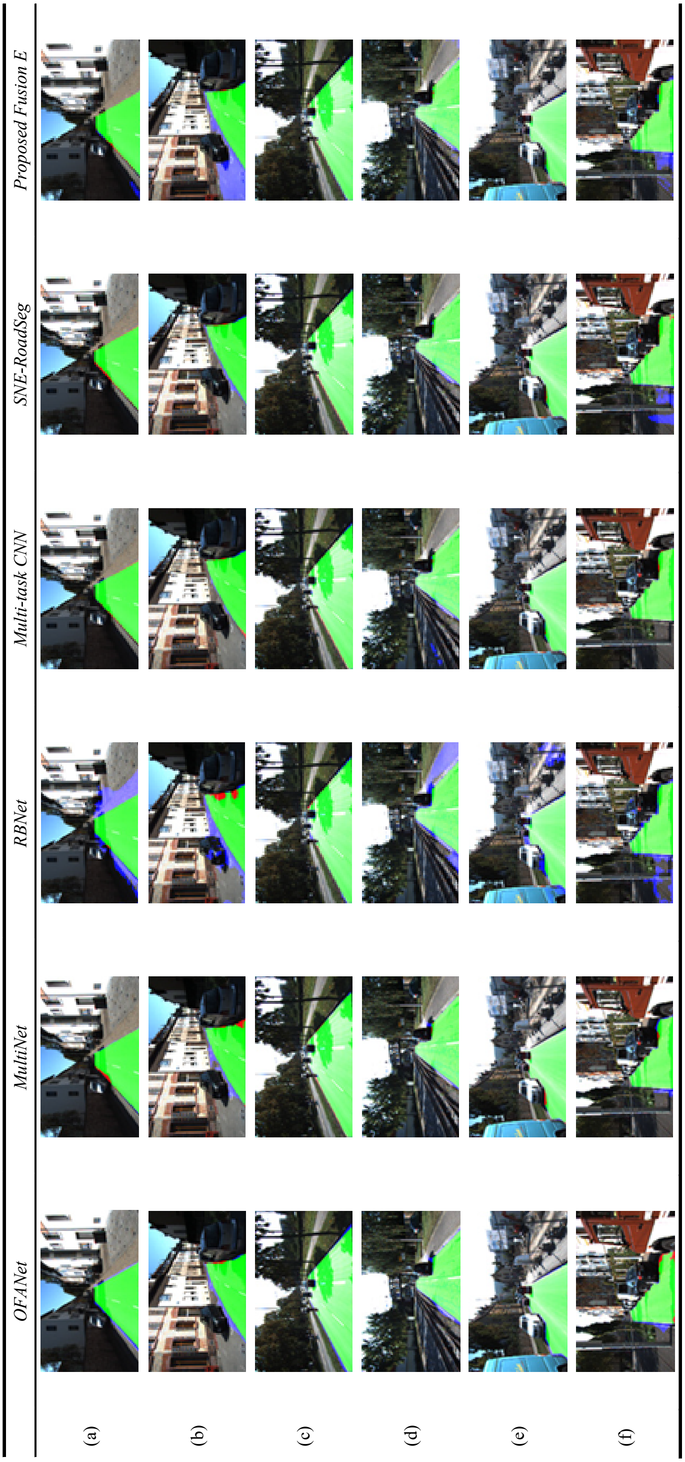
Existing 2D road segmentation methods mostly use data information of lidar to supplement RGB image information, while crossover method 3 can supplement both characteristic information of the two channels, placing the two sensor data in an equally important position. The combined form features of original features increase the feature dimension, improve the accuracy of target segmentation, and solve the problems of unstable pixel-level fusion subject to environmental noise and time-consuming. The decision level fusion has good error correction, can eliminate the error caused by a single sensor, and has a good segmentation speed. The combination of the two can improve the accuracy of segmentation and have good error correction.

Figure 6 is an example of segmentation results of different fusion methods on the same road map. By comparing multiple groups of images, it can be seen that the fusion E segmentation result proposed in this paper is closest to the truth graph, and the road contour segmentation is relatively complete without too many false detection areas. For the pavement at the same level, the distant intersection area and the area around the vehicle, fusion E can eliminate the non-road area cleanly.

4.2 Comparison with other methods

Figure 7 shows the test results for several typical scenarios in the KITTI dataset. The proposed fusion method (fusion E) is compared with OFANet, MultiNet, RBNet, multi-task CNN, and SNE-RoadSeg. The first column is the segmentation result of OFANet. The second column is the segmentation result of MultiNet. The third column is the result of RBNet. The fourth column is the segmentation result of Multi-task CNN. The fifth column is the segmentation result of SNE-RoadSeg, and the sixth column is the segmentation result of fusion E. Figures 7(a) and 7(b) are UM scenarios, Figures 7(c) and 7(d) are UMM scenarios, and Figures 7(e) and 7(f) are UU scenarios. The green area is the correct driving area (true positive). The blue areas correspond to missing driving areas (false detection areas). The red areas represent areas of false driving (false detection).

Figure 7 Example of KITTI dataset experimental results (see online version for colours)



For Figure 7(a), OFANet detects that the green area is more complete, and there are few red error detection areas, but there is a circle of blue error detection areas at the edge of the road. SNE-RoadSeg has the least number of blue error detection areas and a small number of red error detection areas. Fusion E has a small number of blue error detection areas in the shadow, a small number of red error detection areas in the position close to the vehicle, and a relatively complete green area. As for Figure 7(b), although fusion E has misjudged the pedestrian area under the vehicle, the green area is the most complete and the junction with the vehicle on the right is also well handled. Other methods all have a small amount of red or blue areas. For Figure 7(c), the detection results of all methods are relatively ideal with very few error detection and error detection areas. For Figure 7(d), the fusion E, OFANet, RBNet, Multi-task CNN have the best results, and the rail area is basically completely eliminated.

For Figure 7(e), it can be seen that fusion E handles the junction between vehicle and road very well. The green road area is around the edge of the vehicle, and there is basically no red error detection area. Other methods more or less have some error areas or error areas. Figure 7(f) is the same. The detection on the right part is relatively complete, although there is a small amount of wrong detection in the pedestrian area on the left. Although the detection results of multi-task CNN are relatively complete, there are too many blue error detection areas. Overall consideration, fusion E is very good for the junction between road and vehicle.

In the fusion scheme E, trainable parameters are cross-fused to carry out feature-level fusion of image and normal data, and the segmentation information of two sensors is fused by comprehensive use of dense texture information of image data and direction information of normal data, which effectively reduces the false detection rate of road segmentation.

The proposed fusion method (fusion E) in this paper is compared with the above five methods in different scenarios. Table 2 gives a quantitative comparison of several methods on the test set. As can be seen from the data in Table 2, OFANet and multi-task CNN (based on image segmentation) methods have high recall, which can exceed 98% in UMM scenario, but they are unsatisfactory in precision. It shows that there are many correct road pixels detected by image-based segmentation methods, but there are many misjudgments. The segmentation method based on point cloud image fusion has good performance in MaxF(Max F1-score), AP (average precision) and precision. Recall is slightly inferior, indicating that a higher proportion of roads detected by the multi-data fusion model are real roads with a small number of missed detection. The results show that multi-data fusion can significantly reduce road misjudgement.

In the segmentation method based on point cloud image fusion, compared with SNE-RoadSeg using feature fusion, all aspects of performance of fusion E (cross method 3) in UM and UU scenarios are improved, while AP is improved by 0.28% in UMM scenarios. Recall increases by 0.22%, Precision decreases by 0.95%, and MaxF decreases by 0.37%. Compared with SNE-RoadSeg, fusion E method has the highest AP value in all scenarios. In the UU scenario, recall is similar, but other aspects are insufficient. Precision reflects the specific gravity of real roads in positive examples judged by the model, reflecting the accuracy of detection. The precision of fusion E is lower than that of SNE-RoadSeg, indicating that there are many misjudgements in pixels judged as roads. Recall reflects the proportion of positive examples correctly judged as roads in the total real roads, reflecting the integrity of detection. Both methods are 96.05%, indicating that the number of pixels correctly judged as roads is basically the same. For road detection tasks, the higher AP denotes the fewer detection errors. The decrease of Precision indicates that fusion E method has road misdetection. As can be seen from Figures 7(a) and 7(b), in the UM scenario, vehicles appear in non-road areas with the same height as the road, and the detection results show serious deviations. It can be seen from Figures 7(c) and 7(d) that in UMM scenario, the detection results are relatively good when the road surface is complicated. In addition, the AP of fusion E is improved, indicating that our fusion E method has the situation of road misdetection. As can be seen from Figures 7(a) and 7(b), in the UM scenario, vehicles appear in non-road areas with the same height as the road, and serious deviations occur in the detection results. It can be seen from Figures 7(c) and 7(d) that in UMM scenario, the detection results are relatively good when the road surface is complicated. In addition, fusion E is improved in AP, indicating that crossover method 3 improves the model performance, but it is still insufficient in the case of individual road and sidewalk heights are the same and there are confounding factors.

Table 2 The KITTI road benchmark results

	Method	*MaxF/%*	*AP/%*	*P/%*	*R/%*
UM	OFANet	92.19	83.84	87.98	96.83
	MultiNet	94.10	93.35	94.62	93.59
	RBNet	94.88	91.53	95.27	94.48
	Multi-task CNN	86.06	81.39	77.51	96.75
	SNE-RoadSeg	96.53	93.78	96.70	96.37
	Fusion E	*95.83*	*95.23*	*95.98*	*95.70*
UMM	OFANet	95.54	89.21	92.89	98.35
	MultiNet	96.26	95.47	95.90	96.62
	RBNet	96.17	93.60	95.91	96.42
	Multi-task CNN	91.26	87.56	85.19	98.26
	SNE-RoadSeg	97.58	95.74	97.43	97.72
	Fusion E	*96.82*	*95.90*	*96.44*	*97.21*
UU	OFANet	92.73	83.23	89.08	96.69
	MultiNet	93.80	92.66	94.35	93.25
	RBNet	93.32	89.29	92.92	93.71
	Multi-task CNN	80.56	75.98	68.74	97.30
	SNE-RoadSeg	96.14	93.14	96.33	95.94
	Fusion E	*95.49*	*93.34*	*95.06*	*95.94*

5 Conclusions

This paper studies the road segmentation method based on the fusion of point cloud and image data, and designs a variety of pixel-level, feature-level and decision-level fusion schemes, especially four cross fusion schemes in feature-level fusion. The KITTI data set is used to carry out the experimental verification of various fusion methods. The fusion scheme E can better obtain the feature information of images and normals, and has the best road segmentation effect. Compared with other road detection methods, the optimal fusion method proposed in this paper has the advantage of average detection accuracy and better overall performance.

Acknowledgements

This work was supported by project Name: Research on road scene segmentation based on CNN, No. 222102210318.

References

Caltagirone, L., Bellone, M., Svensson, L. et al. (2019) 'LIDAR-camera fusion for road detection using fully convolutional neural networks', *Rob. Auton. Syst.*, Vol. 111, pp.125–131, https://doi.org/10.1016/j.robot.2018.11.

002.

Chen, Z. and Chen, Z.J. (2017) 'RBNet: a deep neural network for unified road and road boundary detection', *Proceedings of the 24th International Conference on Neural Information Processing*.

Chen, L.C., Zhu, Y., Papandreou, G., Schroff, F. and Adam, H. (2018) 'Encoder-decoder with atrous separable convolution for semantic image segmentation', in Ferrari, V., Hebert, M., Sminchisescu, C. and Weiss, Y. (Eds.): *Computer Vision – ECCV 2018. ECCV 2018. Lecture Notes in Computer Science()*, Vol. 11211. Springer, Cham, https://doi.org/10.1007/978-3-030-01234-2_49.

Chen, Z., Zhang, J. and Tao, D. (2019) 'Progressive LiDAR adaptation for road detection', in *IEEE/CAA Journal of Automatica Sinica*, May, Vol. 6, No. 3, pp.693–702, DOI: 10.1109/JAS.2019.1911459.

Fan, R., Wang, H., Cai, P. and Liu, M. (2020) 'SNE-RoadSeg: incorporating surface normal information into semantic segmentation for accurate freespace detection', in Vedaldi, A., Bischof, H., Brox, T. and Frahm, JM. (Eds.): *Computer Vision – ECCV 2020. ECCV 2020. Lecture Notes in Computer Science()*, Vol. 12375, Springer, Cham, https://doi.org/10.1007/978-3-030-58577-8_21.

Gao, H., Chen, F., Hao, Z. et al. (2021) 'Adaptive finite-time trajectory tracking control of autonomous vehicles that experience disturbances and actuator saturation', *IEEE Intelligent Transportation Systems Magazine*, No. 99, pp.2–13.

Huang, G., Liu, Z., Van Der Maaten, L. and Weinberger, K.Q. (2017) 'Densely connected convolutional networks', *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp.2261–2269, DOI: 10.1109/CVPR.2017.243.

Jaritz, M., Charette, R.D., Wirbel, E., Perrotton, X. and Nashashibi, F. (2018) 'Sparse and dense data with CNNs: depth completion and semantic segmentation', *2018 International Conference on 3D Vision (3DV)*, pp.52–60, DOI: 10.1109/3DV.2018.00017.

Kim, Y., Patel, S., Kim, H. et al. (2021) 'Ultra-low power and high-throughput SRAM design to enhance AI computing ability in autonomous vehicles', *Electronics*, Vol. 10, No. 3, p.256.

Oeljeklaus, M., Hoffmann, F. and Bertram, T. (2018) 'A fast multi-task CNN for spatial understanding of traffic scenes', *Proceedings of the 2018 21st International Conference on Intelligent Transportation Systems*.

Schlosser, J., Chow, C.K. and Kira, Z. (2016) 'Fusing LIDAR and images for pedestrian detection using convolutional neural networks', *Proceedings of 2016 IEEE International Conference on Robotics and Automation*.

Shi, Q., Yin, S., Wang, K., Teng, L. and Li, H. (2022) 'Multichannel convolutional neural network-based fuzzy active contour model for medical image segmentation', *Evolving Systems*, Vol. 13, No. 4, pp.535–549, https://doi.org/10.1007/s12530-021-09392-3.

Teichmann, M., Weber, M., Zöllner, M. et al. (2018) 'MultiNet: real-time joint semantic reasoning for autonomous driving', *Proceedings of 2018 IEEE Intelligent Vehicles Symposium*.

Van Gansbeke, W., Neven, D., De Brabandere, B. and Van Gool, L. (2019) 'Sparse and noisy LiDAR completion with RGB guidance and uncertainty', *2019 16th International Conference on Machine Vision Applications (MVA)*, pp.1–6, DOI: 10.23919/MVA.2019.8757939.

Wang, T-H., Hu, H-N., Lin, C.H., Tsai, Y-H., Chiu, W-C. and Sun, M. (2019) '3D LiDAR and stereo fusion using stereo matching network with conditional cost volume normalization', *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp.5895–5902, DOI: 10.1109/IROS40897.2019.8968170.

Wang, Y., Yan, G., Zhu, H. et al. (2021) 'VC-Net: deep volume-composition networks for segmentation and visualization of highly sparse and noisy image data', *IEEE Transactions on Visualization and Computer Graphics*, Vol. 27, No. 2, pp.1301–1311.

Xiao, X., Lian, S., Luo, Z. and Li, S. (2018) 'Weighted Res-UNet for high-quality retina vessel segmentation', *2018 9th International Conference on Information Technology in Medicine and Education (ITME)*, pp.327–331, DOI: 10.1109/ITME.2018.00080.

Xu, H., Yang, M., Deng, L. et al. (2021) 'Neutral cross-entropy loss based unsupervised domain adaptation for semantic segmentation', *IEEE Transactions on Image Processing*, Vol. 30, pp.4516–4525, DOI: 10.1109/TIP.2021.3073285 [online] https://ieeexplore.ieee.org/document/9409715.

Yin, S., Meng, L. and Liu, J. (2019) 'A new Apple segmentation and recognition method based on modified fuzzy C-means and Hough transform', *Journal of Applied Science and Engineering*, Vol. 22, No. 2, pp.349–354.

Yu, J., Li, H. and Yin, S. (2020) 'Dynamic gesture recognition based on deep learning in human-to-computer interfaces', *Journal of Applied Science and Engineering*, Vol. 23, No. 1, pp.31–38.

Zhang, S., Zhang, Z., Sun, L. et al. (2019) 'One for all: a mutual enhancement method for object detection and semantic segmentation', *Applied Sciences*, Vol. 10, No. 1, p.13.

Zhang, Y. and Funkhouser, T. (2018) 'Deep depth completion of a single RGB-D image', *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp.175–185, DOI: 10.1109/CVPR.2018.00026.

Zhou, Z., Rahman Siddiquee, M.M., Tajbakhsh, N. and Liang, J. (2018) 'UNet++: a nested U-Net architecture for medical image segmentation', in *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support. DLMIA ML-CDS 2018 2018. Lecture Notes in Computer Science()*,Vol. 11045, Springer, Cham, https://doi.org/10.1007/978-3-030-00889-5_1.