Bladder Semantic Segmentation

Vadim Chernyshev[1[0000-0002-8713-9026]], Alexander Gromov[1,3[0000-0001-9818-3770]], Anton Konushin [1,2 [0000−0002−6152−0021]], and Anna Mesheryakova [3[0000-0002-2409-0018]]

[1] Lomonosov Moscow State University, Moscow, Russia
{vadim.chernyshev, alexander.gromov, anton.konushin}@graphics.cs.msu.ru
[2] NRU Higher School of Economics, Moscow, Russia
[3] Third Opinion Platform LLC, Moscow, Russia
{alexander.gromov, ceo}@3opinion.ai

Abstract. Obtaining information about the shape and volume of the bladder plays a significant role in determining the pathologies of this or- gan. To collect the relevant data, the first thing to do is to separate the bladder from the background on the ultrasound image. The article is de- voted to automation this process using an algorithm based on the Unet architecture with a pretrained imagenet encoder (encoder – ResNet50). The article gives a comparative analysis of some well-known methods in literature that improve the accuracy of the proposed algorithm. The qual- ity of the basic architecture has been improved by more than 4 percent on the PR AUC metric (from 84.49% to 89.62%) in the series of exper- iments with the help of automatic annotation of previously unmarked data. In addition, there are two important results showing practical ef- fectiveness of using the data from another medical task (which raised the accuracy to 88.50%) and using time dependent sequence of frames inside the video (raised the quality to 88.19%).

Keywords: Semantic segmentation · Pseudo Labeling · Bladder ultrasound · 3D convolution · Time dependency.

1 Introduction

The bladder is an organ that performs a very important function in a human body. Its walls are elastic and can stretch or contract depending on certain factors. As a result, parameters such as the shape and volume of the organ itself change. The analysis of these parameters plays a key role in determining the pathologies of the bladder.

Performing the analysis, most clinics use a transabdominal ultrasound image, which shows the entire organ and the surrounding anatomy. To collect information about the

* Publication is supported by RFBR grant № 19-07-00844.

volume and the shape of the bladder, the image of the organ itself must be separated from the background. Physicians with appropriate qualifications are the only people who can do this job. However, this task can be solved using automatic semantic segmentation methods that would allow significant reduction of the physicians' workload and could help them to devote more time to treating patients. At the same time, using an automatic system may reduce the risk of human error caused by fatigue and monotonous work.

The most promising approach for solving the problems of semantic segmentation of medical images is deep convolutional neural networks. It should be noted that the best results are currently obtained using algorithms based on the Unet [1] architecture, which was originally developed for this purpose. The main limitation of such models (especially in the medical field) is that they re- quire very large amounts of reliable training data. These data are accurate and tightly annotated images, which creation requires substantial human labor and experience.

The purpose of this work was to review and compare methods that can potentially improve the accuracy of the classical Unet network. We collected 400 videos of ultrasound of the bladder, each lasting 10 seconds (and 10 fps). Every 5 frames from these videos are taken and marked by specialists (physicians). The studied methods are primarily aimed at smart use of the provided data:

- There is a connection between the marked frames, they are linked in time. Therefore, it makes sense to try to use this dependency. There are series of experiments related to volumetric convolutions.
- Unannotated frames also have a feature – they are located between the marked frames of the video. Thus, they are very similar to annotated images, since unannotated frames are only a few fractions of a second away from them. This means that even a overfitted network would be able to mark them very well. As a result, we get a lot of new maximally realistic annotated frames (because they are fragments of a real ultrasound video). It should be mentioned that the error of the marking is close to an error that might have occurred if the physician had worked manually.
- What should we do if there are no redundant data? It is proved experimentally that encoder pretraining on ImageNet has a positive effect on the final accuracy and speed of network convergence. In the current work, the possibility of pretraining architecture on a dataset of a similar medical problem was considered. We also studied the effect of applying these data directly during network training.

In addition to implementation of the main ideas, experiments on the choice of augmentations, input resolution, optimizer, and approaches to changing the pace of learning were conducted.

2 Preparation for experiments

2.1 Prepared data

The reference collection used in this work was provided by Third Opinion Platform [9]. It contains 5270 annotated ultrasound images of the bladder taken from 400 videos. The example of bladder image and picture of its mask which were taken from a total sample is shown in Fig. 1. The annotation contains not only bladder images and its masks but also the information about the videos from which that images were taken. It also includes information about the position of the frame inside this video and information about the presence or absence of the bladder. Moreover, we have ultrasound videos themselves, containing about 18000 unannotated frames. For further experiments, all videos and corresponding frames were randomly allocated either to the training sample or to the test sample.

It is important to emphasize that all frames from the same video necessarily belong to only one of the samples, otherwise the purity of the experiment would be disturbed owning to false quality improvement of the algorithm work which occurs as a result of testing on images that are similar to the training ones.

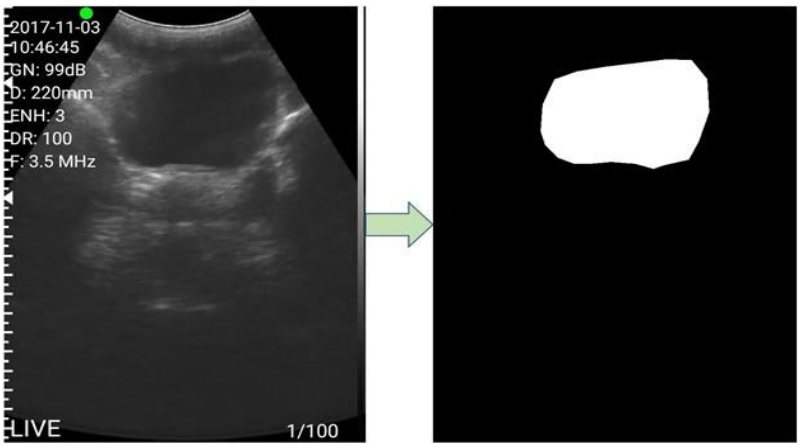

Fig. 1. On the left is the ultrasound image, on the right is the bladder mask.

2.2 Problem statement

The main purpose is to obtain the mask of the bladder. In practice, our algorithm should mark the bladder directly while the physician is working. Since the physician makes a whole ultrasound video, we have an opportunity to submit to the input of the algorithm not only the frame itself that needs to be marked (Fig. 2a), but also the whole series of pictures taken from the video. So, we can predict masks for both cases: masks for all frames (Fig. 2b) or mask only for the central one (Fig. 2c). However, the usage of more than one frame as an in- put reduces the scope of the algorithm. For example, in the case when it would be necessary to annotate only a single image, our algorithm (which

must use many frames as input) will not be able to manage this task. Then we consider all methods mentioned above, since the main aim of the work is to obtain an increase in accuracy.

All experiments were conducted on one video card GeForce GTX 1080 Ti.

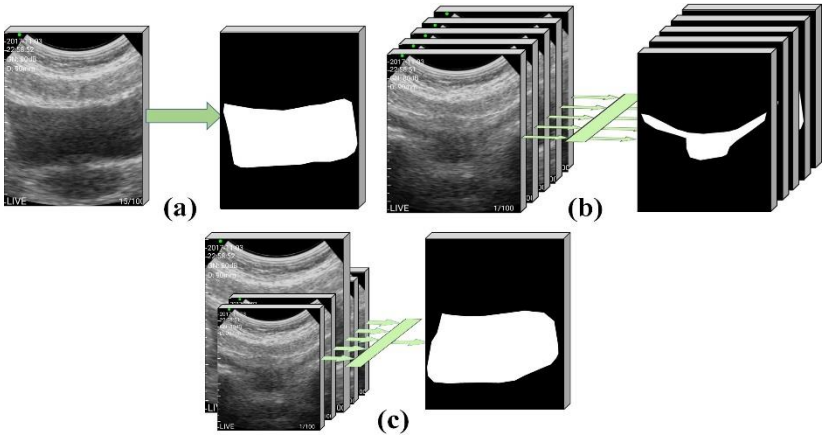

Fig. 2. a) The algorithm accepts one ultrasound image as input and gets a mask of this image as output. b) The algorithm accepts a series of images taken from the entire video at a fixed frequency as input and receives their masks as output. c) The algorithm accepts a frame and several surrounding images taken from the video as input and gets a mask for only the central image as output

2.3 Metrics

In the beginning, we considered the selected metrics which are necessary for evaluating the algorithm performance. One of the most well-known metrics used for evaluating the solution of the binary semantic segmentation problem is IoU. This metric is based on calculating truly classified positive pixels (TP), false- positive pixels (FP) and false-negative pixels (FN) (1):

$$IoU = \frac{TP}{TP + FP + FN}. \quad (1)$$

Alternatively, F-score, which is calculated as the harmonic mean between recall and precision, calculated by pixel, can be used (2):

$$Recall = \frac{TP}{TP + FN}, \quad (2)$$

$$Precision = \frac{TP}{TP + FP}, \quad (3)$$

$$F_{score} = \frac{2 \cdot Precision \cdot Recall}{Precision + Recall}. \quad (4)$$

However, all these metrics significantly depend on the threshold at which a decision is made about whether or not a pixel belongs to the bladder after the mask exits the neural network. To avoid binding to this parameter when choosing the best solution, we rely on generalization over all thresholds: on the size of the area under the Precision-Recall graph (PR AuC - Fig. 3).

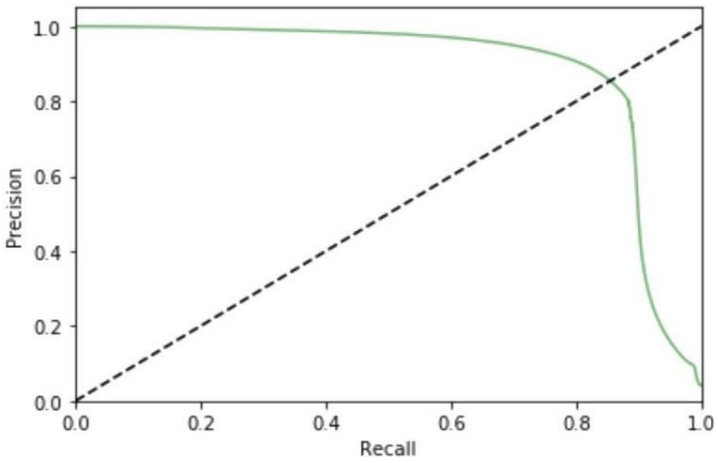

Fig. 3. Precision-Recall curve.

3 Related work

3.1 Basic architecture

During neural network training we usually change the network architecture itself or the data feed strategy, but we can also vary many other parameters, such as the optimizer, the learning rate, input resolution, and some others. It takes a tremendous amount of time to perform an enumeration of all possible combi- nations of these variables in each experiment. And the result of that work does not give a significant effect. So, to avoid these problems, the basic character- istics were selected experimentally. Such characteristics will be used in further experiments. Thus, a Unet-like architecture with a classifier in the form of a pretrained ResNet50 [8], a Novograd[2] optimizer (sometimes AdamW[3] was used instead), a cosine learning rate[4] and also with an input resolution of 128 * 128 or 256 * 256 pixels was chosen as the baseline.

3.2 Using Pseudo labels

One of the main problems arising with the training of neural networks is the difficulty of obtaining a training sample that covers the entire range of possible situations that may occur in the future when using the network in practice. We can significantly expand our sample by adding almost 18 thousand frames that can be parsed from the video. However, they are not marked, and their manual annotation is extremely difficult (it

takes a lot of time and physicians must be involved). At first glance, the annotation using our own network looks questionable – the accuracy of our best basic architecture reaches 85.68% for PR AUC (84.37% F-score).

It is logical to assume that using data annotated with an accuracy of 84% will not significantly increase the accuracy above this number. However, it should be noted that the data we want to annotate will be taken only from training videos (since using data from test videos can undeservedly improve the result on the test sample). These new frames are very similar to the surrounding training frames. Indeed, each new frame is separated from the annotated one by no more than 0.1 seconds (see Fig. 4), and the shape of the bladder smoothly changes over time (just like time dependence of the position of any other physical body).

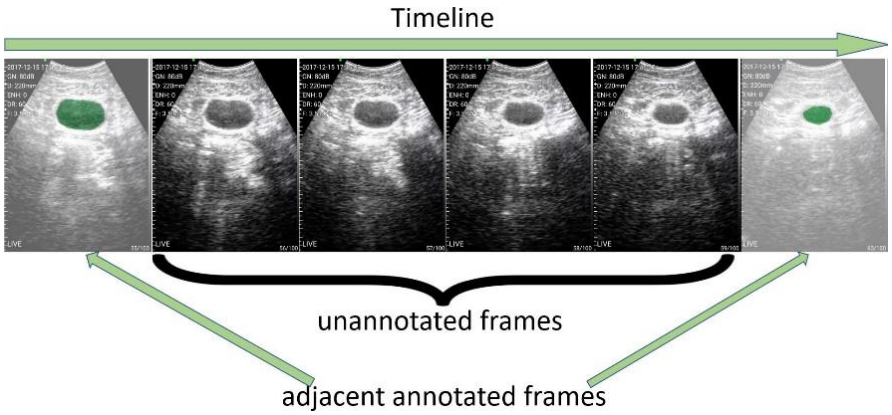

Fig. 4. Six consecutive frames from a single ultrasound video. Two frames are annotated (located on the edge), the central ones are to be annotated.

As a result, our network should annotate new frames with an accuracy close to the accuracy of its work on the training sample, which in turn is equal to 95.04%. In total, when we annotate new data with our network, we know that the accuracy of their annotation will lie in the range of [84.37, 95.04] percent. Moreover, it most likely tends to the right border of the value. This means that their use in further training should help to raise the accuracy of the algorithm to a value that lies within the presented range. By marking about 18000 images using a better network, we applied them while learning the same basic architecture. We added them to each batch of data in fixed portions. The result is shown in Table 1.

The table shows that the accuracy of work increases with any percentage of new data, up to very large values of 50 percent. This does not support the theory that the data was annotated with an accuracy close to 95%, but in any case, it confirms the effectiveness of usage of pseudo labels [10].

Table 1. Using Pseudo labels. This table presents a comparison of the accuracy of trained neural networks depending on what percentage of pictures in each batch was occupied by received pseudo labels.

Pseudo label (%)	PR AuC	Best F-score	F-score	IoU
0	0.8568	0.8513	0.8419	0.7364
10	0.8789	0.8665	0.8481	0.7432
20	0.8845	0.8710	0.8504	0.7466
25	0.8871	0.8732	0.8468	0.7431
30	0.8824	0.8693	0.8458	0.7398
35	0.8871	0.8729	0.8539	0.7513
40	0.8851	0.8717	0.8513	0.7482
45	0.8837	0.8713	0.8468	0.7410
50	0.8798	0.8680	0.8471	0.7418

3.3 Using data from another medical task

There are two main reasons for using data from another medical task. The first is the positive experience of usage classifiers, pretrained on the ImageNet task, in the Unet network. We would like to observe an increase in this effect when our algorithm is retrained on the most approximate medical task. The second is a potential solution to the problem of overfitting. During training, the accuracy of our network in the training sample reaches 95 percent and continues to grow, while the accuracy in the control sample begins to decrease over time. The reason for this is memorization of training sample data, that is, overfitting. We suggest that stirring each batch during training with data from a similar medical task may slightly weaken this effect. Behind this is the following heuristics: the network receives every batch which contains the data that differ in meaning from an ultrasound of the bladder. If we assume that the network is overfitted so much that it begins to poorly annotate slightly different frames from the test sample of the bladder ultrasound, then it should be even worse to annotate the pictures of another task. This means that it will receive a fine in the form of a large loss function. Otherwise, a similar medical task (for example, abdominal ultrasound snaps) can help the network identify new useful patterns. These patterns can increase the final accuracy and they will help avoid a heavy load such as marking something very new. For example, training the neural network in two completely different tasks - marking the bladder in ultrasound images and brain tumor in MRI is unlikely to be effective (since the neural network should simultaneously know signs that are practically unrelated to each other)

An open data set from the task of finding the circumference of the fetal head on an ultrasound image of the abdominal cavity was taken for further experiments [5][6].
This set contains 1000 training frames on which the circumference of the fetal head is marked. We approximated new data closer to our task by manually re-marking entire head area on new data (Fig. 5).

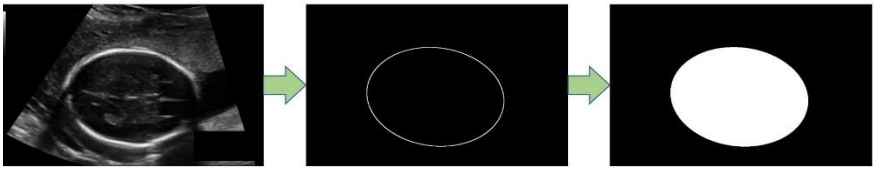

Fig. 5. a) an ultrasound image of the abdominal cavity taken from the new dataset b) corresponding masks c) Re-marked masks (these modification makes new data similar to ours).

It was decided to choice this set, because it is both visually as close as possible to our data, and as similar as possible in the medical sense (abdominal area, ultrasound). The results of a series of experiments are shown in Table 2. It should be noted that the best results were obtained by combining two approaches: pretraining on new data and their further use during training.

Table 2. New dataset. This table presents a comparison of algorithms obtained through various uses of the new dataset. The last line also shows an algorithm trained using pseudo labels.

Experiment	PR AuC	Best F-score	F-score	IoU
-	0.8568	0.8513	0.8419	0.7363
pretraining on new data	0.8613	0.8611	0.8357	0.7275
diluting by new data	0.8811	0.8695	0.8453	0.7398
pretraining + diluting	0.8850	0.8702	0.8474	0.7475
diluting by pseudo labels	0.8871	0.8729	0.8539	0.7513

3.4 Using a time dependency

In medicine it is important not only to mark a two-dimensional image but also to construct a volumetric segmentation map. For example, BraTS Brain Tumor Segmentation. One of the best solutions of such tasks is methods based on the 3D Unet architecture [7]. The main difference between 3D Unet architecture and the classical Unet is the replacement of two-dimensional convolutions with volumetric ones (Fig. 6). These convolutions allow to determine spatial dependencies in all three directions - width, height and depth.

However, the last of these spatial directions can be replaced by a temporal one. So, we have videos in which all annotated frames are separated from each other by certain time intervals. We tried to use a tensor consisting of all annotated video frames arranged in a row as an input to the 3D Unet network. And we got a tensor of the same dimension containing masks of all submitted frames as an output (Fig. 2b). The results of the experiment are presented in Table 3, experiment 1. It should be noted that the lack of quality improvement could be due to the following factors:

- reducing the number of input data units by 20 times. Now there is only 1 tensor containing all 20 images used as an input earlier;
- lack of pretraining and classic Unet architecture, while in the 2D experiments pretrained ResNet50 were used as encoder.

We slightly changed the training strategy to avoid these negative aspects: we decided to submit not all annotated frames from a single video, but a single image and a certain number of frames going in front of this image in the video and the same number of frames going after it. So, now we need to get the mask only for the main frame (see Fig. 2c) and because of that surrounding area might be unannotated. And now the number of input data units is the same as before. The results are presented in Table 3, experiment 2. As you can see, the final quality has become better.

Table 3. 3D convolutions. This table compares three experiments based on the idea of three-dimensional convolution (multiple frames are fed to the networks). "Step" means the distance within the video between the taken adjacent frames.

experiment	step	PR AuC	Best F-score	F-score	IoU
1	-	0.7170	0.7152	0.7072	0.5546
2	1	0.8239	0.8239	0.7847	0.6765
	2	0.8394	0.8450	0.7774	0.6735
	4	0.8501	0.8546	0.7917	0.6835
	8	0.8311	0.8217	0.7817	0.6817
3	4	0.8819	0.8632	0.8397	0.7320

However, the second problem was not solved - we still had the lack of pre- training. To deal with it, we had to abandon 3D Unet and use 2D Unet with the ResNet50 encoder. So, to determine time dependence we added 3D base, which consists of a certain number of 3D convolutions and processes the original tensor, making it two-dimensional. Then the processed tensor is fed to 2D Unet. The results are presented in Table 3, experiment 3. This modification gave an even greater increase in quality and made it possible to prove that application of time dependence could be useful in similar tasks.

4 Analysis of the obtained algorithms

We have reviewed some methods that can improve the accuracy of the semantic segmentation of the bladder. Now we would like to show the most advantageous combinations of these methods and show the visual difference in their work.

The best accuracy was achieved by using pretraining on another medical task in conjunction with the use of pseudo labels during the training (“Best our 2D network”, table 4, line 1).

Another experiment also deserves attention. It uses a series of 3D convolutions, the output of which was eventually fed into a 2D Unet (“Best our 3D network”, table 4, line 2). This approach gives lower accuracy on the test sample (Fig. 6a), however, it has some advantages. So, for example, the network using one frame as an input (“Best our

2D network") often mislabels some very complex frames that have several shaded areas. And the 2D Unet with 3D base network ("Best our 3D network"), which also analyzes adjacent frames, marks them more correctly (Fig. 6b).

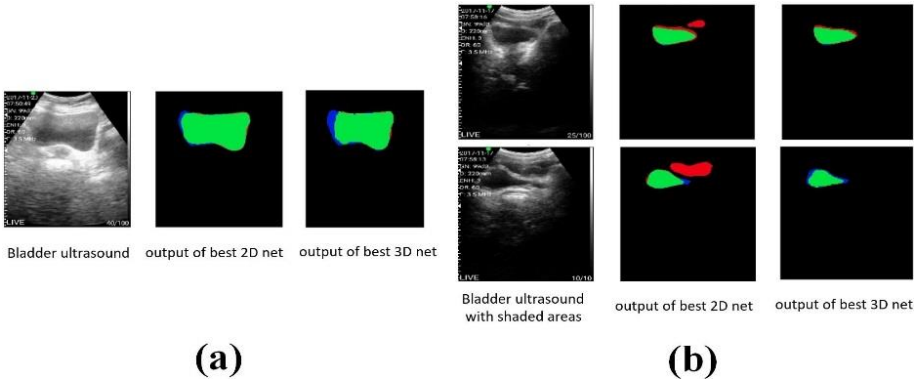

Fig. 6. Red means false-positive pixels, blue - false-negative pixels, green - true- positive, black - true-negative. a) 2D Unet generally performs slightly better results than 3D network. b) The complex samples for 2D Unet, but acceptable for 3D Unet

Table 4. Best our 2D network: encoder - resnet50, optimizer - NovoGrad, augmentation - random rotation (10 degrees), cosine learning rate, pretraining on a similar task, pseudo data 35%. Best our 3D network: encoder - resnet50, optimizer - NovoGrad, augmentation - random rotation (10 degrees), cosine learning rate.

Experiment	PR AuC	Best F-score	F-score	IoU
Best our 2D network	0.8962	0.8747	0.8654	0.7651
Best our 3D network	0.8819	0.8632	0.8397	0.7320
experiment	PR AuC	Best F-score	F-score	IoU

5 Conclusion

To sum up, in our work dedicated to bladder semantic segmentation we carried out a comparative analysis of well-known methods in literature that improve the accuracy of classical Unet network. Pseudo labels for unlabeled frames of the video were generated using a baseline trained on annotated frames from the same video. It was found that their further use during training of the same model provides a significant increase in quality of work by more than 4 percent. Another important conclusion is not only the potential usefulness of pretraining on data from a similar medical task, but also improving the quality of the bladder segmentation by adding this data directly to training.

References

1. Olaf Ronneberger and Philipp Fischer and Thomas Brox: U-Net: Convolutional Networks for Biomedical Image Segmentation, 2015.
2. Boris Ginsburg and Patrice Castonguay and Oleksii Hrinchuk and Oleksii Kuchaiev and Ryan Leary and Vitaly Lavrukhin and Jason Li and Huyen Nguyen and Yang Zhang and Jonathan M. Cohen: Training Deep Networks with Stochastic Gradient Normalized by Layerwise Adaptive Second Moments, 2020.
3. Ilya Loshchilov and Frank Hutter: Decoupled weight decay regularization, 2019.
4. Leslie N. Smith : Cyclical Learning Rates for Training Neural Networks, 2017.
5. Thomas L. A. van den Heuvel and Dagmar de Bruijn and Chris L. de Korte and Bram van Ginneken.: Automated measurement of fetal head circumference using 2D ultrasound images, 2018.
6. Dong-Hyun Lee: Pseudo-Label : The Simple and Efficient Semi-Supervised Learning Method for Deep Neural Networks, 2013.
7. Thomas L. A. van den Heuvel and Dagmar de Bruijn and Chris L. de Korte and Bram van Ginneken.: Automated measurement of fetal head circumference using 2D ultrasound images [Data set], 2018.
8. Ozgun C¸ i¸cek and Ahmed Abdulkadir and Soeren S. Lienkamp and Thomas Brox and Olaf Ronneberger: 3D U-Net: Learning Dense Volumetric Segmentation from Sparse Annotation, 2016.
9. Pavel Yakubovskiy: Segmentation Models Pytorch, 2020.
10. ”Third Opinion Platform” Limited Liability Company. URL: https://thirdopinion.ai/