Improved PSP and U-Net architectures for forest segmentation in remote sensing pictures

Vadym Slyusar
Central Research Institute of Armaments and Military Equipment of Armed Forces of Ukraine
Kyiv, Ukraine
swadim@ukr.net

Ihor Sliusar
Department of information systems and technologies
Poltava State Agrarian University
Poltava, Ukraine
islyusar2007@ukr.net

Pavlenko Anatolii
Department of information systems and technologies
Poltava State Agrarian University
Poltava, Ukraine
tolikpavlenko11@gmail.com

***Abstract* — Various PSP and U-Net architectures for forest segmentation in remote sensing pictures have been proposed and investigated. The main improvements of the proposed architectures are based on using the BathNormalization layers, replacing MaxPool2D layers with AveragePooling2D, changing Conv2DTranspose to UpSampling2D blocks, etc. For the training of neural networks was used modified dataset of 128x128 pictures based on the dataset from Kaggle. As a result of improving architecture was given the maximum segmentation accuracy of 80.8 % on the validation set of pictures.**

Keywords — Semantic segmentation, Convolutional Neural Networks, Fully Convolutional Network, U-Net, Pyramid Scene Parsing (PSP)

I. Introduction

When using satellite [1] and high-altitude platforms [2] for remote sensing of the Earth, one of the important tasks is the monitoring of forests. As you know, the state of the forest ecosystem can be influenced by many factors, the main of which should be considered: climate change, the intensity of fires, the degree of forest restoration after a fire, the composition of the fauna in a given habitat, illegal logging or other options for human impact. In addition, it is quite a difficult task to monitor fires, including those of natural origin, that occur in forests with a heterogeneous topography, which are geographically remote from settlements.

All this testifies to the expediency of using images obtained based on remote sensing of the Earth for the operational assessment of the forest ecosystem. Such images very often contain diverse homogeneous regions, for which the intraclass standard deviations of their characteristics are often comparable to the spread between classes. In addition, they may have low resolution. As a result, classical segmentation methods do not guarantee the required result.

To solve this problem, artificial intelligence is increasingly being used, namely, neural networks [3] - [7]. They provide more accurate estimates of automatic feature identification and terrain classification.

II. Analysis of Recent Studies and Publications, which Discuss the Problem

The solution to the problems of remote sensing of the Earth by aerospace-based means is mainly implemented based on neural network technologies Semantic segmentation [8] (determines whether the sets of pixels in the image belong to certain classes of objects) or Instance segmentation (each object within one class is allocated by separate segments). In turn, the combination of these approaches (Semantic and Instance segmentation) gives rise to Panoptic segmentation technology.

Semantic segmentation can use several types of Convolutional Neural Networks (CNN) [9]-[13]. They are followed by Fully Convolutional Network (FCN) [14], U-Net and modifications [15]-[19], as well as Pyramid Scene Parsing (PSP) [20], etc.

According to [20], although FCNs have an efficient underlying architecture, they suffer from several drawbacks. First, the presence of staggered artifacts associated with non-uniform overlap of outputs in the transposed convolution operation. Secondly, low resolution at the edges is due to loss of information in the encoding process. To eliminate them, several solutions have been developed, for example, U-Net, DeepLab, and PSP, as well as their derivatives and modifications.

At the same time, in the U-Net model, the architecture is improved by using narrowing convolutions for context capture, expanding convolutions for localization, as well as direct links between convolutions at the same levels. The PSP combines features from several scales without significantly increasing the number of parameters. This allows you to study a more general context.

However, the implementation of deep learning models often requires not only a significant amount of training data, but also the corresponding computing power. The latter factor can play a decisive role in their implementation on unmanned robotic platforms or other AI IoT solutions.

As a consequence, it is advisable to target low-resolution images. At the same time, the properties of the different types of CNNs used affect the level of quality of the results obtained. The analysis of existing works indicates the dominance of U-Net studies, while PSP requires more detailed study.

III. The Aim of Research

The aim of the work is to analyze the properties of deep learning models of various CNN architectures based on the results of mathematical modeling in the interests of the problem of forest segmentation based on low-resolution images.

IV. The Main Results of the Study

As you know, when implementing a deep learning model, the formation of a dataset plays an important role. To create it, we used data from the set from Kaggle [21]. This dataset was derived from the land cover classification track in the DeepGlobe Challenge. In [21], the images in the dataset were changed to 256x256 images to create more sample images.

Their number was increased to 5108 pieces.

In this work, during the research, the resources of Google Colab Pro + with GPU Tesla V100-SXM2-16GB and Ceras were used.

At the same time, to work out the learning technology, the specified dataset was modified by switching to 128x128 images. As a result, instead of the original size of 1.2 GB with a format of 256x256, the size of the compressed set was 476.14 MB, and it was also possible to increase the batch, which became equal to 16. When preparing the dataset, two folders were created separately. One of them contains the original images, and the other contains the mask files. The following structure was placed in the root directory:

```
# label:color_rgb:parts:actions
Forest:0,0,0::
Field:255,255,255::
```

The training sample was formed based on the full dataset, divided into a ratio of 70:30.

At the same time, the percentage of space in the training sample was: 38% for the first class (field) and 62% for the second class (forest), and the percentage of space in the test sample was: 40% for the first class (field) and 60% for the second class (forest). In total, in the training sample, the 1st class was represented on 3565 photographs, and the 2nd class - on 3570. In the test sample, the 1st class accounted for 1527 photographs, and the 2nd class for 1529.

Unfortunately, some of the masks in the dataset were made quite roughly and did not always correspond to the real picture of the distribution of forest stands. Therefore, the input control of the dataset for compliance with the masks was additionally performed. The Adam optimizer was used to optimize the loss function.

In the first stage, several variants of the PSP architecture were evaluated (for example, Fig. 1 and 2).

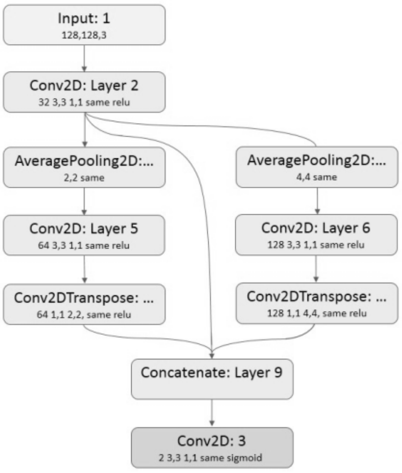

Fig. 1. The modified architecture of the “small” PSP based on the PSPsmall in the Terra AI framework [8] (all MaxPool2D layers replaced with AveragePooling2D).

On their basis, modifications were created that made it possible to evaluate the influence of the architecture on the result obtained. For example, an improvement in accuracy can be achieved if:

– immediately after the input block add the BathNormalization layer;

– replace all MaxPool2D layers with AveragePooling2D of the appropriate size and add them immediately after the input or BathNormalization.

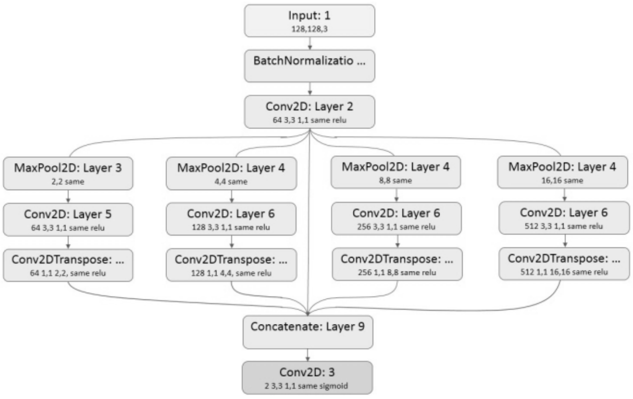

Fig. 2. Architecture of the “big” PSP from the Terra AI framework.

As you know, MaxPool2D provides image size reduction. It takes a parameter called the pool (kernel) size and gets the first p x p pixels of the image. It then finds the maximum value in those pixel values and stores it as the first pixel value for the output image. The layer continues the process for the entire image, moving through the image and outputting the image with the same data in a more compressed form. In turn, AveragePool2D uses the arithmetic mean instead of the maximum.

During training, the maximum test accuracy was obtained at 40 epochs and was 77.2% with a learning rate of 0.001 (Fig. 3). Some results of using PSP are shown in Fig. 4.

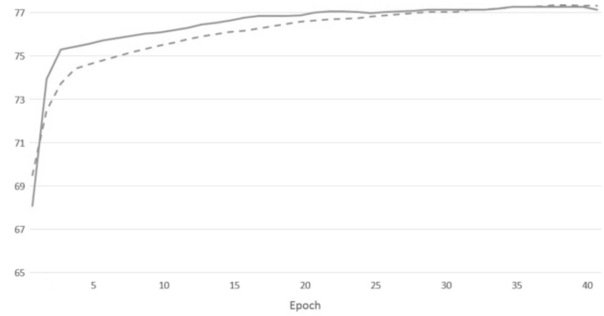

Fig. 3. Learning Outcomes (the best result on the training sample 77.4%, the best result on a provocative sample 77.2%): solid - provocative sample; dash - training sample.

At the next stage, the influence of the variation in the dimension of the kernels in the Conv2DTranspose layers was studied. At the same time, it was found that increasing the dimension according to the scaling factor leads to a more uniform filling of the data matrix. In turn, the elimination of empty points in Conv2DTranspose made it possible to increase accuracy, but the rendering of fine details deteriorated.

In general, this approach made it possible to achieve an accuracy of the 98th epoch - 80% on the test sample (the learning rate given to 0.001).

Thus, we can conclude that it is necessary to increase the filling of voids in Conv2DTranspose, but not raise it to 16 or even limit it to level 8 (4). That is, acceptable accuracy is maintained when the kernel values in Conv2DTranspose are rolled back to 1x1 and 2x2 values, while only 1x1 kernels were used in the original scheme (Fig. 5).

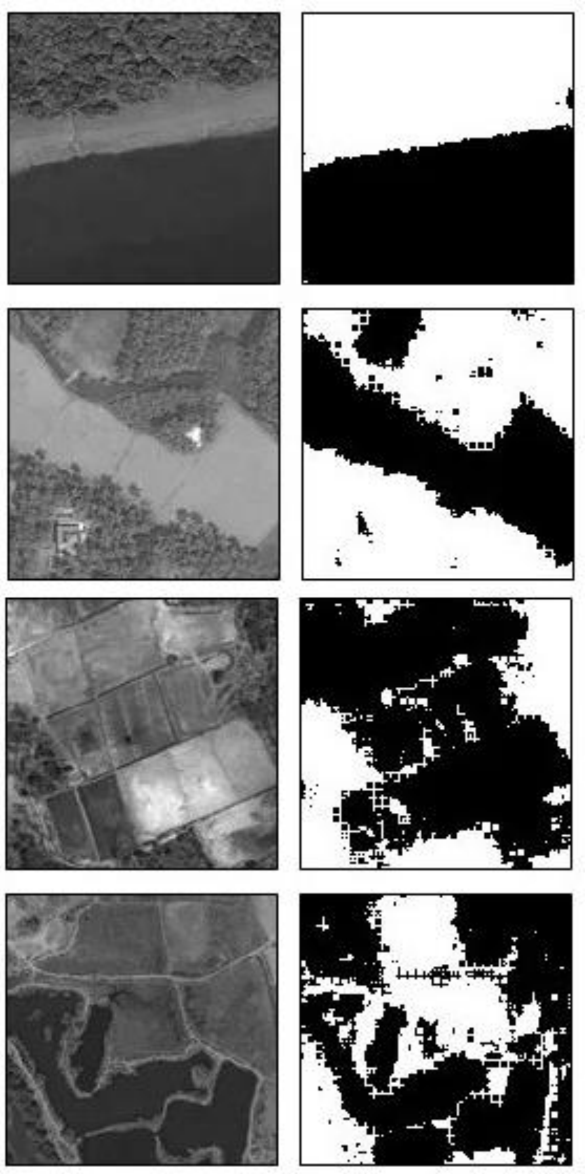

Fig. 4. Results of Using Deep Learning Models for Forest Segmentation Based on the PSP Architecture.

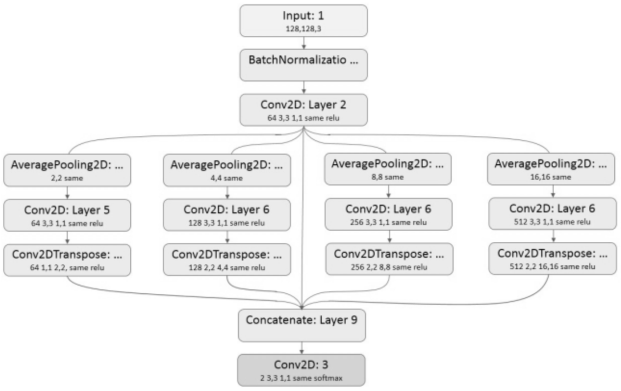

Fig. 5. PSP architecture with cores in Conv2DTranspose 1x1 and 2x2.

The next stage of research was the modification of the PSP architecture by replacing the Conv2DTranspose layers with UpSampling2D with the same multiplier (Fig. 6).

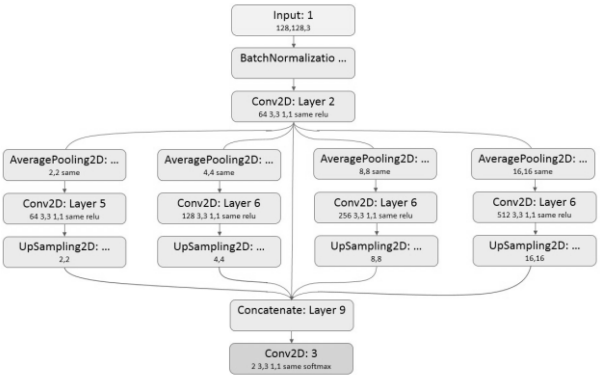

Fig. 6. Replacing Conv2Dtranspose layers with UpSampling2D with the same multiplier.

These are two common types of layers that can be used to increase the size of arrays. As you know, Conv2DTranspose performs UpSampling2D and convolution (transposed convolution). At the same time, theoretically, it can lead to the appearance of “chessboard” artifacts. UpSampling2D is like a pool in that it repeats the rows and columns of the input.

Formally, switching to UpSampling2D gave better accuracy rates, almost 80% already at the 20th epoch (Fig. 7). But this PSP architecture produces more visual mismatches and follows error masks more strictly without highlighting fine details (Fig. 8). This is because UpSampling2D, unlike Conv2DTranspose, does not contain training weights, and therefore Conv2DTranspose is more adaptive for solving the segmentation problem.

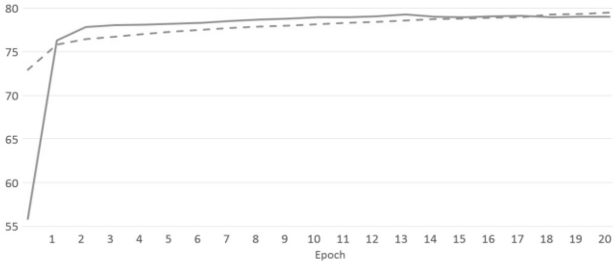

Fig. 7. Results of training a PSP model with UpSampling2D (the best result on the training sample 79.5%, the best result on a provocative sample 79.3%): solid - provocative sample; dash - training sample.

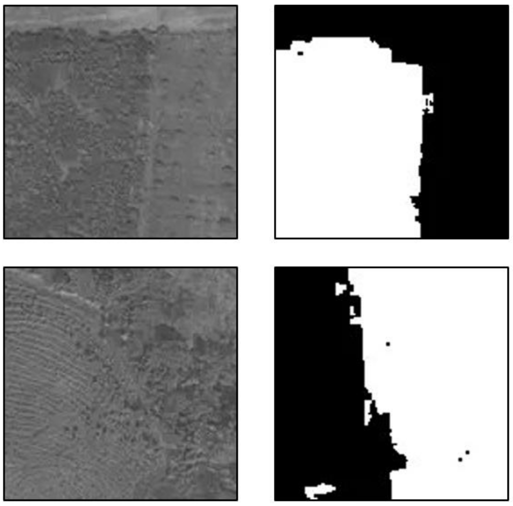

Fig. 8. Results of using deep learning models for forest segmentation based on PSP architecture with UpSampling2D blocks.

In conclusion, the work carried out an analysis of the impact on the final result of the U-Net architecture and its comparative assessment with PSP. An example of the “small” U-Net architecture is shown in Fig. 9. During the research, the following results were obtained:

– “small” U-Net gives an accuracy of 99% on the training sample and 80% on the test sample;

– A “large” U-Net gives an accuracy of 87% on the training set and 70% on the test set.

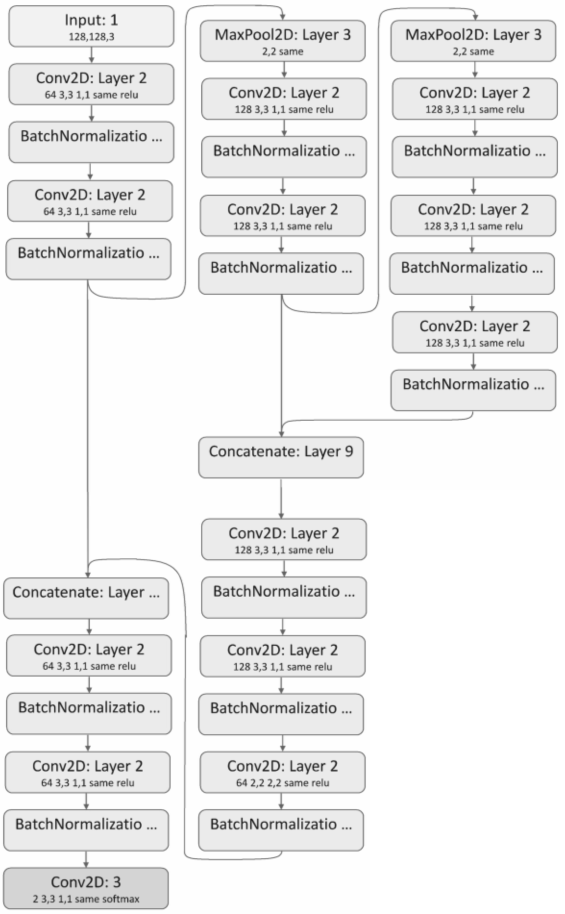

Fig. 9. Architecture “Small” U-Net from the Terra AI framework.

The use of more complex “U-Net++” (fig. 10) and “U-Net^2” (fig. 11) [21] architectures is accompanied by their rapid retraining and does not allow a significant increase in the training accuracy. In particular, on the specified modified dataset “U-Net++” allowed to achieve the maximum accuracy of 79.7%, and “U-Net^2” – 80.8%. At the same time, in the “U-Net^2” architecture, AveragePooling2D layers were used instead of MaxPooling2D in intercascade connections, and Conv2Transpose layers were used instead of UpSampling2D.

Unfortunately, due to text size limitations, the architectures of these “large” U-Nets are given here without proper detail.

In general, the results obtained allow us to conclude that the U-Net renders better shading than the small PSP. But, in turn, the PSP on the considered dataset works more accurately.

In conclusion, need to highlight that the proposed PSP and U-Net architectures overcome the main drawback of the FCN framework for semantic segmentation of remote sensing pictures without regard to contextual information. It is possible based on the extraction of complete feature presentations and gives improved segmentation accuracy.

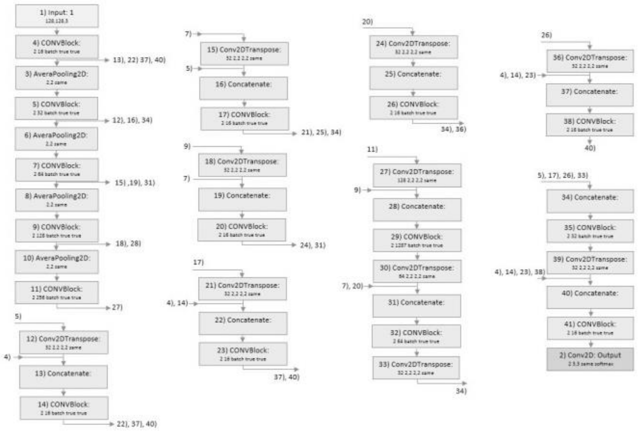

Fig. 10. The modified architecture “U-Net++”.

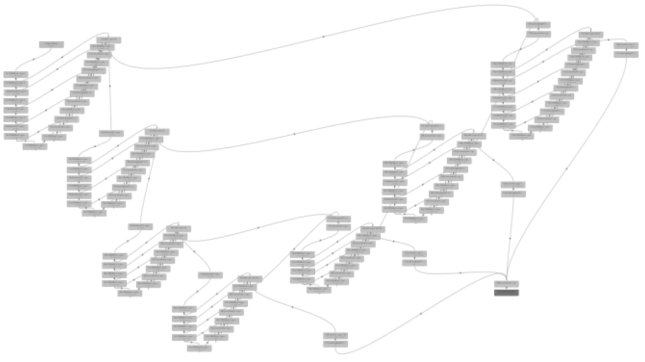

Fig. 11. Architecture “U-Net^2” from the Terra AI framework.

V. Perspectives of further Reseach

Further research may be aimed at evaluating hybrid architectures, for example, using pre-trained networks in branches with different scales [22]. Sets of pre-trained models are provided by Tensorflow, PyTorch, Keras, and Caffe2. In addition, need to evaluate the possibilities of using TensorFlow Lite to deploy the developed architectures on mobile, embedded and peripheral devices, etc.

To increase the accuracy of segmentation results you can use a combination of pre-trained networks with an attention mechanism [24].

VI. Conclusions

The analysis of the Earth's remote sensing imagery can play an important role in areas such as wildlife ecology, deforestation monitoring, land cover assessment and geological mapping.

The proposed approach based on a modification of PSP or U-Net architectures can be expanded to the use of radars and infrared or other spectral diapasons pictures.

The intensive development of aerospace-based Earth remote sensing means, the increase in the volume and information content of aerospace data leads to a continuous expansion of the range of thematic tasks solved on their basis.

REFERENCES

[1] I. Demir, K. Koperski, D. Lindenbaum, et al., “A Challenge to Parse the Earth Through Satellite Images,” in *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2018. [Online]. Available: https://arxiv.org/pdf/1805.06561.pdf. [Accessed September 09, 2022].

[2] V.M. Vishnevsky, A.I. Liachov, S.L. Portnoj et al., *Shirokopolosnye besprovodnye seti peredachi informacii [Broadband wireless communication networks]*, Moscow, Russia: Technosphera, 2005, pp. 524-526. (In Russian).

[3] Z. Wang, P. Yang, H. Liang et al., “Semantic segmentation and analysis on sensitive parameters of forest fire smoke using smoke-UNet and LandSat-8 imagery,” *Remote Sensing*, 2021, vol. 14, no. 1, p. 45.

[4] M. Umar, Lakshmi B. Saheer and J. Zarrin, “Forest Terrain Identification using Semantic Segmentation on UAV Images”, in *38th Int. Conf. on Machine Learning*, 2021, 6 p.

[5] D. Filatov, G. Nabi and A. Yar, “Forest and Water Bodies Segmentation Through Satellite Images Using U-Net.” [Online]. Available: https://arxiv.org/pdf/2207.11222.pdf. [Accessed September 09, 2022].

[6] D. Kislov, K. Korznikov, J. Altman, A. Vozmishcheva, and P. Krestov, “Extending deep learning approaches for forest disturbance segmentation on very high-resolution satellite images,” *Remote Sensing in Ecology and Conservation*, 2021, vol. 7, no. 3, pp. 355-368.

[7] Z. Wang, T. Peng and Z. Lu, ‘Comparative Research on Forest Fire Image Segmentation Algorithms Based on Fully Convolutional Neural Networks,” *Forests*, 2022, no. 13, p. 1133. DOI: 10.3390/f13071133.

[8] V. Slyusar, M. Protsenko, A. Chernukha, V. Melkin, O. Petrova, M. Kravtsov, S. Velma, N. Kosenko, O. Sydorenko and M. Sobol, “Improving a neural network model for semantic segmentation of images of monitored objects in aerial photographs,” *Eastern-European Journal of Enterprise Technologies*, 2021, vol. 2, no. 6 (114), pp. 86-95. DOI: 10.15587/1729-4061.2021.248390.

[9] V. Slyusar, M. Protsenko, A. Chernukha, P. Kovalov, P. Borodych, S. Shevchenko, O. Chernikov, S. Vazhynskyi, O. Bogatov and K. Khrustalev, “Improvement of the object recognition model on aerophotos using deep conventional neural network,” *Eastern-European Journal of Enterprise Technologies*, 2021, vol. 5, no. 2 (113). pp. 6-21. DOI: 10.15587/1729-4061.2021.243094.

[10] V. Slyusar and I. Sliusar, “The Lions of the Neural Network Zoo,” in *Int. Conf. Neural network technologies and their applications*, Kramatorsk, 2021, pp. 128-133.

[11] V. Slyusar, “The tensor-matrix version of LeNet5,” in *IVth International scientific-practical conference dedicated to the 50th anniversary of the Department of Information Systems and Technologies «Integration Information Systems and Intelligent Technologies in the Conditions of Information Society Transformation»*, Poltava, 2021, pp. 114-119. DOI: 10.32782/978-966-289-562-9.

[12] S. Naumenko, I. Sliusar and V. Slyusar, “Neural network for recognition of handwritten digits,” in *IVth International scientific-practical conference dedicated to the 50th anniversary of the Department of Information Systems and Technologies «Integration Information Systems and Intelligent Technologies in the Conditions of Information Society Transformation»*, Poltava, 2021, pp. 141-143. DOI: 10.32782/978-966-289-562-9.

[13] V. Slyusar, M. Protsenko, A. Chernukha, S. Gornostal, S. Rudakov, S. Shevchenko, O. Chernikov, N. Kolpachenko, V. Timofeyev, R. Artiukh, “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*, 2021, vol. 4, no. 9(112), pp. 65-77. DOI: 10.15587/1729-4061.2021.238601.

[14] F. Yang, Q. Sun,, H. Jin and Z. Zhou, “Superpixel segmentation with fully convolutional networks,” in *IEEE Conf. on Computer Vision and Pattern Recognition*, 2020, pp. 13964-13973.

[15] O. Ronneberger, P. Fischer and T. Brox, “U-Net: Convolutional Networks for Biomedical Image Segmentation.” [Online]. Available: https://arxiv.org/pdf/1505.04597.pdf. [Accessed September 09, 2022].

[16] W. Jwaid, Z. Al-Husseini and A. Sabry, “Development of brain tumor segmentation of magnetic resonance imaging (MRI) using U-Net deep learning,” *Eastern-European Journal of Enterprise Technologies*, 2021, vol. 4, no. 9(112), pp. 23-31. DOI: 10.15587/1729-4061.2021.238957.

[17] N. Singh and K. Nongmeikapam, “Semantic segmentation of satellite images using deep-UNet,” *Arabian Journal for Science and Engineering*, 2022, pp. 1-13.

[18] A. Soni, R. Koner, and V. G. K. Villuri, “M-Unet: Modified U-Net segmentation framework with satellite imagery,” in *Proceedings of the Global AI Congress 2019*, Springer, 2020, pp. 47-59.

[19] E. Irwansyah, Y. Heryadi, and A. Gunawan, “Semantic image segmentation for building detection in urban area with aerial photograph image using U-Net models,” in *2020 IEEE Asia-Pacific Conf. on Geoscience, Electronics and Remote Sensing Technology (AGERS)*, 2020, pp. 48-51.

[20] H. Zhao, J. Shi, X. Qi, X. Wang and J. Jia, “Pyramid Scene Parsing Network.” [Online]. Available: https://arxiv.org/abs/1612.01105. . [Accessed September 09, 2022].

[21] Forest Aerial Images for Segmentation. [Online]. Available: https://www.kaggle.com/datasets/quadeer15sh/augmented-forest-segmentation?resource=download. [Accessed September 09, 2022].

[22] V. Slyusar, “Multimodal quasi-fractal neural networks,” in *Int. Conf. Neural network technologies and their applications*, Kramatorsk, 2021, pp. 134-137.

[23] V. Slyusar, “Architectural and mathematical foundations of improving neural networks for image classification,” *Artificial Intelligence*, 2022, no .1, pp. 127-138. DOI: 10.15407/jai2022.01.127.

[24] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. arXiv preprint, arXiv:1409.0473, 2014, 15 p.