This is the author's version of an article published in Artificial Intelligence in Medicine. The final authenticated version is available online at: https://doi.org/10.1016/j.artmed.2022.102407.

Weakly Supervised Learning using Attention gates for colon cancer histopathological image segmentation

A. Ben Hamida[1], M. Devanne[2], J. Weber[2], C. Truntzer[3], V. Derangère[3], F. Ghiringhelli[3], G. Forestier[2], C. Wemmert[1]

[1] *ICube, University of Strasbourg, France*
[2] *IRIMAS, University of Haute-Alsace, France*
[3] *Platform of Transform in Biological Oncology, Dijon, France*

Abstract

Recently, *Artificial Intelligence* namely *Deep Learning* methods have revolutionized a wide range of domains and applications. Besides, *Digital Pathology* has so far played a major role in the diagnosis and the prognosis of tumors. However, the characteristics of the *Whole Slide Images* namely the gigapixel size, high resolution and the shortage of richly labeled samples have hindered the efficiency of classical *Machine Learning* methods. That goes without saying that traditional methods are poor in generalization to different tasks and data contents. Regarding the success of *Deep learning* when dealing with Large Scale applications, we have resorted to the use of such models for histopathological image segmentation tasks. First, we review and compare the classical UNet and Att-UNet models for colon cancer WSI segmentation in a sparsely annotated data scenario. Then, we introduce novel enhanced models of the Att-UNet where different schemes are proposed for the *skip connections* and *spatial attention gates* positions in the network. In fact, *spatial attention gates* assist the training process and enable the model to avoid irrelevant feature learning. Alternating the presence of such modules namely in our Alter-AttUNet model adds robustness and ensures better image segmentation results. In order to cope with the lack of richly annotated data in our AiCOLO colon cancer dataset, we suggest the use of a multi-step training strategy that also deals with the WSI sparse annotations and unbalanced class issues. All proposed methods outperform state-of-the-art approaches but Alter-AttUNet generates the best compromise between accurate results and light network. The model achieves 95.88% accuracy with our sparse AiCOLO colon cancer datasets. Finally, to evaluate and validate our proposed architectures we resort to publicly available WSI data: the NCT-CRC-HE-100K, the CRC-5000 and the Warwick colon cancer histopathological dataset. Respective accuracies of 99.65%, 99.73% and 79.03% were reached. A comparison with state-of-art approaches is established to view and compare the key solutions for histopathological image segmentation.

Keywords: Digital pathology, Colon cancer, Weak supervision, Attention gates, Deep learning, Image segmentation

1. Introduction

Image segmentation is a key task of image processing. Over the last few years, its approaches have tremendously evolved and have become a hotspot in the research field. The main purpose of such task is to group similar regions of the image and assign their respective class labels. In fact, image segmentation combines both localization and classification steps. Its applications cover a wide range of domains like computer vision [1], remote sensing [2], medical imaging [3], etc. Actually, the emergence of different medical imaging tools has catalyzed the efforts to enhance the image processing techniques [4]. Medical image segmentation is a crucial step for many other related tasks namely pathology diagnosis, surgical planning and mass detection. Traditionally, the segmentation process relied on the pathologists experience to extract the aimed information such as organs, tissues and nuclei [5]. However, such procedure is both time and effort consuming. Medical images also introduce a high level of complexity compared with natural scene images and other computer vision data. Most of the medical images include many components with high visual resemblance and confusing boundaries [6]. As a matter of fact, the emergence of the digital Whole Slide Images (WSI) has introduced new challenges for image segmentation. Histopathological images are usually gigapixel slides with complex clinical features. They often suffer from a lack of richly annotated reference data for accurate segmentation tasks. Thanks to the advances in Artificial Intelligence (AI) and computational resources, the segmentation of these gigantic slides is possible and serves as a key tool for pathological diagnosis, prognosis, and therapeutic response prediction [7, 8].

Despite the promising potential of Deep Learning (DL) tools, the segmentation of WSIs is still a challenging task. Novel approaches need to cope with the particular traits of histopathological data and ensure accurate tissue seg-

mentation [9].

This paper shines the light on the segmentation of histopathological images. First, a review of the role of AI in WSI image segmentation is presented along with the challenges that hinder its success. Then, a colon cancer WSI segmentation is executed in a weakly supervised scenario. Therefore, UNet model [10] is presented and evaluated on the AiCOLO dataset. Att-UNet [11] is also used for the same task. Novel enhanced versions of the Att-UNet are introduced for segmentation of colon cancer histopathological images. The proposed models are compared with state-of-the-art semantic segmentation models namely FCN8s, FCN16s, FCN32s and DeepLabv3+. Moreover, evaluation of the models is performed with different public datasets (CRC-5000 , NCT-CRC-HE-100K- and Warwick). Finally, the proposed approaches are assessed as regards to state of the art methods in colon cancer digital pathology tasks. The main contributions of the paper are:

- The establishment of a novel multi-step training strategy. This approach enables the model to deal with the lack of richly labeled samples, the sparse annotation of the images and the unbalanced class representation in histopathological data.
- The use of the Att-UNet model for colon cancer histopathological image segmentation.
- The introduction of novel Att-UNet inspired models for better feature learning. The proposed models are low-cost and ensure the focus on the relevant information in the histopathological data.

2. Artificial Intelligence for Medical Image segmentation

2.1. The challenges

The use of AI for histopathological image segmentation is disrupted by many challenges as detailed in [12] namely:

Insufficiency of annotated samples:. Most of the DL models in a digital pathology context require an important amount of good quality, curated and representative training images. Therefore, pathologists need to thoroughly label the WSI and highlight the Regions Of Interest (ROI) based on the targeted application. However, such task is very time consuming and requires a huge share of involvement especially when dealing with large images acquired at different resolutions and staining techniques. Consequently, WSI datasets are often lacking of annotated samples and balanced classes representation. Actually, most the current WSI processing tasks use private data [13]. Therefore, the trained DL architectures are suffering from limited practicality and restrained utility in different applications.

Color Variation and Artifacts:. Histopathological images are a result of a multi-task workflow with many interferes from different fields. Therefore, many undesirable effects may appear at each step of the process. For example, bending and wrinkling of the tissue may generate blurry unsolvable regions. Moreover, color variation could occur during the staining process as a result for the different manufacturers of staining reagents and scanners, thickness and tearing in tissues and every lab staining conditions. The presence of such artifacts in the WSI can mislead the feature learning process and hinder accurate tissue segmentation. In order to cope with such issue, most of the previously proposed approaches apply prior processing or augmentation to the histopathological slides [14, 15].

2.2. Related work

The approaches of WSI segmentation has evolved from hand-crafted to semi-automatic models and recently to the fully automatic segmentation as depicted in [16]. A variety of graph-based methods were developed to segment and highlight targeted tissues in WSI namely as suggested in [17]. However, these methods are highly dependent from a predefined set of features. Thus, their generalization to different applications and datasets is very restricted as detailed in [18]. Therefore, the interest was deflected toward DL models regarding their efficiency in automatic feature extraction. Different methods combined graph-based approaches and DL as detailed in [19]. One of the main trials to use Fully Convolutional networks (FCN) for WSI semantic segmentation was evoked by Long *et al.* in [20]. The proposed model is trained via end-to-end back-propagation to generate a pixel-wise segmentation map. A deep contour-aware network (DCAN) was introduced in [21] based on a multi level FCN for colorectal WSI segmentation. Authors in [22] also introduce a FCN-8s model that combines multi-level localization and feature information for inflammatory colon disease detection in bowel biopsies. Later on, different variants of the FCN were suggested as detailed in [23, 24, 25]. In fact, the use of FCN models for histopathological image segmentation has also been used as a key step for different applications. For instance, the FCN network with a VGG-16 backbone as presented in [26], has inspired authors in [27] and [28] for respective Ovarian cancer bio-markers identification and Thyroid cancer diagnosis. The VGG-16 based FCN was also used to execute foreground segmentation in [29]. The model is combined with an edge detection CNN for multi-channel image segmentation. In fact, a great share of the histopathological image segmentation literature is dedicated to Convolutional neural Networks (CNN) as they ensure accurate feature learning with low computational complexity. The authors in [30] combined the outputs of different CNN models in order to generate a gland segmentation map for the Histology Images Challenge Contest (GlaS) histopathological data [31]. In the same context, Xu et al introduce a mutli-CNN framework for complex multichannel information, location, and

boundary cues fusion. Different CNN architectures were deployed in the context of colorectal cancer namely the VGG-19 [32], 5-layer CNN [33] and a CNN-LSTM dual model [34]. LeNet-5 architecture is also used for the same context in [35]. A hybrid approach was introduced by Qaiser *et al.* in [36] where they combine both CNN extracted features and mathematical feature representations of the training data for accurate segmentation of colon cancer.

These approaches ensure reasonable performance rates for colon WSI segmentation but are computationally expensive and are at high risk of gradient vanishing while training. Therefore, the trials to cope with the high dimensionality of WSI has generated a plethora of CNN-based models. These networks are mainly deeper yet lighter namely the ResNet . Residual models [37] come with the hallmark of reusing the learnt feature for accurate and less expensive learning as detailed in [38]. Thus, The same concept was integrated in a DenseNet architecture to segment digital pathology images in [39]. In this context, the residual blocks are replaced with dense blocks were identity mapping is replaced by dense concatenation connections in order to reinforce the feature re-usage. However, histopathological slides usually encompass different shapes and sizes for the same objects and neighboring tissues which makes it hard to distinguish. As a remedy for such problem, a recent work focuses on the use of encoder-decoder models such as the SegNet and the UNet as depicted in [40]. The UNet model is also used in [41] for stain separation in $H\&E$ images to obtain the H-stain, E-stain, and background stain intensity maps. Colorectal cancer nuclei are then segmented on the H-stain map. A Multi magnification version of the encoder-decoder models is introduced in [42] for multi-class segmentation in WSIs. Authors in [43] combine the classical UNet architecture while inserting residual connections in both blocks to ensure accurate feature learning throughout the entire process. A dense-UNet model was also established in [44] for the same purpose. However, these models rely on a progressively down-sampled feature map grid. This way the model is not assigning any priority to the contextual features and is incapable of reducing false predictions. As a remedy to this issue several papers have established a 2-step procedure where the segmentation and localization modules are independent [45, 46]. To simplify the task, authors in [11] propose the use of the so-called *Soft Attention mechanism*. In the context of image processing, *soft-Attention* refers to the learning process where exclusively relevant information are highlighted. Consequently, the network cuts the computational cost of irrelevant activations and gains more generalisation properties. Soft-attention mechanisms are applied to transfer information between two components of the network (encoder and decoder) unlike self-attention which are usually used at modeling dependencies between different parts of a sequence input. In other words, soft attention of one layer focus on the activation of other layers while self-attention looks for the activation of the same layer where it's applied as detailed in [47]. Therefore, different attention-based models have emerged for medical image segmentation within the last few years as authors in [48] combine spatial and spectral attention gates for MRI, CT and Endoscopy image segmentation. For lumbar MRI image segmentation, a three module framework is presented in [49]. It combines a full feature fusing block followed by a combination of ResNet and attention mechanism. The final unit is a Generative Adversarial Network (GAN). The duality of residual and attention blocks is also introduced in a enhanced efficient UNet model where segmentation of otoscopic images is executed as detailed in [50]. In histology, a similar model was used in [51] where a residual-inception-channel attention-Unet (RIC-Unet) enable accurate nuclei segmentation of few Cancer Genomic Atlas (TCGA) WSIs. Authors in [52] introduce a weakly supervised multi-module framework where a first CNN model is used to detect Regions of interest (ROI) in histopathological images. Then, attention units are inserted in the second CNN model to refine the feature extraction process and enhances the slides classification process. In fact, most of the attention-based models for histopathological image segmentation rely on multi-step hybrid networks as detailed in [53], [54] and [55]. Although they enable good performance rates, such models do not deal with the main problem of WSIs which is the poor sparse annotation of the histopathological slides. Trials to deal with such issue have been presented and discussed in [56], [57] and [58]. However, most of the proposed solutions rely on independent pre-processing modules to enhance the available annotations before training the network. Recently, many review papers present opportunities and challenges of the use of DL for WSI image analysis [59, 60, 13].

3. Proposed methodology

In this section, we propose the use of enhanced versions of the UNet model for colon cancer histopathological image segmentation. First, we rely on the classical UNet then we introduce the use of spatial-attention blocks to enhance the segmentation accuracy.

3.1. Description of The Architectures

UNet

Regarding the variety of features included in each WSI, we resort to the use of skip connections for multi-level feature representation as detailed in [61]. UNet [10] has the advantage of combining high resolution features with high semantic reused ones. Both the contracting and expanding paths are symmetric which ensures accurate learning of not only the content of the image but also its localization. In fact, this network encompasses three main components: an encoder, a bottleneck, and a decoder as seen in Figure 1.

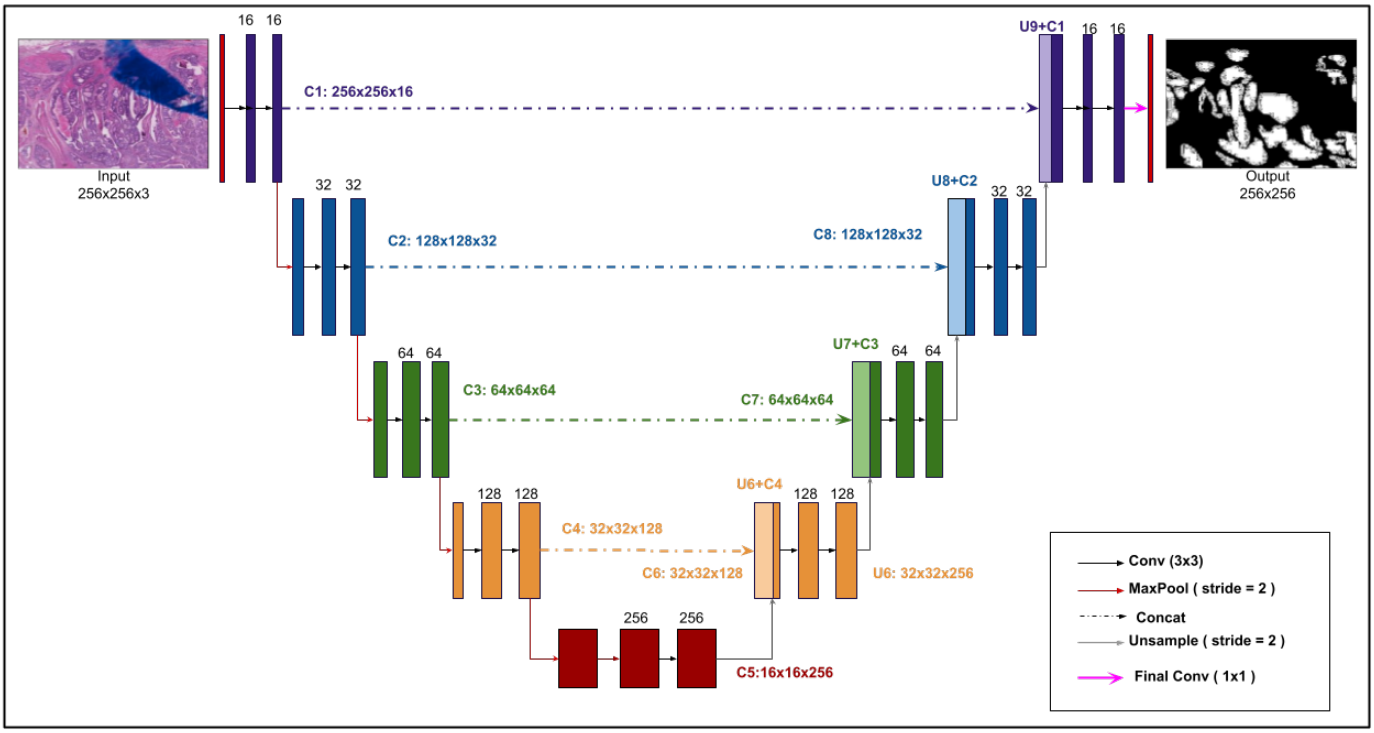


Figure 1: Architecture of the proposed UNet model.

The Encoder. is a classical stack of convolutional layers as seen in the CNN models. It ensures the mapping of the inputs into a feature vector in order to grasp the context presented in the original images. While decreasing the spatial dimensions in every layer and increasing the channels, the different convolutional layers progressively learn the key features. The proposed architecture presents a 4-convolutional block encoder as detailed in Table 1. Each block is a stacking of 2 convolutional layers with 3×3 filters and a 1×1 stride. For each convolutional layer, a ReLU activation function is used along with batch normalization and max-pooling for progressive feature map size reduction.

The bottleneck. links the down-sampling block to the decoding units. It consists of one convolutional block with 2 Conv2d layers with Batch Normalization and ReLu Activation for each. The main purpose behind using bottleneck layers is to create a compressed version of the input that only contains useful information for the reconstruction process.

The Decoder. is the up-sampling path which enables the re-construction of the input image. The purpose behind this procedure is to enable precise localization using deconvolution. It also includes 4 deconvolutional blocks where each block is a stacking of 2 up-sampling layers. In the classical UNet model, transposed convolutions are used with 3×3 filters and a stride equal to 2×2 in order to halve the features map number and double their size. Here, we replace the deconvolutional layers with non-trainable up-sampling filters that execute nearest neighbor interpolation of factor 2. Consequently, we reduce the number of trainable parameters while ensuring smooth image reconstruction. The core of UNet is the use of skip connections to concatenate the input of each deconvolutional block with its corresponding feature map from the contracting path. The final layer is a 1×1 convolution to map the channels to the desired number of classes.

Att-UNet

Here, we suggest the joining of all skip connections in the UNet model with *Spatial Attention Gates* as shown Figure 3. The proposed Att-UNet model encompasses the same *Encoder, Decoder* and *bottleneck* as seen in the UNet architecture. Both the encoding and decoding paths are 4-convolutional blocks with a final 1D-convolutional layer to map the outputted binary mask. Each convolutional block is the succession of 2 convolutional layers with their corresponding Batch Normalization and Activation non-linearity ReLU.

Spatial Attention Gates. Instead of simply concatenating spatial information from the *Encoder* path with the *Decoder* path, attention gates introduce a selective feature learning procedure. In fact, the role of attention gates is to essentially weighting the different regions of the image and assign the largest weights for the most relevant parts. These modules are trainable and are applied to every patch of the image which ensures progressive weights learning and increasing focus on the key areas. We define

	Type	Size	Feature maps : Input	Feature maps : Output	# Param
Encoding Path	Conv Block* 1	$16 \times 3 \times 3$	$3 \times 256 \times 3$	$16 \times 256 \times 256$	2832
	Max Pool 1	–	$16 \times 256 \times 256$	$16 \times 128 \times 128$	0
	Conv Block 2	$32 \times 3 \times 3$	$16 \times 128 \times 128$	$32 \times 128 \times 128$	14016
	Max Pool 2	–	$32 \times 128 \times 128$	$32 \times 64 \times 64$	0
	Conv Block 3	$64 \times 3 \times 3$	$32 \times 64 \times 64$	$64 \times 64 \times 64$	55680
	Max Pool 3	–	$64 \times 64 \times 64$	$64 \times 32 \times 32$	0
	Conv Block 4	$128 \times 3 \times 3$	$64 \times 32 \times 32$	$128 \times 32 \times 32$	369536
	Max Pool 4	–	$128 \times 32 \times 32$	$128 \times 16 \times 16$	0
Bottleneck	Conv Block	$256 \times 3 \times 3$	$128 \times 16 \times 16$	$256 \times 16 \times 16$	886272
Decoding Path	Upsample 1	–	$256 \times 16 \times 16$	$256 \times 32 \times 32$	0
	Conv Block 1	$128 \times 3 \times 3$	$256 \times 32 \times 32$	$128 \times 32 \times 32$	590592
	Upsample Block 2	–	$128 \times 32 \times 32$	$128 \times 64 \times 64$	0
	Conv Block 2	$64 \times 3 \times 3$	$128 \times 64 \times 64$	$64 \times 64 \times 64$	147840
	Upsample Block 3	–	$64 \times 64 \times 64$	$64 \times 128 \times 128$	0
	Conv Block 3	$32 \times 3 \times 3$	$64 \times 128 \times 128$	$32 \times 128 \times 128$	37056
	Upsample Block 4	–	$32 \times 128 \times 128$	$32 \times 256 \times 256$	0
	Conv Block 4	$16 \times 3 \times 3$	$32 \times 256 \times 256$	$16 \times 256 \times 256$	11664
	Final Conv Block	$2 \times 1 \times 1$	$16 \times 256 \times 256$	$2 \times 256 \times 256$	34

Table 1: Layout and number of parameters of the proposed UNet model.* A Conv Block encompasses 2 Conv2d layers with Batch Normalization and ReLU Activation for each.

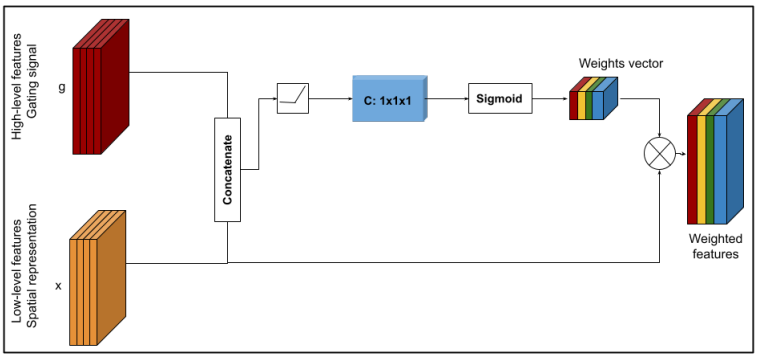


Figure 2: Attention mechanism: inside an attention Gate

two main inputs for each spatial gate: 1) the gating signal g which is the output of the previous low level layer and 2) the vector signal x which is the encoder vector from the same hierarchical level. Simply put, g represents the high level-features since it comes from deeper in the network and x provides the spatial information since it comes from the encoding path. The two elements are brought to the same size and then summed element-wise. Aligned weights are then emphasized while inharmonious weights are penalised. As detailed in Figure 2, the summed vector is fed to a ReLU activation and a 1×1 convolutional layer. A sigmoid layer is added in order to scale the vector into a [0,1] range. The outputted 1D vector holds the attention weights where weights closer to 1 implies more relevant features. This attention vector is then applied to the signal x to generate a weighted feature map which is fed to the Att-UNet convolutional block.

Enhancing Att-UNet

We propose here enhanced models of the Att-UNet architecture through new schemes for both the Attention Gate and Skip Connection positions in the network. The main goal of adding spatial attention gates in the UNet is to enhance the model focus on the crucial features and discard the useless information. However, the problem arises when useful information are judged irrelevant from the first levels and vice-versa. Then, the model has no opportunity to re-adjust its learnt feature maps. To cope with this issue, we introduce novel Att-UNet models as detailed in Table 2:

Alternate Attention in Att-UNet*:* Alter-AttUNet*:*. As seen in Figure 4a, the idea is to eliminate the Attention Gates from certain positions to add features that might be useful and discarded by the previous layer. The base model is the same architecture as described in 3.1. Differently,

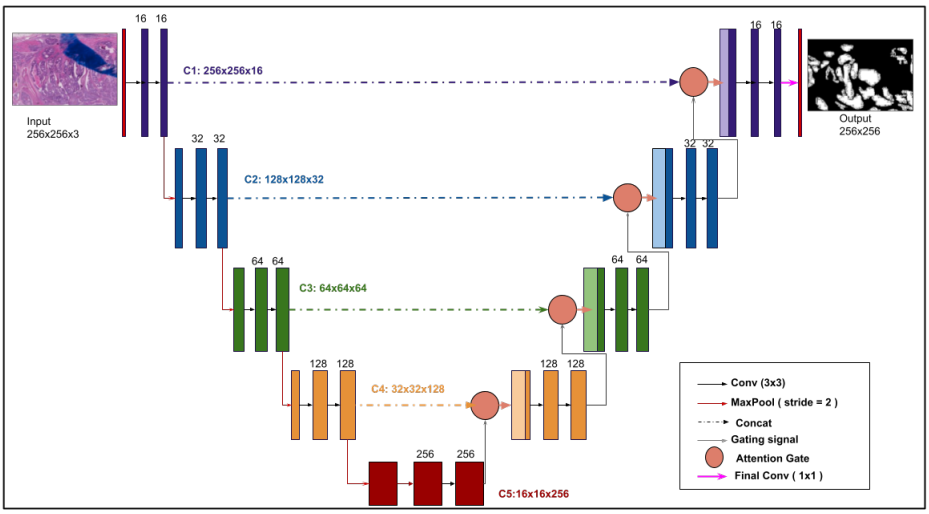


Figure 3: The original Att-UNet Model

Model	Figure	Nbr Attention Gates	position Attention Gates	Nbr Skip Con	position Skip Con
UNet	1	0	[None]	4	[1,2,3,4]
Att-UNet	3	4	[1,2,3,4]	4	[1,2,3,4]
Alter-AttUNet	4a	2	[2,4]	4	[1,2,3,4]
Alter-SkipUNet	4b	2	[2,4]	2	[2,4]
AutoEncoder-AttUNet	4c	2	[3,4]	2	[3,4]
AttUNet-AutoEncoder	4d	2	[1,2]	2	[1,2]

Table 2: Summary of the the Att-UNet, Alter-AttUNet,Alter-SkipUNet, AutoEncoder-AttUNet and AttUNet-AutoEncoder models.

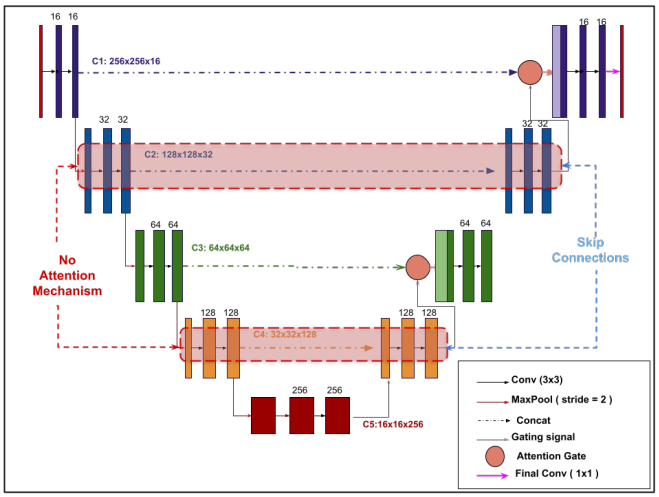
Spatial Attention Gates are inserted in positions 1 and 3 from the Decoding Path while convolutional blocks 2 and 4 are connected simply through skip connections to the Encoder.

Alternate Both Attention and Skip connections in Att-UNet*:* Alter-SkipUNet*:.* Although skip connections are the core of the UNet model, the combination of both Attention gates and skip connections merges low-level features from the encoder with high semantic features from the decoder. The semantic gap between these two feature representation levels could mislead the learning process. Therefore, we propose a novel Alter-SkipUNet model where skip connections are deleted in the absence of Attention Gates. As seen in Figure 4b, the same Alter-AttUNet model is maintained where convolutional blocks of position 1 and 3 take respectively the output of the previous encoding layer as entry.

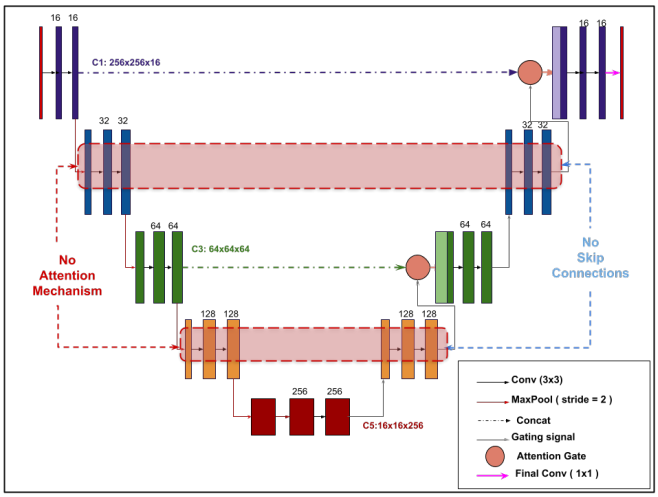
Merging Auto-Encoders and Att-UNet*:* AutoEncoder-AttUNet*and* AttUNet-AutoEncoder*:.* An *auto encoder* learns to capture as much information as possible rather than as much relevant information as possible. Therefore, we combine both the *Spatial Attention* UNet and the classical *auto-encoder* model in novel models: the AutoEncoder-AttUNet and the AttUNet-AutoEncoder which respectively encompass *Attention Gates* in position [1,2] and [3,4] as detailed in Figure 4c and Figure 4d. The other share of the models is a classical encoder/decoder duality with the absence of any skip connections or attention mechanisms. That way, we loosen the control degree over the model and instate it to learn more features.

3.2. Training Strategy: Learning from sparsely annotated WSI

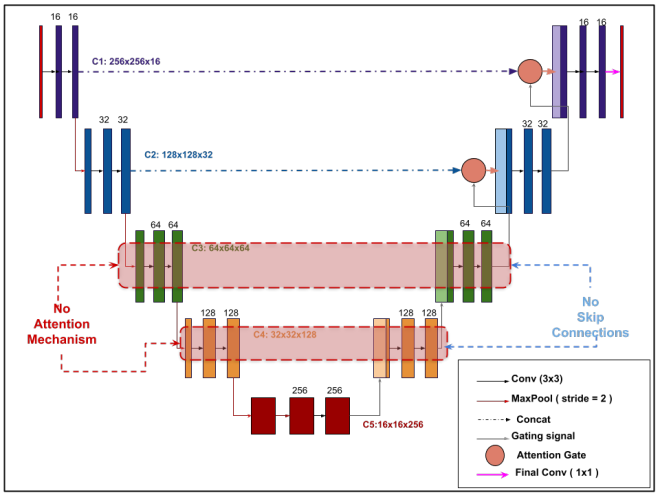
In supervised DL segmentation tasks, an important amount of labeled clean data is usually required to achieve accurate results. However, such condition is hard to accomplish in the pathology field as detailed in the previous section 2.1. Therefore, experts tend to only highlight some specific regions or points in the WSI to simply describe the content of the image. This procedure is described as *"sparse annotation"* where an important amount of the pixels is left unlabeled. Regardless from its rapid annotation, this method generates reference images that lack localization and boundaries information of the classes. Having to deal with the shortage of labeling information, we propose a weakly supervised procedure for training our DL models for WSI segmentation. The 3-steps strategy is described as follows.



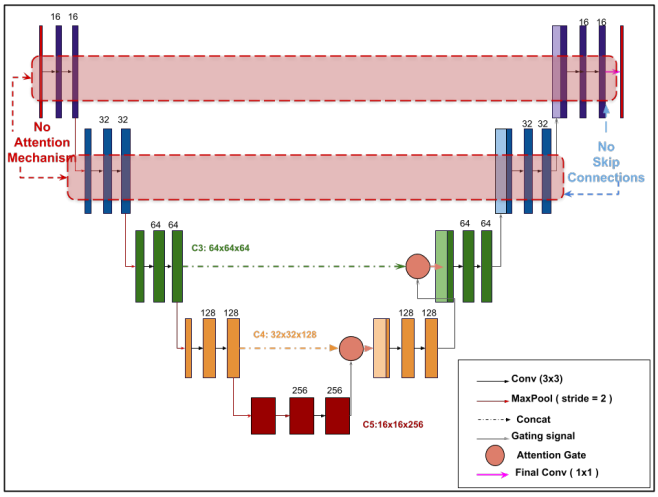
(a) The ALTER-ATTUNET architecture



(b) The ALTER-SKIPUNET architecture



(c) The AUTOENCODER-ATTUNET architecture



(d) The ATTUNET-AUTOENCODER architecture

Figure 4: The proposed enhanced ATT-UNET models. *The highlighted red blocks represent our proposed modified parts from the original ATT-UNET architecture.

Valid patches only:. In order to avoid the influence of unlabeled regions, we choose to select small richly annotated patches. For both the learning and evaluation passes, the model is fed only with patches containing a majority of annotated pixels. The used patches usually encompass one class at a time since the sparsely annotated regions are small and distant. Trying to include different classes in one patch injects lots of doubtful information in the learning process and biases the results. However, this stage isn't capable of compensating the unbalanced class representation.

Weighted cross entropy loss:. The sparse annotation of histopathological data usually implies unbalanced class representation. Since the main focus is on the tumour tissues, the rest of the classes are unequally presented namely stroma, tissues background and fat. The classical way of evaluating a model is using a cross-entropy loss where for each class we assign a true label $\hat{y}$ and a predicted label y. The loss value is calculated as:

$$Loss(y, \hat{y}) = -\sum_{i=1}^{n} \hat{y}_i \log y_i \tag{1}$$

However, the same importance is given to all classes regardless from their presence rate in the data. Therefore, we propose to add a weighted factor w_i that assigns a different value to each class according to its representation in the WSI as detailed in (2).

$$Loss(y, \hat{y}) = -w_i \sum_{i=1}^{n} \hat{y}_i \log y_i, \qquad \sum_{i=1}^{n} w_i = 1 \tag{2}$$

where $Loss(y, \hat{y})$ is the cross entropy loss evaluating the difference between the predicted probability $y_i \in [0, 1]$

and the target label $\hat{y}_i \in [0,1]$ ($w_i \in [0,1]$ is the weight of each class i and n the class number).

Boundary-aware loss:. Histopathological images represent different neighboring tissue types with bulk regions and infrequent edge pixels. In that situation, DL models have tendency to mainly focus on the continuous more presented tissue blocks. As a remedy to this issue, we add a penalty for mistaken border pixel prediction. The proposed approach is inspired from the original UNet model [10] where sophisticated morphological functions are used to generate the edge-aware weights. Here, we propose a less complicated method to re-adjust the feature maps with highlighted edges. A binary morphological dilatation of the border pixels is used where boundaries of tumour pixels are gradually enlarged. Consequently, this regions are more noticeable and less confusing for the learning process. The dilatation is applied for each region centered at an edge pixel (i,j) with value equal to 1 in the binary masks. All neighboring pixels are equally set to 1 to further highlight the boundaries. The new weighted boundary-aware loss function $Loss_{wba}(y, \hat{y})$ is then computed as follows:

$$Loss_{wba}(y, \hat{y}) = {w_e}^{w_i} \sum_{i=1}^{n} \hat{y}_i \log y_i, \tag{3}$$

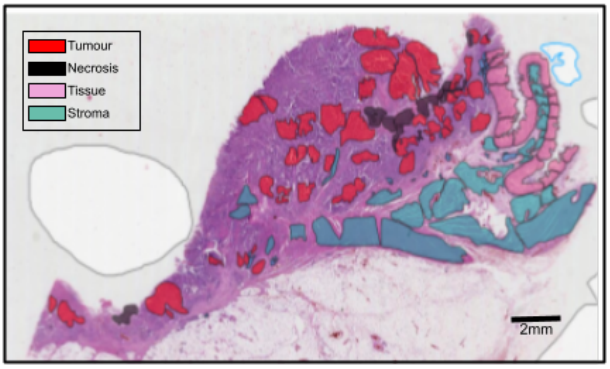
where w_e is the new edge dependent weight map and $w_i \in [0,1]$.

4. Experimental Settings

Publicly available Colon cancer WSI datasets are used to evaluate and compare the proposed Att-UNetmodels with state of the art approaches. Linux operating system is used with an Intel(R) Xeon(R) Bronze 3204 1.9 GHz processors and 62GB RAM. All DL models were implemented using The Pytorch framework. Tests were executed on a Nvidia Quadro RTX 5000 GPU with 16GB memory.

4.1. Data

The* AiCOLO *dataset. includes 396 colon cancer WSIs. All images are stained with Haematoxylin, Eosin, and Natural Saffron. The dataset is created using a Hamamatsu photonics scanner at a 0.454µm/pixel spatial resolution. The number of pixels per slide varies between 4 and 5 billions. A slide sample is shown in Figure 5. Only 15% of the WSIs were sparsely labeled by pathologists from CFGL (Dijon, France).The dataset includes 8 different classes namely tumour, stroma, fat, necrosis, immune, healthy tissue, artifacts and background as seen in Figure 5b. 256 × 256 patches are extracted from the labelled regions using the Cytomine [62] image retrieval tools. The final set of patches is split into Training and Testing sub-sets as seen in Table 3. For binary image segmentation tasks, "Tumour" patches represent the positive class while the rest of the 7 classes-all joint together-represent the negative class of "Normal Tissues". Consequently, the dataset is made of 1454 "Tumour" samples and 3727 "Non-Tumour" patches.



(a) Sparse annotations of the classes of interest

(b) Binary mask for the class "Tumour"

Figure 5: Sample of sparsely annotated WSI/mask from the our dataset.

	Training samples	Testing samples
Tumour	976	478
Necrosis	387	193
Immune	301	150
Stroma	642	320
Fat	75	37
Tissue	477	238
Artifacts	280	139
Background	326	162

Table 3: Number of samples in training and testing sets in our AiCOLO patch-based dataset.

The 100,000 histological images dataset. The NCT-CRC-HE-100K dataset encompasses 86 $H\&E$ stained colon cancer WSIs from both the NCT Biobank (National Center for Tumor Diseases, Heidelberg, Germany) and the UMM pathology archive (University Medical Center Mannheim, Mannheim, Germany). A total of 100.000 224×224 patches

were extracted from the digital slides including 9 classes namely tumours tissues and healthy epithelium regions.

The Colorectal Histology MNIST. The CRC-5000 dataset includes 5000 histopathological images using the Aperio ScanScope scanner at a 20× magnification. The 150 × 150 patches come from the archive of the Institute of Pathology, University Medical Center Mannheim, Heidelberg University,). The dataset represents colon cancer adenocarcinoma along with other 8 normal tissue types.

The GlaS (Gland Segmentation in Colon Histology Images Challenge): Warwick. The WARWICK dataset was first created for the GlaS challenge including T3 and T4 colon tumour adenocarcinoma. The original 16 *H&E* stained histopathological slides are cropped into 825 patches of 150 × 150 pixels.

As detailed in Table 4, AICOLO dataset introduces different challenges when compared with state of the art datasets including NCT-CRC-HE-100K, CRC-5000 and WARWICK. The AICOLO slides suffer from many artifacts namely out of focus regions, tears and cuts in the tissues. Besides, the number of samples per class is completely unbalanced where tissues like Immune are very poorly represented as seen in 3. This issue occurs with AICOLO binary segmentation. Indeed, the class "tumour" represents approximately 2.5× less surface than "non tumour" tissues. In contrast, CRC-5000, NCT-CRC-HE-100K and WARWICK datasets are all composed of a balanced set of tissue types, and their patches are clean (no staining problems or artifacts). Thus, AiCOLO dataset presents a high level of difficulty for training DL models.

4.2. Data augmentation

Data augmentation is helpful to enhance the performance of DL models by providing new and different data samples for the training process. In fact, a rich data is crucial to ensure high accuracy in this context. In the absence of richly annotated datasets in our case of study, we resort to augmentation techniques. Thus, collecting and labeling histopathological images can be exhausting and costly processes as already detailed in previous sections. Transformations in datasets by using data augmentation techniques allow us to reduce these operational costs while creating a wide range of image variations. In order to reproduce the different pathologist perspectives, spatial alterations are applied to each WSI and its respective binary mask namely arbitrary axial flips, center and resized crops and rotations. Furthermore, Each WSI is converted to gray-scale with random brightness, saturation and contract values.

4.3. Evaluation Criteria

Our models have been evaluated and compared with state of the art approaches, using *accuracy, specificity, sensitivity and F1-score*:

$$\textbf{Accuracy} = \frac{TN + TP}{TN + TP + FN + FP}$$

$$\textbf{Specificity} = \frac{TN}{TN + FP}$$

$$\textbf{Sensitivity} = \frac{TP}{TP + FN}$$

$$\textbf{F1-score} = \frac{2\ TP}{2\ TP + FP + FN}$$

where

TP = True Positives: Correctly classified "Tumour" Pixels.

TN = True Negatives: Correctly classified "Non-Tumour" Pixels.

FP = False Positives: "Non-Tumour" pixels that are Miss-classified as "Tumour".

FN = False Negatives: "Tumour" pixels that are Miss-classified as "Non-Tumour".

4.4. Model Parameters

To train the models we only use 256 × 256 patches derived from the AICOLO dataset as detailed in Table 3. For sufficient representation of the data content, we use a batch size of 32 for optimization and weight update. The Stochastic Gradient Descent algorithm is used with an initial learning rate of value 0.0001 and a 0.9 momentum. For more adapted training, the learning rate value is divided by 10 each 25 epochs. All models including the UNET, ATT-UNET, ALTER-ATTUNET, ALTER-SKIPUNET, AUTOENCODER-ATTUNET and ATTUNET-AUTOENCODER are trained for 200 epochs. The ReLu activation function is deployed for all convolutional layers in all models. The different FCN models along with DEEPLABV3+ are trained for 100 epochs. Note that each model is trained 10 times with random train and test splits.

5. Results and Discussion

UNet *vs.* Att-UNet:. We tested two different UNET schemes. The first UNET is detailed in Table 1 where the number of filters ranges from 16 to 256 progressively at each convolutional block. The model ensures ≈ 9% higher accuracy and F1-score than the SEGNET suggested in [63]. The number of trained parameters is also around 3× less important than the SEGNETas seen in Table 5. A heavier UNET architecture is used on our AICOLO dataset. The number of filters vary between 64 to 1024 at each convolutional block. Although this model provides ≈ 3% higher accuracy rates it still introduces a dramatically important computational cost since it trains about 17× more parameters. In order to enhance both the accuracy rate and the cost, we propose the evaluation of the integration of *spatial attention gates* in the light UNET model. In

Data	#images	Image size	Annotation-type	Balanced classes	Pre-processed	Artifacts
AiCOLO	5181	256×256	Sparse	No	None	Yes
NCT-CRC-HE-100K	100,000	224×224	Dense	Yes	Yes	No
CRC-5000	5000	150×150	Dense	Yes	Yes	No
Warwick	825	150×150	Dense	Yes	Yes	No

Table 4: Comparison between the AiCOLO,NCT-CRC-HE-100K, CRC-5000 and Warwick datasets.

other words, we use a low number of filters and integrate the attention mechanism to focus on the relevant regions only. As seen in Table 5, the Att-UNet model has an accuracy rate of 95.02% and a F1-score of 93.28%. As shown in Figure 7, the integration of *attention gates* enhances the model performances of the segmentation on the colorectal cancer AiCOLO slides. The Att-UNet also guarantees accurate results with a light model since it only trains 2.18M parameters while SegNet trains 7.6M and the heavy UNet generates 34.53M parameters.

Although the three models rely on the encoder/decoder duality, they still introduce different feature learning strategies. The SegNet uses multi-level convolutional filters along with trainable weights in the pooling layers to ensure multi-scale semantic data learning. However, the only connection between the encoder and the decoder blocks are the pooling layer learnt weights which alone are insufficient to have a thorough insight into the data content in the decoding process. The classical UNet model as presented in Table 1 comes with the hallmark of using skip connections to link the encoder and decoder blocks unlike the SegNet model. The learning strategy relies on the merge of low-level features from the encoder with high deep features from the decoder. Despite its ability to enhance the segmentation precision as detailed in Table 5, the combination of features from different semantic levels can bias the learning process. Simply put, the semantic gap between the encoder and decoder features generates incompatible sets of features to learn from and misleads the focus of the convolutional filters in the decoding path. Consequently, the use of Att-UNet guarantees the link between the encoding and decoding path unlike the SegNet model while compensating the semantic gap between the fused features generated by the classical UNet skip connections. Actually, the *attention mechanism* as seen in Figure 2 uses trainable weights. As a result, the spatial attention filters are updated to make the model progressively focuses on the relevant regions. The Att-UNet model is then capable of generating more precise tumour segmentation in the AiCOLO slides as show in Figure 6 where less false negatives are introduced.

***Enhancing* Att-UNet**. As detailed above, the introduction of *spatial attention gates* enhances the segmentation results on the AiCOLO dataset. Figure 6 shows an example of an AiCOLO slide segmentation results where one can clearly notice that although the Att-UNet is capable of successfully tracing the tumor tissues it still suffers from the presence of false positives. This result is reflected in Table 6 and in Figure 7 where a specificity of 93.15% is reached versus a sensitivity of 96.06%. In other words, the Att-UNet model is able to detect positive tumor tissues in 96.06% of cases but still confuses 7% of the negative pixels with tumor. As a remedy to this issue, we propose the different schemes of the *attention based models.* As seen in Table 6, the Alter-AttUNet model ensures not only higher accuracy and F1 rates but especially a 3% higher specificity which indicates a better tumor segmentation and less false positives in the resulting mask. The deletion of skip connections in the absence of attention gates also guarantee better performance than the Att-UNet as detailed in Figures 7a, 7b, 7c and 7d. As detailed in Table 6, three different architectures are proposed where the skip connections and the attention gates are deleted in different positions in the network. First, when inserting skip connections and attention gates only in the first levels of the model (positions 1 and 2 of the decoder), the AttUNet-AutoEncoder model simulates a combination of an Att-UNet followed by a classical Autoencoder. Then, we propose a similar yet reversed model were skip connections and attention gates come in the final levels of the model (positions 3 and 4). Both AttUNet-AutoEncoder and AutoEncoder-AttUNet models generate similar performances where the accuracy rates are $\approx 13\%$ greater than SegNet[63] segmentation results. However, these models still introduce relatively a more important number of false negatives which is obvious with the respective 90.36% and 92.08% sensitivity rates for both the AttUNet-AutoEncoder and the AutoEncoder-AttUNet. Finally, the Alter-SkipUNet is presented where skip connections and attention gates come in alternated positions (positions 2 and 4 of the decoder). This model ensures very close accuracy and F1 rates to the Alter-AttUNetwith similar number of parameters $= 2.18M$.

As seen in Figures 6 and 7, the introduction of attention gates in the learning process can fill in the semantic gap between the encoder and decode features. However, inserting attention gates in all positions of the model as seen in Figure 3 can mislead the learning process. In fact, if the model judges a region as irrelevant in some stage of the learning process, it will eventually be discarded for the rest of the procedure. Therefore, as seen in the performance results alternating between attention gates and

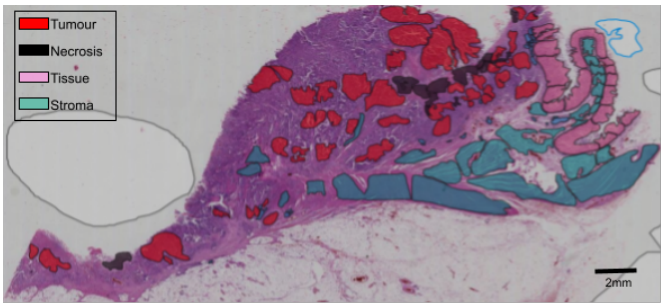
Model	#filters	Accuracy	Specificity	Sensitivity	F1-score	#params
UNet	[64 to 1024]	92.52 ± 0.06	91.85 ± 0.04	92.79 ± 0.06	89.47 ± 0.02	34.53M
UNet	[16 to 256]	89.87 ± 0.08	90.82 ± 0.06	89.17 ± 0.06	87.48 ± 0.05	2.23M
Att-UNet	[16 to 256]	$\mathbf{95.02} \pm 0.04$	$\mathbf{93.15} \pm 0.07$	$\mathbf{96.06} \pm 0.04$	$\mathbf{93.28} \pm 0.06$	**2.18M**
SegNet[63]	[16 to 256]	81.22 ± 0.02	80.70 ± 0.06	81.40 ± 0.02	$75.53 \pm 0.03.$	7.6M

Table 5: Accuracy, specificity, sensitivity rates and F1-score (in %) for UNet, Att-UNet and SegNet.

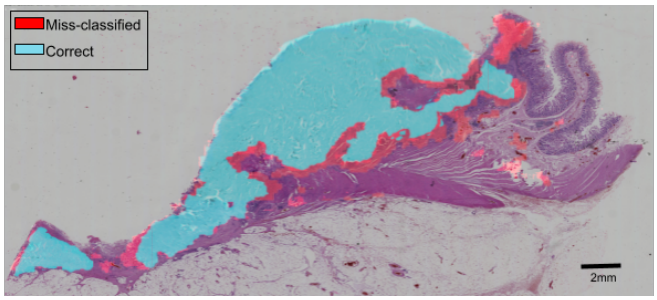
simple skip connections guarantee the link between the encoder and the decoder to extract positions of the pixels while using the attention mechanism to re-adjust the relevant regions to learn from. Although the elimination of skip connections in the absence of attention gates can ensure decent results, it still is problematic when dealing with WSI. Limitations of Alter-SkipUNet, AutoEncoder-AttUNet and AttUNet-AutoEncoder models come from the gap between "highly framed learning" to "free learning". Simply put, layers that encompass both skip connections and attention gates provide information about "where" and "what" to look for into the data. Layers where no encoder/decoder links are included forces the model to reconstruct the data with no prior knowledge about the position and the content of the features. Therefore, it is better performing than Att-UNet but slightly less accurate than Alter-AttUNet where skip connections are present in all levels of the model.

***Enhanced* Att-UNet *VS* FCNs *and* DeepLabv3+.** In order to evaluate our proposed model, we compare it with widely used deep learning-based semantic segmentation models namely FCN and DeepLabv3+ networks. We rely on the FCN model that was first introduced in [26] and lately used in [27, 28]. The architecture uses a padding layer, VGG-16 as a backbone followed by deconvolutional and cropping layers. FCN32s up-samples the output with no prior spatial information. FCN16s and FCN8s fuse the final output with up-sampled outputs from encoding layers using element-wise addition as detailed in [26]. DeepLabv3+ is another encoder-decoder CNN based model [64]. The highlights of this model are the use of dilated convolutions combined with atrous spatial pyramid pooling (ASPP) to encode multi-scale contextual information [65].
As seen in Table 6 and Figure 7, the proposed Att-UNet models ensure the highest performance rates when trained with the AiCOLO dataset. In the absence of spatial information, FCN32s generates rough output maps that lack accuracy (84.98%) and sensitivity (73.89%). Although FCN16s and FCN8s add more spatial information to enhance the results, they still generate $< 90\%$ accuracy rates. Besides, these models suffer from low true positive rates ($< 80\%$) sensitivity compared with our enhanced Att-UNet schemes($> 90\%$). In fact, the different FCN models only rely on classical up-sampling layers with no trainable filters which results in loosing spatial information when going deeper. This issue is solved by adding both Attention Gates and Skip Connections in certain positions as detailed in our introduced models and shown in Figure 6. In the same context, DeepLabv3+ uses dilated separable convolutions to learn the spatial resolution of the outputted feature maps. Therefore, the model extracts dense feature maps that covers spatial information at multiple scales. However, histopathological images usually encompass low-level features with spatially limited regions like seen in our AiCOLO dataset. Models like DeepLabv3+ introduce a high level of complexity that doesn't fit with the colon cancer segmentation tasks. Therefore, the proposed Att-UNet models outperform DeepLabv3+ with $> 8\%$ accuracy and sensitivity. That goes without saying, that the enhanced Att-UNet networks ensure not only high performances rates but also low computational costs. As a matter a fact, the proposed models train $\approx 8\times$ less parameters than the FCN architectures and $\approx 27\times$ less than the DeepLabv3+.

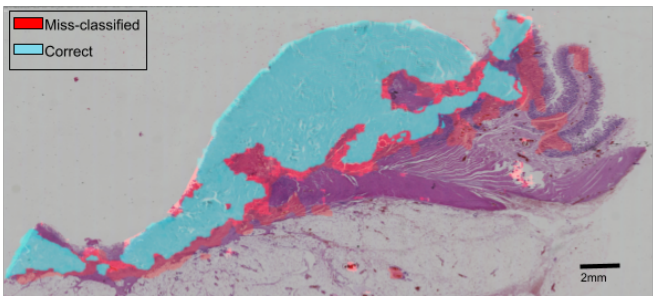
Comparison with state of the art methods. In order to evaluate and compare our proposed models with state-of-the-art methods, we used three different publicly available histopathological data including the CRC-5000, NCT-CRC-HE-100K and the Warwick datasets. As seen in Table 7, our proposed enhanced model Alter-AttUNet ensures the best performance among all models when trained with the three different datasets. First, when using the CRC-5000 images, the Alter-AttUNet reaches an overall performance rates $> 99\%$ including accuracy, specificity, sensitivity and F1 scores. Actually, the model super-pass the approaches proposed in [63] and [66] where the authors introduce a combination of different texture filters for binary WSI segmentation. Then, we use the NCT-CRC-HE-100K data to train and evaluate our models. Here again our Alter-AttUNet achieves the best segmentation accuracy compared with the Ensemble DNN proposed by the authors in [67]. In fact, the methods relies on an Ensemble Deep Neural network composed of DenseNet-121, InceptionResNetV2, Xception and a custom feed forward CNN. Despite its $> 98\%$ performance rates, the approach in [67] introduces a complex model for automatic data learning which limits it generalization properties. Finally, we deploy the Warwick dataset where the GlaS MICCAI 2015 challenge winners suggested in



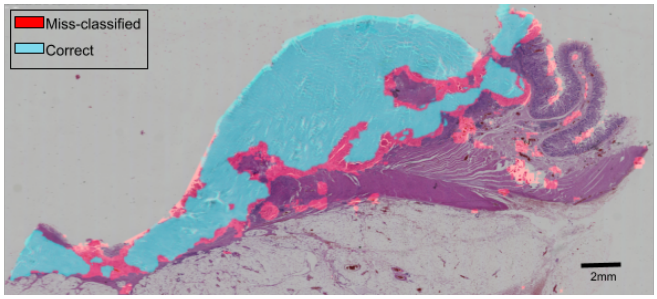
(a) Sparse annotated WSI



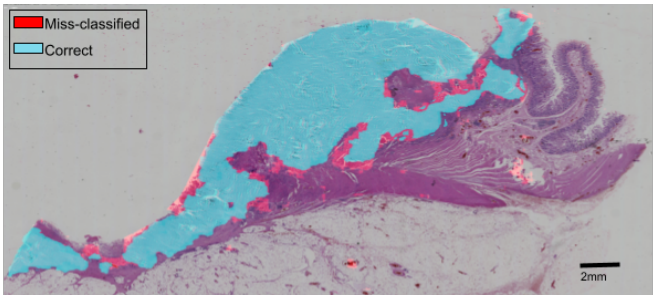
(b) FCN8s segmentation result



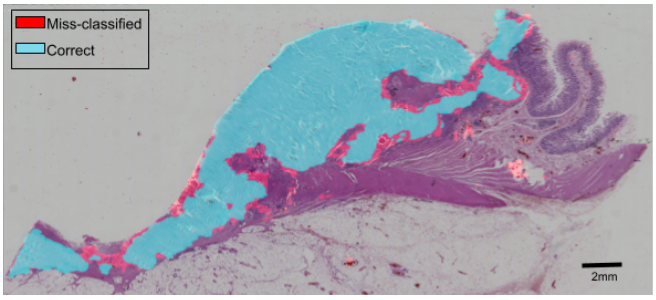
(c) DeepLabv3+ segmentation result



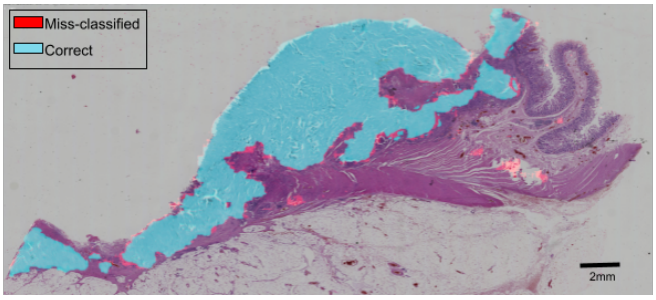
(d) UNet segmentation result



(e) Att-UNet segmentation result



(f) AttUNet-AutoEncoder segmentation result



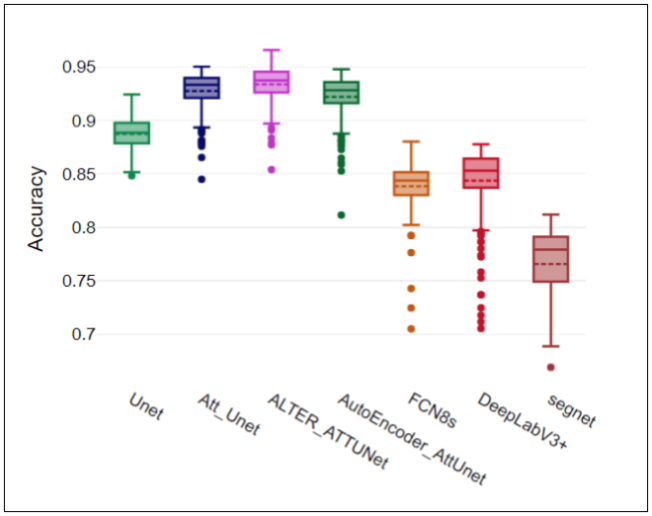
(g) Alter-AttUNet segmentation result

Figure 6:
Segmentation maps of the FCN8s, DeepLabv3+,UNet, Att-UNet, AutoEncoder-AttUNet and Alter-AttUNet models.

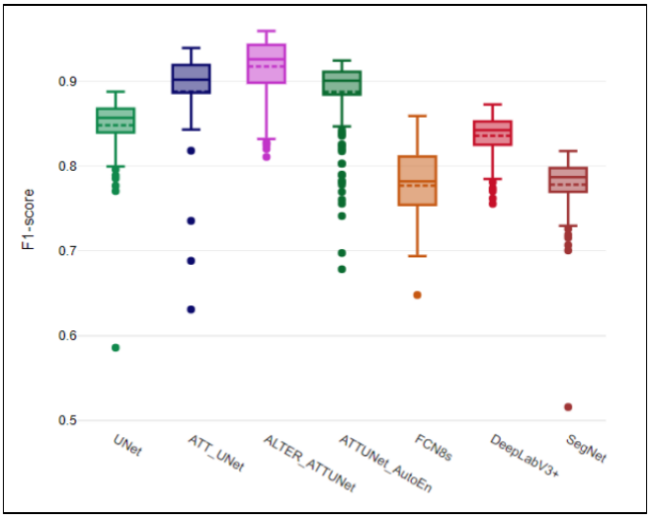
[35], a multi-level CNN model. The architecture uses a first CNN as a classifier to highlight the glands from the background and then a second CNN is used for gland segmentation based on weighted total variation. The outputted result is then the regularization of the CNNs predictions. The hallmark of this approach is a high sensitivity of 73%. CUMedVision2 is a deep contour-aware network that generates multi-level feature representations using an FCN model The architecture achieved an F1-score of 76.9% as presented in the Glas challenge [24]. However, our proposed Alter-AttUNet ensures high accuracy and F1 rates $> 78\%$ while increasing the sensitivity

Model	Figure	pos Attention Gates	pos Skip Con	Accuracy	Specificity	Sensitivity	F1-score	#params
Att-UNet	3	[1,2,3,4]	[1,2,3,4]	95.02 ± 0.04	93.15 ± 0.07	95.06 ± 0.04	93.28 ± 0.06	2.185M
Alter-AttUNet	4a	[2,4]	[1,2,3,4]	**95.88** ± 0.03	**96.12** ± 0.04	**95.05** ± 0.03	**95.18** ± 0.02	**2.180M**
Alter-SkipUNet	4b	[2,4]	[2,4]	95.73 ± 0.06	96.00 ± 0.04	95.06 ± 0.06	94.78 ± 0.04	2.180M
AutoEncoder-AttUNet	4c	[3,4]	[3,4]	94.98 ± 0.07	95.84 ± 0.06	92.1 ± 0.07	92.08 ± 0.08	2.183M
AttUNet-AutoEncoder	4d	[1,2]	[1,2]	94.44 ± 0.08	96.11 ± 0.08	90.36 ± 0.06	92.23 ± 0.07	2.163M
SegNet[63]	-	-	-	81.22 ± 0.02	80.70 ± 0.06	81.40 ± 0.02	75.53 ± 0.03.	7.6M
FCN8s	-	-	-	88.05 ± 0.09	95.23 ± 0.02	78.37 ± 0.06	85.98 ± 0.05	18.6M
FCN16s	-	-	-	85.23 ± 0.09	91.38 ± 0.03	74.13 ± 0.07	81.85 ± 0.04	18.6M
FCN32s	-	-	-	84.98 ± 0.08	89.17 ± 0.03	73.89 ± 0.05	80.81 ± 0.04	18.6M
DeepLabv3+	-	-	-	87.53 ± 0.06	88.13 ± 0.03	87.02 ± 0.08	87.57 ± 0.02	59.3M

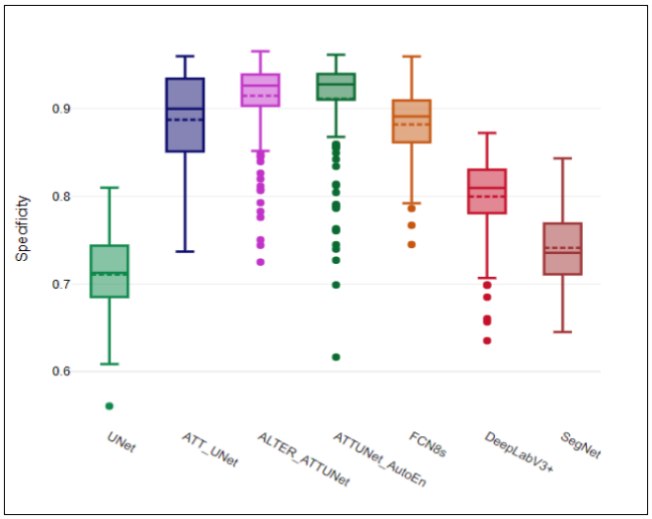
Table 6: Accuracy, specificity, sensitivity rates and F1-score (in %) for Att-UNet, Alter-AttUNet,Alter-SkipUNet, AutoEncoder-AttUNet, AttUNet-AutoEncoder, FCN8s, FCN16s, FCN32s and DeepLabv3+.



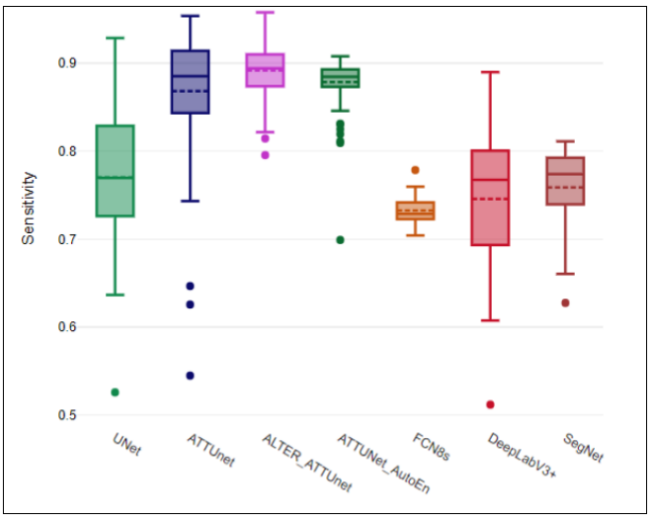
(a) Accuracy



(b) F1-Score



(c) Specificity



(d) Sensitivity

Figure 7: Statistical performance analysis of the UNet, Att-UNet, Alter-AttUNet, AutoEncoder-AttUNet, FCN8s, DeepLabv3+ and SegNet models.

to $> 82\%$. Alter-AttUNet architectures are not only lighter than available techniques but can also cope with different histopathological datasets in different contexts.

Training Strategy for weakly supervised Learning. As detailed above, we rely on a three-step training strategy to cope with the sparse annotation of the AiCOLO colon cancer dataset. In order to evaluate the impact of such procedure on the segmentation results, we have executed different tests. We resorted to the classical training process where all patches contribute in the same way and no special interest is dedicated to boundaries. For all models including SegNet, UNet, Att-UNetand the different Alter-AttUNet versions, the models are incapable of correctly learning and classifying the "Tumour" pixels in all AiCOLO histopathological slides. In fact, in the absence of the first step the models are fed with random patches that could include too small sparse annotated regions and mislead the learning process. Tests show that accuracy rates collapse to under $\approx 40\%$ when patches aren't precisely selected while rises up to $\approx 90\%$ when only valid patches are used for the training. Moreover, both the weighted and boundary aware losses enable the enhancement of the accuracy rates of all models. The Alter-AttUNet model for example witnesses a 5% better accuracy and F1-scores when trained with weighted boundary aware loss.

6. Conclusion

In this paper, we have proposed the use of novel enhanced models inspired from the Att-UNet. First, we compared and highlighted the role of *spatial attention gates* in enhancing feature learning from histopathological colorectal data. Then, we introduced different schemes of attention-based UNet models. The Att-UNet, Alter-AttUNet, Alter-SkipUNet, AutoEncoder-AttUNet and AttUNet-AutoEncoder architectures all perform well in an AiCOLO colon cancer WSI segmentation task. The models outperform state-of-the-art semantic segmentation models namely FCN8s, FCN16s, FCN32s and DeepLabv3+. The enhanced Att-UNet models enable the simultaneous feature learning and spatial localization at different hierarchical levels. The hallmark of such networks is the ability to focus on relevant information without exploding the computational cost. Furthermore, the Alter-AttUNet proposed model outperform state-of-the-art methods when dealing with publicly available datasets namely the NCT-CRC-HE-100K, CRC-5000 and Warwick.

Introducing a new pattern for Att-UNet represents an appealing solution for histopathological image segmentation. Delving into the details of the model one can easily notice that a $> 99\%$ rate has been reached for accuracy, sensitivity, specificity and F1-score when processing NCT-CRC-HE-100K and CRC-5000 datasets. Less rich WSI collections like Warwick and AiCOLO suffer from lower performance rates. As a matter of fact, the lack of richly annotated data and a balanced class representation hinder the efficiency of Alter-AttUNet. The incorporation of a special training strategy is capable of enhancing the segmentation results to a certain extent. However, datasets that encompass a low number of annotated samples with an important share of biased, unbalanced and full of artifacts are one of the main obstacle toward accurate feature learning and successfully accomplished WSI segmentation tasks.

Acknowledgment

This work was supported by the AiCOLO project (ASC-19050MSA) funded by INSERM/Plan Cancer.

References

[1] Minaee, S., Boykov, Y.Y., Porikli, F., Plaza, A.J., Kehtarnavaz, N., Terzopoulos, D.. Image segmentation using deep learning: A survey. IEEE Transactions on Pattern Analysis and Machine Intelligence 2021;.

[2] Yuan, X., Shi, J., Gu, L.. A review of deep learning methods for semantic segmentation of remote sensing imagery. Expert Systems with Applications 2020;:114417.

[3] Liu, X., Song, L., Liu, S., Zhang, Y.. A review of deep-learning-based medical image segmentation methods. Sustainability 2021;13(3):1224.

[4] Nogales, A., García-Tejedor, Á.J., Monge, D., Vara, J.S., Antón, C.. A survey of deep learning models in medical therapeutic areas. Artificial Intelligence in Medicine 2021;112:102020.

[5] Withey, D.J., Koles, Z.J.. Medical image segmentation: Methods and software. In: 2007 Joint Meeting of the 6th International Symposium on Noninvasive Functional Source Imaging of the Brain and Heart and the International Conference on Functional Biomedical Imaging. IEEE; 2007, p. 140–143.

[6] Zheng, H., Zhang, Y., Yang, L., Wang, C., Chen, D.Z.. An annotation sparsification strategy for 3d medical image segmentation via representative selection and self-training. In: Proceedings of the AAAI Conference on Artificial Intelligence; vol. 34:04. 2020, p. 6925–6932.

[7] Zeiser, F.A., da Costa, C.A., de Oliveira Ramos, G., Bohn, H.C., Santos, I., Roehe, A.V.. Deepbatch: A hybrid deep learning model for interpretable diagnosis of breast cancer in whole-slide images. Expert Systems with Applications 2021;185:115586.

[8] Yao, J., Zhu, X., Jonnagaddala, J., Hawkins, N., Huang, J.. Whole slide images based cancer survival prediction using attention guided deep multiple instance learning networks. Medical Image Analysis 2020;65:101789.

[9] Janowczyk, A., Madabhushi, A.. Deep learning for digital pathology image analysis: A comprehensive tutorial with selected use cases. Journal of pathology informatics 2016;7.

[10] Ronneberger, O., Fischer, P., Brox, T.. U-net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical image computing and computer-assisted intervention. Springer; 2015, p. 234–241.

[11] Schlemper, J., Oktay, O., Schaap, M., Heinrich, M., Kainz, B., Glocker, B., et al. Attention gated networks: Learning to leverage salient regions in medical images. Medical image analysis 2019;53:197–207.

[12] Tizhoosh, H.R., Pantanowitz, L.. Artificial intelligence and digital pathology: challenges and opportunities. Journal of pathology informatics 2018;9.

[13] Pacal, I., Karaboga, D., Basturk, A., Akay, B., Nalbantoglu, U.. A comprehensive review of deep learning in colon cancer. Computers in Biology and Medicine 2020;:104003.

	CRC-5000				NCT-CRC-HE-100K				Warwick			
	Acc.	Spec.	Sens.	F1.	Acc.	Spec.	Sens.	F1.	Acc.	Spec.	Sens.	F1.
UNet	98.82 ± 0.18	99.08 ± 0.09	98.32 ± 0.08	98.12 ± 0.12	99.23 ± 0.09	99.54 ± 0.04	99.07 ± 0.10	99.28 ± 0.09	78.81 ± 0.09	76.73 ± 0.13	81.87 ± 0.09	75.03 ± 0.09
Att-UNet	99.06 ± 0.07	98.67 ± 0.07	97.97 ± 0.06	96.85 ± 0.10	99.56 ± 0.03	99.58 ± 0.01	99.09 ± 0.09	99.23 ± 0.07	78.92 ± 0.09	76.69 ± 0.10	82.02 ± 0.08	78.07 ± 0.06
Alter-AttUNet	$\mathbf{99.65} \pm 0.09$	$\mathbf{99.78} \pm 0.07$	$\mathbf{99.02} \pm 0.08$	$\mathbf{99.01} \pm 0.07$	$\mathbf{99.73} \pm 0.03$	$\mathbf{99.61} \pm 0.04$	$\mathbf{99.23} \pm 0.02$	$\mathbf{99.31} \pm 0.02$	$\mathbf{79.03} \pm 0.07$	$\mathbf{76.84} \pm 0.06$	$\mathbf{82.13} \pm 0.07$	$\mathbf{78.31} \pm 0.04$
SegNet[63]	98.66 ± 0.08	99.02 ± 0.12	98.14 ± 0.08	98.38 ± 0.04	99.12 ± 0.08	99.56 ± 0.07	98.36 ± 0.13	98.73 ± 0.04	78.39 ± 0.24	76.09 ± 0.16	81.93 ± 0.08	74.48 ± 0.05
Texture Analysis[66]	98.60	-	-		-	-	-		-	-	-	-
Ensemble DNN[67]	-	-	-	-	96.16	-	-	-	-	-	-	-
CNN[35]	-	-	-	-	-	-	-	-	-	57.00	73.00	61.0
CUMedVision1 [24]	-	-	-	-	-	-	-	-	-	-	-	76.90

Table 7: Accuracy rates, specificity, sensitivity and F1-score (in %) for Att-UNet, Alter-AttUNet,Alter-SkipUNet, AutoEncoder-AttUNet and AttUNet-AutoEncoder with NCT-CRC-HE-100K, CRC-5000 Warwick datasets

[14] Magee, D., Treanor, D., Crellin, D., Shires, M., Smith, K., Mohee, K., et al. Colour normalisation in digital histopathology images. In: Proc Optical Tissue Image analysis in Microscopy, Histopathology and Endoscopy (MICCAI Workshop); vol. 100. Citeseer; 2009, p. 100–111.

[15] Brieu, N., Gavriel, C.G., Harrison, D.J., Caie, P.D., Schmidt, G.. Context-based interpolation of coarse deep learning prediction maps for the segmentation of fine structures in immunofluorescence images. In: Medical Imaging 2018: Digital Pathology; vol. 10581. International Society for Optics and Photonics; 2018, p. 105810P.

[16] Srinidhi, C.L., Ciga, O., Martel, A.L.. Deep neural network models for computational histopathology: A survey. Medical Image Analysis 2021;67:101813.

[17] Gunduz-Demir, C., Kandemir, M., Tosun, A.B., Sokmensuer, C.. Automatic segmentation of colon glands using object-graphs. Medical image analysis 2010;14(1):1–12.

[18] Janowczyk, A., Madabhushi, A.. Deep learning for digital pathology image analysis: A comprehensive tutorial with selected use cases. Journal of pathology informatics 2016;7.

[19] Ahmedt-Aristizabal, D., Armin, M.A., Denman, S., Fookes, C., Petersson, L.. A survey on graph-based deep learning for computational histopathology. arXiv preprint arXiv:210700272 2021;.

[20] Long, J., Shelhamer, E., Darrell, T.. Fully convolutional networks for semantic segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. 2015, p. 3431–3440.

[21] Chen, H., Qi, X., Yu, L., Heng, P.A.. Dcan: deep contour-aware networks for accurate gland segmentation. In: Proceedings of the IEEE conference on Computer Vision and Pattern Recognition. 2016, p. 2487–2496.

[22] Ma, Z., Swiderska-Chadaj, Z., Salemi, H., McGovern, D., Knudsen, B., Gertych, A., et al. Semantic segmentation of colon glands in inflammatory bowel disease biopsies. In: International Conference on Information Technologies in Biomedicine. Springer; 2018, p. 379–392.

[23] Kumar, N., Verma, R., Anand, D., Zhou, Y., Onder, O.F., Tsougenis, E., et al. A multi-organ nucleus segmentation challenge. IEEE transactions on medical imaging 2019;39(5):1380–1391.

[24] Sirinukunwattana, K., Pluim, J.P., Chen, H., Qi, X., Heng, P.A., Guo, Y.B., et al. Gland segmentation in colon histology images: The glas challenge contest. Medical image analysis 2017;35:489–502.

[25] Seth, N., Akbar, S., Nofech-Mozes, S., Salama, S., Martel, A.L.. Automated segmentation of dcis in whole slide images. In: European Congress on Digital Pathology. Springer; 2019, p. 67–74.

[26] Long, J., Shelhamer, E., Darrell, T.. Fully convolutional networks for semantic segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. 2015, p. 3431–3440.

[27] Wang, C.W., Lee, Y.C., Chang, C.C., Lin, Y.J., Liou, Y.A., Hsu, P.C., et al. A weakly supervised deep learning method for guiding ovarian cancer treatment and identifying an effective biomarker. Cancers 2022;14(7):1651.

[28] Lin, Y.J., Chao, T.K., Khalil, M.A., Lee, Y.C., Hong, D.Z., Wu, J.J., et al. Deep learning fast screening approach on cytological whole slides for thyroid cancer diagnosis. Cancers 2021;13(15):3891.

[29] Xu, Y., Li, Y., Wang, Y., Liu, M., Fan, Y., Lai, M., et al. Gland instance segmentation using deep multichannel neural networks. IEEE Transactions on Biomedical Engineering 2017;64(12):2901–2912.

[30] Li, W., Manivannan, S., Akbar, S., Zhang, J., Trucco, E., McKenna, S.J.. Gland segmentation in colon histology images using hand-crafted features and convolutional neural networks. In: 2016 IEEE 13th international symposium on biomedical imaging (ISBI). IEEE; 2016, p. 1405–1408.

[31] Sirinukunwattana, K., Pluim, J.P., Chen, H., Qi, X., Heng, P.A., Guo, Y.B., et al. Gland segmentation in colon histology images: The glas challenge contest. Medical image analysis 2017;35:489–502.

[32] Kather, J.N., Pearson, A.T., Halama, N., Jäger, D., Krause, J., Loosen, S.H., et al. Deep learning can predict microsatellite instability directly from histology in gastrointestinal cancer. Nature medicine 2019;25(7):1054–1056.

[33] Sena, P., Fioresi, R., Faglioni, F., Losi, L., Faglioni, G., Roncucci, L.. Deep learning techniques for detecting preneoplastic and neoplastic lesions in human colorectal histological images. Oncology letters 2019;18(6):6101–6107.

[34] Bychkov, D., Linder, N., Turkki, R., Nordling, S., Kovanen, P.E., Verrill, C., et al. Deep learning based tissue analysis predicts outcome in colorectal cancer. Scientific reports 2018;8(1):1–11.

[35] Kainz, P., Pfeiffer, M., Urschler, M.. Segmentation and classification of colon glands with deep convolutional neural networks and total variation regularization. PeerJ 2017;5:e3874.

[36] Qaiser, T., Tsang, Y.W., Taniyama, D., Sakamoto, N., Nakane, K., Epstein, D., et al. Fast and accurate tumor segmentation of histology images using persistent homology and deep convolutional features. Medical image analysis 2019;55:1–14.

[37] He, K., Zhang, X., Ren, S., Sun, J.. Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. 2016, p. 770–778.

[38] Graham, S., Chen, H., Gamper, J., Dou, Q., Heng, P.A., Snead, D., et al. Mild-net: Minimal information loss dilated network for gland instance segmentation in colon histology images. Medical image analysis 2019;52:199–211.

[39] Riasatian, A., Babaie, M., Maleki, D., Kalra, S., Valipour, M., Hemati, S., et al. Fine-tuning and training of densenet for histopathology image representation using tcga diagnostic slides. Medical Image Analysis 2021;70:102032.

[40] Hamida, A.B., Devanne, M., Weber, J., Truntzer, C., Derangère, V., Ghiringhelli, F., et al. Deep learning for colon cancer histopathological images analysis. Computers in Biology and Medicine 2021;136:104730.

[41] Fu, J., Zhang, K., Zhang, P.. Poorly differentiated colorectal

gland segmentation approach based on internal and external stress in histology images. In: 2020 5th International Conference on Computer and Communication Systems (ICCCS). IEEE; 2020, p. 338–342.

[42] Ho, D.J., Yarlagadda, D.V., D'Alfonso, T.M., Hanna, M.G., Grabenstetter, A., Ntiamoah, P., et al. Deep multi-magnification networks for multi-class breast cancer image segmentation. Computerized Medical Imaging and Graphics 2021;88:101866.

[43] Shah, N.A., Gupta, D., Lodaya, R., Baid, U., Talbar, S.. Colorectal cancer segmentation using atrous convolution and residual enhanced unet. arXiv preprint arXiv:210309289 2021;.

[44] Binder, T., Tantaoui, E.M., Pati, P., Catena, R., Set-Aghayan, A., Gabrani, M.. Multi-organ gland segmentation using deep learning. Frontiers in medicine 2019;6:173.

[45] Khened, M., Kollerathu, V.A., Krishnamurthi, G.. Fully convolutional multi-scale residual densenets for cardiac segmentation and automated cardiac diagnosis using ensemble of classifiers. Medical image analysis 2019;51:21–45.

[46] Roth, H.R., Lu, L., Lay, N., Harrison, A.P., Farag, A., Sohn, A., et al. Spatial aggregation of holistically-nested convolutional neural networks for automated pancreas localization and segmentation. Medical image analysis 2018;45:94–107.

[47] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., et al. Attention is all you need. In: Advances in neural information processing systems. 2017, p. 5998–6008.

[48] Khanh, T.L.B., Dao, D.P., Ho, N.H., Yang, H.J., Baek, E.T., Lee, G., et al. Enhancing u-net with spatial-channel attention gate for abnormal tissue segmentation in medical imaging. Applied Sciences 2020;10(17):5729.

[49] Gong, H., Liu, J., Chen, B., Li, S.. Resattengan: Simultaneous segmentation of multiple spinal structures on axial lumbar mri image using residual attention and adversarial learning. Artificial Intelligence in Medicine 2022;:102243.

[50] Pham, V.T., Tran, T.T., Wang, P.C., Chen, P.Y., Lo, M.T.. Ear-unet: A deep learning-based approach for segmentation of tympanic membranes from otoscopic images. Artificial Intelligence in Medicine 2021;115:102065.

[51] Zeng, Z., Xie, W., Zhang, Y., Lu, Y.. Ric-unet: An improved neural network based on unet for nuclei segmentation in histology images. Ieee Access 2019;7:21420–21428.

[52] Del Amor, R., Launet, L., Colomer, A., Moscardó, A., Mosquera-Zamudio, A., Monteagudo, C., et al. An attention-based weakly supervised framework for spitzoid melanocytic lesion diagnosis in whole slide images. Artificial intelligence in medicine 2021;121:102197.

[53] He, H., Zhang, C., Chen, J., Geng, R., Chen, L., Liang, Y., et al. A hybrid-attention nested unet for nuclear segmentation in histopathological images. Frontiers in Molecular Biosciences 2021;8:6.

[54] Dabass, M., Vashisth, S., Vig, R.. Attention-guided deep atrous-residual u-net architecture for automated gland segmentation in colon histopathology images. Informatics in Medicine Unlocked 2021;27:100784.

[55] Shi, T., Li, C., Xu, D., Fan, X.. Fine-grained histopathological cell segmentation through residual attention with prior embedding. Multimedia Tools and Applications 2022;:1–15.

[56] Xu, Y., Zhu, J.Y., Eric, I., Chang, C., Lai, M., Tu, Z.. Weakly supervised histopathology cancer image segmentation and classification. Medical image analysis 2014;18(3):591–604.

[57] Qu, H., Wu, P., Huang, Q., Yi, J., Yan, Z., Li, K., et al. Weakly supervised deep nuclei segmentation using partial points annotation in histopathology images. IEEE Transactions on Medical Imaging 2020;39(11):3655–3666.

[58] Xu, G., Song, Z., Sun, Z., Ku, C., Yang, Z., Liu, C., et al. Camel: A weakly supervised learning framework for histopathology image segmentation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. 2019, p. 10682–10691.

[59] Komura, D., Ishikawa, S.. Machine learning methods for histopathological image analysis. Computational and structural biotechnology journal 2018;16:34–42.

[60] Srinidhi, C.L., Ciga, O., Martel, A.L.. Deep neural network models for computational histopathology: A survey. Medical Image Analysis 2020;:101813.

[61] Simard, P.Y., Steinkraus, D., Platt, J.C., et al. Best practices for convolutional neural networks applied to visual document analysis. In: Icdar; vol. 3. 2003,.

[62] Marée, R., Rollus, L., Stévens, B., Hoyoux, R., Louppe, G., Vandaele, R., et al. Collaborative analysis of multi-gigapixel imaging data using cytomine. Bioinformatics 2016;32(9):1395–1401.

[63] Hamida, A.B., Devanne, M., Weber, J., Truntzer, C., Derangère, V., Ghiringhelli, F., et al. Deep learning for colon cancer histopathological images analysis. Computers in Biology and Medicine 2021;136:104730.

[64] Chen, L.C., Zhu, Y., Papandreou, G., Schroff, F., Adam, H.. Encoder-decoder with atrous separable convolution for semantic image segmentation. In: Proceedings of the European conference on computer vision (ECCV). 2018, p. 801–818.

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

[66] Kather, J.N., Weis, C.A., Bianconi, F., Melchers, S.M., Schad, L.R., Gaiser, T., et al. Multi-class texture analysis in colorectal cancer histology. Scientific reports 2016;6(1):1–11.

[67] Ghosh, S., Bandyopadhyay, A., Sahay, S., Ghosh, R., Kundu, I., Santosh, K.. Colorectal histology tumor detection using ensemble deep neural network. Engineering Applications of Artificial Intelligence 2021;100:104202.