Progressive Alternating Attention for Bio-medical Image Segmentation

Abhishek Srivastava
Computer Vision & Pattern Recognition Unit
Indian Statistical Institute
Kolkata, India
abhisheksrivastava2397@gmail.com

Sukalpa Chanda
Department of Computer Science and Communication
Østfold University College
Halden, Norway
sukalpa@ieee.org

Debesh Jha
SimulaMet
UiT The Arctic University of Norway)
Tromsø, Norway
debesh@simula.no

Michael A. Riegler
SimulaMet
UiT The Arctic University of Norway
Tromsø, Norway
michael@simula.no

Pål Halvorsen
SimulaMet
UiT The Arctic University of Norway
Tromsø, Norway
paalh@simula.no

Dag Johansen
UiT The Arctic University of Norway
Tromsø, Norway
dag.johansen@uit.no

Umapada Pal
Computer Vision & Pattern Recognition Unit
Indian Statistical Institute
Kolkata, India
umapada@isical.ac.in

***Abstract*—Medical image segmentation can provide detailed information for clinical analysis. The detailed location of the disease can play a vital role in treatment and decision-making. The emergence of convolutional neural network (CNN) based encoder-decoder techniques have advanced the performance of automated medical image segmentation systems. Several such CNN-based methodologies utilize techniques such as spatial- and channel-wise attention to enhance performance. Another technique that has drawn attention in recent years is residual dense blocks (RDBs). The successive convolutional layers in densely connected blocks are capable of extracting diverse features with varied receptive fields and thus, enhancing performance. However, consecutive stacked convolutional operators may not necessarily generate features that facilitate the identification of the target structures. In this paper, we propose a progressive alternating attention network (PAA-Net). We develop progressive alternating attention dense (PAAD) blocks, which construct a guiding attention map (GAM) after every convolutional layer in the dense blocks using features from all scales. The GAM allows the following layers in the dense blocks to focus on the spatial locations relevant to the target region. Every alternate PAAD block inverts the GAM to generate a reverse attention map which guides ensuing layers to extract boundary and edge-related information, refining the segmentation process. Our experiments done on three different biomedical image segmentation datasets exhibit that our PAA-Net achieves favorable performance when compared to other state-of-the-art methods.**

***Index Terms*—Medical image segmentation, convolutional neural network, attention, semantic segmentation**

I. Introduction

Deep learning based medical image segmentation based methods have garnered a lot of attention during the past few years. Manual annotation of images to identify and locate the region-of-interest is a time consuming task. The accuracy of such annotation depends upon the expertise of the medical professionals making it prone to undesired oversights. Convolutional Neural Network(CNN) based techniques for delineation of desired anatomical regions from medical images have been proven to be effective. Methods like U-Net [1], U-Net++ [2], ResUNet++ [3], PraNet [4], Attention U-Net [5] have served as baselines capable of accurate segmentation performance. Attention mechanisms have served as an integral component in such methods. Attention U-Net uses the deep features from lower levels of the decoder to generate spatial attention maps, which are in turn used to prune irrelevant features from skip-connections. ResUnet++ [3] used the squeeze and excitation block [6] to model inter-dependencies between the channels to suppress irrelevant and enhance relevant channels. FED-Net [7] introduced Feature Fusion blocks which applied a combination of spatial- and channel-wise attention to increase the network's segmentation ability. Different combinations of channel and spatial attention at various stages of the encoder-decoder structure has also been used in FocusNet [8] and MSRF-Net [9]. PNS-Net [10] designed self-attention mechanism for video polyp segmentation to utilize both temporal and spatial features. PraNet [4] used reverse attention by generating an initial global guiding map which was later used for mining boundary cues. Another familiar technique for image segmentation in both medical imaging and natural computer vision is residual dense blocks(RDBs) [11]–[14]. The main advantage offered by RDBs is the combination of features obtained by both high- and low-receptive fields. These advantages motivated many works to incorporate RDB's in an encoder-decoder based architectures [15]–[17]. An important aspect of RDB's is that multiple convolutional layers with a smaller number of output channels are stacked on top of each other. This allows to progressively increase the receptive field while maintaining relevant low-level features. In this paper, we introduce a novel progressive alternating attention dense block (PAAD). After each convolutional layer in the dense block, a mini-decoder is used to generate a guiding attention map (GAM). The successive layers utilize this segmentation map to to prune features impertinent to identifying the region of interest. Additionally, we use reverse attention in every alternative PAAD block which allows layers to further mine peripheral features allowing the network to accurately capture the variation in shape and size of the region of interest. Since the GAM is created after every convolutional layer of the PAAD blocks, they are updated progressively which further refines the quality of feature maps, prune irrelevant features and allow the later convolutional layers to produce only meaningful features. We validate our method on three different biomedical datasets: Data Science Bowl(DSB) 2018, ISIC 2018 skin lesion segmentation, Kvasir-Instruments.

II. Method

In this section ,We elaborate upon the encoder, progressive alternating attention dense(PAAD) block used in our PAA-Net. The architecture of our PAA-Net is illustrated in Figure 1. The input image is encoded using ResNet-50 pretrained on ImageNet, which has served as a standard backbone for medical image segmentation. Let all features from different levels of the encoder be denoted as E_v, where $v \in 1, 2, 3, 4$.

A. Progressive Alternating Attention Dense Blocks

Each feature set from the encoder blocks is fed into the PAAD blocks. Equation 1 describes how the feature maps are generated for each convolutional layer within PAAD blocks. Here, F_v^c denotes the features generated by the layer number c for the $v'th$ resolution scale, $\oplus$ represents the concatenation operator and Con represents a 3×3 convolutional operator. P_v^0 is initially set to E_v. The architecture of the PAAD block is shown in Figure 1, for clarity we have only selected two distinct resolution scales in the figure to represent the operations and the feature flow within the PAAD block.

$$F_v^c = Con(P_v^{c-1} \oplus P_v^{c-2} \cdots \oplus P_v^0) \tag{1}$$

1) Mini-Decoder: We demonstrate the functioning of our Mini-Decoder block(see Figure 1) in Equation 2. For all scales, F_v^c is upscaled to the size of the ground truth map. They are concatenated before being processed by a convolutional layer with kernel size 3×3. Finally, sigmoid function is used to transform the feature maps within the range of [0,1]. The guiding segmentation map is supervised using the ground truth.

$$G^c = \sigma(Con(F_1^c \oplus F_2^c \oplus F_3^c \oplus F_4^c)) \tag{2}$$

TABLE I
Results on the 2018 Data Science Bowl

Method	**DSC**	**mIoU**	**Recall**	**Precision**
U-Net [1]	0.9080	0.8314	0.9029	0.9130
U-Net++ [2]	0.7705	0.5265	0.7159	0.6657
ResUNet++ [3]	0.9098	0.8370	0.9169	0.9057
Deeplabv3+ (Xception) [18]	0.8857	0.8367	0.9141	0.9081
Deeplabv3+ (Mobilenet) [18]	0.8239	0.7402	0.8896	0.8151
HRNetV2-W18-Smallv2 [19]	0.8495	0.7585	0.8640	0.8398
HRNetV2-W48 [19]	0.8488	0.7588	0.8359	0.8913
ColonSegNet [20]	0.9197	0.8466	0.9153	**0.9312**
ResUNet++ + CRF [21]	0.7806	0.7322	0.7534	0.6308
PraNet [4]	0.8751	0.7868	0.9182	0.8438
MSRF-Net	0.9224	0.8534	**0.9402**	0.9022
PAA-Net(Ours)	**0.9244**	**0.8627**	0.9319	0.9208

2) Progressive Alternating Attention: In this stage, the feature maps generated by the layer c in the PAAD block are multiplied by the guidance map G as described in Equation 3. The guidance maps is downscaled appropriately to the spatial dimensions of the v'th scale. In the first PAAD Block, we use spatial attention to allow convolutional layers to progressively prune irrelevant features while focusing on the region of interest as deemed by the guidance maps. In the next PAAD Block, the GAM is inverted as $G_v^c = 1 - G_v^c$, to allow all layers of the PAAD to extract further boundary and edge information which may have been omitted in the first PAAD block. This results in further refining the feature map and restricting the flow of extraneous features. We maintain informative deep level and shallow level features throughout the process consequently improving accuracy of our proposed PAA-Net. Further skip connections are added from the input to improve gradient flow.

$$P_v^c = F_v^c \otimes G_v^c \tag{3}$$

B. Decoder

The final output from the PAAD is denoted as P_v. The output from each decoder level is upscaled to the spatial resolution dimensions of the next decoder level as shown in Equation 4. The skip-connections from the corresponding scale is concatenated to the upscaled features before being fused using a 3×3 convolutional layer.

$$P_v = Con(Upscale(P_{v-1}) \otimes P_v) \tag{4}$$

Here, $Upscale$ represents a strided 4×4 transpose convolution layer. Each P_v where $v \in 2, 3, 4$ is upscaled to the spatial resolution of the ground truth and deeply supervised. Equal parts IoU loss and binary-cross entropy loss is used for supervision.

III. Experiments

To conduct experiments and determine the effectiveness of our method, we use three different biomedical image segmentation datasets with different types of region-of-interest. We use 2018 Data Science Bowl(DSB) Challenge which contains 670 segmented nuclei images. Next, we use KVASIR-Instruments which is a diagnostic and therapeutic tool segmentation dataset in gastro-intestinal endoscopy dataset [22].

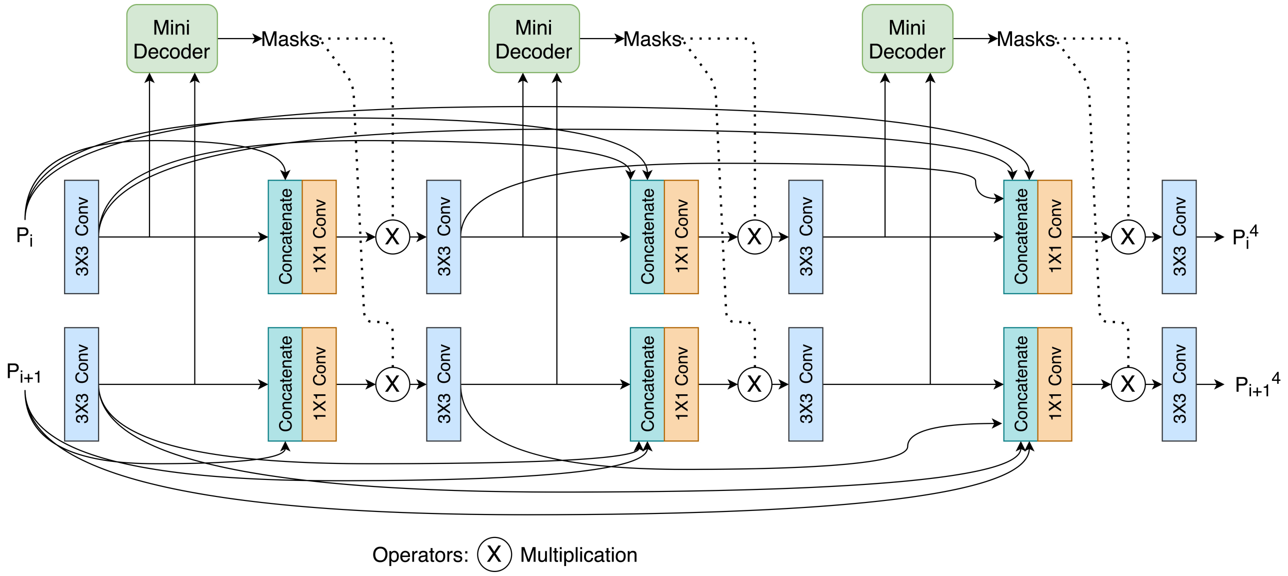

Fig. 1. The architecture of our progressive alternative attention dense block

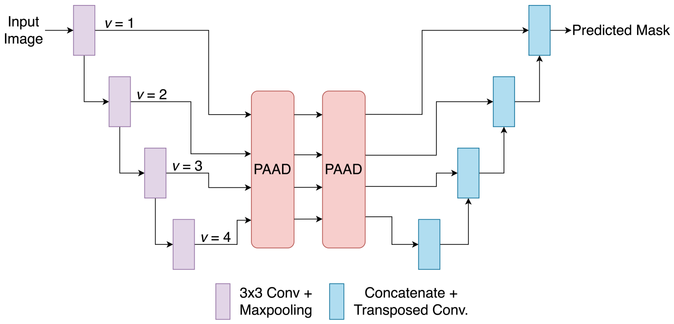

Fig. 2. Overview of the complete PAA-Net architecture.

TABLE II
RESULTS ON THE ISIC-2018 SKIN LESION SEGMENTATION CHALLENGE

Method	DSC	mIoU	Recall	Precision
U-Net [1]	0.8554	0.7847	0.8204	**0.9474**
U-Net++ [2]	0.8094	0.7288	0.7866	0.9084
ResUNet++ [3]	0.8557	0.8135	0.8801	0.8676
Deeplabv3+ (Xception) [18]	0.8772	0.8128	0.8681	0.9272
Deeplabv3+ (Mobilenet) [18]	0.8781	0.8236	0.8830	0.9244
HRNetV2-W18-Smallv2 [19]	0.8561	0.7821	0.8556	0.8974
HRNetV2-W48 [19]	0.8667	0.8109	0.8584	0.9155
ResUNet++ + CRF [21]	0.8688	0.8209	0.8826	0.8736
MSRF-Net [9]	0.8824	**0.8373**	0.8893	0.9348
PAA-Net (Ours)	**0.8912**	0.8219	**0.9019**	0.9054

TABLE III
RESULTS ON THE KVASIR-INSTRUMENTS

Method	DSC	mIoU	Recall	Precision
U-Net [1]	0.9158	0.8578	0.9487	0.8998
U-Net++ [2]	0.8808	0.8453	0.8623	0.9173
HRNetV2-W18-Smallv2 [19]	0.9272	0.8822	0.9244	0.9438
HRNetV2-W48 [19]	0.9306	0.8867	0.9294	0.9429
Deeplabv3+ (Xception) [18]	0.8998	0.8615	0.9012	0.9272
Deeplabv3+ (Mobilenet) [18]	0.9079	0.8635	0.9075	0.9468
ColonSegNet [20]	0.9201	0.8820	0.9169	0.9317
MSRF-Net [9]	0.9379	0.8990	**0.9661**	0.9283
PAA-Net(Ours)	**0.9495**	**0.9160**	0.9475	**0.9571**

Finally, we experiment on ISIC-2018 Challenge for segmentation of skin lesions. DSB and ISIC-2018 is divided into training, validation and testing splits which contains 80%, 10% and 10% of the data, respectively. For Kvasir-Instruments we follow the official train-test split. Adam optimizer was used with a learning rate of $1e-4$. All models were trained for 30 epochs with a batch size of 8. We have compared our work with state-of-the-art(SOTA) medical image segmentation methods such as U-Net [1],U-Net++ [2], ResUNet++ + CRF [21]. Further comparisons were made with standard semantic segmentation methods like HR-Net [23] and Deeplabv3+ [18].

The nuclie images found in DSB 2018 were captured under varying conditions like different cell size, magnification and imaging modality. This variation within the distribution makes segmentation of nuclie images, a challenging problem. Table I reports the results obtained by our PAA-Net on DSB 2018, we can observe that our PAA-Net avhieves a dice coefficient (DSC) of 0.9244 and mean intersection over union (mIoU) of 0.8627, outperforming SOTA MSRF-Net in both metrics.

Skin lesion segmentation assists in melanoma detection, melanoma being the most serious forms of skin cancer warrants an automatic skin lesion segmentation system. From Table II we can observe that PAA-Net reports a DSC of 0.8912 and recall of 0.9019, outperforming MSRF-Net by 0.88% in terms of DSC and 1.26% in terms of recall.

Tool segmentation in gastrointestinal images allows tracking of crucial instruments used during endoscopy and assist in robotic and non-robotic surgeries. Therefore developing such automated segmentation system may help in real-time complex surgeries in gastrointestinal tract organs. In Table III, it can be noted that PAA-Net attains an improvement of 1.16% in DSC and 1.70% in mIoU over best performing MSRF-Net. The

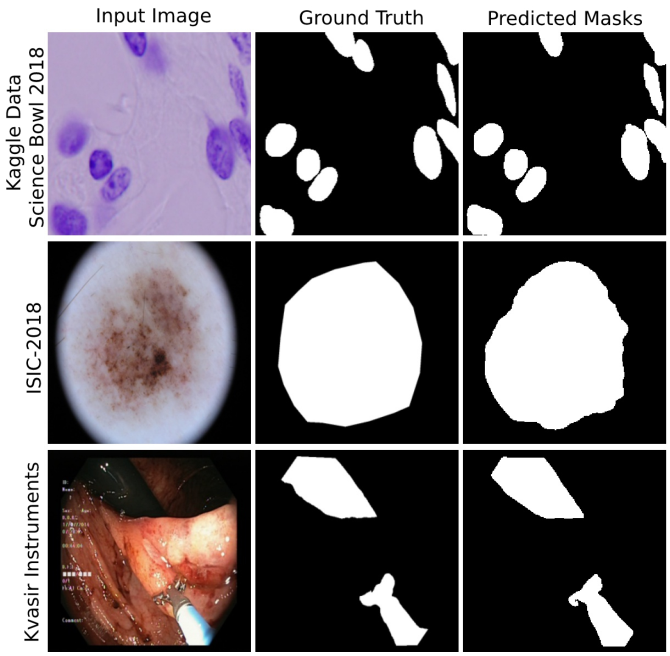

Fig. 3. Qualitative results of our proposed PAA-Net.

qualitative results of our approach can be observed in Figure 3. Overall, the progressive alternating attention mechanism used to control and limit the features contributed by convolutional layers in residual dense blocks enhances the segmentation ability of our PAA-Net. The consecutive GAM's which are progressively updated enables the generation of feature maps germane to the region-of-interest and subsequent mining of boundary cues generates detailed and spatially accurate segmentation maps.

IV. Conclusion

In this our work, we propose PAA-Net which employs a novel progressive alternating attention dense block. The PAAD uses an attention mechanism which alternatively guides the features contributed by the convolutional layers in dense block. First, the attention maps enables layers to focus the spatial locations pertinent to the target structure, then the reverse attention guides the layers to capture the boundary and edge information which further refines the features and allows generation of finer and spatially precise segmentation maps. Experiments done on three different biomedical image segmentation datasets validates our approach and posts new benchmarks. Our future work will comprise of modifying our PAA-Net for medical image classification.

References

[1] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: convolutional networks for biomedical image segmentation," in *Proc. of Internat. Confer. on Med. Ima. Compu. Comput.-Assis. Interven.*, 2015, pp. 234–241.

[2] Z. Zhou, M. M. R. Siddiquee, N. Tajbakhsh, and J. Liang, "UNet++: Redesigning skip connections to exploit multiscale features in image segmentation," *IEEE Trans. Med. Imag.*, vol. 39, no. 6, pp. 1856–1867, 2019.

[3] D. Jha *et al.*, "ResUNet++: An advanced architecture for medical image segmentation," in *Proc. of Internat. Sympos. Multime.*, 2019, pp. 225–230.

[4] D.-P. Fan *et al.*, "PraNet: parallel reverse attention network for polyp segmentation," in *Proc. of Internat. Confer. on Med. Ima. Compu. Comput.-Assis. Interven.*, 2020, pp. 263–273.

[5] O. Oktay *et al.*, "Attention U-Net: learning where to look for the pancreas," *arXiv preprint arXiv:1804.03999*, 2018.

[6] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. of Comput. Vis. and Patt. Recogn.*, 2018, pp. 7132–7141.

[7] X. Chen, R. Zhang, and P. Yan, "Feature fusion encoder decoder network for automatic liver lesion segmentation," in *Proc. of internat. sympos. biomed. imag.*, 2019, pp. 430–433.

[8] C. Kaul, S. Manandhar, and N. Pears, "Focusnet: An attention-based fully convolutional network for medical image segmentation," in *Proc. of Internat. Sympo. on Biomed. Imag.*, 2019, pp. 455–458.

[9] A. Srivastava, D. Jha, S. Chanda, U. Pal, H. D. Johansen, D. Johansen, M. A. Riegler, S. Ali, and P. Halvorsen, "Msrf-net: A multi-scale residual fusion network for biomedical image segmentation," *arXiv preprint arXiv:2105.07451*, 2021.

[10] G.-P. Ji, Y.-C. Chou, D.-P. Fan, G. Chen, H. Fu, D. Jha, and L. Shao, "Progressively normalized self-attention network for video polyp segmentation," *arXiv preprint arXiv:2105.08468*, 2021.

[11] Y. Zhang, Y. Tian, Y. Kong, B. Zhong, and Y. Fu, "Residual dense network for image super-resolution," in *Proc. of Comput. Vis. and Patt. Recogn.*, 2018.

[12] Z. Zhang, Q. Liu, and Y. Wang, "Road extraction by deep residual U-Net," *IEEE Geosci. and Remo. Sens. Lett.*, vol. 15, no. 5, pp. 749–753, 2018.

[13] J. Dolz, K. Gopinath, J. Yuan, H. Lombaert, C. Desrosiers, and I. B. Ayed, "Hyperdense-net: a hyper-densely connected cnn for multi-modal image segmentation," *IEEE Trans. Med. Imag.*, vol. 38, no. 5, pp. 1116–1126, 2018.

[14] N. Ibtehaz and M. S. Rahman, "Multiresunet: Rethinking the u-net architecture for multimodal biomedical image segmentation," *Neur. Networ.*, vol. 121, pp. 74–87, 2020.

[15] X. o. Yang, "Road Detection via Deep Residual Dense U-Net," in *Pro. of Internat. Joi. Conf. on Neu. Netwo.*, 2019, pp. 1–7.

[16] P. L. K. Ding, Z. Li, Y. Zhou, and B. Li, "Deep residual dense U-Net for resolution enhancement in accelerated MRI acquisition," in *Proc. of Medi. Imag. 2019: Ima. Proce.*, vol. 10949, 2019, p. 109490F.

[17] A. Srivastava, N. Sharma, S. Gupta, and S. Chandra, "Residual dense u-net for segmentation of lung ct images infected with covid-19," in *International Advanced Computing Conference*. Springer, 2020, pp. 17–30.

[18] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *Proc. of the Europ. conf. comput. vis.*, 2018, pp. 801–818.

[19] J. Wang and other, "Deep high-resolution representation learning for visual recognition," *IEEE Trans. on Patt. Analy. Mach. Intelli.*, p. 1–1, 2020.

[20] D. Jha *et al.*, "Real-Time Polyp Detection, Localisation and Segmentation in Colonoscopy Using Deep Learning," *IEEE Acc.*, 2021.

[21] D. Jha and Others, "A Comprehensive Study on Colorectal Polyp Segmentation with ResUNet++, Conditional Random Field and Test-Time Augmentation," *IEEE J. Biomed. Health Inform*, 2021.

[22] D. Jha, S. Ali, K. Emanuelsen, S. A. Hicks, V. Thambawita, E. Garcia-Ceja, M. A. Riegler, T. de Lange, P. T. Schmidt, H. D. Johansen *et al.*, "Kvasir-instrument: Diagnostic and therapeutic tool segmentation dataset in gastrointestinal endoscopy," in *International Conference on Multimedia Modeling*. Springer, 2021, pp. 218–229.

[23] J. Wang *et al.*, "Deep high-resolution representation learning for visual recognition," *IEEE trans. patt. analy. mach.*, 2020.