Robust Interactive Semantic Segmentation of Pathology Images with Minimal User Input

Mostafa Jahanifar   Neda Zamani Tajeddin   Navid Alemi Koohbanani

Nasir Rajpoot

Tissue Image Analytics Centre, Department of Computer Science, University of Warwick, UK

mostafa.jahanifar@warwick.ac.uk

Abstract

From the simple measurement of tissue attributes in pathology workflow to designing an explainable diagnostic/prognostic AI tool, access to accurate semantic segmentation of tissue regions in histology images is a prerequisite. However, delineating different tissue regions manually is a laborious, time-consuming and costly task that requires expert knowledge. On the other hand, the state-of-the-art automatic deep learning models for semantic segmentation require lots of annotated training data and there are only a limited number of tissue region annotated images publicly available. To obviate this issue in computational pathology projects and collect large-scale region annotations efficiently, we propose an efficient interactive segmentation network that requires minimum input from the user to accurately annotate different tissue types in the histology image. The user is only required to draw a simple squiggle inside each region of interest so it will be used as the guiding signal for the model. To deal with the complex appearance and amorph geometry of different tissue regions we introduce several automatic and minimalistic guiding signal generation techniques that help the model to become robust against the variation in the user input. By experimenting on a dataset of breast cancer images, we show that not only does our proposed method speed up the interactive annotation process, it can also outperform the existing automatic and interactive region segmentation models.

1. Introduction

Annotated data in computational pathology (CPath) plays an important role in developing algorithms for cancer diagnosis [1] and prognosis [2]. Measurement of tissue attributes can also help make the diagnostic process more objective and reproducible. Furthermore, annotations in digital pathology can be leveraged to extract hidden information embedded in images which cannot be perceived by the hu-

man eye but may be of significant diagnostic value [3].

In digital pathology with the rise of utilizing digital scanners in hospitals and cancer diagnosis centres, digital images are generated at an unprecedented pace. However, providing accurate and reliable labels for these large data repositories is not feasible because it requires expert knowledge, is very time consuming and labour-intensive. Pixel-wise (dense) annotation of tissue regions, which is necessary for quantification and increasing the explainability of models [4], is one of the hardest type of annotations to obtain in digital pathology due to the large size of digital whole slide images (WSIs) that increases the search area for the annotator. Besides, it has been shown that pixel-wise concordance between tissue type annotations from different human annotators in the same region is not very high [5], which portend the challenging nature of the task.

CPath can facilitate this procedure by identifying and segmenting different tissue elements such as cells [6], nuclei [7, 8, 9], glands [10], cancer regions [11], etc, for downstream analysis tasks [12]. However, deep learning (DL) models at the forefront of CPath algorithms require a large amount of annotated data to be trained effectively, otherwise, they cannot generalize well on unseen data [13]. Recently, human-in-the-loop annotation approaches have been proposed in the literature [14] where a DL model generates initial annotations and then the results are reviewed and refined by pathologists. The DL model is then tuned on the refined data and applied to the images to propose a new and presumably better set of annotations [14]. Nonetheless, implementing this approach for tissue region annotation task would not be feasible yet as there is no large publicly available dataset with tissue regions annotated, consequently, the initial DL model may not work well and the need for results refinement will increase. Therefore, the overall human-in-the-loop annotation process might take more time than manual annotation.

One way of facilitating and expediting the process of providing dense annotation is developing platforms and software for interactive segmentation of digital images

which demand minimum human effort and time [15, 16]. Utilizing such software can help us to save time and also generate a large number of annotations for downstream analysis or quantification [17]. Benefiting from human input as guiding signals before segmentation, makes interactive models very accurate and robust in segmenting objects in images from new datasets or domains [15].

In this paper, we propose an interactive segmentation model, in which the user provided squiggles guide the model toward semantic segmentation of tissue regions. Particularly, our main contributions in this work are listed here:

- **Efficient-UNet**: a light-weight, scaleable, and efficient network architecture that benefits from a novel Residual Multi-Scale (RMS) block to better recognize regions with various sizes.
- **Minimalistic guiding signals**: four novel techniques are proposed to allow automatic generation of human-drawn-like guiding signal for training of interactive semantic segmentation model. These methods help the model to become robust against the variations in user input.
- State-of-the-art performance on the largest publicly available dataset [18] for semantic tissue region segmentation.

2. Related work

In various works interactive object segmentation is formulated as energy minimization on a graph defined over the objects [19, 20]. However, it has been shown in [15] that these kinds of methods cannot handle the complexity that might appear in histology primitives (glands or tissue regions) and their performance in comparison to supervised and DL-based methods is relatively poor.

Deep learning models have also been extensively used for interactive segmentation [19, 21, 22, 23, 24] among which DEXTRE [22] is a powerful approach that utilizes extreme points as an auxiliary input to the network. First, the annotator clicks four points on the extreme positions of objects then a heat map (Gaussian map for each point where points are at the centres of Gaussian maps) channel is created from these clicks which are attached to the input and serve as guiding signals. Although DEXTRE performs very well on natural images [22], this kind of approach does not segment different tissue types in histology images for several reasons. First, there may be several separate regions belonging to the same region type, where selecting extreme points for all of them together is not feasible. Furthermore, tissue regions from different tissue types are usually entangled or nested (see Fig. 4 Ground Truth column). Therefore, the extreme points for different regions may overlap with or fall close to (within) each other which may confuse DEXTRE in recognizing different regions.

There are other methods in the literature that require the user to draw a bounding box around the desired object as a guiding signal for the model. Wang *et al*. [25] applied a deep network on a cropped image based on a bounding box to obtain segmentation. In a correction phase, this approach also takes squiggles from the user to indicate the foreground and background. However, this model is not practical for segmentation of multiple objects (like nuclei) or amorphous objects (like glands or tissue region) in the histology domain. Some methods combine bounding box annotations with Graph Convolutional Network (GCN) to achieve image segmentation [23, 24] where the selected bounding box is cropped from the image and fed to a GCN to predict polygon/spline around the object. Again, similar to extreme point-based signals, methods relying on bounding box signals cannot address the shape complexity, nested objects and entangled regions challenges that are present in histology images.

To the best of our knowledge, NuClick [15] is the only interactive segmentation approach for extracting objects in histology images in the literature that deals with these challenges by introducing the use of squiggle based guiding signals. In the original NuClick [15], a random point inside the GT mask and morphological skeleton of the GT mask was used for nucleus and gland segmentation tasks, respectively.

Although the morphological skeleton of the original mask may work well in the case of gland instance segmentation [15], for semantic region segmentation task the skeleton based on the original mask may get complicated due to the noise and variations in the boundaries of the original mask. This effect can be seen in Fig. 2(a), where the morphological skeleton of the original mask is extracted using Lee *et al*. method [26] and it has many small branches that are over-fitted to the GT mask. We have addressed this issue in this work by introducing *minimalistic guiding signals*.

3. The proposed method

The pipeline of the proposed interactive semantic segmentation framework is illustrated in Fig. 1. Our segmentation network follows an encoder-decoder paradigm where two auxiliary maps termed as the "*guiding signals*" are concatenated to the input image. One of these maps point to the desired object in the image (*inclusion map*) and another one refers to the other objects that are present in the same field of view (FOV) but are not of interest (*exclusion map*). During the training phase of the network, these guiding signals are automatically generated based on the ground truth mask whereas in the inference phase we expect the user to draw the guiding signals interactively. In either case, we expect a fine segmentation of the desired object, which is marked by the inclusion signal, in the model output. Outputs related to all guiding signals are then assembled to form the final semantic segmentation map. In the current work, we
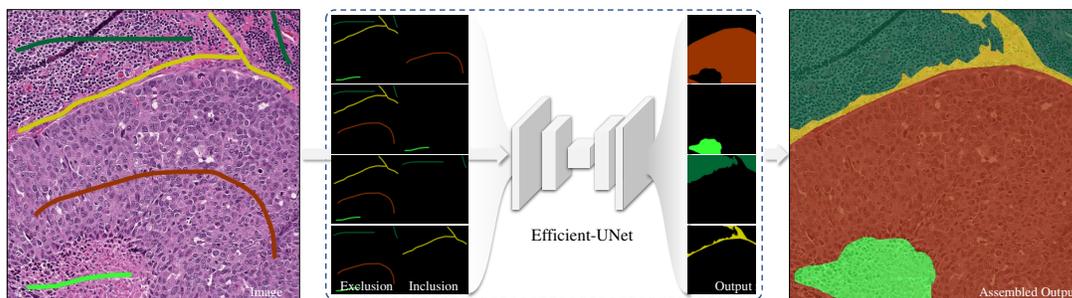
Figure 1. Overview of the proposed interactive segmentation framework. Squiggles in different colours specify different region types, based on which *inclusion* and *exclusion* guiding signals are generated to be fed into the network and generate segmentation map.

propose four techniques to make the model more robust against the variation in the guiding signals by modifying the extracted morphological skeleton during the training (Section 3.1) and introduce an efficient network architecture that is fast and accurate in segmentation task (Section 3.2).

3.1. Minimalistic guiding signals

3.1.1 Mask approximating

One way to reduce the noise on the mask boundaries is to approximate the mask polygon to a similar but more simplistic polygon. For this goal, we extract the boundary co-ordinates of the binary mask to form its polygon, then apply Douglas–Peucker algorithm [27] to iteratively decimate the original polygon to a similar polygon with fewer points. Then, we convert the approximated polygon back to a binary mask. Doing so, we filter out small changes (noise or details) on the mask boundary to obtain a much more simple skeleton, as depicted in Fig. 2(b). To better follow the changes related to mask approximation in Fig. 2, the original mask, the approximated (modified) mask, and their intersection are showed in green, red, and yellow colours, respectively, while the blue line segments depict the generated guiding signal (skeleton) based on the modified mask. During the training, we randomly change the approximation accuracy (the maximum distance between the original polygon and its approximation) to obtain a different guiding signal in each epoch.

3.1.2 Mask smoothing

The idea of image smoothing prior to the medial axis (skeleton) extraction has been investigated before [28] as it would result in less noisy skeletons. In this research, we adopt the same idea and apply smoothing filters on the original mask to reduce the noise in the boundaries and then attain a better morphological skeleton. In our implementation, Gaussian and Median filters with variable scales are incorporated. By changing the filter scale (kernel size and sigma) during the training phase, the resulted modified mask (and the gener-

ated skeleton) would be different in each epoch. The result of applying a Gaussian filter with kernel size of 25×25 and $\sigma = 15$ is shown in Fig. 2(c).

3.1.3 Mask partitioning

Due to the approach used for morphological skeleton generation [26], the resulted guiding signal usually align with the major axis of the source mask or falls on the centre-line of it (see Fig. 2(a)-(c)). In real-world applications, however, the user may draw the guiding signals perpendicular to the main axis or even in segmented lines that do not follow the same orientation. To mimic this behaviour during the guiding signal generation, we propose to partition the mask into smaller parts and calculate the morphological skeleton for each part separately. In our design, each connected component in the mask is treated separately and partitioned adaptively i.e., partitioning direction (horizontal or vertical or both directions) and the number of partitioning lines is decided based on the sizes of that connected component in each axis. Partitioned regions do not necessarily follow the same orientation as the main mask, hence, different parts would result in different skeleton representations. You can see the mask partitioning effect on the generated skeletons in Fig. 2(d) where the biggest object (connected component) in the original mask has been partitioned using a grid pattern (both horizontal and vertical partitioning). The generated guiding signals via this method are not only minimalistic but also vary significantly which will make the network robust against the vast variation in the user input during the inference.

3.1.4 Mask distance transform thresholding

Calculating the distance transformation map of the input mask, randomly thresholding it, and then extracting the guiding signal is the only technique proposed in the NuClick [15]. This method proved to be very effective as not only does it smooth the original mask to obtain a more simple and real guiding signal, but also it shortens the length
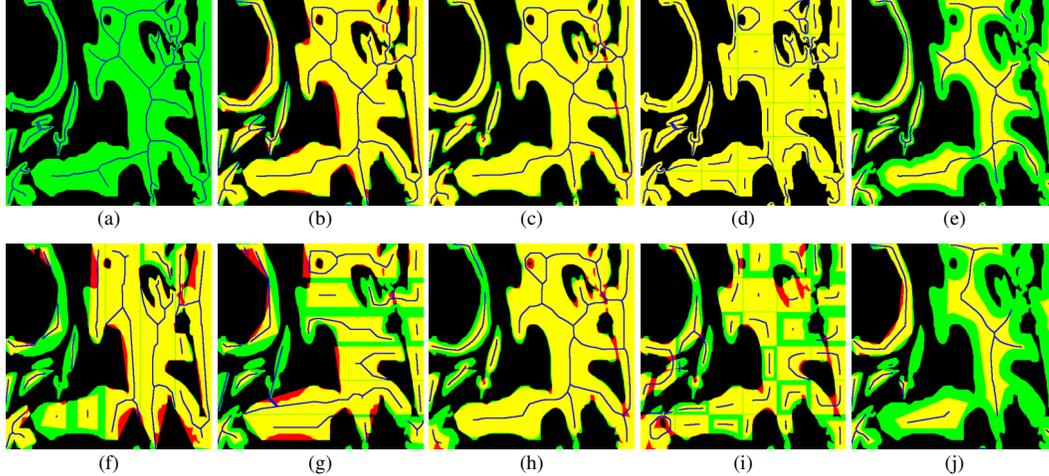
Figure 2. Minimalistic guiding signal generation techniques. (a) shows the original (GT) tissue mask in green and its corresponding skeleton in blue colour. Images (b)-(e) illustrate the effect of mask approximation, mask smoothing, mask partitioning, and distance transformation thresholding on the output guiding signal, respectively. In all these images red and yellow regions indicate the modified mask (used for skeleton extraction) and its intersection with the ground truth, respectively. (f)-(e) are five examples of random combination of mask modification techniques and their resulting skeleton.

of the generated skeleton which is a valid variation in real-world guiding signals. Having said that, in NuClick [15], the distance map was thresholded based on the statistics of all connected components in the mask which was only one gland whereas in tissue region segmentation we have multiple connected components (regions) with various sizes related to the same class and therefore presented in the same mask. If we estimate the threshold based on the statistics of all objects together, that might result in eliminating some objects from the deformed mask and hence no guiding signal would be generated for those objects. Therefore, we proposed to apply distance transformation and thresholding in an adaptive manner to each connected component in the mask, separately. A sample of applying this method to a multi-object mask is depicted in Fig. 2(e).

3.1.5 On-the-fly signal generation

Our implementation of the above-mentioned techniques allows us to incorporate a combination of them for automatic guiding signals generation (both inclusion and exclusion maps) during the training phase. In particular, we apply this ordered sequence of mask approximating, smoothing, partitioning, and distance transform thresholding techniques with probabilities of 0.75, 0.75, 0.5, and 0.5, respectively, and after all, we generate the morphological skeleton [26] of the modified mask as the guiding signal. The combination of these mask modification techniques will guarantee the generation of unique minimalistic guiding signals in each epoch, as illustrated in Fig. 2(f)-(j). It is important to note that although a copy of the original mask is modified

to generate the guiding signal, the original mask is used for network training as the expected output.

3.2. Model and loss function

Inspired by EfficientNet [29], we propose an encoder-decoder network architecture for end-to-end segmentation in which skip connections between the encoder and decoder part have been utilized, similar to UNet [30]. We keep the idea of network scalability as suggested in EfficientNet [29] and mainly used the same convolutional blocks proposed in that work. However, to make our model robust against variation in object scales, we propose a novel Residual Multi-scale (RMS) block inspired by multi-scale blocks introduced in [31]. In the following, we explain in detail how these network constructing blocks work, and how we use them to build a scalable segmentation network architecture.

3.2.1 Network constructing blocks

Mobile Inverted Residual block with Squeeze and Excitation (MIRSE): These blocks as the main part of the network are directly taken from EfficientNet [29]. This block is constructed by embedding a squeeze-and-excitation (S&E) mechanism [32] into a residual block from MobileNetsV2 [33]. As depicted in Fig. 3(a), our implementation of MIRSE blocks requires setting three parameters K, F, and S that are convolution kernel size, the number of feature maps, and squeeze ratio for the S&E (Fig. 3(b)), respectively. Note that in each block and after convolution layers, we use Batch Normalization [34] (BN) and self-gated ac-
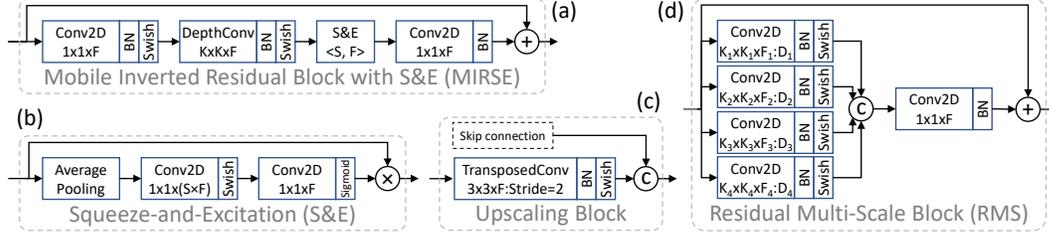
Figure 3. Constructing blocks of the Efficient-UNet segmentation model.

tivation function [35] (Swish) layers. Also, note that the squeeze parameter for all MIRSE blocks in our network has been set to $S = 0.25$.

Upscaling block: This block consists of a 2D transposed convolution layer with a kernel size of 3 and stride of 2 to increase the resolution of feature maps in each stage of the decoding path. This block is also responsible for concatenating the feature maps from the encoder part of the network and creating skip connections.

Residual Multi-Scale block (RMS): This block consists of 4 convolution layers (followed by BN and Swish layers) at different scales. Each layer is a *atrous* convolution [36] with different kernel size (K) and dilation rate (D). In comparison to MSC block [31], the proposed RMS block combines the output feature maps from different scales using a 1×1 convolution and incorporates a residual connection from its input to both keep the effect of original input and ease the flow of gradient during backpropagation [37].

3.2.2 Efficient-UNet Architecture

Similar to [29], we first introduce a baseline model architecture and then scale its width (number of channels or feature maps in constructing blocks) and depth (number of block repetition in each stage of network) uniformly using w and d scaling factors, respectively. These factors are calculated using a compound scaling method and are directly adopted from [29]. Note that unlike [29] we did not scale the network for resolution because the concept of resolution in digital pathology depends on the optical magnification and changing image size or FOV will affect the problem dramatically.

The architecture of our baseline segmentation model, Efficient-UNet-B0, is described in Table 1, in which operations and their configurations in the encoding and decoding paths of the network are outlined. The encoder part contains only MIRSE blocks with different configurations and numbers of repetition, on the other hand, the proposed decoder part encompasses upscaling and RMS blocks as well. As we have four layers of *atrous* convolution in RMS block, this block is configured using kernel size vector $\mathbf{K} = [K_1, K_2, K_3, K_4]$ and dilation rate

Encoder	Decoder
Conv2D (K=3, F=32, Stride=2)	Upscale (F=320) MIRSE (K=5, F=192, R=3)
MIRSE ((K=3, F=16, R=1)	RMS (**K**=[3,5,5,7], **D**=[3,3,5,7], F=192) MIRSE (K=5, F=112, R=3)
MIRSE (K=3, F=24, R=2)*	Upscale (F=112) MIRSE (K=3, F=80, R=3)
MIRSE (K=5, F=40, R=2)*	RMS (**K**=[3,3,5,5], **D**=[1,3,3,5], F=80)
MIRSE (K=3, F=80, R=3)*	Upscale (F=80) MIRSE (K=5, F=40, R=2)
MIRSE (K=5, F=112, R=3)	Upscale (F=40) MIRSE (K=3, F=24, R=2)
MIRSE (K=5, F=192, R=5)*	Upscale (F=24) MIRSE (K=3, F=16, R=1)
MIRSE (K=3, F=320, R=1)	Conv2D (K=1, F=1)

Table 1. Architecture of the baseline network **Efficient-UNet-B0**. For each operation, its parameters are outlined in the parenthesis.

$\mathbf{D} = [D_1, D_2, D_3, D_4]$. From Table 1 note that in stages with *-marked MIRSE block, the first convolution layer would be applied with a stride of 2 to decrease the spatial resolution of feature maps by a factor of 2, similar to max-pooling operation.

During the network scaling, the number of feature maps (F network width) in all blocks (except for the last Conv2D which is responsible for segmentation map generation) and the number of MIRSE block repetitions in each stage (R network depth) will be scaled by w and d factors whereas the size of all kernels and number of Upscale and RMS blocks stay the same.

3.2.3 Loss function

We use a hybrid loss function which consists of soft Dice and binary cross entropy (BCE) parts [38]:

$$\mathcal{L} = 1 - \frac{2\sum_{i,j} p_{ij} g_{ij} + \varepsilon}{\sum_{i,j} p_{ij}^2 + \sum_{i,j} g_{ij}^2 + \varepsilon} - \frac{1}{N}\sum_{i,j} g_{ij}\log(p_{ij}), \quad (1)$$

where p_{ij} and g_{ij} are the values of every (i,j) pixel in the prediction and GT mask, respectively, $\varepsilon = 1$ is used to avoid division by zero, and N is the total number of pixels. The

Dice part of the loss function is robust against the pixels imbalanced class population whereas BCE penalizes the loss for even small pixel-wise errors [38].

4. Experiments and results

4.1. Dataset, preprocessing, and metrics

The dataset used for model training and validation is from the work of Amgad *et al*. [18] which contains 151 H&E stained tissue image regions extracted from WSIs of the same number of triple-negative breast cancer cases acquired from the Cancer Genome Atlas. Tissue regions in this dataset have been annotated with the help of 25 annotators. In total, there are 20340 regions annotated in 20 categories that are merged into 5 classes: tumour, lymphocytic regions, stroma, necrosis or debris, and others (blood, fat, vessel, etc.) [18]. In this work, we followed the same region labelling paradigm and used the same image regions and rendered masks in Amgad *et al*. [18] which are split into 82 slides for training and 43 slides for validation.

Following [18], images are stain normalized using Reinhard's method [39]. The original images are captured at 0.25 micron per pixel (MPP) resolution (which is equal to 40x magnification) with various scanners. However, to keep enough context during the training of our interactive segmentation model, we extract 512×512 patches from image regions and their corresponding masks at 10x magnification (1 MPP resolution). Having guiding signals in the input of the interactive segmentation model, we can use lower resolution images to speed up the region marking and processing time. However, extracted patches were confirmed by a pathologist to show enough contextual and detailed information for tissue region annotation.

For the evaluation of the segmentation performance, we calculate the DICE similarity coefficient, accuracy, and Area Under ROC curve (AUC) of pixel-wise classification. These metrics are reported for each region type separately as well as the overall performance for all tissue types.

4.2. Validation experiments

In this section, we report the results of applying the proposed method and other state-of-the-art (SOTA) interactive or automatic segmentation models on the validation set of the Amgad dataset [18] in Table 2, where the first column indicates the segmentation model, the second column designate the guiding signal used for that model training, and the rest six columns report the evaluation metrics on overall performance or each tissue category. Please note that during the tests on the validation set, for all interactive models, similar guiding signals are provided to make fair comparisons.

Results in Table 2 suggest that interactive segmentation models like NuClick [15] and the proposed method can outperform SOTA automatic segmentation models like UNet [30], DeepLab v3 [36], and the baseline method [18] by a large margin as they are provided with guiding signals in the input. Particularly, our best performing model, **Efficient-Unet-B3**, achieves overall Dice, accuracy, and AUC of 0.875, 0.984, 0.995, respectively. In comparison to the SOTA automatic segmentation models, our proposed approach performs about 14% and 11% better than UNet [30] and DeepLab v3 [36] in terms of overall Dice score, respectively. The same trend can be seen not only for overall accuracy and AUC metrics but also for all the metrics reported for different tissue types in Table 2. Note that the lower performance for some region types in comparison to the overall (average) performance can be associated with the higher noise in GT annotations of those regions. Higher noise in GT arises from ambiguity in the boundaries of these regions which makes it hard to separate them from other regions as reported in [18]. Although the original NuClick [15] performs better than all other automatic segmentation models (overall Dice score 0.773), it still shows lower performance metrics than the proposed method i.e., Efficient-Unet-B3 segmenting 10% better in terms of Dice score. Nevertheless, it can be seen that when we train the NuClick model with the proposed minimalistic signals (Section 3.1), overall Dice scores rises to 0.835, which shows the effectiveness of the proposed minimalistic guiding signal generation.

The four last rows of Table 2 are reporting the effect of increasing the scale of the proposed Efficient-UNet model architecture. In each row, number of features maps F and repetition of MIRSE blocks R (see Table 1) are multiplied by the introduced w and d factors. Scaling up the network from Efficient-Unet-B0 variant to the B3 variant does elevate the performance level of interactive segmentation, however, this raise is not significant. Specifically, the largest model (Efficient-Unet-B3) performs only 0.8% better than the baseline model (Efficient-Unet-B0) in overall Dice.

4.3. Experiment with human-drawn guiding signals

To test the robustness of the proposed method against the variations in human-provided guiding signals, we presented a subset of validation set (25 images) to 2 non-expert annotators and asked them to draw guiding signals on images. To make the process consistent between the two annotators, for each image and each region type we overlaid the ground truth annotation on images and asked the annotators to draw their squiggle markers (guiding signals) inside the annotated region. By doing this, we are able to compare the performance of the same model using minimalistic (auto-generated) signals (Section 3.1) versus using human-drawn guiding signals as reported in Table 3.

The trained Efficient-Unet-B0 performs accurately whether the input guiding signals are automatically generated or provided by a human annotator. In Table 2, the

Model	Guiding Signal	Overall			Tumour			Stroma			Inflamatory			Necrosis			Others		
		DICE	Acc.	AUC	DICE	Acc.	AUC	DICE	Acc.	AUC	DICE	Acc.	AUC	DICE	Acc.	AUC	DICE	Acc.	AUC
Amgad *et al.* [18]	-	0.750	0.783	0.898	0.851	0.804	0.941	0.8	0.824	0.881	0.712	0.743	0.917	0.723	0.872	0.864	0.666	0.67	0.885
UNet [30]	-	0.733	0.756	0.880	0.833	0.772	0.913	0.788	0.798	0.873	0.701	0.722	0.907	0.713	0.856	0.842	0.629	0.634	0.866
DeepLab v3 [36]	-	0.760	0.794	0.904	0.862	0.822	0.956	0.812	0.841	0.893	0.725	0.76	0.921	0.729	0.875	0.865	0.671	0.673	0.887
NuClick [15]	NuClick	0.773	0.957	0.968	0.855	0.941	0.972	0.793	0.926	0.957	0.724	0.958	0.960	0.810	0.974	0.983	0.681	0.986	0.967
NuClick	Minimalistic	0.835	0.967	0.991	0.895	0.951	0.989	0.839	0.944	0.981	0.824	0.970	0.991	0.851	0.979	0.996	0.768	0.990	0.995
Efficient-Unet-B0 ($w=1, d=1$)	Minimalistic	0.867	0.978	0.994	0.931	0.973	0.995	0.877	0.961	0.987	0.854	0.978	0.995	0.883	0.986	0.997	0.790	0.991	0.996
Efficient-Unet-B1 ($w=1, d=1.1$)	Minimalistic	0.869	0.980	0.994	0.933	0.973	**0.996**	0.875	0.961	0.987	0.855	0.980	0.995	0.889	0.991	**0.997**	0.792	0.993	0.996
Efficient-Unet-B2 ($w=1.1, d=1.2$)	Minimalistic	0.871	0.981	0.994	**0.935**	0.977	**0.996**	0.878	0.962	0.987	0.853	0.977	0.995	0.888	0.992	**0.997**	0.801	0.995	**0.997**
Efficient-Unet-B3 ($w=1.2, d=1.4$)	Minimalistic	**0.875**	**0.984**	**0.995**	**0.935**	0.978	**0.996**	**0.881**	**0.969**	**0.990**	**0.859**	0.982	**0.996**	**0.891**	0.993	**0.997**	**0.809**	0.996	**0.997**

Table 2. Evaluation results on validation set of Amgad *et al.* [18] dataset using different automatic and interactive segmentation methods.

Model	Guiding Signal	Overall		
		DICE	Acc.	AUC
Efficient-Unet-B0	Minimalistic	0.874	0.982	0.994
Efficient-Unet-B0	Human #1	0.887	0.986	0.995
Efficient-Unet-B0	Human #2	0.870	0.981	0.994

Table 3. Comparison of proposed model performance on a subset of 25 validation images when used with auto-generated minimalistic guiding signals versus human-provided guiding signals.

model provided with guiding signals drawn by annotator "Human #2" interestingly performs better than the same model with auto-generated signals, achieving an overall Dice score of 0.887 over 0.874. Qualitative results for three different samples in Fig. 4 (where each region type has been coloured differently and guiding signals for tissue region are depicted with dark blue lines) also show that the trained models can handle variations in the guiding signals and segment tissue regions quite perfectly even if provided with overly simplified guiding signals.

5. Discussion

Evaluation results in Table 2 show that our proposed segmentation models, Efficient-UNet B0-B3, perform considerably better than SOTA interactive segmentation model, NuClick [15], using the same minimalistic guiding signal (Section 3.1) and increase the overall Dice metric by 3%–4% as the network scale grows. One may argue that despite the increase of network scale, its performance does not elevate considerably i.e., less than 1% increase in all performance metrics throughout all tissue types when increasing the Efficient-UNet scale from B0 to B3. With the growth of network scale, our models capacity for learning feature representations also expands [13], however, bigger networks usually need more training data to be trained optimally whereas we only used a limited number of training images in this work. Therefore, we believe that one of the main reasons for not having a considerable performance boost while increasing network scale is the lack of training data. Nevertheless, we kept the design of the network as scaleable as possible so it can be used in future experiments or other applications with more data.

The efficiency of the proposed segmentation network is not only related to its accuracy but also its speed. The baseline network Efficient-UNet-B0 is twice faster as compared to NuClick while gaining 3% higher overall Dice metric. Particularly, our network can generate the annotation map for each guiding signal in about 0.1 seconds (using an Nvidia Tesla V100 GPU). This means that the proposed network fits in an interactive annotation setup where it can respond to the annotator's markers in real-time and make the interactive process more user friendly.

The positive effect of introducing proposed minimalistic signal generation techniques during the training phase is visible in validation results of Table 2 and Table 3. In Table 2, one can see that by only using the proposed minimalistic guiding signal generation method, overall NuClick performance (Dice metric) increases by more than 6% that is more prominent than the effect of improving network architecture discussed in the previous paragraphs (3% improvement when using Efficient-Unet instead of NuClick). Results in Table 3 also testify to the effectiveness of the proposed method for guiding signal generation, where a given model performs at the same level or even better when it is provided human-drawn guiding signals instead of GT-based extracted ones. This shows that the proposed techniques in Section 3.1 can generate guiding signals that resemble human-drawn markers and capture the variations they might show. This claim is supported by the qualitative results illustrated in Fig. 4, where even overly-simplified squiggles as guiding signals can lead to high quality segmentation outputs. This also implies that the proposed interactive segmentation model is insensitive to variations in the user input. This experiment expresses the adaptiveness of the proposed interactive segmentation method, where the more carefully and detailed guiding signals are drawn, the more accurate the segmentation output will be. Having near perfect initial annotations by only providing a simple guiding signal in the beginning will save a considerable amount of annotator time during the annotation review and correction [5].

Although we have trained and tested our interactive segmentation models on image-level information, it is only a matter of implementation to extend this idea to work on WSIs. Coupled with WSI viewer and annotation tools like

| Ground Truth | Minimalistic (auto) Signal | Human #1 signal | Human #2 signal |

● Tumour　　● Stroma　　● Lymphocytic　　● Necrosis/Debris　　● Blood　　● Fat　　● Plasma cells　　● Vessel
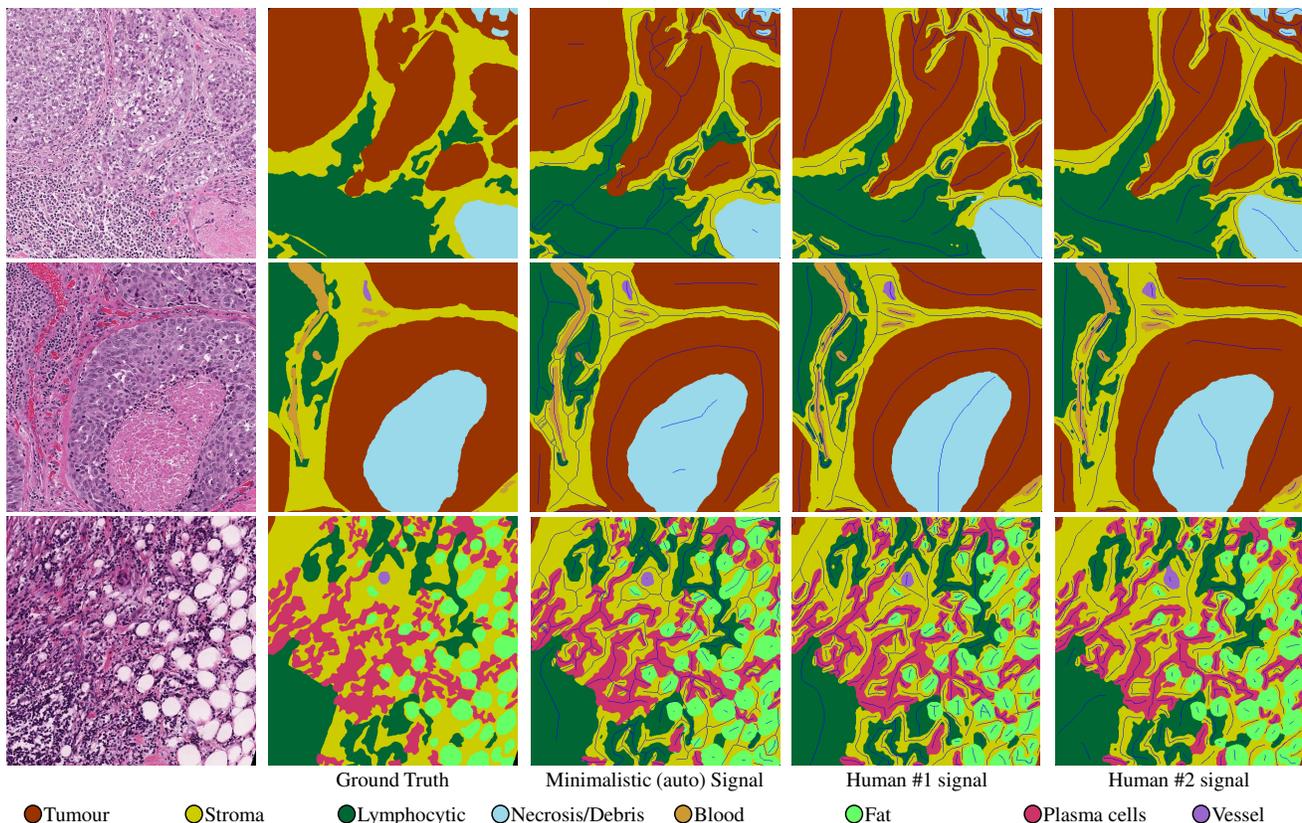
Figure 4. Results of applying proposed baseline interactive segmentation model, **Efficient-UNet-B0**, on three samples using minimalistic guiding signals (described in Section 3.1), and guiding signals provided by two human annotators.

ASAP [40], the proposed interactive method can speed up the annotation curation process considerably. Annotators only need to view the WSI at low-level magnification and draw squiggles with different indices (colours) inside different tissue types, then we only need to extract patches based on the drawn markers and feed them into the proposed interactive segmentation model to get the initial segmentation map for each path. Results of the patches are then stitched and overlaid on the WSI in the form of polygons so they can be edited by the annotators.

6. Conclusions and future directions

In this work, we proposed an interactive semantic segmentation model for robust tissue region annotation in a semi-automated manner. Our approach follows the work of [15] in which for segmentation of each object in the image, two guiding signal maps (inclusion and exclusion maps) are concatenated to the image in the input of the network. Here, we improved the network architecture by introducing Efficient-UNet which is 2 times faster and 4% more accurate in terms of overall Dice metric than the original NuClick. We have also proposed four novel techniques

to automatically extract minimalistic and human-drawn-like guiding signals from GT masks, so they can be used during the training of the model. We showed that using the proposed minimalistic guiding signals makes the models robust against variations in the user input and can improve the overall Dice metric by another 6%. We believe this AI tool can be embedded in a whole slide image viewer/annotation software efficiently to effectively accelerate large-scale region annotation acquisition projects.

There are several ideas that can be pursued to extend this work. Testing with the generalizability of the proposed method to new unseen datasets (e.g. samples from other organs) or checking the adaptability of the trained model to new domains (like IHC stained samples) can be of interest as an extension of this work. In that case, we would like to investigate the effect of incorporating unsupervised pre-training methods, such as SelfPath [41], on the performance of the proposed interactive method for tissue region segmentation in new domains. Furthermore, it is possible to add a classification head to the proposed network architecture to enable proposing a category label for each region which may also improve the segmentation quality synergistically.

References

[1] Hanya Mahmood, Muhammad Shaban, Nasir Rajpoot, and Syed A Khurram. Artificial intelligence-based methods in head and neck cancer diagnosis: an overview. *British journal of cancer*, 124(12):1934–1940, 2021.

[2] Najah M Alsubaie, David Snead, and Nasir M Rajpoot. Tumour nuclear morphometrics predict survival in lung adenocarcinoma. *IEEE Access*, 9:12322–12331, 2021.

[3] Neda Zamanitajeddin, Mostafa Jahanifar, and Nasir Rajpoot. Cells are actors: Social network analysis with classical ml for sota histology image classification. *arXiv preprint arXiv:2106.15299*, 2021.

[4] Alexander Binder, Michael Bockmayr, Miriam Hägele, Stephan Wienert, Daniel Heim, Katharina Hellweg, Masaru Ishii, Albrecht Stenzinger, Andreas Hocke, Carsten Denkert, et al. Morphological and molecular breast cancer profiling through explainable machine learning. *Nature Machine Intelligence*, 3(4):355–366, 2021.

[5] Noorul Wahab, Islam M Miligy, Katherine Dodd, Harvir Sahota, Michael Toss, Wenqi Lu, Mostafa Jahanifar, Mohsin Bilal, Simon Graham, Young Park, et al. Semantic annotation for computational pathology: Multidisciplinary experience and best practice recommendations. *arXiv preprint arXiv:2106.13689*, 2021.

[6] Mostafa Jahanifar, Meisam Hasani, and Seyed Jamal Khaleghi. Automatic zone identification in blood smear images using optimal set of features. In *2016 23rd Iranian Conference on Biomedical Engineering and 2016 1st International Iranian Conference on Biomedical Engineering (ICBME)*, pages 128–133. IEEE, 2016.

[7] Navid Alemi Koohababni, Mostafa Jahanifar, Ali Gooya, and Nasir Rajpoot. Nuclei detection using mixture density networks. In *International Workshop on Machine Learning in Medical Imaging*, pages 241–248. Springer, 2018.

[8] Neeraj Kumar, Ruchika Verma, Deepak Anand, Yanning Zhou, Omer Fahri Onder, Efstratios Tsougenis, Hao Chen, Pheng-Ann Heng, Jiahui Li, Zhiqiang Hu, et al. A multi-organ nucleus segmentation challenge. *IEEE transactions on medical imaging*, 39(5):1380–1391, 2019.

[9] Navid Alemi Koohbanani, Mostafa Jahanifar, Ali Gooya, and Nasir Rajpoot. Nuclear instance segmentation using a proposal-free spatially aware deep learning framework. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 622–630. Springer, 2019.

[10] Simon Graham, Hao Chen, Jevgenij Gamper, Qi Dou, Pheng-Ann Heng, David Snead, Yee Wah Tsang, and Nasir Rajpoot. Mild-net: Minimal information loss dilated network for gland instance segmentation in colon histology images. *Medical image analysis*, 52:199–211, 2019.

[11] Quoc Dang Vu, Simon Graham, Tahsin Kurc, Minh Nguyen Nhat To, Muhammad Shaban, Talha Qaiser, Navid Alemi Koohbanani, Syed Ali Khurram, Jayashree Kalpathy-Cramer, Tianhao Zhao, et al. Methods for segmentation and classification of digital microscopy tissue images. *Frontiers in bioengineering and biotechnology*, 7:53, 2019.

[12] Fang Zhang, Su Yao, Zhi Li, Changhong Liang, Ke Zhao, Yanqi Huang, Ying Gao, Jinrong Qu, Zhenhui Li, and Zaiyi Liu. Predicting treatment response to neoadjuvant chemoradiotherapy in local advanced rectal cancer by biopsy digital pathology image features. *Clinical and translational medicine*, 10(2):e110, 2020.

[13] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *nature*, 521(7553):436–444, 2015.

[14] Samuel Budd, Emma C Robinson, and Bernhard Kainz. A survey on active learning and human-in-the-loop deep learning for medical image analysis. *Medical Image Analysis*, page 102062, 2021.

[15] Navid Alemi Koohbanani, Mostafa Jahanifar, Neda Zamani Tajadin, and Nasir Rajpoot. Nuclick: a deep learning framework for interactive segmentation of microscopic images. *Medical Image Analysis*, 65:101771, 2020.

[16] Mostafa Jahanifar, Navid Alemi Koohbanani, and Nasir Rajpoot. Nuclick: From clicks in the nuclei to nuclear boundaries. In *MICCAI 2019 Workshop COMPAY*, 2019.

[17] Jevgenij Gamper, Navid Alemi Koohbanani, Ksenija Benes, Simon Graham, Mostafa Jahanifar, Syed Ali Khurram, Ayesha Azam, Katherine Hewitt, and Nasir Rajpoot. Pannuke dataset extension, insights and baselines. *arXiv preprint arXiv:2003.10778*, 2020.

[18] Mohamed Amgad, Habiba Elfandy, Hagar Hussein, Lamees A Atteya, Mai AT Elsebaie, Lamia S Abo Elnasr, Rokia A Sakr, Hazem SE Salem, Ahmed F Ismail, Anas M Saad, et al. Structured crowdsourcing enables convolutional segmentation of histology images. *Bioinformatics*, 35(18):3461–3467, 2019.

[19] Hiba Ramadan, Chaymae Lachqar, and Hamid Tairi. A survey of recent interactive image segmentation methods. *Computational visual media*, pages 1–30, 2020.

[20] Carsten Rother, Vladimir Kolmogorov, and Andrew Blake. Grabcut: Interactive foreground extraction using iterated graph cuts. In *ACM transactions on graphics (TOG)*, volume 23, pages 309–314. ACM, 2004.

[21] Ning Xu, Brian Price, Scott Cohen, Jimei Yang, and Thomas Huang. Deep grabcut for object selection. *arXiv preprint arXiv:1707.00243*, 2017.

[22] Kevis-Kokitsi Maninis, Sergi Caelles, Jordi Pont-Tuset, and Luc Van Gool. Deep extreme cut: From extreme points to object segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 616–625, 2018.

[23] Huan Ling, Jun Gao, Amlan Kar, Wenzheng Chen, and Sanja Fidler. Fast interactive object annotation with curve-gcn. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5257–5266, 2019.

[24] Lluis Castrejon, Kaustav Kundu, Raquel Urtasun, and Sanja Fidler. Annotating object instances with a polygon-rnn. In

Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 5230–5238, 2017.

[25] Guotai Wang, Wenqi Li, Maria A Zuluaga, Rosalind Pratt, Premal A Patel, Michael Aertsen, Tom Doel, Anna L David, Jan Deprest, Sébastien Ourselin, et al. Interactive medical image segmentation using deep learning with image-specific fine tuning. *IEEE transactions on medical imaging*, 37(7):1562–1573, 2018.

[26] Ta-Chih Lee, Rangasami L Kashyap, and Chong-Nam Chu. Building skeleton models via 3-d medial surface axis thinning algorithms. *CVGIP: Graphical Models and Image Processing*, 56(6):462–478, 1994.

[27] David H Douglas and Thomas K Peucker. Algorithms for the reduction of the number of points required to represent a digitized line or its caricature. *Cartographica: the international journal for geographic information and geovisualization*, 10(2):112–122, 1973.

[28] Daniel S Fritsch, Stephen M Pizer, Bryan S Morse, David H Eberly, and Alan Liu. The multiscale medial axis and its applications in image registration. *Pattern recognition letters*, 15(5):445–452, 1994.

[29] Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International Conference on Machine Learning*, pages 6105–6114. PMLR, 2019.

[30] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.

[31] Mostafa Jahanifar, Neda Zamani Tajeddin, Navid Alemi Koohbanani, Ali Gooya, and Nasir Rajpoot. Segmentation of skin lesions and their attributes using multi-scale convolutional neural networks and domain specific augmentations. *arXiv preprint arXiv:1809.10243*, 2018.

[32] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7132–7141, 2018.

[33] 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.

[34] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International conference on machine learning*, pages 448–456. PMLR, 2015.

[35] Prajit Ramachandran, Barret Zoph, and Quoc V Le. Swish: a self-gated activation function. *arXiv preprint arXiv:1710.05941*, 7:1, 2017.

[36] Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. Rethinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv:1706.05587*, 2017.

[37] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[38] Mostafa Jahanifar, Neda Zamani Tajeddin, Meisam Hasani, Babak Shekarchi, and Kamran Azema. Automatic recognition of the supraspinatus tendinopathy from ultrasound images using convolutional neural networks. *arXiv preprint arXiv:2011.11777*, 2020.

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

[40] G Litjens. Automated slide analysis platform (asap). http://rse.diagnijmegen.nl/software/asap/, 2017.

[41] Navid Alemi Koohbanani, Balagopal Unnikrishnan, Syed Ali Khurram, Pavitra Krishnaswamy, and Nasir Rajpoot. Self-path: Self-supervision for classification of pathology images with limited annotations. *IEEE Transactions on Medical Imaging*, 2021.