Ingénierie des Systèmes d'Information
Vol. 29, No. 4, August, 2024, pp. 1603-1609
Journal homepage: http://iieta.org/journals/isi

Innovative: A Novel Deep Learning-Based Semantic Segmentation Architecture for Medical Applications

Elmehdi Aniq[1,2*], Mohamed Chakraoui[1], Naoual Mouhni[3]

[1] LS2ME, Polydisciplinary Faculty of Khouribga, Sultan Moulay Slimane University, Khouribga 25000, Morocco
[2] LAMIGEP, EMSI Marrakech, Marrakech 40000, Morocco
[3] GL-ISI, Department of Informatics, Faculty of Sciences and Technics, UMI-Meknes, Errachidia 52202, Morocco

Corresponding Author Email: elmehdi.aniq@gmail.com

https://doi.org/10.18280/isi.290433

Received: 19 December 2023
Revised: 9 July 2024
Accepted: 2 August 2024
Available online: 21 August 2024

Keywords:
medical applications, deep learning, semantic segmentation, encoder-decoder, atrous convolutions

ABSTRACT

Within the field of computer vision and artificial intelligence, the analysis of two-dimensional image data stands as a pivotal domain, specifically in the context of semantic segmentation. This intricate process involves the precise categorization of pixels within a two-dimensional space, thereby enabling nuanced classification at a granular level. In this research endeavor, we present a novel network architecture, denoted as "a-Net," strategically crafted to achieve a delicate balance between computational expeditiousness, operational efficiency, adaptability, and precision for the overarching objective of semantic segmentation in two-dimensional imagery. The a-Net architecture, grounded in the principles of auto-encoding, tactically addresses data loss concerns inherent in segmentation processes. Engineered to adeptly outline objects within two-dimensional spaces, this architecture yields meticulous masks for individual objects, ensuring the generation of high-fidelity segmentation outcomes. The design philosophy of a-Net underscores not only its computational efficacy but also its straightforward implementability and training, thus imparting versatility across a diverse array of applications. Its efficacy spans the resolution of varied challenges within the domain of two-dimensional semantic segmentation, with particular relevance in medical imaging scenarios encompassing objects of both microscopic and macroscopic scales. Our investigative methodology establishes the superior performance of the a-Net architecture relative to alternative two-dimensional semantic segmentation frameworks. This superiority is underscored by commendable outcomes observed across diverse challenges, affirming the a-Net's status as a robust and versatile solution within the evolving landscape of two-dimensional semantic segmentation. This research significantly contributes to advancing the state of the art in the realm of image segmentation, offering a sophisticated and efficient solution that attains optimal precision while preserving computational efficiency.

1. INTRODUCTION

This new field of deep learning has taken off within machine learning, enjoying tremendous acclamation for the capacity to deal with very complicated data and run through complex tasks. It is applied in the medical domain, more so in handling some of the pivotal challenges associated with semantic segmentation, which testifies to this transformative capability it is having. In the medical context, semantic segmentation means the careful classification of every pixel in a medical image into predefined categories, such as organs, tissues, or anomalies. It is a process intended to enhance diagnostic accuracy, aid thoughtful treatment planning, and further medical research.

Deep learning remained an effective solution for semantic segmentation due to its inherent ability for capturing complex features and generating generalized insights from large datasets. Deep Convolutional Neural Networks [1] have been proved to be very powerful on a range of visual recognition tasks. However, more often than not, performance is hamstrung by the requirement of large datasets and a huge number of parameters during training. The unique challenge in semantic segmentation is classifying every pixel independently, which departs from this conventional approach of giving a single class to an image as seen in most image classification paradigms.

Semantic segmentation spans across dimensions: 2D images, videos, and 3D objects. Each one is a very challenging and crucial aspect in computer vision. This means that, in 2D images, all pixels should be classified by understanding the semantic relations between neighboring pixels to delineate an object. More traditionalick systems for semantic segmentation have been manually designed with features fed into a 'flat' classifier like Boosting, Random Forests, or Support Vector Machines. In supervised approaches, pixels were classified using their features, while in clustering algorithms, segmentation of objects was based on defining the number of clusters and then allowing the algorithm to segregate pixels

based on inter-cluster distances. However, features of this kind were relatively of limited expressive power.

Even though deep neural networks have overshadowed all other traditional algorithms of machine learning, the Convolution Neural Networks are still accompanied by certain challenges like reduction in features due to Convolution operations. To deal with it, the dilated convolutional technique creates holes between subsequent elements of the kernel while training and increases its covered area, thereby enhancing the extraction of relevant information. Another challenge that remains is how to deal with objects at multiple scales, surmounted by employing the pyramidal space technique.

The research proposes an a-Net semantic segmentation architecture that combines pyramidal spatial dilated convolution with an encoder-decoder approach. This fusion seeks to retain maximum information from an image, and the ensuing sections will deliberate on separate elements that go into this architecture while testing its influence on variability across different datasets.

2. RELATED WORKS

Large success in deep learning for semantic segmentation was initiated by Long et al., who developed fully convolutional neural networks for this task. This enabled hierarchical feature extraction and learning through CNNs, extending those large classification models like AlexNet [2], VGG-16 [3], GoogLeNet [4], and ResNet [5] into FCNs [6] for generating spatial maps rather than simple classification scores.

Recent research has mostly been centered on modifications and improvements of FCNs in semantic segmentation with good results.

2.1 Encoder-decoder

Another influential methodology apart from the FCN paradigm is that of encoder-decoder. Essentially, it consists of two phases: an encoder generating a feature map, and a decoder reconstructing an image of the same dimensions as the input image from the feature map. Contrary to FCNs, the decoder phase comprises upsampling and convolution, and ends with a softmax operation to assign each pixel a specific class. This approach spawned various different architectures, of which SegNet [7] and U-Net [8] are ones that have enriched the segmentation methodologies landscape. Each of their strengths is different within the computer vision and image analysis spectrum.

2.2 Atrous convolutions

The technique of Dilated Convolution has been outstanding in efficiently extracting features across arbitrary resolutions. Chen et al. [9] introduced the DeepLab architecture that uses dilated convolution with different dilation rates to achieve a larger receptive field without extra computational cost or undersampling of feature maps. Improvements along this line, DeepLab version 3 [10], have gone much beyond the performance without the post-processing Conditional Random Field step of previous versions. This means that Dilated Convolutional Layers can actually improve the extracted features effectively for semantic segmentation.

2.3 Fusion of features

Feature fusion is one of the prevalent techniques for semantic segmentation. It demonstrated very remarkable performance on a large variety of tasks. Chen et al. [11] explained this methodology that refers simply to the concatenation of outputs stemming from several layers within a network. Similarly, Pinheiro et al. [12] proposed a Feature Fusion-based network. This introduced a progressive refinement module that refines smoothly the functionalities from the previous layers into the following ones. This technique further underlines the role and power of feature fusion in strengthening semantic segmentation networks toward more robust adaptation with respect to context-aware segmentation results.

3. OUR METHOD

One of the primary limitations in advanced architectures like the U-Net and DeepLab was their inability to evaluate and extract advanced features. To answer this challenge, much attention had to be put forward on improving this critical phase to enhance model performance and accuracy.

As shown in this section, our architecture, a-Net (Figure 1) [13, 14], is constructed as illustrated, with Figure below, showing a-Net incessantly putting emphasis on the roles of atrous convolution and the encoder for efficient feature extraction.

3.1 Stride and Atrous convolution (Features extraction)

Speaking with regard to DCNNs, traditional convolution starts a window for convolution in the top left of the tensor and goes scan line by scan line to the right and downwards. If a stride is 1, then every element is visited one at a time. But for the sake of efficiency or subsampling, strides of two or three elements may be used, skipping intermediate locations. This method combined with pooling will lead to decreasing spatial resolution and increase the number of trainable parameters.

Dilated convolution, also named atrous spatial pyramid sampling, is a strategic solution keeping up—determining spaces between kernel values and explicit control over the density of feature response computation in fully convolutional networks. For instance, a 3×3 kernel with a dilation rate of 3 provides the same view as a 7×7 kernel, thus portraying how the filter's vision will change adaptively with the modification of the dilation rate. In our particular case, with 256×256 image dimensions, we feel the difference between the spatial resolution of the input and output. With a stride of 3 in the first convolution layer, the dimension turns out to be 86×86. Then, applying the same Convolution six times with the same stride is constant and chances of feature decimation are not there. On the other side, atrous convolution keeps the same dimension of the output where r, the dilation rate, implies the sampling step at the input image shown in Figure 2.

It is accomplished through filtering of the image and putting (r-1) zeros between two successive values of filters along each dimension of space.

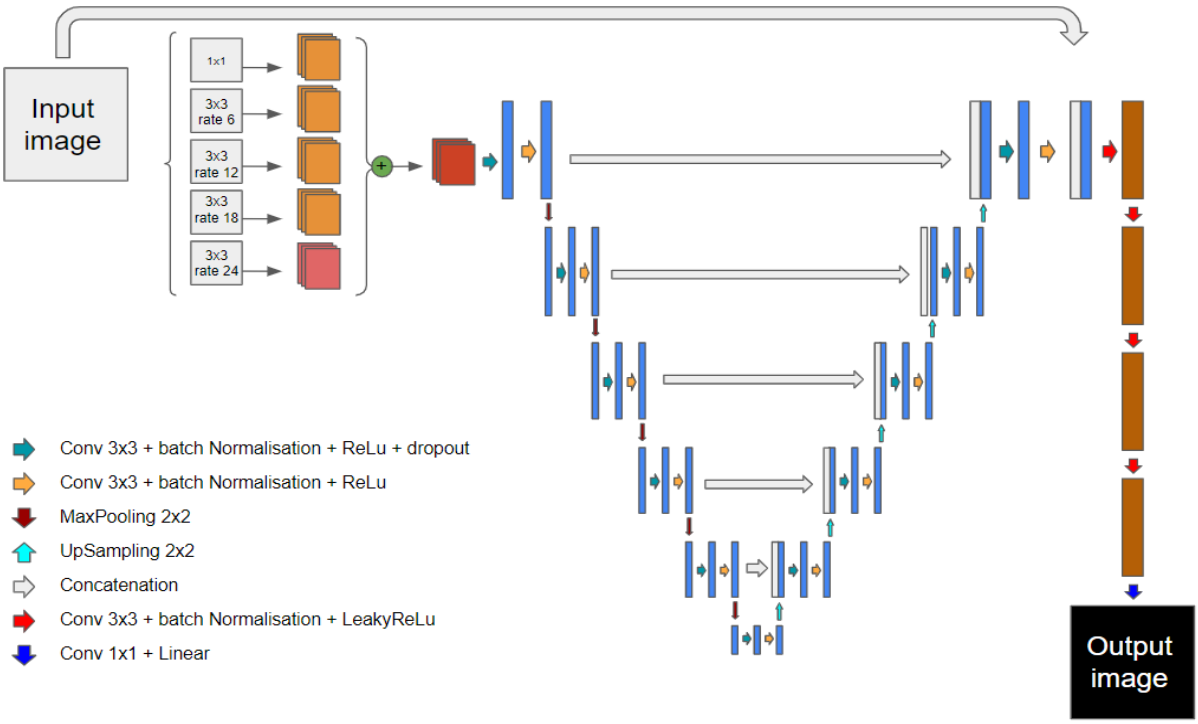

Figure 1. a-Net architecture

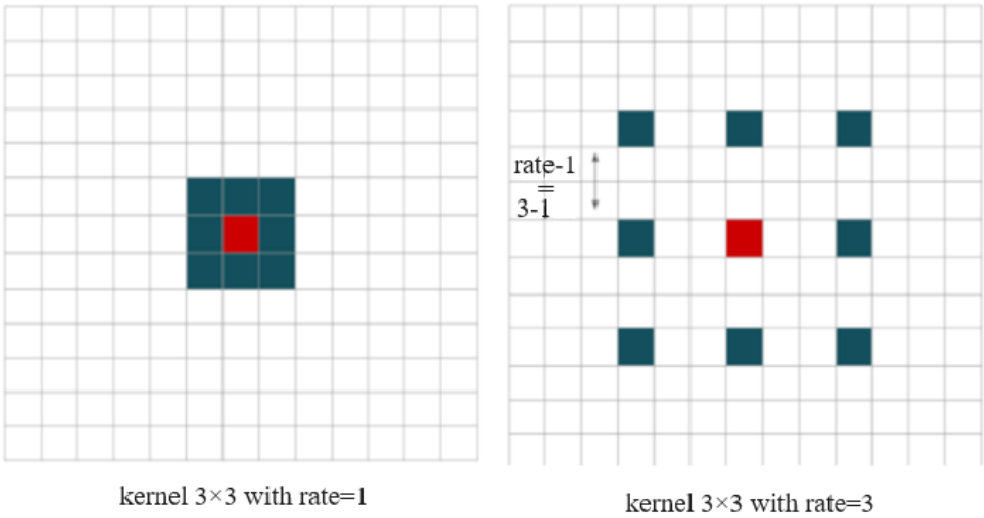

Figure 2. Atrous 3×3 convolution with varying rates

At the outset of our network, we utilize atrous convolution side by side. As elaborated above, using the stride would result in features remaining in the output but resolution decreases which is not desirable for semantic segmentation. Hence, we make five copies of convolution consisting of 1×1 and 3×3 convolutions applied on the input image with different dilation rates ranging from 6 to 24 with an increment of 6 in parallel. This has been inspired by the design proposed by Yu and Koltun [15]. This parallel path will cater to the problems associated with the diminishing resolution and also enhance the effectiveness of semantic segmentation.

3.2 Merging the features

Such integration of outputs from different layers is very important in generating robust feature representations. Using each output individually is challenged by sophisticated mechanisms of images, capture angles, and the strong presence of noise, which make it ever more difficult to interpret the meaning of each target. This was pointed out by Liu et al. [16] Since it incorporates dilated convolution to achieve parallel outputs, it is then of paramount importance to merge these outputs to make the optimization of the deep network easier and increase learning efficiency. However, since each of the five outputs of atrous convolution comprises 16 feature maps, concatenating all these outputs would give an output with a massive number of 80 feature maps, greatly increasing the training time. For this reason, the strategy adopted adds distinct feature maps at a size that will match the outputs from previous layers. The aim here is not only to generate features that are more appropriateness, but also to avoid a huge number of feature maps that increase the challenge, hence optimizing the performance of the network during training.

3.3 Auto-Encoding phases

In the last stage of our architecture chain, this output is further refined from the merge with a two-step contraction and expansion procedure, taken also from U-Net. This encoding step is mainly a series of convolutions where the 3×3 convolutions are followed by ReLU activation ($f(x) = x^+ = max(0, x)$) for dropping the negative values. Setting a 2×2 max-pooling layer, we will further reduce the spatial dimensions. The contractive path extracts a unified feature tensor using 12 convolutional layers, starting from 32 to 1024 feature maps. On the other hand, the successive expansive path [17] is simply a reversal of the former path, which starts from 1024 to 32 feature maps; it uses 2×2 upsampling layers concatenated with their corresponding paths from the first phase and convolves it by 3×3 and ReLU activates it.

It preserves the relevant structure and features by concatenating the output from the encoder to the input image. Finally, five 3×3 convolutions are applied, the first four of which are followed by the Leaky ReLU activation function, defined by ($f(x) = x\ if\ x > 0\ else\ 0.02 * x$), and the last layer utilizes a 1x1 convolution. Leak ReLU ensures that all negative values have a small, non-zero gradient, which helps ensure that all variables are trainable and contributes to a more robust training process. That last set of convolutions maps each of the feature vectors with 8 components the last two layers produced to the number of classes wanted at the end, culminating in a refined class-aware output of segmentation.

U-Net is good at capturing fine details in virtue of the symmetric encoder-decoder structure. However, it sometimes struggles to handle complex and multi-scale features better dealt by DeepLabv3 using atrous convolutions. Thus, DeepLabv3 may lose high spatial resolution in the feature maps. Our a-Net architecture alleviates these challenges by combining atrous convolution and feature fusion techniques to enhance feature extraction without losing spatial details. This approach uses both U-Net and DeepLabv3 strengths while improving on their limitations to give better results. The details in these parts further make it clearer why a-Net can work better than its competitors in certain respects, thus further enhancing the contribution of our work.

4. TRAINING

The training protocol is primarily grounded in utilizing input images paired with their corresponding masks for each image. Kingma and Ba [18] is utilized as the optimization algorithm, enhancing the gradient descent method through the integration of momentum and adaptive learning rate techniques. Its application in semantic segmentation is particularly advantageous, as it provides faster convergence and improved performance in complex image analysis tasks. Momentum aids in determining the optimal direction, while adaptive learning rate facilitates more effective step changes. An initial learning rate of 0.01 is employed for initialization. To maximize GPU memory usage and accelerate convergence speed towards minimum error, the batch size technique is applied, providing a judicious sampling for each epoch during training. Batch normalization layers are incorporated to reinforce training by ensuring speed and stability through the normalization of inputs via rescaling or recentering. Proposed by Ioffe and Szegedy [19], the effectiveness of batch normalization during training is well-documented, even though the exact reasons behind its efficacy remain a topic of discussion.

For the final layer, the softmax activation function is chosen due to its compatibility with the objective of pixel-wise classification for multiclasses. Softmax relies on probabilities and is defined by the formula

$\sigma(z) = \frac{e^{zi}}{\sum_1^k e^{zj}}$, for $i = 1, \ldots, k$, z = $(z_1, \ldots, z_k) \in R^k$,and $k \geq 1$. The a-Net enhances training and computes the model error using the softmax output alongside the ground truth segmentation. Focal loss is employed as the error function due to its proven effectiveness in mitigating class imbalance issues within each image—a persistent challenge in image processing. The focal loss function is defined as $FL(P_i) = -\alpha * (1 - P_i)\gamma * log(P_i)$, where P_i is the predicted probability of the positive class, α controls the weight of positive samples, and γ determines the focus on easy-to-classify ($\gamma = 0$) or hard-to-classify examples ($\gamma > 0$). The output size corresponds to the number of scalar values in the model output.

5. EXPERIMENTAL EVALUATION

In this section on evaluation, we trained our models for 400 epochs to garner the models' performance on different datasets. We used two different image datasets: one of large objects and the other of small objects, evidencing the flexibility and effectiveness of the proposed architecture.

The first dataset concerns 3D semantic segmentation, with the target of large image stacks resulting from electron microscopy recordings, focusing on mitochondria (Figure 3). This dataset is courtesy of EPFL [20], and it contains 165 image slices, each measuring 768×1024×3 pixels. For every image, there exists a ground truth segmentation map in which mitochondria are encoded in white and the background is encoded in black.

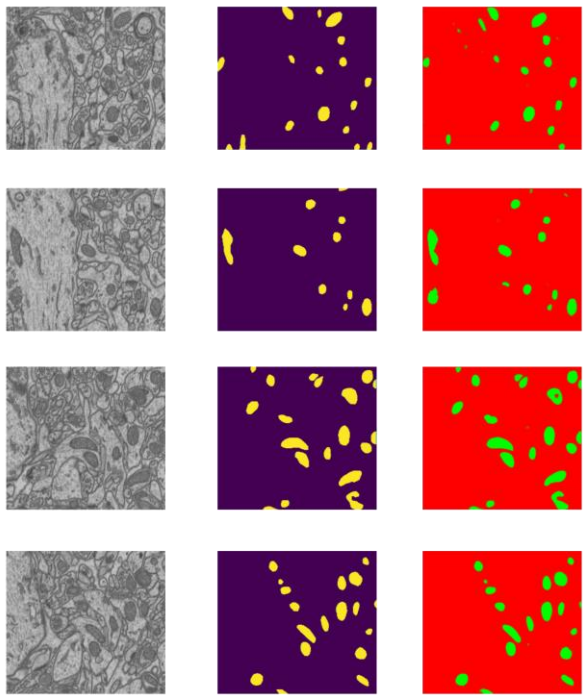

Figure 3. Our predictions in the Electron Microscopy problem (The first column corresponds to the input images, the second to their masks, and finally, our a-Net predictions)

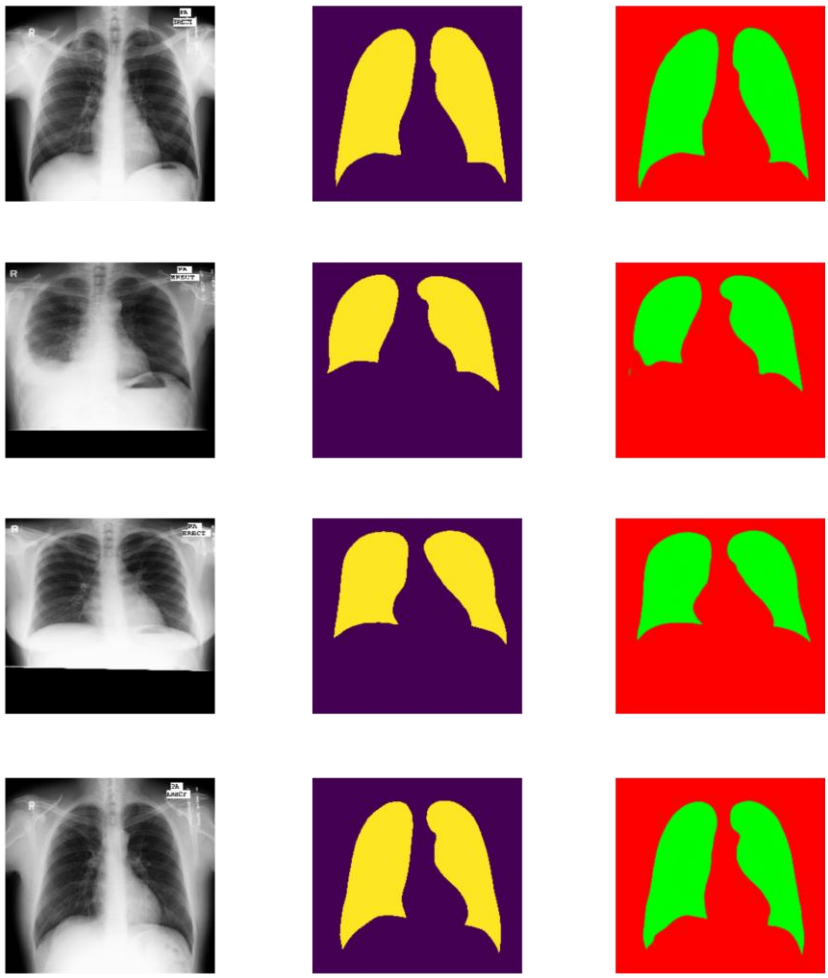

Figure 4. Our predictions in X-ray images from the tuberculosis control program

The second dataset was about segmenting X-ray images, most of which concerned tuberculosis. Figure (Figure 4). This dataset, sourced from the tuberculosis control program of the Montgomery County Department of Health and Human Services, MD, USA [21], consists of 704 radiographs with clinical labels. The dataset was compiled on the Kaggle platform that contained images for training and for validation.

Both of these datasets are considered among the most essential benchmarks with respect to evaluating new models' performance in semantic segmentation. We've used standard semantic segmentation evaluation metrics for both models: IoU, MSE, RMSE, F1-score, and Pixel Accuracy on these two diverse datasets (Table 1 and Table 2). This theoretical evaluation sheds light on the performance and generalizability of our model within two very different medical imaging domains, thus showing flexibility and efficiency of the a-Net architecture.

To this end, detailed preprocessing at the front end had to be done in order to ensure robustness. In the case of an electron microscopy image dataset, normalization and contrast enhancement were done to bring out the features of interest—the mitochondria. While for the X-ray dataset, histogram equalization and resizing were done to introduce uniformity in terms of dimensional input. The datasets were divided into training, validation, and test sets, ensuring that samples are diverse and representative. More specifically, the electron microscopy dataset contained 132 images for training and 16 for validation, leaving 17 for testing. For the X-ray dataset, the split consisted of 560 images for training, 70 for validation, and 74 for testing.

Results obtained from these experiments clearly resonate with the effectiveness of the a-Net architecture in performing a wide range of medical image analysis tasks, underlining its applicability elsewhere within semantic segmentation, with minimal changes in hyperparameters depending on the problem domain.

Table 1. Evaluation table of the three architectures on microscopy images of mitochondria

Name	MSE	RMSE	MIOU	Pixel Accuracy	F1
u-Net	0.0240	0.1549	0.7685	0.9740	0.7033
DeepLabv3	0.0203	0.1421	0.8202	0.9778	0.7889
a-Net	0.0201	0.1415	0.8144	0.9779	0.8977

Table 2. Evaluation table of the three architectures on the x-ray images for tuberculosis

Name	MSE	RMSE	MIOU	Pixel Accuracy	F1
u-Net	0.0213	0.1459	0.9466	0.9780	0.9584
DeepLabv3	0.0392	0.1978	0.9020	0.9601	0.9198
a-Net	0.0156	0.1246	0.9606	0.9837	0.9696

6. CONCLUSIONS

In this research effort, we proposed a new architecture—a-

Net—for semantic segmentation in medical applications. Our method was based on the seamless integration of atrous convolution, feature fusion, and the encoder-decoder technique that improves feature extraction and segmentation efficiency. We proved that a-Net could perform better than existing architectures like U-Net and DeepLab through different semantic segmentation tasks by elaborating very carefully on the different stages involved in our implementation.

First, it introduced parallel atrous convolutions, which retain spatial resolution; second, it had strategically merged feature maps for fastening deep network optimization; lastly, it had a specially designed encoding-decoding process following the U-Net. These components cement a-Net as a strong and versatile framework with the capability of solving complicated semantic segmentation tasks.

Despite these promising results, several limitations remain worthy of further investigation. An improvement in performance may be achieved with more hyperparameter tuning and use of further data augmentation techniques. Overall, the current implementation is computationally demanding and would possibly limit its applications, especially in resource-constrained scenarios—the need for additional research on more efficient architectures.

Future research efforts will be directed toward the adaption of the model with respect to the automation of tumor stage detection in breast cancer, an important step toward practical implementations. If this becomes realized, it would underscore the greater effect our work and its contributions have on the development of medical imaging applications. Other than this, we are going to investigate a-Net further concerning its scalability to other domains of medicine or other fields and increasing its adaptability to even more semantic segmentation problems.

7. LIMITATIONS

Despite the promising performance of our new architecture for semantic segmentation of medical images, several limitations should be acknowledged. Firstly, the architecture's efficacy is highly dependent on the quality and diversity of the training data. Limited availability of annotated medical images can restrict the model's generalizability across different types of medical imagery. Secondly, the computational complexity of the proposed architecture requires substantial processing power and memory, which may not be feasible in resource-constrained environments. Additionally, while our model has shown improved accuracy, it may still struggle with segmenting very small or highly irregular structures within medical images. Finally, the model's performance has primarily been evaluated on a specific set of medical imaging modalities, and its effectiveness across a wider range of imaging techniques remains to be thoroughly investigated. Further research is necessary to address these limitations and to enhance the robustness and versatility of our proposed architecture.

REFERENCES

[1] LeCun, Y., Boser, B., Denker, J.S., Henderson, D., Howard, R.E., Hubbard, W., Jackel, L.D. (1989). Backpropagation applied to handwritten zip code recognition. Neural Computation, 1(4): 541-551. https://doi.org/10.1162/neco.1989.1.4.541

[2] Krizhevsky, A., Sutskever, I., Hinton, G.E. (2017). ImageNet classification with deep convolutional neural networks. Communications of the ACM, 60(6): 84-90. https://doi.org/10.1145/3065386

[3] Simonyan, K., Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. Computer Vision and Pattern Recognition. https://doi.org/10.48550/arXiv.1409.1556

[4] Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A. (2015). Going deeper with convolutions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1-9. https://doi.org/10.1109/CVPR.2015.7298594

[5] He, K., Zhang, X., Ren, S., Sun, J. (2016). Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770-778. https://doi.org/10.48550/arXiv.1512.03385

[6] Long, J., Shelhamer, E., Darrell, T. (2015). Fully convolutional networks for semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3431-3440. https://doi.org/10.48550/arXiv.1605.06211

[7] Badrinarayanan, V., Kendall, A., Cipolla, R. (2017). Segnet: A deep convolutional encoder-decoder architecture for image segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence, 39(12): 2481-2495. https://doi.org/10.1109/TPAMI.2016.2644615

[8] Ronneberger, O., Fischer, P., Brox, T. (2015). U-net: Convolutional networks for biomedical image segmentation. In Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, pp. 234-241. https://link.springer.com/chapter/10.1007/978-3-319-24574-4_28.

[9] Chen, L.C., Papandreou, G., Kokkinos, I., Murphy, K., 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): 834-848. https://doi.org/10.1109/TPAMI.2017.2699184

[10] Chen, L.C., Papandreou, G., Schroff, F., Adam, H. (2017). Rethinking atrous convolution for semantic image segmentation. Computer Vision and Pattern Recognition, arXiv:1706.05587. https://doi.org/10.48550/arXiv.1706.05587

[11] Chen, S.Q., Zhan, R.H., Hu, J.M., Zhang, J. Chen, S.Q., Zhan, R.H., Hu, J.M., Zhang, J. (2017). Feature fusion based on convolutional neural network for SAR ATR. In ITM Web of Conferences, 12: 05001. https://doi.org/10.1051/itmconf/20171205001

[12] Pinheiro, P.O., Lin, T.Y., Collobert, R., Dollár, P. (2016). Learning to refine object segments. In Computer Vision-ECCV 2016: 14th European conference, Amsterdam, The Netherlands, Proceedings, pp. 75-91. https://doi.org/10.1007/978-3-319-46448-0_5

[13] Aniq, E., Chakraoui, M., Mouhni, N. (2024). Artificial intelligence in pathological anatomy: Digitization of the calculation of the proliferation index (Ki-67) in breast

carcinoma. Artificial Life and Robotics, 29(1): 177-186. https://link.springer.com/article/10.1007/s10015-023-00923-6.

[14] Aniq, E., Chakraoui, M., Mouhni, N., Aboulfalah, A., Rais, H. (2023). Breast cancer stage determination using deep learning. In World Conference on Information Systems and Technologies, pp. 550-558. https://link.springer.com/chapter/10.1007/978-3-031-45642-8_53.

[15] Yu, F., Koltun, V. (2015). Multi-scale context aggregation by dilated convolutions. Computer Vision and Pattern Recognition, arXiv:1511.07122. https://doi.org/10.48550/arXiv.1511.07122

[16] Liu, W., Rabinovich, A., Berg, A.C. (2015). Parsenet: Looking wider to see better. Computer Vision and Pattern Recognition. https://doi.org/10.48550/arXiv.1506.04579

[17] Noh, H., Hong, S., Han, B. (2015). Learning deconvolution network for semantic segmentation. In Proceedings of the IEEE International Conference on Computer Vision, pp. 1520-1528. https://doi.org/10.48550/arXiv.1505.04366

[18] Kingma, D.P., Ba, J. (2014). Adam: A method for stochastic optimization. Machine Learning. https://doi.org/10.48550/arXiv.1412.6980

[19] Ioffe, S., Szegedy, C. (2015). Batch normalization: Accelerating deep network training by reducing internal covariate shift. In International Conference on Machine Learning, pp. 448-456. https://doi.org/10.48550/arXiv.1502.03167

[20] Lucchi, A., Li, Y., Fua, P. (2013). Learning for structured prediction using approximate subgradient descent with working sets. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1987-1994. https://doi.org/10.1109/CVPR.2013.259

[21] Jaeger, S., Candemir, S., Antani, S., Wáng, Y.X.J., Lu, P.X., Thoma, G. (2014). Two public chest X-ray datasets for computer-aided screening of pulmonary diseases. Quantitative Imaging in Medicine and Surgery, 4(6): 475. https://doi.org/10.3978%2Fj.issn.2223-4292.2014.11.20