Winter Road Surface Status Recognition Using Deep Semantic Segmentation Network

Caojia Liang[1], Junfeng Ge[1], Wei Zhang[2], Kang Gui[1], Faouzi Alaya Cheikh[3], Lin Ye[1]

[1] *School of Artificial Intelligence and Automation, Huazhong University of Science and Technology (HUST), Wuhan, Hubei 430074, China*

[2] *Wuhan Second Ship Design and Research Institute, Wuhan, Hubei 430205, China*

[3] *Department of Computer Science, Norwegian University of Science and Technology (NTNU), Teknologivegen 22, 2815 Gjøvik, Norway*

LCJ_hust@126.com, gejf@hust.edu.cn, 5709195@qq.com, gk_work@hust.edu.cn, faouzi.cheikh@ntnu.no, lye @hust.edu.cn

Abstract— **Adverse road surface status in winter has a strong impact on traffic safety, mobility, efficiency and maintenance. Slippery road statuses caused by water, ice and snow have led to fatal accidents around the world every year. The timely road surface status information can reduce the potential injuries by early warning and maintenance cost through the right treatments. This paper investigates the application of the deep neural networks especially the semantic segmentation network for detailed road surface status recognition based on images. Unlike the rough road surface status recognition with one status per image through image classification based on convolutional neural network, the proposed approach gives out the detailed road surface status of different road regions using the semantic segmentation network called D-UNet. The D-UNet has the architecture consisting of a contracting path to capture context and a symmetric expanding path that enables precise localization and utilizes the dilated convolutions to increase the receptive field of a network exponentially which capture both the local and contextual information. Results show that the proposed approach has the highest classification performance in comparison to the traditional machine learning techniques under the small size dataset. The testing accuracy with different training dataset sizes is also analyzed, showing the potential of achieving much higher accuracy with a larger training dataset.**

Keywords— ***road surface status recognition, deep learning, semantic segmentation, dilated convolutions***

I. INTRODUCTION

Severe weather conditions can cause the road surface to be in a complex status such as moisture, snow, ice or snow mixed with ice, which decrease the friction between the wheels and the road and then lead to traffic accidents. Statistics from the National Highway Traffic Safety Administration(NHTSA) indicated that an average of 6 million traffic accidents occurred on US highways each year, among which about 1.7 million traffic accidents are caused by adverse road conditions or severe weather, resulting in about 800 thousand injuries and 7000 deaths. Data released by the Swedish National Institute of Road and Transportation (VTI) showed that there were obvious differences in the accident rates under different road conditions. The occurrence of accidents in frost and ice conditions is 3-16 times higher than that in dry road conditions, up to 0.53 million/km as shown in [1]. About a quarter of fatal traffic accidents in Finland are caused by snow and ice roads directly or indirectly. More than 3,800 traffic accidents are caused by wet or snow roads every year in Europe. Therefore, it is urgent to develop a system capable of monitoring and recognizing the road status timely, then issuing early warning signals to avoid potential accidents.

II. PREVIOUS RESERCH

In recent years, the development of image processing technology and the popularity of road monitoring equipment have led to many non-contact road surface status detection and recognition methods which based on image technology. Surveillance cameras are installed in sections where accidents occur frequently. Cameras acquire real-time road images and transmit them to the control centre for road surface status recognition.

A. Traditional machine learning

The usual approach is to intercept the road image from the road surveillance video or vehicle cameras. Their underlying features such as colour, texture and brightness are extracted after a series of pre-processing. Then, machine learning techniques such as support vector machine, k nearest neighbours, Bayes classifier and neural network are used to classify.

Omer et al. [2] investigate the feasibility of classifying winter road surface conditions using images from low cost mounted on regular vehicles. All 400 images are marked as bare road, snowy road and tracks. RGB features along with gradients have been uses as feature vectors. A Support Vector Machine (SVM) is trained using the extracted features and then used to classify the images into their respective categories and have a classification accuracy of over 80%.

Kawai et al. [3] propose a distinction method for road surface at night time. This paper discussed the differences of image features of dry, wet and snowy roads under different light sources to combine the three features of colour, brightness and texture. K nearest neighbour algorithm is used for classification and the accuracy can reach 96.1%, 89.4% and 95.6% respectively.

Jonsson et al. [4] use an infrared camera equipped with a set of optical wavelength filters to obtain the brightness of each pixel as features. The images have primarily been used to develop multivariate data models and also for the classification of road conditions in each pixel. The resulting imaging system can reliably distinguish between dry, wet, icy, or snow covered sections on road surface. This system is a vast improvement on existing single spot road status classification systems.

B. *Deep learning*

In recent years, deep learning technology has made great achievements in fields both of academia and industry, including image classification, recognition and detection. Thereafter, some researchers apply deep learning to the field of recognition of pavement state.

Researchers at Brunswick Polytechnic University in Germany intercept six kinds of road surface images from KITTI in [5] and Robotcar dataset, which are dry, wet, snow and so on. They train and then compare two different models --InceptionV3 and ResNet50 respectively. Then a long-term and short-term memory unit, is added to further improve the accuracy in [6].

Similarly, Pan et al. [7] use pre-trained deep classification models – Inception and ResNet to classify snow cover on the road surface. This method makes full use of the weight of the pre-trained model, and only requires a small scale of pavement state images for fine-tuning and can achieve 90.7% classification accuracy.

In summary, most of the current methods design and select the features manually, such as texture, colour, brightness or other features. Then, combine them to form a feature database. Finally, machine learning algorithms are used to establish classification model. With the development of deep learning, some researchers use pre-trained deep neural network to classify the road surface images directly and have achieved results comparable to traditional machine learning algorithm. However, the status of pavement is complex and changeable. There are many mixed states such as ice, snow and tracks in a region and simple classification cannot meet the requirements of practical application. Therefore, the recognition of road surface condition should be refined to intensive prediction of each pixel in the image and give the information of category, that is, image segmentation.

III. Deep Learning for Semantic Segmentation

Semantic segmentation aims to analyse an image at pixel level, which assigns each pixel in the image to an object class.

Before deep learning took over computer vision, people used approaches like TextonForest and Random Forest based classifiers for semantic segmentation. Early algorithms are usually based on prior knowledge of thresholds, edges or regions, which essentially relied on low-level visual information of the images themselves. Since these methods have no training stage, the computational complexity is usually low. Of course, the accuracy is difficult to meet the demand without auxiliary information.

The rise of deep learning has made the trend of using deep neural networks to solve semantic segmentation. The fundamental reason is that deep networks have strong non-linear simulation ability, while the traditional algorithms rely on the prior to extract features can be replaced by the network and even get more abundant features. In 2014, Fully Convolutional Networks (FCN) by Long et al. from Berkeley, popularized CNN architectures for dense predictions without any fully connected layers in [8]. This allowed segmentation maps to be generated for image of any size. Almost all the subsequent state of the art approaches on semantic segmentation adopted this paradigm. Typical recognition nets and their deeper successors ostensibly take fixed-sized inputs and produce non-spatial outputs. The fully connected layers of these nets have fixed dimensions and throw away spatial coordinates. Therefore, the fully connected layers are replaced by the convolution layers.

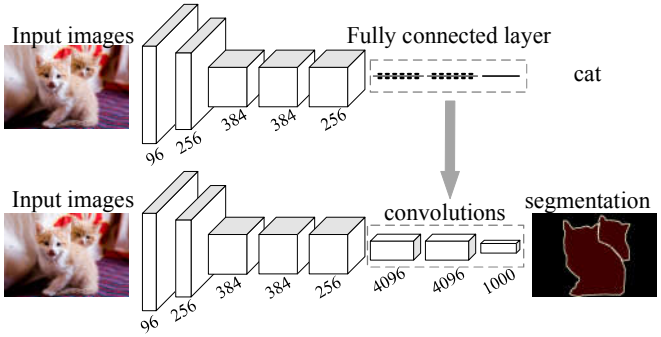


Fig. 1 Fully Convolution Network(FCN).

Apart from fully connected layers, one of the main problems with using CNNs for segmentation is pooling layers. Pooling layers increase the receptive field and are able to aggregate the context while discarding the 'where' information to be preserved. The segmentation task needs to align the class label with the original image, so the location information should be reintroduced. Two different classes of architecture evolved in the literature to tackle this issue.

First one is encoder-decoder architecture. Encoder gradually reduces the spatial dimension with pooling layers and decoder gradually recovers the object details and spatial dimension. There are usually shortcut connections from encoder to decoder to help decoder recover the object details better.

Architecture in the second class use what called as dilated in [9] or atrous convolutions in [10] and do away with pooling layers. Dilated convolutions allows for exponential increase in receptive field without decrease of spatial dimensions.

IV. Network Architecture

We propose a new architecture with cascaded dilated convolutions called D-UNet. Encoder-decoder, a commonly used structure of semantic segmentation models is adopted, and several cascaded dilated convolutions are added between the encoder and decoder to better preserve the global abstract features and local details at the same time.

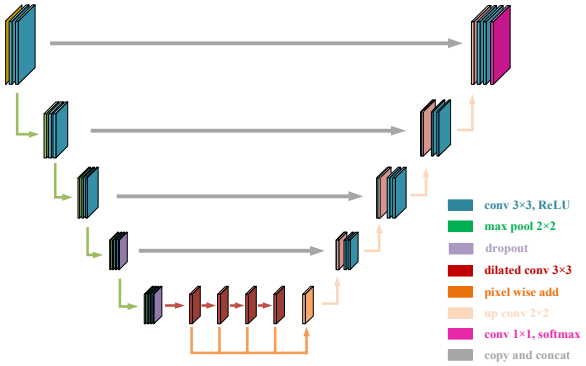


Fig.2 Architecture of D-UNet

The architecture consists of a contracting path to capture context and symmetric expanding path that enables precise localization. The features of different levels are fused by shortcut connections and dilated convolutions.

A. *Encoder*

The structure of encoder is similar to classical classification network VGG. Two 3×3 small convolutions are stacked repeatedly. After each convolution, the ReLU activation function is used to increase the non-linear simulation ability. Finally, the maximum pooling operation of 2×2 is performed as a complete down-sampling operation. In each down-sampling, the number of feature channels is doubled in order to extract more diverse features. Several consecutive 3×3 convolutions can deepen the network and enhance the expression ability. Meanwhile, it can reduce the model parameters and accelerate the training. In encoder, there are 4 down-sampling operations. The resolution of features is reduced to 1/16 of the original image to extract the local details fully. In order to increase the generalization ability of the network and avoid over-fitting, 50% of the parameters are randomly dropout after the third and fourth down-sampling.

B. Decoder

The decoder includes up-sampling operations equal to the number of down-sampling in the encoder. The number of feature map channels is gradually halved and the resolution is restored to the original image size. Transportation convolution is used for up-sampling, and the output of layer incorporates a down-sampled output with the same resolution from the encoder, which is shortcut connection. The shortcut connection fuses the low-level features extracted from the encoder with the high-level features extracted from the decoder, forming a richer and more comprehensive description of features.

C. Dilated convolution

Apart from the shortcut connections to merge the feature map with the same resolution into hierarchical features, there are also dilated convolutions with different dilated rate superimposed to form more expressive comprehensive features.

Receptive field refers to the size of region in which the features in the CNNs are mapped to the input space. And the larger, the more context is included. In segmentation task, the more information contains, the more likely the pixels will be correctly classified. Deepening the network can increase the receptive field, but will also increase the complexity as whole and make training difficult. Or using the pooling layers first to reduce the image size to capture larger receptive field, then sample the feature map to restore the resolution of the original image. However, pooling will discard details. Dilated convolutions balance the contradiction between information loss and increased receptive field.

Take the 3×3 convolution kernel as an example, when a traditional convolution operation is performed, a convolution kernel is multiplied by pixel in a continuous 3×3 region of input tensor and summed point by point, as shown in Fig. 3(a). The red dots are input pixels corresponding to the kernel, and the green grids ate their receptive fields in the input. The dilated convolution is to convolution with a dilated filter, that means, to convolute the kernel by a number of pixels spaced from the 3×3 region of the input tensor. Fig. 3(b) is produced from Fig. 3(a) by a 2-dilated convolution and each element has a receptive field of 7×7. Fig. 3(c) is produced from Fig. 3(a) by a 4-dilated convolution and each element has a receptive field of 15×15. The number of parameters associated with each layer is identical. The receptive field grows exponentially while the number of parameters grows linearly. Systematic dilation supports exponential expansion of the receptive field without loss of resolution or coverage.

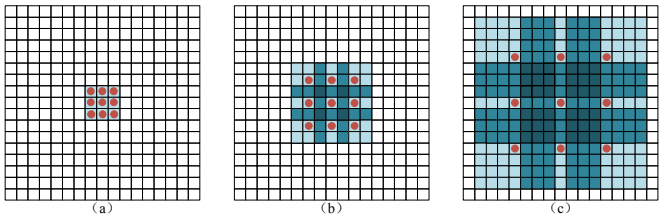


Fig.3 The receptive field when dilated rate are 1,2 and 4 respectively.

In the D-UNet network, after the down-sampling of encoder, the dilated convolution with rate of 1,2,3,5 is performed on the feature maps in turn and the features on multi-scale are merged. The shallower feature maps have smaller receptive fields, focusing on learning local detail features, while the deeper feature maps have larger receptive fields and can learn more abstract information which is helpful to improve the recognition performance.

D. Loss function

Every algorithm in deep learning has a loss function to measure the difference between the prediction and ground truth. The closer the difference is to zero, the smaller the deviation between the prediction and the real result. Learning and updating the parameters are guided by back propagation of errors.

Semantic segmentation can be abstracted as a dense multi-classification problem. So, we choose cross-entropy loss, which is commonly used in multi-classification tasks:

$$Loss_{cross\_entropy} = -\frac{1}{n}\sum_{n}\left[y\log(a)+(1-y)\log(1-a)\right]$$

V. EXPERIMENT

This chapter will introduce the experimental process in detail. The dataset used in the experiment, the training of the model, experimental results and comparative analysis will be introduced one by one.

A. Dataset

The algorithm based on video or image for the recognition of non-contact road surface state takes characteristics of wet and snow road as the object for recognition. Datasets used by researchers in various countries are usually obtained through road monitoring video of traffic departments and vehicle-mounted cameras of non-special vehicles, and then marked by themselves according to research needs. Now, there is no public and unified special dataset for pavement state recognition.

According to the requirements of the paper, we rely on the existing hardware resources of the laboratory and collected a large number of images of the campus road.

The road surface condition recognition algorithm based on deep learning needs a large number of image samples as support, so it is necessary to collect data through multiple ways. Due to the lack of diversity of road surface image samples collected in campus roads, it is difficult to obtain appropriate images of ice and snow in campus. In order to

expand the size of the dataset and ensure the richness and diversity of samples, while using the winter pavement images of 2017 and 2018 obtained by the laboratory camera, multiple sources such as highway video monitoring resources and network resources were added to ensure the size of the data set meets the requirements.

TABLE I. ROAD SURFACE STATUS RECOGNITION DATASET

label	status	samples	pixels	label color
0	background	——	——	
1	Dry	186	4.52e+7	
2	Wet	487	1.56e+8	
3	Snow	461	7.68e+7	
4	Ice	136	3.98e+7	
5	Water	473	7.44e+7	
6	Tracks	337	1.23e+8	

Finally, the self-made dataset of mixed road surface status recognition used in this paper is formed, which includes seven state categories： dry, wet, snow, ice, water, track surfaces run over by wheels, and unrelated background. Besides a very small area of background, each image contains one to three pavement states such as an entire area with a muddy rutted surface, or a wet-water mixed surface, or a snow-ice mixed surface, etc. All image data were randomly divided into training set, verification set and test set according to the hierarchical sampling method, with 804, 100 and 100 samples respectively. Some examples are shown in Fig.4.

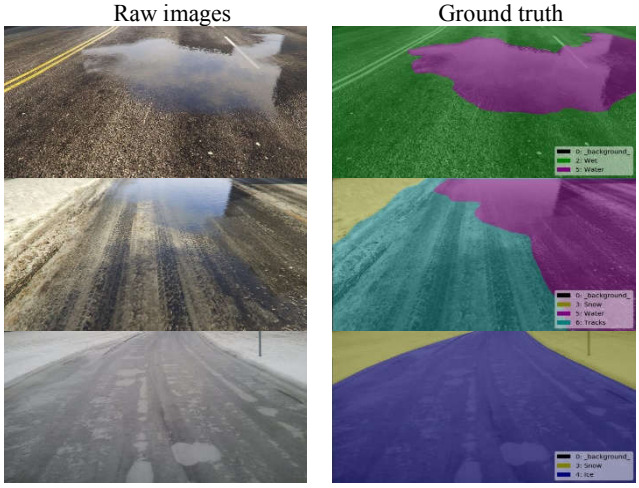


Fig.4 Raw images and ground truth.

B. Training

Deep learning model involves a lot of complex matrix operation and float calculation in training, which needs to be iterated gradually through continuous attempts to find the optimal solution among thousands of variables. Therefore, higher requirements are put forward for the experimental environment.

The hardware configuration of the experimental platform is one GPU of GeForce GTX 1080 Ti and CPU of E5-2650 with 64GB memory. In terms of software configuration, the experimental platform is based on the 64-bit operating system of Ubuntu18.04. The current mainstream deep learning framework TensorFlow and Python are used to build the network model, and the parallel computing architecture CUDA and the GPU acceleration library cuDNN are used to conduct high-performance parallel computing. In addition, Numpy, OpenCV, PIL, matplotlib and other commonly used computing library or visual library are used.

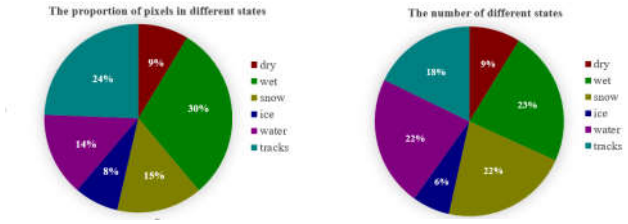
This section conducts experiments on the proposed D-UNet with cascade dilated convolutions, which mainly explores the influence of training parameters on the convergence speed of training and the final accuracy. Through experiments, it is found that batch size, learning rate and the weights of different classes in loss function have significant influence on the accuracy of the algorithm.

1) *Batch size:* This section sets different Batch sizes to verify the impact of the Batch size on the algorithm. Due to the limitation of the experimental platform hardware, only the Batch size of 8 was verified. Table Ⅱ. shows the experimental results of the D-UNet model in the test set after 200,000 iterations. The average pixel accuracy and mean IoU are both improved with the increase of the Batch size.

TABLE Ⅱ. BATCH SIZE EFFECT ON MODEL ACCURACY

Batch size	2	4	6	8
mPA/%	49.25	82.32	83.94	84.67
mIoU/%	23.28	62.98	71.25	74.83

Batch size represents the number of samples involved in the calculation in each iteration, and the larger, the more it can reflect the overall distribution of the dataset. As shown in Fig. 5, pavement state recognition is an unbalanced task, and even can be said to image segmentation task itself so, not only embodies the disequilibrium distribution in each category of the whole dataset, also includes the distribution of all categories of pixels in each sample is not balanced. Therefore, the larger the batch size is, the more accurately and comprehensively the whole data set can be summarized, making the feature expressions learned in each iteration of the model more accurate.



(a) The proportion of pixels in different status

(b) The number of different status

Fig.5 The distribution of different status in road condition dataset.

2) *Decay of learning rate:* Through the experiment in the previous section, the batch size is selected as 8. In this case, this section mainly explores the influence of learning rate decay. The learning rate adjusts the weight of the network according to the gradient of the loss function. The smaller the value is, the slower the change speed of the loss function will be. Although it can ensure that no local minimum is missed, it also means that more time is needed. In the process of network training, an appropriate learning rate will accelerate the convergence of the model, while an unsatisfactory learning rate will directly lead to the loss of the model objective function, oscillation or even explosion. This section compares the effects of constant learning rate, exponential decay

learning rate and piecewise decay learning rate on the training process.

The learning rate with exponential decay is that the learning rate decreases exponentially with the number of iterations, and its calculation formula is:

$$decayed\_lr = init\_lr \times decay\_rate^{\left(global\_steps / decay\_per\_steps\right)}$$

The initial learning rate is set as 0.0001, the decay rate is 0.95, and the number of decay steps is 50000. The curve of learning rate changing with iteration times is shown in the green curve in Fig.6.

The learning rate of piecewise decay is the initial value of the learning rate and the value of the subsequent decay in the defined piecewise interval. The initial learning rate is maintained in the first 100,000 iterations of training, and is set as $5e-5$ and $1e-5$ for each subsequent 50,000 iterations, as the red curve shown in Fig.6.

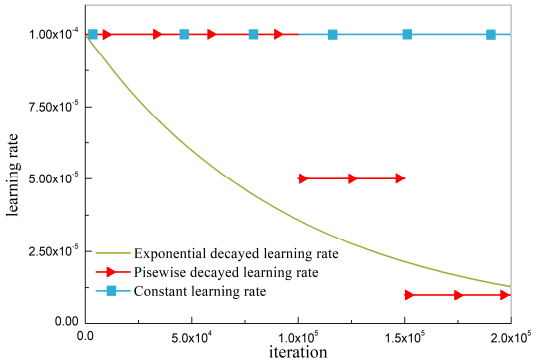


Fig.6 Curve of learning rate with iterations.

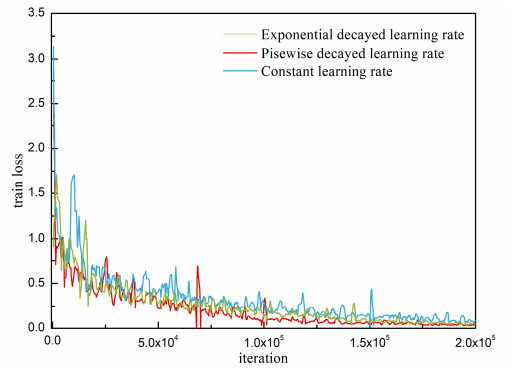


Fig.7 Training loss under different learning rate

Fig.7 is the loss of the objective function under three different learning rates. It shows that different learning rates can eventually make the model converge. Among them, the loss of the model training under the fixed constant learning rate is slightly larger than that under the learning rate decay mechanism. In the training process, the learning rate with dynamic changes according to the number of training rounds can make the model better converge to the minimum value, and the loss curve of the objective function should be in the form of a slide in the ideal learning rate. Therefore, the learning rate of exponential attenuation or piecewise attenuation can meet the demand. In the subsequent experiments, piecewise attenuation strategy will be adopted.

3) *Weight adjustment of loss function:* Studies show that the friction coefficient between wheels and road surface is different under different road surface conditions, so the probability of causing traffic accidents is also different. In practical application, more attention should be paid to icy, water and muddy rutted road surface, while the possibility of accidents caused by dry and wet road surface is relatively low. Therefore, different degrees of attention should be paid to each pavement state, and the weight of the state with high risk coefficient and difficult to be recognized in the loss function should be increased. Then, the weighted cross entropy function is:

$$Loss_{weighted\_loss} = -\frac{1}{n}\sum_{i=0}^{n} w_i \left[y\log(a) + (1-y)\log(1-a) \right]$$

The weight of loss function corresponding to different pavement states is adjusted appropriately according to the friction coefficient between wheels, traffic accident rate and accuracy, as shown in Table Ⅲ.

TABLE Ⅲ. WEIGHT OF DIFFERENT PAVEMENT STATUS IN LOSS FUNCTION

status	weight	Acc/%	IoU/%
background	0.05	--	--
dry	0.1	84.87	82.73
wet	0.1	92.08	84.95
snow	0.2	88.25	79.75
ice	0.15	89.54	87.14
water	0.25	76.07	62.85
tracks	0.15	92.11	78.39
sum/average	1	87.15	79.30

After adjusting the weight of loss function, the mean pixel accuracy and mean IoU are improved on the whole, especially for the three states: snow cover, ice and snow, which are difficult to identify. The effect of recognition is also improved significantly, which verifies the effectiveness of rational weight allocation.

C. Results

Finally, the batch size is determined to be 8, the piecewise decayed learning rate and the weight combination are determined to be [0.05, 0.1, 0.1, 0.2, 0.15, 0.25, 0.15], which are adopted in subsequent experiments. This section mainly discusses the effect of cascade dilated convolutions module on improving the accuracy of pavement state recognition.

The cascade dilated convolution module contains four layers of dilated convolution, and different rates are set to obtain a wider range of context information. In the experiment, the dilated rates are set as [1,2,3,5], [1,2,2,1], and [1,2,4,8] respectively. They can completely cover all pixels in the feature graph, the last layer of the encoder is connected. And is compared with the original U-Net network.

TABLE Ⅳ. INFLUENCE OF DIFFERENT DILATED RATE

model	U-Net	[1,2,3,5]	[1,2,2,1]	[1,2,4,8]
Acc/%	84.67	87.15	85.89	85.75
IoU/%	74.83	79.30	77.76	75.52

Results in Table Ⅳ show that the model with cascade dilated convolution module is relatively accurate, which also

confirms the theoretical analysis above. When the dilated rate is set as [1,2,3,5], the theoretical effective void convolution is the largest, and the highest segmentation accuracy can be achieved.

The semantic segmentation task based on deep learning can be considered as a dense pixel classification problem. Compared with the classification or recognition task of ordinary whole graph, the object to be predicted is each pixel, and there is a strong correlation between the pixels in different ranges of its surrounding areas. In order to make a correct judgment on the category corresponding to a pixel, the classification should not only be based on the limited set of pixels in a small range around it, but also consider the semantic information expressed as a whole. In view of the characteristics of state recognition of hybrid pavement which has large similarity and irregular shape among the scene categories, the improvement scheme is proposed: In the original U-Net network, concatenated hollow convolution modules are added to obtain context information of different ranges. The mean pixel accuracy and mean IoU are improved to 87.15% and 79.30%. The identification of three states (water, ice, snow) which has lower accuracy in original U-Net model improves significantly.

TABLE Ⅴ. RESULTS OF ORIGINAL U-NET AND D-UNET

status	U-Net		D-UNet	
	Acc/%	IoU/%	Acc/%	IoU/%
dry	84.95	81.28	84.87	82.73
wet	92.24	80.00	92.08	84.95
snow	82.46	74.19	88.35	79.75
ice	81.32	79.77	89.44	87.14
water	72.74	57.37	76.07	62.85
tracks	91.85	79.20	92.11	78.73
mean	84.67	74.83	87.15	79.30

Fig.8 shows the results in the test set. After improving the original U-Net network, the overall accuracy of recognition has been significantly improved, and the segmentation of small targets and boundary positioning are also more accurate. Fig.8(b-0) shows that in the mixed road surface with snow and tracks, there is a small area of long track road surface between the two snow-covered roads. U-Net ignores the small area and that is 'undersegmentation'. However, the improved D-UNet can accurately classify this 'small target' in the large background. In Fig.8(c-0) and Fig.8 (e-0), it can be intuitively seen that the improved network has stronger ability for boundary characterization and the segmentation at the boundary of different states is smoother and more accurate.

VI. CONCLUSION

The road surface status recognition algorithm based on deep semantic segmentation model in this paper has achieved considerable results on the self-made dataset. Considering that the road surface status recognition is different from the general semantic segmentation in that it is very similar between different categories. In order to improve the accuracy of recognition, the large receptive field needs to provide overall global information to reduce the probability of misclassification. Therefore, the improvement is conducted from the aspect of integrating context information, and the effectiveness of the improvement is verified through experiments. The mean accuracy of pixel and mean ratio of intersection can reach 87.15% and 79.30%, which provides a new idea for solving the problem on recognition of road surface condition.

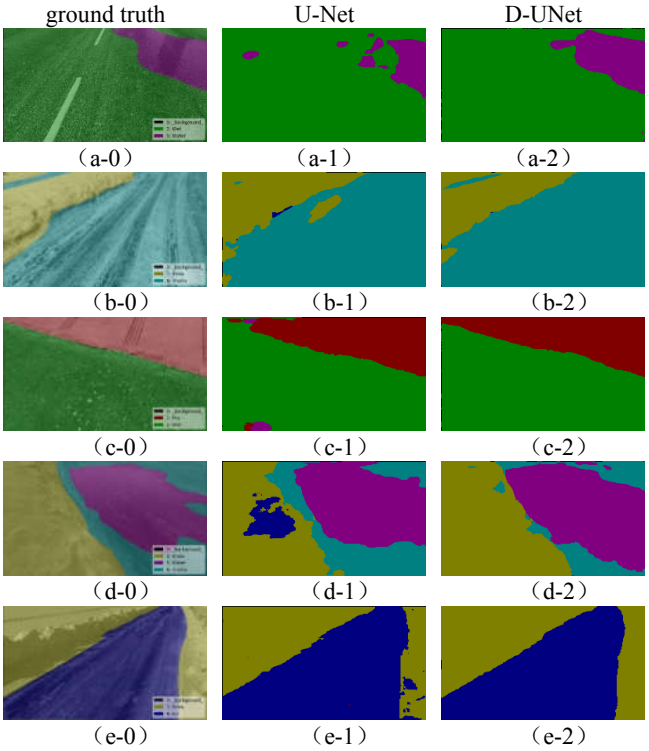


Fig.8 Ground truth and segmentation results of U-Net, D-UNet

REFERENCES

[1] Norrman J, Eriksson M, Lindqvist S. *Relationships between road slipperiness, traffic accident risk and winter road maintenance activity*[J]. Climate Research, 2000, 15(3): 185-193.

[2] Omer R, Fu L. *An automatic image recognition system for winter road surface condition classification*[C]. in: IEEE Conf. on Intelligent Transportation Systems, 2010: 1375-1379.

[3] Kawai S, Takeuchi K, Shibata K, et al. *A method to distinguish road surface conditions for car-mounted camera images at night-time*[C]. in: Int'l Conf. on ITS Telecommunications, IEEE, 2012: 668-672.

[4] Jonsson P, Casselgren J, Thornberg B. *Road Surface Status Classification Using Spectral Analysis of NIR Camera Images*[J]. IEEE Sensors Journal, 2015, 15(3): 1641-1656.

[5] Geiger A, Lenz P, Urtasun R. *Are we ready for autonomous driving? The KITTI Vision Benchmark Suite*[C]. in: IEEE Conf. on Computer Vision and Pattern Recognition. 2012: 3354-3361.

[6] Nolte M, Kister N, Maurer M. *Assessment of Deep Convolutional Neural Networks for Road Surface Classification*[J]. arXiv preprint arXiv, 2018, 2: 1804.08872.

[7] Pan G, Fu L, Yu R, et al. *Winter Road Surface Condition Recognition Using a Pre-trained Deep Convolutional Neural Network*[C]. in: Transportation Research Board 97th Annual Meeting, 2018: 838-855.

[8] Long J, Shelhamer E, Darrell T. *Fully convolutional networks for semantic segmentation*[C]. in: IEEE Conf. on Computer Vision and Pattern Recognition. 2015: 3431-3440.

[9] Yu F, Koltun V. *Multi-Scale Context Aggregation by Dilated Convolutions*[C]. in: Int'l Conf. on Learning Representations, 2016, 1511(7122): 1-13.

[10] Papandreou G, Kokkinos I, Savalle P A. Modeling local and global deformations in Deep Learning: Epitomic convolution, Multiple Instance Learning, and sliding window detection[C]. in: IEEE Conf. on Computer Vision and Pattern Recognition, 2016: 390-399.