ISSN 1846-6168 (Print), ISSN 1848-5588 (Online)
https://doi.org/10.31803/tg-20220329114254

Original scientific paper
Received: 2022-03-29, Accepted: 2022-04-11

Automated Semantic Segmentation for Autonomous Railway Vehicles

Oğuzhan Katar*, Erkan Duman

Abstract: With the development of computer vision methods, the number of areas where autonomous systems are used has also increased. Among these areas is the transportation sector. Autonomous systems in the transportation sector are mostly developed for road vehicles, but highway rules and standards different between countries. In this study, models capable of semantic segmentation have been developed for autonomous railway vehicles with the help of the public dataset. Four different U-Net models were trained with 8500 images for four different scenarios. The model trained for binary semantic segmentation reached mean Intersection over Union (mIoU) value of 89.1%, while the models trained for multi-class semantic segmentation reached 83.2% mIoU, 79.7% mIoU and 29.6% mIoU. Information about the inclusion of high-resolution images in model training and performance metrics in semantic segmentation studies shared.

Keywords: autonomous systems; deep learning; railway vehicles; semantic segmentation; U-Net

1 INTRODUCTION

As a result of the increase in the processing capacity of computer hardware, there have been significant developments in artificial intelligence and computer vision [1]. Thanks to these developments, autonomous systems have become widespread. The main purpose of autonomous systems is to minimize the human factor. Due to the advanced decision-making mechanisms, autonomous systems do not need any direction during movement [2]. One of the most basic benefits of autonomous technologies is the opportunity to save time and energy in the area where it is used. One of the areas where time and energy saving is most needed is the transportation sector. For this reason, the use of autonomous systems in the field of transportation is inevitable [3].

Transportation is one of the important components of the world economy [4]. Because in order for the production to take place, first the raw material must be transported and then the product must be shipped after the production is completed [5]. It is desired that all these processes take place in a low cost and reliable way. Therefore, rail transport is preferred. Railway transportation is used not only for freight transport but also for passenger transfer. With all these features, railway transportation has an important potential for autonomous systems, but most studies are carried out by researchers for highways. Among the reasons for this is that the scope of public datasets is limited to highways.

The most vital function of autonomous systems in transportation is the ability to vision. With the help of a camera, the ability to vision by using pixel-based segmentation in the processing of the image can be brought to the system. In the segmentation process, instead of using classical methods, deep learning-based approaches should be used [6]. In this way, the system will ensure that the image is interpreted and the relevant action is taken in return, without the need for any human contribution. With such approaches, undesirable events that occur due to human factors such as inattention, fatigue, and thoughtlessness can be prevented.

The main purpose of this study is to provide vision capability to autonomous railway vehicles with deep learning-based segmentation method. With its implementation in the field, it is aimed to increase efficiency and safety in railway transportation. The rest of this paper is organized as follows. Section 2 includes other studies in the literature on deep learning techniques for image segmentation. The details of the model, dataset and performance metrics is described in Section 3. The analysis of test results and experimental results of U-Net model are given in Section 4. Conclusion part of the study is in Section 5.

2 RELATED WORKS

In this section, some studies from computer vision in railway are examined. The main idea behind these studies is to contribute to railway transportation with various methods.

Edge detection methods are frequently used in computer vision studies for railway transportation due to the texture and shape of the rail-tracks [7] [8]. In addition to the edges extracted with edge detection, the Hough transform is applied [9]. In edge detection methods used with classical image processing, there may be factors such as the angle of light that will affect the removal of the relevant part and the determination of its area. In order to overcome such problems, deep learning-based approaches are used.

In [10], a model capable of deep learning-based railroad track segmentation has been proposed. This model, called RailNet, consists of a feature extraction network and a segmentation network. In order to train the specified model, a private dataset consisting of 3000 images was created by the researchers. This model, which was designed only for binary segmentation, reached a mIoU value of 89.8% as a result of the tests performed.

In [11], rail track regions are determined from the images taken with the help of unmanned aerial vehicles. The U-Net based model is trained for the binary segmentation task. The model, called Rail-UNet, reached 95.9% mIoU.

3 MATERIAL AND METHOD

3.1 Dataset

In this study, the dataset RailSem19 [12] was used. Most of the studies on gaining autonomous vision features of transportation vehicles cover road vehicles. A general reason for this is that the content of the datasets that researchers can use consists of only highway images. In order to overcome this deficiency in railway transportation studies, RailSem19 was made publicly available to researchers by the Austrian Institute of Technology employees in 2019.

RailSem19 dataset consists of 8500 unique high-resolution images taken from the machinist point of view of a railway vehicle and mask images prepared to be used in segmentation studies of these images. Samples of images and masks in the dataset are given in Fig. 1.

Figure 1 Samples of images and masks in RailSem19

There are 19 different labels in the mask images in the RailSem19 dataset. Each label represents a class, so class distribution calculations in the dataset can be made. These labels are represented by various pixel values in 8-bit and single-channel '.png' images. Pixel values, labels and percentage of frames containing pixels with the label information are given in Tab. 1.

Table 1 Dataset pixel details

Label Name	Pixel Value	In Frames (%)
road	0	48.1
sidewalk	1	57.2
construction	2	72.1
tram-track	3	51.7
fence	4	48.9
pole	5	60.0
traffic-light	6	37.1
traffic-sign	7	32.3
vegetation	8	83.3
terrain	9	61.2
sky	10	94.5
human	11	6.0
rail-track	12	86.2
car	13	13.8
truck	14	4.6
trackbed	15	87.6
on-rails	16	15.4
rail-raised	17	87.2
rail-embedded	18	14.6

3.2 Pre-processing

Before defining and training the model, various pre-processes were applied to the RailSem-19 dataset samples for this study. The common objectives of the pre-processes applied are to use our resources more efficiently, to increase the training accuracy and time efficiency.

3.2.1 Creating Sub-datasets

Firstly, the dataset consisting of 8500 images and 8500 mask images was divided into four sub-datasets by preserving the image and mask unity. While performing the division process, it is important to determine the appropriate visuals for the scenarios they will represent, rather than separating them in equal numbers. The sub-datasets are named with the 'SubDS' tag and the detail of the sub-datasets is shown in Tab. 2.

Table 2 Sub-datasets details

Dataset Name	Image Count	Mask Count	Resolution
SubDS1	1870	1870	1920×1080 px
SubDS2	2850	2850	1920×1080 px
SubDS3	2250	2250	1920×1080 px
SubDS4	1530	1530	1920×1080 px

No random selection was made while creating the sub-datasets. This is because each dataset represents a scenario. Samples are separated according to the labels and distribution of pixel values that each scenario needs. Detailed information about the scenarios is as follows.

SubDS1 represents the scenario in which railway vehicles are able to segment only rail-tracks, and the dataset is designed for input to the binary segmentation model. In the relevant dataset, the pixels with the rail-track label in the mask images are given a value of '1'. Pixel values for all other labels are assigned '0'. SubDS1 sample and its mask image are given in Fig. 2.

(a) Original Image (b) Original Mask (c) SubDS1 Mask

Figure 2 SubDS1 dataset sample

SubDS2 is designed for the segmentation of rail-raised areas in addition to rail-track during the movement of railway vehicles. It is an example of multi-class segmentation scenarios that contains the least number of class labels. Because it consists of three labels in total. In this dataset, as in binary segmentation pixels with rail-track label are given the value '1', pixels with rail-raised label are given the value '2' and all other labels are given the value '0'. SubDS2 sample and its mask image are given in Fig. 3.

(a) Original Image (b) Original Mask (c) SubDS2 Mask

Figure 3 SubDS2 dataset sample

SubDS3 dataset is designed for scenarios where it is necessary to detect whether there is a person on the rail-track or rail-raised in autonomous driving. There are four different pixel values and labels in the dataset. As in SubDS2, pixels with rail-track label are represented by a value of '1', pixels with a rail-raised label are represented by a value of '2'. In addition, the pixels in which people are located are assigned the value of '3' and the value of the ones other than the specified labels is '0'. SubDS3 sample and its mask image are given in Fig. 4.

(a) Original Image (b) Original Mask (c) SubDS3 Mask

Figure 4 SubDS3 dataset sample

SubDS4 dataset has the default labels and pixel values of the RailSem-19 dataset. It is the most challenging multi-class segmentation example among the scenarios designed. SubDS4 sample and its mask image are given in Fig. 5.

(a) Original Image (b) Original Mask (c) SubDS4 Mask

Figure 5 SubDS4 dataset sample

3.2.2 Data Splitting

Considering the unity of the images and mask images in the datasets, they were randomly divided into 70% train, 20% validation and 10% test. Data splitting is shown in Fig. 6.

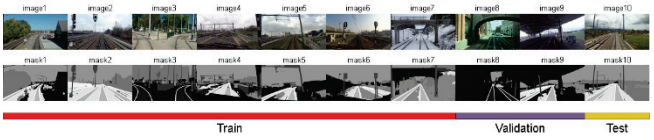

Figure 6 Data splitting method

3.2.3 Image Cropping

As it is known, the size of the dataset samples is 1920×1080 px. It is not a logical approach to input such high-resolution images directly into the model, considering our hardware features. Since the input size of the default U-Net model is 256×256, resizing our images to this value may be a solution, but resize is not recommended in segmentation studies [13]. The details of the algorithm we applied to solve the related problem are as follows.

First, the reference image to be cropped is determined. 1920×1080 px image is cropped in a square format, starting from the first pixel, with edges on the *x*-axis of 256 pixels and the *y*-axis of 256 pixels. This process is repeated until it covers the size values of the image. A total of 40 256×256 cropped images can be created for 1920×1080 px images. In Fig. 7, the cropping process and numbering of rows and columns are given.

Figure 7 Image cropping method

When the first step is completed, the squares shown in white in Fig. 7. are smoothly cropped sub-images. The squares indicated in red there are pixels that cannot complete the 256×256 size. In order not to lose the information in these pixels, the values (0, 0, 0) were added with the padding methods. The result of the padding method is given in Fig. 8.

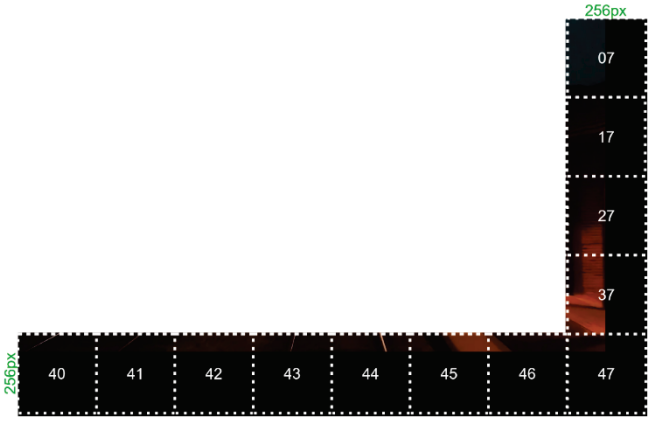

Figure 8 Result of padding method

The same steps applied for the images were applied for the mask images, and a sample mask image cropping output is given in Fig. 9.

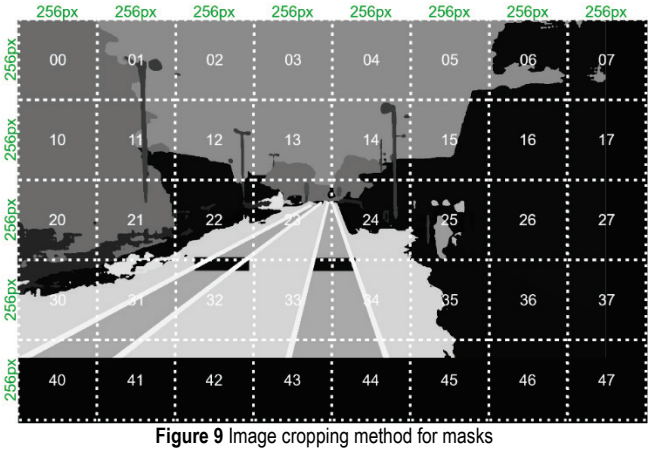

Figure 9 Image cropping method for masks

3.3 U-Net Model

U-Net is a kind of artificial neural network that contains a series of convolutional layers and non-convolutional layers to perform the image segmentation task. It is one of the most popularly used approaches to any semantic partitioning task today. U-Net gets its name from its U-like architecture, as seen in Fig. 10.

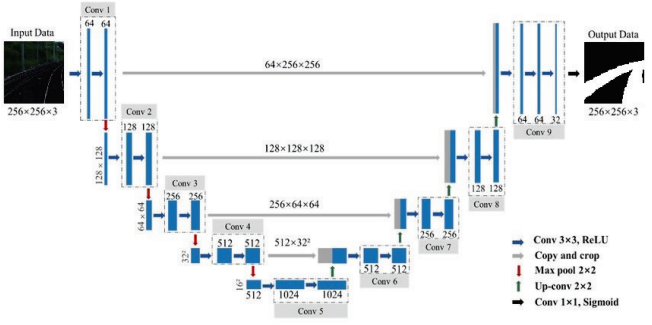

Figure 10 U-Net architecture

U-Net is an encoder-decoder network architecture consisting of four encoder blocks and four decoder blocks connected via bridge. The encoder is designed to extract spatial features from the original image.

3.3.1 U-Net Encoder Network

The encoder network acts as a feature extractor and learns an abstract representation of the input image through a sequence of encoder blocks. Each encoder block consists of two 3×3 convolutions where each convolution is followed by a Rectified Linear Unit (ReLU) activation function. ReLU introduces non-linearity to the network, which helps to better generalize the training data. Then comes the 2×2 maximum pooling, in which the height and width of the feature maps are reduced by half. This reduces the calculation cost by reducing the number of trainable parameters.

3.3.2 U-Net Decoder Network

Decoder network is used to take the abstract representation and create a semantic segmentation mask. The decoder block starts with a 2×2 transposed convolution. It is then combined with the corresponding feature map from the encoder block. These links provide features from previous layers that are sometimes lost due to the depth of the network. Two 3×3 convolutions are then used, where each convolution is followed by a ReLU activation function. At the output of the final decoder, sigmoid is used for binary segmentation scenarios, while softmax activation function is used for multi-class segmentation scenarios.

3.4 Model Training

Model training was carried out with the help of the workstation computer, the details of which are given in Tab. 3.

Table 3 Computer specifications

CPU	GPU	RAM	Operating System
İntel Xeon E5-1603	Quadro P1000	16 GB	Windows 10

Python was chosen as the programming language. The Keras library was used for the implementations of the models. Epoch number is 500, batch size is 16, learning rate is 0.001, Adam is used as optimizer function.

3.5 Performance Metrics

In this study, pixel accuracy and jaccard index evaluation criteria were used to measure the prediction success of the segmentation model (code available at https://github.com/OguzhanKATAR23). Predictions in machine learning studies are examined under four categories: true positives (TP), false positives (FP), true negatives (TN), and false negatives (FN). These categories are also valid for segmentation studies, but they are more difficult to distinguish compared to standard classification studies.

3.5.1 Pixel Accuracy

Pixel Accuracy (PA) is the ratio of the number of pixels known to be correctly classified because of the pixel-based comparison of the ground truth mask and the predicted mask to the total number of pixels. It is calculated according to the mathematical equation specified in Eq. (1).

$$Pixel\ Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

Pixel Accuracy seems useful due to its easy computation and complexity of performance metrics, but it can produce deceptive results. High resolution images may cause deceptive results as stated. For example, an image with 1920×1080 size has a total of 2073600 pixels. Assuming 90% of pixels with background value in ground truth masks, the remaining 10% for other labels. Such cases mean that the class pixel distribution is extremely uneven. Even if the model fails to predict the mask and only returns an image with pixel values of zeros, the pixel accuracy value will be 90% due to the large true negatives value. However, the model could not produce results for the areas to be segmented in this scenario. Therefore, pixel accuracy evaluation should be used for images with balanced distribution instead of images with uneven class distribution. The extraction of false values, which are important in the calculation of the pixel accuracy metric is shown in Fig. 11.

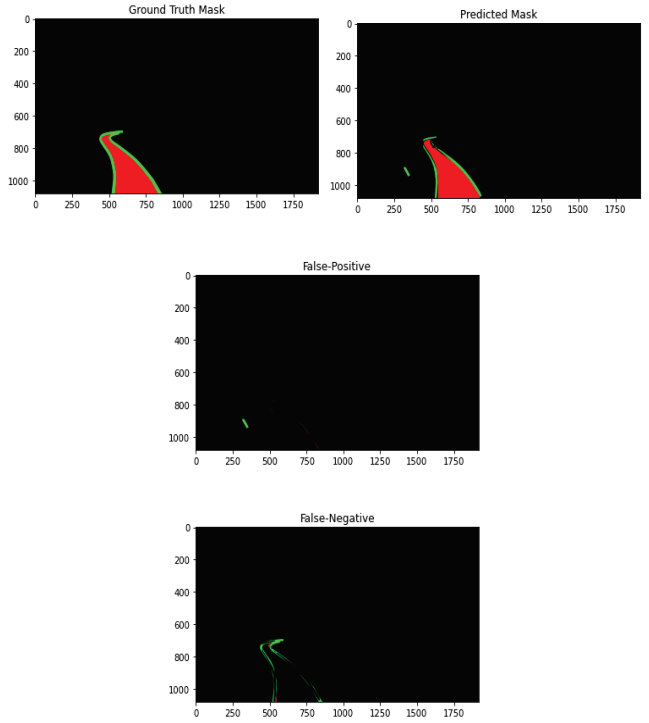

Figure 11 False values in pixel accuracy calculation

Various additional evaluation criteria can be calculated using the four categories required to reveal the pixel accuracy value. These evaluation criteria and mathematical Eq. (2), Eq. (3) and Eq. (4) are as follows.

$$Precision = \frac{TP}{TP + FP} \tag{2}$$

$$Recall = \frac{TP}{TP + FN} \tag{3}$$

$$Specificity = \frac{TN}{TN + FP} \tag{4}$$

3.5.2 Jaccard Index

The jaccard index, which indicates the similarity between the two given images, is also called Intersection over Union (IoU). IoU value is calculated by the mathematical equations given in Eq. (5) at what rate the estimated mask pixels overlap with the original mask.

$$IoU(A, B) = \frac{A \cap B}{A \cup B} \tag{5}$$

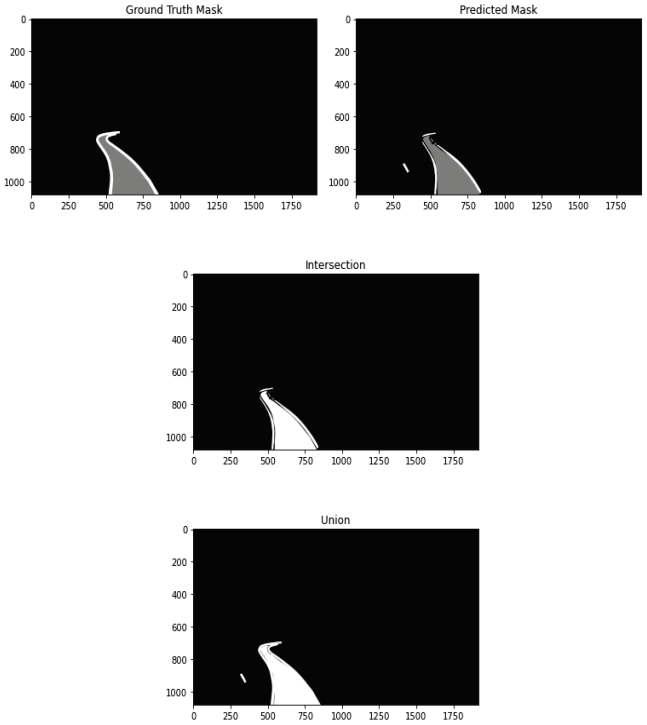

Figure 12 Intersection and Union values in IoU calculation

It gives more reliable results than pixel accuracy. Therefore, it is one of the most important performance metrics considered in segmentation studies. Examples of intersection and union required for calculating the IoU value are given in Fig. 12.

In order to compare the pixel accuracy and jaccard index methods more clearly, the same ground truth mask and predicted mask were used in both methods. In Tab. 4, the numerical values of the related methods are given in detail.

Table 4 Pixel accuracy vs jaccard index

Metric	Score	Equation with pixel counts
PA	99.6%	(67227+1998187) / (67227+1998187+1361+6825)
IoU	87.4%	65576 / 75002

4 EXPERIMENTAL RESULTS

In this study, we developed a fully automatic method to provide vision ability to railway vehicles. Our U-Net models were trained for 500 epochs with the help of the four sub-datasets. The mean Intersection over Union (mIoU) values obtained as a results of the relevant trainings are given in Tab. 5.

Table 5 Results of the trainings

Dataset Name	Model Name	Number of Class	mIoU (%)
SubDS1	U-Net	2	89.1
SubDS2	U-Net	3	83.2
SubDS3	U-Net	4	79.7
SubDS4	U-Net	19	29.6

The loss graph of the model, which was designed to detect the rail trail of autonomous railway vehicles with the multi-class segmentation method and trained with the SubDS2 dataset, is shown in Fig. 13.

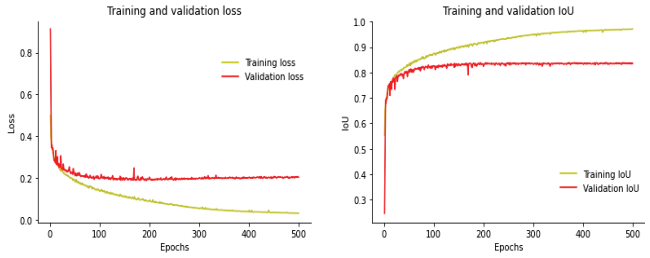

Figure 13 Loss and IoU graphics

In the SubDS1 dataset, the mask was estimated with an *IoU* value of 95.2% for the randomly selected image among the samples reserved for testing. Intersection pixel count is 153935, union pixel count is 161549. The prediction mask is given in Fig. 14.

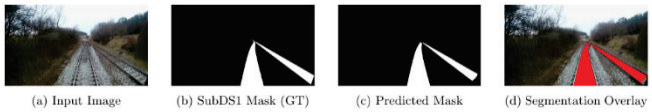

Figure 14 Prediction of the model trained with SubDS1

In the SubDS2 dataset, the mask was estimated with an *IoU* value of 91.9% for the randomly selected image among the samples reserved for testing. Intersection pixel count is 141705, union pixel count is 154038. The prediction mask is given in Fig. 15.

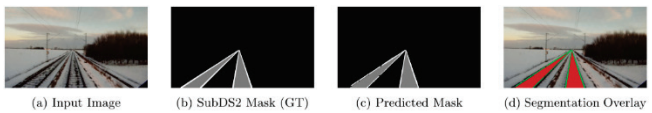

Figure 15 Prediction of the model trained with SubDS2

In the SubDS3 dataset, the mask was estimated with an *IoU* value of 83.6% for the randomly selected image among the samples reserved for testing. Intersection pixel count is 222384, union pixel count is 265786. The prediction mask is given in Fig. 16.

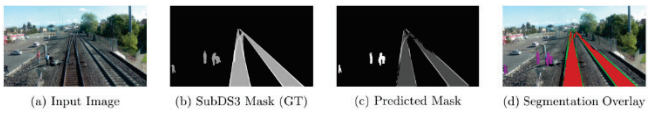

Figure 16 Prediction of the model trained with SubDS3

In the SubDS4 dataset, the mask was estimated with an *IoU* value of 45.4% for the randomly selected image among the samples reserved for testing. Intersection pixel count is 941415, union pixel count is 2073600. The prediction mask is given in Fig. 17.

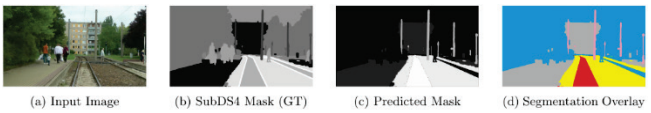

Figure 17 Prediction of the model trained with SubDS4

5 CONCLUSION

The increase in the usage areas of autonomous systems has also led to the realization of various studies in the transportation sector. In this study, an artificial intelligence-based system was proposed to improve the vision features of autonomous driving vehicles in railway transportation. A public railway dataset was customized for this study. It was aimed to reach high IoU values by training the U-Net models with the specified dataset. In order to increase the prediction accuracy values of our models, which are capable of multi-class segmentation and binary segmentation, the number of samples used in training and the scope of the samples should be increased. For autonomous railway vehicles that will perform vital tasks such as passenger transport, the margin of error should be close zero. In addition to the operations that need to be done on the datasets to reduce this margin of error it is necessary to develop deep learning-based models used. The proposed system can be easily integrated into the technology used in the field of railway transportation today. Designing such artificial intelligence-based studies and using them in the field will create the future of the transportation sector. Revealing similar studies can help researchers working on rail systems and smart transportation.

Notice

The paper was presented at the International Congress of Electrical and Computer Engineering (ICECENG'22), which took place in Bandırma (Turkey), on February 9-12, 2022. The paper will not be published anywhere else.

6 REFERENCES

[1] Hager, G. D., Bryant, R., Horvitz, E., Mataric, M., & Honavar, V. (2017). Advances in artificial intelligence require progress across all of computer science. arXiv preprint arXiv:1707.04352.

[2] Chen, H., Wen, Y., Zhu, M., Huang, Y., Xiao, C., Wei, T., & Hahn, A. (2021). From automation system to autonomous system: An architecture perspective. *Journal of Marine Science and Engineering, 9*(6), 645. https://doi.org/10.3390/jmse9060645

[3] Hutchins, N. & Hook, L. (2017). Technology acceptance model for safety critical autonomous transportation systems. *In 2017 IEEE/AIAA 36th Digital Avionics Systems Conference (DASC),* 1-5. https://doi.org/10.1109/DASC.2017.8102010
[4] Banister, D. (Ed.). (1995). *Transport and urban development*. Taylor & Francis.
[5] Lababidi, H. M., Ahmed, M. A., Alatiqi, I. M., & Al-Enzi, A. F. (2004). Optimizing the supply chain of a petrochemical company under uncertain operating and economic conditions. *Industrial & Engineering Chemistry Research, 43*(1), 63-73. https://doi.org/10.1021/ie030555d
[6] Tabernik, D., Šela, S., Skvarč, J., & Skočaj, D. (2020). Segmentation-based deep-learning approach for surface-defect detection. *Journal of Intelligent Manufacturing, 31*(3), 759-776. https://doi.org/10.1007/s10845-019-01476-x
[7] Gschwandtner, M., Pree, W., & Uhl, A. (2010). Track detection for autonomous trains. *In International Symposium on Visual Computing,* Springer, 19-28. https://doi.org/10.1007/978-3-642-17277-9_3
[8] Singh, A. K., Swarup, A., Agarwal, A., & Singh, D. (2019). Vision based rail track extraction and monitoring through drone imagery. *Ict Express, 5*(4), 250-255. https://doi.org/10.1016/j.icte.2017.11.010
[9] Aly, M. (2008). Real time detection of lane markers in urban streets. *In 2008 IEEE Intelligent Vehicles Symposium*, 7-12. https://doi.org/10.1109/IVS.2008.4621152
[10] Wang, Y., Wang, L., Hu, Y. H., & Qiu, J. (2019). RailNet: a segmentation network for railroad detection. *IEEE Access, 7,* 143772-143779. https://doi.org/10.1109/ACCESS.2019.2945633
[11] Mammeri, A., Siddiqui, A. J., & Zhao, Y. (2021). UAV-assisted Railway Track Segmentation based on Convolutional Neural Networks. *In 2021 IEEE 93rd Vehicular Technology Conference (VTC2021-Spring),* 1-7. https://doi.org/10.1109/VTC2021-Spring51267.2021.9448887
[12] Zendel, O., Murschitz, M., Zeilinger, M., Steininger, D., Abbasi, S., & Beleznai, C. (2019). Railsem19: A dataset for semantic rail scene understanding. *In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*. https://doi.org/10.1109/CVPRW.2019.00161
[13] Liu, Y., Ren, Q., Geng, J., Ding, M., & Li, J. (2018). Efficient patch-wise semantic segmentation for large-scale remote sensing images. *Sensors, 18*(10), 3232. https://doi.org/10.3390/s18103232

Authors' contacts:

Oğuzhan Katar, Research Assistant
(Corresponding author)
Firat University,
Department of Software Engineering,
23119 Elazig/Turkey
okatar@firat.edu.tr

Erkan Duman, Assistant Prof. Dr
Firat University,
Department of Computer Engineering,
23119 Elazig/Turkey
erkanduman@firat.edu.tr