

UDC 004.932

DOI: 10.15587/1729-4061.2021.248390

IMPROVING A NEURAL NETWORK MODEL FOR SEMANTIC SEGMENTATION OF IMAGES OF MONITORED OBJECTS IN AERIAL PHOTOGRAPHS

Vadym Slyusar

Doctor of Technical Sciences, Professor
Research Institute Group*

Mykhailo Protsenko

PhD, Senior Researcher
Office of Special Forces*

Anton Chernukha

Corresponding author
PhD

Department of Service and Training
National University of Civil Defence of Ukraine
Chernyshevska str., 94, Kharkiv, Ukraine, 61023
E-mail: an_cher@nuczu.edu.ua

Vasyl Melkin

PhD
Organizational and Scientific Division*

Olena Petrova

PhD, Associate Professor
Department of Technology of Processing, Standardization and Certification of Livestock Products
Mykolayiv National Agrarian University
Heorhiya Honhadze str., 9, Mykolayiv, Ukraine, 54020

Mikhail Kravtsov

PhD, Associate Professor
Department of Metrology and Industrial Safety
Kharkiv National Automobile and Highway University
Yaroslava Mudroho str., 25, Kharkiv, Ukraine, 61002

Svitlana Velma

Department of Educational and Information Technologies
National University of Pharmacy
Pushkinska str., 53, Kharkiv, Ukraine, 61002

Nataliia Kosenko

PhD
Department of Project Management in Urban Economy and Construction
O.M. Beketov National University of Urban Economy in Kharkiv
Marshala Bazhanova str., 17, Kharkiv, Ukraine, 61002

Olga Sydorenko

PhD **

Maksym Sobol

PhD**

*Central Scientific Research Institute of Armament and Military Equipment of the Armed Forces of Ukraine
Povitroflotsky ave., 28, Kyiv, Ukraine, 03049

**Department of Computer Science and Intellectual Property
National Technical University "Kharkiv Polytechnic Institute"
Kyrpychova str., 2, Kharkiv, Ukraine, 61002

This paper considers a model of the neural network for semantically segmenting the images of monitored objects on aerial photographs. Unmanned aerial vehicles monitor objects by analyzing (processing) aerial photographs and video streams. The results of aerial photography are processed by the operator in a manual mode; however, there are objective difficulties associated with the operator's handling a large number of aerial photographs, which is why it is advisable to automate this process. Analysis of the models showed that to perform the task of semantic segmentation of images of monitored objects on aerial photographs, the U-Net model (Germany), which is a convolutional neural network, is most suitable as a basic model. This model has been improved by using a wavelet layer and the optimal values of the model training parameters: speed (step) – 0.001, the number of epochs – 60, the optimization algorithm – Adam. The training was conducted by a set of segmented images acquired from aerial photographs (with a resolution of 6,000×4,000 pixels) by the Image Labeler software in the mathematical programming environment MATLAB R2020b (USA). As a result, a new model for semantically segmenting the images of monitored objects on aerial photographs with the proposed name U-NetWavelet was built.

The effectiveness of the improved model was investigated using an example of processing 80 aerial photographs. The accuracy, sensitivity, and segmentation error were selected as the main indicators of the model's efficiency. The use of a modified wavelet layer has made it possible to adapt the size of an aerial photograph to the parameters of the input layer of the neural network, to improve the efficiency of image segmentation in aerial photographs; the application of a convolutional neural network has allowed this process to be automatic

Keywords: semantic segmentation of images, convolutional neural network, aerial photograph, unmanned aerial vehicle

Received date 18.10.2021

How to Cite: Slyusar, V., Protsenko, M., Chernukha, A., Melkin, V., Petrova, O., Kravtsov, M., Velma, S., Kosenko, N., Sydorenko, O.,

Accepted date 03.12.2021

Sobol, M. (2021). Improving a neural network model for semantic segmentation of images of monitored objects in aerial photographs.

Published date 29.12.2021

Eastern-European Journal of Enterprise Technologies, 6 (2 (114)), 86–95. doi: <https://doi.org/10.15587/1729-4061.2021.248390>

1. Introduction

The use of unmanned aerial vehicles (UAVs) makes it possible to accelerate the process of monitoring critical

infrastructure objects [1]. Such facilities include industrial enterprises [2], energy plants [3], chemically hazardous industries [4], and other strategic objects [5]. Disruption of the functioning of these facilities can threaten the national inter-

ests of people's lives [6, 7]. With the help of UAVs, objects are monitored by processing (analyzing) aerial photographs and a video stream. One of the types of image processing is its segmentation. Segmentation of aerial photographs involves dividing it into areas according to certain criteria. The result of segmentation is a set of areas that cover the entire aerial image. Therefore, the development of new and improvement of existing neural network models for segmenting images of monitored objects (MOs) in aerial photography is of particular relevance.

2. Literature review and problem statement

Paper [8] shows that traffic surveillance using UAVs has gained great popularity in civilian applications and remote sensing tasks. Due to its high mobility and large field of vision, as well as the ability to cover large areas at different altitudes, UAVs have become a sought-after surveillance tool in recent years. The option of counting vehicles with the elimination of the problem of excessive counting of information in sequential frames of video from UAVs is proposed in the cited paper. However, it did not address issues related to the segmentation of MO images.

Study [9] proposed various models based on convolutional neural networks (CNNs) to collect information obtained using a segmentation network; a generative adversarial network based on Pixel2Pixel was suggested. The discriminator employed CNN to distinguish between the results of segmentation of the generated model and the Expert Advisor. The results showed that the network model could provide effective automatic segmentation of the hippocampus and is of practical importance for the correct diagnosis of diseases such as Alzheimer's disease. The disadvantage of that method is its high computational complexity, its lack of adaptability for the segmentation of MO images on aerial photographs.

In work [10], a fast clustering algorithm based on super pixels for segmentation of radar images with a synthesized aperture is proposed. Experimental results of two real images of the synthetic aperture radar show that the proposed method is superior to other modern methods both in terms of segmentation accuracy and in terms of computational efficiency. The disadvantage of that model is its lack of adaptability for the segmentation of MO images in aerial photographs.

Paper [11] shows that malware detection methods based on deep learning are generally highly accurate. However, when malware families with a high degree of similarity are detected, the detection accuracy is seriously compromised due to the lack of obvious training functions. To resolve that issue, the cited paper proposes a method for detecting malicious code that is based on image segmentation and deep CNN. A disadvantage of the model is its high computational complexity and maladjustment for the segmentation of MO images in aerial photographs.

Paper [12] proposes a multiscale model of semantic segmentation in real time. Experimentally, it has been shown that the proposed model could be used to solve many recognition problems, has a good decoding ability. Despite this, the issues of automating the process of segmenting MO images in aerial photographs were not considered.

Study [13] proposes a new classification scheme for hyperspectral images of remote sensing of the Earth. The proposed model is able to increase intraclass similarity by locally suppressing spectral variations, while promoting

cross-class variability on a global scale, resulting in recovery with more distinguishable pixels. Experimental results on three test datasets demonstrate a significant superiority of the proposed method over modern ones. The disadvantage of that model is its lack of adaptability for the segmentation of MO images in aerial photographs.

Paper [14] considers obtaining accurate multiscale semantic information from images for high-quality semantic segmentation. A model called cross fusion net (CF-Net) is proposed for fast and efficient extraction of multiscale semantic information. The model is able to encode more accurate semantic information from small-scale objects, and, accordingly, improve the accuracy of segmentation of small-scale objects. The disadvantage of the model is its computational complexity.

Our review of the literature [8–14] revealed the following shortcomings in the above models (methods):

- the computational complexity of the segmentation of MO images on aerial photographs obtained from UAVs;
- the lack of neural network models that solve the problem of segmenting MO images in aerial photographs.

All this suggests that it is advisable to conduct a study to improve a neural network model for the semantic segmentation of images of monitored objects in aerial photographs, which would significantly improve the accuracy and efficiency of the segmentation of MO images in aerial photographs.

3. The aim and objectives of the study

The aim of this study is to improve the neural network model for the segmentation of MO images in aerial photographs involving the choice of its learning parameters. This will make it possible to automate the process of analysis (processing) of aerial photographs.

To accomplish the aim, the following tasks have been set:

- to investigate the effectiveness of MO image segmentation using CNN;

- to evaluate the effectiveness of segmenting MO images on aerial photographs by the proposed U-NetWavelet model.

4. The study materials and methods

Suppose that a digital camera is installed on board a UAV. In this case, aerial photographs are transmitted through the communication channel to the computer of the ground control point. There, they are stored digitally as a file. Segmentation is important for the tasks of analyzing images of monitored objects in aerial photographs. Semantic segmentation describes the process of connecting each pixel in an image to a class label (color).

The mathematical statement of the problem of semantic segmentation of images is to assign each pixel of the MO image in the aerial photograph $S(x,y,z)$ to the label (color) of each pixel of class (object) B_i :








$$P[\|S\|] = B_i, \quad (1)$$

where P is the operator that characterizes the work of CNN.

In the proposed model, an RGB aerial photograph is fed to the CNN input; dimension, $6,000 \times 4,000 \times 3$; JPEG format; the output is the label (color) of each pixel of the class (object): Table 1.

Table 1

Label (color) of each pixel of the class (object)

Class	Class name	Label	The color of each pixel of the class (object)
1	Helicopter	Helicopter	
2	Airplane	Airplane	
3	Tank	Tank	
4	Vehicle tractor	Vehicle tractor	
5	Truck	Truck	
6	Car	Car	
7	Bus	Bus	

Our study of object recognition in aerial photographs was conducted using CNN methods in combination with the selection of optimal training parameters.

To automate the process of semantic segmentation of MO images in aerial photographs, it is proposed to use the U-Net model as a basic one, which demonstrated high efficiency in solving biomedical problems.

The architecture of CNN U-Net is discussed in [15, 16] and is shown in Fig. 1. CNN uses a weight matrix in convolution operations. The convolution layer sums up the results of the element-by-element product of each fragment of the image onto a matrix – the core of the convolution.

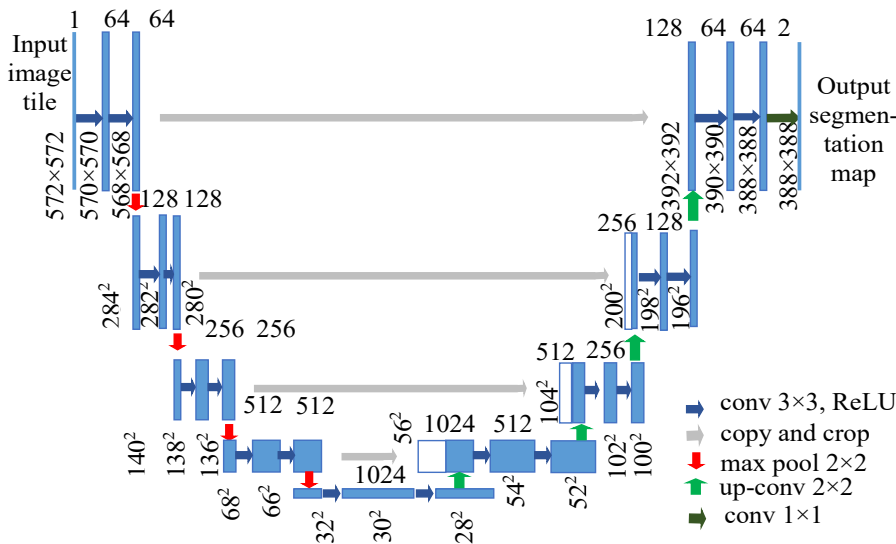


Fig. 1. U-Net architecture [15] (example for 32×32 pixels at the lowest resolution). Each blue field corresponds to a map of multichannel functions. The number of channels is indicated at the top of the window. The size of x, y is indicated in the lower left corner of the field. The white rectangles are copied function maps. Arrows indicate different operations

U-Net consists of a tapering path (left side) and an expanding path (right side). It consists of the use of two convolutions 3×3 (incomplete convolutions), each of which is followed by a positively linear ReLU function and a maximum unification (pooling) operation of 2×2 in steps 2 to lower the sampling. At each stage of down-sampling, the number of functional channels is doubled. Each step of

the extended path consists of an upscaling discretization of the feature map, followed by a convolution 2×2 (“convolution up”), which halves the number of feature channels. Each step of the tapering path consists of a downscaling of the feature map followed by a convolution of 3×3, each followed by a ReLU.

Cropping is necessary because of the loss of edge pixels with each convolution. On the last layer, a 1×1 convolution is used to map each 64-component feature vector to the desired number of classes. In total, there are 23 convolutional layers in the network.

Features of the ReLU activation function, its mathematical notation is described in detail in [17, 18]; the implementation of the operation of maximum unification (pooling) – in [17].

Training U-Net.

U-Net is trained in stochastic gradient descent based on input images and their corresponding segmentation maps. Because of convolutions, the output image is smaller than the input signal by a constant border width. Applied pixel-by-pixel, the Softmax function, which calculates energy from the final feature map along with the cross-entropy function. The Softmax function is defined in [15] as:

$$p_k(x) = \frac{\exp(a_k(x))}{\sum_{k=1}^{k=K} \exp(a_k(x))}, \quad (2)$$

where $p_k(x)$ is the value of the function approaching 1, when k has the maximum activation $a_k(x)$, which represents the activation channel of the function k pixel position ($x \in \Omega$) и ($\Omega \subset Z^2$); k denotes the number of classes.

The cross-entropy at each point shows the deviation and is defined in [15] as:

$$E = \sum_{x \in \Omega} w(x) \log(p_{\ell(x)}(x)), \quad (3)$$

where $\ell: \Omega \rightarrow \{1, \dots, K\}$ is the true label of each pixel;

$w: \Omega \rightarrow \mathbb{R}$ is the weight map, which is introduced to give some pixels a greater value during training.

The separation boundary is calculated using morphological operations. The calculation of the map of weighting coefficients is carried out according to the formula given in [15]:

$$w(x) = w_c(x) + w_0 \cdot \exp\left(-\frac{(d_1(x) + d_2(x))^2}{2\sigma^2}\right), \quad (4)$$

where $w_c: \Omega \rightarrow \mathbb{R}$ is the weight map for balancing class frequencies;

$d_1: \Omega \rightarrow \mathbb{R}$ is the distance to the border of the nearest cell;
 $d_2: \Omega \rightarrow \mathbb{R}$ is the distance to the border of the second closest cell;

$w_0=10$ and $\sigma=5$ pixels have been experimentally established [11].

Justification of the architecture and the mathematical apparatus used for the implementation proposed by CNN.

Our review of the literature [15, 16] showed that the U-Net model demonstrates high efficiency for the semantic segmentation of images of objects of different shapes and positions.

The advantages of U-Net and neural networks based on it are:

- high efficiency for solving the problems of segmentation of medical images [14, 15];
- information from large scales (upper layers) allows the model to be better at classification;
- information from smaller scales (deep layers) helps the model segment better;
- increasing dimensionality by increasing the number of feature channels allows the CNN to distribute contextual information to layers of greater resolution;
- symmetrical network strategy makes it possible to process large images (snapshots) such as aerial photographs, hyperspectral images, images for orthophoto plans;
- using a small number of images [15] for training and obtaining good accuracy.

To solve the problem of semantic segmentation of images of monitored objects on aerial photographs for 7

classes and improve the efficiency of segmentation, it is proposed to use a modified wavelet layer as an input layer, and CNNU-Net as a basic model. The training of the model was carried out by a set of images prepared from aerial photographs.

The architecture of CNN (Basic U-Net) is shown in Fig. 2. The task solved by CNN is the semantic segmentation of images of monitored objects into 7 classes.

Neural network layers (Fig. 2):

1. Input – 1) Input Input 1.
2. Convolutional – 3) Conv2D: Entry block (filters=16, kernel_size=[3, 3], strides=[1, 1], padding='same', activation='relu').
3. Normalization – 4) BatchNormalization.
4. Convolutional – 5) Conv2D: Layer 2 (filters=16, kernel_size=[3, 3], strides=[1, 1], padding='same', activation='relu').
5. Normalization – 6) BatchNormalization.
6. MaxPool – 7) MaxPool2D: 1st Do...(pool_size=[2, 2], padding='same').
7. Convolutional – 8) Conv2D: Layer 6 (filters=32, kernel_size=[3, 3], strides=[1, 1], padding='same', activation='relu').
- ...
8. Concatenate – 30) Concatenate: Layer.
- ...
56. Output – 2) Conv2D: Выход2 (filters=2, kernel_size=[3, 3], strides=[1, 1], padding='same', activation='sigmoid').



Fig. 2. The architecture of the proposed CNN (basic U-Net) in the Terra AI framework

Python code snippet using the Keras library for a neural network is shown in Fig. 2:

```

from tensorflow.keras.layers import Input
from tensorflow.keras.layers import Conv2D
from tensorflow.keras.layers import BatchNormalization
from tensorflow.keras.layers import MaxPool2D
from tensorflow.keras.layers import Conv2DTranspose
from tensorflow.keras.layers import Concatenate
from tensorflow.keras.models import Model
input_1 = Input(shape=(512, 512, 3), name='1')
x_3 = Conv2D(filters=16, kernel_size=[3, 3], strides=[1, 1], padding='same', activation='relu', data_format='channels_last', dilation_rate=[1, 1], groups=1, use_bias=True, kernel_initializer='glorot_uniform', bias_initializer='zeros', kernel_regularizer=None, bias_regularizer=None, activity_regularizer=None, kernel_constraint=None, bias_constraint=None, name='Conv2D_3')(input_1)
x_4 = BatchNormalization(axis=-1, momentum=0.99, epsilon=0.001, center=True, scale=True, beta_initializer='zeros', gamma_initializer='ones', moving_mean_initializer='zeros', moving_variance_initializer='ones', beta_regularizer=None, gamma_regularizer=None, beta_constraint=None, gamma_constraint=None, name='BatchNormalization_4')(x_3)
...
output_2 = Conv2D(filters=7, kernel_size=[3, 3], strides=[1, 1], padding='same', activation='softmax', data_format='channels_last', dilation_rate=[1, 1], groups=1, use_bias=True, kernel_initializer='glorot_uniform', bias_initializer='zeros', kernel_regularizer=None, bias_regularizer=None, activity_regularizer=None, kernel_constraint=None, bias_constraint=None, name='2')(x_5)
model = Model([input_1], [output_2])
    
```

Input layer is used to load the input data (image).

The convolutional layer is the main layer of a convolutional neural network. The convolution layer includes for each channel its own filter, the core of the convolution that processes the previous layer into fragments (summing up the results of the element-by-element product for each fragment).

Normalization layer is necessary so that different elements in different places of the same feature map (the image of the convolution operation) are normalized in the same way.

MaxPool layer is necessary to speed up the learning process and reduce the computing resources used.

Unifying layer combines the outputs of the neural network layers.

The output layer is the last layer of the neural network, which gives the output data (result) of the neural network.

The architecture of the proposed CNN is similar to U-Net, the difference is the dimensionality of the input and output layer of the network.

As performance indicators that characterize the process of training and evaluating the effectiveness of CNN, the following ones are chosen in [18]:

- accuracy is the ratio of correctly segmented objects to the total number of expected and true objects [18]:

$$Accuracy_{val} = \sum_{t=1}^{N_{val}} \frac{N_{TP_t}}{N_{TP_t} + N_{FP_t}} \cdot 100\%, \quad (5)$$

where N_{TP} is the number of correctly segmented objects in the aerial photograph; N_{FP} is the number of erroneously segmented objects in the aerial photograph; N_{val} is the number of aerial photographs in the test sample; t is the current aerial image;

- sensitivity is the ratio of correctly segmented objects to the total number of objects in an aerial photograph [18]:

$$Sensitivity_{val} = \sum_{t=1}^{N_{val}} \frac{N_{TP_t}}{N_{TP_t} + N_{FN_t}} \cdot 100\%, \quad (6)$$

where N_{FN} is the number of erroneously unsegmented objects in the aerial image.

We tested the CNN models using a computer ACPI X64 (China), equipped with a Tesla GPU card of 12 GB and a RAM capacity of 8 GB.

To prepare aerial photographs for the training sample, the Image Labeler software from the mathematical programming environment MATLAB R2020b (USA) was used. The preparation (marking) of aerial photographs of objects “Truck”, “Car” is shown in Fig. 3.

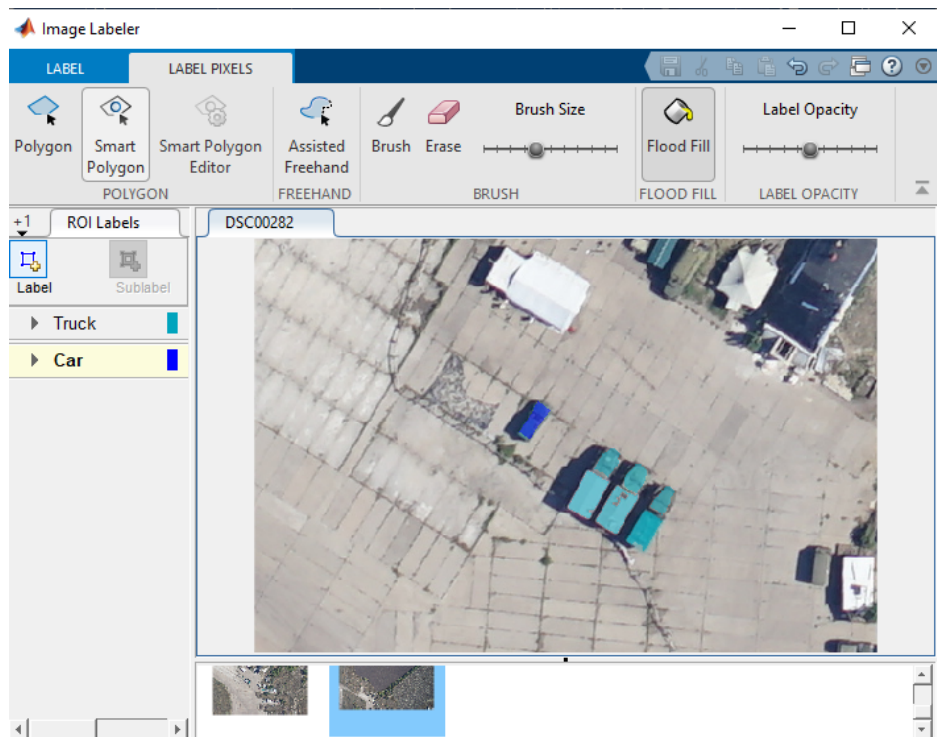


Fig. 3. Preparation of a segmented image of a training sample of objects “truck, passenger car” from an aerial photograph by the Image Labeler software in the mathematical programming environment MATLAB R2020b

Their studies were carried out under the following assumptions and limitations:

- a digital camera is installed on board the UAV and shoots in the view range in the daytime;
- an aerial photograph in digital form is transmitted through a communication channel to a ground control point;
- the process of the semantic segmentation of images of objects in an aerial photograph is carried out using a computer of the ground control point of the unmanned aerial system.

5. Results of studying the effectiveness of segmenting images of monitored objects in aerial photographs using CNN

5.1. Investigating the effectiveness of segmentation of images of monitored objects using CNN

The efficiency of segmentation of MO images using CNNs of the following models U-Net (Fig. 3), PSPsmall (Fig. 4), U-Netaverage (Fig. 5) was investigated. To train and test the models, a set of aircraft images with a dimension of 128x160x3, RGB type, JPEG format, was used. Training sample includes 800 images, verification – 140 images.

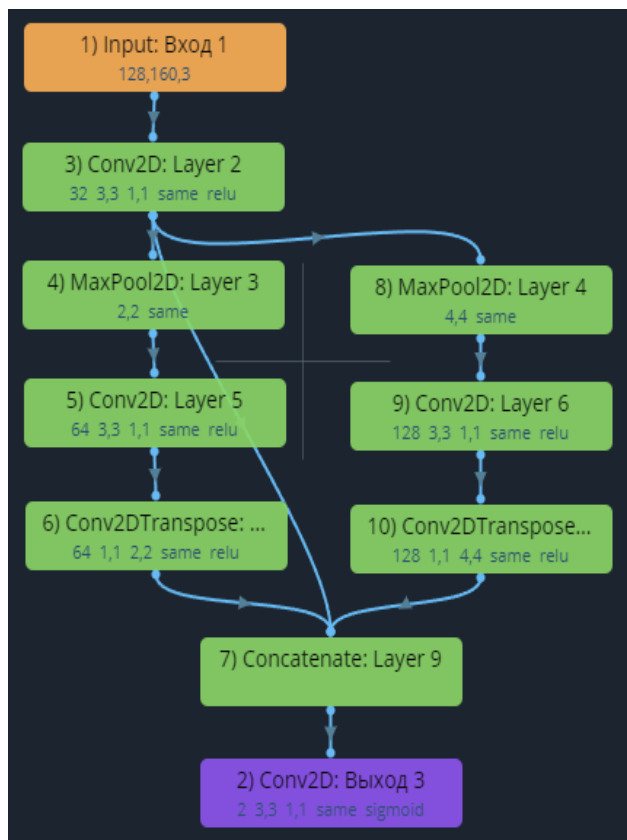


Fig. 4. The PSPsmall model architecture in the Terra AI framework

The neural network’s layers (Fig. 4):

1. Input – 1) Input 1 input image size 128, 160, 3 pixels.
2. Convolutional – 3) Conv2D: Layer 2 (filters=32, kernel_size=[3, 3], strides=[1, 1], padding='same', activation='relu').

3. MaxPool – 4) MaxPool2D: Layer 3 (pool_size=[2, 2], padding='same').

4. Convolutional – 5) Conv2D: Layer 5 (filters=64, kernel_size=[3, 3], strides=[1, 1], padding='same', activation='relu').

5. Convolutional – 6) Conv2D: Layer 2 (filters=64, kernel_size=[3, 3], strides=[1, 1], padding='same', activation='relu').

8. Output – 2) Conv2DOutput3 (filters=2, kernel_size=[3, 3], strides=[1, 1], padding='same', activation='sigmoid').

The neural network’s layers (Fig. 5):

1. Input – 1) Input 1 input image size 128, 160, 3 pixels.

2. Convolutional – 3) Conv2D: Entry block (filters=64, kernel_size=[3, 3], strides=[1, 1], padding='same', activation='relu').

3. Normalization – 4) BatchNormalization.

4. Convolutional – 5) Conv2D: Layer 3 (filters=64, kernel_size=[3, 3], strides=[1, 1], padding='same', activation='relu').

42. Output layer – 2) Conv2DOutput2 (filters=2, kernel_size=[3, 3],

- strides=[1, 1], padding='same', activation='sigmoid').

Accuracy and sensitivity were chosen as indicators of the effectiveness of semantic segmentation of images of objects by CNN. The parameters of CNN training are the duration of training (number of epochs), the optimization algorithm, the speed of learning (learning step). The physical meaning of the learning speed (learning step) of CNN is set out in [18].

The parameters for training and testing CNN models for the semantic segmentation of object images are shown in Table 2. In this case, the comparison was carried out for three types of neural networks: PSPsmall, U-Netaverage, and U-Net.

Table 2

CNN training and testing parameters

Model	Number of epochs	Time	Optimization algorithm	Packet (batch) size	Speed
PSPsmall	30	4 m 37 s	Adam	20	0.001
U-Netaverage	30	6 m 00 s	Adam	20	0.001
U-Net	30	28 m 15 s	Adam	20	0.001

For modeling, the Terra AI framework, and the MATLAB R2020b mathematical modeling environment were used.

Fig. 6 shows the accuracy plots on a test sample of PSPsmall, U-Netaverage, U-Net models.

Fig. 6 shows that the U-Net model demonstrates the best accuracy (91 %) on the test sample. After epoch 20, the accuracy of the model varies in the range from 90 % to 91 %.

Fig. 7 shows sensitivity check plots on a test sample of PSPsmall, U-Netaverage, U-Net models.

Fig. 7 shows that in the test sample, the U-Net model demonstrates the best sensitivity (87 %), which, after epoch 10, stabilizes and changes in the range from 84 % to 87 %.

Fig. 8 shows the result of the semantic segmentation of “airplane” images by the U-Net model in the Terra AI framework. In Fig. 8, during segmentation, 2 areas are highlighted: ■ – “airplane”, ■ – “sky”.

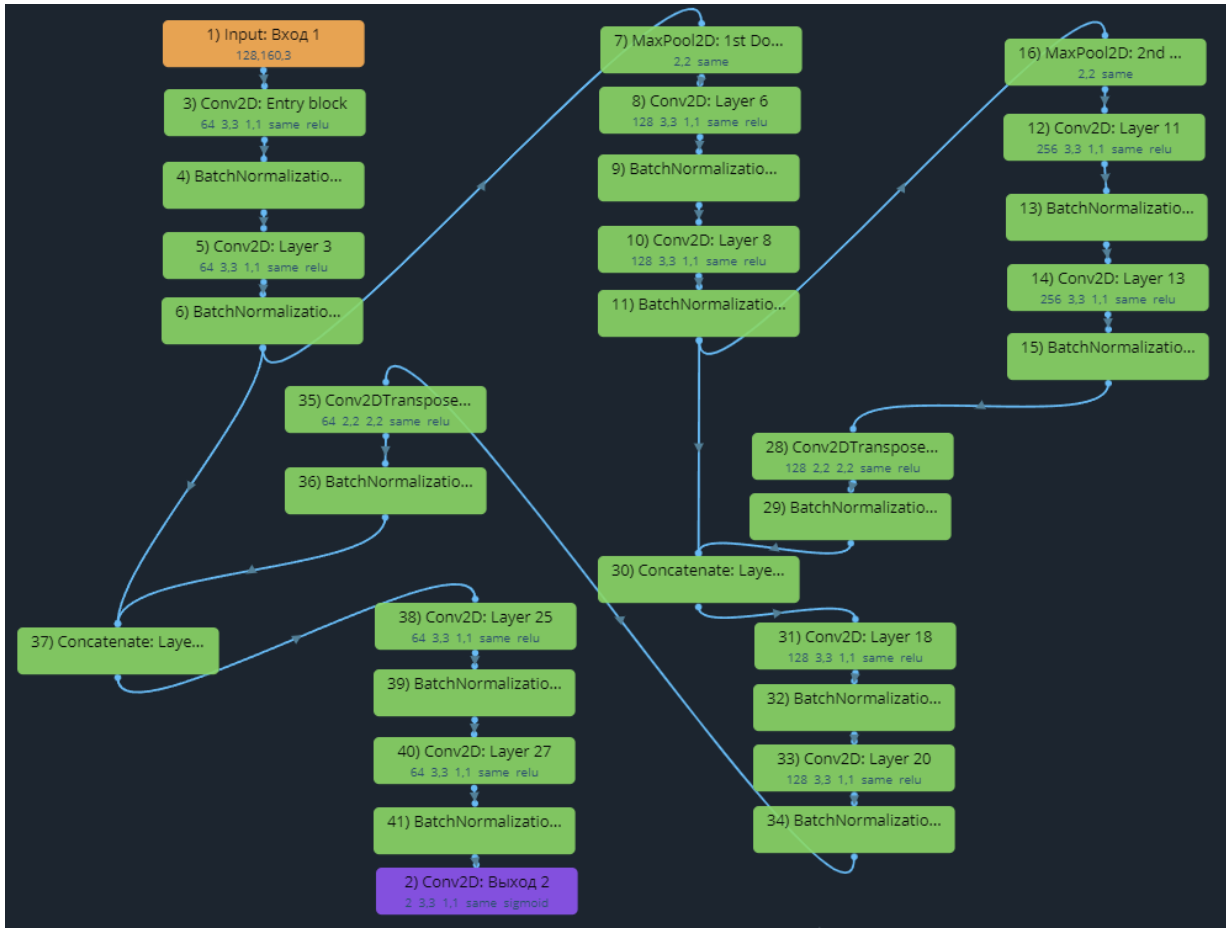


Fig. 5. The U-Netaverage model architecture in the Terra AI framework

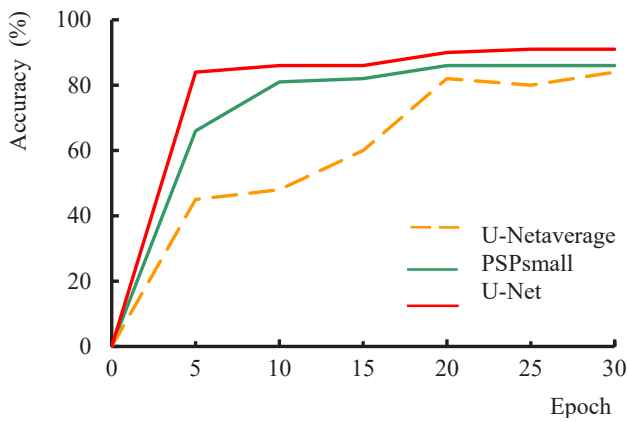


Fig. 6. Plots of accuracy change on the test sample depending on the epoch for the PSPsmall, U-Netaverage, U-Net models

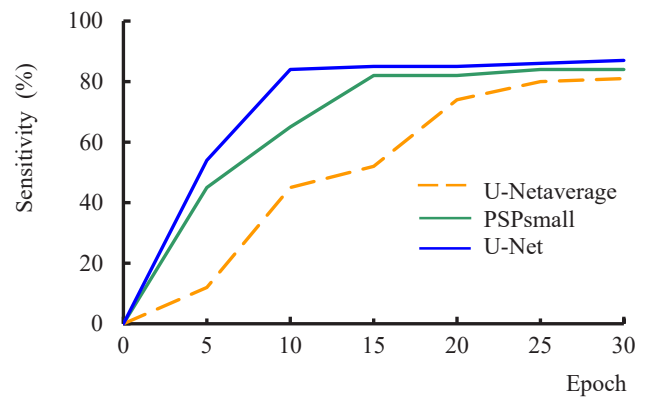


Fig. 7. Plots of sensitivity change on the test sample depending on the epoch for the PSPsmall, U-Netaverage, U-Net models



Fig. 8. Semantic segmentation of “airplane” images in the Terra AI framework using the U-Net model

Our analysis of the results reveals that the best performance indicators are shown by the U-Net model: accuracy (91 %), sensitivity (87 %), maximum error value (0.232), minimum error value (0.0132).

5.2. Evaluation of the effectiveness of segmentation of MO images on aerial photographs proposed by the U-NetWavelet model

To investigate the effectiveness of the segmentation of MO images in aerial photographs, aerial photographs of training and verification samples were prepared. 100 aerial photographs were used as a training sample. The total number of classes for semantic segmentation was 7 (helicopter, airplane, tank, tractor, truck, car, bus). The type of training and test samples (the same) is an aerial photograph of 6,000×4,000 pixels; JPEG format. 80 aerial photographs were used as a test sample.

The segmentation of images of monitored objects in aerial photographs using CNN was carried out at a ground control point. For the shooting, a UAV was used, which is equipped with a Sony ILCE-7M2 camera. This camera took aerial photographs under the following mode:

- shutter speed, 1/1,600 s;
- focal length, 55 mm;
- aerial image size (pixels): 6,000×4,000 (24M).

The aerial photograph was taken by a Sony ILCE-7M2K digital camera aboard a UAV at an altitude of 1,100 meters; it is shown in Fig. 9.



Fig. 9. Aerial photograph, taken by a Sony ILCE-7M2K digital camera

The study procedure (modeling) using an example of the proposed U-NetWavelet model:

- Step 1. Download aerial photographs: 6,000×4,000×3 pixels.
- Step 2. Split aerial photographs of 6,000×4,000×3 pixels into 1,000×1,000×3 images; a total of 24 for each aerial photograph.
- Step 3. Apply a wavelet layer to a snapshot of 1,000×1,000×3 (implemented on a modified Haar transform – the value of the adjacent two pixels is summed up and divided by two) and adapted to the dimension of 512×512×3.
- Step 4. Split data into training and validation datasets.
- Step 5. Training and validation of the network.
- Step 6. Segmentation of verification sample snapshots.
- Step 7. Evaluate the accuracy of the segmentation of the test sample.
- Step 8. Evaluate the sensitivity of the model on a test sample.
- Step 9. Assemble the segmented aerial photograph of 3,072×2,048×3.

We trained the proposed U-NetWavelet model by using the optimal values of the parameters, which were obtained experimentally:

- learning speed – 0.001;
- learning duration (the number of epochs) – 60;
- packet (batch) size – 20;
- optimization algorithm – Adam.

As a result, a new model with the proposed name U-Net-Wavelet was built. The results of checking the accuracy and sensitivity of this neural network are shown in Fig. 10.

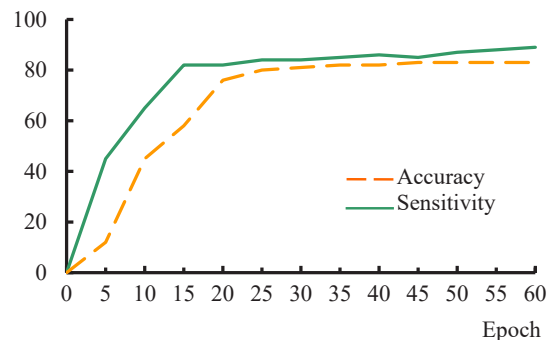


Fig. 10. Epoch-dependent plots of changes in the accuracy and sensitivity of the U-NetWavelet model on a test sample

Fig. 10 shows that for the U-NetWavelet model, the accuracy on the test sample stabilizes after epoch 15, and, after the end of epoch 60, the accuracy reaches – 89 %, sensitivity – 83 %.

Fig. 11 shows a fragment of the segmented aerial photograph using the U-NetWavelet model.

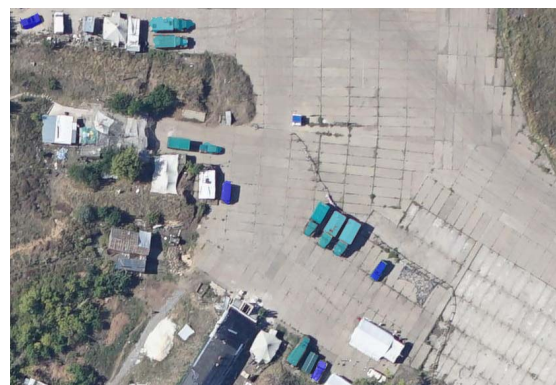


Fig. 11. Fragment of the segmented aerial photograph using the U-NetWavelet model

In Fig. 11, two types of objects are distinguished during segmentation: ■ – “passenger car”, ■ – “truck”.

A comparison of the new U-NetWavelet model was made with the FCN, SegNet models. 80 aerial photographs were used as a test sample to evaluate the U-NetWavelet model for convergence, adequacy, and validity.

Convergence. A CNN shows convergence provided that the error decreases with each epoch. The convergence of the CNN model is influenced by three components: the completeness of the database (aerial photographs); the correct choice of architecture; selection of CNN training parameters.

Fig. 12 shows the U-NetWavelet convergence score on a test sample.

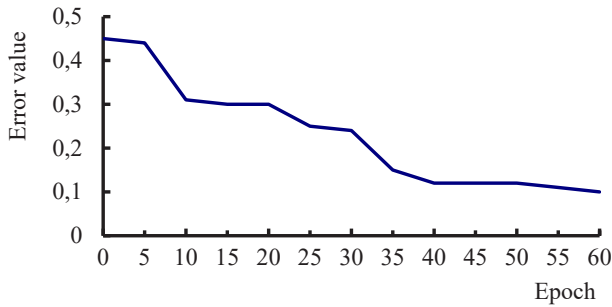


Fig. 12. Estimation of convergence of the proposed model U-NetWavelet

Our analysis of Fig. 12 reveals that the proposed U-Net-Wavelet model does have convergence.

Adequacy. A neural network is adequate if the learning outcomes converge to close values, a necessary condition that there is a dependence between the output and input data, which is implemented by CNN.

The most recommended way to test a CNN model for adequacy is to compare the results with known models.

The results of checking on the test sample (80 aerial photographs) are shown in Table 3.

Table 3

Results of model accuracy and sensitivity assessment

Model	Accuracy, %	Sensitivity, %	Error maximal value	Error minimal value
FCN	83	79	0.741	0.187
SegNet	85	82	0.536	0.124
Proposed model U-NetWavelet	89	83	0.451	0.102

Table 3 shows that in comparison with the FCN, SegNet models, the proposed U-NetWavelet model demonstrates the best efficiency indicators: accuracy (89 %), sensitivity (83 %), maximum error value (0.451), minimum error value (0.102).

6. Discussion of results of studying the semantic segmentation of images of objects in aerial photographs using CNN

It is proposed to use the U-Net CNN [15, 16] to segment images of objects in aerial photographs. To improve the efficiency of the neural network, this model was trained by a set of aerial photographs (Fig. 9) with the selection of optimal parameters (speed (step) of training, the number of epochs, packet size (batch), optimization algorithm). As a result, a new model with the proposed name U-NetWavelet was constructed (Fig. 2).

Due to the use of a modified wavelet layer, the size of the aerial photograph adapts to the parameters of the input layer of the neural network; the efficiency of segmentation of images in aerial photographs increases. The use of the U-NetWavelet CNN makes it possible to increase the performance and automate the process of semantic segmentation of MO images.

Using the proposed model allows us to solve the following issues [8–14]:

- the computational complexity of segmentation of MO images on aerial photographs obtained from UAVs;

- the lack of neural network models that solve the task of the segmentation of MO images in aerial photographs.

Note the following limitations in the proposed model:

- the segmentation of MO images on aerial photographs is carried out within 7 classes (Table 1);
- the orientation of MOs in the images is not taken into consideration;
- the resolution of aerial photographs for the classification of MOs is 6,000×4,000 pixels;
- the CNN’s transmission invariance is not taken into consideration;
- aerial photography is carried out in the visible range in the daytime.

The proposed model is constrained by that it is adapted to segment objects in an aerial photograph into seven classes. The CNN training was conducted on aerial photographs of high contrast, clarity (Fig. 6). The shooting was carried out in the daytime, the time of year was summer. Therefore, high values of accuracy and sensitivity of the segmentation of object images were obtained (Table 3). For other types of images of objects (shooting conditions), the accuracy, sensitivity of the segmentation of MO images by class may vary, which requires additional research.

It is planned to advance the proposed model by:

- increasing the base of marked (segmented) aerial photographs for a training sample;
- exploring the proposed and other models [19–21] (PSPNet, DenseNet, DeepLab, DilatedNet, etc.) for different conditions of aerial photography;
- optimizing the proposed model in terms of computational complexity to increase performance;
- building method for counting the number of objects in aerial photographs by class;
- devising a method for detecting and identifying objects in the video stream received by the UAV video camera.

Our model is proposed to be used at a ground control point of UAV when processing aerial photographs, ortho-photo plans; in systems with artificial intelligence; in MO control systems; when designing robots; in unmanned vehicle systems.

7. Conclusions

1. The indicators of efficiency of PSPsmall, U-Netaverage, U-Net models have been studied. Verification of the effectiveness of these models was carried out on the basis of images of aircraft (800 images in a training sample, 140 in a test sample). It has been established that the best indicators are shown by the U-Net model: accuracy (91 %), sensitivity (87 %), maximum error value (0.232), minimum error value (0.0132). The lowest accuracy (84 %) and sensitivity (81 %) are shown by the U-Netaverage model.

2. The effectiveness of the proposed U-NetWavelet model (based on images prepared from aerial photographs) was evaluated. The model has the best efficiency indicators in comparison with the FCN, SegNet models: accuracy (89 %), sensitivity (83 %), maximum error value (0.451), minimum error value (0.102). The obtained values of the performance indicators of the U-NetWavelet model allow us to assert the correctness of the choice of the CNN architecture and the selection of its training parameters: the learning rate is 0.001; the duration of training (number of epochs) is 60; the optimization algorithm is Adam.

References

1. Pospelov, B., Andronov, V., Rybka, E., Krainiukov, O., Maksymenko, N., Meleshchenko, R. et. al. (2020). Mathematical model of determining a risk to the human health along with the detection of hazardous states of urban atmosphere pollution based on measuring the current concentrations of pollutants. *Eastern-European Journal of Enterprise Technologies*, 4 (10 (106)), 37–44. doi: <https://doi.org/10.15587/1729-4061.2020.210059>
2. Semko, A. N., Beskrovnaya, M. V., Vinogradov, S. A., Hritsina, I. N., Yagudina, N. I. (2014). The usage of high speed impulse liquid jets for putting out gas blowouts. *Journal of Theoretical and Applied Mechanics*, 52 (3), 655–664.
3. Chernukha, A., Teslenko, A., Kovalov, P., Bezuglov, O. (2020). Mathematical Modeling of Fire-Proof Efficiency of Coatings Based on Silicate Composition. *Materials Science Forum*, 1006, 70–75. doi: <https://doi.org/10.4028/www.scientific.net/msf.1006.70>
4. Vambol, S., Vambol, V., Kondratenko, O., Suchikova, Y., Hurenko, O. (2017). Assessment of improvement of ecological safety of power plants by arranging the system of pollutant neutralization. *Eastern-European Journal of Enterprise Technologies*, 3 (10 (87)), 63–73. doi: <https://doi.org/10.15587/1729-4061.2017.102314>
5. Vambol, S., Vambol, V., Sobyna, V., Koloskov, V., Poberezhna, L. (2018). Investigation of the energy efficiency of waste utilization technology, with considering the use of low-temperature separation of the resulting gas mixtures. *Energetika*, 64 (4), 186–195. doi: <https://doi.org/10.6001/energetika.v64i4.3893>
6. Pospelov, B., Rybka, E., Meleshchenko, R., Borodych, P., Gornostal, S. (2019). Development of the method for rapid detection of hazardous atmospheric pollution of cities with the help of recurrence measures. *Eastern-European Journal of Enterprise Technologies*, 1 (10 (97)), 29–35. doi: <https://doi.org/10.15587/1729-4061.2019.155027>
7. Dadashov, I., Loboichenko, V., Kireev, A. (2018). Analysis of the ecological characteristics of environment friendly fire fighting chemicals used in extinguishing oil products. *Pollution Research*, 37 (1), 63–77. Available at: <http://repositsc.nuczu.edu.ua/handle/123456789/6849>
8. Holla, A., Pai, M., Verma, U., Pai, R. M. (2020). Efficient Vehicle Counting by Eliminating Identical Vehicles in UAV aerial videos. 2020 IEEE International Conference on Distributed Computing, VLSI, Electrical Circuits and Robotics (DISCOVER), 246–251. doi: <https://doi.org/10.1109/discover50404.2020.9278095>
9. Deng, H., Zhang, Y., Li, R., Hu, C., Feng, Z., Li, H. (2022). Combining residual attention mechanisms and generative adversarial networks for hippocampus segmentation. *Tsinghua Science and Technology*, 27 (1), 68–78. doi: <https://doi.org/10.26599/tst.2020.9010056>
10. Jing, W., Jin, T., Xiang, D. (2021). Fast Superpixel-Based Clustering Algorithm for SAR Image Segmentation. *IEEE Geoscience and Remote Sensing Letters*, 1–1. doi: <https://doi.org/10.1109/lgrs.2021.3124071>
11. Xin, L., Chao, L., He, L. (2021). Malicious code detection method based on image segmentation and deep residual network RESNET. 2021 International Conference on Computer Engineering and Application (ICCEA), 473–480. doi: <https://doi.org/10.1109/ICCEA53728.2021.00099>
12. Xie, B., Yang, Z., Yang, L., Luo, R., Wei, A., Weng, X., Li, B. (2021). Multi-Scale Fusion With Matching Attention Model: A Novel Decoding Network Cooperated With NAS for Real-Time Semantic Segmentation. *IEEE Transactions on Intelligent Transportation Systems*, 1–11. doi: <https://doi.org/10.1109/tits.2021.3115705>
13. Yang, S., Hou, J., Jia, Y., Mei, S., Du, Q. (2021). Superpixel-Guided Discriminative Low-Rank Representation of Hyperspectral Images for Classification. *IEEE Transactions on Image Processing*, 30, 8823–8835. doi: <https://doi.org/10.1109/tip.2021.3120675>
14. Peng, C., Zhang, K., Ma, Y., Ma, J. (2021). Cross Fusion Net: A Fast Semantic Segmentation Network for Small-Scale Semantic Information Capturing in Aerial Scenes. *IEEE Transactions on Geoscience and Remote Sensing*, 60, 1–13. doi: <https://doi.org/10.1109/tgrs.2021.3053062>
15. Ronneberger, O., Fischer, P., Brox, T. (2015). U-Net: Convolutional Networks for Biomedical Image Segmentation. *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, 234–241. doi: https://doi.org/10.1007/978-3-319-24574-4_28
16. Jwaid, W. M., Al-Husseini, Z. S. M., Sabry, A. H. (2021). Development of brain tumor segmentation of magnetic resonance imaging (MRI) using U-Net deep learning. *Eastern-European Journal of Enterprise Technologies*, 4 (9 (112)), 23–31. doi: <https://doi.org/10.15587/1729-4061.2021.238957>
17. Slyusar, V., Protsenko, M., Chernukha, A., Gornostal, S., Rudakov, S., Shevchenko, S. et. al. (2021). Construction of an advanced method for recognizing monitored objects by a convolutional neural network using a discrete wavelet transform. *Eastern-European Journal of Enterprise Technologies*, 4 (9 (112)), 65–77. doi: <https://doi.org/10.15587/1729-4061.2021.238601>
18. Slyusar, V., Protsenko, M., Chernukha, A., Kovalov, P., Borodych, P., Shevchenko, S. et. al. (2021). Improvement of the model of object recognition in aero photographs using deep convolutional neural networks. *Eastern-European Journal of Enterprise Technologies*, 5 (2 (113)), 6–21. doi: <https://doi.org/10.15587/1729-4061.2021.243094>
19. Long, J., Shelhamer, E., Darrell, T. (2015). Fully convolutional networks for semantic segmentation. 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). doi: <https://doi.org/10.1109/cvpr.2015.7298965>
20. Badrinarayanan, V., Kendall, A., Cipolla, R. (2017). SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39 (12), 2481–2495. doi: <https://doi.org/10.1109/tpami.2016.2644615>
21. Zhao, H., Shi, J., Qi, X., Wang, X., Jia, J. (2017). Pyramid Scene Parsing Network. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). doi: <https://doi.org/10.1109/cvpr.2017.660>