Road pavement crack detection using deep learning with synthetic data

I A Kanaeva and Ju A Ivanova
Division for Information Technology, Tomsk Polytechnic University, Tomsk, Russia
E-mail: iap15@tpu.ru

Abstract. The improvement of road system quality is a critical task. The mechanism to address such important issue is close monitoring of road pavement condition. Traditional approach requires manual identification of damages. Taking into account considerable length of road system it is essential to create an effective automatic pavement defects detection tool. This approach will extremely reduce time for monitoring of current road state. In this paper global experience in solution of detection issues of road pavement's distress is reviewed. The article includes information about the existing datasets of road defects, which are commonly used for detection and segmentation. The present work is based on deep learning approach with the use of synthetic generated training data for segmentation of cracks in driver-view image. The novelty of the approach lies in creating synthetic dataset for training state-of-the-art deep learning frameworks. The relevance of the research is emphasized by processing of wide-view images in which heterogeneous pixel intensity, complex crack topology, different illumination condition and complexity of background make the task challenging.

1. Introduction

Infrastructure in modern economic model is a sufficiently important key point. From the perspective of Russian Government statistical data analyses the 67.1% of all freights in 2018 were transported by road. Moreover in recent years the volume of carriage steadily increase as well as the amount of private vehicles. All of these factors considerably influence the road pavement condition. Deterioration of road system in turn lead to decrease in shipment transport speed and also increase the number of road accidents. This is precisely why the task of road system quality monitoring is practically essential.

Furthermore, every third car accident in Russian Federation occurs due to unsatisfactory maintenance and furnishing of road system. According to the data of World Health Organization (WHO) each year over 1.35 million people die on roads and more than 50 million get injures. Analogous research in European region shows that 120 thousand of road users die and 2.4 million get injures each year. In Russia 20 thousand traffic deaths and 221 thousand injuries were reported in 2016. The last data in 2019 shows that 17 thousand people died and 210 thousand were injured. Despite the steady decline of numbers of death accidents from 2012 till nowadays the death rate in Russian federation is still beyond than average in Europe. Recently, in the Russian Federation, the national project "Safe and high-quality roads" has been carried and now it is actively developing. The goal of this project is to improve the quality of significant regional roads and road network of urban agglomerations up to the standard state. The released program is directed not only to traffic improvement, but also to reduction of deaths numbers on roads in 3.5 times.

Considering that fact that Russian road system is 5th longest in the world, the road diagnostic system critically need to automate routine tasks such as defects detection. Due to the sufficient length of the

road, some aspects of the diagnostic process require analysis of a considerable amount of video data. For this aim the computer vision may be effectively implemented.

Current progress in computer vision is based on deep learning and has been achieved mostly due to large image data sets with careful annotations. Such spheres as for example autonomous driving systems are actively developing due to accurate semantic segmentation datasets, such as Cityscapes [1], Wilddash [2] and KITTI [3]. Manually labeling process of such kind of data is expensive and labor-consuming.

Due to the need of computer, processing of high-quality video data of roads in the road industry there is requirement for developing automatic algorithm of road surface defects detection by images. The development of an effective algorithm of pavement defects identification is notably important task. On one hand, it is able to improve quality and speed of the monitoring process. On the other hand, repair roadworks carried in time not only allow to maintain the standard road condition, but also preclude significant amount of accidents, inclusively resulting in death.

2. Related works

Current methods of monitoring road pavement can be divided into three types: 3D scanning-based, vibration-based and 2D image vision-based. The first group of methods are based on different sensors such as inertial measurements, 3D scanners, and optical sensors. A high cost of LIDAR system and complexity of exploitation and data post-processing are limit their use. Accelerometer is a principal component of vibration-based methods. It measures vibrations caused in a vehicle due to hitting a pothole or other road damages.

Over the past decade, many different techniques of image processing have been proposed in the field of automatic detection, classification and segmentation of pavement distress. A. Mohan and S. Poobal in [4] reviewed 50 research papers in the area and provided the collective survey of the different image processing techniques used for the detection of the cracks in the engineering structures.

In a review [5] on automated pavement distress detection methods authors conducted that the highest need for further research is demanded from the surface defects: raveling, polished aggregate and bleeding. German researchers in paper [6] divided the algorithms developed for evaluation of the pavement surface into three major groups: crack threshold filtering, patch-based classification, and depth-based algorithms.

The first group of methods designed explicitly for crack detection, mostly by applying image processing methods that segment distress textures by threshold filtering. In order to reduce illumination artifacts image-preprocessing algorithms are preceded. Due to pixels of cracks have minimum intensity threshold filtering is applied after preprocessing. On last stage, the detection is refined by morphological image operations and by searching for connected components. Next papers present the aforementioned approach.

In [7] the research results were implemented as a CrackIT software tool for segmentation of cracks in an image taken directly above the road surface. The CrackTree [8] toolbox is based on the construction of probability map of pixels belonging to a crack on the image that previously was preprocessed by geodesic shadow-removal algorithm. Paper [9] proposes a new unsupervised multi-scale fusion crack detection algorithm that works on a series of images smoothed by different-scale Gauss filters and combines the resulting masks. Gabor filters in [10] are used for searching areas as cracks.

The algorithms of the second group apply various types of classifiers to patch of the image to determine distress areas or cracks. One part of the researchers initially selects a certain feature vector from the considered image region, and then use it as input of the classifier. The advantage of this approach is that the size of the region is not fixed, but initially you need to divide the image into a sufficient number of regions. In [11] regions are defined using the over-segmentation algorithm SLIC. Then support vector machines were used as a binary classifier. Despite the fact that the method does not have high accuracy, it allows to calculate the ratio of damaged and non-damaged pavement. In addition, this method can be expanded for different defects detection, such as pothole, manhole and road marking.

With the advent of public available datasets of road images with pavement distress, such as GAPs [5] and CRACK500 [12], many researchers have used deep learning approaches for the problem. For example, in [13] a truncated CNN VGG16 is used for extract the feature vector from the input image. Next, a neural network with one hidden layer of 256 neurons classifies the feature vector.

3. Existing datasets

Consider the most popular and public available datasets for road distress classification, segmentation and detection tasks:

- GAPs dataset [5]: includes total 1,969 gray valued images with resolution is 1,920×1,080 pixels. These images are divided into 64×64 patches and each patch is labeled as a crack or not. The pictured surface material contains pavement of three different German federal roads.
- CRACK500 dataset [14]: consists of 500 RGB images of pavement cracks of size around 2,000×1,500 pixels that were collected on main campus of Temple University using cell phones. Each crack image has a pixel-level annotated binary map.
- CrackTree200 dataset [8]: includes 206 pavement images of size 800×600 with various types of cracks that were annotated in pixel-wise level. The images have complex asphalt context with shadows, occlusions, low contrast and noise.
- CFD dataset [15]: have 118 images of annotated road crack of size 480×320 pixels. These images was captured on urban Beijing roads. The images contain a significant amount of noisy pixels like oil spots and water stains, and some of them are under the poor illumination condition.
- Road Damage dataset [16]: consist of 9,053 labelled road images of size 600×600, acquired from a smartphone camera installed on the dashboard of a car. The main aim of this is capturing general front view from driver's position, as opposed to capturing images above the road surface. It decreases difficulty of capturing image process and increases practical applicability of such images. The dataset has 15,435 bounding boxes of damages in total, annotated for the dataset.

The Road Damage dataset was collected in seven Japan municipalities and has eight classes of pavement distress: five types of cracks, two types with wear of road marking, one class that combines other damages like rutting, bump, pothole and separation. The dataset has a PASCAL VOC [17] format and was presented at the IEEE International Conference on Big Data Cup in 2018.

The dataset of Japanese scientists has revived interest in solving the challenge of automatic damage detection on wide-view road images using convolutional neural networks. This was facilitated by sufficiently large scale of the dataset and presence of different distress types, not only cracks. Disadvantages of the dataset is that bounding boxes may include a lot of redundant information due to various forms and sizes of damage on image, especially diagonal cracks on road. For maintains of road pavement condition pixel-level segmentation task is more appropriate and allows localize damage and estimate distress size.

Creating dataset for road distress segmentation with high quality pixel-level annotations is a laborious and time-consuming process due to manual labeling of object's pixels in the image. Another approach to deal with this problem is a synthetic data generation. In next section, we propose algorithm for generation instance-level synthetic dataset for crack segmentation based on well-known collections with a marked road, such as KITTI and Cityscapes dataset.

4. Related work

4.1. Synthetic dataset creation

At current stage of machine learning evolution, the formation of a set of training data have paramount importance for the successful solution of the tasks of detection and segmentation. However, the meticulous manual targeting of several thousands of images is an enormous and sufficiently labor-consuming process, so quite important task is to develop methods for obtaining a representative synthetic samples.

A synthetic dataset is a repository of data that is generated programmatically and cannot be collected by any real-life survey or experiment. For creating the synthetic road crack dataset, we decided to use three publically available sets: KITTI and Cityscapes dataset as images of the road scene, CFD as a source of cracks marked at the pixel level (Figure 1).

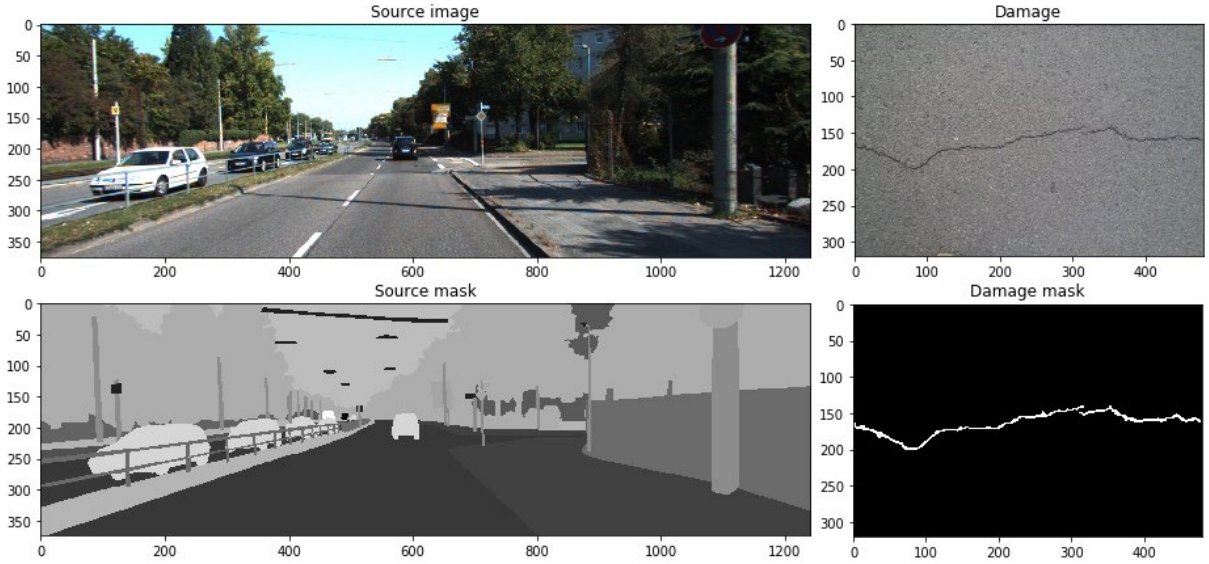


Figure 1. Road and crack images with masks.

To determine the part of the image corresponding to the roadway, all the pixels of the road mask are chosen. Then, the connected components algorithm with 8-connectivity is applied to the resulting binary mask for determine connected areas. As a result, the area with the maximum number of pixels is taken as the main road mask (Figure 2, main road mask is highlighted in gray).

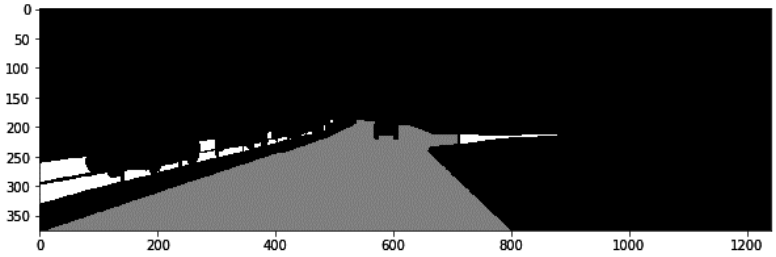


Figure 2. Maximal area of roadway.

Then the original image of the crack and its mask are cropped at the minimum bounding rectangular for the purpose of reducing following calculations. After that, the image of crack with the mask is scaled and rotated randomly. The result is an image of the crack D and its mask D^{mask} (Figure 3).

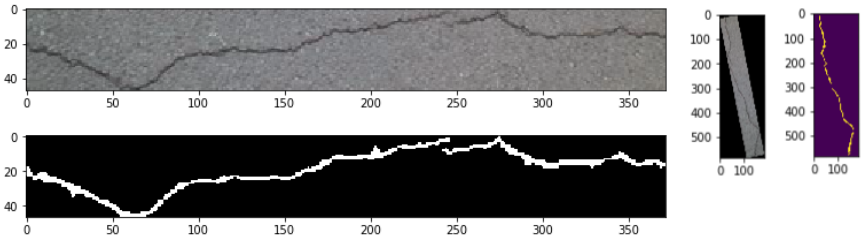


Figure 3. Cropping, rotating and scaling of crack.

In the next step a point, which is the center of the overlapping area, is selected inside the mask of the main area of the road, and the S^{mask} is cut out of the original image mask, equal in size to D^{mask}. To mix two images in mask area, the mean asphalt value $\overline{D_c}$ are calculated in each channel c for the image of the crack that does not lie under the mask:

$$\overline{D_c} = \frac{1}{k}\sum_p D_c(p)\cdot\left(1-D^{mask}(p)\right), c \in \{R,G,B\}, \quad (1)$$

where $D_c(p)$ – pixel value p of crack image D in channel c, $D^{mask}(p)$ – pixel value p in binary crack mask, k – number of pixels for which $D^{mask}(p)=0$. This calculation process is possible due to the homogeneous texture of the asphalt on the image of the defect. Then, when a crack is applied to an image of a road, only the values under the mask of the roadway and the cracks are considered:

$$M(p) = D^{mask}(p)\cdot S^{mask}(p) \quad (2)$$

The crack is added by changing the pixel values of the original image S. These pixels are located under the common mask M and their values are multiplied by the ratio of the calculated mean asphalt surface to the crack pixel values:

$$S_c(p) = S_c(p)\cdot M(p)\cdot D_c(p)/\overline{D_c} \quad (3)$$

The algorithm result is shown in Figure 4.

Figure 4. The result of synthetic crack generation.

For the purpose of increasing the informational capacity of the training image sample, from 1 to 5 cracks, which can intersect were generated on one image. As a result, the training set contains 1524 images, and the test dataset includes 505.

4.2. Object Segmentation System

To solve the crack detection problem in pixel level as segmentation task, we decided to use two modern convolution neural network systems: Mask R-CNN [18] and U-Net [19]. Consider its structures and main principles of operation.

4.2.1. Mask R-CNN is a state-of-the-art framework for detecting objects in an image while simultaneously generating a high-quality segmentation mask for each instance. Mask R-CNN extends object detection algorithm Faster R-CNN by adding a module for predicting an object mask. Simplify Mask R-CNN structure is illustrated in Figure 5. Mask R-CNN have a complex, flexible and powerful block architecture with two stage: generating object proposals and classifying proposals to generate bounding boxes and masks in parallel.

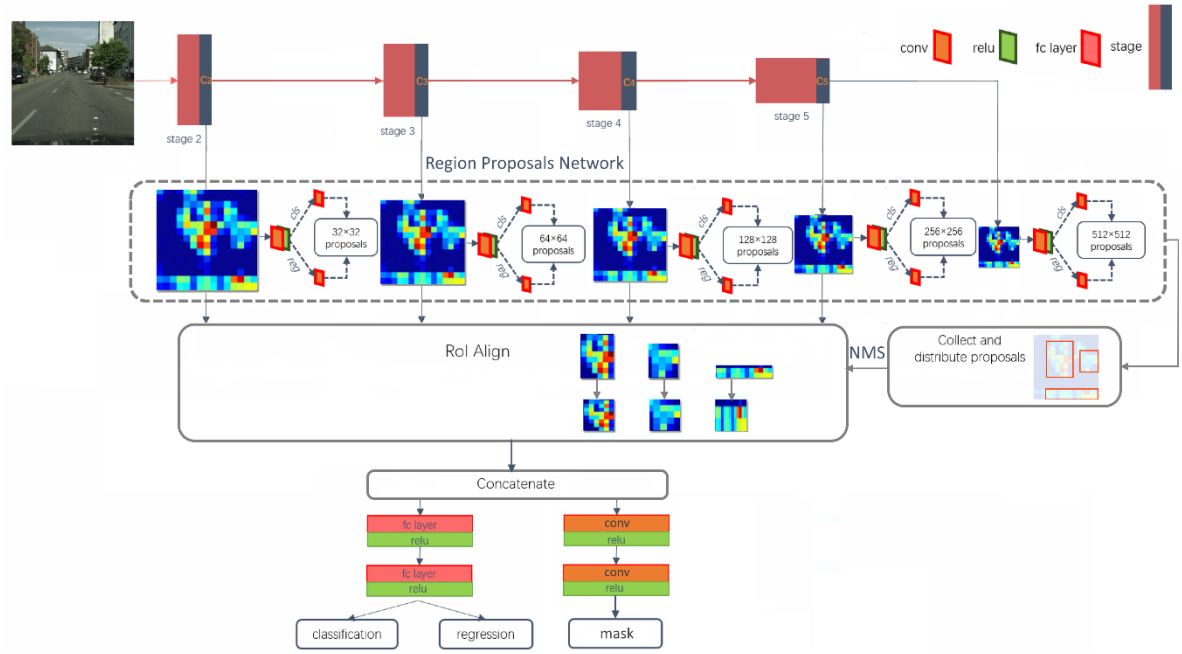


Figure 5. Mask R-CNN architecture.

Initially, the image is fed to the input of Mask R-CNN to produce a feature map, which often uses pre-trained VGG16 or ResNet50/101 with excluded layers responsible for classification and named backbone. One of the improvements in this framework is using Feature pyramid networks (FPN) for generation multi-scale feature maps. Sequential layers of FPN with decreasing dimension are considered as a hierarchical pyramid, in which the lower level maps have high resolution, and the upper level maps have high generalizing, semantic ability.

The resulting feature maps are processed in CNN Region Proposals Network (RPN), whose task is to create the regions of interests (RoIs) which may contain objects. For this purpose, each feature map is scanned by lightweight neural network with a 3×3-convolution layer. Output of RPN is based on k anchors - set of boxes with predefined locations and scales relative to images. For each anchor, RPN generates a probability of a proposal having the target object, and a refinement of the coordinates of the bounding box of the object. The purpose of this stage is to identify regions of interests that may contain objects. At the end, duplicate proposal regions are discarded due to non-maximum suppression operation.

Then the proposals are mapped from corresponding feature map levels, extracted from its and resized to the same size using the RoI Align operation. According to RoIs, the final operations of classification, refinement of the bounding box's coordinates and mask prediction are performed at the second stage. The output mask has a greatly reduced size, but contains real values, which allow to obtain sufficient accuracy by scaling the mask to the size of the selected object's bounding box.

4.2.2. U-Net model is a fully convolutional network that outputs a classification of each pixel in the image to generate a segmentation mask. U-Net architecture consists of a contracting path to capture context and of a symmetrically expanding path that enables precise localization. The contracting path follows the typical architecture of a convolutional network with alternating convolution and pooling operations and progressively down-samples feature maps, increasing the number of feature maps per layer at the same time. Every step in the expansive path consists of an up-sampling of the feature map followed by a convolution.

Typically, U-Net is trained from scratch starting with randomly initialized weights. In order to account that our training dataset is synthetically generated dataset, it only models plain cracks on road surface without perspective distortion and light aspects, we use transfer learning similar to TernausNet

[20]. We used U-Net type architecture improved by the using of the pre-trained VGG16 on ImageNet as the encoder. Initialized weights from the pre-trained network are frozen. This approach allows significantly reducing the number of trainable parameters and shows better performance than training from scratch. To construct an encoder, we kept only first four convolution blocks with last convolutional layer with 512 channels. This U-Net architecture is illustrated in Figure 6.

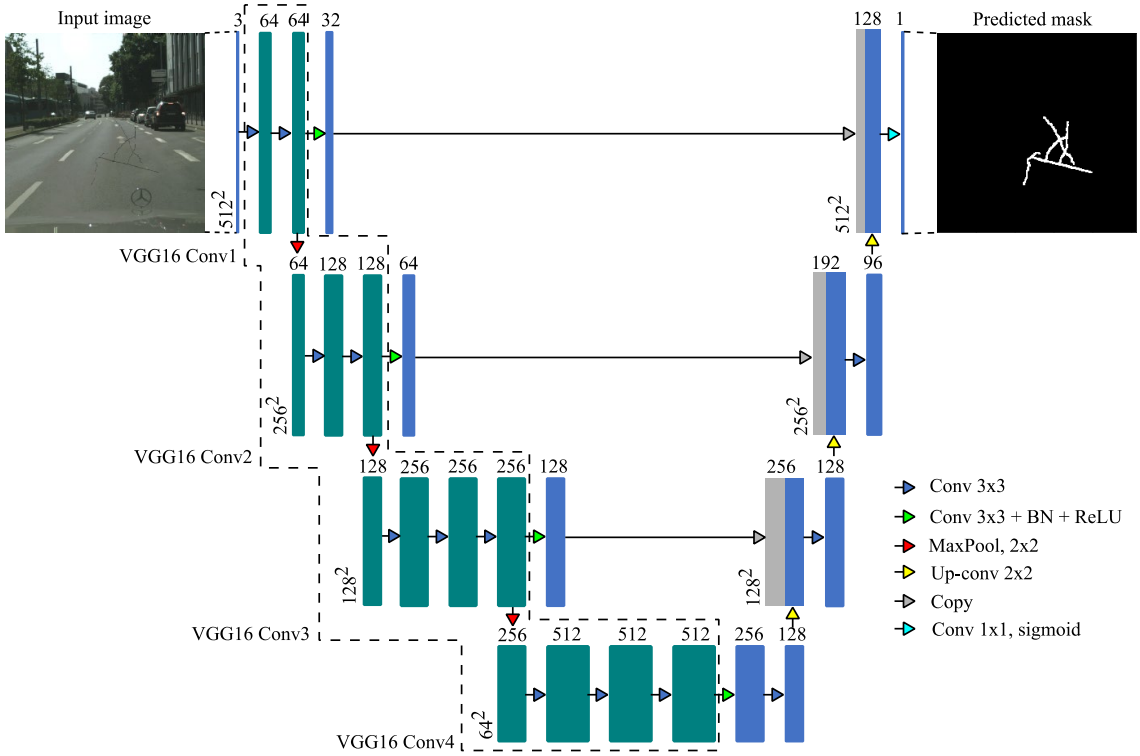


Figure 6. VGG16 + U-Net architecture.

5. Experiments

The obtained synthetic set of images with per-pixel labeled road cracks was used to train modern convolutional neural networks Mask R-CNN and U-Net. For our research we used Abdulla et al. [21] Tensorflow + Keras implementation of Mask R-CNN with modifications to perform our dataset. ResNet101 with FPN forms backbone for feature maps construction. Due to the fact that the synthetic dataset cannot contains all characteristics of real asphalt cracks, the transfer learning technology was used in combination with pre-trained model ResNet101 on the MS-COCO dataset. Transfer learning allows accelerating the training process and improving the performance of a model.

We use RGB images of size 1024×1024 as input of the Mask R-CNN. In addition, the following values are used as anchor scales: 0.33, 0.5, 1, 2, 3. The training took out 40 epochs of 400 iterations using mini-masks with size 56 × 56 pixels to optimize the used computer memory. We get the best result using learning rate annealing by starting with 0.001 and decreasing it by a factor of 10 every 10 epochs. Removing detection results of the same class happen if there is more than 0.7 value of IoU among bounding boxes.

VGG16 + U-Net realization in input uses RGB image with 512×512 resolution and output prediction mask of the same size. Pre-trained VGG16 on ImageNet is used as the encoder. Images from the synthetic dataset resized to 472×472 size and padded to the input size. This operation allows to avoid losing border pixel in the input due to convolution sequence. In these experiments, we use Adam optimizer with a batch size of 8, momentum of 0.9 and a learning rate of 0.001 with decay of 0.000001 and trained the models for 17 epochs.

All the experiments have been conducted on a workstation with a NVIDIA Tesla K80 GPU graphics card machine with 13 GB DDR5X memory on Google Colaboratory platform.

6. Results & evaluation

Instance segmentation using Mask R-CNN often is evaluated in PASCAL VOC style. The best Mask R-CNN's result estimated by the average precision (AP) metric with the value IoU = 0.5 on the validation synthetic subset is 78.1%.

Jaccard index (Intersection over Union) is chosen as evaluation metric for segmentation task. It reflects similarity measure between a finite numbers of sets. The measure between two sets A and B is defined as following:

$$IoU(A,B) = \frac{|A \cap B|}{|A \cup B|} = \frac{|A \cap B|}{|A| + |B| - |A \cap B|} \tag{4}$$

Pixel accuracy metric not used in the evaluation stage, because it can provide misleading results for the reason of small crack class representation within the image in contrast of negative case.

To evaluate the relevance of the training on our synthetic dataset we constructed a small dataset of 67 real images with cracks on carriageways that were manually labeled. These real images were obtained by digital camera mounted at a roof of a vehicle. The collected real-image dataset consists of consecutive frames of two street and has different cracks, potholes, shadows, road marking, manholes, road facilities and equipment. The presence of these elements on the images leads to the conclusion that the real-image dataset is representative.

Finally, the IoU score for segmentation results on synthetic and real image dataset was calculated and summarized in table 1.

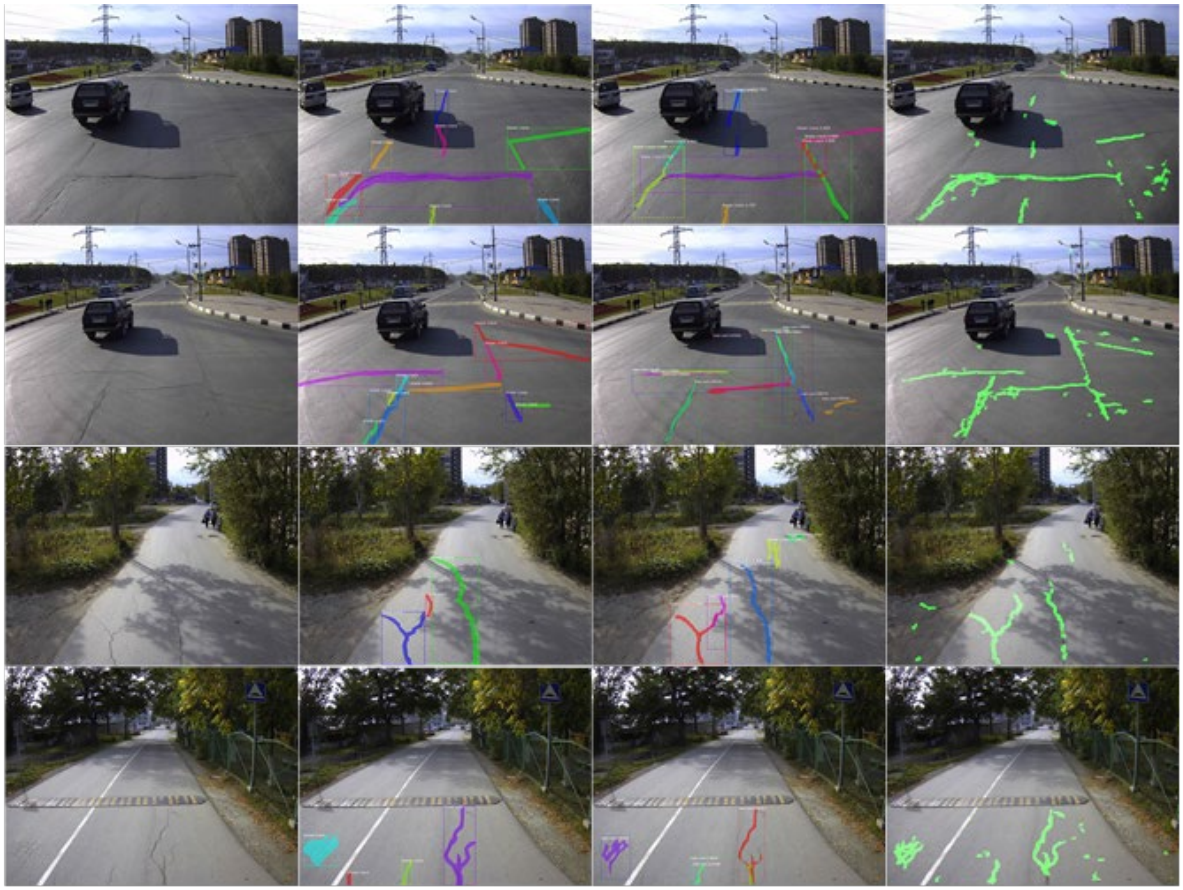
Table 1. Intersection over Union evaluation

Model	Dataset		
	Synthetic		Real-image
	training	validation	
Mask R-CNN	0.79	0.59	0.46
U-Net	0.81	0.56	0.47

Figure 7 shows four examples from real-image dataset and corresponding results of manual human annotation, Mask R-CNN detection and VGG16+U-Net segmentation.

7. Conclusions

The automatization of road condition control is current trend for practical application of computer vision methods. The conducted analytical review of road distress detection problem has shown that crack detection in pixel level in the driver-view image is a priority and challenging task. We have proposed a novel deep learning approach based on synthetic training data generation for segmentation of cracks in the images with road pavement. The synthetic data generation algorithm allows to easily obtain crack training dataset of any size for instance segmentation task. It makes possible to use state-of-the-art Mask R-CNN-based and U-Net-based segmentation model. The models, trained on synthetic crack data, give acceptable outcomes over 47% of IoU metrics on real images with surfaces' cracks. It should be noted that the proposed approach is weak sensitive to shadows, road marking and light condition. Improving the detection results is possible with additional training of models with attention to road technical facilities, such as manholes and restoration patches.



a) b) c) d)

Figure 7. Segmentation results: (a) examples of pavement surface crack from real-image dataset, (b) their groundtruth labels, (c) instance-segmentation result by Mask R-CNN, (d) pixel-wise segmentation results by U-Net.

Acknowledgment
This research was supported by Tomsk Polytechnic University Competitiveness Enhancement Program and was funded by RFBR according to the research project № 18-08-00977 A.

References

[1] Cordts M, Omran M, Ramos S, Rehfeld T, Enzweiler M, Benenson R, Franke U, Roth S and Schiele B 2016 *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)* 3213–3223
[2] Zendel O, Honauer K, Murschitz M, Steininger D and Dom´ınguez G F 2018 *Proceeding of Computer Vision – ECCV 2018* 407–421
[3] Fritsch J, K¨uhnl T and Geiger A 2013 *Proceeding of 16th Int. IEEE Conf. on Intelligent Transportation Systems (ITSC 2013)* 1693–1700
[4] Mohan A and Poobal S 2018 *Alexandria Engineering J.* **57** 787–798
[5] Coenen T B and Golroo A 2017 *Cogent Engineering* **4** 1–23
[6] Eisenbach M, Stricker R, Seichter D, Amende K, Debes K, Sesselmann M, Ebersbach D, Stoeckert U and Gross H 2017 *Proceeding of Int. Joint Conf. on Neural Networks (IJCNN)* 2039–2047
[7] Oliveira H and Correia P L 2014 *Proceeding of IEEE Int. Conf. on Image Processing (ICIP)* 798–802
[8] Zhang L, Yang F, Zhang Y D and Zhu Y J 2016 *Proceeding of IEEE Int. Conf. on Image*

Processing (ICIP) 3708–3712
[9] Li H, Song D, Liu Y and Li B 2019 *IEEE Transactions on Intelligent Transportation Systems* **20** 2025–2036
[10] Salman M, Mathavan S, Kamal K and Rahman M 2013 *Proceeding of 16th Int. IEEE Conf. on Intelligent Transportation Systems (ITSC 2013)* 2039–2044
[11] Varadharajan S, Jose S, Sharma K, Wander L and Mertz C 2014 *IEEE Winter Conf. on Applications of Computer Vision* 115–22
[12] Liu Y, Yao J, Lu X, Xie R and Li L 2019 *Neurocomputing* **338** 139–153
[13] Gopalakrishnan K, Khaitan S K, Choudhary A and Agrawal A 2017 *Construction and Building Materials* **157** 322–330
[14] Yang F, Zhang L, Yu S, Prokhorov D, Mei X and Ling H 2020 *IEEE Transactions on Intelligent Transportation Systems* **21** 1525–1535
[15] Shi Y, Cui L, Qi Z, Meng F and Chen Z 2016 *IEEE Transactions on Intelligent Transportation Systems* **17** 3434–4345
[16] Maeda H, Sekimoto Y, Seto T, Kashiyama T and Omata H 2018 *Computer-Aided Civil and Infrastructure Engineering* **33** 1127–1241
[17] Everingham M, Eslami S M, Van Gool L 2015 *Int J Comput* **111** 98–136
[18] He K, Gkioxari G, Doll´ar P and Girshick R 2017 *Proceeding of IEEE International Conf. on Computer Vision (ICCV)* 2980–2988
[19] Ronneberger O, Fischer P and Brox T 2015 *LNCS* **935**1234–1241
[20] Iglovikov V and Shvets A 2018 TernausNet: U-Net with VGG11 Encoder Pre-Trained on ImageNet for Image Segmentation *Preprint* 1801.05746
[21] Abdulla W 2017 Mask r-cnn for object detection and instance segmentation on keras and tensorflow. Available at: https://github.com/matterport/MaskRCNN