ISSN 2959-6157

Semantic segmentation based on multimodal information is used for object detection during autonomous driving at night

Haoqi Wu

Mechanical Design, Manufacturing and Automation, Shaanxi University of Technology, Hanzhong, Shaanxi, 732000, China

Tutor: Zhiqiang Liu, School of Mechanical Engineering, Shaanxi University of Technology

Email: 13892060678@163.com

Abstract:

Currently, it is difficult for autonomous driving to detect objects accurately enough at night, mainly due to the poor light at night. The laser camera's captured RGB image information is less effective than during the day, and the image content is seriously affected. The blurred outline of the object and the decreased accuracy of semantic segmentation lead to missed detections. In view of this situation, this study proposes the following solutions: introducing point cloud data collected by lidar during data acquisition, combining it with the image information from the camera, to constitute precise multimodal three-dimensional information. Then, using a dual adversarial network to preprocess the data and U-net semantic segmentation to segment the multimodal 3D information. The simulation experiment is then used to test and evaluate the segmentation effect. After the completion of the training, the object information detection is applied during the autonomous driving of the vehicle. This method, when compared with the general method, has accurate detection, is independent of the intensity of the light, and has the advantage of good universality.

Keywords: Autonomous driving, multimodal information semantic segmentation, object detection, Dual U-net network, Dual-way adversarial network

1 Introduction

In today's society, autonomous driving technology, as an important part of the intelligent transportation system, is gradually changing the way we travel and life.[1] With the rapid development of artificial intelligence and machine learning technology, autonomous vehicles have moved from proof-of-concept to practical application. However, despite significant advances in autonomous driving technology, object detection during night driving remains an extremely challenging problem. Due to the poor light conditions at night, the effect of the traditional laser camera is limited when capturing images, resulting in the unclear outline of the object [2], which affects the accuracy of semantic segmentation and increases the risk of missed and false detection.

Object detection in autonomous driving systems is crucial to ensure driving safety. Accurate object detection not only helps autonomous vehicles identify and avoid potential obstacles, but also optimizes path planning and improves driving efficiency. However, existing technologies tend to perform poorly in low-light environments, which limits the use of autonomous vehicles at night or in bad weather conditions. Therefore, developing a technology that can accurately detect objects under various light conditions is a key step in improving the overall performance of autonomous driving systems.

In this study, we propose a solution based on multimodal information. By integrating data from lidar and cameras, we were able to build accurate 3 D information models. Moreover, using the data preprocessing and U-net semantic segmentation algorithm for multi-modal 3 D information [3][4], this study aims to improve the accuracy and robustness of object detection. After testing and evaluating the proposed method through simulation experiments, we find that this method can effectively improve the accuracy of object detection at night, not affected by the strength of light, and has a good universality.

In conclusion, the research content of this study is not only expected to solve a long-standing problem in autonomous driving —— the accuracy of nighttime object detection [5], but also provide new ideas and methods for the future development of autonomous driving technology. By introducing multimodal information processing technology and advanced deep learning algorithms, we be-

lieve that future autonomous vehicles will be able to show higher safety and reliability in a variety of environmental conditions.

2 Theoretical analysis

Convolutional Neural Network (CNN) is a deep learning algorithm that has achieved remarkable success in image recognition, video analysis, natural language processing and other fields. CNN is inspired by biological nervous systems, especially the visual cortex, which are able to effectively process grid-like data, such as images (2D grids) and time series data (1D grids).

The following are the key components and features of the CNN: 2.1-1. Input layer: This is the starting point of the network and usually receives the input image of the original pixel value. 2.1-2. Convolutional layers: These layers extract features by applying a convolution kernel or filter to the input. Each filter is responsible for detecting specific low-level features such as edges or color patches from the input volume. 2.1-3. Activation function: A nonlinear activation function, such as the ReLU (Rectified Linear Unit), is usually applied after the convolution. This helps introduce nonlinearity that allows the CNN to capture complex patterns. 2.1-4. Pooling layer: The pooling layer follows the convolution layer, and is used to reduce the spatial dimension (length and width) of the feature map , improve the calculation efficiency, and control the overfitting. The most common pooling operations are maximum pooling and average pooling. 2.1-5. Fully connected layer: After multiple convolution and pooling layers, the fully connected layer is used to learn high-level combinations between features. This eventually leads to the last layer of the network, usually the output layer for classification. 2.1-6. Output layer: The design of the output layer will vary according to the task. For multi-class classification problems, the output layer usually contains the number of neurons equal to the number of classes and uses the softmax activation function to output the probability distribution. 2.1-7. Loss function: During the training process, the network is optimized by minimizing the loss function. For classification tasks, common loss functions include cross-entropy loss . 2.1-8. Backpropagation: By calculating the loss function gradient about the network weights, and using these gradients to update the weights to reduce the difference between predicted and true labels. 2.1-9. Batch normalization: This is a common technique to accelerate the training and improve the generalization ability of the model by using the data within the normalization layer. 2.1-10. Jump connections: In some deep networks, jump connections allow information to bypass some layers, help solve the gradient disappearance problem, and promote the training of deeper networks. The advantage of CNN lies in their ability to process high-dimensional data without sacrificing computational efficiency, while maintaining translation invariance to extract low-level to high-level features of images layer by layer, and finally perform tasks such as classification or regression.[7] (features can be correctly identified no matter where they appear in the input).

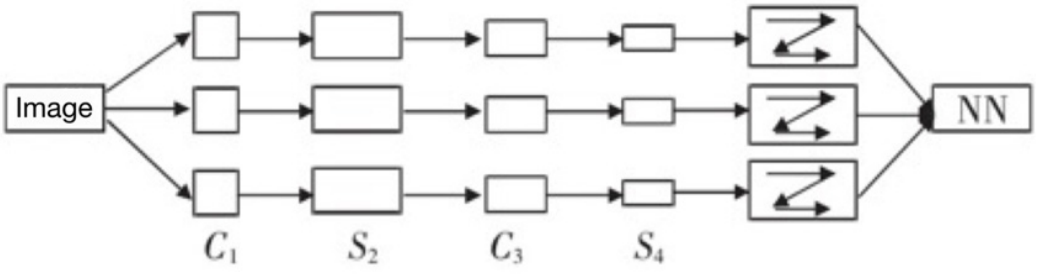
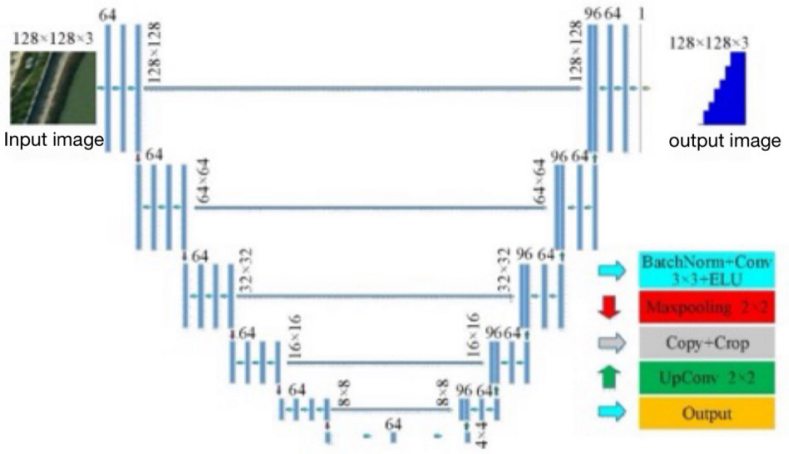


Figure 1 CNN schematic diagram [6]

In the field of image semantic segmentation, U-net is a classic image semantic segmentation network [8], which adopts a unique encoder-decoder structure and maintains the context information from encoder to decoder through jump connection (also known as identity mapping). This design enables U-net to efficiently utilize spatial information when processing images, especially suitable for biomedical image segmentation tasks with distinct boundaries and hierarchical structures. The main feature of the U-net is its “U” -shaped structure, including a contracted encoding path and an extended decoding path. In the encoding stage, the network gradually reduces the spatial size of the input image by a series of convolution layers. In the decoding stage, these features are gradually restored to the original spatial size by upsampling operation and stitched with the corresponding layers in the encoding stage to retain more spatial details. Moreover, U-net also adopts symmetric jump connections, which directly connect the activation map in the encoder to the corresponding layer in the decoder. This design allows the network to access a wider range of background information during the decoding phase, contributing to improve accuracy and

robustness of segmentation. Overall, U-net has become a breakthrough model in the field of image semantic segmentation due to its simple and effective structure. It has achieved excellent performance in a variety of complex image segmentation tasks, especially in the tasks that require precise positioning and segmentation. The network adopts the encoder-decoder structure and retains more spatial information through jumping connection, which has achieved remarkable achievements in the fields of medical image segmentation, and has wide application prospects for object detection during autonomous driving [9].



[10]

Figure 2 A Schematic diagram of the basic U-net image semantic segmentation network

2.3 dual against network: dual against network is a kind of technology for data preprocessing, through the introduction of confrontation learning, can effectively extract the useful features of data and information, early often used in accurate information preprocessing of face recognition, and the visible light and invisible light multimodal information fusion has greater help [11], reference dual against network will provide better for subsequent segmentation task input.

3 Design and optimization of the network model

3.1 model design

3.1.1 Model design idea

First, the objectives and performance indicators of the model need to be defined. These goals may include classification accuracy, detection speed, or generalization ability. After these targets are determined, you can start to choose the appropriate network layer type according to the characteristics of the problem, such as the convolution layer for image processing and the loop layer for sequence data processing. Also, determines the number of parameters per layer, the type of activation function, and whether to use pooling or normalization operations.

3.1.2 . Feature extraction and fusion

For the images captured by the camera, we can extract the feature [12] using the CNN. By adjusting the convolution kernel size, stride size and packing strategy, we can affect the size and receptive fields of the feature map, thus extracting features at different scales. For the point cloud data collected by lidar, PointNet can be used to use the point cloud processing algorithm (such as PointNet, PointNet + +, etc.) [13] to extract 3 D features, and convert the lidar data into point cloud representation. Series of structures to extract the local features in the space. Moreover, the voxelization of point cloud data is also an important step to transform the data in 3 D space into a two-dimensional tensor form that the model is able to handle.

3.1.3 . Multimodal information fusion

After obtaining feature expression from different sensors, the next step is to design a mechanism to combine this information effectively. We can use simple splicing operations, or we use more complex methods (such as attention mechanisms) to determine which features are more important to the current decision.

3.1.4 . Construction of encoder and decoder

The encoder is responsible for capturing the global features of the spatial information and squeezing them into a compact representation. In the U-net structure, the encoder reduces the resolution layer by layer to capture a larger range of spatial context information. The decoder, on the other hand, starts with the output of the encoder, gradually recovering the resolution of the original input and generating fine detailed features. In the dual U-net structure, each decoder is connected to the corresponding encoder, allowing for information transfer and sharing across different levels.

3.1.5 . Skip connection

The skip connection is one of the key components of U-net, which enables the semantic information at the high level

to be propagated directly out to the lower level for detail recovery. This cross-layer connectivity not only helps to maintain the integrity of the spatial information, but also improves the accuracy of the segmentation tasks.

3.1.6 . Optimization strategy and regularization

Choosing the appropriate loss function and optimization algorithm is crucial during training. Common loss functions include cross-entropy loss, mean square error (MSE), etc. The optimization algorithm can be randomized gradient descent Meanwhile, to prevent overfitting and accelerate convergence, we can also introduce some regularization techniques, such as Dropout, L1 / L2 regularization, etc. These techniques help to stabilize the training process and improve the generalization ability of the model.

3.1.7 . Hyperparameter adjustment and early stop strategy

During the training process, the hyperparameters also need to be constantly adjusted to accommodate different task conditions and data set characteristics. These hyperparameters include the learning rate, batch size, iteration number, etc. The early stop strategy is also a common optimization method, stopping training when the performance on the validation set no longer improves to avoid the occurrence of overfitting phenomenon. Through the detailed analysis and design of the above steps, an accurate and robust network model is constructed, which provides strong technical support for the autonomous driving system.

3.2 Optimization ideas Improvement in the semantic segmentation of U-net

The dual U-net network model is based on an improved version of U-net and combines the structure of two U-net networks, which can improve the performance and effect of image semantic segmentation task. The following are the design modules of the dual U-net network model:

Encoder: The dual U-net network contains two identical encoder parts, each encoder consists of multiple convolutional layers and pooling layers to extract image features and gradually reduce spatial resolution. Each encoder is structured similar to the encoder part [14] of the U-net.

3.2.1 . Decoder

The dual U-net network also contains two same decoder parts, each decoder consists of multiple upsampling layers and convolution layers, which is used to restore the feature map extracted by the encoder to the segmentation result of the original input image size. The structure of each decoder is also similar to the decoder part of U-net, and the upsampling layer and the convolution layer can be adopted to gradually restore the spatial resolution [15] of the feature map.

3.2.2 . Jump connection

In a dual U-net network, the jump connection is used to connect the corresponding layer of the encoder and the decoder to retain more spatial information and detailed features. Each decoder layer will be connected to the corresponding encoder layer, combining high-level semantic features with low-level semantic features through jumping connection, which helps to improve the accuracy of segmentation results.

3.2.3 Output layer

The output layer of a dual U-net network usually uses a convolutional layer, which outputs the segmentation results of the same size as the input image. The output layer can classify each pixel with the appropriate activation function (such as the sigmoid function) to generate the final segmentation result.

Through the above design module, the dual U-net network model can effectively combine the advantages of the two U-net networks to improve the performance and effect of the image semantic segmentation task.

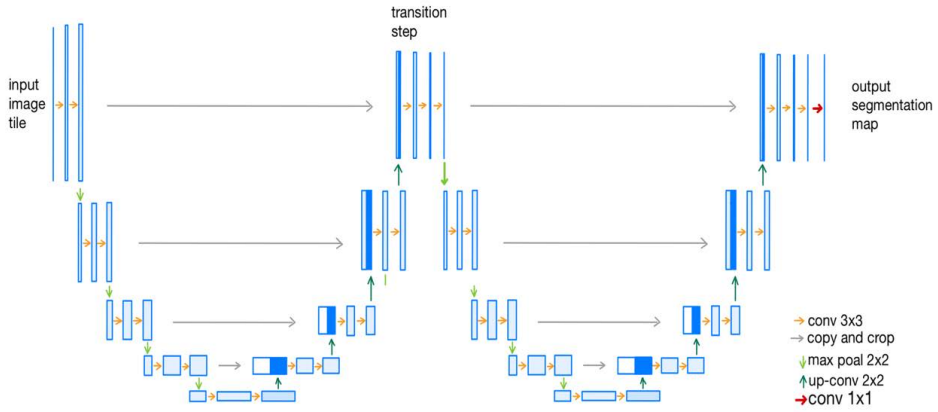


Figure 3 Schematic diagram of dual U-net semantic segmentation

4 Conclusion

In this study, we propose a solution based on multimodal information for the insufficient accuracy of object detection in nighttime autonomous driving scenarios. By introducing the point cloud data collected by lidar and combining it with the image information captured by the camera, an accurate multimodal three-dimensional information is constructed. Using the dual-way adversarial network for data preprocessing and U-net semantic segmentation algorithm, thus improving the accuracy and robustness of object detection. In addition, we also improve the U-net semantic segmentation algorithm to design a dual U-net network model, which significantly improves the performance and effect of the image semantic segmentation task through the symmetric design of the encoder and decoder. Each decoder layer is connected to the corresponding encoder layer, combining high-level semantic features with low-level semantic features through jumping connections, thus significantly improving the accuracy of segmentation results. Overall, this study presents a set of effective solutions to address the challenges of object detection in night autonomous driving through in-depth analysis and innovative design.

5 Reference

[1] Zhao Xiangmo National Key Research and Development Program (2021YFB2501200) team. Research progress in autonomous driving testing and evaluation technology [J]. Transportation engineering,2023,23(06):10-77.DOI:10.19818/j.cnki.1671-1637.2023.06.002.

[2] Yu Lijiao, Yu Bo, Li Chungeng, et al. Optimize the nighttime human and vehicle detection and identification of convolutional networks and low-resolution thermal imaging [J]. Infrared technology, 2020,42(07):651-659.

[3] Wang Bin, Chen Zhanlong, Wu Liang, et al. Road extraction of high-resolution remote sensing images of U-Net network with both connectivity [J]. Journal of Remote Sensing,2020,24(12):1488-1499.

[4] Jin Yufeng, Tao Ben. A Transformer-based fusion information-enhanced 3D object detection algorithm [J]. Journal of Instrumentation,2023,44(12):297-306.DOI:10.19650/j.cnki.cjsi.J2311940.

[5] Wang Zhongyu, Ni Xianyang, Shang Zhendong. Semantic segmentation of autonomous driving scenes using a convolutional neural network [J]. Optical and precision engineering,2019,27(11):2429-2438.

[6] Sun Zhijun, Xue Lei, Xu Yangming, et al. Review of deep learning studies [J]. Computer application research,2012,29(08):2806-2810.

[7] Zhang Xinyu, Gao Hongbo, Zhao Jianhui, et al. Summary of autonomous driving techniques based on deep learning [J]. Journal of Tsinghua University (Natural Science Edition),2018,58(04):438-444.DOI:10.16511/j.cnki.qhdxxb.2018.21.010.

[8] Yang Kang, Chen Li. Research on pedestrian detection based on convolutional neural network in autonomous driving [J]. Computer knowledge and technology,2020,16(25):22-24+30.DOI:10.14004/j.cnki.ckt.2020.2958.

[9] Zhang Shengjian, Mo Zewen. Autonomous driving road detection based on remote sensing images: a histogram equalization strategy [J]. Sensor world,2023,29(08):28-31.DOI:10.16204/j.sw.issn.1006-883X.2023.08.005.

[10] Su Jianmin, Yang Lanxin, Jing Weipeng. Semantic segmentation method for high-resolution remote sensing images based on U-Net [J]. And Computer Engineering and Application,2019,55(07):207-213.

[11] Tang Lili, Liu Gang, Xiao Gang. Infrared and visible light image fusion method based on the two-way cascade confrontation mechanism [J]. Photonics,2021,50(09):321-331.

[12] Zhou Feiyan, Jin Linpeng, Dong Jun. A Review of Convolutional Neural Network Research [J]. The Journal of Computer Science,2017,40(06):1229-1251.

[13] Jing Zhuangwei, Guan Haiyan, Zang Yufu, et al. Review of semantic segmentation research on point clouds based on deep learning [J]. And Computer Science and Exploration,2021,15(01):1-26.

[14] Cao Jingang, Yang Guotian, Yang Xiyong. Deep learning pavement crack detection based on the attention mechanism [J]. The Journal of Computer-Aided Design and Graphics,2020,32(08):1324-1333.

[15] Canina Wang, Chen Yi. Enhanced semantic dual decoder generation model for image repair [J]. Chinese Journal of Image and Graphs,2022,27(10):2994-3009.