Computer Modeling in Engineering & Sciences

Tech Science Press

DOI: 10.32604/cmes.2023.025119

check for updates

ARTICLE

Image Semantic Segmentation for Autonomous Driving Based on Improved U-Net

Chuanlong Sun, Hong Zhao*, Liang Mu, Fuliang Xu and Laiwei Lu

College of Mechanical and Electrical Engineering, Qingdao University, Qingdao, 266071, China

*Corresponding Author: Hong Zhao. Email: qdlizh@163.com

Received: 04 July 2022 Accepted: 06 September 2022

ABSTRACT

Image semantic segmentation has become an essential part of autonomous driving. To further improve the generalization ability and the robustness of semantic segmentation algorithms, a lightweight algorithm network based on Squeeze-and-Excitation Attention Mechanism (SE) and Depthwise Separable Convolution (DSC) is designed. Meanwhile, Adam-GC, an Adam optimization algorithm based on Gradient Compression (GC), is proposed to improve the training speed, segmentation accuracy, generalization ability and stability of the algorithm network. To verify and compare the effectiveness of the algorithm network proposed in this paper, the trained network model is used for experimental verification and comparative test on the Cityscapes semantic segmentation dataset. The validation and comparison results show that the overall segmentation results of the algorithm network can achieve 78.02% MIoU on Cityscapes validation set, which is better than the basic algorithm network and the other latest semantic segmentation algorithms network. Besides meeting the stability and accuracy requirements, it has a particular significance for the development of image semantic segmentation.

KEYWORDS

Deep learning; semantic segmentation; attention mechanism; depthwise separable convolution; gradient compression

1 Introduction

With the combination of Artificial Intelligence (AI) and automobile transportation, autonomous driving [1] has become one of the development strategies in many countries, which involves Global Positioning System (GPS), Computer Vision (CV) [2] and other advanced technologies. The perception system [3,4] is one of the indispensable parts of autonomous driving vehicles. Its perceptual adaptability and real-time performance in the environment can directly affect the safety and reliability of autonomous driving vehicles. While image semantic segmentation is one of the main tasks in the perception system, its effectiveness will directly affect the decision quality of the autonomous driving vehicle.

In recent years, due to the progress of large datasets, powerful computing power, complex network architectures and optimization algorithms, the application of deep learning in the field of image semantic segmentation has achieved major breakthroughs [5]. At present, the semantic segmentation

methods for autonomous driving mainly include traditional semantic segmentation methods and deep learning-based semantic segmentation methods.

Most of the traditional semantic segmentation methods are the early semantic segmentation methods, which were first applied to the medical field with simple scenes and obvious differences between the background objects. The main researches are: segmentation methods based on threshold [6–8], which classifies the image gray histogram by setting different gray thresholds, and the pixels whose gray values are in the same gray range are considered to belong to the same class and have a certain similarity, so as to achieve semantic segmentation; the edge-based image segmentation method [9], which compares the gray value difference between adjacent pixels, regards the points with large differences as boundary points and detects these points. The pixel points at the boundary are connected to form edge contours to achieve the segmentation of different regions; the region-based image segmentation method [10] segmented the image by obtaining the spatial information of the image. It classifies the pixels by the similarity features of the pixels and forms the region; image segmentation methods based on graph theory [11–13] convert the segmentation problem into graph division and complete the segmentation process by optimizing the objective function. Most of the traditional semantic segmentation methods use the surface information of images, which is not suitable for segmentation tasks that require a lot of semantic information and cannot meet the actual needs. Most of the traditional semantic segmentation methods only use the shallow-level information of the image, which cannot meet the needs of current research. At present, it is often used as a preprocessing step in image processing to obtain key feature information of the image and improve the efficiency of image analysis.

In the field of semantic segmentation based on deep learning, convolutional neural network has become an important means of image processing, which can fully utilize the semantic information of images to achieve semantic segmentation. To cope with the increasingly complex challenges of image segmentation scenarios, a series of deep learning-based image semantic segmentation methods have been proposed to achieve more accurate and efficient segmentation, and further promote the application scope of image segmentation. Image semantic segmentation based on region classification and image semantic segmentation based on pixel classification are the current mainstream deep learning-based semantic segmentation methods. The former divides the image into a series of target candidate regions and classifies the target region through the deep learning algorithm, which can avoid the generation of superpixels and improve the efficiency of image segmentation effectively. The former is represented by MPA [14], DeepMask [15], etc. Through pixel classification, the latter directly uses deep neural networks of an end-to-end structure for segmentation, which avoid the problems caused by the defects of the candidate-regions algorithm. The latter is represented by DeepLab [16], ICNet [17], U-Net [18], etc.

As one of the representatives in the field of semantic segmentation algorithms, U-Net uses the "encoder-decoder" structure to perform feature fusion between feature maps, so that the shallow convolutions can focus on texture features and the deep convolutions focus on image essential features. This paper selects the U-Net semantic segmentation algorithm as the basic algorithm for research. In recent years, in the study of U-Net semantic segmentation network, Huang et al. [19] and others used full-scale skip connections to replace the long connections of the U-Net model, which combined high-level semantic information with low-level semantic information to obtain more segmentation accuracy. Zhong et al. [20] introduced the DenseNet module and applied it to the convolutional layer to improve the network's ability to extract features in small areas and avoid the problem of gradient disappearance. The CRF 3D-U-Net network proposed by Hou et al. [21] used 3D-U-Net and a fully connected conditional random field to segment images coarsely and finely, respectively, which enables

the network to improve the correlation between pixels. Although the improvement of the U-Net model in the above research has a positive effect, it increases the complexity and running cost of the algorithm model. It does not take advantage of the relationship between each feature map. Meanwhile, because of the redundancy of U-Net itself, it is easy to have low segmentation and positioning accuracy when used in autonomous driving scenarios. Therefore, to meet the complex environment of autonomous driving scenes and real-time requirements, this paper does the following work based on U-Net:

(1) Change the convolution method and update the standard convolution to depthwise separable convolution, which reduces the calculation parameters and realizes the separation of channels and regions.

(2) The attention mechanism is introduced in this paper, which enables the network to learn weight information from the feature channel dimension: the weight of the feature channel with good network performance is improved, and the weight of the feature channel with poor network performance is suppressed, so that the training efficiency can be improved.

(3) For the segmentation accuracy and generalization ability of the semantic segmentation algorithm, this paper operates on the gradient directly and smoothes the gradient curve by using a suitable gradient compression method. Meanwhile, gradient compression can regularize the weight space and output feature, thereby improving the performance of the detection algorithm.

2 Network Structure

2.1 Lightweight Feature Extraction Network

Fig. 1 shows the structure of the improved U-Net lightweight feature extraction network in this paper. The feature extraction network is mainly composed of the Depthwise Separable Convolution block DSC-R (Depthwise Separable Convolution-ReLu), SE Attention block. Maximum Pooling layer, the Upsampling block, and the Skip Connection. Among the component, compared with the basic U-Net algorithm structure, the improvements made in this paper mainly include: the original standard convolution method is replaced by depthwise separable convolution, and the SE attention mechanism module is introduced, which can improve the accuracy of feature extraction while realizing the lightweight of the algorithm model.

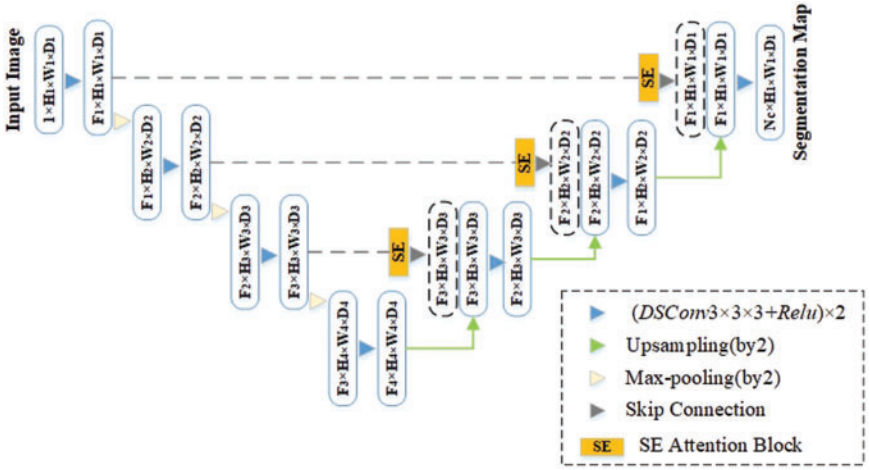


Figure 1: Improved U-NET feature extraction network

2.2 Activation Function

Activation function plays an essential role for neural network models to learn and understand the complex and nonlinear input characteristics. In this paper, the widely-used LeakyReLU function is used as the activation function.

Although the traditional ReLU activation function has a faster calculation speed and convergence speed, however, when the input is negative, the neuron cannot update the parameters because of its 0 value output. As is shown in Fig. 2, compared with traditional ReLU function, the Leaky ReLU function introduces the Leaky value in the negative half of the input, avoiding 0 value derivatives, which can cause neurons to fail to update parameters, when the input is negative.

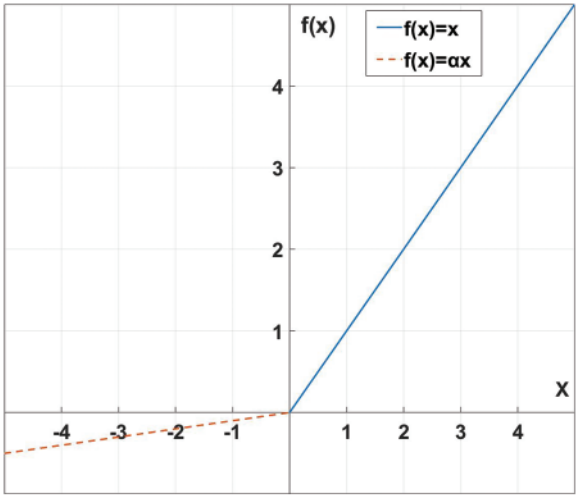


Figure 2: Image and formula of Leaky ReLU activation function

3 Algorithm Improvement

3.1 Depthwise Separable Convolution

The basic assumption of Depthwise Separable Convolution [22] is that the spatial and channel (depth) dimensions of feature maps in convolutional neural networks can be decoupled. Standard convolution computations use weight matrices to achieve joint mapping of spatial and channel-dimensional features, but at the cost of high computational complexity, high memory overhead, and a large number of weight coefficients. Conceptually, the Depthwise Separable Convolution reduces the number of weight coefficients while basically retaining the representation learning ability of the convolution kernel by mapping the spatial dimension and the channel dimension respectively and combining the results.

Fig. 3 shows the process of the Depthwise Separable Convolution. The convolution process is divided into depthwise convolution and pointwise convolution. The former uses a convolution kernel for each channel of the input feature map, and splices the output results of each channel. Then a 1×1 point-by-point convolution to get the final output result. Compared with the standard convolution method, the depthwise separable convolution can reduce the operation cost and improve the calculation speed. At the same time, the spatial feature relationship of the image and the feature relationship between channels can be independently calculated, thereby improving the performance of the semantic segmentation network.

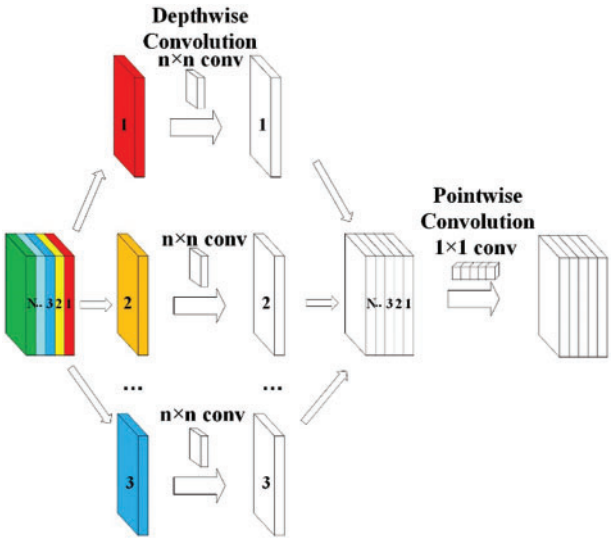


Figure 3: Depthwise separable convolution

3.2 Attention Mechanism Module

The relationship between the feature map channels is particularly important in image semantic segmentation, especially in autonomous driving. Therefore, this paper introduces the SE [23] lightweight attention mechanism module. When the up-sampling feature map and the down-sampling feature map are used for feature fusion, due to the introduction of SE attention mechanism module, the fusion results can focus on the spatial relationship between each feature channel, meanwhile, the network can start from the global information, improving the feature channel weight parameters that are beneficial to network performance, suppressing the feature channel weight parameters that are not conducive to network performance. It can achieve the dynamic calibration of channel information, and the performance of semantic segmentation network can be improved. Fig. 4 shows the structure of the SE Attention Mechanism module.

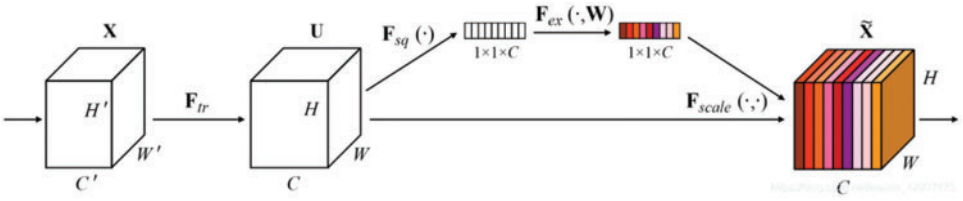


Figure 4: SE attention mechanism module

The main operations of the SE module are: Squeeze and Excitation. The SE module compresses the input feature map to obtain channel-level global features first, then performs excitation operations on the global features. While learning the relationship between each feature channel, it also obtains the weights of different feature channels, and finally multiplied with the input features. Finally, multiply the feature map with the input feature map to get the final map. The Squeeze operation can be expressed as follows:

$$z = F_{sq}(f) = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} f(i,j) \tag{1}$$

In the formula, F_{sq} is the Squeeze operation function; $f \in \mathbf{R}^{H \times W}$ is the two-dimensional feature map set, $f(i,j)$ is one of its elements, H and W is the size of the feature map respectively, and z is the output of the compression operation.

The Excitation operation can be expressed as follows:

$$s = F_{ex}(z, W) = \sigma[W_2 \delta(W_1 z)] \tag{2}$$

In the formula, F_{ex} is the excitation operation function; σ and δ represent the Sigmoid and ReLu activation functions, respectively; $\mathbf{W_1} \in \mathbf{R}^{\frac{C}{r} \times C}$,$\mathbf{W_2} \in \mathbf{R}^{C \times \frac{C}{r}}$,$\mathbf{W_1}$ and $\mathbf{W_2}$ are one of the elements, respectively, C is the number of dimensions of the feature map,r is the dimensionality reduction coefficient; s is the output of the excitation operation.

After the above operations, the output weight of the excitation operation is multiplied by the original input feature and the output of the SE module is:

$$x = F_{scale}(f, s) = s \cdot f(i,j) \tag{3}$$

In the formula: F_{scale} is the scale operation; x is one of the value in $\mathbf{X}$ which is the final output of the SE module, $\mathbf{X} = [x_1, x_2, \ldots, x_c]$.

3.3 Gradient Compression

Optimization techniques are of great significance for improving the performance of neural networks. Currently, the optimization methods used in the field of semantic segmentation algorithms mainly include BN (batch normalization), which works in the activation function and WS (weight standardization), which operates on weights [24]. In addition to operating in these two aspects, this paper considers directly improving the gradient part to make the training process more effective and stable, thereby improving the generalization ability and segmentation accuracy of the semantic segmentation network.

Among the optimization algorithms that operate on gradients in the field of semantic segmentation algorithms, the most common methods are to calculate the momentum of the gradient. The main optimization algorithms are Stochastic Gradient Descent with Momentum (SGDM) [25] and Adaptive Moment Estimation (Adam) [26]. After reference to the literature, the Adam optimization algorithm can dynamically adjust the update step size by using the first-order moment estimation of the gradient (that is, the mean value of the gradient) and the second-order moment estimation (and the variance of the gradient), making it more efficient than the SGDM algorithm. In order to further improve the performance of the algorithm and facilitate the operator to use the method in this paper, a method is proposed to automatically update the gradient according to the training epoch on the Adam optimizer, which is called Gradient Compression, and the improved optimizer is referred to as Adam-GC for short.

The formula for Gradient Compression is as follows:

$$\Phi_{GC}(\nabla w_i L) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{\mu^2}{2}} \left(\nabla w_i L - \frac{1}{N} \sum_{j=1}^{N} \nabla w_{i,j} L \right) \tag{4}$$

In the above formula, w_i represents the weight vector, $\nabla w_i L$ represents the gradient of the loss function to the weight vector, μ is the ratio of the current training times t to the total training times epoch. In the formula, $\varphi(u) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{\mu^2}{2}}$ is the gradient smoothing curve, which is used to smooth the update process of the weight parameters and the image of the curve is shown in Fig. 5 when $\sigma = 0.4$.

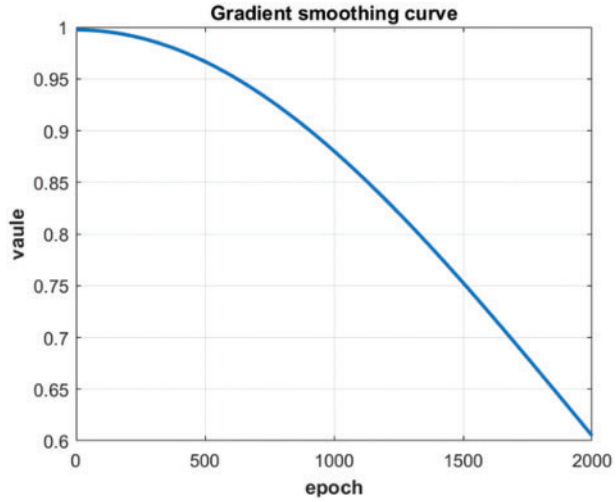


Figure 5: Gradient smoothing curve

As long as the network obtain the mean of the gradient matrix, subtract the mean value from the column vector of each gradient, and then multiply it by the gradient smoothing coefficient, it can get the update direction of the optimal weight. The calculation of this method is relatively simple, and it does not require too much computational cost when applied to the Adam optimization algorithm. Experiments show that it only takes about 0.5 s more per epoch when using the LeNet convolutional neural network model to train the Mnist handwritten digit recognition dataset.

The above formula can be written in matrix form as follows:

$$\Phi_{GC}(\nabla \mathbf{w} L) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{\mu^2}{2}} (\mathbf{P} \nabla \mathbf{w} L) \tag{5}$$

$$\mathbf{P} = \mathbf{I} - \mathbf{i}\mathbf{i}^{\mathrm{T}} \tag{6}$$

In the above formula, $\mathbf{P}$ represents the projection matrix of the hyperplane whose weight space normal vector is $\mathbf{e}$, $\mathbf{i} = 1/\sqrt{\mathbf{N}}$ is the N dimensional unit vector, $\mathbf{I}$ is the N dimensional unit matrix, $\mathbf{P}\nabla \mathbf{w} L$ is the projected gradient on the plane [27], the projected gradient on the hyperplane will compress the weight space, And the range of the gradient smoothing curve is between 0.6–1, which will further reduce the projected gradient and compress the weight space. The gradient compression method in this paper can be simply implemented in the Adam optimization algorithm. Table 1 shows the process of the algorithm.

In Table 1, ε is a small constant, $\frac{\hat{\mathbf{m}}^t}{(\sqrt{\hat{\mathbf{v}}^t} + \varepsilon)}$ represents the weight update direction based on the projected gradient. The gradient compression method in this paper can also be explained from the perspective of projected gradient. Fig. 6 shows the geometric explanation of the Adam optimization algorithm using gradient compression. The gradient is first projected into the hyperplane determined by $\mathbf{i}^T(\mathbf{w} - \mathbf{w}^0) = \mathbf{0}$, then the weights are updated along the update direction determined by the gradient.

Table 1: The process of the GC algorithm

Input: Weight vector $\mathbf{w}^0$; Learning rates α; Exponential decay rates β_1,β_2; Initialize 2nd moment vector $\mathbf{v}_0$	
Traning step:	
for t in epoch:	4. $\mathbf{v}^t = \beta_2 \mathbf{v}^{t-1} + (1-\beta_2)\,\hat{\mathbf{g}}^{t^2}$
1. $\mathbf{g}^t = \nabla \mathbf{w}^t L$	5. $\hat{\mathbf{m}}^t = \mathbf{m}^t / \left(1-\beta_1^t\right)$
2. $\hat{\mathbf{g}}^t = \Phi_{GC}\left(\mathbf{g}^t\right)$	6. $\hat{\mathbf{v}}^t = \mathbf{v}^t / \left(1-\beta_2^t\right)$
3. $\mathbf{m}^t = \beta_1 \mathbf{m}^{t-1} + (1-\beta_1)\,\hat{\mathbf{g}}^t$.	7. $\mathbf{w}^{t+1} = \mathbf{w}^t - \dfrac{\alpha \hat{\mathbf{m}}^t}{\left(\sqrt{\hat{\mathbf{v}}^t} + \varepsilon\right)}$
	8. **end for**

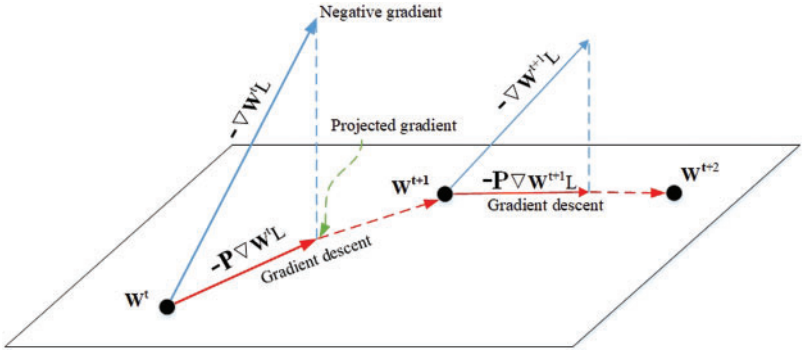


Figure 6: Geometric interpretation of Adam-GC

4 Algorithmic Network Training Settings

4.1 Semantic Segmentation Dataset

To better verify the effectiveness of our algorithm, this paper uses the Cityscapes dataset, the urban landscape dataset, which contains various stereoscopic video sequences recorded from street scenes of 50 different cities, except for a larger 20,000 weakly annotated frames. In addition, there are high-quality 5000-frame pixel-level annotations. The Cityscapes dataset has two sets of evaluation criteria: fine and coarse. The former provides 5,000 finely annotated images, and the latter provides 5,000 finely annotated images plus 20,000 coarsely annotated images.

The Cityscapes dataset is designed to: (1) Evaluate the performance of vision algorithms on the main tasks of semantic urban scene understanding: pixel-level, instance-level and panoramic semantic labels; (2) Support research aimed at leveraging large amounts of (weakly) annotated data, for example for training deep neural networks. Fig. 7 shows its data file.

(a) Training image (b) Grayscale of the label (c) Lable (Ground Truth)

Figure 7: Cityscapes dataset

4.2 Loss Function

MIoU (Mean Intersection over Union) is the average intersection and union ratio, which is the current standard measure of semantic segmentation. It calculates the interaction ratio of the two sets. In the semantic segmentation problem, the two sets are the ground truth and the predicted value. The formula is as follows:.

$$MIoU = \frac{1}{k+1}\sum_{i=0}^{k}\frac{p_{ii}}{\sum_{i=0}^{k}p_{ij}+\sum_{i=0}^{k}p_{ji}-p_{ii}} \tag{7}$$

In the formula, k is the number of semantic segmentation categories, i is the real value, j is the predicted value, and p_{ij} represents that i is predicted to be j.

5 Test Verification and Result Analysis

The algorithms in this paper are built under the framework of Pytorch 1.2. The training and detection are based on the hardware configuration of the CPU is Intel(R) Core(TM) i7-9700 CPU@3.00 GHz, the GPU is NVIDIA GeForce RTX 2070 SUPER, 8 G video memory, and the number of CUDA cores is 2560, the running memory is 16 G, and the operating system is Windows10 computer platform.

5.1 Loss Function

According to the above-mentioned improvements to the algorithm, they are combined to verify their effectiveness, and the changes in Loss during the training process are recorded. At the same time,

the early-stop method in pytorch is used to prevent overfitting of the training model, which leads to poor model generalization ability. The maximum number of training iterations is set to 1000, and the model weights are saved every 50 generations. The Loss of the training process is shown in Fig. 8:

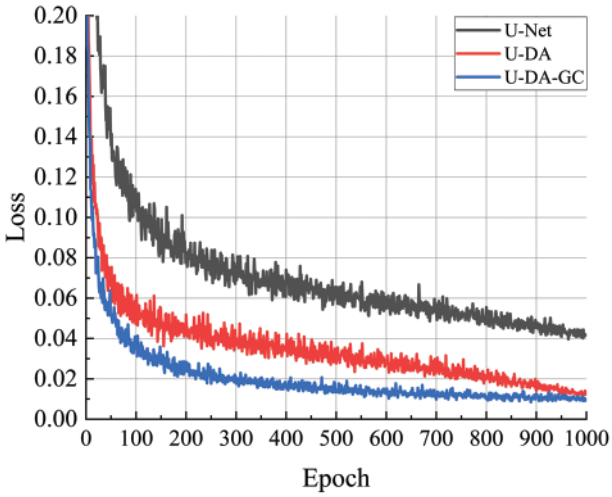


Figure 8: Traing loss curves

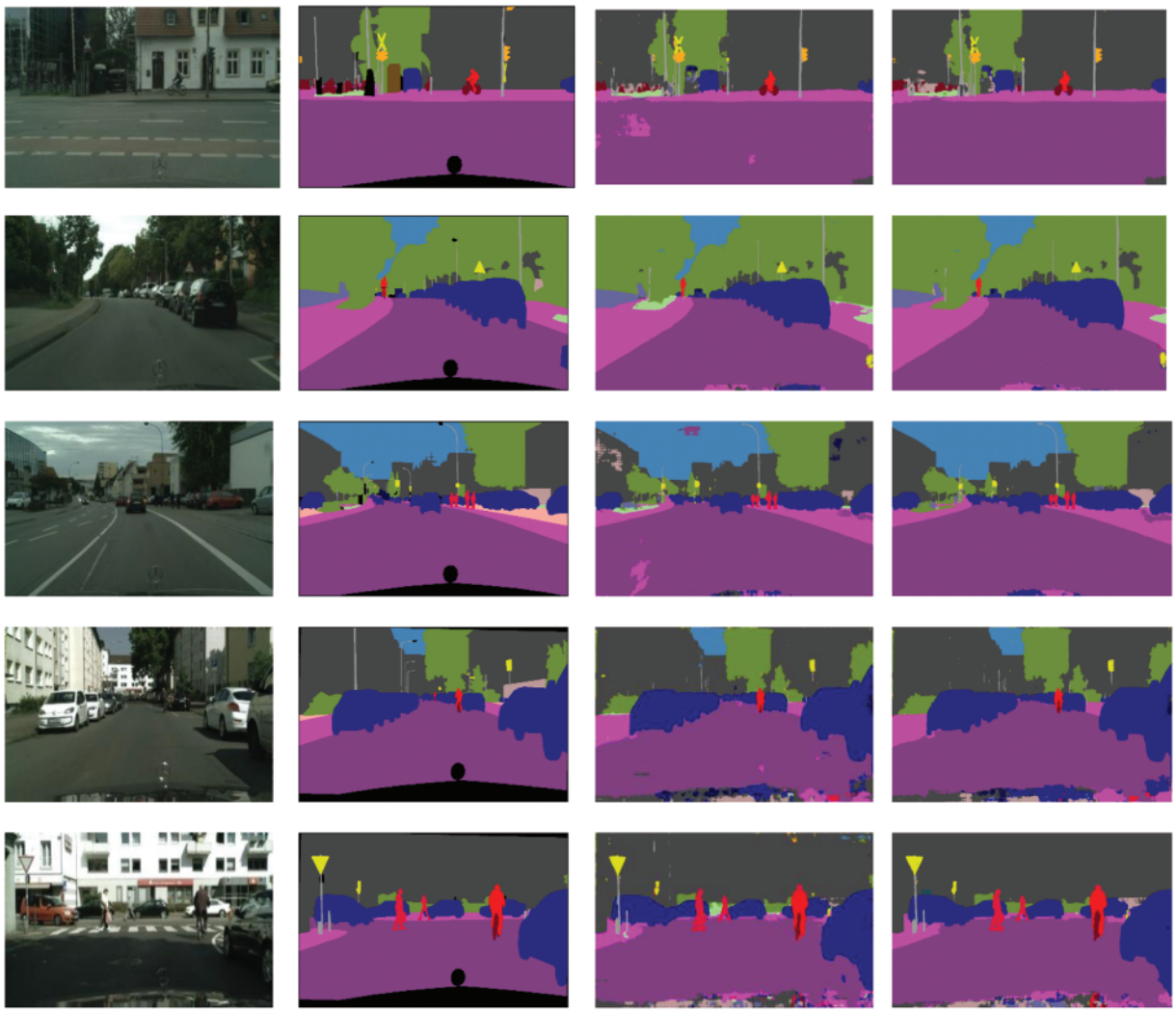
In Fig. 8, U-DA represents the basic U-Net algorithm network added with Depthwise Separable Convolution and SE Attention Mechanism.U-DA-GC represents the U-DA algorithm network added with GC. Fig. 8 shows that: Compared with the basic U-Net algorithm, both U-DA and U-DA-GC have a relatively stable and lower-Loss training process, especially U-DA-GC, which also a faster loss-convergence speed.

5.2 Validation Test

In order to verify the effectiveness of the improved algorithms and training method in this paper, each experiment is performed on the Cityscapes training set, then each accuracy index test is performed on the validation set. The parameter settings are consistent with the overall accuracy test experiment. The visual comparison of some segmentation results is shown in Fig. 9.

By comparing the segmentation effects of the basic U-Net algorithm model and the U-DA-GC algorithm model, it is clear that the latter has an overall excellent performance in classification accuracy and positioning accuracy, which is significantly improved compared with the former. The segmentation effect on categories such as pedestrians, trees, vehicles, and roads are excellent, and the category to which it belongs can be basically identified. At the same time, the segmentation edge is also relatively smooth and accurate. However, due to the relatively low maximum number of iterations set, both of them cannot segment small objects or object edges well in the face of long-distance and complex scene segmentation, which is also an inevitable problem in segmentation area.

Meanwhile, due to the multiple network improvements and methods for the basic U-Net semantic segmentation network, it is necessary to verify the effectiveness of each part, so that its effect on the overall network performance of the model can be quantitatively observed. The resulting data are shown in Table 2.



(a) Input image (b) Ground Truth (c) U-Net (d) U-DA-GC

Figure 9: Validation of efficiency

Table 2: Validation of efficiency

U-Net	DA	GC	MIoU/%
√			73.67
√	√		76.68
√	√	√	**78.02**

Note: The bold part is the best value of this experiment.

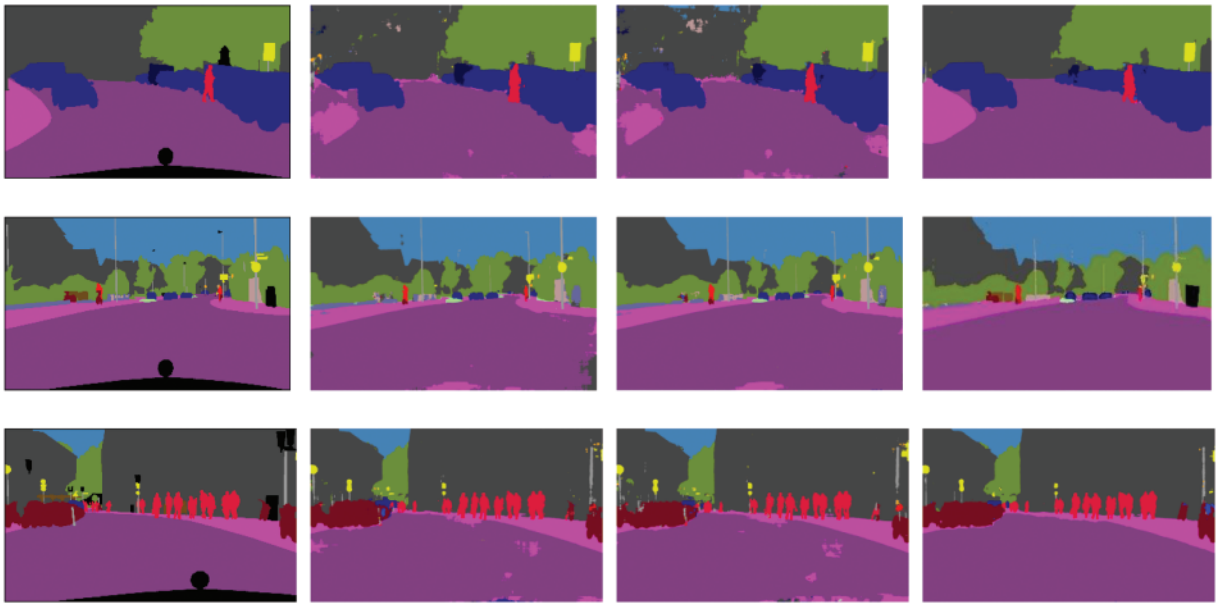
Table 2 shows that the MIoU of the U-DA algorithm network after adding DA can reach 76.68% compared with the original network, which is 4.1% higher than the basic network performance. Due to the SE attention mechanism module and the depthwise separable convolution are introduced into the

basic U-Net algorithm network, so that the improved network can focus on the spatial relationship between each feature channel, and then it can start from the global information to achieve better performance.

Meanwhile, the MIoU of the U-DA-GC algorithm network can reach 78.02%, compared with the basic network and the U-Net algorithm network after adding DA, which is 5.9% higher than the basic U-Net, and 1.7% than U-DA, it can be concluded that since the GC optimization algorithm can make the training process more stable and effective, the ability of the trained model to learn image features has also been enhanced, and the improved U-Net algorithm can achieve better results.

5.3 Comparative Test

In order to further verify the effectiveness of the improved algorithms and training method in this paper, this paper selects U-DA-GC and other latest semantic segmentation algorithm network: Deeplabv3 and SegNet to conduct comparative experiments. All experiments are carried out in the same experimental environment. And each experiment is performed on the Cityscapes training set, then each accuracy index test is performed on the validation set. The parameter settings are consistent with the overall accuracy test experiment. The visual comparison of some segmentation results and results data are in Fig. 10 and Table 3.



(a) Ground Truth (b) Deeplabv3 (c) SegNet (d) U-DA-GC

Figure 10: Comparison of efficiency

Table 3: Comparison of efficiency

Algorithm network	MIoU/%
Deeplabv3	74.53
SegNet	75.25
U-DA-GC	**78.02**

As is shown in Fig. 10, the overall performance of U-DA-GC is excellent in classification accuracy and positioning accuracy, which is significantly better than Deeplabv3 and SegNet, especially in vehicles and roads, which are important categories of autonomous driving scenario.

Table 3 shows that compared with Deeplabv3 and SegNet, the U-DA-GC algorithm network can achieve 78.02% MIoU on the validation set, which is 4.6% and 3.8% higher than Deeplabv3 and SegNet, respectively.

6 Conclusion

In this paper, depthwise separable convolution and attention mechanism are introduced on the basis of the basic network U-Net, and a new training adjustment strategy of gradient compression is proposed at the same time. Through a series of experimental verifications, the following conclusions are obtained:

(1) The improvement methods in this paper can meet the demand for a lightweight semantic segmentation network in the autonomous driving perception system, reduce the operation cost and improve the operation speed. It also provides support for the road condition analysis and real-time segmentation of the autonomous driving perception system.

(2) The training optimization algorithm proposed in this paper can not only improve the generalization ability and segmentation accuracy of the training model but also has strong algorithm adaptability that can be easily added to other optimization algorithms.

(3) Compared with the basic algorithm and the other latest semantic segmentation algorithms, the improved method in this paper has a considerable improvement in the segmentation accuracy of common road objects, especially the segmentation effect on the driving area, which is an important segmentation target in the autonomous driving system.

(4) The data set used in this paper has less training data, and all of them are in the daytime traffic flow with a good line of sight. The segmentation effect for other weather or nighttime needs to be further researched.

(5) In the process of segmentation, the problem of low segmentation accuracy is easy to occur when facing more complex driving scenes. To solve this problem, it is necessary to conduct more deep research on the feature extraction network.

Funding Statement: This work is supported by Qingdao People's Livelihood Science and Technology Plan (Grant 19-6-1-88-nsh).

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

1. Huang, M. (2017). A brief introduction on autonomous driving technology. *Science & Technology Information, 15(27),* 1–2 (in Chinese).
2. Mu, L., Zhao, H., Li, Y., Liu, X. T., Qiu, J. Z. et al. (2021). Traffic flow statistics method based on deep learning and multi-feature fusion. *Computer Modeling in Engineering & Sciences, 129(2),* 465–483 DOI 10.32604/cmes.2021.017276.

3. Wang, S. F., Dai, X., Xu, N., Zhang, P. F. (2017). Overview on environmental perception technology for unmanned ground vehicles. *Journal of Changchun University of Science and Technology (Natural Science Edition), 40(1),* 1–6.
4. Chen, Q., Xie, Y., Guo, S., Bai, J., Shu, Q. (2021). Sensing system of environmental perception technologies for driverless vehicle: A review of state of the art and challenges. *Sensors and Actuators a Physical, 319,* 112566. DOI 10.1016/j.sna.2021.112566.
5. Devi, K. G., Rath, M., Linh, N. T. D. (2020). *Artificial intelligence trends for data analytics using machine learning and deep learning approaches*. USA: CRC Press.
6. Otsu, N. (1979). A threshold selection method from gray-level histograms. *IEEE Transactions on Systems, Man, and Cybernetics, 9(1),* 62–66. DOI 10.1109/TSMC.1979.4310076.
7. Pun, T. (1980). A new method for grey-level picture thresholding using the entropy of the histogram. *Signal Processing, 2(3),* 223–237. DOI 10.1016/0165-1684(80)90020-1.
8. Yen, J. C., Chang, F. J., Chang, S. (1995). A new criterion for automatic multilevel thresholding. *IEEE Transactions on Image Processing, 4(3),* 370–378. DOI 10.1109/83.366472.
9. Khan, J. F., Bhuiyan, S. M. A., Adhami, R. R. (2011). Image segmentation and shape analysis for road-sign detection. *IEEE Transactions on Intelligent Transportation Systems, 12(1),* 83–96. DOI 10.1109/TITS.2010.2073466.
10. Huang, P., Zheng, Q., Liang, C. (2020). Overview of image segmentation methods. *Journal of Wuhan University (Science Edition), 66(6),* 519–531.
11. Boykov, Y. Y., Jolly, M. P. (2001). Interactive graph cuts for optimal boundary region segmentation of objects in N-D images. *Proceedings Eighth IEEE International Conference on Computer Vision*, pp. 105–112. Vancouver, Canada.
12. Rother, C., Kolmogorov, V., Blake, A. (2004). "GrabCut": Interactive foreground extraction using iterated graph cuts. *International Conference on Computer Graphics and Interactive Techniques, 23(3),* 309–314.
13. Tang, M., Gorelick, L., Veksiler, O., Boykov, Y. (2013). GrabCut in one cut. *Proceedings of the 2013 IEEE International Conference on Computer Vision*, pp. 1769–1776.
14. Liu, S., Qi, X., Shi, J., Zhang, H., Jia, J. (2016). Multi-scale patch aggregation (MPA) for simultaneous detection and segmentation. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3141–3149. Las Vegas, NV, USA.
15. Pinheiro, P., Collobert, R., Dollar, P. (2015). *Learning to segments objects candidates advances in neural information processing systems*. Montreal: NIPS.
16. Chen, L. C., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A. L. (2014). Semantic image segmentation with deep convolutional nets and fully connected CRFs. *Computer Science, 2014(4),* 357–361.
17. Zhao, H., Qi, X., Shen, X., Shi, J., Jia, J. (2017). ICNet for real-time semantic segmentation on high-resolution images. *Lecture Notes in Computer Science,* 418–434. DOI 10.48550/arXiv.1704.08545.
18. Ronneberger, O., Fischer, P., Brox, T. (2015). U-Net: Convolutional networks for biomedical image segmentation. *Lecture Notes in Computer Science, 9351,* 234–241. DOI 10.48550/arXiv.1505.04597.
19. Huang, H., Lin, L., Tong, R. (2020). UNet 3+: A full-scale connected UNet for medical image segmentation. *2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1055–1059. Barcelona International Convention Centre, Spain.
20. Zhong, S., Guo, X., Zheng, Y. (2020). Improved U-NET network for pulmonary nodule segmentation. *Computer Engineering and Applications, 56(17),* 203–209.
21. Hou, T. X., Zhao, J. J., Qiang, Y., Wang, S. H., Wang, P. (2020). CRF 3D-UNet pulmonary nodule segmentation network. *Computer Engineering and Design, 41(6),* 1663–1669.
22. Chollet, F. (2017). Xception: Deep learning with depthwise separable convolutions. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1800–1807. Honolulu, USA.

23. Hu, J., Shen, L., Sun, G. (2018). Squeeze-and-excitation networks. *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 7132–7141. Salt Lake City, USA.
24. Mu, L., Zhao, H., Li, Y., Qiu, J. Z., Sun, C. L. et al. (2022). Vehicle recognition based on gradient compression and YOLO v4 algorithm. *Chinese Journal of Engineering, 44(5),* 940–950.
25. Cheng, K. Y., Tao, F., Zhan, Y. Z., Li, M., Li, K. (2020). Hierarchical attributes learning for pedestrian re-identification via parallel stochastic gradient descent combined with momentum correction and adaptive learning rate. *Neural Computing and Applications, 32(10),* 5695–5712. DOI 10.1007/s00521-019-04485-2.
26. Daniel, O. M., Luige, V. (2020). Facial expressions recognition for human–robot interaction using deep convolutional neural networks with rectified adam optimizer. *Sensors, 20(8),* 2393. DOI 10.3390/s20082393.
27. Gupta, H., Jin, K. H., Nguyen, H. Q., McCann, M. T., Unser, M. (2018). CNN-Based projected gradient descent for consistent CT image reconstruction. *IEEE Transactions on Medical Imaging, 37(6),* 1440–1453. DOI 10.1109/TMI.42.