DOI 10.15276/imms.v15.no4.496
UDC 004.932.1

Informatics and Mathematical Methods in Simulation
Vol.15 (2025), No. 4, pp. 496-506

MODIFICATION OF THE MASK R-CNN ARCHITECTURE FOR IMAGE DETECTION AND SEGMENTATION

N. Volkova, M. Shvandt

National Odesa Polytechnic University
1, Shevchenko Ave., Odesa, Ukraine, 65044
Emails: volkova.n.p@op.edu.ua, maxim.shvandt@gmail.com

The paper addresses the problem of detecting and segmenting animal images. An analysis of neural network architectures for detecting and segmenting objects is carried out. A neural network architecture is proposed for detecting and segmenting images with a fuzzy background and partial overlapping of objects based on a modification of the Mask R-CNN architecture, which demonstrates a sufficiently high evaluation indicators of accuracy and quality of segmentation and is able to utilize additional image features of multichannel images. The main elements of the proposed architecture are a double branch of a feature extractor with feature fusion that uses additionally obtained image features. The proposed architecture was tested on a set of test images of experimental animals. The results of detecting and segmenting experimental animals by the proposed architecture and several basic Mask R-CNN variants were compared. The segmentation quality was assessed using the accuracy (Accuracy, Precision), completeness (Recall) metrics. Based on experimental studies, it was determined that training a modification of the Mask RCNN architecture for 50 epochs allows obtaining sufficiently high indicators of quality and accuracy of detection and segmentation, namely: accuracy (Accuracy) - 0.9, precision (Precision) - 0.92, completeness (Recall) - 0.92, while maintaining basic operability. Mask R-CNN variants with ResNet18/34 feature extractors have lower accuracy, and basic Mask R-CNN with ResNet50/101 have significantly larger sizes without the possibility of using additional image features. Thus, the architecture proposed in the work is effective for tasks of detection and segmentation of objects that require high accuracy and quality of their localization in the image.

Keywords: neural network; Mask R-CNN, architecture; object detection; segmentation, object tracking; evaluation indicators

Introduction. In recent decades, the systematic study of animal behavior has assumed increasing importance across disciplines including neuroscience, ethology, environmental sciences, and agricultural research. Behavioral responses to environmental stimuli provide not only key indicators of animal health and welfare but also serve as sensitive proxies for broader ecological dynamics [1-3]. Parallel to these scientific needs, advances in computational technologies, most notably in computer vision, have enabled the quantification and analysis of behavior with levels of precision, reproducibility, and scalability that were previously unattainable [4-5].

Historically, behavioral studies relied heavily on manual observation or basic sensor systems. Such approaches, while being foundational, were very labor-intensive, constrained in temporal and spatial resolution, and susceptible to observer bias [4]. The emergence of high-resolution imaging, sophisticated motion analysis, and automated tracking algorithms has transformed this landscape, allowing continuous and non-invasive monitoring of animal activity across diverse experimental and natural contexts. Although deep learning has become one of the leading tools in the field, many influential studies continue to use classic computer vision techniques such as background subtraction, contour detection, and trajectory clustering, which remain effective and interpretable for a wide range of behavioral analyses [7,8].

However, the adoption of neural networks has rapidly expanded the scope of computational ethology. Deep learning frameworks, especially, convolutional and recurrent

architectures, are increasingly used for pose estimation, fine-grained action recognition, and multi-animal tracking. These methods allow researchers to capture subtle behavioral nuances, integrate multimodal data streams, and generalize across species and contexts [9]. Toolkits such as DeepBehavior demonstrate how deep learning can be applied to both animal and human behavior imaging data, providing accessible pipelines for neuroscience and ethology [9]. Recent comprehensive studies have highlighted the breadth of applications of deep learning in the study of animal behavior, ranging from bioacoustics to video tracking, and have outlined challenges and opportunities for future research [10]. Neural networks are also being leveraged to link behavioral signatures with physiological or genetic data, offering a powerful bridge between observable actions and underlying biological mechanisms.

The growing popularity of automated behavioral analysis is driven by a combination of technological readiness and scientific necessity. Global challenges such as climate change, disease emergence, and food security demand real-time tools to assess animal responses to environmental pressures [11]. Moreover, behavior often reflects underlying physiological or cognitive states more rapidly than biochemical markers can detect. Automated behavioral metrics therefore provide a non-invasive view into animal welfare and can even serve as early -warning systems for stress or illness [12,13]. Together, these developments underscore the central role of computational methods, ranging from traditional vision algorithms to state-of-the-art neural networks, in shaping the future of animal behavior research.

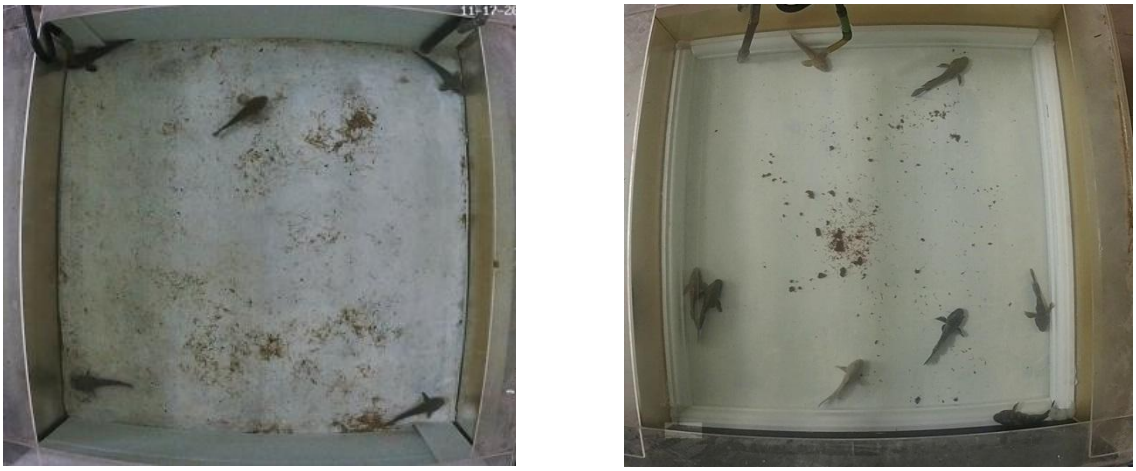
Analysis of recent publications and formulation of the problem. As noted earlier, mice and fish are among the most commonly employed model organisms in ecotoxicology and ethology. A substantial portion of such research relies on the study of animal behavior under controlled, enclosed conditions, particularly during the initial stages of experimentation. For rodents, the test setup involves a perforated test box designed to stimulate exploratory activity. A fixed overhead camera is positioned above the apparatus to continuously record the animals over extended periods, ranging from 30 minutes to several hours. Behavioral observations in this context often focus on quantifying movements between holes and documenting investigatory attempts to peer into them, which are interpreted as indicators of exploratory drive and territorial assessment (Fig. 1a–1b).

a *b*

Fig. 1. Mice/rats behaviour study: *a* – screenshot of a video with a lab rat, *b* – screenshot of a video with lab mice

In our particular case the behavior of fish (especially bullheads) is of considerable interest to the researchers. The enclosed test environment is represented by a square aquarium filled with real sea water to simulate the real environmental conditions. The aquarium is also supplied with oxygen via pipes, and a camera is mounted above the aquarium and records the bullheads over a period of time (from several hours to one day). The video recording is divided into 30-minute video segments to facilitate further analysis (Fig. 2a-2b).

In terms of the experiment here are up to 10 subjects supposed to be placed inside the aquarium. The behaviour patterns that currently are of greatest interest to the researchers among other ones are: the general number of movements (position changes) for each subject; the general number of movements for complete test group; number of attacks/conflicts between 2 subjects (a movement of subject A towards subject B that results in subject B escaping in other direction); keeping each object`s ID during the complete recording session. This is particularly important in case of placement of 2 or 3 different kinds of bullheads into the aquarium.



a *b*

Fig. 2. Fish (bullheads) behaviour study: *a, b* – screenshots from videos with bullheads in the test environment (aquarium)

Currently, analysis of video sequences is performed manually by laboratory personnel, which is a time-consuming and insufficiently accurate process, since behavioral patterns are formed subjectively. This highlights the need for an automated algorithm capable of detecting, tracking, and analysing subject movements with greater efficiency and precision. Developing such a algorithm, however, is complicated by several factors. Although the position of a subject typically does not shift dramatically between two or three consecutive frames, its direction of movement can change abruptly, making trajectory estimation challenging. In addition, while the experimental environment is generally stable, the aquarium background is subject to dynamic changes. Food can introduced before and during test sessions, and the by-products of bullhead activity accumulate on the aquarium floor. These sediments often migrate in the form of clumps under the influence of water currents generated by swimming fish and aeration from oxygen tubes. Because these clumps may resemble the fish in colour and can grow up to considerable size, distinguishing the animals when they swim above such deposits becomes difficult. This variability also precludes the use of simple strategies such as colour-based tracking. Therefore, all these aspects must be taken into account when developing a robust and reliable tracking algorithm.

Another important problem during detection and tracking before performing object identification is the separation of objects that may merge into one spot when approaching each other. That is why it was decided to add a neural detector capable of segmenting objects to the proposed object detection and tracking algorithm, along with basic detection methods such as background subtraction [14].

Object detection with segmentation, often referred to as instance segmentation, is a central task in computer vision. Unlike pure object detection, which predicts bounding boxes and class labels, instance segmentation requires delineating each object at the pixel level. This dual requirement makes the task more challenging but also more useful in domains such as

autonomous driving, medical imaging, and video surveillance. Over the past decade, researchers have proposed a variety of neural network architectures to address this problem. This overview examines four representative families: Mask R‑CNN, YOLACT, SOLO/SOLOv2, and DetectoRS, while also briefly noting related approaches such as CondInst and Panoptic FPN. Each architecture embodies different trade‑offs between accuracy, speed, and complexity.

Mask R-CNN [15] is the most widely recognized architecture for instance segmentation. It extends Faster R-CNN [16], a two-stage object detector, by adding a parallel branch dedicated to mask prediction. The model begins with a convolutional backbone, typically ResNet-50 or ResNet-101, augmented with a Feature Pyramid Network (FPN) [17] to capture multi-scale features. A Region Proposal Network (RPN) then generates candidate object regions, which are refined through RoIAlign, a pooling method that avoids the misalignments caused by quantization in earlier designs. For each region of interest, the network predicts class labels, bounding box coordinates, and a binary mask. This modular design allows Mask R-CNN to achieve high accuracy across benchmarks such as COCO, and its flexibility has made it a standard baseline in both research and industry. However, the two-stage nature of the model makes it computationally heavy and relatively slow, which can be a limitation in real-time applications. In contrast to the two-stage paradigm, YOLACT [18] demonstrates that instance segmentation can be achieved in real time. Instead of predicting masks for each region of interest, YOLACT generates a set of prototype masks for the entire image and then linearly combines them with per-instance coefficients predicted by the detection branch. This one-stage design is conceptually simpler and avoids the overhead of region proposals and RoI operations. The result is a system that can run at real-time speeds on modern GPUs, making it attractive for applications such as robotics or video analysis where latency is critical. The trade-off is that YOLACT generally achieves lower accuracy than two-stage methods, particularly on small or overlapping objects, and its prototype mask approximation can blur fine boundaries. Nevertheless, it still represents a step toward balancing speed and accuracy in instance segmentation.

SOLO (Segmenting Objects by Locations) [19] and its successor SOLOv2 [20] take a different approach by reformulating instance segmentation as a direct classification problem on a spatial grid. The image is divided into grids, and each grid cell predicts whether it belongs to an object instance and outputs a mask kernel to generate the segmentation. This anchor‑free, grid‑based formulation eliminates the need for proposals or RoI pooling, making the pipeline more straightforward. SOLOv2 improves upon the original by introducing dynamic convolution, which enhances both speed and accuracy. These models demonstrate that instance segmentation can be addressed in an end‑to‑end manner without the complexities of proposal generation. However, the grid‑based formulation can be sensitive to object scale and placement, and dense scenes with many overlapping objects remain challenging. Despite these limitations, SOLO and SOLOv2 highlight the potential of anchor‑free methods and have influenced subsequent research. As an alternative, DetectoRS [21] represents the high‑accuracy, high‑complexity end of the spectrum. Built upon Cascade Mask R‑CNN [22], it introduces two key innovations: Recursive Feature Pyramid (RFP) and Switchable Atrous Convolution (SAC). RFP enhances multi‑scale feature representation by feeding FPN outputs back into the backbone, allowing recursive refinement of features. SAC adapts receptive fields dynamically by switching between different atrous rates, enabling the network to capture varied contextual information. Together, these innovations significantly improve performance, and DetectoRS has achieved state‑of‑the‑art accuracy on COCO

benchmarks. The cost of this performance is computational expense: the model is very resource‑intensive, requires significant GPU memory, and is unsuitable for real‑time applications. Its complexity also makes it harder to train and deploy compared to simpler designs. Nonetheless, DetectoRS demonstrates how architectural innovations can push the boundaries of accuracy in instance segmentation.

Instance segmentation has progressed rapidly, moving from the proposal‑based, two‑stage paradigm of Mask R‑CNN to real‑time one‑stage models like YOLACT, to anchor‑free formulations like SOLOv2, and to advanced recursive designs like DetectoRS. Each architecture reflects a different balance between accuracy, speed, and complexity, and each has influenced subsequent research. The diversity of approaches underscores the richness of the problem and the ongoing search for architectures that can deliver both precision and efficiency. As applications of computer vision expand, the demand for models that can perform accurate, real‑time instance segmentation will continue to drive innovation, likely leading to architectures that combine the strengths of current paradigms while mitigating their weaknesses.

The considered models are initially designed for typical images with 3 channels without additional features. Also, not all architectures have a flexible modular structure. Therefore, the goal of the study is to create an architecture that could use the advantages of the presence of additional image features while maintaining the overall accuracy of the basic options with a smaller structure/size of the model (compared, for example, to the basic Mask R-CNN with ResNet101) along with preserving the relative modularity for further modifications if necessary. To achieve the set goal, it is necessary to solve the following main tasks: analysis of the architectures of object detection and segmentation networks; development of the basis of a feature extractor that will use additional image feature channels; experimental study of the developed architecture.

Main material. Image processing within the proposed approach [14] includes several sequential stages: preprocessing, primary segmentation (for further background filtering), detection, and analysis. During image preprocessing and earlier detection stages, such as background subtraction [14], we are able to obtain additional image features in form of new image channels, such as the result of background subtraction/image with CLAHE or shadow correction. These channels contain additional image representations with highlighting specific image aspects. Thus it was decided to utilize these additional channels to enhance object detection and segmentation in terms of combined detector. The basic Mask R-CNN was chosen for it general accuracy and primarily, for its modularity since currently it is not required to adopt the detection for real-time usage. Also there was an aim to create an architecture that with the utilization of additional channels would be smaller in terms of basic Mask R-CNN with ResNet101 backbone or even ResNet50.

Since using a separate backbone for additional image data is already a valid approach for extra data handling [23,24], it was decided to use to parallel feature extraction branches in the backbone, one for main image channels (Branch A) and one for feature (engineered) channels (Branch B). Since additional channels are refined and filtered from noise, to enhance feature constructions the feature fusion was applied from branch B into branch A. Among different fusion technics [25,26] it was currently decided to utilize SE-based fusion [27] since using it one will stick to data-aware mixing: after concatenation, the SE gate learns which channels (from A or B) are useful per stage and per image, instead of blindly summing. Also it is robust to scale/statistics mismatch as the channel attention can down-weight noisy or poorly scaled auxiliary features (e.g., depth/engineered bands) before they enter the top-down pathway and is Lightweight & stable, as the squeeze-excite MLP adds very few parameters/compute and preserves spatial structure, which is good for plugging into FPN

stages without upsetting optimization. Table 1 and Figure 3 show the general idea of merging two branches within the backbone.

The SE-based fusion performs:

a) concatenation of features from the main (A) and aux (B) branches at the same spatial scale:

$$U = \mathrm{Concat}\left(A, B\right) \in \mathrm{R}^{H\times W\times C},\ C = C_A + C_B,$$

where *H,W,C* are height, width and number of channels.

b) feature squeezing (global context per channel):

$$z_c = \frac{1}{HW}\sum_{i=1}^{H}\sum_{j=1}^{W} U_{i,j,c} \Rightarrow z \in \mathrm{R}^{C};$$

c) learning of channel importances with a bottleneck set by *r* (SE-ratio)

$$s = \sigma\left(W_2\delta\left(W_1 z\right)\right),\ W_1 \in \mathrm{R}^{\frac{C}{r}\times C},\ W_2 \in \mathrm{R}^{C\times\frac{C}{r}},$$

where *r* is SE-ratio, $\delta = \mathrm{ReLU}$, $\sigma = \mathrm{Sigmoid}$.

d) channels reweighting:

$$\tilde{U}_{i,j,c} = s_c \cdot U_{i,j,c}.$$

Table 1.

General dual-branch fused feature extractor backbone

Stage	Spatial stride	A-branch (main)	B-branch (aux)	Fusion points (B→A)	Output for FPN
Input	1×	Split 6-ch → A: ch0–2, B: ch3–5	same	–	–
C1 (stem)	2×	conv7×7/2, BN, ReLU → 64ch; maxpool/2	same	–	–
C2 (layer1)	4×	ResNet-18 layer1 → 64ch	64ch	–	C2(A) → goes to FPN lateral 1×1
C3 (layer2)	8×	128ch	128ch	Stage-entry fuse at C3: B→A (se_concat)	C3(A′) → FPN lateral
C4 (layer3)	16×	256ch	256ch	Blockwise fuse inside C4 (after blocks)	C4(A′) → FPN lateral
C5 (layer4)	32×	512ch	512ch	Blockwise fuse inside C5 (after blocks)	C5(A′) → FPN lateral

Research results. To assess overall performance and applicability in more realistic conditions, all training and testing sessions were conducted on locally available hardware. Test experiments were conducted on a laptop with an Intel Core i9-13980HX CPU, 64GB RAM and a single NVidia GeForce RTX 4090 Laptop GPU. As a relative primary test the basic Mask RCNN variant with several single backbones and training parameters were trained and tested. They include the standard configuration with ResNet50 and ResNet101 backbones, as well as a version with smaller backbone, such as Resnet18 and ResNet34. Due to GPU memory limitations the batch size has to be altered in order to perform training on available hardware. The number of epochs was fixed at value when model accuracy stopped increasing significantly.

For basic Mask RCNN architectures a dataset with JPG-images was used. It was separated into 340 train and 13 validation images (the dataset was augmented using image rotations). The example benchmarks are presented on table 2. The architecture implementation from [28,29] was used as an additional source of implementation example.

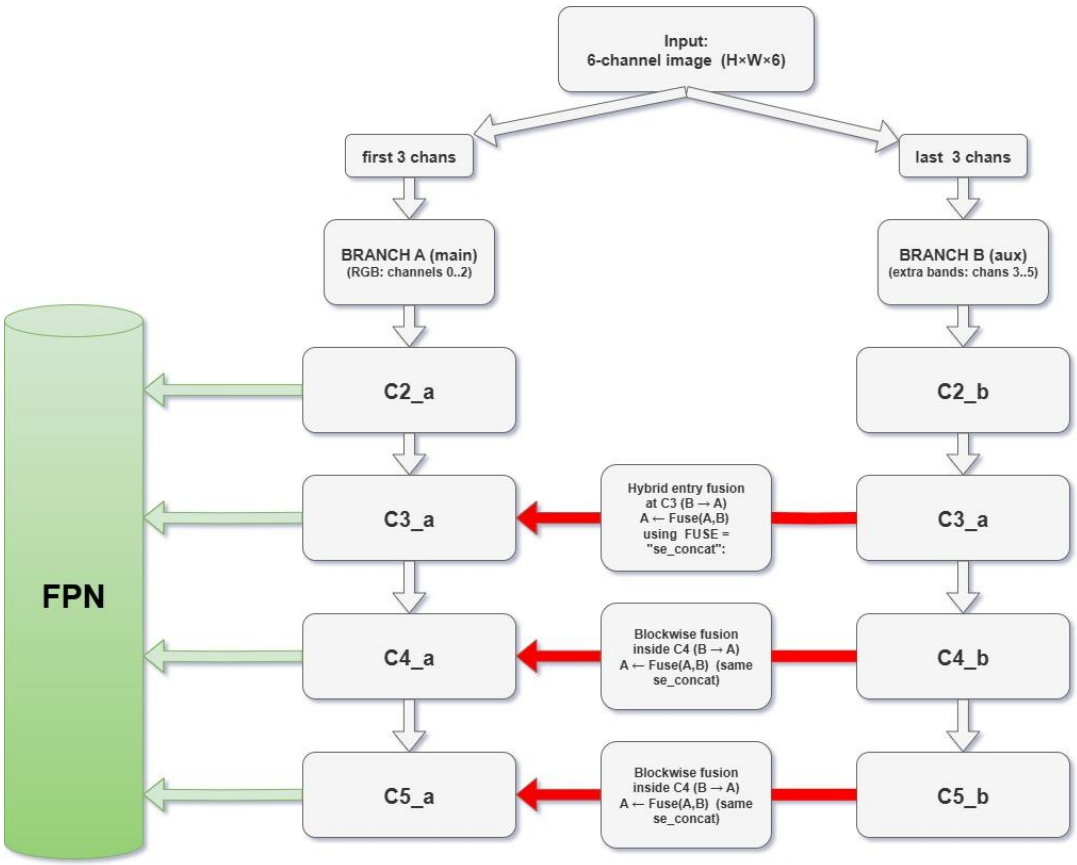


Fig. 3. Visualization of the general scheme of fusion of two branches inside the backbone

The elements of the confusion matrix were used to calculate segmentation quality scores: Accuracy (1), Precision (2), Recall (3) [30]:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}, \tag{1}$$

$$\text{Precision} = \frac{TP}{TP + FP}, \tag{2}$$

$$\text{Recall} = \frac{TP}{TP + FN}. \tag{3}$$

Table 2.

Basic Mask R-CNN architecture training and benchmarking

Architecture	Backbone	Batch size	Training epochs	mAP50	AVG Inference ms (excluding model 1st sample warm-up)	Precision	Recall
Mask R-CNN	ResNet101	2	30	0.92	102.5	0.963	0.939
Mask R-CNN	ResNet50	4	30	0.92	101.8	0.974	0.927
Mask R-CNN	ResNet34	6	30	0.84	~91	0.89	0.89
Mask R-CNN	ResNet18	4	30	0.83	104.9	0.86	0.902

For the developed dual-branch feature extractor architecture, the number of training epochs was increased to 50. The benchmark is shown on table 3. Figure 3 shows the example detection result.

Table 3.

Dual Fused ResNet18 backbone Mask R-CNN architecture benchmarking

Architecture	Backbone	Batch size	Training epochs	mAP50	AVG Inference ms (excluding model 1st sample warm-up)	Precision	Recall
Dual Branch Fused backbone Mask R-CNN	ResNet18	4	50	0.9	105.5	0.915	0.915

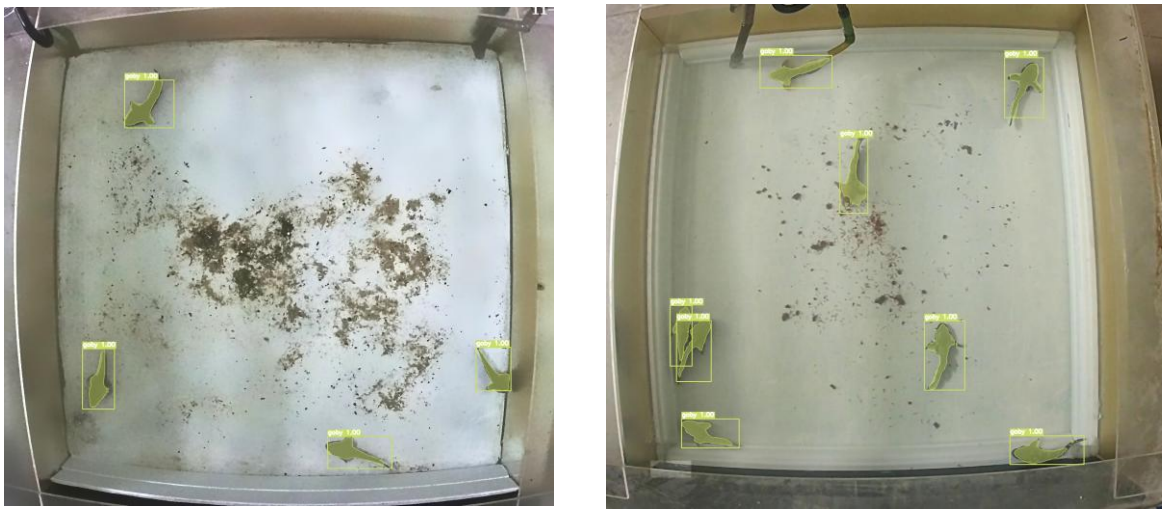


Fig. 3. Dual fused backbone detection visualization

Conclusions. This research proposes a neural network architecture for detecting and segmenting images with a fuzzy background and partial overlapping of objects based on a modification of the Mask R-CNN architecture. The accuracy and quality of object detection and segmentation in images using the proposed modified architecture of the Mask R-CNN neural network with a dual feature extractor on multi-channel images were studied. To increase the volume of the training sample, the paper performed a data augmentation procedure, the volume of which was increased by 4 times. The segmentation quality was assessed using an confusion matrix, based on the elements of which segmentation accuracy indicators (Accuracy, Precision, Recall) were calculated. Comparative analysis showed that the proposed modified architecture of the Mask R-CNN network with a smaller number of parameters compared to the variants with one ResNet101/ResNet50 shows accuracy and speed at the level of these architectures. Also, the indicators of the modified architecture exceed the accuracy indicators for the variants with ResNet18/34. The presence of an additional branch for additional features also allows working with images with more than 3 channels for a wider use of all image features. Further research will be aimed at additional improving of the architecture accuracy and to execution of object identification.

Akwnoledgments. Special thanks for the research assistance and provided test videos and images of lab animals to Faculty of Biology of Odesa I.I. Mechnikov National University.

References

1. Tinbergen N. On aims and methods of ethology. *Zeitschrift für Tierpsychologie*. 1963. 20(4). 410‑433. URL:https://doi.org/10.1111/j.1439-0310.1963.tb01161.x

2. Dawkins M. Behavior as a tool in welfare assessment. *Applied Animal Behaviour Science*. 2004. V.86(3-4). P.227-233. URL:https://doi.org/10.1016/j.applanim.2004.02.001
3. Guo C., Chen Y., Ma C., Hao S., Song J. A survey on AI-driven mouse behavior analysis applications and solutions. *Bioengineering*. 2024. V.11(11). P.1121. URL:https://doi.org/10.3390/bioengineering11111121
4. Feng J.-X., Li P., Liu Y., Liu L., Li Z.H. A latest progress in the study of fish behavior: Cross-generational effects of behavior under pollution pressure and new technologies for behavior monitoring. *Environmental Science and Pollution Research*. 2024. V.31. P. 11529-11542. URL:https://doi.org/10.1007/s11356-024-31885-2
5. Dell A., Bender J., Branson K., Couzin I., Polavieja G. Automated image-based tracking and its application in ecology. *Trends in Ecology & Evolution*. 2014. V.29(7). P.417-428. URL:https://doi.org/10.1016/j.tree.2014.05.004
6. Anderson, D., Perona, P. Toward a science of computational ethology. *Neuron*. 2014. V.84(1). 18-31. https://doi.org/10.1016/j.neuron.2014.09.005
7. Yin Z., Xiao L., Ma R., Han Z., Li Y. Detecting abnormal animal behaviors using optical flow and background subtraction. *Computers and Electronics in Agriculture*. 2020. 174. P.105471. URL:https://doi.org/10.1016/j.compag.2020.105471
8. Beyan C., Fisher R. Animal behavior recognition using spatio-temporal features. *Pattern Recognition*. 2018. No.76. P.12-22. URL:https://doi.org/10.1016/j.patcog.2017.10.008
9. Arac A., Zhao P., Dobkin B. H., Carmichael S. T., Golshani P. DeepBehavior: A deep learning toolbox for automated analysis of animal and human behavior imaging data. *Frontiers in Systems Neuroscience*. 2019. No. 13. P. 20. URL: https://doi.org/10.3389/fnsys.2019.00020
10. Fazzari E., Romano D., Falchi F., Stefanini C. Animal behavior analysis methods using deep learning: A survey. *arXiv preprint*. 2023. URL: https://arxiv.org/abs/2405.14002
11. Manteuffel G., Puppe B., Schön P., Bruckmaier R., Janssen D. Sensor-based analysis of animal behavior. *Animal*. 2009. No.3(9). P.1197 - 1204. URL: https://doi.org/10.1017/S1751731109004526
12. Neethirajan S. Recent advances in wearable sensors for animal health management. *Sensors and Biosensors Research*. 2017. No.20. P.1 - 11. URL: https://doi.org/10.1016/j.sbsr.2018.02.004
13. Spampinato, C., Palazzo, S., Boom, B., Lin, H., Wei, J., et al. Understanding fish behavior during typhoon events in real-life underwater environments. *Multimedia Tools and Applications*. 2014. 70(1). 199-236. https://doi.org/10.1007/s11042-012-1101-5
14. Volkova, M., Shvandt. Segmentation-based approach for object detection. *Proceedings of Odessa Polytechnic University*. 2025. 1(71). P. 145 - 156. https://doi.org/10.15276/opu.1.71.2025
15. He K., Gkioxari G., Dollár P., Girshick R. Mask R-CNN. *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*. 2017. URL: https://doi.org/10.48550/arXiv.1703.06870
16. Ren S., He K., Girshick R., Sun J. Faster R-CNN: Towards real-time object detection with region proposal networks. *Advances in Neural Information Processing Systems (NeurIPS)*. 2015. URL: https://doi.org/10.48550/arXiv.1506.01497

17. Lin T.Y., Dollár P., Girshick R., He K., Hariharan B., Belongie S. Feature pyramid networks for object detection. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2017. URL: https://doi.org/10.48550/arXiv.1612.03144
18. Bolya D., Zhou C., Xiao F., Lee Y.J. YOLACT: Real‑time instance segmentation. *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*. 2019. URL: https://doi.org/10.48550/arXiv.1904.02689
19. Wang X., Kong T., Shen C., Jiang Y., Li L. SOLO: Segmenting objects by locations. *Proceedings of the European Conference on Computer Vision (ECCV)*. 2020. URL: https://doi.org/10.48550/arXiv.1912.04488
20. Wang X., Zhang R., Kong T., Li L., Shen C. SOLOv2: Dynamic and fast instance segmentation. *Advances in Neural Information Processing Systems (NeurIPS)*. 2020. URL: https://doi.org/10.48550/arXiv.2003.10152
21. Qiao S., Chen L.C., Yuille A. DetectoRS: Detecting objects with recursive feature pyramid and switchable atrous convolution. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2021. URL: https://doi.org/10.48550/arXiv.2006.02334
22. Cai Z., Vasconcelos N. Cascade R‑CNN: Delving into high quality object detection. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2018. URL: https://doi.org/10.48550/arXiv.1712.00726
23. Gupta S., Girshick R., Arbelaez P., Malik J. Learning rich features from RGB‑D images for object detection and segmentation. *European Conference on Computer Vision (ECCV)*. 2014. P. 345‑360. URL: https://doi.org/10.1007/978-3-319-10584-0_23
24. Zhang Z., Zhang J., Bailo O., Vázquez D., Xu J., López A.M. A two‑branch feature fusion framework for RGB‑D pedestrian detection. *Remote Sensing*. 2022. No.14(3). P.645. URL: https://doi.org/10.3390/rs14030645
25. Woo S., Park J., Lee J.Y., Kweon I. S. CBAM: Convolutional block attention module. *European Conference on Computer Vision (ECCV)*. 2018. P.3 ‑ 19. URL: https://doi.org/10.1007/978-3-030-01234-2_1
26. Perez E., Strub F., de Vries H., Dumoulin V., Courville A. FiLM: Visual reasoning with a general conditioning layer. *International Conference on Learning Representations (ICLR)*. 2018. URL: https://doi.org/10.48550/arXiv.1709.07871
27. Hu J., Shen L., Sun G. Squeeze‑and‑excitation networks. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2018. P. 7132‑7141. URL: https://doi.org/10.1109/CVPR.2018.00745
28. Mask R‑CNN for object detection and instance segmentation on Keras and TensorFlow. *GitHub repository*. 2017. URL: https://github.com/matterport/Mask_RCNN
29. Waleed A., Mask R. CNN: Train on the Balloon Dataset and Run Color Splash. *Medium*. 2018. URL: https://engineering.matterport.com/splash-of-color-instance-segmentation-with-mask-r-cnn-and-tensorflow-7c761e238b46
30. Sathyanarayanan S., Roopashri-Tantri B. Confusion Matrix ‑ Based Performance Evaluation Metrics. *African Journal of Biomedical Research*, 2024. No.27(4S). P.4023-4031. DOI: 10.53555/AJBR.v27i4S.4345

N. Volkova, M. Shvandt

МОДИФІКАЦІЯ АРХІТЕКТУРИ MASK R-CNN ДЛЯ ДЕТЕКТУВАННЯ ТА СЕГМЕНТАЦІЇ ЗОБРАЖЕНЬ

Н. Волкова, М. Швандт

Національний університет «Одеська політехніка»
1, Шевченка пр., Одеса, 65044, Україна
Emails: volkova.n.p@op.edu.ua, maxim.shvandt@gmail.com

У роботі розглянуто задачу детектування та сегментації зображень тварин. Проведено аналіз архітектур нейронних мереж для детектування та сегментації об’єктів. Запропоновано архітектуру нейронної мережі для детектування та сегментації зображень з нечітким заднім фоном та частковим перекриттям об’єктів на основі модифікації архітектури Mask R-CNN, яка демонструє достатньо високі показники точності та якості сегментації та здатна використовувати додаткові особливості багатоканальних зображень. Основними елементами запропонованої архітектури є подвійна гілка екстрактора ознак із злиттям ознак, що використовує додатково отримані ознаки зображення. Запропоновану архітектуру апробовано на наборі тестових зображень піддослідних тварин. Проведено порівняння результатів детектування та сегментації піддослідних тварин запропонованою архітектурою та декількома базовими варіантами Mask R-CNN. Оцінку якості сегментації виконано з використанням метрик точності (Accuracy, Precision), повноти (Recall). На основі експериментальних досліджень визначено, що навчання модифікації архітектури Mask R-CNN протягом 50 епох дозволяє отримати достатньо високі показники якості та точності детектування та сегментації, а саме: точність (Accuracy) - 0.9, точність (Precision) - 0.92, повнота (Recall) - 0.92, при зберіганні базової оперативності. Варіанти Mask R-CNN із екстракторами ознак ResNet18/34 мають меншу точність, а базові Mask R-CNN з ResNet50/101 мають значно більші розміри без можливості використання додаткових ознак зображення. Таким чином, запропонована в роботі архітектура є ефективною для задач детектування та сегментації об’єктів, які потребують високої точності та якості їхньої локалізації на зображенні.

Ключові слова: нейронна мережа; Mask R-CNN, архітектура; детектування об'єктів; сегментація; відстеження об'єктів; показники якості