EN RU PROGRAM SYSTEMS: THEORY AND APPLICATIONS ISSN 2079-3316
Research Article APPLIED SOFTWARE SYSTEMS

UDC 004.932.72:004.89
10.25209/2079-3316-2025-16-1-3-44

Using the Mask R-CNN model for segmentation of real estate objects in aerial photographs

Igor Victorovich **Vinokurov**
Financial University under the Government of the Russian Federation, Moscow, Russia
✉*igvvinokurov@fa.ru*

Abstract. The mass appearance of illegal and unregistered in the Unified State Register of Real Estate (USRRE) real estate objects complicates cadastral registration for many entities at the territorial and administrative levels. Traditional methods of identifying objects of this type, based on manual analysis of geospatial data, are labor-intensive and time-consuming.

To improve the efficiency of this process, it is proposed to automate the detection of objects in aerial photographs by solving the instance segmentation problem using the Mask R-CNN deep learning model. The article describes the preparation of a dataset for this model, examines the main quality metrics, and analyzes the results obtained. The efficiency of the Mask R-CNN model in practice is shown for solving the problem of detecting construction projects that are not registered in the USRRE. (*Linked article texts in English and in Russian*).

Key words and phrases: Cadastral registration, aerial photography analysis, instance segmentation, Mask R-CNN, PyTorch

2020 *Mathematics Subject Classification*: 68T20; 68T07, 68T45

For citation: Igor V. Vinokurov. *Using the Mask R-CNN model for segmentation of real estate objects in aerial photographs*. Program Systems: Theory and Applications, 2025, **16**:1(64), pp. 3–44. (*In English, in Russian*).
https://psta.psiras.ru/read/psta2025_1_3-44.pdf

Introduction

Cottage settlements, dacha cooperatives and gardening associations have a high density of development, which complicates the automatic detection of illegally built real estate objects and objects not registered in the Unified State Register of Real Estate on aerial photographs (USRRE).

For the implementation of aerial photography, the Roskadastr PPC uses quadcopters. The resulting photographs are used to search for real estate objects subject to mandatory registration. Such objects include:

- country houses intended for permanent residence,
- objects standing on a foundation (baths, terraces, summer kitchens),
- buildings on plots for individual housing construction.

Manual search limits the efficiency and accuracy of this process. With the development of deep learning models such as RetinaNet [1], SOLO [2], YOLACT [3], Mask R-CNN [4] and others, it becomes possible to significantly reduce the time and improve the quality of image analysis due to automatic detection of objects. The models listed above are a powerful tool for implementing instance segmentation, which allows not only to identify objects in an image, but also to accurately highlight their boundaries. This is especially important for tasks related to the analysis of dense buildings, where objects can be partially overlapped or located in close proximity to each other. The use of instance segmentation implemented by these models is best suited for analyzing geospatial data, since the presence of masks and bounding boxes around the found object instances allows not only to anticipate the presence of search objects, but also to estimate their sizes. The latter can be useful when assessing changes in the size of objects associated with their reconstruction.

This paper presents the results of a study of the applicability of the Mask R-CNN (Mask Region-based Convolutional Neural Network) model for detecting real estate in cottage settlements, summer cottage cooperative and gardening associations. The peculiarity of the proposed approach is the processing of aerial photographs in order to identify the bounding frames of the detected objects, compared with certain geographic coordinates, and the comparison of their presence or significant change in size with the results of similar processing of aerial photographs for the previous period, usually a year. This approach allows us to identify changes in development and record buildings that may not have official registration in USRRE.

1. Overview of works on object recognition in aerial photographs using Mask R-CNN

Mask R-CNN is an extension of the Faster R-CNN model, designed to solve the problems of object detection and image segmentation [5]. A distinctive feature of Mask R-CNN from Faster R-CNN is the creation of accurate masks for each of the detected objects in the image. The model consists of two main components – object detection and object segmentation [4].

To extract features from the input image, Mask R-CNN uses the ResNet architecture [6] or another convolutional neural network. These features are then passed to a Region Proposal Network (RPN) [4, 6], which generates Regions of Interest (RoI) – predicted object locations. All obtained RoIs are passed to the main classifier, which determines the class of the object and specifies its boundaries.

Unlike Faster R-CNN, Mask R-CNN has an additional branch responsible for creating segmentation masks. This branch uses a small convolutional network to generate binary masks for all objects in the RoI. The mask size corresponds to the size of the RoI.

The model is trained using a loss function that combines the losses from classification, coordinate regression, and segmentation, which allows the model to simultaneously optimize all three tasks [4].

Mask R-CNN copes well with occluded objects and complex backgrounds due to its ability to generate accurate masks. However, its performance may depend on the quality of the training dataset and the chosen model architecture. Overall, Mask R-CNN is a strong tool for object detection and segmentation tasks in images, providing high accuracy and flexibility in application.

Currently, there are quite a lot of works on recognizing objects of various types in photographs and aerial photographs using this model, which indicates the effectiveness of its application for solving practical problems.

The paper [4] presents the basic architecture of Mask R-CNN and its application to various segmentation problems, including instance segmentation of objects.

In [7], the authors apply Mask R-CNN to detect buildings in high-resolution satellite images. High segmentation accuracy and advantages of using the Mask R-CNN model for solving such problems are demonstrated.

Detection of buildings in photographs from natural disaster zones using the Mask R-CNN model and pre-processing of the dataset, in order to improve the efficiency of its use, is given in [8]. The paper demonstrates the advantages of this model over other models that implement object detection in photographs.

The paper [9] solves the problem of localizing building polygons in high-resolution satellite images. The efficiency of the Mask R-CNN model for obtaining real building contour boundaries in photographs with high object density is demonstrated.

A method for detecting and segmenting ships at the pixel level using the Mask R-CNN model is proposed in [10]. The advantages of using this model are noted and the efficiency of the proposed method is evaluated.

In [11], it is shown that extracting building contours from satellite images is a complex task due to differences in scale, structure and types of buildings. To solve this problem, it is proposed to use the Mask R-CNN model. The efficiency of its use is demonstrated. The results of extracting individual buildings from satellite images are proposed to be used for applications that automate population assessment, implement urban planning and others.

Detection of modern urban architecture using the Mask R-CNN model trained on a dataset consisting of elements of modern architectural styles is given in [12]. As a result of comparing the obtained results, the efficiency of using Mask R-CNN is shown in comparison with other models.

The work [13] shows the difficulty of detecting buildings and various types of structures on satellite images due to illumination, building density, different types of terrain and other factors. To effectively solve this problem, it is proposed to use the Mask R-CNN model and our own dataset with improved image augmentation.

The use of the Mask R-CNN model for recognizing various objects on satellite images and aerial photographs for the purpose of automating mapping and maintaining local maps up to date is described in the work [14].

In [15], the Mask R-CNN model is proposed to be used to maintain the accuracy of terrain maps for effective response to natural disasters and catastrophes. It is proposed to update terrain maps based on aerial and satellite images. The work obtained more than acceptable results for photographs of different quality, resolutions and color channels.

In [16, 17], a hybrid approach to extracting building contours from low-resolution satellite images using the Mask R-CNN model is proposed. This approach opens up prospects for the development of automated tools for processing satellite images, and their effective use in land use monitoring and disaster response.

2. Justification for the choice of the model

Based on the above review, the Mask R-CNN model can be reasonably chosen for implementing instance segmentation of real estate objects in aerial photographs, since it has the following advantages.

(*1*) Accuracy

- The Mask R-CNN model demonstrates higher accuracy in detecting objects compared to other models [9, 10].
- The two-stage approach of the Mask R-CNN model (region proposal and segmentation [4–6]) allows for more accurate localization and segmentation of objects while simultaneously generating their masks.

(*2*) Versatility

- Mask R-CNN can be easily adapted to solve a variety of problems, including object detection and contouring, instance segmentation and semantic segmentation [4–6].
- The modular architecture of this model allows you to easily add or remove components as needed.

(*3*) Robustness

- Mask R-CNN is more robust to noise and distortion in images [15].
- The two-stage approach of this model helps reduce the number of false positives and improves the overall robustness [4–6].

(*4*) Support for objects of different sizes

- Mask R-CNN can effectively segment objects of different sizes, from small to large [4, 8].

- The region proposal mechanism [4] allows this model to detect objects regardless of their type and size (e.g. COCO [5]).

(5) Less overfitting issues
- Mask R-CNN is less prone to overfitting [11].
- Using multiple loss functions and a two-stage approach helps the model to generalize better [4–6, 18]

Mask R-CNN models also have some drawbacks that are not critical when solving the problem of automating image processing, these are relatively low speed of operation and complexity of implementation [19].

In addition to the models discussed above, the YOLO model [20], which has become quite popular recently, can be used to solve the instance segmentation problem. Despite the high speed of obtaining results, this model has significant drawbacks – low quality of recognition of groups of small objects due to the limited number of candidates for bounding boxes (two) and the possibility of duplicating bounding boxes for the same object [21]. A comparison of the results of real estate recognition with similar results obtained using the latest version of the YOLO model is given in p. 6

3. Purpose of the work

As noted above, photographing the territories under study by the «Roscadastr» PLC control center is carried out using quadcopters. The quadcopter (at the time of writing, it is Phantom 4 RTK) implements movement along a predetermined trajectory. The camera automatically takes pictures upon reaching certain points along the route, forming either its full image or images of route fragments (in most cases, with overlap).

The purpose of this work is to study the applicability of the Mask R-CNN model for the implementation of automatic or automated processing of the obtained images and the identification of illegally built or unregistered real estate objects on them. To achieve the stated goal, the following tasks were solved in the work:

(1) creation of our own dataset for training the model,

(2) analysis of the quality of the model,

(3) analysis of the results of using the model to detect illegal construction sites or real estate objects not registered in the USRRE.

TABLE 1. Object types and their labels

Object type (class)	Label name and class segmentation color
Cottage or summer house	*building*
Greenhouse (hothouse)	*greenhouse*
Outbuilding	*outbuilding*
Vehicle	*vehicle*
Swimming pool	*swimming*

4. Dataset creation

4.1. Image annotation

The main task in generating a dataset used for training and testing the model is labeling or annotating the original images, which is a polygonal contouring of objects recognized in photographs. The accuracy of the contouring determines the accuracy of object detection.

Currently, there is a large amount of software that implements image annotation. To generate the dataset, open source software LabelMe was selected.

The annotation process involves assigning a label to each recognizable object of a particular type. When annotating aerial photographs, labels were assigned to the following 5 types of objects, Table 1.

The main type of objects is a cottage or a summer house. Other types of objects are used to prevent false recognition of the latter and for the purpose of possible continuation of work in this direction.

The annotation results were saved in the JSON format supported by LabelMe. When forming the dataset, 435 photographs obtained from a quadcopter were annotated. An example of a photograph of a fragment of a summer house cooperative and the polygonal contouring of 2 classes of objects – summer house and greenhouse in LabelMe implemented during the annotation process are shown in Figure 1, 2*a* and 2*b*, respectively.

The dataset formed for training and testing the model is a collection of the above number of aerial photographs and the same number of JSON files in the LabelMe software format, containing arrays of coordinates of polygonal contouring points and names for objects of 5 types.

For training and testing the model, 80% and 20% of the elements of this dataset were randomly selected, respectively. The distribution

FIGURE 1. Photograph of a fragment of a summer cottage cooperative taken by a Phantom 4 RTK quadcopter

(*a*) A summer cottage and a greenhouse (*b*) Contouring of objects of 2 types

FIGURE 2. Dataset element – image of objects and their polygonal contouring on a fragment of an image

diagram of objects belonging to these classes on aerial photographs is shown in Figure 3.

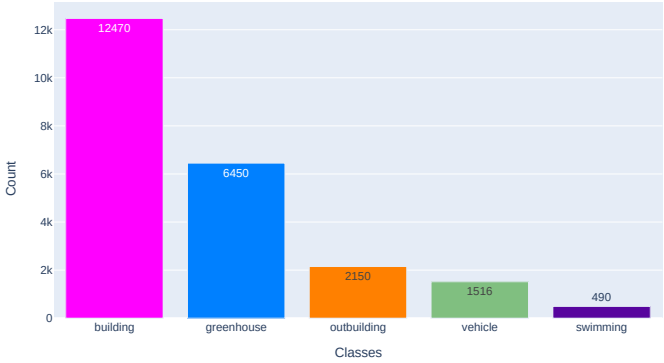


Figure 3. Distribution of classes in the dataset photographs

4.2. Dataset augmentation

To improve the generalization ability of the model, a pipeline of transformations implemented in the PyTorch package was formed and applied to the images from the training dataset. At the very beginning, based on the IoU (Intersection over Union) metric, images were cropped while preserving the ROI. Then, stochastic color transformations were applied, including variations in brightness, contrast, saturation, and hue, as well as random grayscale transformation (RandomGrayscale) and histogram equalization (RandomEqualize).

Further increase in diversity consisted in reducing the number of color levels and horizontal reflection of images using the methods RandomPosterize and RandomHorizontalFlip respectively. In addition, scaling of images to the maximum size while maintaining the aspect ratio, padding to a square shape, and resizing to the required size using antialiasing were implemented. The final step was to convert the data type to torch.float32 with scaling of pixel values and validation of bounding box coordinates (SanitizeBoundingBoxes).

5. Model creation and exploration

5.1. Model creation

Mask R-CNN is characterized by a relatively high complexity of implementation. As a result, successful application of this model requires careful tuning of the parameters and network architecture. A fairly fast and

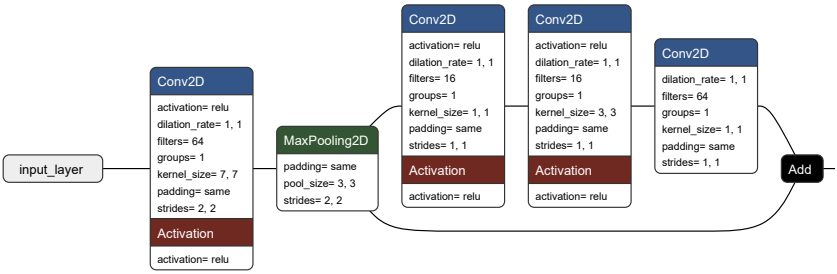


FIGURE 4. The initial convolutional layer in ResNet-101 and the first residual block [22]

accurate object detection model maskrcnn_resnet50_fpn_v2 from the torchvision package was chosen in this work. To improve the efficiency of feature extraction (backbone), the ResNet-101 architecture with Feature Pyramid Network (FPN) [4] was chosen in this model. The general principles of its formation and training are described quite well in [22] and [23].

The model is pre-trained on the COCO dataset [5]. For this model, the optimal number of epochs for additional training was experimentally found, the optimizer Adam and the scheduler OneCycleLR (the mechanisms that determine the decrease in weights during training and the speed of this process) were selected, the number of output channels was changed, and the partitioning for anchor boxes – pre-defined rectangular frames that are used to propose potential RoIs when detecting objects in an image was configured. The code for generating and exploring the model was written using the PyTorch machine learning library [24].

The initial convolutional layer of ResNet-101 was modified to efficiently implement instance segmentation of objects of different scales. Several parallel convolutional layers with kernels of different sizes were added before the layer with a sufficiently large kernel size (7×7), Figure 4 and Figure 5.

This potentially allows the model to simultaneously take into account both contextual features (kernel 5×5) and detailed characteristics of objects (kernels 1×1 and 3×3). In addition, parallel convolutional layers allow accelerating the convergence of the model.

Attempts to vary the residual blocks of this model did not lead to any significant improvement in object detection.

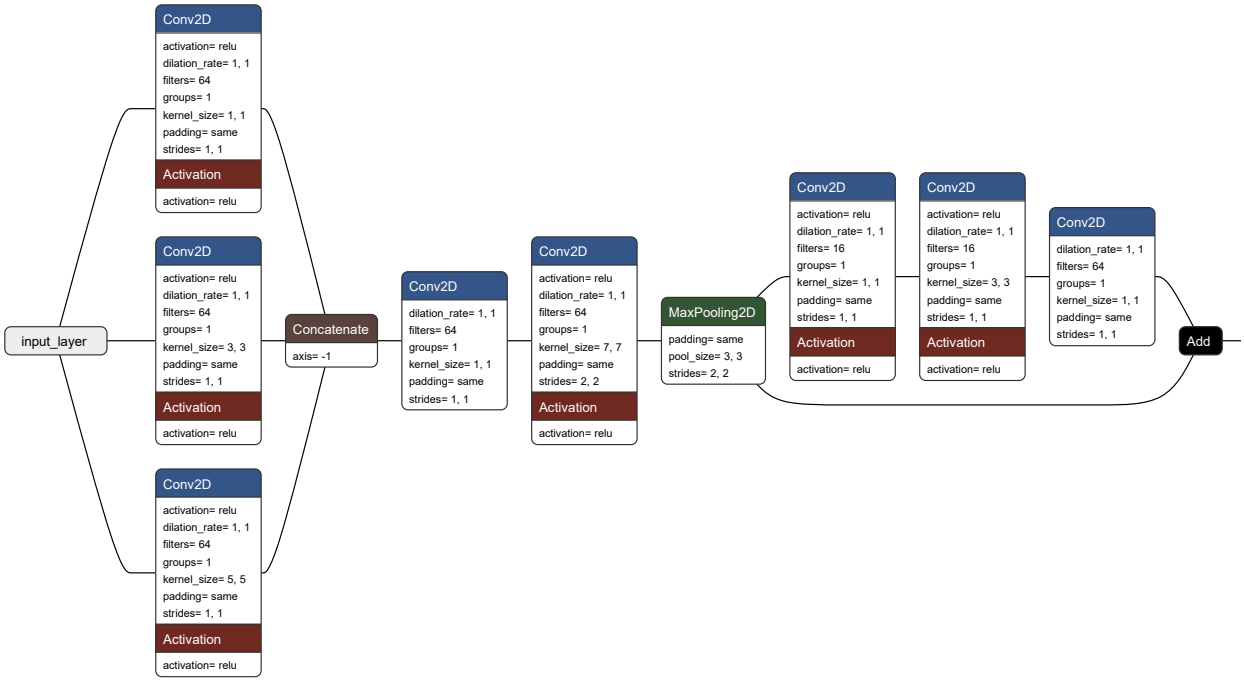


FIGURE 5. New parallel convolutional layers in ResNet-101

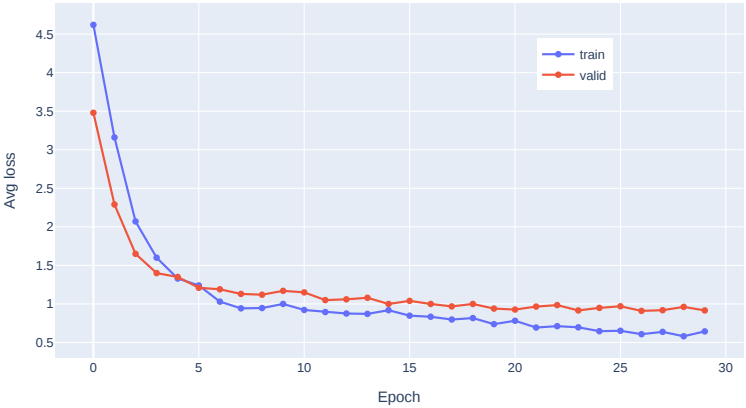


FIGURE 6. Change in the mean value of the model loss function

5.2. Model exploration

To analyze the accuracy of the model during the experimental studies, 3 key metrics were calculated – Loss, mAP and CS.

Loss in Mask R-CNN allows you to evaluate how accurately the model predicts classification, coordinate regression and object segmentation. The closer the values of this metric to 0 (0%), the more accurate the model's result. The change in the average value of this metric for all 5 classes from the training and testing (valid) datasets depending on the training epoch is shown in Figure 6.

The mAP (mean Average Precision) metric is the average value of the precision metric over all classes of objects. A similar metric mAP50-95 is calculated for different thresholds of overlapping bounding boxes IoU (Intersection over Union) – from 50% to 95%. Both metrics take into account both precision (precision) and recall (recall) at different IoU levels. The closer the values of each of these metrics are to 1 (100%), the more accurate the model's output is. The dependence of the mAP and mAP50-95 metrics for all classes from the training and valid datasets on the training epoch is shown in Figure 7.

The Confidence Score (CS) metric represents the probability that an object belongs to a certain class. CS is a measure of the model's confidence that the predicted object is actually present in the bounding

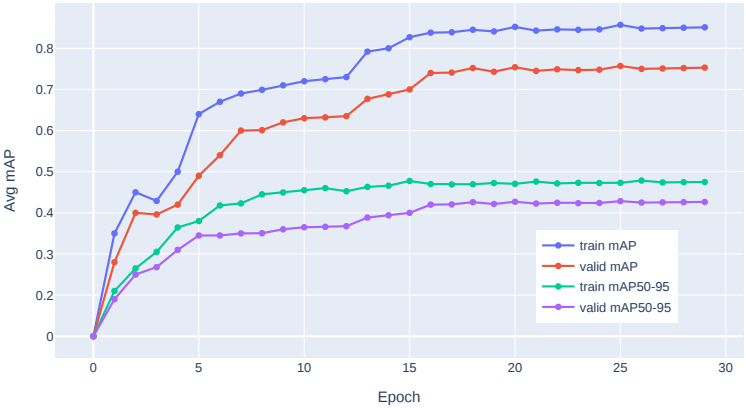


FIGURE 7. Change in the average value of mAP and mAP50-95 metrics in bounding box detection

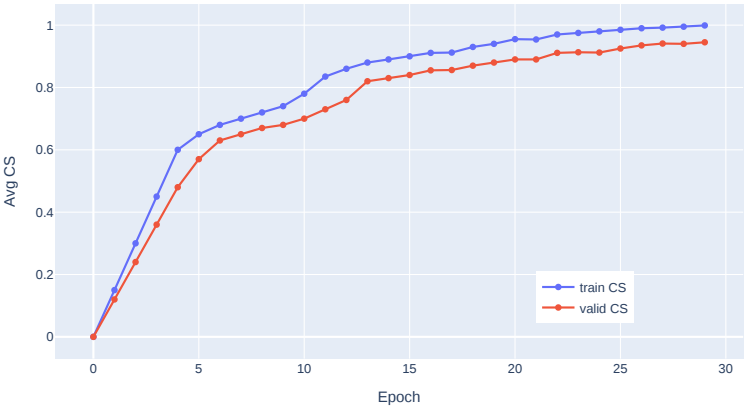
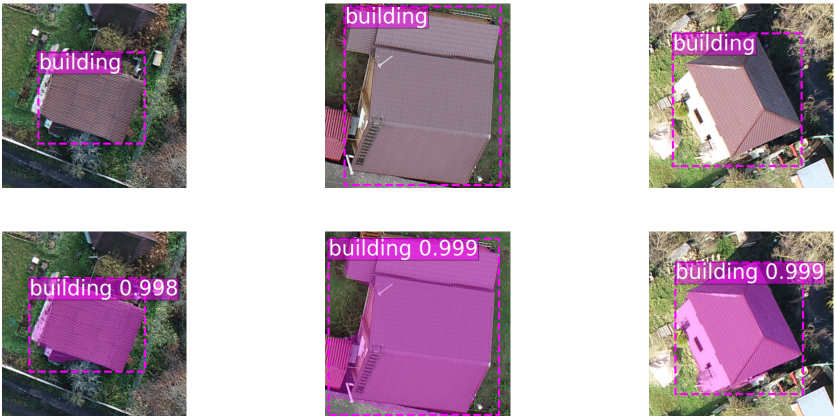


FIGURE 8. Change in the average value of the CS metric during object detection

box. This metric is one of the most important for models implementing object detection. CS values range from 0 to 1 (100%); higher values indicate greater model confidence in the correctness of its prediction. The dependence of this metric for all classes from the training and valid datasets on the training epoch is shown in Figure 8.

TABLE 2. Accuracy metric values at the last epoch of model training

Dataset	Loss	mAP	mAP50-95	CS
train	0.741	0.851	0.475	0.999
valid	0.915	0.753	0.416	0.991



(*a*) CS=0.998 (*b*) CS=0.999 (*c*) CS=0.999

FIGURE 9. Detecting real estate objects in images from the test dataset

Table 2 shows the values of the accuracy metrics for the training and test datasets at the last epoch of model training, see Figure 6–8.

Calculating accuracy metric values in Google Colab Pro in the T4 GPU runtime took about 80-100 minutes. From the graphs and table above, it follows that the Mask R-CNN model trained on its own dataset has more than acceptable accuracy in detecting real estate objects of interest to us in aerial photographs. Several examples of detecting summer houses from the datasets for training the model and for testing it are shown in Figure 9, 10, respectively.

When using this model, minor errors were detected in 3-5% of cases, manifested in the form of incorrect identification objects or incorrect segmentation of these boundaries. The occurrence of these errors is associated with the density of the segmented objects, with an insufficiently representative dataset, the presence of noise in the images and insufficient

(*a*) CS=0,999 (*b*) CS=0,997 (*c*) CS=0,999

FIGURE 10. Detection of real estate objects in images from the test dataset

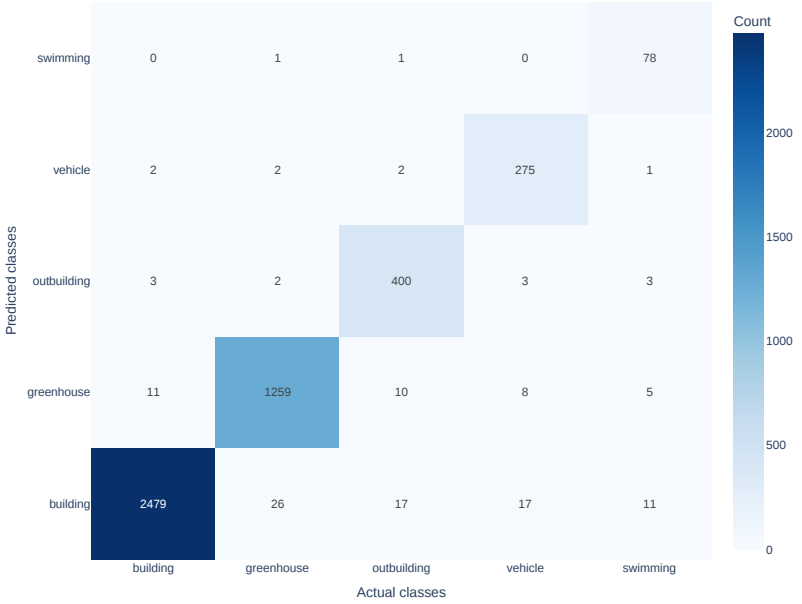


FIGURE 11. Model confusion matrix for the test dataset

sharpness of the image area. The matrix of errors in detecting real estate objects in images from the test dataset is shown in Figure 11.

Example incorrect identification for several objects of the *greenhouse*, *outbuilding* and *vehicle* classes is shown in Figure 12. It is easy to see that all objects of the *building* class, due to their large number in the training dataset, are identified correctly, which is quite sufficient for the above stated purpose of the work.

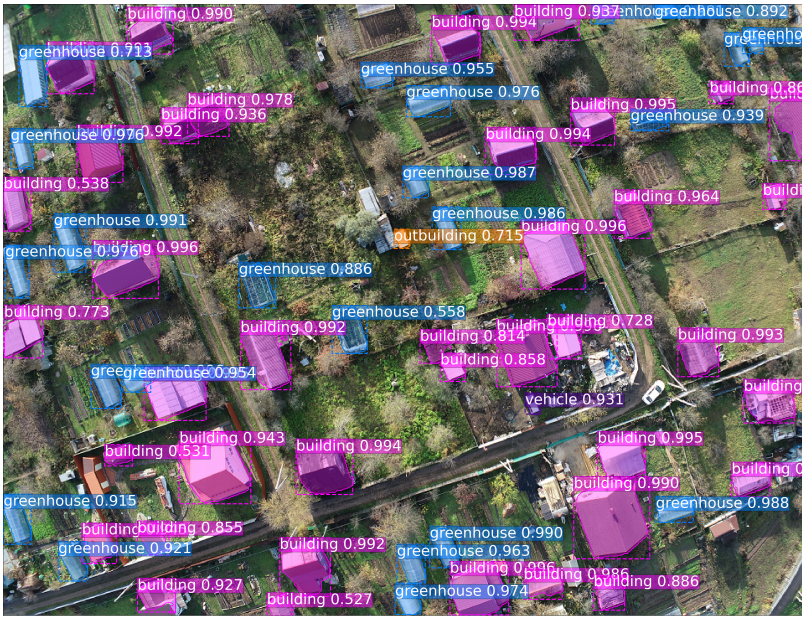


FIGURE 12. Results of detection of objects of different types in a photograph of a fragment of a summer cottage cooperative (see Figure 1)

6. Comparison with the YOLO model results

The YOLO deep learning model is currently being intensively developed and can be used to implement real estate detection in aerial photographs along with Mask R-CNN. Studies similar to those above were also conducted for the yolo11n-seg model from Ultralytics. Figure 13 shows the dependence of the mAP and mAP50-95 accuracy metrics on the training epoch when detecting real estate bounding boxes. This figure shows that these metrics achieve values comparable to those obtained by the Mask R-CNN model over a larger number of epochs.

Quite often, when analyzing the obtained results, cases were found where the detected object was related to different classes. In Figure 14, which displays the detection results similar to those shown in Figure 12, several such cases are visible (for example, with objects of the *building* and *outbuilding* classes). When studying the results obtained using Mask R-CNN, no similar cases were found. In addition, when analyzing a fairly large number of detection results, a more accurate (up to 10%-15%, see

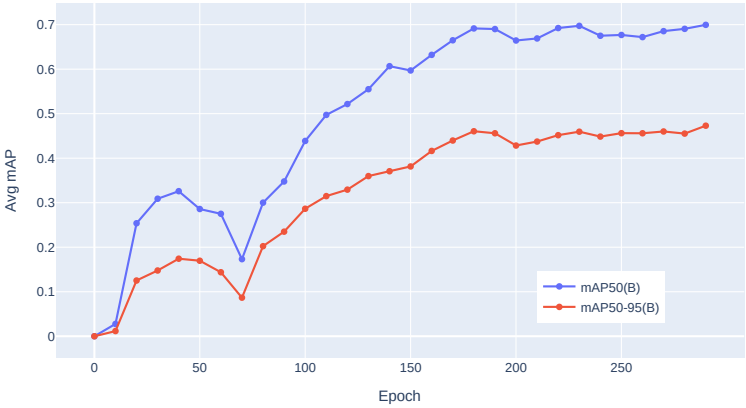


Figure 13. Change in the average value of the mAP and mAP50-95 metrics of the YOLO model when detecting bounding boxes

Figure 12,14) definition of object masks by the Mask R-CNN model was noted compared to YOLO. Incorrect (inaccurate) definition of an object mask can be critical for the «Roscadastr» PLC information system.

Overall, the YOLO model from Ultralystics left a favorable impression with its ease of use and the availability of ready-made functionality for conducting experiments and analyzing the results obtained. In some cases, depending on the specifics of the problem being solved and the dataset, using YOLO models allows you to get a slightly better result compared to Mask R-CNN [**25**].

7. Practical implementation of the results

The results obtained in the course of work to improve the efficiency of detecting cottages and summer houses on aerial photographs using the Mask R-CNN model were implemented in the beta version of one of the subsystems of the of the «Roscadastr» PLC information system. The main goal of this subsystem is to determine the presence of registration of construction objects in the USRRE. This subsystem implements a process that includes orthotransformation, creating a digital terrain model (DTM), obtaining reference points and calculating the coordinates of real estate objects [**26**].

A photograph taken from a quadcopter is divided into a number of fragments (tiles), which are analyzed for the presence or absence

FIGURE 14. Results of detecting objects of different types using YOLO

of unregistered objects. The tile size in pixels is determined by the scale of the image and the detection capabilities of the model. Dividing a photograph into fragments is justified by its large size, which can be several hundred thousand pixels and a size reaching 1T. An example of detecting an unregistered summer house on one of the aerial photograph fragments is shown in Figure 15.

Conclusion

This paper assessed the possibility of using the Mask R-CNN model with a modified backbone for detection and segmentation of cottages and summer houses on aerial photographs. To train the model, a custom dataset was created, including images of objects and their annotations. The studies showed that the Mask R-CNN model successfully copes with the task of instance segmentation, demonstrating acceptable accuracy for the Loss, mAP, mAP50-95 and CS metrics.

The developed beta version of the subsystem of the «Roscadastr»

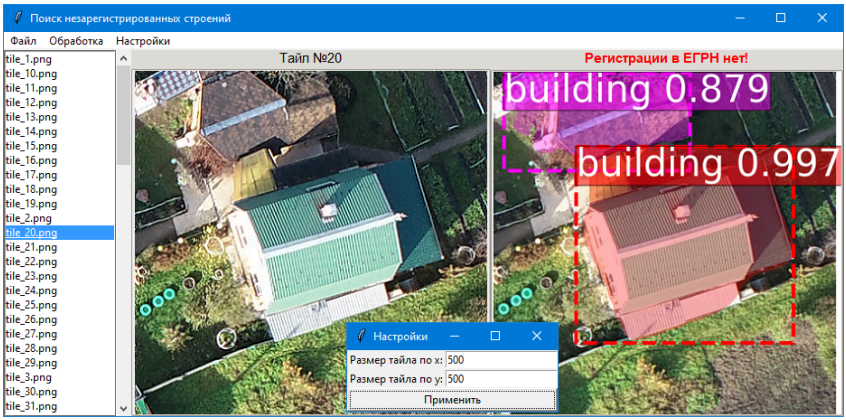


Figure 15. An example of detecting an unregistered summer house on an aerial photograph fragment

PLC information system based on Mask R-CNN uses this model to recognize objects in images obtained from a quadcopter, which significantly reduces the time and resources spent on identifying illegal and unregistered construction, and increases the efficiency of land use control.

Further research on the application of the Mask R-CNN model will be aimed at increasing its accuracy by expanding and optimizing the training dataset, as well as integrating this model into existing land use monitoring and control systems. In addition, a promising direction is the development of algorithms for automatic verification of the model's results based on comparison with the USRN data.

References

[1] T.-Y. Lin, P. Goyal, R. B. Girshick, K. He, P. Dollár. *Focal loss for dense object detection*, Computing Research Repository (CoRR), 2017, 10 pp. arXiv 1708.02002 ↑4

[2] X. Wang, T. Kong, Ch. Shen, Y. Jiang, L. Li. *SOLO: Segmenting objects by locations*, Computing Research Repository (CoRR), 2019, 19 pp. arXiv 1912.04488 ↑4

[3] K. Duda, A. Ivanov. "On decidability of amenability in computable groups", *Archive for Mathematical Logic*, **61** (2022), pp. 891–902. ↑4

[4] K. He, G. Gkioxari, P. Dollár, R. B. Girshick. *Mask R-CNN*, Computing Research Repository (CoRR), 2017, 12 pp. arXiv 1703.06870 ↑4, 5, 7, 8, 12

[5] S. Ren, K. He, R. B. Girshick, J. Sun. *Faster R-CNN: Towards real-time object detection with region proposal networks*, Computing Research Repository (CoRR), 2015, 14 pp. arXiv 1506.01497 ↑5, 7, 8, 12

[6] K. He, X. Zhang, S. Ren, J. Sun. *Identity mappings in deep residual networks*, Computing Research Repository (CoRR), 2016, 15 pp. arXiv 1603.05027 ↑5, 7, 8

[7] T.-Y. Lin, M. Maire, S. J. Belongie, L. D. Bourdev, R. B. Girshick, J. Hays, P. Perona, D. Ramanan, P. Dollár, C. L. Zitnick. *Microsoft COCO: Common objects in context*, Computing Research Repository (CoRR), 2014, 15 pp. arXiv 1405.0312 ↑6

[8] Y. Xu, L. Wu, Z. Xie, Z. Chen. "Building extraction in very high resolution remote sensing imagery using deep learning and guided filters", *Remote. Sens.*, **10**:1 (2018), id. 144, 18 pp. ↑6, 7

[9] Q. Han, Q. Yin, X. Zheng, Z. Chen. "Remote sensing image building detection method based on Mask R-CNN", *Complex Intell. Syst.*, **8** (2022), pp. 1847-–1855. ↑6, 7

[10] K. Zhao, J. Kang, J. Jung, G. Sohn. "Building extraction from satellite images using Mask R-CNN with building boundary regularization", *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)* (18–22 June 2018, Salt Lake City, UT, USA), IEEE, ISBN 9781538661017, id. 242, 4 pp. ↑6, 7

[11] X. Nie, M. Duan, H. Ding, B. Hu, E. K. Wong. "Attention Mask R-CNN for ship detection and segmentation from remote sensing images", *IEEE Access*, **8** (2020), pp. 9325–9334. ↑6, 8

[12] M. Jenila Vincent, P. Varalakshmi. "Extraction of building footprint using MASK-RCNN for high resolution aerial imagery", *Environmental Research Communications*, **6**:7 (2024), id. 075015, 17 pp. ↑6

[13] X. Zhu, L. Hu, J. Wang. "Urban modern architecture recognition based on Mask-RCNN and ECA attention mechanism", Fifth International Conference on Geoscience and Remote Sensing Mapping (ICGRSM 2023) (13–15 October 2023, Lianyungang, China), Proc. SPIE, vol. **12980**, 2024, ISBN 9781510672789, id. 129801D. ↑6

[14] R. Raghavan, D. Chander Verma, D. Pandey, R. Anand, B. Kumar Pandey, H. Singh. "Optimized building extraction from high-resolution satellite imagery using deep learning", *Multimedia Tools and Applications*, **81**:29 (2022), pp. 42309–42323. ↑6

[15] D. Ulanov, A. Syrov. "Building footprint extraction based on RGBD satellite imagery", CS230 Deep Learning (Winter 2020, Stanford University, CA), 2020, 11 pp. URL ↑7

[16] A. Solanki, R. K. Singh, B. Demeneze. "Aerial pictures semantic segmentation applying deep learning", *International Journal of Trendy Research in Engineering and Technology*, **5**:1 (2021), pp. 42–48. ↑7

[17] A. NourEldeen, M. E. Wahed. "Enhanced building footprint extraction from satellite imagery using Mask R-CNN and PointRend", *Bulletin of Electrical Engineering and Informatics*, **5**:13 (2024), pp. 3601–3608. ↑7

[18] K. He, X. Zhang, S. Ren, J. Sun. *Deep residual learning for image recognition*, Computing Research Repository (CoRR), 2015, 12 pp. arXiv 1512.03385 ↑8

[19] Ch. J. Mills. *PyTorch Mask R-CNN tutorial*, GitHub repository, 2023. ↑8
[20] J. Redmon, S. Divvala, R. B. Girshick, A. Farhadi. *You Only Look Once: Unified, real-time object detection*, Computing Research Repository (CoRR), 2015, 10 pp. arXiv 1506.02640 ↑8
[21] R. Khanam, M. Hussain. *YOLOv11: An overview of the key architectural enhancements*, 2024, 9 pp. arXiv 2410.17725 ↑8
[22] T.-Y. Lin, P. Dollár, R. B. Girshick, K. He, B. Hariharan, S. J. Belongie. *Feature pyramid networks for object detection*, Computing Research Repository (CoRR), 2016, 10 pp. arXiv 1612.03144 ↑12
[23] A. Waleed. *Mask R-CNN for object detection and instance segmentation on Keras and TensorFlow*, GitHub repository, 2017. ↑12
[24] E. Stevens, L. Antiga, T. Viehmann. *Deep Learning with PyTorch*, Manning Publications, New York, 2020, ISBN 9781617295263, 520 pp. ↑12
[25] R. Sapkota, A. Dawood, M. Karkee. "Comparing YOLOv8 and Mask R-CNN for instance segmentation in complex orchard environments", *Artificial Intelligence in Agriculture*, **13**:1 (2024), pp. 84–99. ↑19
[26] V. F. Bulavitsky. "Use of drones to obtain aerial photographas terrain", *Elektronnoe nauchnoe izdanie «Uchyonye zametki TOGU»*, **4**:4 (2013), pp. 1747–1755 (in Russian). ↑19

Received 21.10.2024;
approved after reviewing 24.12.2024;
accepted for publication 11.01.2025;
published online 31.01.2025.

Recommended by *dr. V. P. Fralenko*

Information about the author:

Igor Victorovich Vinokurov

Candidate of Technical Sciences (PhD), Associate Professor at the Financial University under the Government of the Russian Federation. Research interests: information systems, information technologies, data processing technologies

0000-0001-8697-1032
e-mail: igvvinokurov@fa.ru

The author declare no conflicts of interests.

RU EN ПРОГРАММНЫЕ СИСТЕМЫ: ТЕОРИЯ И ПРИЛОЖЕНИЯ ISSN 2079-3316
научная статья ПРИКЛАДНЫЕ ПРОГРАММНЫЕ СИСТЕМЫ

УДК 004.932.72:004.89
10.25209/2079-3316-2025-16-1-3-44

Использование модели Mask R-CNN для сегментации объектов недвижимости на аэрофотоснимках

Игорь Викторович **Винокуров**✉
Финансовый Университет при Правительстве Российской Федерации, Москва, Россия
✉*igvvinokurov@fa.ru*

Аннотация. Массовое появление незаконных и незарегистрированных в Едином Государственном Реестре Недвижимости (ЕГРН) объектов недвижимости осложняет кадастровый учёт для многих субъектов территориального и административного уровня. Традиционные методы выявления объектов подобных типов, основанные на ручном анализе геопространственных данных, трудоёмки и требуют значительного времени.

Для повышения эффективности этого процесса предлагается автоматизировать обнаружение объектов на аэрофотоснимках путём решения задачи инстанс-сегментации с использованием модели глубокого обучения Mask R-CNN. В статье описана подготовка набора данных для этой модели, исследованы основные метрики качества и проанализированы полученные результаты. Показана эффективность модели Mask R-CNN при обнаружении объектов недвижимости, не имеющих регистрации в ЕГРН. (*Связанные тексты статьи на английском и на русском языках*)

Ключевые слова и фразы: Кадастровый учёт, анализ аэрофотоснимков, инстанс-сегментация, Mask R-CNN, PyTorch

Для цитирования: Винокуров И. В. *Использование модели Mask R-CNN для сегментации объектов недвижимости на аэрофотоснимках* // Программные системы: теория и приложения. 2025. **Т. 16**. № 1(64). С. 3–44. (*Англ.+русс.*)
https://psta.psiras.ru/read/psta2025_1_3-44.pdf

Введение

Коттеджные поселения, дачные кооперативы и садовые товарищества обладают высокой плотностью застройки, осложняющей автоматическое обнаружение на аэрофотоснимках незаконно построенных объектов недвижимости и объектов, не зарегистрированных в ЕГРН.

Для реализации аэрофотосъёмки в ППК «Роскадастр» используются квадрокоптеры. На полученных фотографиях ищутся объекты недвижимости, подлежащие обязательной регистрации. К таким объектам относятся:

- дачные дома, предназначенные для постоянного проживания,
- объекты, стоящие на фундаменте (бани, террасы, летние кухни),
- строения на участках для индивидуального жилищного строительства.

Ручной поиск ограничивает эффективность и точность этого процесса. С развитием моделей глубокого обучения, таких как RetinaNet [1], SOLO [2], YOLACT [3], Mask R-CNN [4] и других, появляется возможность значительно сократить время и повысить качество анализа изображений за счёт автоматического детектирования объектов. Перечисленные выше модели позволяют не только идентифицировать объекты на изображении, но и точно выделять их границы. Последнее особенно важно для анализа плотных застроек, где объекты могут быть частично перекрыты или находится в близком соседстве друг с другом. Использование инстанс-сегментации, реализуемой этими моделями, наилучшим образом подходит для анализа геопространственных данных, поскольку наличие масок и ограничивающих рамок вокруг найденных экземпляров объектов позволяет не только обнаружить объекты, но и оценить их размеры. Это помогает заметить изменения размеров в результате перестройки объекта.

В работе приведены результаты исследования применимости модели Mask R-CNN (Mask Region-based Convolutional Neural Network) для обнаружения объектов недвижимости в коттеджных поселениях, дачных кооперативах и садовых товариществах. Особенностью предлагаемого подхода является обработка аэрофотоснимка с целью выявление ограничивающих рамок обнаруженных объектов, сопоставленных с определёнными географическими координатами, и сравнение их наличия или значимого изменения размеров с результатами аналогичной обработки аэрофотоснимков за прошлый период, как правило год. Такой подход позволяет выявлять изменения в застройке и постройки, не имеющие официальной регистрации.

1. Обзор работ по распознаванию объектов на аэрофотоснимках с использованием Mask R-CNN

Mask R-CNN является расширением модели Faster R-CNN, предназначенной для решения задач детектирования объектов и сегментации изображений [5]. Отличительной особенностью Mask R-CNN от Faster R-CNN является создании точных масок для каждого из обнаруженных на изображении объектов. Модель состоит из двух основных компонентов – обнаружение объектов и их сегментация [4].

Для извлечения признаков из входного изображения Mask R-CNN использует архитектуру ResNet [6] или другую свёрточную нейронную сеть. Эти признаки затем передаются в сеть предположения региона (RPN, Region Proposal Network) [4, 6], которая генерирует области интереса (RoI, Region of Interest) – предполагаемые местоположения объектов. Все полученные RoI передаются в основной классификатор, который определяет класс объекта и уточняет его границы.

В отличие от Faster R-CNN, в модели Mask R-CNN реализована ещё одна ветвь, отвечающая за создание масок сегментации. Эта ветвь использует небольшую свёрточную сеть с целью генерации бинарных масок для всех объектов в RoI. Размер маски соответствует размеру RoI.

Для обучения модели используется функция потерь, которая объединяет потери от классификации, регрессии координат и сегментации, что позволяет модели одновременно оптимизировать все три задачи [4].

Mask R-CNN хорошо справляется с перекрывающимися объектами и сложными фонами благодаря способности генерировать точные маски. Однако её производительность может зависеть от качества обучающего набора данных и выбранной архитектуры модели. В целом, Mask R-CNN является сильным инструментом для задач детекции и сегментации объектов в изображениях, обеспечивая высокую точность и гибкость в применении.

В настоящее время существует достаточно большое количество работ по распознаванию объектов различных типов на фотографиях и аэрофотоснимках с использованием этой модели, что свидетельствует об эффективности её применения для решения практических задач.

В работе [4] представлена основная архитектура Mask R-CNN и её применение к различным задачам сегментации, включая инстанс-сегментацию объектов.

В [7] авторы применяют Mask R-CNN для обнаружения зданий на спутниковых снимках высокого разрешения. Показывается высокая

точность сегментации и преимущества использования модели Mask R-CNN для решения подобных задач.

Обнаружение зданий на фотографиях из зон стихийных бедствий с использованием модели Mask R-CNN и предварительной обработки набора данных, с целью повышения эффективности его использования, приведено в [8]. В работе показаны преимущества этой модели перед другими моделями, реализующими детектирование объектов на фотографиях.

В работе [9] решается задача локализации полигонов зданий на спутниковых снимках высокого разрешения. Показывается эффективность применения модели Mask R-CNN для получения реальных границ контуров зданий на фотографиях с высокой плотностью объектов.

Метод обнаружения и сегментации кораблей на уровне пикселей с использованием модели Mask R-CNN предлагается в [10]. Отмечается преимущества использования этой модели и оценивается эффективность предлагаемого метода.

В [11] показывается, что извлечение контуров зданий из спутниковых снимков является сложной задачей из-за различий в масштабах, структурах и типов зданий. Для решения этой задачи предлагается использовать модель Mask R-CNN. Показывается эффективность её использования. Результаты извлечение отдельных зданий из спутниковых снимков предлагается использовать для приложений, автоматизирующих оценку населения, реализующих городское планирование и других.

Детектирование современной городской архитектуры с использованием модели Mask R-CNN, обученной на наборе данных, состоящим из элементов современных архитектурных стилей, приведено в [12]. В результате сравнения полученных результатов показана эффективность использования Mask R-CNN по сравнению с другими моделями.

В работе [13] показывается сложность обнаружения зданий и различного типа построек на спутниковых снимках из-за освещённости, плотности застроек, различных типов рельефов местности и других факторов. Для эффективного решения этой задачи предлагается использовать модель Mask R-CNN и собственный набор данных с усовершенствованной аугментацией изображений.

Применение модели Mask R-CNN для распознавания различных объектов на спутниковых снимках и аэрофотоснимках с целью автоматизации картографирования и поддержания карт местностей в актуальном состоянии описано в работе [14].

В [15] модель Mask R-CNN предлагается использовать для поддержания точности карт местностей с целью эффективного реагирования

на стихийных бедствия и катастрофы. Актуализировать карты местностей предлагается на основе аэрофотоснимков и спутниковых снимков. В работе получены более чем приемлемые результаты для фотографий с различным качеством, разрешениями и цветовыми каналами.

В [16, 17] предлагается гибридный подход к извлечению контуров зданий из спутниковых снимков низкого разрешения с использованием модели Mask R-CNN. Такой подход открывает перспективы для разработки автоматизированных инструментов обработки спутниковых снимков, и их эффективному использованию при мониторинге землепользования и реагирования на стихийные бедствия.

2. Обоснование выбора модели

Для реализации инстанс-сегментации объектов недвижимости на аэрофотоснимках, исходя из проведённого выше обзора, может быть обоснованно выбрана модель Mask R-CNN, поскольку она обладает следующими преимуществами.

(1) Точность
- Модель Mask R-CNN демонстрирует более высокую точность детектирования объектов по сравнению с другими моделями [9, 10].
- Двухэтапный подход модели Mask R-CNN (предложение региона и сегментация [4–6]) позволяет более точно локализовать и сегментировать объекты с одновременным формированием их масок.

(2) Универсальность
- Mask R-CNN можно легко адаптировать для решения различных задач, включая обнаружение и контуризацию объектов, их интсанс- и семантическую сегментацию [4–6].
- Модульная архитектура этой модели позволяет легко добавлять или удалять её компоненты по мере необходимости.

(3) Надежность
- Mask R-CNN более устойчива к шумам и искажениям в изображениях [15].
- Двухэтапный подход этой модели помогает уменьшить количество ложных срабатываний и улучшить общую надежность [4–6].

(4) Поддержка объектов разных размеров
- Mask R-CNN может эффективно сегментировать объекты разных размеров, от маленьких до больших [4, 8].

- Механизм предложения региона [4] позволяет этой модели обнаруживать объекты независимо от их типа и размера (например, COCO [5]).

(5) Меньше проблем с переобучением.

- Mask R-CNN менее склонна к переобучению [11].
- Использование нескольких функций потерь и двухэтапный подход помогают модели лучше обобщать результаты [4–6, 18]

Модели Mask R-CNN присущи и некоторыми недостатками, которые не являются критичыми при решении задачи автоматизации обработки изображений, – это относительно невысокая скорость работы и сложность реализации [19].

Помимо рассмотренных выше моделей, для решения задачи инстанс-сегментации может быть использована и достаточно популярная в последнее время модель YOLO [20]. Несмотря на высокую скорость получения результатов, эта модель обладает существенными недостатками – невысоким качеством распознавания групп небольших объектов из-за ограниченного числа кандидатов для ограничивающих рамок (две) и возможностью дублирования ограничивающих рамок для одного и того же объекта [21]. Сравнение результатов распознавания объектов недвижимости, с аналогичными результатами, полученных с использованием последней версии модели YOLO, приведено в п. 6

3. Постановка задачи

Как уже было отмечено выше, фотографирование исследуемых ППК «Роскадстр» территорий осуществляется с использованием квадрокоптеров. Квадрокоптер (на момент написания статьи – это Phantom 4 RTK) реализует движение по заранее заданной траектории. Камера автоматически делает снимки при достижении определенных точек маршрута, формируя либо его полное изображение, либо изображения фрагментов маршрута (в большинстве случав с перекрытием).

Целью данной работы является исследование применимости модели Mask R-CNN для реализации автоматической или автоматизированной обработки полученных снимков и выявления на них незаконно построенных или незарегистрированных объектов недвижимости. Для достижения поставленной цели в работе решались следующих задач:

(1) формирование собственного набора данных для обучения модели,
(2) анализа качества работы модели,
(3) анализ результатов применения модели для обнаружения объектов незаконного строительства или объектов недвижимости, не зарегистрированных в ЕГРН.

4. Формирование набора данных

4.1. Аннотирование изображений

Основной задачей при формировании набора данных, используемого для обучения и тестирования модели, является отметка или аннотирование исходных изображений, представляющее собой полигональную контуризацию распознаваемых на фотографиях объектов. Точность контуризации определяет точность обнаружения объектов.

В настоящее время существует большое количество ПО, реализующего аннотирование изображений. Для формирования набора данных было выбрано ПО с открытым исходным кодом LabelMe.

Процесс аннотирования предполагает присваивание метки каждому распознаваемому объекту того или иного типа. При аннотировании аэрофотографий метки присваивались следующим 5-ти типам объектов, таблица 1.

Таблица 1. Типы объектов и их метки

Тип (класс) объекта	**Имя метки и цвет сегментации класса**
Коттеджный или дачный домик	*building*
Теплица (парник)	*greenhouse*
Хозяйственная постройка	*outbuilding*
Транспортное средство	*vehicle*
Бассейн	*swimming*

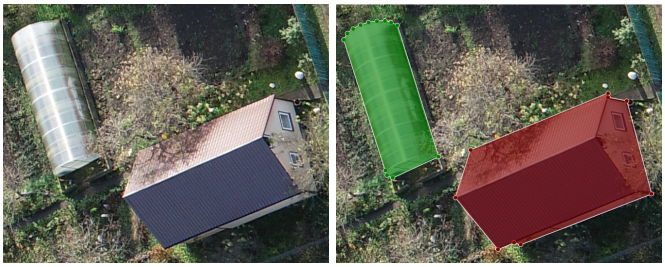
Основным типом объектов является коттеджный или дачный домик. Остальные типы объектов используются с целью предотвращения ложного распознавания последних и с целью возможного продолжения работы в данном направлении.

Результаты аннотирования сохранялись в формате JSON, поддерживаемом LabelMe. При формировании набора данных было осуществлено аннотирование 435 полученных с квадрокоптера фотографий. Пример фотографии фрагмента дачного кооператива и реализованная в процессе аннотирования полигональная контуризация 2-х классов объектов – дачного домика и теплицы в LabelMe приведены на рисунке 1, 2*а* и 2*б* соответственно.

Сформированный для обучения и тестирования модели набор данных представляет собой совокупность из указанного выше количества аэрофотоснимков и такого же количества JSON-файлов в формате ПО LabelMe, содержащих массивы координат точек полигональной контуризации и имена для объектов 5-ти типов.

Для обучения и тестирования модели было случайным образом выбрано 80% и 20% элементов этого набора данных соответственно.

Рисунок 1. Фотография фрагмента дачного кооператива с квадрокоптера Phantom 4 RTK



(*а*) Фрагмент изображения с дачным домиком и теплиц

(*б*) Контуризация объектов 2-х типов

Рисунок 2. Элемент набора данных – изображение объектов и их полигональная контуризация

Диаграмма распределения на аэрофотоснимках объектов, принадлежащих этим классам, приведена на рисунке 3.

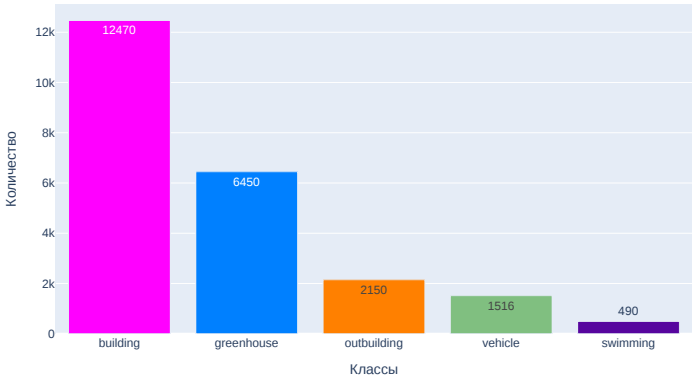


Рисунок 3. Распределение классов на фотографиях набора данных

4.2. Аугментация набора данных

Для повышения обобщающей способности модели сформирован и применён к изображениям из обучающего набора данных конвейер преобразований, реализованных в пакете PyTorch. В самом начале на основе метрики IoU (Intersection over Union) осуществлялось обрезка изображений с сохранением ROI. Затем применялись стохастические преобразования цвета, включающие вариации яркости, контрастности, насыщенности и цветового тона, а также случайное преобразование в градации серого (RandomGrayscale) и выравнивание гистограммы (RandomEqualize).

Дальнейшее увеличения разнообразия заключалось в уменьшении количества цветовых уровней и горизонтального отражения изображений с использованием методов RandomPosterize и RandomHorizontalFlip соответственно. Помимо этого было реализовано масштабирование изображений до максимального размера с сохранением соотношения сторон, дополнения до квадратной формы, и изменении размера до требуемого с использованием антиалиасинга. На заключительном этапе осуществлялось преобразование типа данных в torch.float32 с масштабированием значений пикселей и валидация координат ограничивающих рамок (SanitizeBoundingBoxes).

5. Формирование и исследование модели

5.1. Формирование модели

Mask R-CNN характеризуется относительно высокой сложностью реализации. Как следствие, успешное применение этой модели требует тщательной настройки параметров и архитектуры сети. В работе была

выбрана достаточно быстрая и точная модель обнаружения объектов maskrcnn_resnet50_fpn_v2 из пакета torchvision. Для повышения эффективности извлечения признаков (backbone) в этой модели была выбрана архитектура ResNet-101 с Feature Pyramid Network (FPN) [4]. Общие принципы её формирования и обучения достаточно хорошо описаны в [22] и [23].

Модель предварительно обучена на наборе данных COCO [5]. Для этой модели было экспериментально найдено оптимальное количество эпох дообучения, выбраны оптимизатор Adam и шедулер OneCycleLR (механизмы, определяющие уменьшение весов во время обучения и скорость этого процесса), изменено количество выходных каналов и настроено разбиение для якорных боксов – заранее определенных прямоугольных рамок, которые используются для предложения потенциальных RoI при обнаружении объектов на изображении. Код формирования и исследования модели написан с использованием библиотеки машинного обучения PyTorch [24].

Для эффективной реализации инстанс-сегментации объектов различных масштабов, первоначальный свёрточный слой ResNet-101 был модифицирован. Перед слоем с достаточно большим размером ядра (7×7) были добавлены несколько параллельных свёрточных слоёв, имеющих ядра различных размеров, рисунок 4, 5.

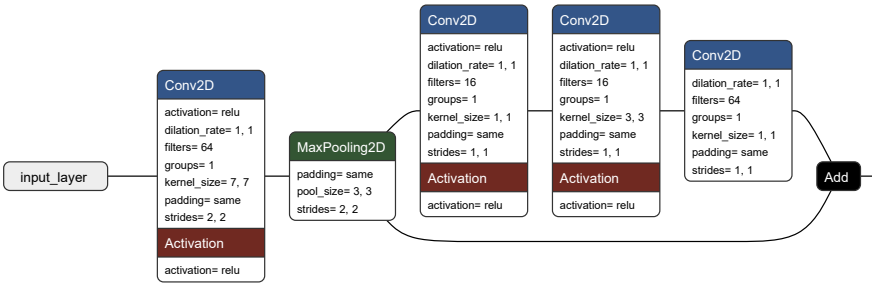


Рисунок 4. Начальный свёрточный слой в ResNet-101 и первый residual-блок [22]

Это потенциально позволяет модели одновременно учитывать как контекстуальные особенности (ядро 5×5), так и детальные характеристики объектов (ядра 1×1 и 3×3). Помимо этого, параллельные свёрточные слои позволяют ускорить сходимость модели.

Попытки вариативного изменения residual-блоков этой модели к какому-либо значимому улучшению детектирования объектов не привели.

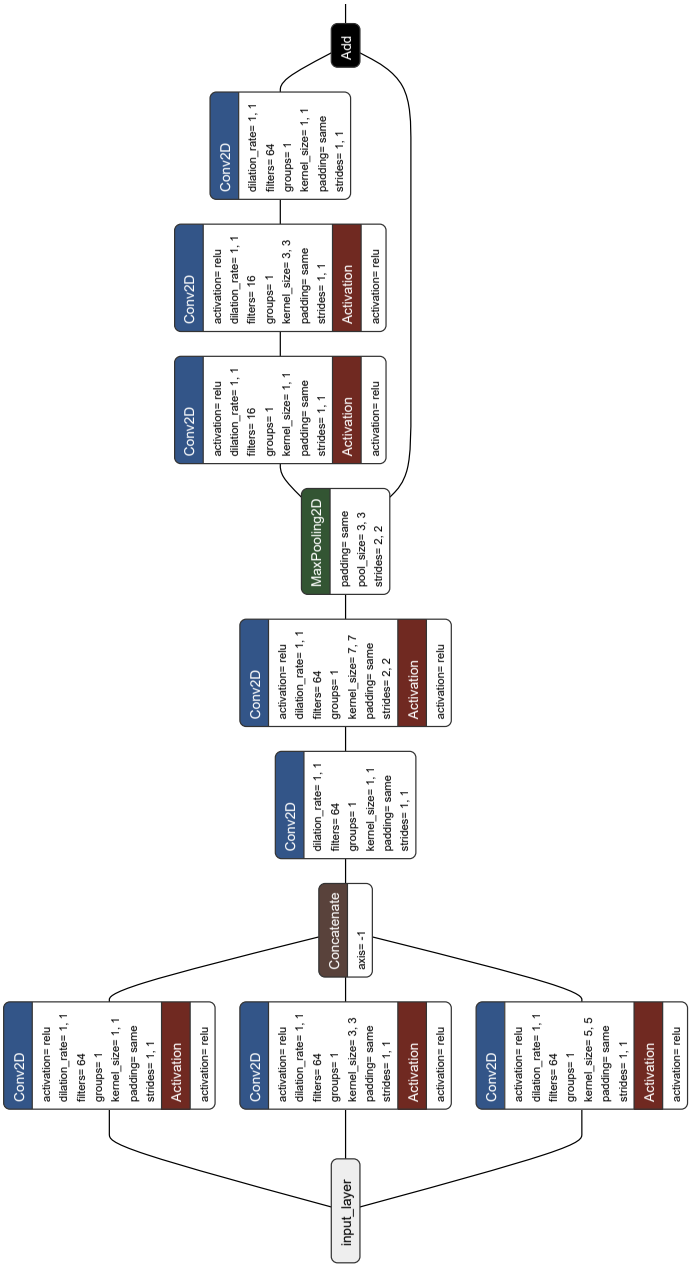


Рисунок 5. Параллельные свёрточные слои в ResNet-101

5.2. Исследование модели

Для анализа точности работы модели в процессе проведения экспериментальных исследований осуществлялось вычисление трёх ключевых метрик – Loss, mAP и CS.

Loss в Mask R-CNN позволяет оценить, насколько точно модель предсказывает классификацию, регрессию координат и сегментацию объектов. Чем ближе значения этой метрик к 0 (0%), тем более точным является результат работы модели. Изменение среднего значения этой метрики для всех 5-ти классов из наборов данных для обучения (train) и тестирования (valid) в зависимости от эпохи обучения приведено на рисунке 6.

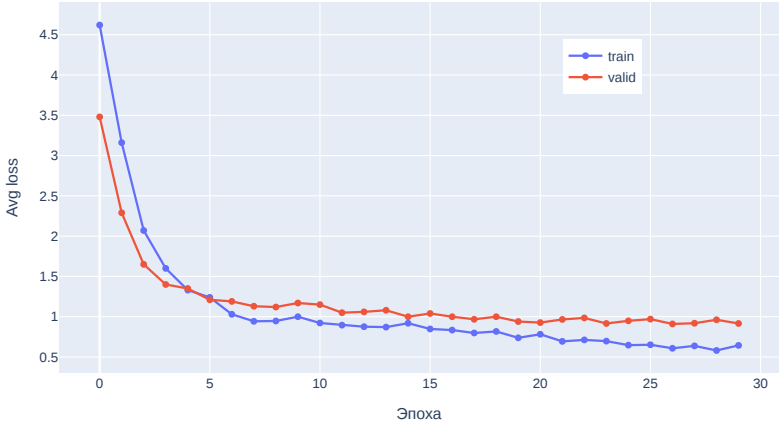


Рисунок 6. Изменение среднего значения функции потерь модели

Метрика mAP (mean Average Precision) является средним значением метрики точности по всем классам объектов. Аналогичная метрика mAP50-95 рассчитывается при различных порогах перекрытия ограничивающих рамок IoU (Intersection over Union) – от 50% до 95%. Обе метрики учитывают как точность (precision), так и полноту (recall) на различных уровнях IoU. Чем ближе значения каждой из этих метрик к 1 (100%), тем более точным является результат работы модели. Зависимости метрик mAP и mAP50-95 для всех классов из наборов данных обучения (train) и тестирования (valid) от эпохи обучения приведена на рисунке 7.

Метрика CS (Confidence Score) представляет собой вероятность того, что объект принадлежит к определенному классу. CS является мерой уверенности модели в том, что предсказанный объект действительно

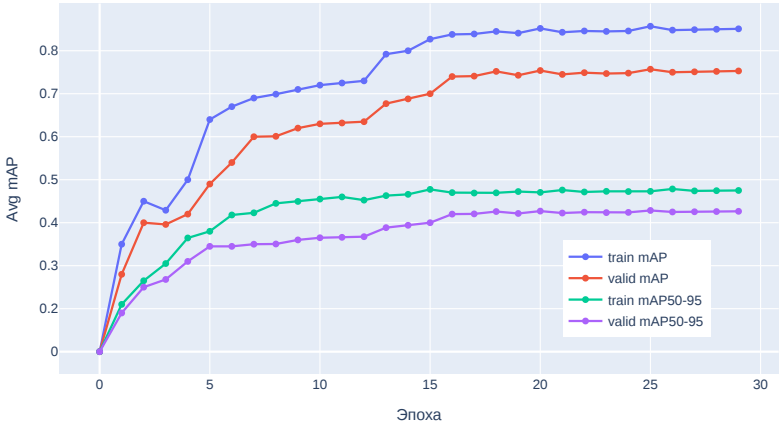


Рисунок 7. Изменение среднего значения метрик mAP и mAP50-95 при детектировании ограничивающих рамок

присутствует в ограничивающей рамке. Эта метрика особенно важна для моделей, реализующих детектирование объектов. Значения CS варьируются от 0 до 1 (100%); более высокие значения указывают на большую уверенность модели в правильности своего предсказания. Зависимость этой метрики для всех классов из наборов данных обучения (train) и тестирования (valid) от эпохи обучения показано на рисунке 8.

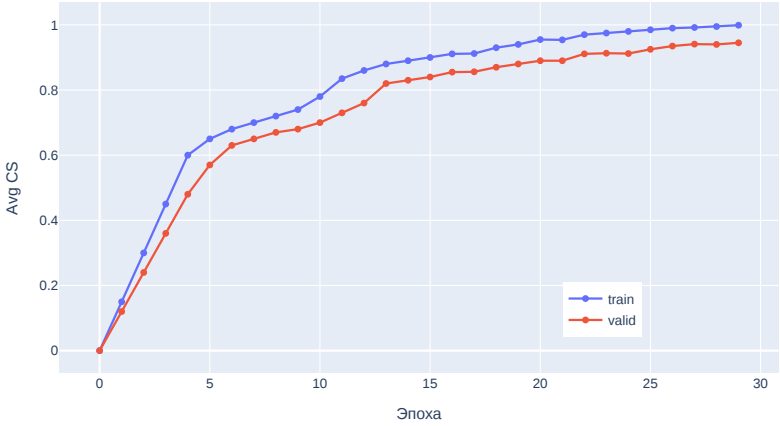


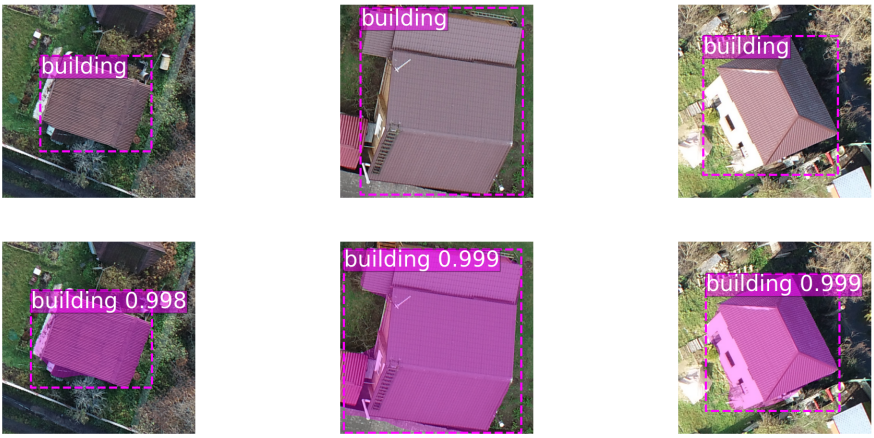
Рисунок 8. Изменение среднего значения метрики CS при детектировании объектов

В таблице 2 приведены значения метрик точности для обучающего и тестового наборов данных на последней эпохе обучения модели, см. рисунки 6–8.

Таблица 2. Значения метрик точности на последней эпохе обучения модели

Набор	**Loss**	**mAP**	**mAP50-95**	**CS**
train	0.741	0.851	0.475	0.999
valid	0.915	0.753	0.416	0.991

Вычисление значений метрик точности в Google Colab Pro в среде выполнения T4 GPU занимало прядка 80-100 минут. Из приведенных выше графиков и таблицы следует, что обученная на собственном наборе модель Mask R-CNN обладает более чем приемлемой точностью выявления интересующих нас объектов недвижимости на аэрофотоснимках. Несколько примеров детектирования дачных домиков из наборов данных для обучения модели и для её тестирования приведены на рисунках 9 и 10 соответственно.



(*а*) CS=0.998 (*б*) CS=0.999 (*в*) CS=0.999

Рисунок 9. Детектирование объектов недвижимости на изображениях из тестового набора данных

При использовании этой модели в 3-5% случаев были выявлены незначительные ошибки, проявляющиеся в виде неверной идентификации объектов или неправильной сегментации границы. Появление этих ошибок связано с плотностью расположения сегментируемых объектов, с недостаточно представительным набором данных, наличием шумов на снимках

(*а*) CS=0,999 (*б*) CS=0,997 (*в*) CS=0,999

Рисунок 10. Детектирование объектов недвижимости на изображениях из тестового набора данных

и недостаточной резкостью области изображения. Матрица ошибок детектирования объектов недвижимости на изображениях из тестового набора данных показана на рисунке 11.

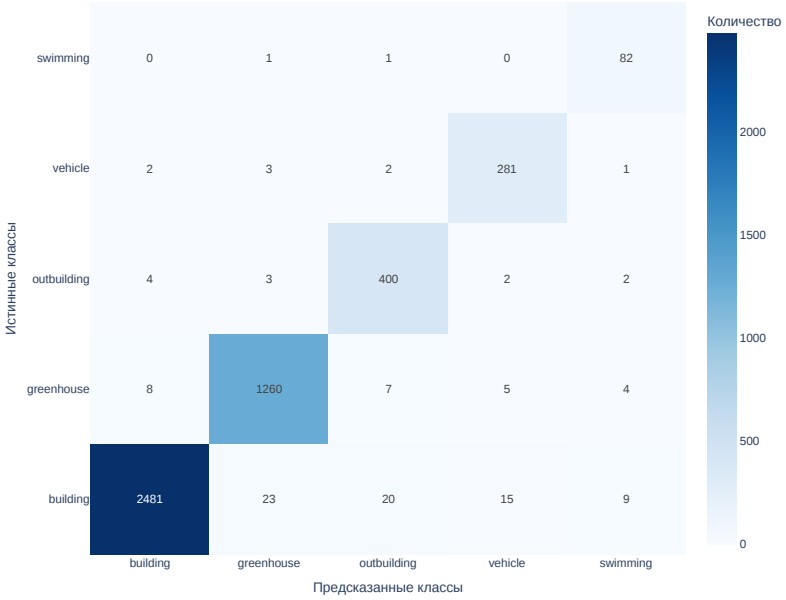


Рисунок 11. Матрица ошибок модели для тестового набора данных

Пример неверной идентификации для нескольких объектов классов *greenhouse*, *outbuilding* и *vehicle* приведён на рисунке 12. Легко заметить, что все объекты класса *building*, из-за достаточно большого их количества в обучающем наборе данных, идентифицируются верно, что является вполне достаточным для поставленной выше цели работы.

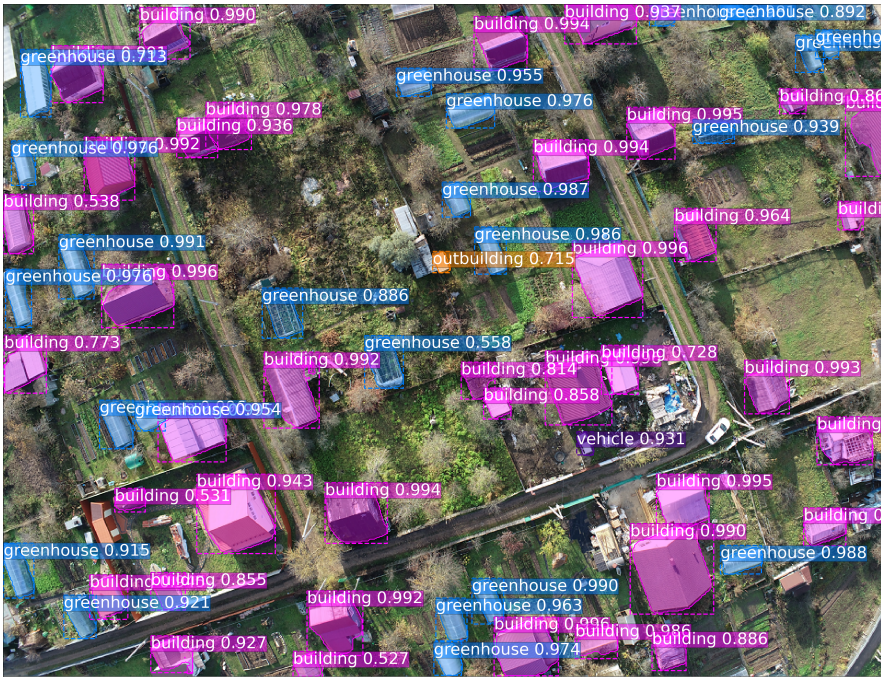


Рисунок 12. Результаты детектирования объектов разных типов на фотографии фрагмента дачного кооператива (см. рисунок 1)

6. Сравнение с моделью YOLO

Модель глубокого обучения YOLO в настоящее время интенсивно развивается и наравне с Mask R-CNN может быть использована для реализации детектирования объектов недвижимости на аэрофотоснимках. Исследования, аналогичные приведённым выше, были проведены и для модели yolo11n-seg от Ultralyticsc. На рисунке 13 приведены зависимости значений метрик точности mAP и mAP50-95 при детектировании ограничивающих рамок объектов недвижимости от эпохи обучения. Из этого рисунка видно, что эти метрики достигают значений, сопоставимых с полученными моделью Mask R-CNN, на большем количестве эпох.

Достаточно часто при анализе полученных результатов были выявлены случаи соотнесения обнаруженного объекта с разными классами. На рисунке 14, отображающем результаты детектирования, аналогичные приведённым на рисунке 12, видно несколько таких случаев (например, с объектами классов *building* и *outbuilding*). При исследовании результа-

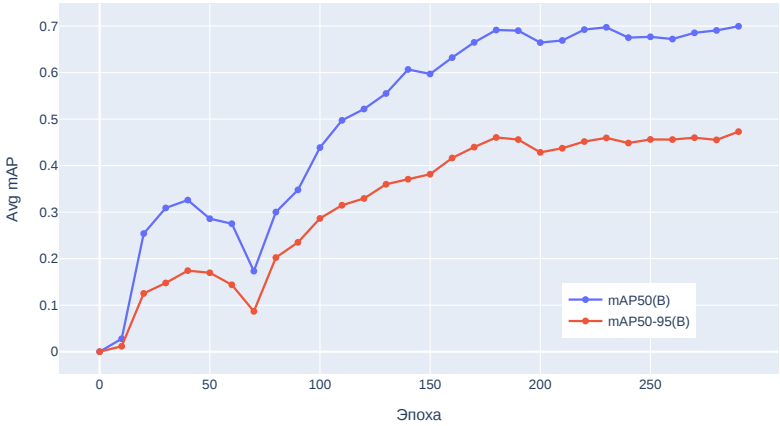


Рисунок 13. Изменение среднего значения метрик mAP и mAP50-95 модели YOLO при детектировании ограничивающих рамок

тов, полученных с использованием Mask R-CNN ни одного аналогичного случая выявлено не было. Кроме этого, при анализе достаточно большого количества результатов детектирования, было замечено более точное (до 10%-15%, см. рисунки 12 и 14) определение масок объектов моделью Mask R-CNN по сравнению с YOLO. Неправильное (неточное) определение маски объекта может быть критичным для ИС ППК «Роскадастр».

В целом, модель YOLO от Ultralystics оставила благоприятное впечатление простотой использования и наличием готового функционала для проведения экспериментов и анализ полученных результатов. В ряде случаев, в зависимости от специфики решаемой задачи и набора данных, использование моделей YOLO позволяет получить немного лучший результат, по сравнению с Mask R-CNN [25].

7. Практическая реализация результатов

Результаты, полученные в ходе работ по повышению эффективности обнаружения коттеджных и дачных домиков на аэрофотоснимках с использованием модели Mask R-CNN, были реализованы в бета-версии одной из подсистем информационной системы (ИС) ППК «Роскадастр». Основная цель этой подсистемы заключается в определении наличия регистрации объектов строительства в ЕГРН. В этой подсистеме реализуется процесс, включающий ортотрансформирование, создание цифровой модели местности (ЦММ), получение точек привязки и расчёт координат объектов недвижимости [26].

Рисунок 14. Результаты детектирования объектов разных типов с использованием YOLO

Фотография, сделанная с квадрокоптера, разбивается на некоторое количество фрагментов (тайлов), которые анализируются на предмет наличия или отсутствия незарегистрированных объектов. Размер тайла в пикселях определяется масштабом изображения и детектирующими способностями модели. Разбиение фотографии на фрагменты обосновывается её большим размером, который может составлять нескольких сотен тысяч пикселей и размера, достигающим 1Т. Пример обнаружения незарегистрированного дачного домика на одном из фрагментов аэрофотоснимка приведён на рисунке 15.

Заключение

В данной работе была проведена оценка возможности использования модели Mask R-CNN с изменённым backbone для обнаружения и сегментации на аэрофотоснимках коттеджных и дачных домиков. Для обучения модели был создан собственный набор данных, включающий в себя изображения объектов и их аннотации. Проведенные исследования показали, что модель Mask R-CNN успешно справляется с задачей

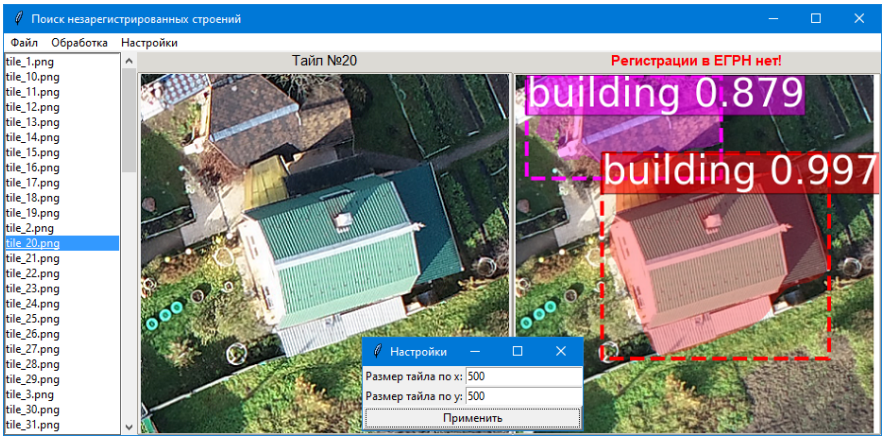


Рисунок 15. Пример выявления незарегистрированного дачного домика на фрагменте аэрофотоснимка

инстанс-сегментации, демонстрируя приемлемую точность по метрикам Loss, mAP, mAP50-95 и CS.

Разработанная бета-версия подсистемы ИС ППК «Роскадастр» на базе Mask R-CNN использует эту модель для распознавания объектов на изображениях, полученных с квадрокоптера, что существенно сокращает время и ресурсы, затрачиваемые на выявление нелегальной и незарегистрированной застройки, и повышает эффективность работы по контролю за использованием земельных ресурсов.

Дальнейшие исследования по применению модели Mask R-CNN будут направлены на повышение её точности в результате расширения и оптимизации обучающего набора данных, а также на интеграцию этой модели в существующие системы мониторинга и контроля за использованием земельных ресурсов. Кроме того, перспективным направлением является разработка алгоритмов автоматической верификации результатов работы модели, основанных на сравнении с данными ЕГРН.

Список использованных источников

[1] Lin T.-Y., Goyal P., Girshick R. B., He K., Dollár P. *Focal loss for dense object detection.*– Computing Research Repository (CoRR).– 2017.– 10 pp. arXiv 1708.02002 ↑25

[2] Wang X., Kong T., Shen Ch., Jiang Y., Li L. *SOLO: Segmenting objects by locations.*– Computing Research Repository (CoRR).– 2019.– 19 pp. arXiv 1912.04488 ↑25

[3] Duda K., Ivanov A. *On decidability of amenability in computable groups* // Archive for Mathematical Logic.– 2022.– Vol. **61**.– Pp. 891–902. ↑25

[4] He K., Gkioxari G., P. Dollár, Girshick R. B. *Mask R-CNN.*– Computing Research Repository (CoRR).– 2017.– 12 pp. arXiv 1703.06870 ↑25, 26, 28, 29, 33

[5] Ren S., He K., Girshick R. B., Sun J. *Faster R-CNN: Towards real-time object detection with region proposal networks.*– Computing Research Repository (CoRR).– 2015.– 14 pp. arXiv 1506.01497 ↑26, 28, 29, 33

[6] He K., Zhang X., Ren S., Sun J. *Identity mappings in deep residual networks.*– Computing Research Repository (CoRR).– 2016.– 15 pp. arXiv 1603.05027 ↑26, 28, 29

[7] Lin T.-Y., Maire M., Belongie S. J., Bourdev L. D., Girshick R. B., Hays J., Perona P., Ramanan D., P. Dollár, Zitnick C. L. *Microsoft COCO: Common objects in context.*– Computing Research Repository (CoRR).– 2014.– 15 pp. arXiv 1405.0312 ↑26

[8] Xu Y., Wu L., Xie Z., Chen Z. *Building extraction in very high resolution remote sensing imagery using deep learning and guided filters* // Remote. Sens.– 2018.– Vol. **10**.– No. 1.– id. 144.– 18 pp. ↑27, 28

[9] Han Q., Yin Q., Zheng X., Chen Z. *Remote sensing image building detection method based on Mask R-CNN* // Complex Intell. Syst.– 2022.– Vol. **8**.– Pp. 1847-–1855. ↑27, 28

[10] Zhao K., Kang J., Jung J., Sohn G. *Building extraction from satellite images using Mask R-CNN with building boundary regularization* // *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)* (18–22 June 2018, Salt Lake City, UT, USA).– IEEE.– ISBN 9781538661017.– id. 242.– 4 pp. ↑27, 28

[11] Nie X., Duan M., Ding H., Hu B., Wong E. K. *Attention Mask R-CNN for ship detection and segmentation from remote sensing images* // IEEE Access.– 2020.– Vol. **8**.– Pp. 9325–9334. ↑27, 29

[12] Jenila Vincent M., Varalakshmi P. *Extraction of building footprint using MASK-RCNN for high resolution aerial imagery* // Environmental Research Communications.– 2024.– Vol. **6**.– No. 7.– id. 075015.– 17 pp. ↑27

[13] Zhu X., Hu L., Wang J. *Urban modern architecture recognition based on Mask-RCNN and ECA attention mechanism*, Fifth International Conference on Geoscience and Remote Sensing Mapping (ICGRSM 2023) (13–15 October 2023, Lianyungang, China), Proc. SPIE.– vol. **12980**.– 2024.– ISBN 9781510672789.– id. 129801D. ↑27

[14] Raghavan R., Chander Verma D., Pandey D., Anand R., Kumar Pandey B., Singh H. *Optimized building extraction from high-resolution satellite imagery using deep learning* // Multimedia Tools and Applications.– 2022.– Vol. **81**.– No. 29.– Pp. 42309–42323. ↑27

[15] Ulanov D., Syrov A. *Building footprint extraction based on RGBD satellite imagery*, CS230 Deep Learning (Winter 2020, Stanford University, CA).– 2020.– 11 pp. ↑27, 28

[16] Solanki A., Singh R. K., Demeneze B. *Aerial pictures semantic segmentation applying deep learning* // International Journal of Trendy Research in Engineering and Technology.– 2021.– Vol. **5**.– No. 1.– Pp. 42–48. ↑28

[17] NourEldeen A., Wahed M. E. *Enhanced building footprint extraction from satellite imagery using Mask R-CNN and PointRend* // Bulletin of Electrical Engineering and Informatics.– 2024.– Vol. **5**.– No. 13.– Pp. 3601–3608. ↑28

[18] He K., Zhang X., Ren S., Sun J. *Deep residual learning for image recognition.*– Computing Research Repository (CoRR).– 2015.– 12 pp. arXiv 1512.03385 ↑29
[19] Mills Ch. J. *PyTorch Mask R-CNN tutorial.*– GitHub repository.– 2023. ↑29
[20] Redmon J., Divvala S., Girshick R. B., Farhadi A. *You Only Look Once: Unified, real-time object detection.*– Computing Research Repository (CoRR).– 2015.– 10 pp. arXiv 1506.02640 ↑29
[21] Khanam R., Hussain M. *YOLOv11: An overview of the key architectural enhancements.*– 2024.– 9 pp. arXiv 2410.17725 ↑29
[22] Lin T.-Y., P. Dollár, Girshick R. B., He K., Hariharan B., Belongie S. J. *Feature pyramid networks for object detection.*– Computing Research Repository (CoRR).– 2016.– 10 pp. arXiv 1612.03144 ↑33
[23] Waleed A. *Mask R-CNN for object detection and instance segmentation on Keras and TensorFlow.*– GitHub repository.– 2017. ↑33
[24] Stevens E., Antiga L., Viehmann T. *Deep Learning with PyTorch.*– New York: Manning Publications.– 2020.– ISBN 9781617295263.– 520 pp. ↑33
[25] Sapkota R., Dawood A., Karkee M. *Comparing YOLOv8 and Mask R-CNN for instance segmentation in complex orchard environments* // Artificial Intelligence in Agriculture.– 2024.– Vol. **13**.– No. 1.– Pp. 84–99. ↑40
[26] Булавицкий В. Ф. *Применение беспилотных летательных аппаратов для оперативного получения аэрофотоснимков местности* // Электронное научное издание «Учёные заметки ТОГУ».– 2013.– Т. **4**.– № 4.– С. 1747–1755. ↑40

Поступила в редакцию 21.10.2024;
одобрена после рецензирования 24.12.2024;
принята к публикации 11.01.2025;
опубликована онлайн 31.01.2025.

Рекомендовал к публикации *к.т.н. В. П. Фраленко*

Информация об авторе:

Игорь Викторович Винокуров

Кандидат технических наук (PhD), ассоциированный профессор в Финансовом Университете при Правительстве Российской Федерации. Область научных интересов: информационные системы, информационные технологии, технологии обработки данных

0000-0001-8697-1032
e-mail: *igvvinokurov@fa.ru*

Декларация об отсутствии личной заинтересованности: *благополучие автора не зависит от результатов исследования.*