Road Damage Detection and Classification Using Mask R-CNN with DenseNet Backbone

Qiqiang Chen[1, *], Xinxin Gan[2], Wei Huang[1], Jingjing Feng[1] and H. Shim[3]

Abstract: Automatic road damage detection using image processing is an important aspect of road maintenance. It is also a challenging problem due to the inhomogeneity of road damage and complicated background in the road images. In recent years, deep convolutional neural network based methods have been used to address the challenges of road damage detection and classification. In this paper, we propose a new approach to address those challenges. This approach uses densely connected convolution networks as the backbone of the Mask R-CNN to effectively extract image feature, a feature pyramid network for combining multiple scales features, a region proposal network to generate the road damage region, and a fully convolutional neural network to classify the road damage region and refine the region bounding box. This method can not only detect and classify the road damage, but also create a mask of the road damage. Experimental results show that the proposed approach can achieve better results compared with other existing methods.

Keywords: Road damage detection, road damage classification, Mask R-CNN framework, densely connected network.

1 Introduction

Roads are essential city infrastructure. Therefore, keeping roads in good condition is of critical importance for both vehicle and pedestrian's safety. Thus, monitoring the road damage and classifying it is an important aspect of road maintenance. But it is also very costly, time-consuming, labor-intensive, and it is often dangerous due to exposure to on-going traffic. Auto road damage detection methods, such as the laser scanner method [Yu and Salari (2011)] have been proposed to reduce the effort involved. However, those require special auto detection equipment, which is considerably expensive and has yet to be widely adopted. Thanks to the continuous development of image processing and machine vision, automatic road damage detection can now be performed with image processing technology.

[1] School of Computer and Communication Engineering, Zhengzhou University of Light Industry, Zhengzhou, 450000, China.

[2] SIPPR Engineering Group Co., Ltd., Zhengzhou, 450000, China.

[3] College of Electronic and Electrical Engineering, Sungkyunkwan University, Suwon, 440746, Korea.

[*] Corresponding Author: Qiqiang Chen. Email: 2015048@zzuli.edu.cn.

Received: 24 April 2020; Accepted: 08 June 2020.

CMC. doi:10.32604/cmc.2020.011191 www.techscience.com/journal/cmc

The traditional road damage detection image processing method is to check the image intensity and image gradient feature of every image pixel, and then determine which pixels in the image are the road damage. These methods include crack detection algorithm based on thresholding, noisy pavement images segmentation method based on iterative clipping and pavement image threshold based on adjacent difference histogram. These threshold-based algorithms are simple and easy to perform, but they cannot easily adapt to the noise, to the light condition and shadow in the image. According to the morphology, Jahanshahi et al. [Jahanshahi, Masri, Padgett et al. (2013)] proposed a method based on depth perception to detect and quantify the cracks. Hu et al. [Hu and Zhao (2010)] proposed a novel local binary patterns-based operation for pavement crack detection. Zou et al. [Zou, Cao, Li et al. (2012)] developed CrackTree, which uses a geodesic shadow-removal algorithm to remove the pavement shadows while preserving cracks, and build a crack probability map using tensor voting, sampling a set of crack seeds and using recursive tree-edge pruning to identify desirable cracks. Given the complex background and harsh lighting conditions of the road images, the hand crafted road damage features are not sufficiently discriminative to differentiate the road damage from normal road correctly.

Due to the on-going development of artificial intelligence and deep learning, the image features of objects can be learned by deep neural networks now [Krizhevsky, Sutskever and Hinton (2012); Girshick, Donahue, Darrell et al. (2014); Zhang, Wang, Li et al. (2018); Wang, Jiang, Luo et al. (2019)]. In recent years, the application of deep neural networks to object feature extraction in the road damage detection and classification task have been proposed. For example, Zhang et al. [Zhang, Yang, Zhang et al. (2016)] proposed a supervised deep convolutional neural network that was trained to classify each image patch. Compared with features extracted with hand-craft methods, the learned deep features provide superior crack detection performance. Fan et al. [Fan, Bocus, Zhu et al. (2019)] also use a deep convolutional neural network to determine whether an image contained cracks, and uses an adaptive thresholding method to extract cracks from the road surface. However, no common dataset exists for deep learning for road damage detection. Maeda et al. [Maeda, Sekimoto, Seto et al. (2018)] prepared a large-scale road damage data set, which contains 9053 road damage images captured using smartphone, and uses a convolutional neural networks to train a damage detection model and classify eight types damage. Ale et al. [Ale, Zhang and Li (2018)] proposed a method of road damage detection using RetinaNet, which is a fast one-stage model that can detect road damages with relatively high accuracy. In 2018, the Damage Detection and Classification Challenge was held. Given an image containing the road, it is necessary to detect and classify all the damaged areas. Chen et al. [Chen, Zhang, Zhang et al. (2019)] use the road damage image dataset to train an object detection model based on a deep convolutional neural network, and deployed it on a low-cost embedded platform to form an embedded system. Wang et al. [Wang, Wu, Yang et al. (2018)] applied faster region convolutional neural network (Faster R-CNN) and data augmentation techniques to detect and classify the damage road. Singh et al. [Singh and Shekhar (2018)] used mask regional convolutional neural network (Mask R-CNN) to automate detection and classification of damage in road image. They demonstrated that their methods can perform the task quickly and effectively.

In the detection and classification of road damage, most deep learning neural networks use many convolutional layers to obtain high-level image features. There are some drawbacks, such as the vanishing gradient problem with the deeper network layers at the model training step. In addition, too many parameters in the model can lead to high computational complexity. To overcome these drawbacks, this paper introduces a densely connected convolution networks (DenseNet) into the Mask R-CNN framework. Mask R-CNN is an object instance segmentation model for object detection and classification [He, Gkioxari, Dollár et al. (2017)]. DenseNet connects each layer of convolutional layer to every other layer in a feed-forward fashion. The feature-maps of all preceding layers are used as inputs of each bellowed layer, and these feature-maps are used as inputs for all subsequent layers. DenseNet has the following advantages: it alleviates the vanishing-gradient problem, enhances the feature propagation, encourages feature reuse, and greatly reduces the number of parameters [Huang, Liu, Van Der Maaten et al. (2017)]. And Wang et al. [Wang, Li, Zou et al. (2020)] introduced DenseNet into MobileNet to reduce the parameters even further and increase classification accuracy. In the Mask R-CNN framework, we use DenseNet as the backbone to replace the VGG [Simonyan and Zisserman (2014)] or ResNet [He, Zhang, Ren et al. (2016)], to obtain the image feature maps effectively and accurately. We also use a feature pyramid network to combine the multi-scale features. After the backbone, the region proposal network is used to generate the road damage proposal region. Then road damage is classfied and the road damage's region bounding box is refined, using the region proposal results and the image feature maps, by fully convolutional neural networks.

The rest of paper is organized as follows: In Section 2, we present the proposed work, including the Mask R-CNN framework and the DenseNet feature extraction model. While Section 3 reports the training data set and the experimental results. Finally, in Section 4 we describe our conclusion.

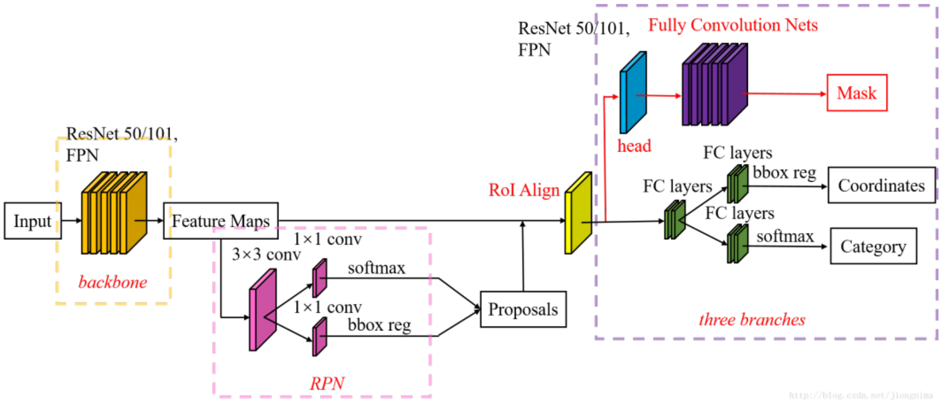


Figure 1: The traditional Mask R-CNN framework which contains feature extraction network (backbone), region proposal network (RPN) and three branches

2 Road damage detection algorithm

2.1 Mask R-CNN framework with DenseNet backbone

Road damage detection and classification can be regarded as an object detection and classification problem. We analyzed the current general object detection methods, especially the deep convolutional neural network methods. The Mask R-CNN performs well in object detection and classification. Zhang et al. [Zhang, Chang and Bian (2020)] proposed an improved Mask R-CNN for vehicle damage detection and segmentation, which performs well. Therefore, we chose Mask R-CNN [He, Gkioxari, Dollár et al. (2017)] as the basic framework for road damage detection and classification, as shown in Fig. 1. The traditional Mask R-CNN framework includes an image features extraction network, a region proposal network and three branches. The three branches are object masking, object bounding box refinement and object classification.

In the traditional Mask R-CNN framework, ResNet 50/101 is used as an image feature extraction network for road damage detection [Singh and Shekhar (2018)]. ResNet employs skip-connections, and uses identity functions to bypass the non-linear transformations. Therefore, the gradient can flow directly from back layers to the front layers through the identity function. The architecture of the ResNet 50/101 is shown in Tab. 1. However, it contains many parameters, may impede the flow information in the network, and may have the vanishing gradient problem. In order to overcome these drawbacks, we introduced a densely connected convolution networks (DenseNet) as the backbone of the traditional Mask R-CNN framework, replacing the ResNet 50/101 with DenseNet. The proposed road damage detection and classification framework is shown in Fig. 2. Compared with ResNet, DenseNet has the advantage of narrower network layers and fewer parameters. This is primarily due to the design of the dense connected block. As will be discussed later, the number of output feature maps per convolution layer in the dense block is very small, as opposed to hundreds of or thousands of widths of other networks. The output of each convolution layer is directly connected as input to all subsequent layers. At the same time, this connection makes the transmission of features and gradients more efficient, make full use of the feature maps of different levels, and the network is easier to train.

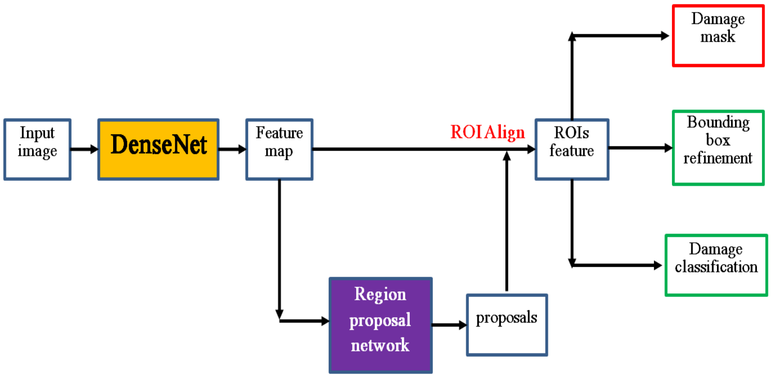


Figure 2: The proposed Mask R-CNN framework with DenseNet

In the proposed Mask R-CNN framework, DenseNet is used as the backbone of road image feature extraction, then different levels of feature maps are obtained. In order to combine the different scale features of the damage image, the feature pyramid network is used to create multiple-scale feature maps. Then the region proposal network is used, take the image feature maps as input and output a set of rectangular object region proposals. Each region proposal will be accompanied with a road damage score. The score is the probability of the road damage in the proposal region. On the feature maps, we can slide some windows, these windows have different scales and shapes (width and height ratio). Each window is called as anchor, and the features of each anchor are inputted into the region proposal network. The region proposal network will calculate the probability that the anchor is a foreground (road damage) or background. The regression parameters of window box can also be computed out. Using the anchor and the regression parameters of the box, it can calculate the position and shape of the region proposal anchor. In these anchor boxes, the region proposal with highest probability of foreground damage is chosen. Then use the NMS (non-maximum suppression) algorithm to remove the redundant and the duplicate anchors. The essence of NMS is to search for local maximum proposals and suppress non-maximum proposals. First, it finds the proposal with highest probability of road damage in the local area, and calculates the Intersection over Union (IoU) between the proposal with the highest score and other local proposals. If the IoU is higher than the threshold, the proposal is a non-maximum proposal and will be suppressed.

Table 1: ResNet with 50/101 convolution layers used for feature extraction in traditional Mask R-CNN framework [Singh and Shekhar (2018)]

Layers	ResNet50	ResNet101
Convolution	7×7 conversion, stride 2	7×7 conversion, stride 2
Pooling	3×3 max pool, stride 2	3×3 max pool, stride 2
Conv 2_x	$\begin{bmatrix} 1\times 1 \text{ conv} \\ 3\times 3 \text{ conv} \\ 1\times 1 \text{ conv} \end{bmatrix}\times 3$	$\begin{bmatrix} 1\times 1 \text{ conv} \\ 3\times 3 \text{ conv} \\ 1\times 1 \text{ conv} \end{bmatrix}\times 3$
Conv 3_x	$\begin{bmatrix} 1\times 1 \text{ conv} \\ 3\times 3 \text{ conv} \\ 1\times 1 \text{ conv} \end{bmatrix}\times 4$	$\begin{bmatrix} 1\times 1 \text{ conv} \\ 3\times 3 \text{ conv} \\ 1\times 1 \text{ conv} \end{bmatrix}\times 4$
Conv 4_x	$\begin{bmatrix} 1\times 1 \text{ conv} \\ 3\times 3 \text{ conv} \\ 1\times 1 \text{ conv} \end{bmatrix}\times 6$	$\begin{bmatrix} 1\times 1 \text{ conv} \\ 3\times 3 \text{ conv} \\ 1\times 1 \text{ conv} \end{bmatrix}\times 23$
Conv 5_x	$\begin{bmatrix} 1\times 1 \text{ conv} \\ 3\times 3 \text{ conv} \\ 1\times 1 \text{ conv} \end{bmatrix}\times 3$	$\begin{bmatrix} 1\times 1 \text{ conv} \\ 3\times 3 \text{ conv} \\ 1\times 1 \text{ conv} \end{bmatrix}\times 3$
Convolution Layer	3×3 conv	3×3 conv

After obtaining the region proposals using the region proposal network, we can use ROI Align to obtain the ROI (Region of Interest) feature of each region proposal. ROI Align cancels the quantization operation, and uses bilinear interpolation method to obtain the image feature value at the position where the coordinates are floating point numbers, thereby transforming the entire feature aggregation process into a continuous operation. It iterates through each region proposal, keeping the boundary of the floating point number box without quantization. The candidate region is divided into $k \times k$ units, and the boundary of each unit is not quantized. It calculates fixed four coordinate positions in each unit, calculates the values of these four positions through bilinear interpolation, and then performs the maximum pooling operation. Therefore, ROI Align does not contain quantization error.

There are three network heads at the end of the Mask R-CNN framework. One head is the road damage classification for each region proposal, which uses the fully connected layer. The result of this head is the probability of which type of road damage belongs to. The second head is bounding-box regression, which also uses the fully connected layer. The result of this head is used to obtain the accurate position of the damaged region and the shape of the bounding box. The third head is used used to obtain the mask of the road damage object through a fully convolutional network, that is, which pixels belong to the road damage in the region proposal box.

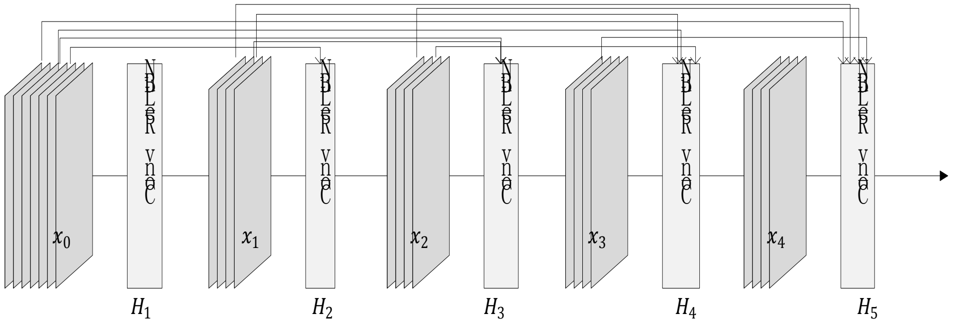


Figure 3: A densely connected module of DenseNet

2.2 Architecture of DenseNet

In the proposed framework of Mask R-CNN, DenseNet has certain advantages, used as a backbone network. It can effectively obtain image feature extraction maps, and then share it with the regional proposal network and the three head networks. DenseNet is a cross-connected model of convolutional neural network model, called concatenated network. Broadly speaking, DenseNet is a convolutional neural network which contains one or more densely connected modules. As shown in Fig. 3, the densely connected module is composed of multiple convolutional layers, that are connected in series through a series of operations, allowing internal cross-layer connections between any two non-adjacent layers. The input of each layer contains the output feature map of all the previous layers, and the output feature maps of each layer are used as the input features of the subsequent layers. The advantage of this structure is that can enhance feature propagation and promote feature reuse.

In the densely connected module, the feature map x_l of the l_{th} layer is calculated by using the feature maps $x_0, x_1, ..., x_{l-1}$ as input, that is, the output of all the previous layers are used as the input of the l_{th} layer. It is expressed as Eq. (1):

$$x_l = H_l([x_0, x_1, ..., x_{l-1}]) \tag{1}$$

Where $[x_0, x_1, ..., x_{l-1}]$ is the result of tensor splicing of the feature maps from the 0_{th} layer to the $(l-1)_{th}$ layer. These feature maps are as the input of the l_{th} layer. The standard $H_l(\ )$ of Eq. (1) is a composite function composed of three operations: batch normalization (BN), rectified linear unit (ReLU) and convolution (Convolution, Conv). Each function $H_l(\ )$ outputs k feature maps. So there are $k \times (l-1) + k_0$ feature maps as the input feature maps of the l_{th} layer, where k_0 is the number of input channels of the densely connected module, and k is the growth rate. In order to control the width of the network and improve the efficiency of the parameters, k is generally limited to a smaller integer. This control of the growth rate not only reduces the parameters of the closed network, but also ensures the performance of the densely connected module. The middle layer connects two densely connected module, and is called the transition layer. It usually composed of a block normalization layer, a 1×1 convolution layer and 2×2 average pooling layer. DenseNet consists of densely connected modules and transition layers, as shown in Fig. 4.

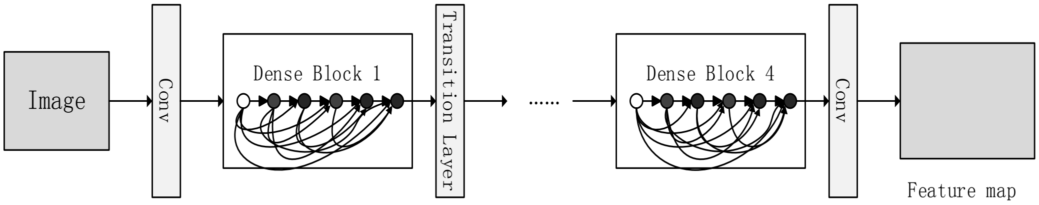


Figure 4: The architecture of the DenseNet, as used in the proposed Mask R-CNN for road image feature extraction

The DenseNet architecture used as backbone in the proposed Mask R-CNN is shown in Fig. 4. This DenseNet contains four densely connected blocks. The number of convolutional layers is similar with ResNet101, which can also contain 101 layers. After inputting the image, the first layer is a 7×7 convolutional layer and 3×3 max pooling with stride equal 2. Then, the first dense connected module is composed of six 1×1 convolutional layers and six 3×3 convolutional layers. After that is the first transition layer. The second densely connected module is composed of twelve 1×1 convolutional layers and twelve 3×3 convolutional layers. Then comes the second transition layer. The third densely connected module is composed of twenty-four 1×1 convolutional layers and twenty-four 3×3 convolutional layers. Then comes the third transition layer. The forth densely connected module is composed of six 1×1 convolutional layers and six

3×3 convolutional layers. The last layer of DenseNet is a 3×3 convolutional layer, and then the image feature maps is generated. Tab. 2 lists the detailed architecture of the densely connected network used. In order to improve the compactness of the model, we also reduce the number of the feature maps at transition layers as reference [Huang, Liu, Van Der Maaten et al. (2017)], that can generate half the number of feature maps using the transition layers.

Table 2: DenseNet with 101 convolution layers used for feature extraction

Layers	DenseNet 101
Convolution	7×7 conv, stride 2
Pooling	3×3 max pool, stride 2
Dense Block (1)	$\begin{bmatrix}1\times1 \text{ conv}\\ 3\times3 \text{ conv}\end{bmatrix}\times 6$
Transition Layer (1)	1×1 conv 2×2 average pool, stride 2
Dense Block (2)	$\begin{bmatrix}1\times1 \text{ conv}\\ 3\times3 \text{ conv}\end{bmatrix}\times 12$
Transition Layer (2)	1×1 conv 2×2 average pool, stride 2
Dense Block (3)	$\begin{bmatrix}1\times1 \text{ conv}\\ 3\times3 \text{ conv}\end{bmatrix}\times 24$
Transition Layer (3)	1×1 conv 2×2 average pool, stride 2
Dense Block (4)	$\begin{bmatrix}1\times1 \text{ conv}\\ 3\times3 \text{ conv}\end{bmatrix}\times 6$
Convolution Layer	3×3 conv

2.3 Loss function and training

For each anchor, the region proposal network will classify it as road damage or background, and also regress the position and size of bounding box. The loss function of region proposal network contains anchor classifier loss and bounding box loss. For each region proposal, the proposal will be passed to the three heads for the road damage classification, bounding box regression, and road damage segmentation. Their losses are road damage classifier loss, bounding box refinement loss, and mask pixel binary cross-entropy loss. It is a multitask training. The loss function of Mask R-CNN is shown in Eq. (2):

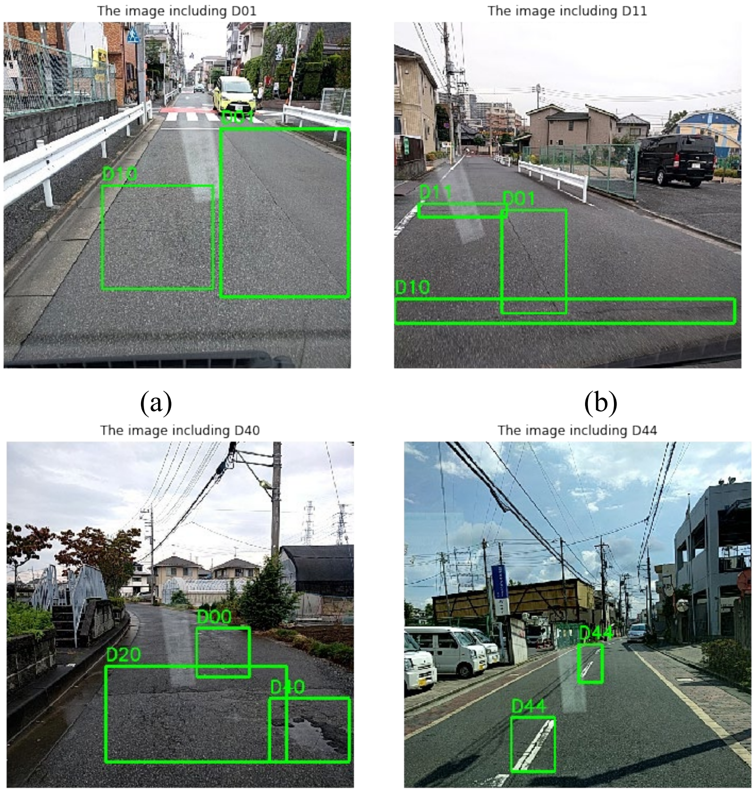
$$L = L_{cls} + L_{bbox} + L_{mask} \tag{2}$$

The first two terms L_{cls} and L_{bbox} are the same as the loss function in the Faster R-CNN model [Wang, Wu, Yang et al. (2018)], which represents the classification error and detection errors respectively. Eq. (3) shows the details of the loss function in Faster R-CNN.

$$L(\{p_i\},\{t_i\}) = \frac{1}{N_{cls}}\sum_i L_{cls}(p_i, p_i^*) + \lambda \frac{1}{N_{reg}}\sum_i p_i^* L_{reg}(t_i, t_i^*) \quad (3)$$

In Eq. (3), p_i represents the predicted probability value of anchor i being an object, p_i^* is 0 if the anchor box is negative, and it is 1 if the anchor box is positive in the region proposal network. L_{reg} is the bounding box regression loss. t_i indicates the 4 parameterized coordinates of the prediction candidate box, t_i^* refers to the 4 parameterized coordinates of the true value object area. L_{mask} in Eq. (2) is the mask branch loss function. The mask branch and the class prediction branch are independent, and a binary mask is independently predicted for each category, without relying on the prediction results of the classification branch. The mask branch has K binary masks for each ROI, and one mask for each of category. When calculating the mask loss, not all the categories calculate the binary cross-entropy loss, but which category the pixel belongs to, sigmoid output of which category is only used to calculate the loss.

For deep learning, a sufficient amount of data is required. However, in some cases it is difficult to obtain sufficient training data for specific problems. In this case, transfer learning can be used to solve the problem. Transfer learning requires source and target domains. Transfer learning can transfer the model parameters that have been trained in the source domain to the new model in the target domain to help train the new model. That is possible because most of the images have similar basic characteristics. Thus, the model pre-training can be done first on the large datasets such as COCO, and then the trained weights are migrated to the road damage datasets. The network parameters are then adjusted and refined based on the road damage datasets, especially the parameters of the region proposal network and the three branches. The training steps of the proposed framework (the improved Mask R-CNN with DenseNet backbone) are similar to the training steps of the traditional Mask R-CNN. Prepare training datasets and validation datasets for the road damage and detection target feature attributes. The COCO datasets are used with the Mask R-CNN with DenseNet as backbone to get the pre-training model weights. Then the pre-trained weights are used together with the road damage datasets to train the region proposal network and the three branches heads network. Finally, the parameters of all Mask R-CNN layers are adjusted through the loss functions.



(a) (b)

(c) (d)

Figure 5: The road image used in the training and testing. There are different types road damage in images. These images are source from Maeda et al. [Maeda, Sekimoto, Seto et al. (2018)]

3 Experiments

We use the Road Damage Detection and Classification Challenge's dataset as our training images and test images. This dataset contains 9053 road images, 15435 bounding boxes and eight types of road damage [Maeda, Sekimoto, Seto et al. (2018)]. The eight types of the road damage are listed in the Tab. 3. They are: wheel mark part, longitudinal linear crack; construction joint part, longitudinal linear crack; lateral equal interval linear crack; construction joint part, lateral linear crack; pavement alligator crack; and other corruption such as rutting, bump, pothole, separation, crosswalk blur, white line blur. Some images of the dataset are shown in Fig. 5. These images are captured using a smart phone installed on the car. The background and light conditions of the images are variable and complex. The image also includes different types of road damage, which are bounded by the green rectangles.

Table 3: Road damage types in the data set and the details definitions [Maeda, Sekimoto, Seto et al. (2018)]

Road damage type	Details	Class name
Linear Crack, Longitudinal	Wheel mark part	D00
Linear Crack, Longitudinal	Construction joint part	D10
Linear Crack, Lateral	Equal interval	D10
Linear Crack, Lateral	Construction joint part	D11
Alligator crack	Partial pavement, overall pavement	D20
Other corruption	Rutting, bump, pothole, separation	D40
Other corruption	Crosswalk blur	D43
Other corruption	White line blur	D44

In the road damage dataset, the annotation files only include the road damage type and damage bounding box. The marking tool Labelme must be used to label the road damage with polygon. Based on the image, its damage type, and damage bounding box, we draw polygons to obtain the truth mask of each road damage instance.

In our experiment, training was performed on a PC by running the Ubuntu 18.04 operating system with an NVIDIA GeForce RTX 2070 Super GPU and 16 GB RAM memory. In order to evaluate the damage detection results, we used the Intersection over Union (IoU). If the IoU of detected damage box bounding and ground truth object box bounding is greater than 0.5, it means that the detection is correct. To evaluate the damage classification results, we used several evaluation indexes. Accuracy rate (Accuracy) mainly refers to the number of samples correctly classified by the model compared to the total number of samples. Recall rate (Recall), which refers to how many positive samples were correctly predicted. Precision refers to how many of the positive samples were detected as positive. The road damage detection and classification results are shown in Fig. 6. The first column shows the datasets images obtained from the Road Damage Detection and Classification Challenge's dataset, which only includes the bounding box. The middle column shows the images labeled by the marking tool Labelme, which can yield a mask of the road damage. The third column shows the detection and classification results of our proposed method. It shows that the proposed method can detect and classify the road damage accurately. It can also correctly segment the road damage mask.

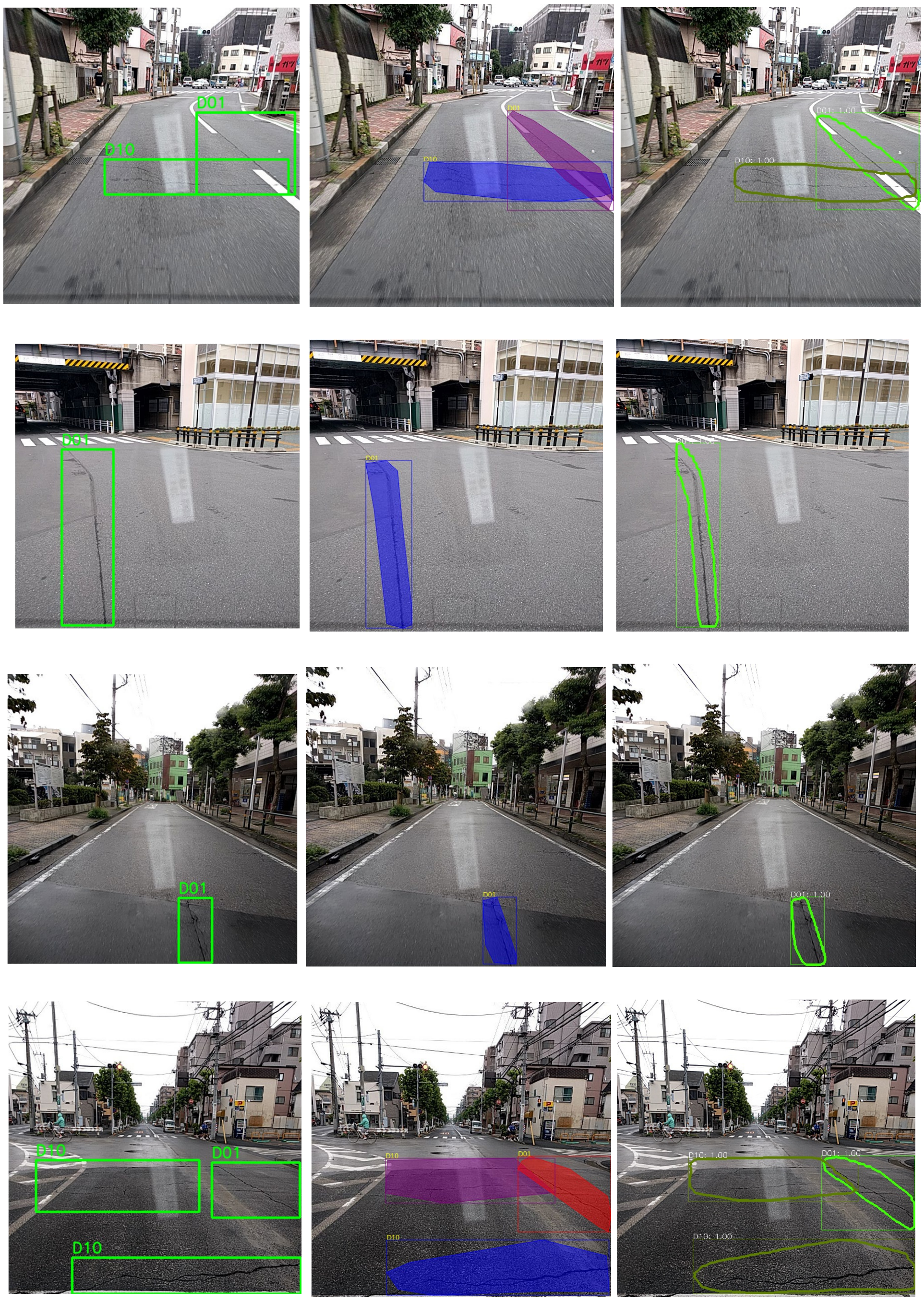


Figure 6: The road damage detection and classification using our proposed method. The first column shows only the datasets with the bounding boxes. The middle column shows the labeled dataset using the mark tool called Labelme. The third column shows the detection and classification of our proposed method

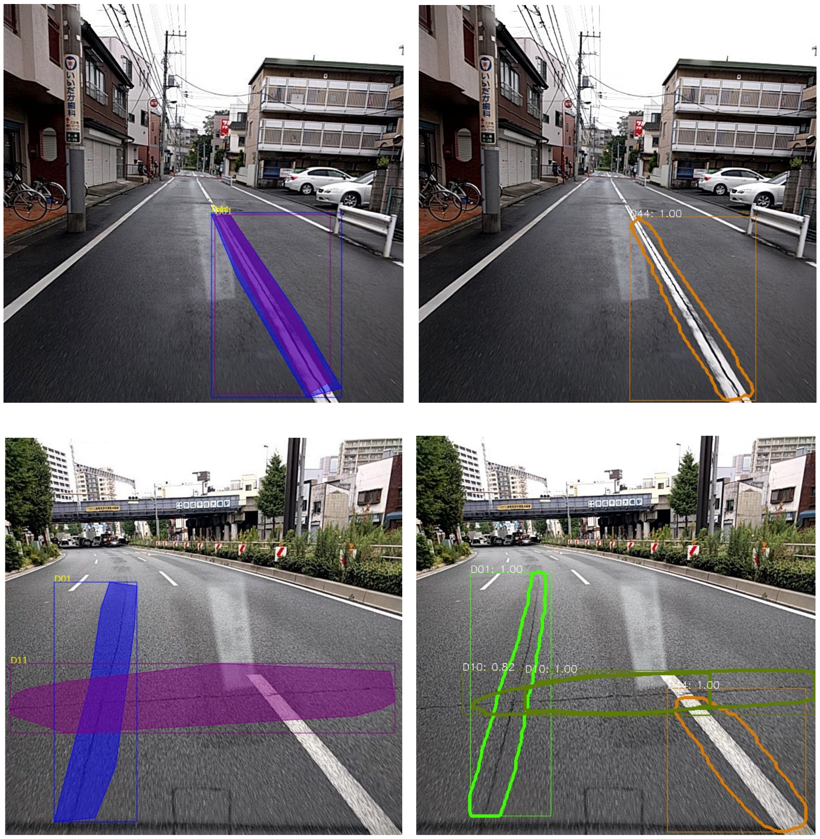


Figure 7: Road damage detection and classification error of our proposed method. The left column shows the truth damages. The right column shows the detections and classifications of our proposed method

However, our proposed method sometimes does errors, as shown in Fig. 7. In the first row, the longitudinal linear crack damage is covered by the white line blur damage. Our proposed method only found out the longitudinal linear crack damage. That is because NMS filters out the IoU greater than the threshold. The error in bottom row is that bright-line is blurred. That is not detected by the handcraft. But there is a little blur. Our proposed method can find out the blur damage accuracy.

4 Conclusion

In this paper, we proposed Mask R-CNN framework with DenseNet as the backbone for feature map extraction, a region proposal network for region proposal generation. And three neural network heads are used for road damage detection, proposal bounding box refinement, and the road damage classification. It can also segment road damage at pixel level. In our approach, we use the dataset from the Road Damage Detection and Classification Challenge in 2018 IEEE Big Data Cup, the images of this dataset were captured using a smartphone installed on the car. We also used the Labelme tool to mark the road damage with polygons. To the best of our knowledge, this is the first time that

DenseNet has been applied to the Mask R-CNN framework for road damage detection and classification. Comprehensive experimental results on the dataset have demonstrated that our approach, using Mask R-CNN framework with DenseNet as backbone, can obtain better results compared with other current methods. Due to insufficient understanding of road damage image type characteristics and un-uniform distribution of various types of road damage data, there is still room for improvement in the trained model. Further research might consider pre-processing before training, such as data augmentation of the data.

Funding Statement: The work was supported by the School Doctoral Fund of Zhengzhou University of Light Industry No. 2015BSJJ051.

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

References

Ale, L.; Zhang, N.; Li, L. (2018): Road damage detection using RetinaNet. *IEEE International Conference on Big Data*, pp. 5197-5200.

Chen, S. Y.; Zhang, Y.; Zhang, Y. H.; Yu, J. J.; Zhu, Y. X. (2019): Embedded system for road damage detection by deep convolutional neural network. *Mathematical Biosciences and Engineering*, vol. 16, no. 6, pp. 7982-7994.

Fan, R.; Bocus, M. J.; Zhu, Y.; Jiao, J.; Wang, L. et al. (2019): Road crack detection using deep convolutional neural network and adaptive thresholding. *IEEE Intelligent Vehicles Symposium*, pp. 474-479.

Girshick, R.; Donahue, J.; Darrell, T.; Malik, J. (2014): Rich feature hierarchies for accurate object detection and semantic segmentation. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 580-587.

He, K.; Gkioxari, G.; Dollár, P.; Girshick, R. (2017). Mask r-cnn. *Proceedings of the IEEE International Conference on Computer Vision*, pp. 2961-2969.

He, K.; Zhang, X.; Ren, S.; Sun, J. (2016): Deep residual learning for image recognition. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770-778.

Hu, Y.; Zhao, C. X. (2010): A novel LBP based methods for pavement crack detection. *Journal of Pattern Recognition Research*, vol. 5, no. 1, pp. 140-147.

Huang, G.; Liu, Z.; Van Der Maaten, L.; Weinberger, K. Q. (2017): Densely connected convolutional networks. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4700-4708.

Jahanshahi, M. R.; Masri, S. F.; Padgett, C. W.; Sukhatme, G. S. (2013): An innovative methodology for detection and quantification of cracks through incorporation of depth perception. *Machine Vision and Applications*, vol. 24, no. 2, pp. 227-241.

Jiongnima (2018): Detailed explanation of the example segmentation model Mask R-CNN: from R-CNN, Fast R-CNN, Faster R-CNN to Mask R-CNN.

https://blog.csdn.net/jiongnima/article/details/79094159.

Krizhevsky, A.; Sutskever, I.; Hinton, G. E. (2012): ImageNet classification with deep convolutional neural networks. *Advances in Neural Information Processing Systems*, vol. 25, no. 2, pp. 1097-1105.

Maeda, H.; Sekimoto, Y.; Seto, T.; Kashiyama, T.; Omata, H. (2018): Road damage detection and classification using deep neural networks with images captured through a smartphone. *Computer Aided Civil and Infrastructure Engineering*, vol. 33, no. 2, pp. 1127-1141.

Singh, J.; Shekhar, S. (2018): Road damage detection and classification in smartphone captured images using Mask R-CNN. arXiv preprint arXiv:1811.04535.

Simonyan, K.; Zisserman, A. (2014): Very deep convolutional networks for large-scale image recognition. *Proceedings of the 3rd International Conference on Learning Representations.* http://arxiv.org/abs/1409.1556.

Wang, W.; Jiang, Y. B.; Luo Y. H.; Li J.; Wang X. et al. (2019): An advanced deep residual dense network (DRDN) approach for image super-resolution. *International Journal of Computational Intelligence Systems*, vol. 12, no. 2, pp. 1592-1601.

Wang, W.; Li, Y. T.; Zou, T.; Wang, X.; You, J. Y. et al. (2020): A novel image classification approach via Dense-MobileNet models. *Mobile Information Systems*, vol. 2020, pp. 1-8.

Wang, W.; Wu, B.; Yang, S.; Wang, Z. (2018): Road damage detection and classification with Faster R-CNN. *IEEE International Conference on Big Data*, pp. 5220-5223.

Yu, X.; Salari, E. (2011). Pavement pothole detection and severity measurement using laser imaging. *Proceedings of IEEE International Conference on Electro/Information Technology*, pp. 1-5.

Zhang, Q.; Chang, X.; Bian, S. B. (2020). Vehicle-damage-detection segmentation algorithm based on improved Mask RCNN. *IEEE Access*, vol. 8, pp. 6997-7004.

Zhang, L.; Yang, F.; Zhang, Y. D.; Zhu, Y. J. (2016): Road crack detection using deep convolutional neural network. *Proceedings of IEEE International Conference on Image Processing*, pp. 3708-3712.

Zhang, Y.; Wang, Q.; Li, Y.; Wu, X. (2018): Sentiment classification based on piecewise pooling convolutional neural network. *Computers, Materials & Continua*, vol. 56, no. 2, pp. 285-297.

Zou, Q.; Cao, Y.; Li, Q.; Mao, Q.; Wang, S. (2012): CrackTree: automatic crack detection from pavement images. *Pattern Recognition Letters*, vol. 33, no. 3, pp. 227-238.