Dual-Decoding Y-Net: A Multi-task Model Using Contour Prediction to Enhance the Semantic Segmentation of Colorectal Cancer Regions

Abstract

Purpose: The detection of tumor regions has been increasingly applied as a key step in pathological image analysis. The ambiguity of the boundary of the region, an essential characteristic of the images of tumors, makes the edges of the tumor region difficult to detect. Traditional methods (such as U-Net) have achieved good results on nuclear segmentation and other issues, but the effect on tumor segmentation needs improving. This paper aims to detect effectively the blurred edge of tumor by using semantic segmentation. **Methods:** A new model is proposed in this paper for tumor region segmentation. We developed a multi-task learning technique for the segmentation based on an improved version of the U-Net, and introduced a parallel contour decoder as an auxiliary task to learn the features of the contour. In addition to using the tumor region masks to supervise the main task, we exported the masks of tumor region contours to supervise the auxiliary task. In order to promote the information fusion of the two tasks, the same encoder was shared and skip connections and spatial gradient fusion were applied between the two decoders. **Results:** Experiments show that our model has achieved satisfying results on the TCGA colorectal cancer pathological image dataset, and the accuracy is 98.20%, which is higher than that of the baseline methods. **Conclusion:** The proposed method can improve the quality of semantic segmentation by enhancing the detection of tumor edge, and can learn the tumor region contours of different subtypes.

Keywords

Semantic Segmentation, Colorectal Cancer, Pathological Images, Multi-Task Learning, Tumor Region Segmentation

1 Introduction

In the field of computer-aided diagnosis (CAD), there is a broad application prospect for the detection of tumor regions in the whole slide images (WSIs). In addition to directly assisting doctors in the diagnosis, the detection of tumor regions can also serve as the basis for many complicated technologies, such as cancer cell segmentation, tumor classification and grade, and medical prognosis.

It is a classic method to treat this task as a classification problem. WSIs are cut into small patches and labeled at patch-level, then an image classifier based on CNN is trained to classify cancer and normal tissue patches. This method is often applied to CAD related research. For instance, Jakob et al. trained a patch-level classifier based on ResNet-18 which can classify gastric tumor patches and colorectal tumor patches while recognizing tumor regions before constructing the prediction model of microsatellite instability (MSI) for gastrointestinal cancer cases [1]. Nicolas et al. introduced a tumor patch classifier based on Inception-v3 to recognize tumor regions and to predict the mutated genes in non-small cell lung cancer [2]. Classification models have been used in related semantic segmentation challenges in recent years. For instance, Wang et al. trained a GoogLeNet model to classify the patches of WSIs to detect cancer metastasis in lymph nodes and won the Camelyon16 Challenge in 2016, organized by IEEE International Symposium on Biomedical Imaging (ISBI) [3]. These segmentation methods based on practical image classification models have proved the effect of tumor region detection successfully. However, patch-level annotations are relatively rough for the tumor region segmentation tasks because it is easy to introduce noise that may lead to incomplete use of information due to patch-level data

cleaning.

In contrast, pixel-level annotations of tumor regions which are suitable for semantic segmentation can maximize the use of information while avoiding noise. Its development prospects are obviously better. Some works have tried semantic segmentation methods on many cancer species. For example, in the task of breast cancer region segmentation, Guo et al. combined the semantic segmentation method with the classification method [4]. The tumor regions are detected in advance by using the classification model Inception-v3, and then the semantic segmentation model DCNN is used for fine segmentation. This work achieved leading points and performance on the Camelyon16 dataset. Chen et al. achieved real-time detection of breast and prostate cancer regions by modifying optical microscopes [5]. By applying FCN to the Inception-v3 architecture, they proposed InceptionV3-FCN, which can achieve fast segmentation by reducing the amount of computation by 75%. The input size of this model is 1000px × 1000px, but the actual field of the view of microscopes is 5120px × 5120px. So, they use the sliding window method to process the entire images.

However, these segmentation models ignore a very important feature when directly learning the pattern of tumor regions, that is, the ambiguity of boundaries. The heterogeneity of tumors makes their boundaries very complicated and variable in morphology because the tumor regions are composed of cancer cells and have complex microenvironments, so there are many unpredictable possibilities. Therefore, the detection and labeling of tumor regions usually require professional pathologists. This also makes the boundary problem of tumor region segmentation of pathological images more complicated than object segmentation on natural images, which needs special attention. In addition, the U-Net proposed by Ronneberger et al. is structurally suitable for medical image semantic segmentation tasks, which inspired many works [6]. It performs well on many medical-related tasks such as cell nucleus segmentation.

Some works have focused on contour perception by introducing a multi-decoder design. Chen et al. proposed DCAN, which uses two decoders to decode feature maps: one of which focuses on the content of the segmented object, and the other focuses on the contour of the segmented object [7]. After being independently trained, the two decoders are combined in the inference stage. However, the skip connections in this model are limited to the encoder and decoder, which leads to insufficient information fusion. Zhen et al. noticed that the classification effect of boundary pixels in semantic segmentation is not satisfactory [8]. On the one hand, the subsampling operation of deep convolution network loses the detail information; on the other hand, cross entropy loss does not make the network pay attention to the classification of boundary pixels. Therefore, they proposed a multi-task model to couple the semantic segmentation task with the semantic edge detection task, and proposed spatial gradient fusion to suppress the non-semantic boundary in the semantic edge detection task, and introduced a loss function including boundary consistency constraint to improve the boundary pixel accuracy. However, although this work introduces a multi-scale pyramid context module, there is a lack of multi-level information fusion between the decoders of the two tasks.

In this paper, we proposed a U-Net-based semantic segmentation model which is named Y-Net. In order to pay more attention to the boundary information of tumor regions, we introduced the contour decoder in addition to the content decoder of tumor regions and developed a multi-task learning technique for tumor region segmentation. In the model structure, in addition to the skip connections between the encoder and the two decoders, there are also skip connections from the contour decoder to the content decoder, which are used to fuse the contour information into the content decoder during each up-sampling. Besides, we introduced spatial gradient fusion to implement the information fusion from the content

decoder to the contour decoder, so as to implement the bidirectional information fusion between the two decoders. We use this innovative structure to distinguish information fusion from content to contour and from contour to content. Furthermore, we use different ground truth and loss functions to supervise the two tasks, so that each loss function focuses more on content or contour information. Experiments show that the proposed model has enhanced the segmentation effect of the multi-subtype tumor regions with an accuracy rate of 98.20%.

2 Materials and Methods

2.1 Whole Slide Image Preprocessing

The pathological images of colorectal cancer we used were from TCGA dataset. 100 cases of colon cancer and 100 cases of rectal cancer were selected from TCGA-COAD project and TCGA-READ project respectively. In each case, a WSI stained with H&E was selected.

The contour of tumor regions in each WSI was annotated by pathologists using the software of Aperio ImageScope in 20× field of view. These contour annotations are composed of vector curves, which are stored as a sequence of key point coordinates in the software. We wrote scripts to reconstruct the contour annotations on the blank canvas, and generated two types of mask: contour mask and content mask. The content mask is the filling of the outline mask. In the contour mask, we set the contour width to 50px, because this width can well contain the morphological differences on both sides of the contour (tumor area and non-tumor area), and avoid irrelevant information. The generated content masks, contour masks and the original WSIs were cut into 1000px × 1000px patches by sliding window method. Finally, the completely blank patches or patches with a small amount of impurities are removed by RGB standard deviation threshold method. The overall preprocessing schematic diagram is shown in Figure 1. Finally, the train set and the test set were divided with a ratio of 4:1 at the case level.

This preprocessing method is based on the widely-used vector curve annotation method, and the proposed process of generating tumor region contour mask is very simple, so this method is easy to migrate to other segmentation tasks, and it has a certain versatility.

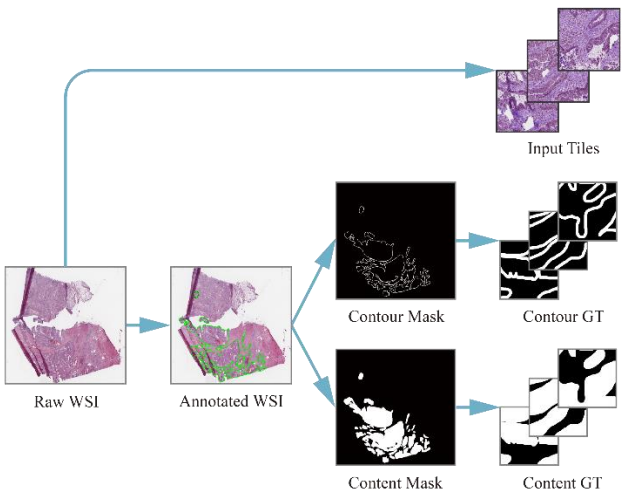

Figure 1. The overall process of WSI preprocessing. The same raw WSI is preprocessed into patches, content masks and contour masks. The latter two are ground truth (marked GT in the figure)

2.2 The Architecture of Y-Net

The proposed Y-Net is improved based on U-Net,which was proposed by Ronneberger et al. in 2015[6].

U-Net is ahead of traditional FCN with two advantages. The first is the symmetrical design of decoder (expanding path) and encoder (contracting path). The number of up-sampling operations in the decoder is equal to the number of down-sampling operations in the encoder. This is helpful to recover the lost resolution in the down-sampling, thus guiding the encoder to extract more relevant features. The second is the encoder-to-decoder skip connections. Skip connections can alleviate the gradient vanishing problem, and make the decoder supplement information from each resolution level of the encoder, so as to recover the information lost by the down-sampling operation. In recent years, a lot of works have been innovating on U-Net, and generally have the following three directions. The first is to innovate in the implementation of the encoder and decoder. For example, Diakogiannis et al. combined the idea of deep residual networks with U-Net backbone, thus designing ResUNet-a for semantic segmentation in the field of remote sensing [9]. The second is to add special layers between the encoder and decoder, such as Gu et al. added a multi-resolution feature extractor based on inception and PSP structure to their model CE-Net and used for medical image segmentation for multiple purposes [10]. The third is the change in the backbone. For instance, Zhou et al. embedded recursive sub-models in U-Net backbone and named it U-Net++, so that the number of U-Net layers and the receptive field can be adjusted flexibly [11]. There are also cases where U-Net is used for multi-task learning. For example, Ke et al. divided the semantic segmentation of food microscopy images into 3 related tasks [12].However, skip connections are still limited in these models to the encoders and decoders.

The proposed Y-Net has changed the encoder, decoder and backbone of U-Net. It has 3 main parts: the joint encoder, the content decoder, and the contour decoder. The main structure is shown in Figure 2. For the encoder, we used the ResNet34 model proposed by He et al. to transform it to avoid gradient vanishing problem while obtaining a larger receptive field and reducing the training time [13]. Consistent with U-Net, four down-sampling operations are performed during the encoding process. The basic unit of the decoder adopts the design corresponding to the encoder. The details of the encoder and decoders are shown in Figure 3.

In order to enable the model to learn the content and contour information of the tumor regions at the same time, we introduce the architecture of multi-task learning, and train two tasks at the same time with joint supervision. Furthermore, the skip connections from the contour decoder to the content decoder are introduced to ensure that the content decoder can obtain the contour information every time the feature map resolution changes, thereby improving the segmentation effect. In the experiments, we put together the feature maps' final output by the two decoders and calculated the loss with the corresponding masks.

In order to further enhance the information fusion between the two tasks, spatial gradient fusion is introduced to realize the information fusion from the content decoder to the contour decoder. Through the derivation of spatial gradient, the semantic boundary can be easily obtained from the semantic segmentation mask output by the content decoder. Based on the work of Zhen et al. [8], we use adaptive pooling to derive the spatial gradient∇M, which is

$$\nabla M[i,j] = |M[i,j] - pool_k\ (M[i,j])$$

where i and j are the current location of the content mask and $pool_k$ is an adaptive average pooling operation with kernel size k, k is used to control the derived boundary width and is set to 10 in our method. As shown in the Figure 2, ∇M will be calculated on the output of the content decoder and superimposed on the feature map of the output of the contour decoder as the final contour mask.

2.3 Loss Function

The common loss functions in semantic segmentation tasks are Dice coefficient loss and cross-entropy loss, etc. In order to supervise the content segmentation task and contour segmentation task, we decided to use different loss functions when we observed, annotated and preprocessed the image data in the early stage.

For the loss function of the content segmentation task, after removing all the blank patches, we found that the area of cancerous region and non-cancerous region had no significant difference in statistics, so the pixel-level binary cross-entropy loss function (BCE loss) was used, which is given by

$$L_{content}(o_1, g_1) = \frac{-\sum_{i,j}(g_1[i,j] \cdot \log(o_1[i,j])}{w \cdot h} + \frac{-\sum_{i,j}(1 - g_1[i,j]) \cdot \log(1 - o_1[i,j])}{w \cdot h}$$

where o_1 is the output feature map of the content decoder, g_1 is the ground truth of the content, i and j are the current location of the feature map, w and h represent size of the feature maps.

For the contour segmentation task, the foreground area is far smaller than the background area, so we use the Dice coefficient loss function which has strong robustness to the problem of pixel category imbalance. The loss can be expressed as

$$L_{contour}(o_2, g_2) = 1 - \frac{2\sum_{i,j} g_2[i,j] \cdot o_2[i,j]}{\sum_{i,j}(g_2[i,j]^2 + o_2[i,j]^2)}$$

where o_2 is the output feature map of the contour decoder, g_2 is the ground truth of the contour.

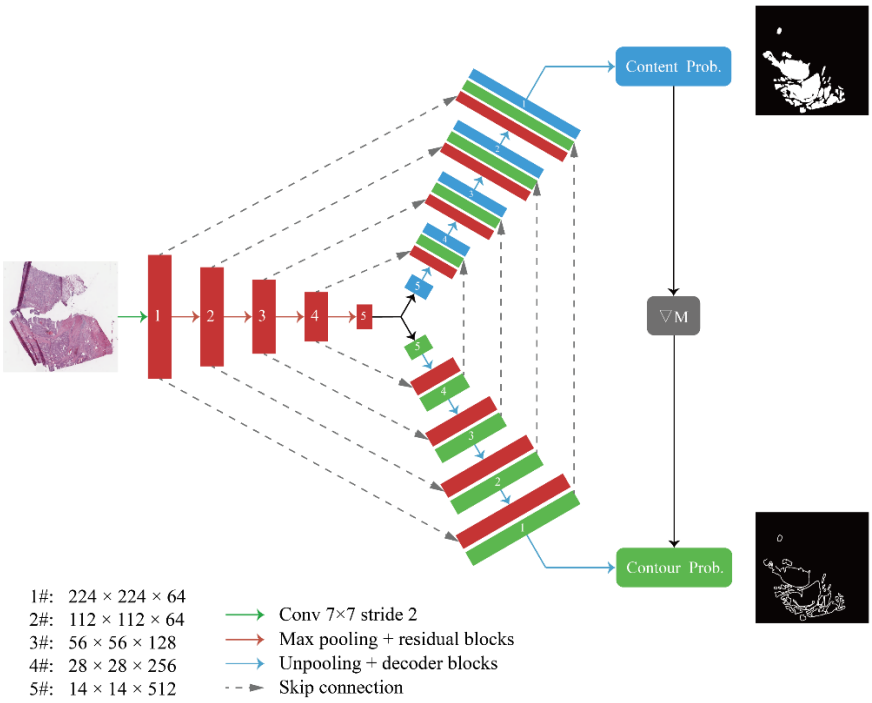

Figure 2. Illustration of the proposed Y-Net. Located below is the joint encoder with an input size of 448px × 448px. The red blocks indicate the feature map output after each down-sampling. The content decoder is in the upper left and the output feature map is marked in blue. The contour decoder is in the upper right and the output feature map is marked in green. Each skip connection sends a feature map from one place to another and performs a concatenating operation on the feature map and the target feature map.

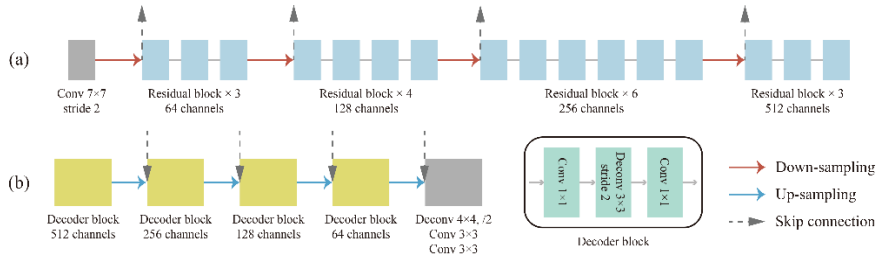

Figure 3. The details of the encoder and decoders. (a) is the concise structure of the encoder, and the residual blocks are the same as the original design in ResNet34. (b) contains the structure of the decoders and the structure of the decoder blocks.

Finally, the loss function of the two tasks is composed by weighted summation, and the objective function is as follows

$$L_{total}(o_1, g_1, o_2, g_2) = L_{content}(o_1, g_1) + \lambda L_{contour}(o_2, g_2) + L_{reg}$$

where L_{reg} is the regularization term to avoid overfitting. λ is the weight coefficient, and the best effect is obtained by taking 1 in the experiment.

3 Results

3.1 Evaluation

We use ACC, AUC and SEN to evaluate the segmentation performance of Y-Net.

(i) Accuracy (ACC)

$$ACC_n = \frac{\sum_{ni}^{h}\sum_{nj}^{w} I(g[ni,nj] = o[ni,nj])}{w * h}$$

$$ACC = \frac{1}{N}\sum_{n=1}^{N} ACC_n$$

where $g[ni,nj]$ means the pixel of $[i,j]$ position in the ground truth of the nth patch, $o[ni,nj]$ means the pixel of $[i,j]$ position in the nth output, N is the total number of test patches.

(ii) Area under curve (AUC)

$$AUC = \frac{1}{N}\sum_{n=1}^{N}\left(\frac{\sum pred_{pos} > \sum pred_{neg}}{positiveNum * negativeNum}\right)$$

The essence of AUC is to randomly select a positive sample (pixel) and a negative sample (pixel), and then use the trained classifier to predict the two samples. AUC is such a probability that the probability of positive samples is greater than that of negative samples.

(iii) Sensitivity (SEN)

$$SEN_n = \frac{\sum_{ni}^{h}\sum_{nj}^{w} I(g[ni,nj] = 1, o[ni,nj] = 1)}{\sum_{ni}^{h}\sum_{nj}^{w} I(g[ni,nj] = 1)}$$

$$SEN = \frac{1}{N}\sum_{n=1}^{N} SEN_n$$

Sensitivity measures the ability to distinguish positive samples (pixels) effectively.

3.2 Controlled experiment

We designed a control experiment to verify the effectiveness of the main ideas of Y-Net. There are three variables that need to be validated: whether to add a contour segmentation task (CT), whether to set the skip connections from the contour decoder to the content decoder (SC), and whether to use the spatial

gradient ∇M fusion (∇M) to fuse the information of the content decoder in the contour decoder.We have conducted experiments with or without the techniques developed in this paper for a comparative study. After training 100 epochs on the unified dataset, the scores on the test set are shown in Table 1. From the table we can see that Y-Net (YN) which has all the characteristics achieves the best results.

Table 1. Experimental results of Y-Net and four variant models.

ID	CT	SC	∇M	AVV	AUC	SEN
YN	Yes	Yes	Yes	0.9820	0.8562	0.9248
M1	Yes	Yes	No	0.9723	0.8323	0.9084
M2	Yes	No	No	0.9614	0.8084	0.8871
M3	No	No	No	0.9642	0.8158	0.8941

3.3 Comparison with the state of art methods

In this section, we compared Y-Net with other classic methods and explored the setting for Y-Net. Since discrete experiments cannot determine the absolute optimal value, we gave a relatively optimal value. As shown in Table 2, when is set as 1, it achieves 0.9820 in ACC, 0.8562 in AUC, 0.9248 in SEN and outperforms other settings displayed in this table.

Table 2. Comparison with other classic methods

Methods	ACC	AUC	SEN
U-Net [6]	0.9649	0.8161	0.8537
DCAN [9]	0.9734	0.8128	0.8198
CE-Net [10]	0.9765	0.8272	0.9020
Y-Net (λ= 0.1)	0.9708	0.8285	0.9066
Y-Net (λ= 0.5)	0.9775	0.8258	0.9154
Y-Net (λ= 1)	0.9820	0.8562	0.9248
Y-Net (λ= 1.5)	0.9799	0.8541	0.9147
Y-Net (λ= 2)	0.9698	0.8395	0.8989

We also compare four sample segmentation results of Y-Net with some of these classic approaches, as shown in Figure 4. In this figure, we randomly choose four samples to perform experiments. From left to right, there are images, the results of U-Net, CE-Net, Y-Net and label. These pictures indicate that our methods can obtain more accurate segmentation results.

3.4 Learning different subtypes of colorectal cancer

Of the 200 cases we obtained, both colon and rectal cancer cases were adenocarcinomas. Adenocarcinoma is divided into several subtypes, such as mucinous adenocarcinoma (mucus outside the cell) and signet-ring cell carcinoma (mucus inside the cell). Among them, mucinous adenocarcinoma is very easy to recognize in morphology, and is very different from other subtypes. In order to study the performance of Y-Net on different subtypes, we extracted 3 cases from mucinous adenocarcinoma and 3 cases from non-mucinous adenocarcinoma. The model predicted all the tiles and finally got more balanced results than the M3 model without the characteristics. The experimental data are shown in Table 3. The experiments have shown that Y-Net can learn contour information of different morphologies (such as boundaries composed of mucus and boundaries composed of tumor cells) and obtain balanced results.

4 Conclusion

Aiming at improving colorectal cancer tumor regions' semantic segmentation, a Y-Net model of which the main idea is using the contour decoder to help the encoder and content decoder to better extract tumor regions' edge information was proposed, thus the segmentation score can be increased. In order to enhance the effect, an improved loss function and the skip connections between two decoders are designed. Experiments show that the design of Y-Net that is based on U-Net can improve its segmentation effect on colorectal cancer, and it can learn the boundaries of different subtypes' tumor regions.

The importance of the tumor regions' contour is demonstrated by this study for semantic segmentation. The preprocessing method of contour masks proposed in this paper is simple and feasible and can be easily adapted to other segmentation tasks. For the further research, we will try other cancer species, test the model using more public datasets, and improve the model to improve the detection effect.

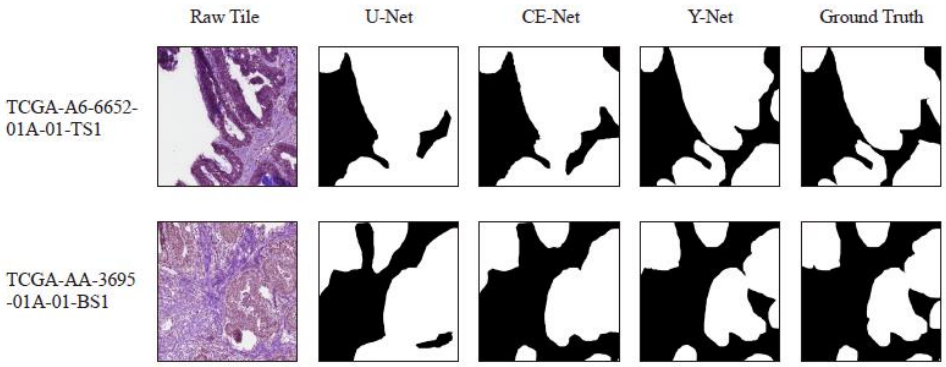

Figure 4. From left to right, there are images, the results of U-Net, CE-Net, Y-Net and label from 2 cases.

Table 3. Experimental results of 6 test cases.

Sample ID	Subtype	Tiles	YN-ACC	M3-ACC	YN-AUC	M3-AUC	YN-SEN	M3-SEN
TCGA-DC-6681-01A	Mucinous	504	0.9788	0.9045	0.8617	0.7946	0.8956	0.8846
TCGA-EI-6507-01A	Mucinous	395	0.9815	0.9284	0.8610	0.8023	0.9221	0.8561
TCGA-AA-3684-01A	Mucinous	79	0.9558	0.9361	0.8574	0.7862	0.9115	0.8504
TCGA-A6-6651-01A	Non-mucinous	142	0.9880	0.9584	0.8418	0.8250	0.9027	0.8495
TCGA-AA-3695-01A	Non-mucinous	78	0.9858	0.9771	0.8851	0.8005	0.8953	0.8606
TCGA-AG-3893-01A	Non-mucinous	63	0.9631	0.9601	0.8625	0.8233	0.9105	0.8558

References

1. Kather J, Pearson A, Halama N, Jäger D, Krause J, Loosen S, Marx A, Boor P, Tacke F, Neumann U, Grabsch H, Yoshikawa T, Brenner H, Chang-Claude J, Hoffmeister M, Trautwein C, Luedde T (2019) Deep learning can predict microsatellite instability directly from histology in gastrointestinal cancer. Nature Medicine 25(7): 1054-1056
2. Ocampo P, Moreira A, Coudray N, Sakellaropoulos T, Narula N, Snuderl M, Fenyö D, Razavian N, Tsirigos, Aristotelis (2018) Classification and mutation prediction from non–small cell lung cancer histopathology images using deep learning. Nature Medicine 24(10): 1559–1567

3. Wang D, Khosla A, Gargeya R, Irshad H, Beck A (2016) Deep learning for identifying metastatic breast cancer. arXiv preprint arXiv.1606.05718
4. Guo Z, Liu H, Ni H, Wang X, Su M, Guo W, Wang K, Jiang T (2019) A Fast and Refined Cancer Regions Segmentation Framework in Whole-slide Breast Pathological Images. Scientific Reports 9(1):882
5. Chen P, Gadepalli K, MacDonald R, Liu Y, Kadowaki S, Nagpal K, Kohlberger T, Dean J, Corrado G, Hipp J, Mermel C, Stumpe M (2019) An augmented reality microscope with real-time artificial intelligence integration for cancer diagnosis. Nature Medicine 25(9):143-1457
6. Ronneberger O, Fischer P, Brox T (2015) U-Net: Convolutional Networks for Biomedical Image Segmentation – MICCAI 2015, vol. 9351, Springer International Publishing, pp 234-241
7. Chen H, Qi X, Yu L, Heng PA (2016) Dcan: deep contour-aware networks for accurate gland segmentation. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp 2487-2496
8. Zhen M, Wang J, Zhou L, Li S, Long Q (2020) Joint semantic segmentation and boundary detection using iterative pyramid contexts. In: 2020 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp 13663–13672
9. Diakogiannis F, Waldner F, Caccetta P, Wu C (2020) ResUNet-a: A deep learning framework for semantic segmentation of remotely sensed data. ISPRS Journal of Photogrammetry and Remote Sensing 16:94–114
10. Gu Z, Cheng J, Fu H, Zhou K, Hao H, Zhao Y, Zhang T, Gao S, Liu J (2019) CE-Net: Context Encoder Network for 2D Medical Image Segmentation. **I**EEE Transactions on Medical Imaging 38(10):2281-2292
11. Zhou Z, Rahman Siddiquee M, Tajbakhsh N, Liang J (2018) UNet++: A Nested U-Net Architecture for Medical Image Segmentation. 4th Deep Learning in Medical Image Analysis (DLMIA) Workshop 11045:3-11
12. Ke R, Bugeau A, Papadakis N, Schuetz P, Schönlieb C (2019) A multi-task U-net for segmentation with lazy labels. arXiv preprint arXiv.1906.12177
13. He K, Zhang X, Ren S, Sun J (2016) Deep Residual Learning for Image Recognition. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp 770-778