Semantic Segmentation for Moon Rock Recognition Using U-Net with Pyramid-Pooling-Based SE Attention Blocks

Antoni Jaszcz[a] and Dawid Połap[b]
Faculty of Applied Mathematics, Silesian University of Technology,
Kaszubska 23, 44-100 Gliwice, Poland

Keywords: Segmentation, U-Net, Semantic Analysis, Image Processing, Moonstones.

Abstract: Analysis of data from the rover's camera is an important element in the proper operation of unmanned vehicles. This is important because of the ability to move, avoid obstacles and even collect samples. In this paper, we propose a new U-Net architecture for rock/boulder recognition on the surface of the moon. For this purpose, architecture is composed of Squeeze and Excitation blocks extended with Pyramid Pooling and Convolution. As a result, such a network can pay attention to individual channels and give them weights based on global data. Moreover, the network analyzes contextual information in terms of local/global features in individual channels which allows for more accurate object segmentation. The proposed solution was tested on a publicly available database, achieving an accuracy of 97.23% and IoU of 0.7905.

SCITEPRESS SCIENCE AND TECHNOLOGY PUBLICATIONS

1 INTRODUCTION

Analysis of the moon and other planets is made with rovers, which are often operated remotely or even autonomously (Liu et al., 2023b; Chen et al., 2023b). These are unmanned vehicles moving on wheels. They have various sensors installed for data analysis. An example is a camera that records images around the rover. This is an important issue from a practical point of view. The rover's movement will be based on moving on an unknown surface. The recorded image can enable obtaining information about the environment and, above all, the location of obstacles that may cause problems with movement or even damage.

The recorded image will most often include part of the surface and sky. Under ideal conditions, the surface will be flat, but it doesn't have to be. Various stones or rocks may appear and should be avoided while moving. During sample collection, small rocks can even be picked up by the rover. Hence, the analysis of the image taken from the camera should be based not only on the location of the stones but also on their properties. In computer science, analyzing images and processing them for precise location and shape is called segmentation (Wu and Castleman, 2023). The incoming image is processed and the output is an image with selected objects. If we only locate stones, the result will be a two-color image, where one color will be the background and the other will represent the found objects. When analyzing a larger number of classes, the located objects are marked with different colors due to other characteristic features (Qureshi et al., 2023).

Semantic segmentation is based on the analysis of various objects within a single class, which automatically makes it a much more difficult task than classic segmentation. The most popular solution in this area are U-Net networks (Chen et al., 2023a), which are based on the architecture of convolutional neural networks. The idea of processing consists of encoding and decoding, i.e. the image is downsampled, which extracts the most important image features while reducing the dimension. Then upsampling is performed which restores the original size. Of course, both mechanisms include layers that process images and reduce/enlarge the size. Additionally, other techniques are introduced, such as context concatenation, which allows for the analysis of various image features, or skip connections, which allow the transfer of information between layers. It is important to note here that there is no single architecture that would allow segmentation for each task. Hence, new models and techniques within these networks are constantly being modeled.

TransAttUnet (Chen et al., 2023a) proposes a segmentation tool based on transformers that use long-distance contextual dependencies. The authors pro-

[a] https://orcid.org/0000-0002-8997-0331
[b] https://orcid.org/0000-0003-1972-5979

Jaszcz, A. and Połap, D.
Semantic Segmentation for Moon Rock Recognition Using U-Net with Pyramid-Pooling-Based SE Attention Blocks.
DOI: 10.5220/0012424600003636
Paper published under CC license (CC BY-NC-ND 4.0)
In *Proceedings of the 16th International Conference on Agents and Artificial Intelligence (ICAART 2024) - Volume 3*, pages 965-971
ISBN: 978-989-758-680-4; ISSN: 2184-433X

Proceedings Copyright © 2024 by SCITEPRESS – Science and Technology Publications, Lda.

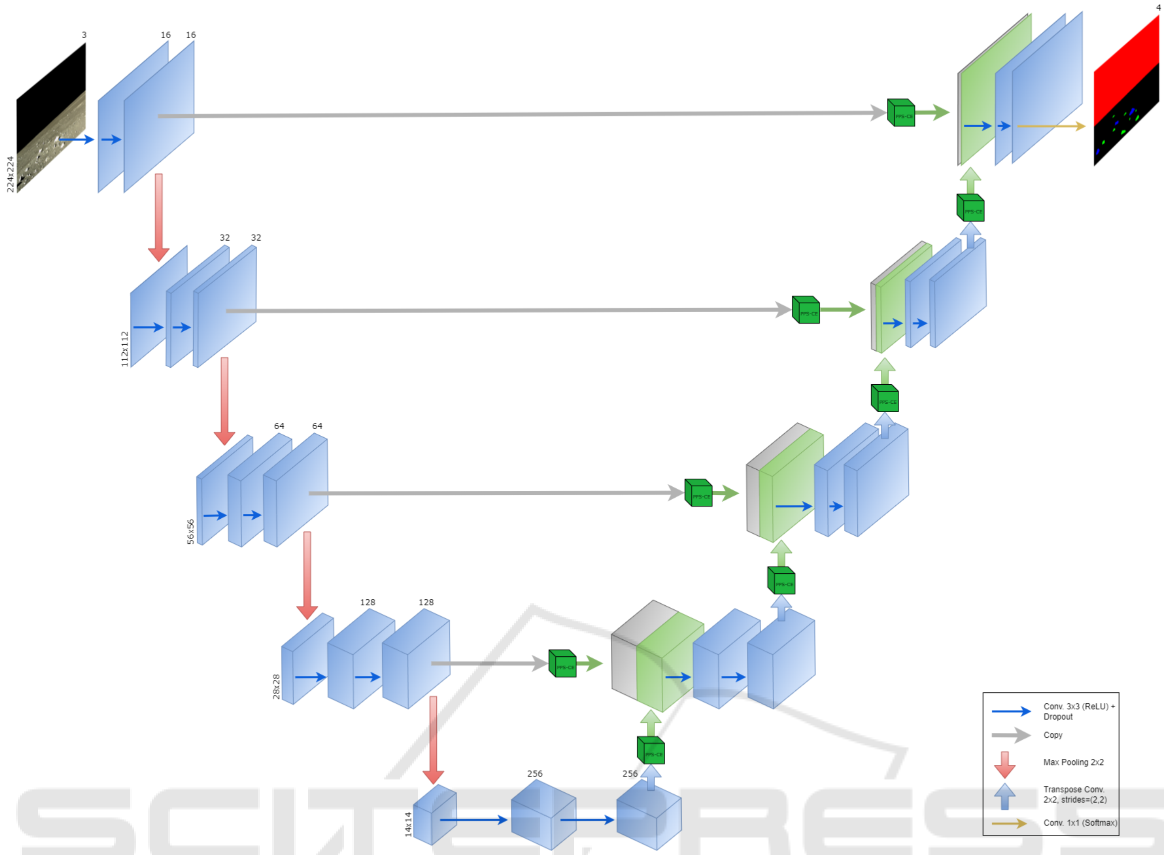


Figure 1: The proposed U-Net model with PPS-CE blocks.

SCITEPRESS
SCIENCE AND TECHNOLOGY PUBLICATIONS

posed this model for image segmentation, where various attention modules were implemented, including the spatial attention module. Another approach is to model an architecture based on transfer learning or even change the color model from RGB to LAB (Zhang and Zhang, 2023). An interesting solution is the fusion of thermal and visual images, as demonstrated by the U-Net segmentation problem (Shojaiee and Baleghi, 2023). The researcher combined the U-Net model with the Fused Atrous Spatial Pyramid Pooling encoder, i.e. the network is adapted to analyze such data through classical layers and atrous convolutional layers. Attention is also paid to the possibility of focusing the network's attention on the importance of selected regions in the image (Zhang et al., 2023). Augmentation is used for extending the datasets, but it can be also used as augmentation in the bottleneck of the u-net model, where attention-augmented convolution was introduced (Rajamani et al., 2023). Recent years have shown that attention modules are an important element of segmentation networks, an example of which is the modeling of new modules or implementation in specific places in the network. Position attention module can be used for feature enhancement (Jiang et al., 2023). Attention allows to direction of the network to specific features or elements, which is crucial when modeling an architecture tailored to a specific problem.

Segmentation analysis of stones was undertaken by building a segmentation network that uses a pretrained model named VGG16 (Li et al., 2023). The research included analysis with other segmentation methods, although neural networks allow for much more accurate results. Segmentation analysis on Mars was processed by the U-Net network with a feature enhancement module and window transformer block (Xiong et al., 2023). This model allowed for the analysis of features at various scales. Again, (Pan et al., 2023) focused on tiny and fracture features. An additional technique was the use of dilated convolution, which focuses on a much larger number of pixels during processing. The results of the research showed that the method can allow for feature extraction even with a complex background. Another solution is the model that will process long-range spatial context (Liu et al., 2023a), which was achieved by introducing a feature refining module between the encoder and decoder.

Based on literature analyses, attention can be drawn to the need for newer models that will enable better data segmentation. In this work, we propose a new solution based on the U-Net network model, including the squeeze and excitation mechanism, which enables the analysis of dependencies between features in feature maps. Additionally, we introduce Pyramid-Pooling to these blocks to take into account information from different scales or sizes of objects and to increase the importance of image context for analysis of the moon's surface. The main contributions of this research are:

- a new U-net model for boulder/rock segmentation,
- a novel block type that combines Pyramid-Pooling with Squeeze and Excitation.

2 METHODOLOGY

In this section, we propose a modified U-Net model enhanced with PPS-CE blocks for multi-class semantic segmentation tasks. The overview illustration of the model is presented in Fig. 1. The contraction path consists of 5 doubled 3 × 3 convolutional layers with (ReLU activation functions) and dropout between them. These are followed by 2 × 2 MaxPooling layers. In the expansive path, to enhance the performance of the model, we propose utilizing PPS-SE blocks after each transpose convolution and copy path. The final output is obtained using 1 × 1 convolution with Softmax activation function (to obtain the probabilistic distribution of the classes).

2.1 Pyramid Pooling

Pyramid pooling is a unique pooling technique that allows the model to gather more contextual information by capturing information at multiple scales. The principle of this method is based on dividing the input feature map into regions of different sizes. Then, for each divided feature map obtained this way, average pooling is performed. The result of each pooling segment is then concatenated, creating a unified representation that carries multi-scale information. Mathematically, this can be presented as processing the input feature map $X = (x_{h,w,c})$ where values h, w, c are accordingly height, width and number of channels of the feature map X. Given the set of scales L, for each l scale in the set, a divided feature map is created according to the following equation:

$$X_l = (x_{h_1,w_1,c}), \tag{1}$$

where $h_1 = \left\lfloor \frac{h}{l} \right\rfloor$ and $w_1 = \left\lfloor \frac{w}{l} \right\rfloor$. For each obtained X_l, the average pooling operation is performed. Lastly, the pooling results at all scales are concatenated, resulting in the final feature map Y, whose dimensionality is presented as:

$$dim(Y) = \left(\sum_{i=1}^{|L|} \left\lfloor \frac{h}{l_i} \right\rfloor \times \sum_{i=1}^{|L|} \left\lfloor \frac{w}{l_i} \right\rfloor \times c \right). \tag{2}$$

As previously mentioned, pyramid-pooling allows the model to gather extended contextual information by utilizing many different perception field sizes. This provides better robustness regarding object scale, which is especially important in semantic segmentation tasks.

2.2 Pyramid-Pooling Squeeze and Convolutional Excitation Blocks

Squeeze-and-excitation (SE) blocks are a mechanism that improves the representational power of the convolutional layers by analyzing the dependencies between various channels in feature maps passed from the convolutional layer and assigning them weights based on the impact they have on the further assessment of the model. This is one of many types of attention mechanisms used in neural networks, highlighting the more influential channels, while also suppressing less informative ones. This process improves the overall feature representation. The basic SE block first performs average global pooling as the squeeze operation, obtaining $1 \times 1 \times c$ (c indicating the number of channels in the input feature map) vector. In the excitation operation, the vector is then passed onto two dense layers with the former having ReLU (introducing non-linearity) and the ladder having a Sigmoid activation function. The output of these layers is then scaled and applied to the original feature map. In this paper, we propose Pyramid-Pooling Squeeze and Convolutional Excitation blocks (PPS-CE), utilizing Pyramid-Pooling in Squeeze operation and double 1×1 convolution instead of dense layers in Excitation operation. The main advantages of this approach are the benefits of using Pyramid-Pooling and convolutional layers having less trainable parameters than dense layers. In squeeze operation, each divided X_l feature map is processed using 1×1 convolution with ReLU activation function. In this paper, we propose that each convolution has the number of output channels equal to $c_{conv} = \lfloor \frac{c}{r} \rfloor$, with r parameter set to 16. The output of each convolutional layer is then concatenated along the channel axis. Next, concatenated feature maps from Pyramid Pooling are passed through two 1×1 convolutional layers, the first of

SCIENCE AND TECHNOLOGY PUBLICATIONS

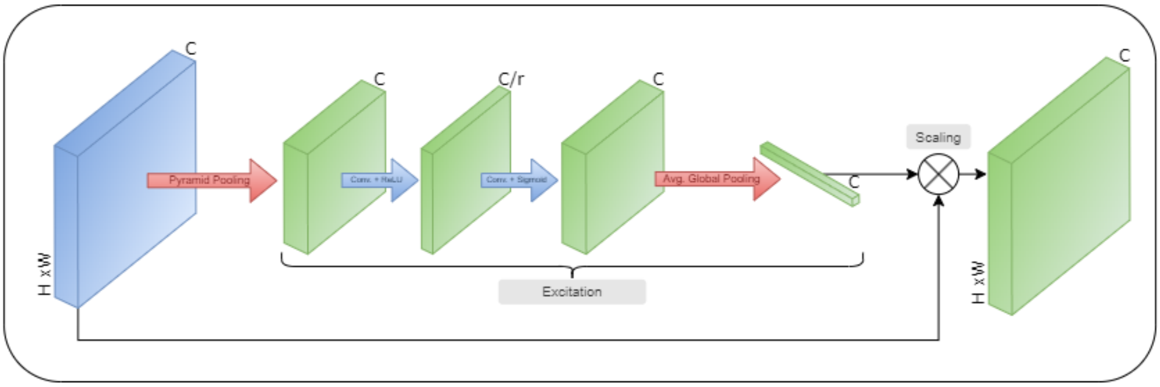


Figure 2: The proposed PPS-CE blocks overview.

which has the number of channels equal to c_{conv} and ReLU activation function, while the other has c channels (same as the input) and Sigmoid activation function. Ending the excitation operation, the global average pooling is applied, ensuring that its shape matches the shape of the inputs and then reshaped into a vector of length c. Lastly, the input tensor is multiplied element-wise by the reshaped output of the excitation operation. The above-described PPS-CE blocks are presented visually in Fig. 2.

2.3 Loss Function

The proposed loss function is based on The Dice coefficient and categorical cross-entropy. The Dice coefficient is a statistic used to measure the similarity between two sets. Given a ground-truth segmentation mask Y and predicted segmentation mask Y' with C classes, a Dice coefficient for i-th class can be calculated as:

$$Dice_i = \frac{2|Y_i' \cap Y_i|}{|Y'| + |Y|}. \tag{3}$$

Then, Dice loss is calculated by the following formula:

$$\begin{aligned} DiceLoss_i &= 1 - Dice_i, \\ DiceLoss &= \frac{1}{C}\sum_{i=1}^{C} DiceLoss_i. \end{aligned} \tag{4}$$

The second one is categorical cross-entropy (CCE), a reliable function loss for multi-class classification tasks, including semantic segmentation. For each corresponding pair of pixels y and y' in the ground truth and predicted masks, CCE is calculated according to the following formula:

$$CCE = -\sum_{i=i}^{C} y_i \cdot log(y_i'). \tag{5}$$

It is worth mentioning, that the labels must be encoded using the one-hot-encoding technique, to resemble probabilistic distribution, for the CCE to work properly. The final loss for classification is the mean CCE of all pixels in the image and $DiceLoss$. This can be described as:

$$L = CCE + DiceLoss. \tag{6}$$

3 EXPERIMENTS

In this section, we describe the dataset used in the experiments and show the results. In the context of evaluation, classic measures such as accuracy, Dice coefficient and IoU were selected.

3.1 Moon Rocks Dataset and Experimental Settings

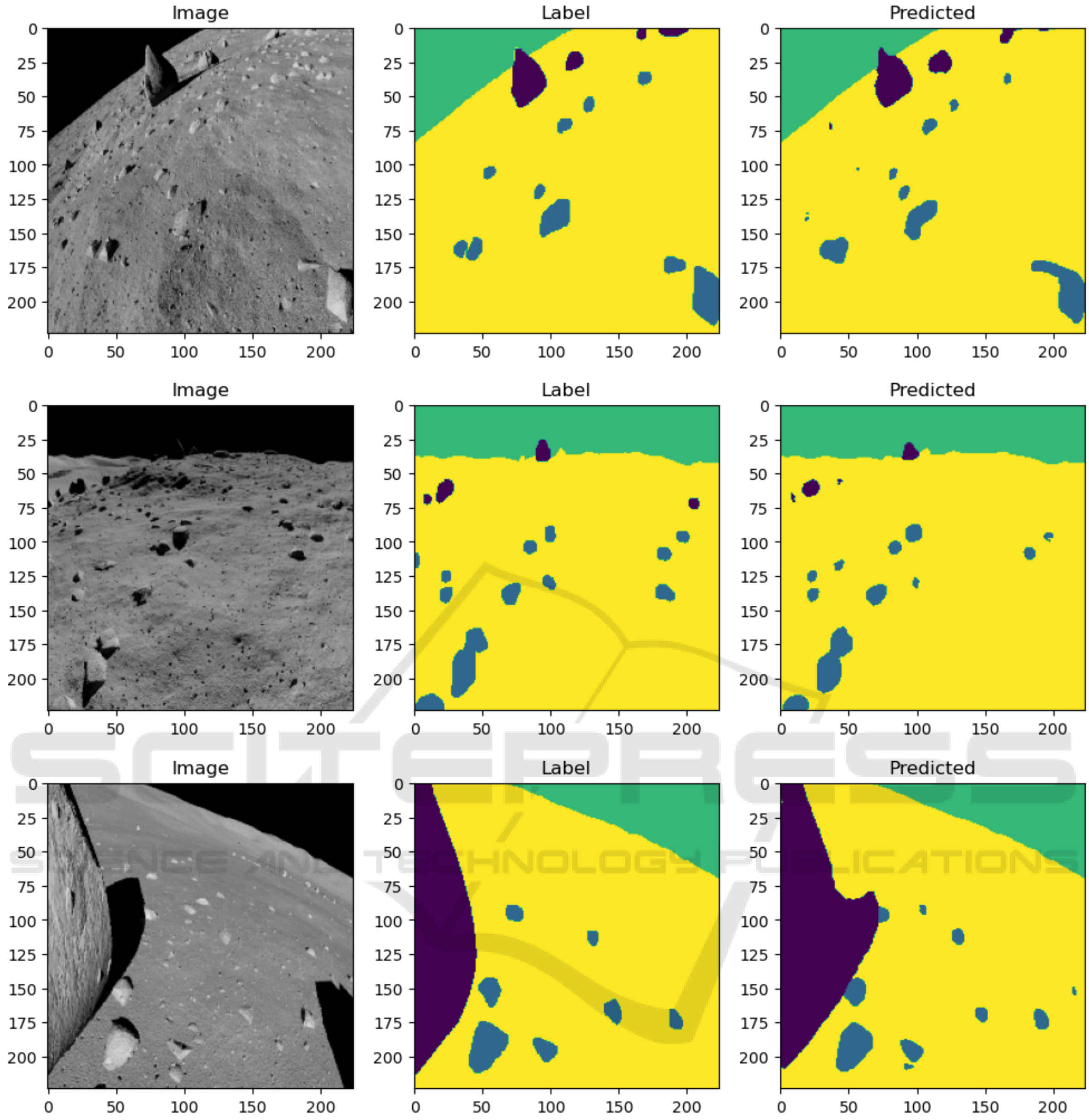
The data used for the training and validation of the model are available publicly at Kaggle. The dataset consists of nearly 10 thousand artificially created lunar landscape RGB images along with the corresponding segmentation masks. The ground truth images from the realistic renders were created using Planetside Software's Terragen. The dataset consists of 4 classes: sky, ground, small rocks and large rocks.

The training, test and validation sets were created in a ratio of 90:5:5. The training set consisted of 8790 images, while the test and validation ones had 488 samples each. Before training, each sample was resized to 224x244 pixels. The model was trained for 50 epochs using mini-batches consisting of 32 samples. As a training algorithm, ADAM was selected with the loss function described in Eq. (6).

3.2 Results

The graph of training the network for 50 epochs is presented in Fig. 4. It shows that the model's accuracy quickly increased to over 94%. Exceeding

SCITEPRESS SCIENCE AND TECHNOLOGY PUBLICATIONS



SCITEPRESS
SCIENCE AND TECHNOLOGY PUBLICATIONS

Figure 3: Sample data with masks and the result returned by the proposed network model.

96% was possible after 13 training iterations. With subsequent epochs, the accuracy was distributed on a training and testing basis. There were small spikes in accuracy during the training process, but they were within 1% point. After epoch 45, accuracy increased linearly, reaching an accuracy of 97.23%. As part of the analysis, other evaluation coefficients were also determined, such as the Dice coefficient, which was 0.8763. This is a measure that determines the quality of image segmentation, the closer to 1, the more accurate the result. The obtained result of 0.8763 indicates that the algorithm maps objects to masks very well, and the differences are minimal. In addition, small objects are also detected by the network, which is an important advantage of the proposed approach due to its potential practical application. The IoU (Intersection over Union) on the validation basis reached 0.7905. This result shows that the segmentation relative to the masks is quite accurate, although there are small areas where the mask does not match the segmentation result. The reason may be shadows or details of objects. Fig. 5 shows the loss value during the training process. The values decrease with subsequent epochs with single value jumps be-

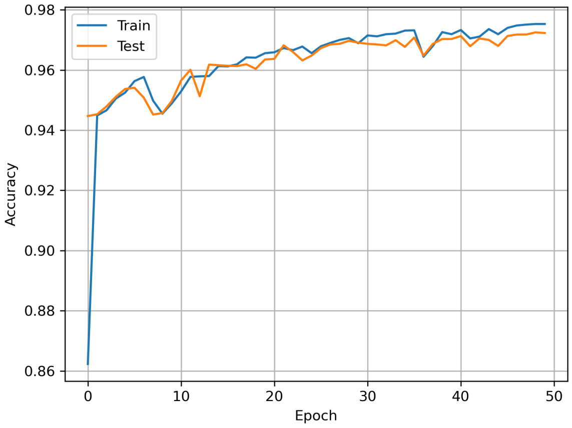


Figure 4: Accuracy plot on the training and test sets during training.

low 0.05 (which is visible primarily after the 35th training epoch).

The original images with masks and segmentation results by the proposed method are shown in Fig. 3. On the presented samples, it can be seen that the network processed a shadow that was not on the mask, or very small stones did not always appear. It is worth noting that in the first row of images, the network detected stones that are in the image but not in the original mask. This shows that the masks themselves are not perfect either. In (Fan et al., 2023), the authors presented the possibility of Combining a Convolutional Neural Network (based on ResNet) and Transformer, where the network achieved an IoU of 78.90% while having nearly 8 million parameters. It should be noted that the model proposed in this paper is based on the extraction of features other than classical solutions. During the analysis, we noticed that the introduced SaE with Pyramid-Pooling blocks allowed for quick achievement of good results, which were improved with the increase in the number of epochs while keeping the number of trained parameters relatively low – less than 2 million.

4 CONCLUSION

Analyzing data from the rover's camera is one of the basic elements when moving to avoid hitting an obstacle. In this work, we proposed a new U-Net model architecture for semantic image segmentation. This operation enables the segmentation of stones with high accuracy, which was 97.23% and an IoU coefficient of 0.7905. The results were made possible by introducing blocks based on the Squeeze and Excitation technique combined with Pyramid Pooling into the U-Net network. As a consequence, this action allowed the network to analyze individual channels and assign them weights. Attention should also be paid to the possibility of analyzing contextual information by processing and focusing on local and global features.

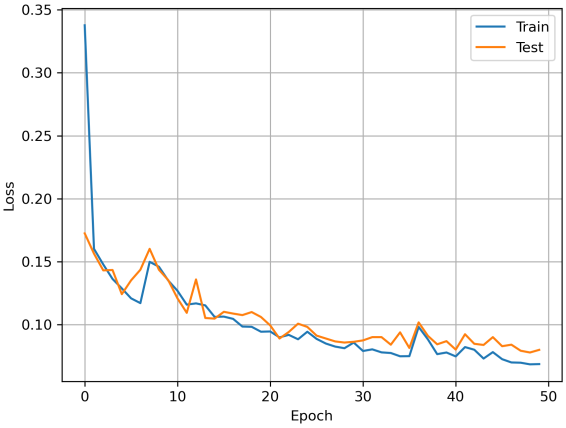


Figure 5: Loss value plot on the training and test sets during training.

In future work, we plan to analyze the possibility of extending the network model to spatial attention modules, which could allow for the analysis of additional features.

ACKNOWLEDGEMENTS

This work was supported by the Rector's mentoring project ”Spread your wings” at the Silesian University of Technology.

REFERENCES

Chen, B., Liu, Y., Zhang, Z., Lu, G., and Kong, A. W. K. (2023a). Transattunet: Multi-level attention-guided u-net with transformer for medical image segmentation. *IEEE Transactions on Emerging Topics in Computational Intelligence*.

Chen, Z., Zou, M., Pan, D., Chen, L., Liu, Y., Yuan, B., and Zhang, Q. (2023b). Study on climbing strategy and analysis of mars rover. *Journal of Field Robotics*.

Fan, L., Yuan, J., Niu, X., Zha, K., and Ma, W. (2023). Rockseg: A novel semantic segmentation network based on a hybrid framework combining a convolutional neural network and transformer for deep space rock images. *Remote Sensing*, 15(16).

Jiang, J., Feng, X., Ye, Q., Hu, Z., Gu, Z., and Huang, H. (2023). Semantic segmentation of remote sensing images combined with attention mechanism and feature enhancement u-net. *International Journal of Remote Sensing*, 44(19):6219–6232.

Li, M., Zhang, P., and Hai, T. (2023). Pore extraction method of rock thin section based on attention u-net. In *Journal of Physics: Conference Series*, volume 2467, page 012016. IOP Publishing.

Liu, H., Yao, M., Xiao, X., and Xiong, Y. (2023a). Rockformer: A u-shaped transformer network for martian rock segmentation. *IEEE Transactions on Geoscience and Remote Sensing*, 61:1–16.

Liu, S., Su, Y., Zhou, B., Dai, S., Yan, W., Li, Y., Zhang, Z., Du, W., and Li, C. (2023b). Data pre-processing and signal analysis of tianwen-1 rover penetrating radar. *Remote Sensing*, 15(4):966.

Pan, D., Li, Y., Lin, C., Wang, X., and Xu, Z. (2023). Intelligent rock fracture identification based on image semantic segmentation: methodology and application. *Environmental Earth Sciences*, 82(3):71.

Qureshi, I., Yan, J., Abbas, Q., Shaheed, K., Riaz, A. B., Wahid, A., Khan, M. W. J., and Szczuko, P. (2023). Medical image segmentation using deep semantic-based methods: A review of techniques, applications and emerging trends. *Information Fusion*, 90:316–352.

Rajamani, K. T., Rani, P., Siebert, H., ElagiriRamalingam, R., and Heinrich, M. P. (2023). Attention-augmented u-net (aa-u-net) for semantic segmentation. *Signal, image and video processing*, 17(4):981–989.

Shojaiee, F. and Baleghi, Y. (2023). Efaspp u-net for semantic segmentation of night traffic scenes using fusion of visible and thermal images. *Engineering Applications of Artificial Intelligence*, 117:105627.

Wu, Q. and Castleman, K. R. (2023). Image segmentation. In *Microscope Image Processing*, pages 119–152. Elsevier.

Xiong, Y., Xiao, X., Yao, M., Liu, H., Yang, H., and Fu, Y. (2023). Marsformer: Martian rock semantic segmentation with transformer. *IEEE Transactions on Geoscience and Remote Sensing*.

Zhang, J., Qin, Q., Ye, Q., and Ruan, T. (2023). St-unet: Swin transformer boosted u-net with cross-layer feature enhancement for medical image segmentation. *Computers in Biology and Medicine*, 153:106516.

Zhang, S. and Zhang, C. (2023). Modified u-net for plant diseased leaf image segmentation. *Computers and Electronics in Agriculture*, 204:107511.

SCITEPRESS SCIENCE AND TECHNOLOGY PUBLICATIONS