The Segmentation of Unnamed Aerial Vehicle-Derived Aerial Photographs to Identify Anthropogenic Changes

Sergey Buzmakov
Perm State University
Perm, Russia
e-mail: lep@psu.ru

Leonid Kuchin
Perm State University
Perm, Russia
e-mail: Kleond@bk.ru

Nikita Permyakov
HSE University
Perm, Russia
e-mail: napermyakov@edu.hse.ru

Elena Zamyatina
HSE University
Perm, Russia
e-mail: ezamyatina@hse.ru

***Abstract* — The paper considers the issues of segmentation of aerial photographs obtained by Unnamed Aerial Vehicles to identify man-made changes. Neural networks are used for this purpose. Based on the orthophotoplan of the oil field area, a dataset was formed of about 4,500 tiles measuring 512 × 512 pixels with 18 classes of man-made zones marked. The images were filtered, balanced, and augmented, after which U-Net, DeepLabv3+, and SegFormer models were trained on them with the training, test, and validation sets divided into 70/15/15%. The best modification of U-Net showed an overall accuracy of 94.4% and mean Intersection over Union of 79.2%, while Intersection over Union values above 80% were obtained for key natural and man-made objects. DeepLabv3+ and SegFormer demonstrated comparable results (mean Intersection over Union about 74%) with better detail for large and rare classes. The proposed method ensures high accuracy and efficiency of analysis, which makes it promising for environmental monitoring.**

***Keywords* — Environmental monitoring, man-made changes, neural networks, segmentation, orthophotoplan, fully convolutional networks, U-net, DeepLabv3+, SegFormer.**

I. Introduction

Industrial development, particularly in the oil industry, has a negative impact on nature. So, there is a need for systematic and prompt environmental monitoring to identify man-made changes in natural objects. Traditional observation methods based on laboratory analysis of soil, water, and air often do not allow for prompt recording of the dynamics of object disturbances or require significant investment of expert time or computing power [1].

One effective solution to this problem is the use of unmanned aerial vehicles (UAVs), which provide prompt and high-precision aerial photography of territories [1, 2].

A research group at Perm State University has developed a comprehensive methodology for analyzing the effects of oil field impacts on natural objects using UAVs, image decoding, and field research, as it is reported [1]. However, there is a need for manual interpretation of images, which requires significant time on the one hand, and the high qualifications of specialists (experts) on the other. Due to the above limitations, more and more attention is paid to the development of tools based on automatic image processing methods.

The article discusses the use of deep learning technologies for analyzing aerial photographs. This solution allows us to speed up data analysis and reduce dependence on subjective factors, which is especially important in the context of analyzing large volumes of images.

So, below we will consider the following neural networks commonly used for semantic segmentation and compare the efficiency of their application for identifying man-made changes in aerial photographs obtained using UAVs: (a) U-Net; (b) DeepLabv3+; (c) SegFormer.

The initial data (aerial photographs of a site with signs of man-made impact) were provided by a research group of Perm State University.

II. Automated Recognition of Techogenic Sites

Deep neural networks are currently a frequently used tool for visual information processing. In areas of computer vision such as image classification, object detection, and segmentation, deep learning methods have demonstrated quality comparable to or superior to human capabilities, especially on large volumes of data. When applied to aerial photographs and remote sensing data, neural networks allow for automated interpretation, eliminating the need for manual labeling of each new image [2]. For environmental monitoring tasks using UAVs, semantic segmentation methods are of greatest interest.

Semantic segmentation is an image classification where each pixel is assigned to a specific class [3, 4]. The result is a mask (or map) of the same dimensions as the original image, where each pixel is labeled, for example, as “technogenic object” or “background”. Unlike object detection, where bounding boxes are allocated around objects, segmentation clearly outlines the shape of objects, which is important for assessing the area of technogenic impact and the exact boundaries of pollution.

Neural network architectures, originally developed for medical imaging and urban video surveillance, have been adapted for satellite imagery. In the case of aerial photographs obtained by UAVs, we are dealing with extremely high-resolution images, where a single photograph may contain many small details, unlike satellite images. In this case, the neural network must process a large amount of data, which

https://doi.org/10.51408/csit2025_51

places high demands on the computing power of the computer and on the algorithms used. In addition, to train the neural network, the data must be labeled (each pixel in the image must be assigned to the correct class).

To solve the problem, we will choose Fully Convolutional Networks (FCN) [5,6]. The structure of such networks is known as encoder-decoder architecture or "convolutional autoencoder"; neural networks with this architecture have shown good results in solving semantic segmentation problems. Next, we will consider: U-Net [6, 7, 8], DeepLabv3+ [6, 7, 9], and SegFormer [6, 7, 10].

Let's consider some issues that were solved when working with aerial photographs obtained using UAVs.

A research group of environmental scientists presented an orthophoto as input data, which has a size of 10,000 × 10,000 pixels. For these input data to be processed using a neural network, it is necessary to split them into fragments (tiles) of 512 × 512 pixels. Each fragment is processed separately, and then the results are combined (stitched). Recognized objects can fall on the border of a fragment, which negatively affects the reliability of the results of recognizing man-made areas. In this case, the fragment sizes should be such that the object is entirely within the fragment.

In addition, annotating aerial photography is a labor-intensive process. Semi-automatic methods are often used to obtain a training sample: for example, pre-clustering the image and then making adjustments manually. In this work, the data were labeled manually by an expert based on visual analysis in the QGIS geoinformation program, which ensures high accuracy, but limits the amount of data for training.

This must be considered when choosing a model. It is advisable to give preference to architectures that can effectively learn from small samples. And finally: the quality of segmentation is assessed by several metrics - Intersection over Union (IoU) (or the Jaccard coefficient) [6] and the average intersection of all classes - mean IoU (mIoU)

The data were obtained at oil fields as a result of aerial photography by a Russian-made Supercam S350F UAV and represent an orthophotomap. The survey was carried out at an altitude of about 400 meters with a resolution of 8.5 cm/pixel [1].

Additionally, vector layers with polygonal markings of areas containing signs of mechanogenesis, bitumenization, and halogenesis were provided. The images were manually marked by environmental specialists using ArcGIS software based on the interpretation of aerial photographs and field surveys. An example of the original orthophotomap image and its markings are shown in Figures 1 and 2.

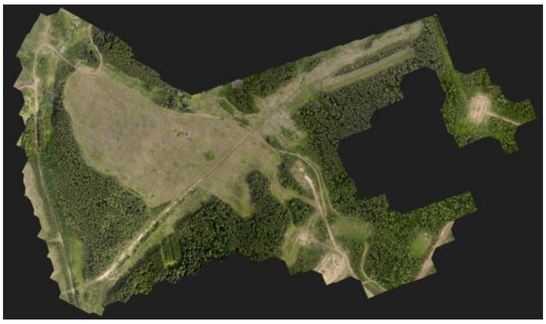

Fig.1. Fragment of the orthophotoplan of the studied deposit

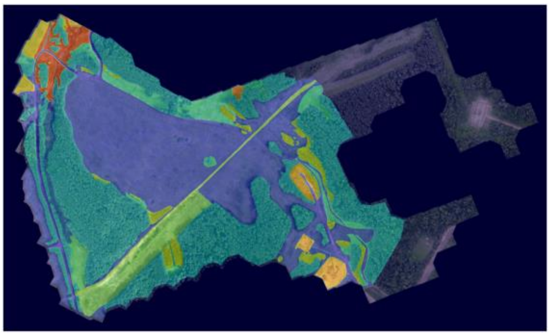

Fig.2. Fragment of the orthophotoplan with markings

Objects in the labeled images can be assigned to one of 17 classes: (a) swamps, (b) roads, (c) overgrown fields, etc. Some classes are insufficiently represented in the labeled images, so it was necessary to augment the data. When preparing the data for training neural networks, the following was performed: (a) filtering of empty fragments (excluding areas where useful information is missing or background prevails); (b) class balancing (control of the number of pixels of each class in the sample for correct training); (c) dividing the sample into training (70%), validation (15%) and test (15%) parts while maintaining the proportions of the classes; (d) data augmentation — artificial increase in the sample by applying image transformations (horizontal and vertical reflections, random 90° rotations, changes in brightness and contrast, adding noise). An example of the application of augmentation is shown in Fig. 3.

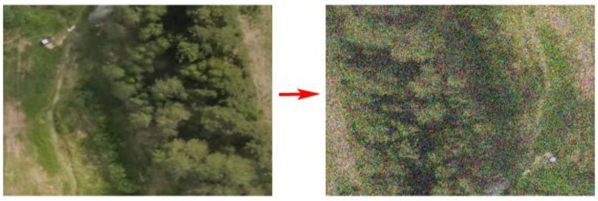

Fig. 3. Example of the use of augmentation: the original image (left) and the image after vertical reflection, rotation, brightness changes and additional noise

When preparing the data, we considered the fact that the data related to different classes are not balanced. For example, areas of mechanical damage were encountered much more often than traces of bitumenization or halogenesis. To compensate for this disproportion, selective augmentation of fragments with rare classes was used.

As a result: (a) the total number of fragments is 4500; (b) the fragment sizes are 512×512 pixels; and a balanced number of areas of each class. An example of a prepared fragment and a fragment with a mask are shown in Fig. 4.

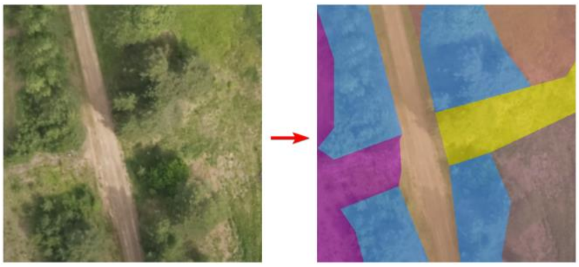

Fig. 4. Example of the prepared image fragment and the fragment with a mask

The PyTorch library [11] was chosen as the main library for creating neural networks. As already mentioned, the following neural networks were selected to solve the problem: U-Net, DeepLabv3+, and SegFormerTiny. Their training is performed according to the following scenario: (a) model initialization depending on the selected architecture; (b) preparation of datasets and data loaders (DataLoader) with the ability to implement class balancing; (c) selection of the loss function (classical weighted cross-entropy, where the class weights are calculated based on the number of pixels of each class, or Focal Loss for more accurate work with rare classes); (d) initialization of the optimizer [10] and learning rate planner (AdamW [12] and CosineAnnealingLR); (d) support for automatic mixed learning (AMP) to speed up training and reduce video memory consumption [13]; (e) support for exponential smoothing of model weights (EMA) to improve the quality of the final results [5, 14]; (g) implementation of the Early Stopping procedure in the absence of improvements in quality metrics when determining validation [15, 16].

To evaluate the quality of the developed models, experimental studies were conducted on the generated dataset of aerial photographs divided into fragments of 512×512 pixels. Each model was trained and tested under the same conditions.

Experimental conditions: (a) fragment size 512×512 pixels; (b) batch size - 8 images for U-Net, DeepLabV3+, 4 images for SegFormerTiny; (c) number of epochs - 100; (d) optimizer: AdamW; (e) scheme for changing the learning rate: cosine attenuation (CosineAnnealingLR); (e) loss function - weighted cross-entropy.

Recognition quality was assessed on a validation sample (~15% of the entire dataset) without the use of augmentations. The following metrics were used to quantitatively assess the quality of models used for semantic segmentation: (a) mean Intersection over Union (mIoU); — the main metric for assessing the quality of segmentation; (b) Per-class IoU — to analyze the quality of recognition of objects of different types; training time of one epoch — as an indicator of the speed of the models.

All models were trained on the same computing platform using GPU acceleration (NVIDIA RTX 4070).

During the experiments, the Early Stopping strategy [15, 16] was used in the absence of growth in the mIoU metric over 10 epochs. Let us consider the results of the experiments in more detail.

A. *U-Net model results*

The U-Net model demonstrated high overall classification accuracy: (a) Overall Accuracy: 94.38%; (b) Mean IoU: 79.15%.

Based on the model evaluation results, the following conclusions can be made: high IoU values were achieved for most key classes: (a) forests (class 6): IoU = 93.42%; (b) overgrown fields (class 3): IoU = 91.61%; (c) ponds (class 14): IoU = 89.63%; power lines (class 9): IoU = 85.50%. The best segmentation was observed for large and clearly defined objects (forests, overgrown fields). Difficulties arose when segmenting classes with low representation in the sample, for example: technical objects (class 5): IoU = 0.00% (not a single object was recognized), transformed rivers (class 18): IoU = 57.84%. Thus, U-Net showed excellent results on the main classes, but problems arose with rare and small objects.

B. *DeepLabV3+ model results*

DeepLabV3+, built on the modified ResNet-50, demonstrated slightly lower performance: (a) Overall Accuracy: 93.25%; (b) Mean IoU: 74.24%. Let's take a closer look at the experimental results. The model demonstrated good segmentation quality for large objects: (a) forests (class 6): IoU = 92.73%; (b) overgrown fields (class 3): IoU = 90.24%. Unfortunately, for some classes, the recognition accuracy is low: (a) swamps (class 1): IoU = 51.87%; (b) transformed rivers (class 18): IoU = 43.96%; (c) technical objects (class 5): IoU = 0.00%.

It is worth noting that DeepLabV3+ showed less robustness on classes with a small number of examples in the training set, despite high overall accuracy on large objects.

C. *SegFormer model results*

The SegFormer model demonstrated the following results: (a) Overall Accuracy: 87.12%; (b) Mean IoU: 73.57%. Even though the model demonstrated the lowest overall accuracy among all three models, it demonstrates high quality for individual classes: (a) Floodplain meadows (class 8): IoU = 87.85%; (b) Power transmission lines (class 9): IoU = 90.55%; (c) Fields (class 12): IoU = 94.75%; (d) Man-made ponds (class 17): IoU = 85.69%.

At the same time, the model was completely unable to correctly identify the background (class 0) IoU = 0.00%. This may be due to the lack of an explicit representation of "absence of an object" in the logic of the transformer architecture or an error in preprocessing. A comparative analysis of the models is presented in Table 1.

Table 1

Metric	U-Net	DeepLabV3+	SegFormer
Overall Accuracy	94.38%	93.25%	87.12%
Mean IoU	79.15%	74.24%	73.57 %
IoU (class)	Forests (93.4%)	Forests (92.7%)	Fields (94.75%)

Based on the comparison results, the following conclusions can be drawn: (a) U-Net demonstrated better values for overall accuracy and average IoU metrics; (b) DeepLabV3+ consistently recognized large elements, but performed worse with small ones; (c) SegFormer demonstrated competitive recognition quality for important small ones on important small classes (man-made objects, ponds), but did not recognize the background.

III. Conclusion

Thus, a system for automated analysis of aerial photographs obtained by UAVs was developed to identify signs of man-made transformation of the natural environment.

Based on the experimental results, it was found that the U-Net model demonstrated the best results in segmentation quality among the tested models, achieving an average mIoU metric of about 79%. The DeepLabV3+ model showed good results on large objects, but had difficulty segmenting rare and small objects. The SegFormerTiny model, despite the lowest overall accuracy (87.1%), showed high accuracy on rare and small classes, such as man-made ponds, fields, power lines, and meadows in floodplains.

Further research suggests increasing the amount of data for training using GAN [5, 6, 17] (generative adversarial networks) to generate new images.

Funding

The study was supported by a grant from the Russian Science Foundation and Perm Krai No. 24-17-20025, https://rscf.ru/project/24-17-20025/.

Conflict Of Interests

The authors of this work declare that they have no conflicts of interest.

Compliance With Ethical Standarts

This work does not contain any studies involving human and animal subjects.

References

[1] S. A. Buzmakov, P. Yu. Sannikov, L. S. Kuchin, E. A. Igosheva and I. F. Abdulmanova, “Application of unmanned aerial photography for diagnostics of technogenic transformation of the natural environment during operation of an oil field”, Zapiski Gornogo Instituta, vol. 260, pp. 180–193, 2023, DOI: 10.31897/PMI.2023.22.

[2] R. Eskandari, M. Mediapart, F. Mohammadimanesh, B. Salehi, B. Brisco and S. Homayouni, “Meta-analysis of unmanned aerial vehicle (UAV) imagery for agro-environmental monitoring using machine learning and statistical models”, Remote Sensing, vol. 12, no. 21, Article 3511, 2020, DOI: 10.3390/rs12213511.

[3] X. Liu, Z. Deng and Y. Yang, “Recent progress in semantic image segmentation”, *Artificial Intelligence Review*, vol. 52, pp. 1089–1106, 2018, DOI: 10.1007/s10462-018-9641-3.

[4] Y. Mo, Y. Wu, X. Yang, F. Liu and Y. Liao, “Review the state-of-the-art technologies of semantic segmentation based on deep learning”, *Neurocomputing*, vol. 493, pp. 626–646, 2022, DOI: 10.1016/j.neucom.2022.01.005.

[5] S. Nikolenko, A. Kadurin and E. Arkhangelskaya, *Deep Learning*. St. Petersburg, Russia: Piter, 2018.

[6] N. Shakla, *Machine Learning and TensorFlow*. St. Petersburg, Russia: Piter, 2019.

[7] D. Joshi and C. Witharana, “Vision transformer-based unhealthy tree crown detection in mixed northeastern US forests and evaluation of annotation uncertainty”, *Remote Sensing*, vol. 17, Article 1066, 2025, DOI: 10.3390/rs17061066.

[8] K. Shah, “U-Net: Convolutional Networks for Biomedical Image Segmentation”, Medium, ProjectPro, 2021. [Online]. Available: https://medium.com/projectpro

[9] S.-H. Tsang, “Review: DeepLabv3+ — Atrous Separable Convolution (Semantic Segmentation)”, Medium, 2019. [Online]. Available: https://medium.com

[10] E. Xie, W. Wang, Z. Yu, A. Anandkumar, J. M. Alvarez and P. Luo, “SegFormer: Simple and efficient design for semantic segmentation with transformers”, *Advances in Neural Information Processing Systems (NeurIPS 2021)*, vol. 34, 2021.

[11] O.-C. Novac, M. C. Chirodea, C. M. Novac, N. Bizon, M. Oproescu, O. P. Stan and C. E. Gordan, “Analysis of the application efficiency of TensorFlow and PyTorch in convolutional neural network”, *Sensors*, vol. 22, Article 8872, 2022, DOI: 10.3390/s22228872.

[12] P. Zhou, X. Xie, Z. Lin and S. Yan, “Towards understanding convergence and generalization of AdamW”, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 46, pp. 6486–6493, 2024, DOI: 10.1109/TPAMI.2024.3382294.

[13] Y. Yang, X. Xing, M. Chen, K. Guo and X. Xu, “Adaptive mixed-precision networks”, SSRN, 2022. [Online]. Available: https://ssrn.com/abstract=4274178

[14] D. Morales Brotons, T. Vogels and H. Hendrikx, “Exponential moving average of weights in deep learning: dynamics and benefits”, *Transactions on Machine Learning Research*, pp. 1–27, 2024, HAL ID: hal-04830859.

[15] L. Prechelt, “Early stopping — but when?”, *Neural Networks: Tricks of the Trade, Lecture Notes in Computer Science*, vol. 1524, Berlin, Germany: Springer, 1996, pp. 53–67, DOI: 10.1007/3-540-49430-8_3.

[16] L. Prechelt, “Automatic early stopping using cross validation: quantifying the criteria”, *Neural Networks*, vol. 11, no. 4, pp. 761–767, 1998, DOI: 10.1016/S0893-6080(98)00010-0.

[17] D. Foster, *Generative Deep Learning: The Creative Potential of Neural Networks*. St. Petersburg, Russia: Piter, 2020.