SN-FPN: Self-attention Nested Feature Pyramid Network for Digital Pathology Image Segmentation

Sanghoon Lee
Kennesaw State University
Marrietta, Georgia, USA
slee297@kennesaw.edu

ABSTRACT

Digital pathology has played a key role in replacing glass slides with digital images, enhancing various pathology workflows. Whole slide images are digitized pathological images improving the capabilities of digital pathology and contributing to the overall turnaround time for diagnoses. The digitized images have been successfully integrated with artificial intelligence algorithms assisting pathologists in many tasks, but there are still demands to develop a new algorithm for a better diagnosis process. In this paper, we propose a new deep convolutional neural network model integrating a feature pyramid network with a self-attention mechanism in three pathways: encoder, decoder, and sel-attention nested for providing accurate tumor region segmentation on whole slide images. The encoder pathway adopts ResNet50 architecture for the bottom-up network. The decoder pathway adopts the feature pyramid network for the top-down network. The self-attention nested pathway forms the attention map represented by the distribution of attention scores focusing on localizing tumor regions and avoiding irrelevant information. The experiment results show that the proposed model outperforms the state-of-the-art deep convolutional neural network models in terms of tumor and stromal region segmentation. Moreover, various encoder networks were equipped with the proposed model and compared with each other. The results indicate that the ResNet series using the proposed model outperforms other encoder networks.

CCS CONCEPTS

• **Computing methodologies → Image segmentation**.

KEYWORDS

breast cancer, deep learning, image segmentation, encoder, decoder

ACM Reference Format:
Sanghoon Lee. 2024. SN-FPN: Self-attention Nested Feature Pyramid Network for Digital Pathology Image Segmentation. In *2024 ACM Southeast Conference (ACMSE 2024), April 18–20, 2024, Marietta, GA, USA.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3564746.35YYYYY

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
ACMSE 2024, April 18–20, 2024, Marietta, GA, USA
© 2024 Association for Computing Machinery.
ACM ISBN 979-8-4007-0237-2/24/04…$15.00
https://doi.org/10.1145/3564746.35YYYYY

1 INTRODUCTION

Traditional pathology examines tissue samples using a microscope by a pathologist to diagnose diseases such as cancer, infectious diseases, and hematologic disorders, determining the characteristics of their development of abnormal cells [1, 2]. However, these examinations have been limited to manual examination of slides, subjective interpretation, and not easily shareable slides, causing time-consuming, subjective, and collaborative challenges. Digital pathology is a branch of pathology assisting traditional pathology by providing high-resolution digital slide images through digital slide scanners and image analysis tools quantifying and interpreting pathology data [3, 4]. Adopting digital pathology can enable pathologists to improve their workflow efficiency, accessibility, and quantitative analysis, advancing biomedical knowledge and research.

Whole slide images (WSIs) are digital slides representing entire pathology slides with high resolutions approximately 1.5 gigabytes per slide allowing detailed examination of tissue structures [5]. WSIs are typically obtained by a process of digital slide scanning converting glass pathology slides stained using various histological stains (i.e., hematoxylin and eosin (H&E)) into digital images at high magnification [6]. Since WSIs are the digitized representation of entire pathology slides, they enable pathologists to explore the detailed information of tissue samples across the entire image, enhancing the examination process of pathology images by collaborating with other pathologists and integrating with image analysis techniques [7]. As WSIs become a key component of digital pathology, it is necessary to develop a new methodology to enhance the whole slide image analysis.

The analysis of WSIs to facilitate research in digital pathology has been developed through advanced image analysis techniques such as object detection, image classification, and image segmentation [8]. Object detection techniques in digital pathology have been utilized for automatically detecting regions of interest such as tumors, immune cells, and nuclei [9, 10]. A typical object detection process includes data preprocessing, training a model, validating a model, indicating the spatial location (i.e., bounding box), and post-processing. Image classification techniques in digital pathology have been utilized for automatically categorizing or classifying entire images into a set of labels [11, 12]. Image classification techniques are very similar to object detection techniques but assign a single label to the entire patch or image. Image segmentation techniques have been used in digital pathology to delineate different regions based on visual context. While object detection and image classification aim to obtain the object or class label, image segmentation is interested in partitioning an image pixel into meaningful

or homogeneous regions [13, 14]. In this paper, we focus on image segmentation techniques in digital pathology.

Traditional image segmentation approaches have mainly utilized hand-engineered models grouping pixels with similar properties or identifying boundaries between regions [15, 16]. These models highly rely on handcrafted features such as edges, corners, and even entropy, which are computationally simple but struggle with scalability and effectiveness on complex features. To remedy these issues on hand-engineered models, deep learning-based models have been introduced for capturing complex representations and contextual information from raw data aiming at accurate segmentation of regions [17, 18].

Deep-learning models, particularly convolutional neural networks (CNNs), have significantly contributed to image segmentation [19, 20]. The convolutional layers in CNNs can enable the deep-learning model to identify local patterns of spatial information in specific regions, the down-sampling layers interspersed between the convolutional layers can down-sample the spatial dimension of the prior map, reducing computational complexity and building spatial hierarchies of features, and the up-sampling layers recover the spatial resolution equal to the input image. CNNs automatically excel at learning hierarchical and spatial features from input images through the combination of convolutional, down-sampling, and up-sampling layers.

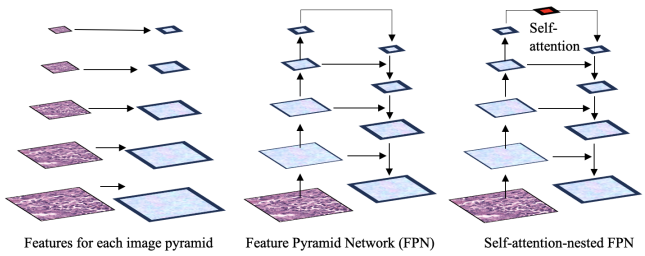

Figure 1: Comparision of Image Pyramid, Feature Pyramid Network, and the Proposed Self-attention Nested Feature Pyramid Network

Feature map generation at different scales using CNNs has been a widely accepted approach for deep learning-guided applications such as object detection, image classification, and image segmentation [21–23]. Feature maps obtained from multi-resolution representations of image pyramids can be independently used for image prediction or the feature maps can be combined with the skip connections producing a single-level feature map through a top-down architecture called Feature Pyramid Network (FPN) [24]. However, using feature maps for every prediction requires a large amount of computation and the top-down model with the skip connections remains to improve the effectiveness of the applications.

In this paper, we propose a feature pyramid network combined with a concept of a self-attention model equipped over the last state of the encoder creating an attention map for the first state of the decoder. We named it the Self-attention Nested Feature Pyramid Network (SN-FPN) for digital pathology image segmentation. SN-FPN takes advantage of self-attention mechanisms handling irregular shapes or contours of image regions and capturing contextual relevance and spatial relationships between pixels in an image. The differences between the proposed model with other models are briefly shown in Figure 1. Features are merely extracted from each image pyramid for predictions (left). Features are extracted by using the feature pyramid network through the bottom-up and top-down pathways; the features on the top-down pathway are merged with the corresponding features on the bottom-up pathway (middle). The proposed self-attention mechanism is nested between the bottom-up and the top-down pathway (right). We describe the details of the proposed SN-FPN model in Section 3.

2 RELATED WORKS

FPN is a multi-resolution feature pyramid network that has been widely studied in various types of models. Mask Region-based Convolutional Neural Network (R-CNN) adopted the concept of multi-scale feature maps of FPN to improve its performance in object detection tasks by capturing object information at different scales [25]. RetinalNet is an object detection model that aims to solve the class imbalance problem by assigning different weights on different scale objects using the Focal Loss [26]. To address the problem, RetinalNet employed FPN as a backbone architecture to enable the top-down architecture with lateral connections thereby capturing multi-resolution semantic information. Cascade R-CNN is an extended version of Faster-RCNN, using a series of detection branches in a cascaded manner refining the results of the prior stage [27]. Both FPN and Cascade R-CNN are similar in terms of multiple stages but Cascade R-CNN follows a cascaded structure. Neural Architecture Search for FPN (NAS-FPN) is an extended version of FPN combined with the concept of Neural Architecture Search. NAS-FPN aims to automatically design feature pyramid architecture through the optimization of the FPN architecture [28]. High-Resolution Network (HRNet) is a multi-scale convolutional neural network that maintains high-resolution images through parallel multi-resolution convolutions and repeated-resolution fusions [29]. Path Aggregation Network (PANet) is an instance segmentation deep convolutional neural network aiming to improve the feature pyramid network by taking accurate localization information at low levels [30]. However, these related works mainly focus on object detection or pose estimation and do not handle attention mechanisms.

Attention Aggregation-based Feature Pyramid Network (A^2-FPN) utilized the attention mechanism for its network pipeline to improve the multi-scale feature aggregation [31]. Multi-attention Object Detection Model (MA-FPN) took the advantage of attention mechanism by adding pixel feature attention structure through the multi-scale convolution branches [32]. Position Attention Guided Connection Network (PAC-Net) emphasized position attention by capturing salient dependencies with accurate location information for an effective 3D image detection model [33]. Although these methods utilized attention mechanisms and improved FPN, their approaches have been limited to adopting attention at multi-resolution. Moreover, the demonstration of the effectiveness of attention-equipped FPN is also limited for histopathology image segmentation.

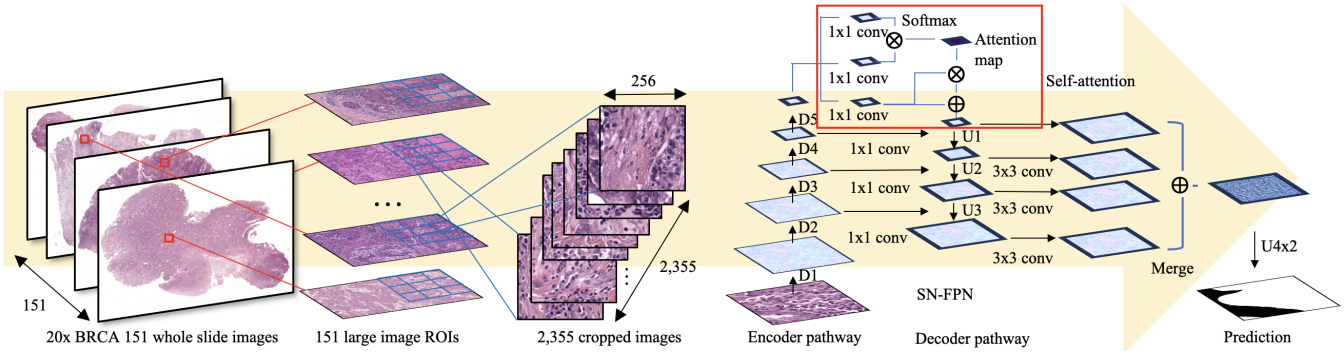

Figure 2: The Overview of the Self-attention Nested Feature Pyramid Network (SN-FPN) for Whole Slide Image Segmentation

3 SELF-ATTENTION NESTED FPN FOR DIGITAL PATHOLOGY IMAGE SEGMENTATION

In this section, we describe the proposed SN-FPN model for digital pathology image segmentation. First, we will describe the preprocessing steps for whole slide image segmentation. This step will explain how the input source of the SN-FPN can be obtained from whole slide images. Second, we will present the encoder pathway of the SN-FPN. Third, we will explain how the self-attention mechanism can be nested over the output of the encoder pathway. Last, we will show the decoder pathway of the SN-FPN.

3.1 Whole slide image segmentation

Whole slide images are high-resolution digital images of entire tissue cells. Segmenting whole slide images can be done by partitioning the images into regions of interest. Since the typical size of whole slide images is approximately 1.5 gigabytes, it is inevitable to divide the image into distinct regions. In this paper, 151 hematoxylin and eosin-stained (H&E) whole slide images provided by the Breast Cancer Semantic Segmentation (BCSS) dataset were used for the input source of the SN-FPN. 2,355 H&E images were cropped from 151 large regions of interest images obtained at 20x magnification [34]. Since these H&E images were created by different laboratories using different scanners, it is necessary to perform a color normalization reducing variations in color and intensity. Color normalization is a process of standardizing image color to avoid irrelevant color variability and is a critical part of whole slide image segmentation. This process ensures that whole slide images are more interpretable when training a deep-learning model, improving the generalization of the model for a new dataset.

Reinhard color normalization was adopted for whole slide image segmentation [35]. First, we converted the RGB image to Ruderman's LAB color space to compute the mean and standard deviation of the image intensities for each channel. Since the source image is the RGB image, the number of channels is three. Next, the mean and standard deviation obtained in the LAB color space were used for transforming the image color characteristics to the standard color characteristics for normalization. In this step, the LAB color space was scaled to unit variance with zero mean and rescaled and recentered to match the reference image mean and standard deviation [36]. 2,355 H&E images were cropped from 151 large image regions of interest and used as the input source of the SN-FPN. The scaled pixel value $\widehat{P_{ij}}$ to unit variance for each channel is defined as:

$$\widehat{P_{ij}} = (P_{ij} - \mu_j)/\sigma_j \tag{1}$$

P_{ij} is the i-th pixel value of the j-th channel in the LAB color space. μ is the mean of the j-th channel. σ is the standard deviation of the j-th channel. The color-normalized image patches were then used as the input source of the encoder pathway. The patches were down-sampled by a factor of 2 for each stage of the encoder pathway and the output of the encoder pathway was used as the input source of the self-attention nested pathway. In the self-attention nested pathway, an attention map was created by using a softmax max function normalizing the attention scores computed by the multiplication of the two feature maps convolved by the filter of 1 by 1 from the output of the encoder pathway. Another feature map convolved by the filter of 1 by 1 was then merged by the attention map multiplied by the convolved feature map. The output of the self-attention nested pathway was used as the input feature map of the decoder pathway. The input feature maps obtained by the attention mechanism were up-sampled by a factor of 2 for each stage of the decoder pathway. These up-sampled feature maps were then merged with the corresponding feature maps with the same resolution convolved by the filter of 1 by 1 in the encoder pathway. These merged feature maps are then used to generate additional feature maps to be merged for the prediction. The details of the whole slide image segmentation using SN-FPN are shown in Figure 2.

3.2 Encoder pathway

The encoder pathway is the bottom-up network extending from one feature map to another. This bottom-up network performs the feed-forward process consisting of several stages where each stage consists of convolutional layers increasing the number of channels followed by pooling layers which reduce the spatial dimension of the prior feature map, avoiding computational complexity. The output feature map is produced at the last stage of the feed-forward process.

We have adopted ResNet50 [37] for our encoder pathway to produce the output feature map which will later be used as the

input source of the self-attention nested pathway. Each stage in the encoder pathway consists of image convolutions followed by pooling or down-sampling by a factor of 2. The number of image convolutions performed in five stages is 1, 3, 4, 6, and 3 respectively. Down-samplings were performed in the four blocks generating 256, 512, 1024, and 2048 feature maps respectively. The details of the encoder pathway are shown in the leftmost blocks in Figure 3.

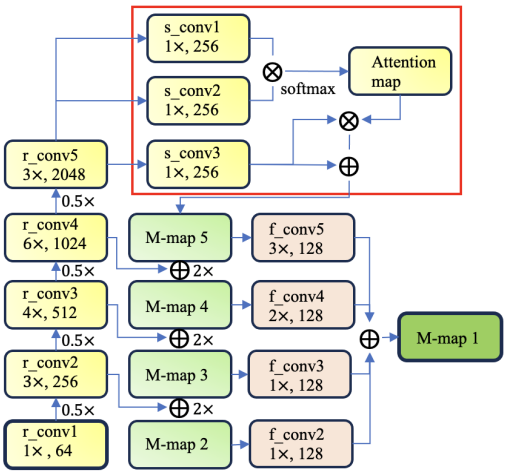

Figure 3: Self-attention Architecture Nested in the Encoder and Decoder Pathways

3.3 Decoder pathway

The decoder pathway is the top-down network extending from one feature map to another. While the encoder pathway down-samples features, the decoder pathway up-samples features mitigating the potential loss of spatial information during down-sampling and facilitating the generation of output images. We have adopted the top-down pathway of the FPN [24], but the output features of the self-attention nested pathway are used as the input features of the decoder pathway. These input features are up-sampled by a factor of 2 and merged with the features of the same size from the encoder pathway. The same-sized features are generated by convolving the filter of size 1 by 1 across the corresponding features from the encoder pathway. The middle blocks in Figure 3 represent the merged maps with the output of the self-attention nested pathway and corresponding feature maps in the encoder pathway.

While iterating this process until we reach the second feature map, four additional feature maps are generated by convolving the filter of size 3 by 3 across the merged feature maps in the decoder pathway. Because these additional feature maps are intentionally to be the same sizes, they are merged into one feature map for prediction. These additional feature maps are shown as four blocks on the right side of Figure 3. The rightmost block in Figure 3 shows the last feature map merged through the four additional feature maps. The last feature map is then up-sampled to be the same resolution as the input image for the prediction.

3.4 Self-attention nested pathway

Self-attention nested pathway is the middle network between the encoder pathway and the decoder pathway. Since the semantic feature values increase as the spatial resolution decreases, we assume that it is necessary to train the model to focus on relevant information rather than on redundant or irrelevant information from the input source thereby leading to accurate and more context-aware prediction results. This process can be done by forming an attention map providing the relatedness of each pixel information in the self-attention nested pathway. The attention map is generated from the attention weights representing the distribution of attention created from the raw attention scores.

To compute the raw attention scores the feature maps in the last block of the encoder pathway were used to derive three feature maps convolved by the filter of size 1 by 1, corresponding to the concepts of query, key, and value in the attention mechanism. The raw attention scores were then computed by matrix multiplication of two feature maps: query and key. After obtaining the raw attention scores, we computed the attention weights by using a softmax function across all feature maps forming an attention map. The attention map was then multiplied by the feature map called value and the results of the multiplication were merged with the value to create the merged feature map used as the input source of the decoder pathway. The merged feature map F_m in the self-attention nested pathway is defined as:

$$F_m = F_v + softmax(F_q F_k^T) \times F_v^T \tag{2}$$

where F_v represents a feature map that plays a role as a value, F_q represents a feature map that plays a role as a query, and F_k represents a feature map that plays a role as a key. The details of the self-attention nested pathway are shown at the top of Figure 3

4 EXPERIMENT RESULTS

In this section, we will perform experiments to verify the effectiveness of the proposed SN-FPN model by comparing it with the baseline of the model and the state-of-the-art models. Moreover, we will explore the effectiveness of the various encoder pathways by adopting deep-learning networks.

4.1 H&E image segmentation for tumor regions

Dataset: The Breast Cancer Semantic Segmentation (BCSS) dataset is a H&E image dataset that contains more than 20,000 segmentation annotations of tissue regions obtained from breast cancer whole slide images provided by The Cancer Genome Atlas (TCGA) [34]. The BCSS dataset was built on the collaborative effort of pathologists including senior and junior regents, and medical students of pathology by using the Digital Slide Archive, a web-based platform for whole-slide digital pathology images. 151 large ROIs of whole-slide images were annotated by them enabling the generation of accurate deep-learning models for tissue region segmentation.

We obtained 151 large ROIs extracted at 20x magnification from the BCSS dataset. A total of 2,355 H&E images were cropped to 256x256 size and resolution from the 151 large image ROIs. These H&E images were used as the dataset for validating the performance of the proposed SN-FPN on the tumor and stroma region segmentation. To perform a tumor region segmentation, we labeled

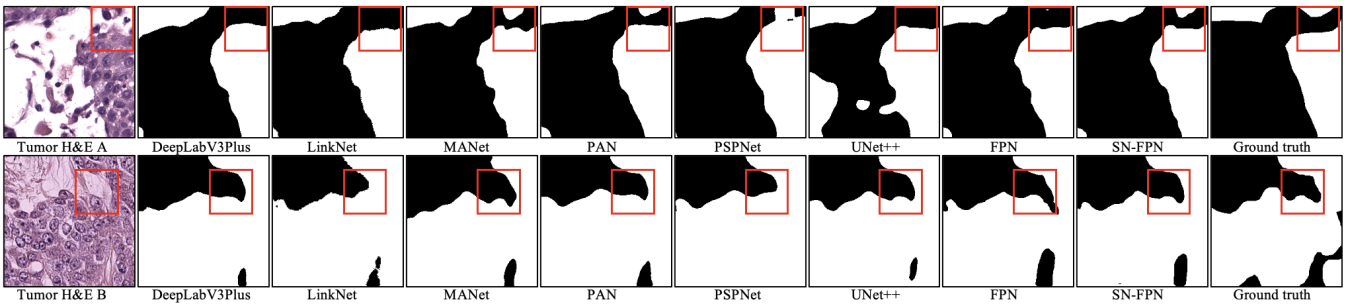

Figure 4: Examples of Tumor Region Predictions using DeepV3Plus, Linknet, MANet, PAN, PSPNet, Unet++, FPN, and SN-FPN

all the pixels as 'non-tumor' if the pixels are not annotated as a tumor. For training SN-FPN, the 2,355 H&E images were randomly divided into three datasets: training, validating, and testing, with 1,413(60%), 417(20%), and 471(20%) respectively.

Baseline and metrics: We compare the proposed SN-FPN with the state-of-the-art deep learning-based semantic segmentation models including DeepLabV3Plus [38], UNet++ [39], LinkNet [40], MANet [41], PAN [42], PSPNet [43], and FPN [24], using ResNet50 as an encoder pathway. We use them as the baselines to validate the effectiveness of the proposed model. DeepLabV3Plus is an extended version of the DeepLabV3 model [44] developed by Google Research teams and used for semantic segmentation. DeepLabV3Plus is well-known for capturing multi-scale contextual information by using dilated convolutions integrating with an atrous spatial pyramid pooling [38]. UNet++ is an extended version of the UNet model [45] used for capturing contextual information through the encoder and decoder networks. UNet++ especially uses an adaptive feature selection to dynamically select relevant features at different resolutions. LinkNet is a semantic segmentation model using the encoder and decoder networks. LinkNet uses some link modules for refining high-resolution information by providing an efficient upsampling method. MANet is a multi-scale attention network developed by Tongle et al [41]. MANet uses a self-attention mechanism for integrating local features by capturing contextual dependencies rather than applying the multi-scale feature combination. PAN is a pyramid attention network presented for exploiting and capturing global contextual information in semantic image segmentation [42]. PSPNet is a pyramid scene-parsing network aggregating multi-region context information and providing pyramid pooling models [43]. We use these models for our baselines.

We use well-known evaluation metrics such as Intersection over Union (*IoU*), F score ($F1$), Accuracy (*ACC*), Precision (*PREC*), and Recall (*REC*). *IoU* is defined as $(I + \epsilon)/U$ where I represents the number of pixels overlapped between the ground truth pixels and the predicted pixels. ϵ is set to $1e - 7$ and added to avoid zero division. U represents the encompassed area by both the ground truth pixels and the predicted pixels. U is defined as $N_g + N_p - I + \epsilon$ where N_g is the number of the ground truth pixels and N_p the number of predicted pixels. We use a confusion matrix including true positive: TP, false positive: FP, false negative: FN, and true negative: TN, to compute the $F1$, *ACC*, *PREC*, and *REC*. *PREC* is defined as (TP+ϵ)/(TP+FP+ϵ), *REC* is defined as (TP+ϵ)/(TP+FN+ϵ), *ACC* is defined as (TP+TN) / (TP+FP+FN+TN), and $F1$ is defined as $((1 + \beta^2)$ * TP + $\beta^2)$ / $((1 + \beta^2)$ * TP + β^2 * FN + FP + $\epsilon)$ where β is set to 1.

Table 1: Performance Results on Tumor Region Segmentation

Types	IoU	F1	ACC	PREC	REC
DeepV3Plus [38]	0.7466	0.8067	0.9116	0.8500	0.8539
Linknet [40]	0.7450	0.8051	0.9095	0.8532	0.8555
MANet [41]	0.7370	0.7976	0.9029	0.8109	0.8953
PAN [42]	0.7425	0.8023	0.9097	**0.8866**	0.8220
PSPNet [43]	0.7383	0.7991	0.9087	0.8388	0.8658
Unet++ [39]	0.7483	0.8085	0.9154	0.8575	0.8510
FPN [24]	0.7564	0.8152	0.9140	0.8403	0.8913
SN-FPN	**0.7740**	**0.8317**	**0.9156**	0.8478	**0.9019**

Results: The performance results on tumor region segmentation comparing the proposed SN-FPN with baseline models are shown in Table 1. The results indicate that the SN-FPN outperforms other models in terms of *IoU*(77.40%), $F1$(83.17%), *ACC*(91.56%), and *REC*(90.19%), and the PAN outperforms other models in terms of *PREC*(88.66%). Since the precision aims to measure the accuracy of the positive predictions indicating the model produces a lower false positive or a higher true positive, it may not be agreeable for other situations where the model requires a lower false negative. Thus, it is necessary to consider the harmonic averages of prediction and recall including both false positive and false negative. To address this consideration we computed the F score in this experiment and the proposed SN-FPN model showed the best result on tumor region segmentation in terms of $F1$. Moreover, the result of the higher *IoU* on the proposed SN-FPN model indicates that the predicted tumor segmentation regions are more aligned with the ground-truth tumor regions showing the best localization and accurately delineating the target tumor regions. The examples of the performance comparison on tumor region segmentation are shown in Figure 4.

4.2 H&E image segmentation for stroma regions

Dataset: The 151 large H&E images from the BCSS dataset were used for stroma region segmentation. As we used in the tumor region segmentation, we randomly split the 2,355 H&E images of 256x256 size into 1,413(60%), 417(20%), and 417(20%) H&E images for training, validating, and testing respectively.

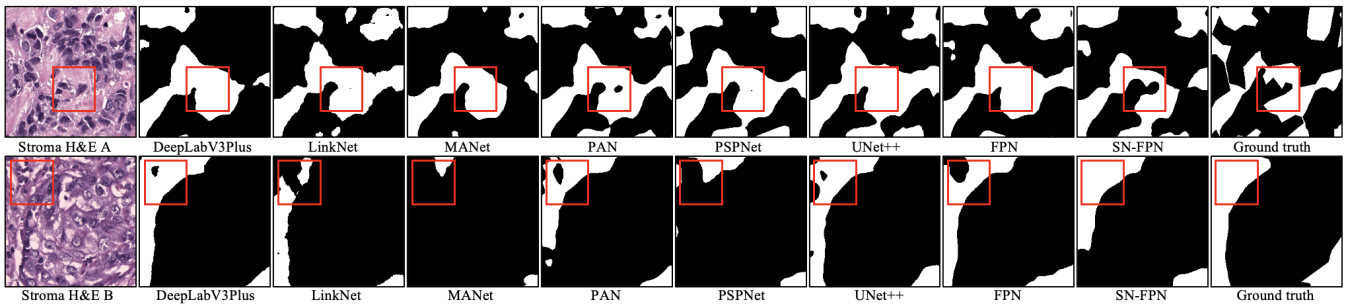

Figure 5: Examples of Stroma Region Predictions using DeepV3Plus, Linknet, MANet, PAN, PSPNet, Unet++, FPN, and SN-FPN

Table 2: Performance Results on Stroma Region Segmentation

Types	IoU	F1	ACC	PREC	REC
DeepV3Plus [38]	0.5404	0.6162	0.8673	0.6884	0.7632
Linknet [40]	0.5372	0.6167	0.8646	0.7108	0.7089
MANet [41]	0.5221	0.5982	0.8534	0.6705	0.7577
PAN [42]	0.5205	0.5908	0.8659	0.6566	**0.7793**
PSPNet [43]	0.5382	0.6173	0.8657	0.7057	0.7193
Unet++ [39]	0.5443	0.6199	0.8647	0.6997	0.7508
FPN [24]	0.5359	0.6097	**0.8677**	0.6701	0.7788
SN-FPN	**0.5452**	**0.6219**	0.8533	**0.7179**	0.7284

Baseline and metrics: The baseline models are the same as the models used in the previous section. We compare the models: DeepLabV3Plus [38], UNet++ [39], LinkNet [40], MANet [41], PAN [42], PSPNet [43], and FPN [24] with the proposed SN-FPN model using the same evaluation metrics: *IoU*, *F*1, *ACC*, *PREC*, and *REC*.

Results: The performance results on stroma region segmentation are shown in Table 2. The results indicate that the SN-FPN outperforms other models in terms of *IoU*(54.52%), *F*1(62.19%), and *PREC*(71.79%). The FPN outperforms other models in terms of *ACC*(86.77%) and the PAN outperforms other models in terms of *REC*(77.93%). Since the accuracy aims to mainly measure the proportion of correctly predicted samples over the total samples, the *ACC* on FPN may not properly provide information about mispredicted samples determined by false positives and false negatives. Moreover, Although *REC* on PAN can provide information about both true positive and false negative samples, it is difficult to mention that PAN outperforms other models because the models should consider the harmonic averages of both prediction and recall. The SN-FPN showed the best *F*1 score combining precision and recall, indicating that the proposed model outperforms other models on stroma region segmentation using the BCSS dataset. Moreover, the *IoU* score obtained from the SN-FPN represents the proposed model outperforms other models in terms of localization and delineation of the stroma region prediction. The examples of the performance comparison on stroma region segmentation are shown in Figure 5.

4.3 Comparison of the encoder pathways on tumor region segmentation

Dataset: We used the same dataset of the tumor region segmentation described in Section 4.1. The 2,355 H&E images were used for the comparison of the encoder pathways on tumor region segmentation and divided into three datasets for training 1,413(60%), validating 417(20%), and testing 417(20%) respectively.

Baseline and metrics: We conduct experiments on tumor region segmentation using various encoder pathways containing ResNet50 [37], ResNet101 [37], ResNet152 [37], MobileNetv2 [46], EfficientNet-b0, b1, b2, b3, b4, b5, b6, and b7 [47], VGG16 [48], and VGG19 [48]. ResNet101 and ResNet152 are deeper versions of ResNet50 extending layers to 101 and 152 respectively. MobileNet-v2 is an extended version of MobileNet providing an efficient inverted residual block using lightweight depthwise convolutions [46]. EfficientNet is a scaling-emphasized architecture addressing network balance based on different depths, widths, and resolutions thereby leading to better performance [47]. This architecture provides different variants such as b0, b1, b2, b3, b4, b5, b6, and b7. VGG16 and VGG19 are both representative deep convolutional neural network architectures presented by the Visual Geometry Group. VGG16 consists of 16 layers while VGG19 consists of 19 layers [48]. We compare them each other using the tumor region segmentation dataset. We use the same evaluation metrics: *IoU*, *F*1, *ACC*, *PREC*, and *REC*.

Results: The experiment results on the comparison of the tumor region segmentation using different encoder pathways are shown in Table 3. The results on the tumor region segmentation indicate that ResNet50 outperforms other pathways in terms of *IoU*(77.40%), *F*1(83.17%), and *REC*(90.19%). VGG16 outperforms other encoder pathways in terms of *PREC*(88.22%) and EfficientNet-b6 outperforms other encoder pathways in terms of *ACC*(91.60%). However, the harmonic average *F*1 of the precision and recall represents that ResNet50 shows better performance on the tumor region segmentation than other pathways in regards to the BCSS dataset.

4.4 Comparison of the encoder pathways on stroma region segmentation

Dataset: We used the same dataset (2,355 H&E images) of the stroma region segmentation described in Section 4.2 for the comparison of the encoder pathways on stroma region segmentation.

Table 3: Experiment Results on Tumor Region Segmentation with Different Encoder Pathways

SN-FPN +	IoU	F1	ACC	PREC	REC
ResNet50 [37]	**0.7740**	**0.8317**	0.9156	0.8478	**0.9019**
ResNet101 [37]	0.7630	0.8189	0.9138	0.8565	0.8751
ResNet152 [37]	0.7633	0.8204	0.9141	0.8313	**0.9019**
Effit.Net-b0 [47]	0.7537	0.8147	0.9103	0.8445	0.8639
Effit.Net-b1 [47]	0.7616	0.8195	0.9103	0.8421	0.8915
Effit.Net-b2 [47]	0.7530	0.8110	0.9071	0.8580	0.8654
Effit.Net-b3 [47]	0.7659	0.8220	0.9146	0.8489	0.8899
Effit.Net-b4 [47]	0.7699	0.8276	0.9144	0.8596	0.8873
Effit.Net-b5 [47]	0.7504	0.8098	0.9128	0.8522	0.8673
Effit.Net-b6 [47]	0.7662	0.8248	**0.9160**	0.8628	0.8706
Effit.Net-b7 [47]	0.7611	0.8191	0.9137	0.8350	0.8904
Mobl.Netv2 [46]	0.7596	0.8166	0.9105	0.8343	0.8994
VGG16 [48]	0.7619	0.8214	0.9051	**0.8822**	0.8562
VGG19 [48]	0.7588	0.8179	0.9066	0.8560	0.8751

Table 4: Experiment Results on Stroma Region Segmentation with Different Encoder Pathways

SN-FPN +	IoU	F1	ACC	PREC	REC
ResNet50 [37]	0.5452	**0.6219**	0.8533	0.7179	0.7284
ResNet101 [37]	0.5357	0.6109	0.8603	0.7082	0.7281
ResNet152 [37]	**0.5495**	0.6218	**0.8740**	0.6617	**0.8093**
Effit.Net-b0 [47]	0.5435	0.6172	0.8654	0.6790	0.7602
Effit.Net-b1 [47]	0.5369	0.6104	0.8674	0.6768	0.7660
Effit.Net-b2 [47]	0.5318	0.6055	0.8553	0.6888	0.7418
Effit.Net-b3 [47]	0.5360	0.6080	0.8688	0.6683	0.7694
Effit.Net-b4 [47]	0.5462	0.6216	0.8659	0.6809	0.7540
Effit.Net-b5 [47]	0.5340	0.6070	0.8667	0.6599	0.7731
Effit.Net-b6 [47]	0.5479	0.6202	0.8665	0.67460	0.7754
Effit.Net-b7 [47]	0.5361	0.6089	0.8674	0.6854	0.7541
Mobl.Netv2 [46]	0.5388	0.6166	0.8712	0.6922	0.7440
VGG16 [48]	0.5358	0.6155	0.8329	**0.7644**	0.6877
VGG19 [48]	0.5306	0.6021	0.8497	0.6737	0.7640

The images were divided into three datasets for training 1,413(60%), validating 417(20%), and testing 417(20%) respectively.

Baseline and metrics: We use the same encoder pathways: ResNet50 [37], ResNet101 [37], ResNet152 [37], MobileNetv2 [46], EfficientNet-b0, b1, b2, b3, b4, b5, b6, and b7 [47], VGG16 [48], and VGG19 [48] for stroma region segmentaion and the same methods: IoU, $F1$, ACC, $PREC$, and REC will be used as the evlauation metrics.

Results: The experiment results on the comparison of the stroma region segmentation using different encoder pathways are shown in Table 4. The results on the stroma region segmentation indicate that ResNet152 outperforms other pathways in terms of IoU(54.95%), ACC(87.40%), and REC(80.93%) and ResNet50 outperforms other pathways in terms of $F1$(62.19%). VGG16 shows the best results on $PREC$(76.44%). Assuming that the overall performance is measured by $F1$ score demonstrating the balance between the precision and the recall, we can determine that ResNet50 shows better performance on the stroma region segmentation than other encoder pathways in regards to the BCSS dataset.

5 CONCLUSION

In this paper, we presented a new image segmentation model integrating a feature pyramid network with a self-attention mechanism using three pathways: the encoder pathway, the decoder pathway, and the self-attention nested pathway, and providing accurate tumor and stroma region segmentation on whole slide images. ResNet50 architecture was used for the encoder pathway, while the feature pyramid network was used for the decoder pathway. The self-attention nested pathway generates the attention map from the distribution of attention scores focusing on relevant information of the feature maps. The experiment results show that the proposed SN-FPN model outperforms the state-of-the-art image segmentation models when performing tumor and stroma region segmentation on the BCSS dataset. Moreover, we compared various encoder networks by using the proposed SN-FPN model. The results indicate that the ResNet50 outperforms other encoder networks on F1 score for the tumor and stroma region segmentation.

6 ACKNOWLEDGMENTS

This research was partially supported by the National Science Foundation under Grant No. 2138260 and Grant No. 2153063.

REFERENCES

[1] Liron Pantanowitz, Paul N. Valenstein, Andrew J. Evans, Keith J. Kaplan, John D. Pfeifer, David C. Wilbur, Laura C. Collins, and Terence J. Colgan. Review of the current state of whole slide imaging in pathology. *Journal of Pathology Informatics*, 2(1):36, 2011.

[2] Sonal Kothari, John H Phan, Todd H Stokes, and May D Wang. Pathology imaging informatics for quantitative analysis of whole-slide images. *Journal of the American Medical Informatics Association*, 20(6):1099–1108, 08 2013.

[3] David A. Gutman, Mohammed Khalilia, Sanghoon Lee, Michael Nalisnik, Zach Mullen, Jonathan Beezley, Deepak R. Chittajallu, David Manthey, and Lee A.D. Cooper. The Digital Slide Archive: A Software Platform for Management, Integration, and Analysis of Histology for Cancer Research. *Cancer Research*, 77(21):e75–e78, 10 2017.

[4] Muhammad Khalid Khan Niazi, Anil V Parwani, and Metin N Gurcan. Digital pathology and artificial intelligence. *The lancet oncology*, 20(5):e253–e261, 2019.

[5] Navid Farahani, Anil V Parwani, and Liron Pantanowitz. Whole slide imaging in pathology: advantages, limitations, and emerging perspectives. *Pathology and Laboratory Medicine International*, pages 23–33, 2015.

[6] Andrew H Fischer, Kenneth A Jacobson, Jack Rose, and Rolf Zeller. Hematoxylin and eosin staining of tissue and cell sections. *Cold spring harbor protocols*, 2008(5):pdb–prot4986, 2008.

[7] Elena Brachtel and Yukako Yagi. Digital imaging in pathology – current applications and challenges. *Journal of Biophotonics*, 5(4):327–335, 2012.

[8] Neofytos Dimitriou, Ognjen Arandjelović, and Peter D. Caie. Deep learning for whole slide image analysis: An overview. *Frontiers in Medicine*, 6, 2019.

[9] Anant Madabhushi and George Lee. Image analysis and machine learning in digital pathology: Challenges and opportunities. *Medical Image Analysis*, 33:170–175, 2016. 20th anniversary of the Medical Image Analysis journal (MedIA).

[10] Andrew Janowczyk and Anant Madabhushi. Deep learning for digital pathology image analysis: A comprehensive tutorial with selected use cases. *Journal of Pathology Informatics*, 7(1):29, 2016.

[11] Weiming Mi, Junjie Li, Yucheng Guo, Xinyu Ren, Zhiyong Liang, Tao Zhang, and Hao Zou. Deep learning-based multi-class classification of breast digital pathology images. *Cancer Management and Research*, 13:4605–4617, 2021. PMID: 34140807.

[12] Morteza Babaie, Shivam Kalra, Aditya Sriram, Christopher Mitcheltree, Shujin Zhu, Amin Khatami, Shahryar Rahnamayan, and Hamid R. Tizhoosh. Classification and retrieval of digital pathology scans: A new dataset. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, July 2017.

[13] Shidan Wang, Donghan M. Yang, Ruichen Rong, Xiaowei Zhan, and Guanghua Xiao. Pathology image analysis using segmentation deep learning algorithms. *The American Journal of Pathology*, 189(9):1686–1698, 2019.

[14] Fuyong Xing and Lin Yang. Robust nucleus/cell detection and segmentation in digital pathology and microscopy images: A comprehensive review. *IEEE Reviews in Biomedical Engineering*, 9:234–263, 2016.

[15] Siyamalan Manivannan, Wenqi Li, Jianguo Zhang, Emanuele Trucco, and Stephen J. McKenna. Structure prediction for gland segmentation with hand-crafted and deep convolutional features. *IEEE Transactions on Medical Imaging*, 37(1):210–221, 2018.

[16] Wenqi Li, Siyamalan Manivannan, Shazia Akbar, Jianguo Zhang, Emanuele Trucco, and Stephen J. McKenna. Gland segmentation in colon histology images using hand-crafted features and convolutional neural networks. In *2016 IEEE 13th International Symposium on Biomedical Imaging (ISBI)*, pages 1405–1408, 2016.

[17] Shervin Minaee, Yuri Boykov, Fatih Porikli, Antonio Plaza, Nasser Kehtarnavaz, and Demetri Terzopoulos. Image segmentation using deep learning: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(7):3523–3542, 2022.

[18] Matthew Lai. Deep learning for medical image segmentation, 2015.

[19] Farhana Sultana, Abu Sufian, and Paramartha Dutta. Evolution of image segmentation using deep convolutional neural network: A survey. *Knowledge-Based Systems*, 201-202:106062, 2020.

[20] Parul Sharma, Yash Paul Singh Berwal, and Wiqas Ghai. Performance analysis of deep learning cnn models for disease detection in plants using image segmentation. *Information Processing in Agriculture*, 7(4):566–574, 2020.

[21] Dong Zhang, Hanwang Zhang, Jinhui Tang, Meng Wang, Xiansheng Hua, and Qianru Sun. Feature pyramid transformer. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXVIII 16*, pages 323–339. Springer, 2020.

[22] Alexander Kirillov, Ross Girshick, Kaiming He, and Piotr Dollár. Panoptic feature pyramid networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6399–6408, 2019.

[23] Selim Seferbekov, Vladimir Iglovikov, Alexander Buslaev, and Alexey Shvets. Feature pyramid network for multi-class land segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 272–275, 2018.

[24] Tsung-Yi Lin, Piotr Dollar, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[25] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017.

[26] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017.

[27] Zhaowei Cai and Nuno Vasconcelos. Cascade r-cnn: Delving into high quality object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6154–6162, 2018.

[28] Golnaz Ghiasi, Tsung-Yi Lin, and Quoc V Le. Nas-fpn: Learning scalable feature pyramid architecture for object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7036–7045, 2019.

[29] Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. Deep high-resolution representation learning for human pose estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5693–5703, 2019.

[30] Shu Liu, Lu Qi, Haifang Qin, Jianping Shi, and Jiaya Jia. Path aggregation network for instance segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8759–8768, 2018.

[31] Miao Hu, Yali Li, Lu Fang, and Shengjin Wang. A2-fpn: Attention aggregation based feature pyramid network for instance segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15343–15352, June 2021.

[32] Xiang Ying, Qiang Wang, Xuewei Li, Mei Yu, Han Jiang, Jie Gao, Zhiqiang Liu, and Ruiguo Yu. Multi-attention object detection model in remote sensing images based on multi-scale. *IEEE Access*, 7:94508–94519, 2019.

[33] Zhenghua Xu, Tianrun Li, Yunxin Liu, Yuefu Zhan, Junyang Chen, and Thomas Lukasiewicz. Pac-net: Multi-pathway fpn with position attention guided connections and vertex distance iou for 3d medical image detection. *Frontiers in Bioengineering and Biotechnology*, 11:1049555, 2023.

[34] Mohamed Amgad, Habiba Elfandy, Hagar Hussein, Lamees A Atteya, Mai A T Elsebaie, Lamia S Abo Elnasr, Rokia A Sakr, Hazem S E Salem, Ahmed F Ismail, Anas M Saad, Joumana Ahmed, Maha A T Elsebaie, Mustafijur Rahman, Inas A Ruhban, Nada M Elgazar, Yahya Alagha, Mohamed H Osman, Ahmed M Alhusseiny, Mariam M Khalaf, Abo-Alela F Younes, Ali Abdulkarim, Duaa M Younes, Ahmed M Gadallah, Ahmad M Elkashash, Salma Y Fala, Basma M Zaki, Jonathan Beezley, Deepak R Chittajallu, David Manthey, David A Gutman, and Lee A D Cooper. Structured crowdsourcing enables convolutional segmentation of histology images. *Bioinformatics*, 35(18):3461–3467, 02 2019.

[35] Erik Reinhard, Michael Adhikhmin, Bruce Gooch, and Peter Shirley. Color transfer between images. *IEEE Computer graphics and applications*, 21(5):34–41, 2001.

[36] Daniel L Ruderman, Thomas W Cronin, and Chuan-Chin Chiao. Statistics of cone responses to natural images: implications for visual coding. *JOSA A*, 15(8):2036–2045, 1998.

[37] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition, 2015.

[38] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation, 2018.

[39] Zongwei Zhou, Md Mahfuzur Rahman Siddiquee, Nima Tajbakhsh, and Jianming Liang. Unet++: A nested u-net architecture for medical image segmentation, 2018.

[40] Abhishek Chaurasia and Eugenio Culurciello. Linknet: Exploiting encoder representations for efficient semantic segmentation. In *2017 IEEE Visual Communications and Image Processing (VCIP)*. IEEE, December 2017.

[41] Tongle Fan, Guanglei Wang, Yan Li, and Hongrui Wang. Ma-net: A multi-scale attention network for liver and tumor segmentation. *IEEE Access*, 8:179656–179665, 2020.

[42] Hanchao Li, Pengfei Xiong, Jie An, and Lingxue Wang. Pyramid attention network for semantic segmentation, 2018.

[43] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network, 2017.

[44] Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. Rethinking atrous convolution for semantic image segmentation, 2017.

[45] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation, 2015.

[46] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4510–4520, 2018.

[47] Mingxing Tan and Quoc Le. EfficientNet: Rethinking model scaling for convolutional neural networks. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 6105–6114. PMLR, 09–15 Jun 2019.

[48] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.