National Research University Higher School of Economics
Faculty of Computer Science
Programme ‘Master of Data Science’

MASTER’S THESIS

Convolutional Neural Networks with Attention Mechanisms for Breast Cancer Segmentation on X-Ray Images

Student: Ivan Vassilenko

Supervisor: Ph.D., researcher Samsung R&D Institute & AI Center Russia, Mikhail Romanov

Moscow, 2022

Contents

Convolutional Neural Networks with Attention Mechanisms for Breast Cancer Segmentation on X-Ray Images

Abstract

Breast cancer is one of the major health problems in the world. X-ray imaging is a primary test for screening and diagnosis of breast cancer, and its effect was fully investigated.

The purpose of our work is to create a reliable convolutional neural network that will be able to do segmentation of a breast X-ray image to find regions that are allegedly cancerous.

We show that the use of different level features and the use of attention mechanism to merge information provides higher segmentation accuracy compared to baseline.

We have built the segmentation model using encoders of the standard backbones with pretrained weights, which makes it possible to use the advantages of transfer learning. Also, our method illustrates full use of the capabilities of a well-trained encoder compared to baseline. As a result, we achieve relatively high segmentation metrics on the small datasets (Dice coefficient of 79.5% on breast cancer segmentation CBIS-DDSM dataset, and mIoU of 81.268% on the validation set of Pascal VOC dataset).

1. Introduction

Breast cancer is one of the major health problems in the world.

Approximately 2.3 million women around the globe were diagnosed with breast cancer and 685 000 of them died in 2020. There were about 8 million women diagnosed with breast cancer in the past 5 years living in the world by the end of 2020. This accounted for the world's most prevalent cancer. Every woman in any given country is at risk of developing breast cancer after puberty. However, they have increased chances of developing one later in life.

In 2020 there were 684,996 deaths from breast cancer globally. In 2021, it is estimated 43,600 deaths from breast cancer in the United States, and 23 130 deaths in Russia in 2021 [1].

The main way to fight cancer is to identify and diagnose it at the earliest stage possible. A vast number of cancer screening programs have been developed in the world; each country has its own program. What these methods have in common is the need for an x-ray study. Full Field Digital Mammography is the only proven cancer screening method to date.

The analysis of the obtained images is a problem since it requires well-trained specialists with extensive expertise to diagnose the disease at the earliest stage possible. Such specialists in almost all countries are overloaded with work, because of this, the time required for making a diagnosis increases, sometimes reaching several weeks. If one needs to get a second opinion, this time stretches to a month.

In this regard, from the earliest times, medical society has been trying to come up with an automatic diagnosis method. The first work on the automation of cancer diagnostics used morphological methods for analyzing X-ray images, based on the selection of bright spots, which are the densest areas of tissue.

Recently, deep learning technology has appeared, which also affects the field of image processing. Deep learning image processing is known to be a good technology for automating X-ray diagnostics of diseases [3].

In this work, we consider the methods of segmenting the breast X-ray images using U-Net with skip connections. Skip-connections are designed to provide a signal to attention mechanism, which merges information from different level features in an efficient way and extracts only significant parts.

We show that our method outperforms baseline U-Net architecture on breast cancer segmentation, achieving Dice coefficient of 79.5%. Proposed method is suitable for primary reading of medical images without expert input because we use full image whereas most other methods use patches, cropped around suspicious regions.

We also tested our proposed architecture on a real-world dataset (Pascal VOC 2012 segmentation challenge), achieved mIoU 81.268%, we did not use additional data or pretrained weights on other segmentation datasets. In this case, our method shows the full use of the context capturing capabilities with a pretrained ResNetV2 Big Transfer (BiT) encoder.

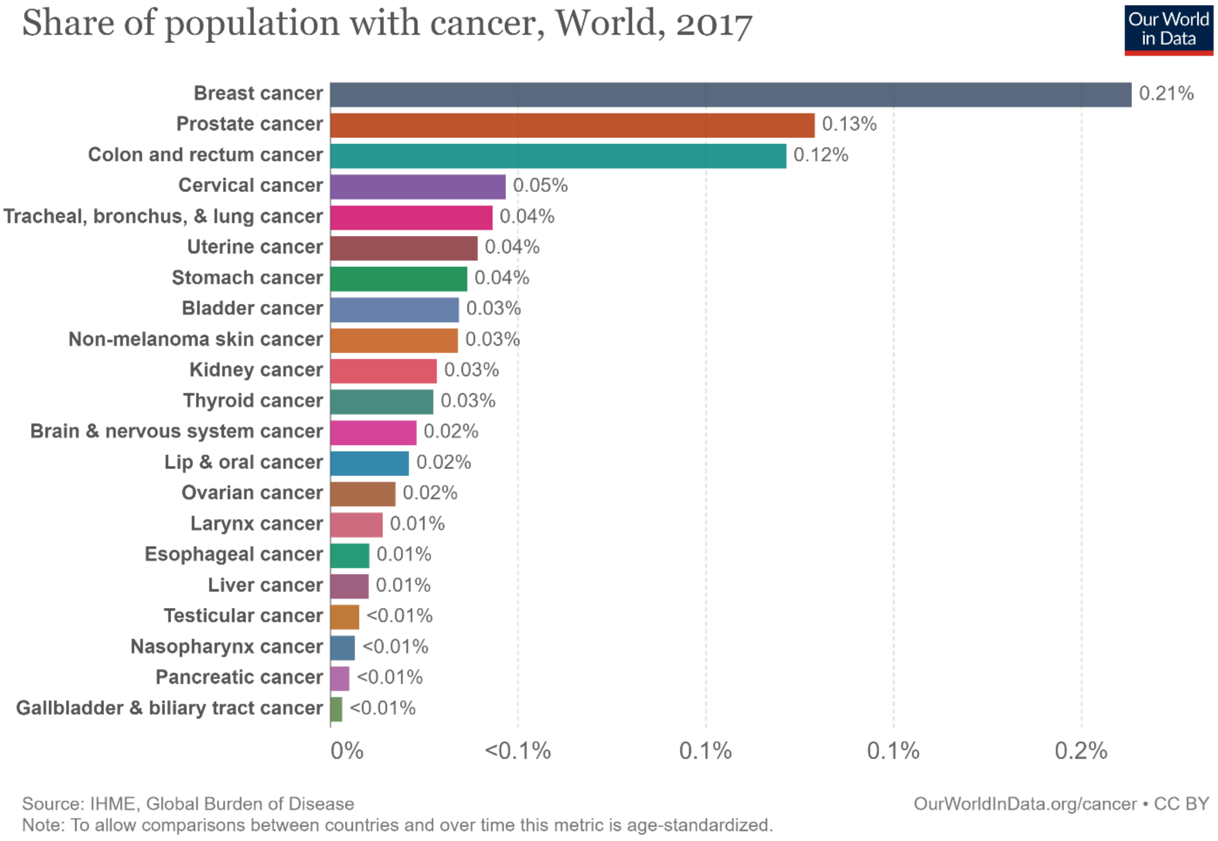


Fig 1. Share of population with cancer worldwide in 2017[2].

2. Problem Statement

We want to explore the possibility of creating a model to automate the diagnosis of breast cancer using deep learning.

It is known that in general there are two ways to train models:

1. Supervised learning, which requires data and associated labels. In this case, the association of data and labels is made by the expert.

2. Unsupervised learning, that only requires data.

Two types of computer vision models may be trained:

a) Classification model.

b) Segmentation model.

We exclude detection methods due to the difficulties of interpretation, which is an important issue in medical image processing.

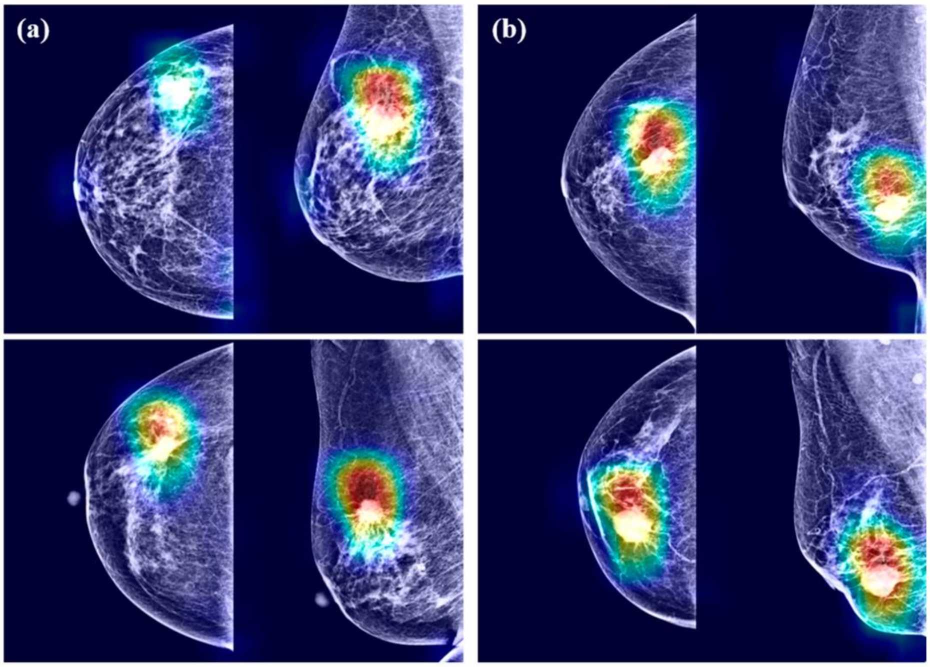


Fig 2. Gradient-weighed class activation mapping for mammograms having breast cancer by (a) DenseNet-169 and (b) EfficientNet-B5, image from [4]. We can see inaccurate highlight regions of interest, covering a large area around the actual allegedly cancerous region.

Training a classification model requires images associated with labels when one image is corresponding to one or more classes. Classification models in medical diagnostics have been studied previously by many researchers with good results. Unfortunately, the classification models, while achieving good accuracy, do not provide an explanation to the physician. To explain the inference, gradient-weighed class activation mapping [39] methods are used, which are not fully accurate since they visualize the area that triggers the decision from the last layer of the neural network. Such methods very often produce false positives, in terms of a classified area.

Therefore, we chose to build a segmentation model. Training the segmentation model requires more complex markup - the expert must associate each pixel of the image from a training set with one class. The training of segmentation models for medical diagnostics is also widely described in many publications (see the Related Work section).

Segmentation models in diagnostic tasks that are based on X-ray images do not require an explanation of the inference. Since classifying each pixel by belonging to one of the classes, we obtain a mask where the areas of interest are highlighted by the trained model.

Many neural architectures for image segmentation are currently known (U-Net [5], FCN [32], FPN [40], etc.).

U-Net architecture is the most common architecture used for medical imaging, which is well described in many publications. U-Net architecture belongs to the family of convolutional neural networks (CNNs). Sometimes, U-Net can be considered as a system of two neural networks - an encoder and a decoder [13, 33]. In this case, a conventional CNN (such as ResNet) can be used as an encoder - a contracting path to capture context. Whereas the decoder presents an expanding path that enables precise localization.

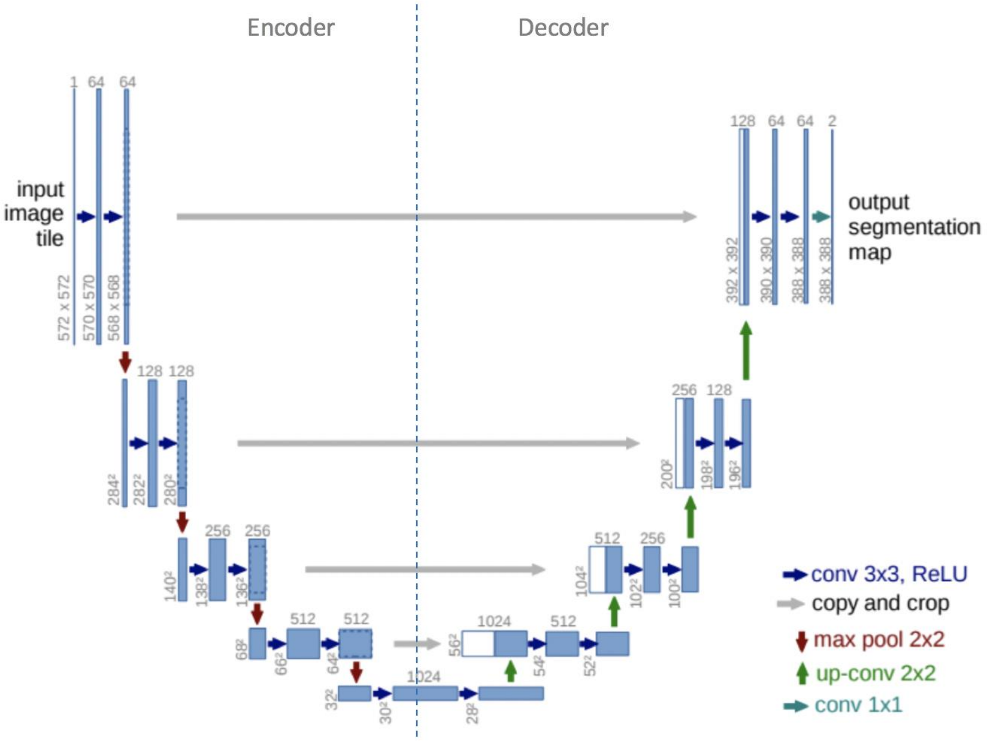


Fig.3. A scheme of the vanilla U-Net architecture [5]. The left part of an architecture is usually referred to as an encoder and the right part can be treated as a decoder.

A review of the available publications showed that the architecture of a neural network based on U-Net can be varied to achieve better segmentation results. In the case of segmentation of X-ray images, we need to reduce the number of false positively classified pixels, while preserving true positives to keep the precision as high as possible. The most promising in that case is improving the decoder since the encoder architectures for feature extraction are very well studied. The decoder is responsible for the upbuilding of the mask by the classification of pixels and utilizing extracted features. Also, keeping the encoder unchanged, we can use pretrained models of convolutional neural networks using the well-known advantages of transfer-learning.

This work aims to use the advantage of fusing high-level and low-level features during decoding, using the attention mechanisms (SE-block, channel attention)- it can help the decoder to focus on informative areas of an input image, and to ignore the non-informative regions. We will discover the possibility of improving the standard decoder architecture using attention mechanisms and compare the performance of the proposed and conventional decoders, but this modification can be done to any other decoder as well.

3. Related Work

Attempts to automate X-ray reading have been undertaken since the 1960s. However significant progress has only been reached recently with the help of Deep Learning methods. In this section, we will briefly review the related work that used U-Net and attention mechanisms for improving breast cancer segmentation.

U-Net is the most common convolutional network architecture utilizing the encoder-decoder principle for medical image segmentation. Based on its encoder-decoder structure, U-Net extracts low-level features that carry the geometry information of the image, while the high-level features carry the semantic information.

One of the first mentions of U-Net for medical image segmentation dates to 2015 - Ronnenberger et al [5] have shown the U-Net architecture applied to the segmentation of microscopic images.

Later, the results of studies were published with modifications of U-Net:

1. 3D U-Net, which enables 3D volumetric segmentation [6,7]
2. Inception U-net, which uses filters of multiple sizes on the same layer in the network [8,9]
3. Residual U-Net, encoder based on the ResNet architecture with corresponding decoder layers [10,11]
4. U-net++, which uses a dense network of skip connections as an intermediary grid between the contracting and expansive paths [12,13]
5. Dense U-net, which employ DenseNet blocks in place of regular layers [14-20]
6. Attention U-net, which uses the attention gate to focus on the specific object [21,22].

A very detailed overview of the use of U-Net for medical segmentation is presented in [23].

The use of the U-Net architecture for segmentation of mammograms also has a long history of publications. Considering the most recent, in [24] authors proposed Vanilla U-Net based model, which was used for precise segmentation of masses in breast X-ray images, comparing the performance of the proposed model with the fully convolutional network (FCN), SegNet, Dilated-Net, original U-Net, and Faster R-CNN models and the conventional region growing (RG) method. Authors of [24] show that the proposed Vanilla U-Net model outperforms the Faster R-CNN model significantly in terms of the inference time and the IOU metric.

The authors of another article [25] show significant superiority of their proposed neural network over all the other architectures: their model yields Dice coefficient of 99.20% and 99.56% and a weighted F1-score of 99.19% and 99.65% for the inference on DDSM and INbreast datasets, respectively. Unfortunately, there is no detailed description of the implementation, which is the biggest challenge for researchers in our area. After reading the paper we found some preprocessing tricks that lead to high segmentation score:

1. Grayscale images were converted to RGB.
2. Training and validation were performed on the images, included healthy (with blank mask).

Those tricks have led to the fact that we cannot compare the results from this article with others.

In the next article [26], the authors use U-Net which combines densely connected blocks with attention gates. Their proposed U-Net consists of an encoder which is a densely connected convolutional network and the decoder with integrated attention gates. Their model was trained and validated on DDSM dataset, reaching F1 coefficient of 82.24% and sensitivity of 77.89% with a specificity of 84.69%, thereby showing overall accuracy at 78.38%.

The combination of low- and high-level features in the U-Net architecture occurs using a simple concatenation, which is not the most effective method, as shown by [31].

Low-level features can be upsampled in several ways:

1. By the trainable deconvolution operation by reversing convolution as in [32] or unpooling as proposed in [33].
2. With fixed bilinear interpolation, the method that is proposed by [34].
3. By applying dense upsampling convolution [35].

The attention mechanism is mainly used in two ways as proposed in [36]:

1. Channel-wise attention.
2. Spatial attention.

Based on reviewed papers, we can conclude that the use of skip-connections in our architecture, which are designed for linking high- and low-level features and the use of attention mechanism can be a promising method.

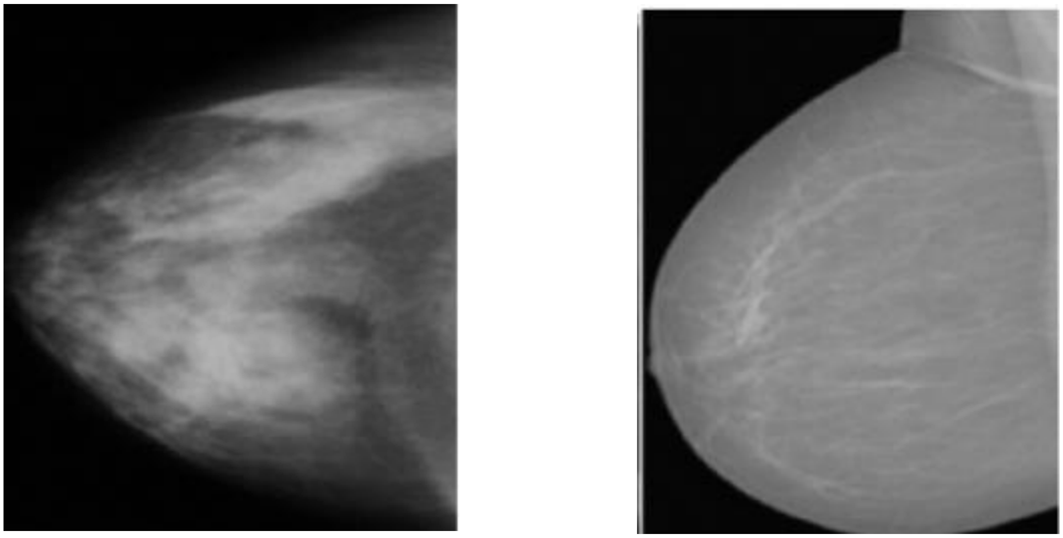
4. Methodology

4.1 Datasets

In all the reviewed articles, the authors demonstrate the results on two available datasets: CBIS-DDSM and IN-breast. Let us briefly describe each of them:

1. CBIS-DDSM [27,28] consists of total 858 images with mass lesions. See Fig. 4(a) for example.

2. IN-breast [29] total of 107 images with the mass lesion. See Fig 4(b) for example.



(a) (b)

Fig 4. Example of CBIS-DDSM dataset (a), and IN-breast dataset (b). The visible difference in image quality is due to the different methods of imaging – CBIS-DDSM is scanned films, and IN-breast is a full-field digital mammography image.

Another dataset named CSAW [30] has recently become available, which, due to its novelty, has not yet had time to be widely used in publications. A distinctive feature of this dataset is the multiclass segmentation labels not only for mass regions but for some physiological parts such as blood vessels, lymph nodes, skin, and muscles.

Most researchers are using private datasets for training and evaluating their classification or segmentation algorithms. Due to the lack of public mammography datasets, more than 60% of the recently published research is hardly reproduced.

4.2 Proposed neural network architecture

Proposed network architecture, based on U-Net with basic ResNet encoder is shown in Fig. 5

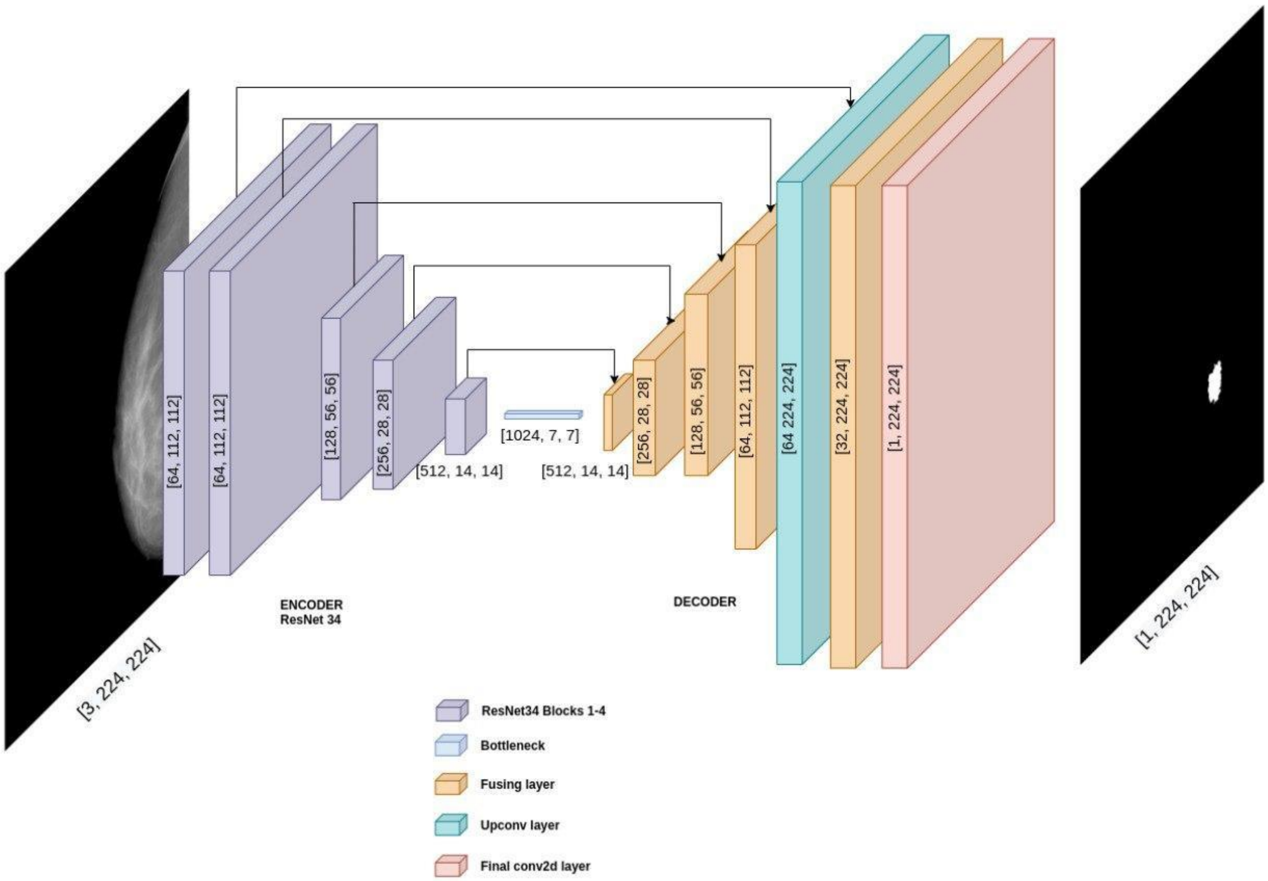


Fig 5. Proposed network architecture with ResNet34 backbone. Upsampling layer is described next.

The main idea of our proposed network is to construct a custom decoder, which, besides taking into account the features from the previous decoder levels, also adds information from corresponding encoder levels, and processes it through the SE-block. The proposed Fusion Layer is shown in Fig. 6.

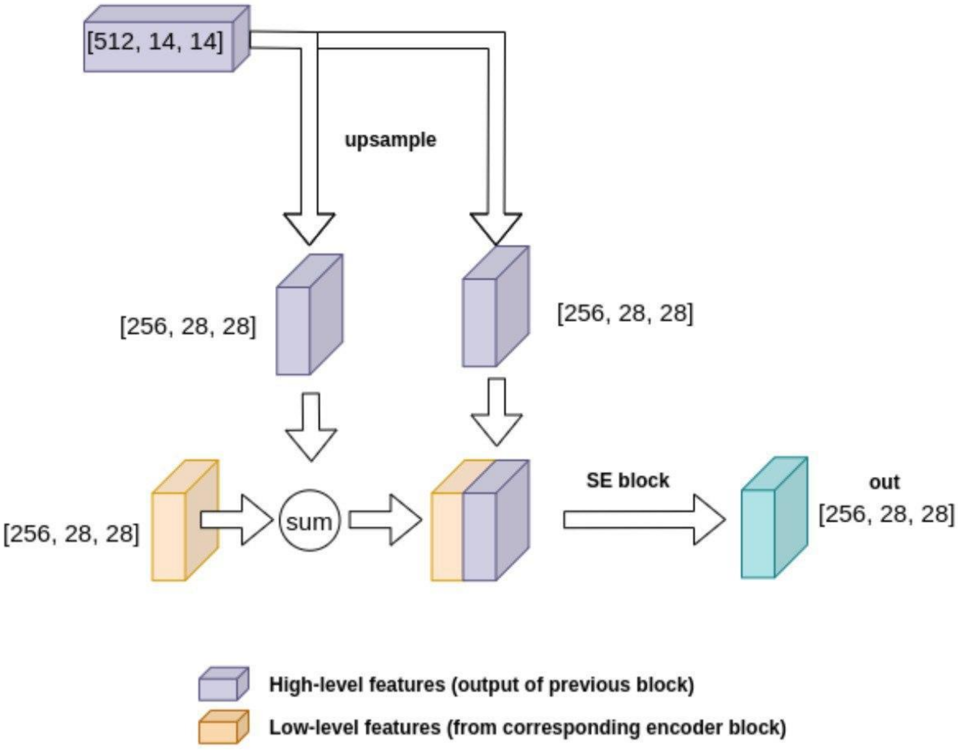

[512, 14, 14]
upsample
[256, 28, 28]
[256, 28, 28]
[256, 28, 28]
sum
SE block
out
[256, 28, 28]
High-level features (output of previous block)
Low-level features (from corresponding encoder block)

Fig. 6. Proposed Fusion Layer, linking high- and low-level features, uses SE-block to merge information in an efficient way to extract only significant parts.

4.3 Implementation details

The proposed architecture was implemented using the Pytorch framework. We use pretrained encoders from Pytorch Image Models (widely known as TIMM) and decoders from Segmentation Models Pytorch (SMP) as a baseline in this work.

Training and testing for all the experiments were performed on the IBM® Power System™ AC922, with 4 Nvidia V100 GPUs with 16 GB VRAM, 2 x POWER9 Processor x 16-core 2.7 GHz, and 570 GB RAM.

To prove the good quality of our proposed method, we used two datasets from different areas:

1. CBIS-DDSM with 256x256 image size for medical usecase.

2. Pascal VOC 2012 Segmentation challenge with 384x384 image size for the general use-case.

With CBIS-DDSM we used the Resnet-34 encoder, and with Pascal VOC – ResNet50_V2 Big Transfer (BiT).

We used AdamW optimizer with amsgrad option on. For the CBIS-DDSM training, we use complex loss function as the sum of Jaccard and Focal Tversky Losses and learning rate of 5E-4, and for the Pascal VOC sum of Cross-Entropy and Focal Tversky Losses and learning rate of 5E-5 were used.

For greater convenience, all implementation details are summarized in Table 1.

Table 1. Implementation details for breast cancer and Pascal VOC segmentation tasks.

Parameter	**CBIS-DDSM**	**Pascal VOC 2012**
Encoder	Resnet-34	ResNet50_V2 Big Transfer (BiT)
Loss Function	Jaccard + Focal Tversky Loss	CE+Focal Tversky Loss
Image Size	256x256	384x384
Learning Rate	5E-4	5E-5
Optimizer	AdamW with amsgrad	AdamW with amsgrad

Scheduler	epochs= [0,40,70,100,120],
	lr_list = [5e-4,1e-4,5e-5,1e-5,1e-6]
		epochs= [0,80],
		lr_list = [0.00005,0.00001])

Num epochs	150	150
Train images	690	1464
Val images	168	1449

5. Experiments

Firstly, we compared our method (A) to the baseline built using Segmentation Models Pytorch library (B and C). Results of the experiment provided in Table 2 (encoder: ResNet34):

Table 2. Experiment #1. Comparison between the baseline and the proposed methods on breast the radiography segmentation task (CBIS-DDSM dataset).

Architecture	IoU (%)	DICE (%)	#Params	FLOPS	Inference time (s)
(A) proposed decoder	67.42	79.5	89.62 M	32.22	0.0149
(B) default SMP decoder	62.31	73.57	24.44 M	3.945	0.00249
(C) Unet++ SMP decoder	61.48	72.11	26.28 M	9.26	0.005

We can clearly see that the proposed method outperforms the baseline methods.

Next, we would like to discover transfer-learning capabilities. For this, we compared the proposed architecture with pretrained and unpretrained encoder. Results of the experiment provided in Table 3.

Table 3. Experiment #2. Transfer-learning advantage on breast radiography segmentation task (CBIS-DDSM dataset)

Architecture	IoU (%)	DICE (%)
(A) ResNet34, encoder pretrained	67.42	79.5
(B) ResNet34, encoder trained from scratch	57.02	67.06

We can see that transfer-learning works for the proposed network on breast cancer radiography images.

Now we want to compare our method on a commonly used photography benchmark segmentation dataset. For such benchmarking, we have selected the Pascal VOC 2012 segmentation dataset, which is the multi-class segmentation task, consisting of 20 classes of real-life images. Results of the experiment provided in Table 4:

Table 4. Experiment #2. Comparison of baseline and proposed method on real-life segmentation task (Pascal VOC 2012 segmentation dataset)

Architecture	mIoU (%)	#Params	FLOPS	Inference time (s)
(A) UNet with proposed decoder	81.268	202.44 M	115.82	0.152
(B) UNet with default SMP decoder	62.111	32.51 M	6.04	0.0845

6. Results discussion

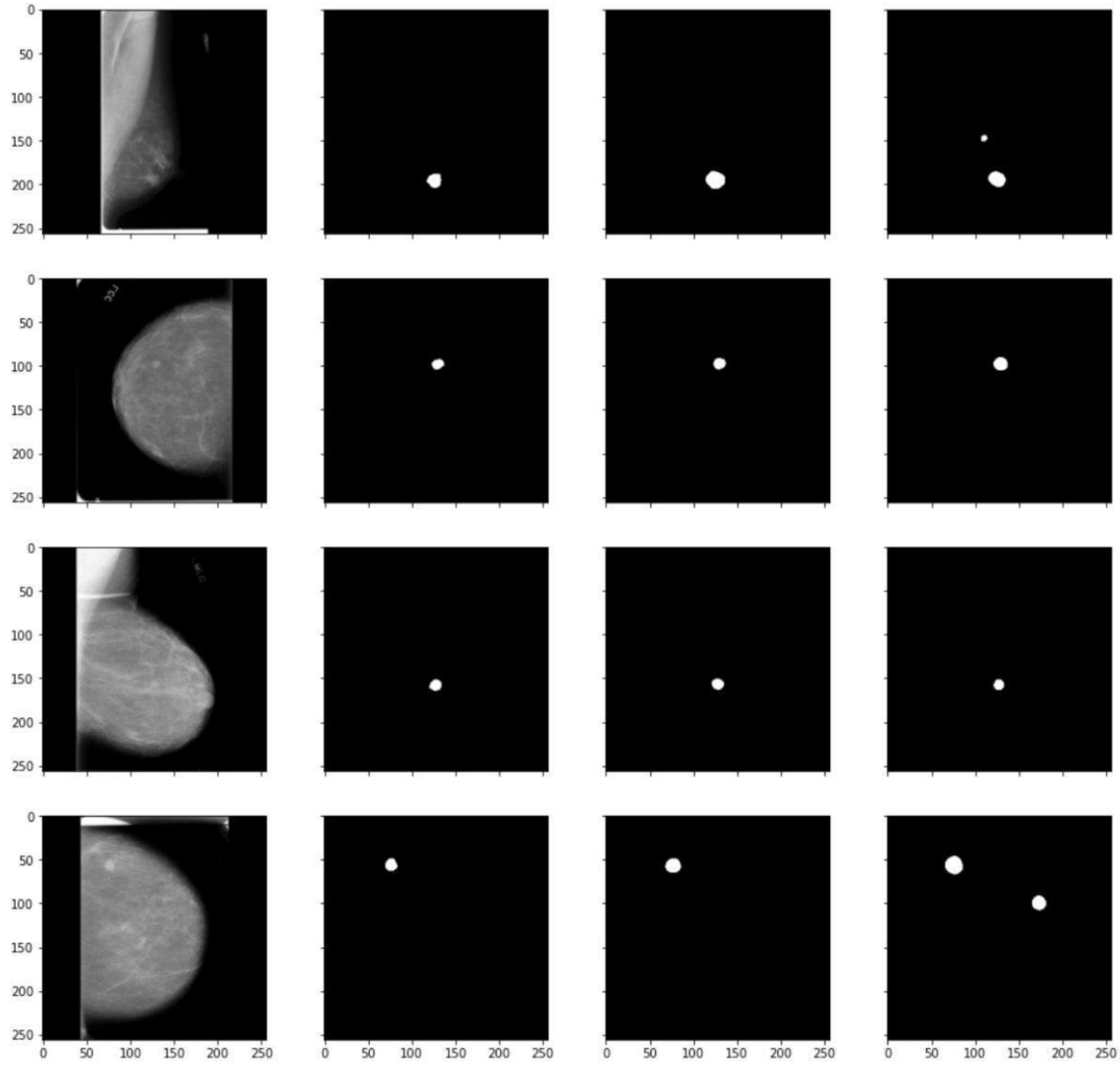
6.1 Breast radiography segmentation task (CBIS-DDSM dataset).

The experiments show the effectiveness of the proposed method in terms of segmentation accuracy – we achieved high DICE and IoU metrics on breast radiography data.

Comparison between the proposed method and baseline shows that our proposed method outperforms baseline by ~5% for both IoU and DICE.

The difference between the proposed method and the benchmark is shown in Fig 7-9.

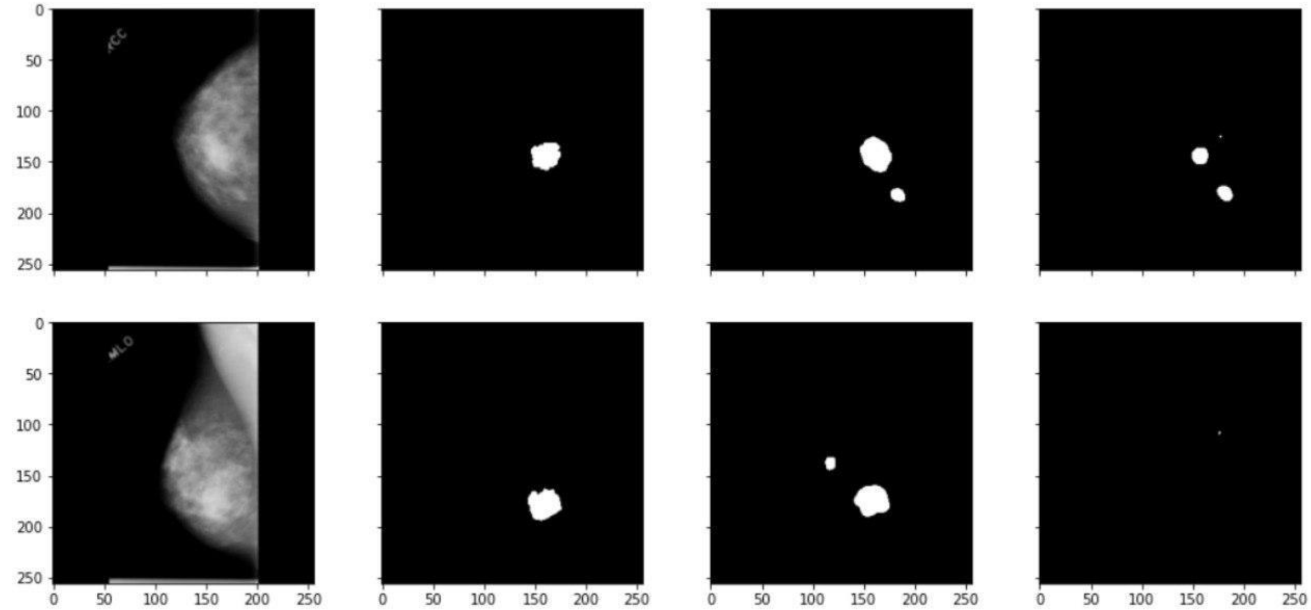
In Fig. 7 we can see that the proposed method finds all the nodules corresponding to the ground truth, however, baseline UNet produces some false positives.



(a) (b) (c) (d)

Fig. 7. Comparison between our proposed method and baseline UNet on breast radiography segmentation task some ordinary cases are shown. Both methods use the pretrained ResNet34 encoder: (a) – original image, (b) – ground truth, (c) – mask produced by proposed method, (d) – mask produced by baseline.

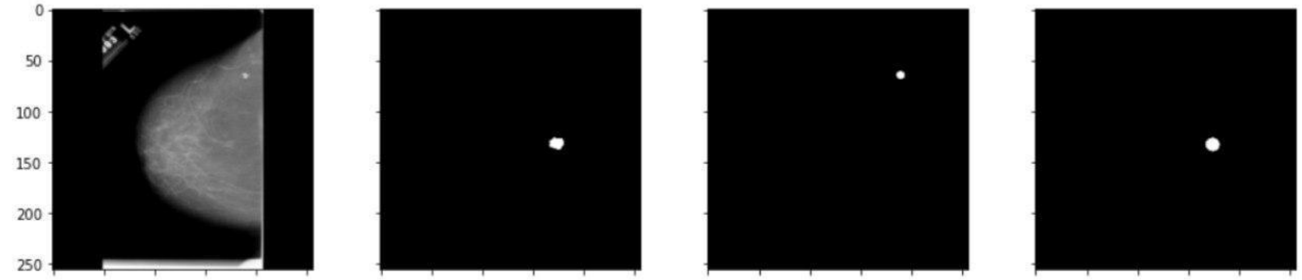
Results on Fig. 8 show that in complicated cases both methods made mistakes, but we can see that the proposed method detects nodules, even with some false-positive detections, while the baseline method fails in size in the first case and fails to detect nodules in the second case.



(a) (b) (c) (d)

Fig. 8. Comparison between our proposed method and baseline UNet on breast radiography segmentation task (medium difficulty cases shown): (a) – original image, (b) – ground truth, (c) – mask produced by proposed method, (d) – mask produced by baseline.

In Fig. 9 we can see the interesting case when the baseline method produces an accurate segmentation mask while the proposed method fails. In a validation set of 168 images, only one such example was found.



(a) (b) (c) (d)

Fig. 9. The case when baseline method produces more accurate segmentation mask, only one in validation sample: (a) – original image, (b) – ground truth, (c) – mask produced by proposed method, (d) – mask produced by baseline.

6.2 Real-life segmentation task (Pascal VOC 2012 segmentation dataset)

Since it’s hard to compare our results with others published in some papers, we want to compare our method on a benchmark segmentation dataset. In the role of such a dataset, we chose Pascal VOC 2012 segmentation challenge, which is the multi-class segmentation task, consists of 20 classes.

Using the benchmark Pascal VOC 2012 segmentation dataset, we ensured the comparability of mIoU metric, even on validation sample, leaderboard is available: https://paperswithcode.com/sota/semantic-segmentation-on-pascal-voc-2012-val

We can see that the proposed method works well, and achieves relatively high-quality mIOU =81.268 %, using only ResNet50-v2 pretrained encoder and image size 384x384 pixels.

For example, nearest in terms of mIOU on leaderboard method achieved 81% mIOU, used COCO and Semantic Boundaries Dataset as pretraining, and image size of 640x640 to 512x512, and two-stage training procedure [37], while we use only Pascal VOC dataset and ImageNet pretrained encoder.

Interestingly, our method illustrates full use of the capabilities of a well-trained ResNet encoder. We achieved high mIOU =81.268% with our proposed decoder, default SMP decoder achieves only mIOU =62.11% when using the same pretrained ResNetV2 Big Transfer (BiT) encoder [38].

In Fig. 10 we can see that the proposed method finds all the classes which somehow correspond to the ground truth, however, baseline UNet fails to assign bottle class in the first case, but segments people very nicely in first and second cases (this most likely happens is due to the class imbalance).

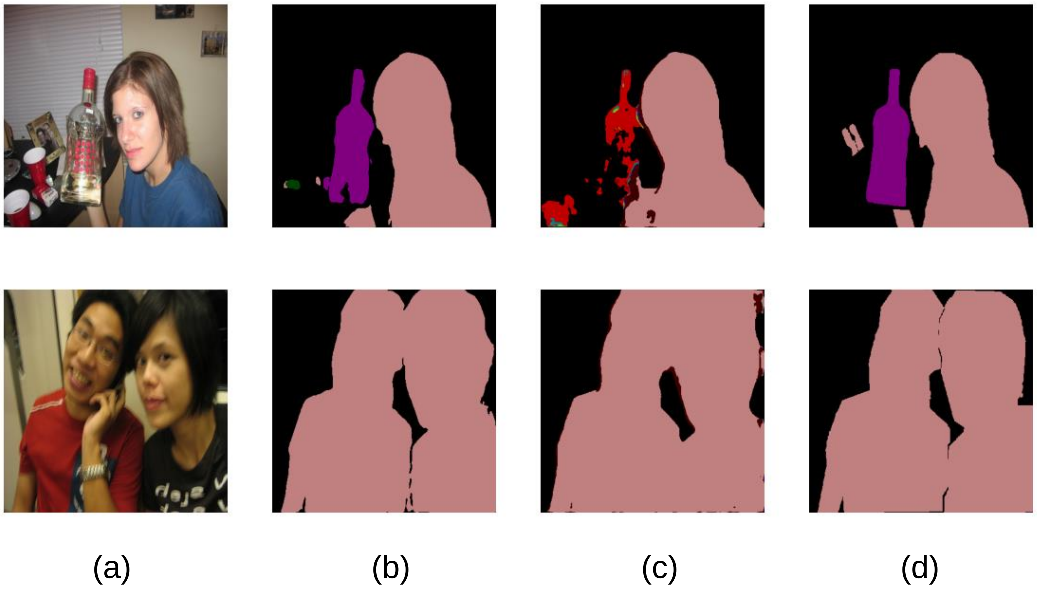


Fig. 10. Segmentation examples from Pascal VOC 2012 validation sample: (a) original image, (b) predictions of proposed UNet, (c) predictions of vanilla SMP UNet, (d) ground-truth.

In Fig. 10 we can see that the proposed method finds all the classes that somehow correspond to the ground truth, however baseline UNet fails assigning correct classes, while the shape was segmented correctly.

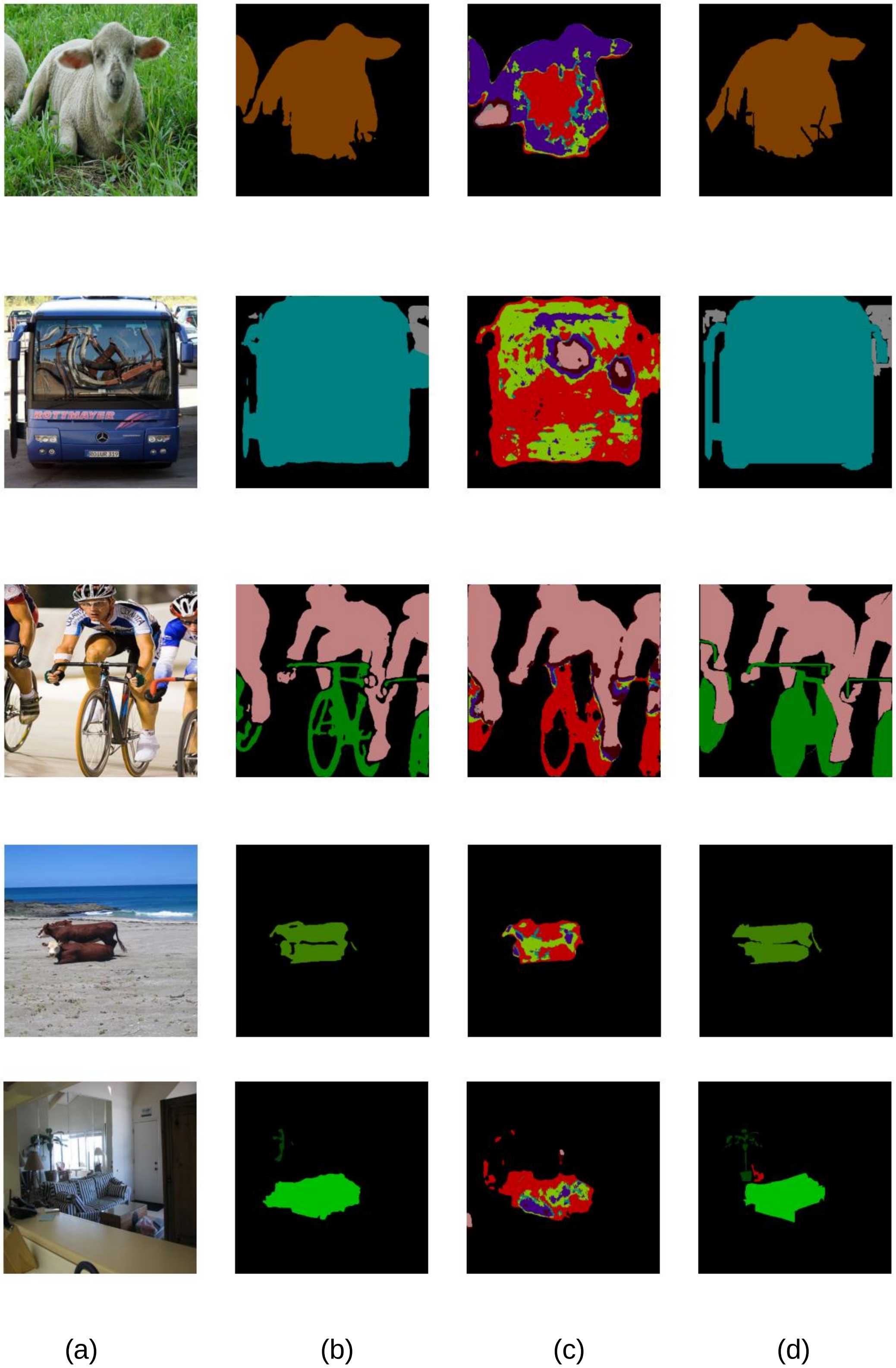


Fig. 11. Segmentation examples from Pascal VOC 2012 validation sample: (a) original image, (b) predict of proposed UNet, (c) predict of vanilla SMP UNet, (d) ground-truth.

7. Summary

Our study presents a promising segmentation method based on high- and low-level feature fusing and attention mechanism, which is suitable for medical image segmentation, particularly breast radiography segmentation using full-frame x-rays. We achieved IOU = 67.42 and DICE = 79.5 on CBIS-DDSM dataset, while not cropping images around suspicious regions, this makes our method suitable for the primary reading of medical images without radiologist input.

Proposed UNet architecture can be effective in real-life segmentation tasks too since we achieved mIOU = 81.268 % on a validation set of Pascal VOC 2012 segmentation dataset. Interesting that we have achieved these mIOU values using only ImageNet pretrained encoder without pretraining on additional segmentation data unlike nearest competitors on the public leaderboard. Nevertheless, we see ways of further improvement of our approach, considering further computational complexity optimization and usage of other encoders. For instance, it is interesting to see how the decoder will work with RefineNet and LightRefineNet decoder.

References

1. Sung, H., Ferlay, J., Siegel, R.L., Laversanne, M., Soerjomataram, I., Jemal, A. and Bray, F., 2021. Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: a cancer journal for clinicians*, *71*(3), pp.209-249.
2. Max Roser and Hannah Ritchie (2015) - "Cancer". Published online at OurWorldInData.org. Retrieved from: 'https://ourworldindata.org/cancer' [Online Resource]
3. Rajpurkar, P., Irvin, J., Zhu, K., Yang, B., Mehta, H., Duan, T., Ding, D., Bagul, A., Langlotz, C., Shpanskaya, K. and Lungren, M.P., 2017. Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning. *arXiv preprint arXiv:1711.05225*.
4. Suh, Y.J., Jung, J. and Cho, B.J., 2020. Automated breast cancer detection in digital mammograms of various densities via deep learning. *Journal of personalized medicine*, *10*(4), p.211.
5. Ronneberger, O., Fischer, P. and Brox, T., 2015, October. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention* (pp. 234-241). Springer, Cham.
6. Zhao, C., Han, J., Jia, Y. and Gou, F., 2018, October. Lung nodule detection via 3D U-Net and contextual convolutional neural network. In *2018 International conference on networking and network applications (NaNA)* (pp. 356-361). IEEE.
7. Zeng, G., Yang, X., Li, J., Yu, L., Heng, P.A. and Zheng, G., 2017, September. 3D U-net with multi-level deep supervision: fully automatic segmentation of proximal femur in 3D MR images. In *International workshop on machine learning in medical imaging* (pp. 274-282). Springer, Cham.
8. Chen, W., Liu, B., Peng, S., Sun, J. and Qiao, X., 2018, September. S3D-UNet: separable 3D U-Net for brain tumor segmentation. In *International MICCAI Brainlesion Workshop* (pp. 358-368). Springer, Cham.
9. Cheng, H., Zhu, Y. and Pan, H., 2019, May. Modified U-Net block network for lung nodule detection. In *2019 IEEE 8th Joint International Information Technology and Artificial Intelligence Conference (ITAIC)* (pp. 599-605). IEEE.
10. Mostafiz, T., Jarin, I., Fattah, S.A. and Shahnaz, C., 2018, December. Retinal blood vessel segmentation using residual block incorporated U-Net architecture and fuzzy inference system. In *2018 IEEE International WIE Conference on Electrical and Computer Engineering (WIECON-ECE)* (pp. 106-109). IEEE.

11. Li, H., Chen, D., Nailon, W.H., Davies, M.E. and Laurenson, D., 2018. Improved breast mass segmentation in mammograms with conditional residual u-net. In *Image Analysis for Moving Organ, Breast, and Thoracic Images* (pp. 81-89). Springer, Cham.

12. Wu, S., Wang, Z., Liu, C., Zhu, C., Wu, S. and Xiao, K., 2019, July. Automatical segmentation of pelvic organs after hysterectomy by using dilated convolution U-Net++. In *2019 IEEE 19th International Conference on Software Quality, Reliability and Security Companion (QRS-C)* (pp. 362-367). IEEE.

13. Zhou, Z., Siddiquee, M.M.R., Tajbakhsh, N. and Liang, J., 2018. Unet++: A nested u-net architecture for medical image segmentation. In *Deep learning in medical image analysis and multimodal learning for clinical decision support* (pp. 3-11). Springer, Cham.

14. Kolarik, M., Burget, R., Uher, V. and Povoda, L., 2019, July. Superresolution of MRI brain images using unbalanced 3D Dense-U-Net network. In *2019 42nd International Conference on Telecommunications and Signal Processing (TSP)* (pp. 643-646). IEEE.

15. Zhang, Z., Wu, C., Coleman, S. and Kerr, D., 2020. DENSE-INception U-net for medical image segmentation. *Computer methods and programs in biomedicine*, *192*, p.105395.

16. Meng, C., Sun, K., Guan, S., Wang, Q., Zong, R. and Liu, L., 2020. Multiscale dense convolutional neural network for DSA cerebrovascular segmentation. *Neurocomputing*, *373*, pp.123-134.

17. Dolz, J., Ayed, I.B. and Desrosiers, C., 2018, September. Dense multi-path U-Net for ischemic stroke lesion segmentation in multiple image modalities. In *International MICCAI Brainlesion Workshop* (pp. 271-282). Springer, Cham.

18. Azad, R., Asadi-Aghbolaghi, M., Fathy, M. and Escalera, S., 2019. Bi-directional ConvLSTM U-Net with densley connected convolutions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops* (pp. 0-0).

19. Li, X., Chen, H., Qi, X., Dou, Q., Fu, C.W. and Heng, P.A., 2018. H-DenseUNet: hybrid densely connected UNet for liver and tumor segmentation from CT volumes. *IEEE transactions on medical imaging*, *37*(12), pp.2663-2674.

20. Wang, Z.H., Liu, Z., Song, Y.Q. and Zhu, Y., 2019, September. Densely connected deep u-net for abdominal multi-organ segmentation. In *2019 IEEE International Conference on Image Processing (ICIP)* (pp. 1415-1419). IEEE.

21. Schlemper, J., Oktay, O., Schaap, M., Heinrich, M., Kainz, B., Glocker, B. and Rueckert, D., 2019. Attention gated networks: Learning to leverage salient regions in medical images. *Medical image analysis*, *53*, pp.197-207.

22. Zhang, Z., Fu, H., Dai, H., Shen, J., Pang, Y. and Shao, L., 2019, October. Et-net: A generic edge-attention guidance network for medical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention* (pp. 442-450). Springer, Cham.

23. Siddique, N., Paheding, S., Elkin, C.P. and Devabhaktuni, V., 2021. U-net and its variants for medical image segmentation: A review of theory and applications. *IEEE Access*.

24. Abdelhafiz, D., Bi, J., Ammar, R., Yang, C. and Nabavi, S., 2020. Convolutional neural network for automated mass segmentation in mammography. *BMC bioinformatics*, *21*(1), pp.1-19.

25. Soulami, K.B., Kaabouch, N., Saidi, M.N. and Tamtaoui, A., 2021. Breast cancer: One-stage automated detection, segmentation, and classification of digital mammograms using UNet model based-semantic segmentation. *Biomedical Signal Processing and Control*, *66*, p.102481.

26. Li, S., Dong, M., Du, G. and Mu, X., 2019. Attention dense-u-net for automatic breast mass segmentation in digital mammogram. *IEEE Access*, *7*, pp.59037-59047.

27. Heath, M., Bowyer, K., Kopans, D., Kegelmeyer, P., Moore, R., Chang, K. and Munishkumaran, S., 1998. Current status of the digital database for screening mammography. In *Digital mammography* (pp. 457-460). Springer, Dordrecht.

28. Lee, R.S., Gimenez, F., Hoogi, A., Miyake, K.K., Gorovoy, M. and Rubin, D.L., 2017. A curated mammography data set for use in computer-aided detection and diagnosis research. *Scientific data*, *4*(1), pp.1-9.

29. Moreira, I.C., Amaral, I., Domingues, I., Cardoso, A., Cardoso, M.J. and Cardoso, J.S., 2012. Inbreast: toward a full-field digital mammographic database. *Academic radiology*, *19*(2), pp.236-248.

30. Dembrower, K., Lindholm, P. and Strand, F., 2020. A multi-million mammography image dataset and population-based screening cohort for the training and evaluation of deep neural networks—the cohort of screen-aged women (CSAW). *Journal of digital imaging*, *33*(2), pp.408-413.

31. Zhang, Z., Zhang, X., Peng, C., Xue, X. and Sun, J., 2018. Exfuse: Enhancing feature fusion for semantic segmentation. In *Proceedings of the European Conference on Computer Vision (ECCV)* (pp. 269-284).

32. Long, J., Shelhamer, E. and Darrell, T., 2015. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 3431-3440).

33. Badrinarayanan, V., Kendall, A. and Cipolla, R., 2017. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE transactions on pattern analysis and machine intelligence*, *39*(12), pp.2481-2495.

34. Chen, L.C., Papandreou, G., Kokkinos, I., Murphy, K. and Yuille, A.L., 2017. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE transactions on pattern analysis and machine intelligence*, *40*(4), pp.834-848.

35. Wang, P., Chen, P., Yuan, Y., Liu, D., Huang, Z., Hou, X. and Cottrell, G., 2018, March. Understanding convolution for semantic segmentation. In *2018 IEEE winter conference on applications of computer vision (WACV)* (pp. 1451-1460). IEEE.

36. Roy, A.G., Navab, N. and Wachinger, C., 2018, September. Concurrent spatial and channel 'squeeze & excitation'in fully convolutional networks. In *International conference on medical image computing and computer-assisted intervention* (pp. 421-429). Springer, Cham.

37. Peng, C., Zhang, X., Yu, G., Luo, G. and Sun, J., 2017. Large kernel matters--improve semantic segmentation by global convolutional network. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 4353-4361).

38. Beyer, L., Zhai, X., Royer, A., Markeeva, L., Anil, R. and Kolesnikov, A., 2021. Knowledge distillation: A good teacher is patient and consistent. *arXiv preprint arXiv:2106.05237*.

39. Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D. and Batra, D., 2017. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision* (pp. 618-626).

40. Jia, Y., Tan, J., Xing, Y., Hong, P. and Zhang, L., 2020, September. Densely Connected Feature Pyramid Network for Image Segmentation. In *2020 8th International Conference on Digital Home (ICDH)* (pp. 93-98). IEEE.