Deep learning Instance Segmentation on Esophageal Squamous Cell Carcinoma detection

Xintong Ren[1], Zhen Zhang[2], Junchao Jiang, Guodong Li[2], Jiahua Jiang[3] and Xinyong Jia[2]

[1]Lily AI
[2]The First Affiliated Hospital of Shandong First Medical University & Shandong Provincial Qianfoshan Hospital
[3]University of Birmingham

ABSTRACT

Background

Esophageal cancer (EC) is one of the most common cancer in the world and is the sixth leading cause of cancer-related mortality, where nearly half of the new EC cases are found in China, and esophageal squamous cell carcinoma (ESCC) is the preponderant type. Detection of early EC is essential to improving a patient's survival. However, EC is a highly aggressive cancer with a poor prognosis, even for experienced endoscopists. Therefore, Computer-aided diagnosis (CAD) has become an important method to help endoscopists accurately and rapidly detect early detection.

Methods

We proposed and developed a computer-assisted diagnosis AI system for early ESCC detection using advanced deep learning algorithms on the accurate endoscope magnified narrow-band imaging (ME-NBI) and the most referenced endoscope Lugol's chromoendoscopy (LCE), which aimed to improve the detection on early ESCC.

Results

All 3,500 images were collected from 800 enrolled patients. The CAD AI system contains 24 experiments and is able to detect the tumor region on both ME-NBI and LCE. Our proposed approach outperforms the current state-of-the-art and achieves 0.9236 mAP(mean average precision) on ME-NBI and 0.8620 mAP on LCE.

Conclusions

The good diagnostic performance of our ESCC AI system has proven the feasibility of deep learning for detecting ESCC in both ME-NBI and LCE and shows great potential in applying the AI system in diagnosing the early stages of EC. Furthermore, collecting and analyzing the data of the patients from wider areas to achieve real-time performance for clinical applications have been added in our future plan.

Keywords: Deep Learning, Instance Segmentation, Object Detection, Gastroenterology, Esophageal Squamous Cell Carcinoma, Esophageal cancer, ME-NBI, LCE, InstaBoost, Deformable Convolutional Networks, Cascade Head RoI, Endoscope

Correspondence to
Junchao Jiang;
jiangjunchao0830@yahoo.com

Xinyong Jia;
jiaxy0505@126.com

INTRODUCTION

Esophageal cancer (EC) is one of the most common cancers in the world and is the sixth leading cause of cancer deaths. There are two predominant histological types of esophageal cancer, esophageal adenocarcinoma (EAC) and esophageal squamous cell carcinoma (ESCC). Almost half of the new EC cases are found in China, most of which are ESCC type. Since ESCC develops in stages and progressively becomes more difficult to cure, early detection is crucial for improving the survival chances of the patients. If detected at an early stage, the overall five-year survival rate is over 80% or greater [1, 2] because cancerous tissue can be removed endoscopically. However, the five-year survival rate for late detection drops to only 20% [3]. This profound improvement in the survival rate motivates the need for effective early detection techniques that enables earlier diagnosis and curative treatment.

However, an easily overlooked case has been emphasized that accurate detection highly relies on experienced endoscopists. Even for advanced endoscopists, it is challenging to distinguish intramucosal neoplasias, especially when they appear as small erosions, flat mucous membranes, or normal mucosa in the biopsy specimen detection. Computer-Aided Diagnosis (CAD) is increasingly important, and much recent effort in this strategy has been rewarded with many promising devel-

XR, ZZ, JJ and GL contributed equally.

opments. Studies in CAD involve several endoscopes, such as white light endoscopy (WLE), narrow-band imaging (NBI), volumetric laser endomicroscopy (VLE), and confocal laser endomicroscopy (CLE), and chromoendoscopy. CAD systems have been widely applied to the analysis of esophageal adenocarcinoma (EAC) [4]. In this work, we propose a novel method for automatic detection of early ESCC using advanced deep neural networks on object segmentation.

Instead of using only one type of endoscope, we apply both magnifying endoscopy narrow-band imaging (ME-NBI) and Lugol's chromoendoscopy (LCE). The randomized controlled studies in [5] show that magnified narrow-band imaging (ME-NBI) is more accurate than the conventional white light endoscopic EEC method since ESCC is challenging to be detected under traditional white light endoscopy (WLE). Lugol's chromoendoscopy (LCE) is a widely used technique to detect esophageal neoplasia in patients with prior ESCC [5]. Another benefit is that the sensitivity of LCE is higher than the HD-WLE (>95%) [6].

This paper thus uses the CAD depiction of the cancerous area on the ME-NBI and LCE to interpret and identify the cancerous and non-cancerous areas, providing objective assistance for diagnosing early esophageal cancer. Furthermore, building a real-time diagnostic support system will be of great help to endoscopists in detecting ESCC. Our results indicate that using a deep learning system to conduct a preliminary study on the endoscopic detection of the ESCC boundary is promising to be applied clinically.

Related Work

In this section, we give a brief review of the related study on esophageal cancer detection. Sestio and Sommen proposed a method to evaluate the different texture features extracted from HD-WLE Barrett esophagus images. The main features include Texture spectrum, Histogram of Oriented Gradient (HOG), Local Binary Pattern (LBP), Gray Level Co-occurrence Matrix (GLCM), Fourier Features, Dominant Neighbor Structure (DNS), and Gabor features. [7] Given the texture features mentioned above, they applied principal component analysis (PCA) in dimension reduction and then Support Vector Machine (SVM) on classification. The combination between Gabor and color features has achieved the best results with 0.86 sensitivity and 0.87 specificity on SVM, plus 0.90 and 0.75 on Random Forest. [8]

Furthermore, deep learning approaches for Esophageal cancer detection have started trending in recent years. Different studies have shown promising results on Esophageal abnormality detection based on various Convolutional Neural Networks (CNNs).

Ghatway et al. adapted and evaluated esophageal adenocarcinoma regions' performance using state-of-the-art deep learning object detection methods like R-CNN, Fast R-CNN, Faster R-CNN, and SSD. Their evaluation used a dataset composed of 100 HD-WLE images from Endoscopic Vision Challenge MICCAI 2015. The SSD method stood out and achieved a 0.70 Average Precision Rate (APR) and 0.79 Average Recall Rate (ARR) on the bounding box level. The Faster R-CNN proved to be another leading and has achieved 0.68 APR and 0.83 ARR. [9]

Ghatway et al. also proposed a new approach in 2019. They built a novel deep learning model based on Faster R-CNN incorporated with the Gabor features in their study. The same dataset, MICCAI 2015, has been used in this study, and at a measurement of 0.5 level IOU, the model achieved a promising 0.84 mAP (mean average precision). [10]

Hashimoto et al. developed and designed a two-step CNN model with a first Xception architecture binary classification model, which assesses the presence of any neoplastic on the image, along with a second YOLO object detection model to locate the lesion with a predicted bounding box. Their study worked on 916 images from 65 histology-proven early esophageal neoplasia patients. The object detection step measured the mAP (mean average precision) at 0.3 threshold IOU and achieved 0.7533. [11]

Wang et al. developed an AI system using the single-shot multibox detector convolutional neural network to diagnose squamous esophageal cancer. 498 white-light imaging (WLI) and 438 narrowband imaging (NBI) images were used as training images, which included three classifications: low-grade squamous dysplasia, high-grade squamous dysplasia, and squamous cell carcinoma, based on the pathological diagnosis. Their system showed good diagnostic performance in detecting and differentiating esophageal neoplasms, and the accuracy can achieve 90%. [12]

METHODOLOGY

Experimental dataset

The data we used in this study were collected from patients with early esophageal cancer from January 2010 to January 2020 in the First Affiliated Hospital of Shandong First Medical University (Shandong Provincial Qianfoshan Hospital).

The Department of Endoscopy and Diagnosis in the First Affiliated Hospital of Shandong First Medical University (Shandong Provincial Qianfoshan Hospital) is among the oldest primary and professional units of the digestive endoscopy department in Shandong Province.

The Department was founded in 2007. Mainly responsi-

XR, ZZ, JJ and GL contributed equally.

ble for the diagnosis and treatment of endoscopy in the outpatient and ward departments of the hospital. The total area is about 2500 square meters. Standardized and scientific clinical and procedural diagnosis and treatment technology, a first-class academic team and advanced international equipment ensure the most standardized achievement of endoscopic medical level and work safety. Our Department has nearly fifty staff, including fifteen endoscopists, six anesthesiologists, and twenty-three staff in nursing teams.

In this work, we mainly focus on the endoscopic images of the patients with early-stage esophageal cancer and define them as high-grade dysplasia or T1 stage adenocarcinoma. The endoscopes used during this period are Olympus series upper endoscopes (260 HQ and 290 HQ; Olympus, Tokyo, Japan), which only include images obtained through conventional endoscopes using standard white light. One image is regarded as non-abnormity if all of the following three criteria are satisfied:

1) The patient has no Esophageal lesions found before the examination.

2) Confirmation from the endoscopists with more than ten years of endoscopy experience.

3) Unanimity needs to be attained with non-abnormity.

Most lesion areas are extremely small, and the size of lesions ranges from 3 mm to 20 mm, making it hard for less experienced endoscopists to keep 100% success in lesion detection. In this pilot study, instead of considering the cases of low-grade dysplasia, we select a large diversity of dysplastic lesion cases for training to optimize the algorithm. Three thousand five hundred images were collected from 800 patients, including numerous images of abnormal hyperplasia confirmed by ESCC histology, with the size of the images 1920*1080.

All the patients with abnormal endoscopic images need to undergo an endoscopic submucosal dissection (ESD) procedure and have obtained corresponding NBI imaging and pathological results, including tumor edge, tumor invasion depth, lymphatic and vascular invasion. Three gastrointestinal pathologists (two with over ten years of experience in pathology and one with over fifteen years experience) completed histological evaluations in the pathology department. The manifestations of ESCC include high-grade intraepithelial neoplasia and neoplasm that invades the mucosa or superficial submucosa (SM1). Three endoscopists with over fifteen years of experience in surgical endoscopy annotated the endoscopy images. In this study, standard single-accessory channel endoscopes (GIF-Q260 and GIF-Q290, Olympus, Tokyo, Japan) and captured mucosal images were mainly used in the inspection process. The annotation process is in the following:

1) Based on the patient's pathological results, the first endoscopist reviewed the Gastroscopy image and produced the annotation(mark the lesion as a polygon).

2) The second endoscopist checked the annotation results and made additional notes from the first afterward.

3) If there is no disagreement, the annotation will be treated as complete. Otherwise, the third endoscopist will join in, and the three will discuss together to make the ultimate decision.

For ME-NBI, we marked the lesion object in two types of approaches:

1) Mark the Real boundary of lesions; label region only covered clear boundaries of lesion areas. We called it the Real boundary class, with 1461 training images and 365 testing images.

2) Mark the Distinct boundary of lesions; label region covering all lesion areas, including blurred borders. We called it the Distinct boundary class, with 1180 training images and 295 testing images.

In ME-NBI imaging, borders of the lesion area sometimes show as blurred compared with the center area. A couple of different reasons could cause blurred borders: 1) hard to differentiate the boundaries between the lesion and normal region when annotation. 2) caused by the imaging process. In deep learning frameworks, the vague area across the borders might show lower confidence. Therefore, We defined the lesion classes in both the Real and Distinct types, and in the Real type, we excluded the vague area. The main benefit of incorporating various types of classes is to help us to identify the border of the lesion area more clearly since the scenario that blurred borders caused by the imaging reconstruction error can be excluded.

We marked the lesion object for LCE by marking the boundary of the iodine-stained lesion regions. We called it the iodine-staining class, with 678 training images and 169 testing images.

When splitting the training and test data sets, we measured the size of the labeled lesion objects to ensure there were no biases. The lesion size (width/length/area) distributions between the training and test set should not differ too much.

Since our data input is only the image itself, compared with the conventional multi-variable prediction models, we do not need to handle problems caused by the missing data. In addition, we do not need to concern about the unadjusted/adjusted association between predictor candidate and outcome.

Besides, the demographic details of the enrolled patients in our study are listed below:

1) The age range of patients is from 35 to 75 years old.

2) The average age of patients is 59 years old.

3) The ratio between men and women is 1.15:1.

Deep learning models

Instance segmentation of natural images is a long-existing fundamental and challenging computer vision problem. Since 2019, there have been published explosive algorithms in this field. For instance, Deformable Convolutional Networks (DCN) [13] realized a free-form deformed sampling grid based on variable convolution and variable RoI pooling. Based on the image random augmentation technique inspired by the stochastic grammar of images, random InstaBoost [14] achieved effective mAP improvement. Furthermore, Cascade RCNN [15], based on multistage target detection architecture, consists of a series of detectors to increase IoU training, improve the hypothetical quality, and declare all detectors of the same size positive training set and minimize the overfitting. Besides, more image instance segmentation methods have been published in the last two years, such as the Hybrid Task Cascade series [16], Mask Scoring R-CNN [17], and the GCNet [18]. This study will implement these approaches to ESCC detection and evaluate the overall results plus overall comparisons.

InstaBoost

InstaBoost augments the training set by using existing instance mask annotations. It improved the training performance by using the pixel redundancy of the background and by simply introducing random jittering to objects without modifying the backbone or network structure. The main idea of random InstaBoost in [14] is to separate the foreground and background with annotations using the matting and inpainting and apply a random transform to generate more diverse augmented images. In addition, InstaBoost is a location probability map-based method to explore possible locations where objects can be placed based on local appearance similarity. InstaBoost can be integrated into the training pipeline for any instance segmentation model and will not affect training and inference efficiency, which is easy to implement and will not increase computational complexity.

Suppose (x0, y0) is the object's original coordinate and (x, y) is the newly transformed coordinate. In order to address the problem that the feasible transformation of (x, y) is restricted in the neighborhood of (x0, y0), the Appearance consistency heatmap was proposed in [14] to refine the position by pasting the new instance. The appearance descriptor encodes the texture of the object neighborhood area to measure the similarity of the object patch based on two locations. The descriptor is defined as the weighted combination of three fixed-width contour areas with different scales. The appearance distance is a metric between pairs of appearance descriptors. By scanning all feasible descriptors and the appearance distance in the image, a heatmap is produced and sampled via the Monte Carlo method to get the pasted position.

Deformable Convolutional Networks (DCN

The keynote of DCN [13] is to introduce two new modules to enhance the transformation modeling capabilities of CNN, namely deformable convolution and deformable RoI pooling. Additional offset is used to augment the spatial sampling locations in the modules and learn the offsets from the target tasks without additional supervision. The new module can easily replace the standard modules in the existing CNN instead of only doing feature sampling from Rectangles. The end-to-end training through standard back-propagation can be easily implemented.

[Figure 1 about here.]

The convolution kernel has the exact spatial resolution and expansion as the current convolution layer; The output offset field has the exact spatial resolution as the input feature map; The channel dimension 2N corresponds to N 2D offsets.

During training, the convolution kernel and offset used to generate output features are learned simultaneously. In addition, The gradient will be back-propagated through bilinear operations, which also learn the offset.

[Figure 2 about here.]

Figure 2 illustrates how to obtain the offset. First, RoI pooling generates pooled feature maps. From the mapping, by element-wise product and the width and height of the RoI, the FC layer generates and transforms the normalized offset through back-propagation learning. The offset normalization right here is necessary to keep the RoI size unchanged from the offset learning.

Cascade Head RoI

The intersection over union (IoU) threshold is usually used to define positive/negative object detection. The commonly used threshold value is generally 0.5; however, the detection performance often decreases for more significant thresholds. The reason is primarily due to the over-fitting caused by the reduction of the high-threshold positive sample size. According to that, a multi-stage object detection architecture, Cascade R-CNN [15], solved this problem. It is composed of a series of detectors trained with increasing IoU thresholds. The detectors are trained in sequence, and the output of the detector is

XR, ZZ, JJ and GL contributed equally.

used in the following training set. This resampling step ensures that all detectors have the exact size of the positive training set and minimizes overfitting.

The architecture of Cascade R-CNN is shown in Figure 3.

[Figure 3 about here.]

The Cascade R-CNN architecture is a multi-stage extension of the Faster R-CNN architecture as a sub-network from RPN. Moreover, the goal of Cascade R-CNN is not limited to the RPN but to achieve high-quality object detection through the combination of cascaded bounding box regression and cascaded detection.

Evaluation Metrics

In this study, we used the mean Average Precision (mAP) to measure the overall performance of the models. In addition, we provided the average IoU (avgIoU), average Dice (avgDice), Average Recall Rate (ARR), and Average Precision Rate (APR) to evaluate the performance of the model detection in lesion regions segmentation. Besides, True Positive (TP), False Positive (FP), False Negative (FN), Precision, Recall, and F1 Score have been used as evaluation metrics at the object level. Additionally, this paper measured all the metrics at the 0.5 IoU level.

Moreover, we compare the performance of the models on both the Bounding boxes level and pixel-based Polygon level. The detailed explanation for evaluation metrics are listed in Chart 1.

[Chart 1 about here.]

$$
\begin{aligned}
IoU &= \frac{B^g \cap B^p}{B^g \cup B^p} \\
avgIoU &= \frac{1}{N}\sum_{i=1}^{N} IoU \\
avgDice &= \frac{1}{N}\sum_{i=1}^{N} \frac{2\times IoU}{IoU+1} \\
APR &= \frac{1}{N}\sum_{i=1}^{N} \frac{B^g \cap B^p}{B^p} \\
ARR &= \frac{1}{N}\sum_{i=1}^{N} \frac{B^g \cap B^p}{B^g} \\
Precision &= \frac{TP}{TP+FP} \\
Recall &= \frac{TP}{TP+FN} \\
F1 &= \frac{2\times Precision \times Recall}{Precision+Recall}
\end{aligned}
$$

EXPERIMENTAL RESULTS

In this study, we proposed and applied a variety of eight Deep Learning frameworks to the segmentation task of early esophageal cancer (ESCC).

The eight frameworks are constructed by combining multiple recently published advanced modules [19]. We obtain the optimal ESCC detection framework by comparing the performance results.

[Chart 2 about here.]

Besides the training pipeline, we also built the validation for performance results and inference for new instance prediction for each experiment. All eight algorithms are tested on the three lesion classes. The results of the total 24 experiments are shown in Chart 3.

The inference time for one experiment is approximately 0.182 seconds per instance on a single GeForce RTX 2080 Ti GPU.

[Chart 3 about here.]

[Chart 4 about here.]

From Chart 3, the bounding box level outperformed the polygon level, as we expected. The object cropped by the bounding box contains the lesion and some nearby background. Besides, the edge of the lesion is hard to distinguish, especially around some noisy and blurred background. According to these two, compared with tracking the intersection area ratio (aka, average IOU) between labeled and predicted lesions, the corresponding ratio between labeled and predicted bounding boxes is more likely to achieve a higher score. That leads to an outcome from the above performance table: the overall performance of instance segmentation based on Bounding boxes level is better than that of polygon pixel level. However, in terms of reliability, the performance of pixel-level Polygon is more convincing. Since it exactly considers the actual area of the tumor lesion object and excludes any background noise interference.

The model's accuracy under the Real boundary is not as good as the Distinct boundary, mainly because the labeled region of the real boundary excluded the blurred borders. Blurred borders are between the Lesion area and the background, with their characteristics being sometimes difficult to distinguish, which leads to an overall performance decrease.

Iodine staining imaging is currently the most widely used and cheapest imaging for processing esophageal cancer images in the medical field. Unlike the Real and Distinct boundary, the Iodine staining class marks the lesion area after the image is stained. From the results, the model of the iodine staining method showed promising performance overall. A few models have achieved more than 0.85 mAp on 0.5 level IoU. Considering the popularity of Iodine staining, this will likewise play a decisive role in promoting esophageal cancer detection technology in the medical community in the future.

The model with both the Hybrid Task Cascade RoI head and Deformable Convolutional Network stands out from others among all three classes. Additionally, the models using

Instaboost in the data augmentation have achieved up to standard performances.

The results of several examples of the model detection are in Figure 4.

[Figure 4 about here.]

[Figure 5 about here.]

[Figure 6 about here.]

In Figure 4, compared with the ground truth region polygon, the majority of the area has been successfully detected and covered in the predicted region—only minor detection differences at the edge of the top right area. Figure 6 indicates that even for those tumor regions with extreme unsmooth borders, the model still performs well even at the polygon level.

Object detection performance

We also further explored the performance at the object level in terms of the objects' Precision and Recall.

Chart 5&6 show the corresponding Precision, Recall, and F1 Scores for all models across three classes, along with the TP, FP, and FN objects we detected on the test image set.

The performance metrics are calculated from the Bounding boxes instance level and Polygon Masks level.

[Chart 5 about here.]

[Chart 6 about here.]

From Chart 5&6, we can see that the eight models all performed well at the bounding box instance level, which has achieved a more than 80% F1 score. Remarkably, their better performance is shown as a result of the lower FP (False Positive) due to selecting the region of interest (RoI) using the Cascade RoI Head. It avoids the model overfitting on the incremental IoU, which makes the model more robust. Models that used Cascade Head or Hybrid Task Cascade Head RoI are more likely to reach a top F1 score.

Another prominent finding is that all models achieved a more approving score on Recall than on Precision. A better Recall is what we expected, primarily because we plan to use these applications on a computer-aided detection (CAD) system. The system aims to show the prediction results to doctors before approaching the final diagnosis of the tumor regions. Based on the application usage, we assume a better Recall model could ensure a lower missing detection rate as much as possible to avoid the omission and negligence of cancer cases.

Relationship between cancer object size and IoU

We also explore the relationship between the size of the tumor object and the detection performance (we used the IoU metric here) to study if the performance on the object level is related to the object size. The correlation coefficient between the size of the ground truth bounding box (width and height) and their corresponding IoU for all experiments is shown in Chart 7.

[Chart 7 about here.]

Based on the correlation coefficient scores for all the 24 experiments, the correlation scores for the majority of the models from both class Real boundary and class Iodine staining showed "No relationship" ($r < 0.25$). Only a few show a slightly "Weak relationship" ($0.25 < r < 0.5$).

However, for the Distinct boundary class, all the experiments achieve at least a Weak correlation. In order to investigate the Distinct boundary class's differentiation, we further explore the distribution of the ground truth object size (take width as an example):

[Figure 7 about here.]

From Figure 7, compared with the Real boundary and Iodine staining classes, the width distribution on the Distinct boundary class shows as extremely left-Skewed—many objects with sizes exceeding 1000 pixels on the Distinct boundary class significantly. Considering that the effective area on our annotated image is close to 1080 * 1080 pixels, those big objects almost cover the entire effective area.

In this case, the labeled lesions in the image are most likely a portion of the entire lesion. Hence, the predicted object can easily cover the entire effective area, leading to a higher IoU and a higher correlation between those big objects' size and their corresponding IoU.

[Chart 8 about here.]

Based on the above assumption, in Chart 8, we remove all the images with object sizes larger than 1000 pixels and regenerate the correlation comparison table. After that, the correlation between object size and IOU from all the algorithms across three classes shows "No relationship" ($r < 0.25$), which indicates that there is no relation between the object performance and object size.

Resident FN cases

From one of our investigations, many FN objects existed among all types of algorithms and classes. Compared with the various ranges of the FP number, the FN number among all models does not change too much. An interesting finding

XR, ZZ, JJ and GL contributed equally.

is that some specific objects appear as FNs across all models. Therefore, we focused on these FNs and observed some commonalities between the "always FN" objects.

The first "FN" commonality is caused by the RGB pixel contrast between the lesion, and normal regions are not apparent. One of the example images is provided in Figure 8.

[Figure 8 about here.]

The left plot shows the ground truth image where the blue area is the labeled tumor region, and the right one shows the predicted image, where the red area stands for the predicted tumor polygon.

From Figure 6, The extra-predicted lesion area in the red area does not have apparent characteristics from both texture and color differences from the original blue area. Thus, it is hard for the model to distinguish, leading to the original blue region as an FN sample and the red becoming a sample of FP.

[Figure 9 about here.]

In Figure 9, we can see that blurry images cause the second "FN" commonality. It is a typical issue caused by the phenomenon of flickering in the acquisition stage. In this example, blurry appears mainly on the bottom part of the image, which makes the prediction hard to distinguish. In addition, blurry is the most common issue for the "always FN" objects, and we will try to avoid such situations in the future data collection process.

DISCUSSION

In this paper, we have developed a new approach for a Computer-aided diagnosis (CAD) system for the early detection of ESCC. New and advanced object segmentation approaches, including Instaboost, DCN, Cascade ROI, GCnet, etc., are used t o construct and build our deep learning frameworks. The final eight frameworks are tested to be reliable and accurate. The model htc_x101_64x4d_fpn_dconv_c3-c5 has proved to be the leading performance model. At a 0.5 threshold IoU level, the Distinct Boundary lesion ME-NBI has achieved 0.9236 mAP, 0.907 APR, and 0.9173 ARR; the Real Boundary lesion ME-NBI has achieved 0.8511 mAP, 0.9059 APR, and 0.9081 ARR; Iodine Staining lesions has achieved 0.8620 mAP, 0.9386 APR and 0.9433 ARR. We also get promising results on 93.98% Precision and 93.05% Recall for Distinct Boundary lesion ME-NBI; 87.34% Precision and 88.98% Recall for Real Boundary lesion ME-NBI; 82.89% Precision and 88.32% Recall for Iodine Staining lesions. With these results, we have proven the feasibility of deep learning for detecting ESCC in both ME-NBI and LCE.

The data collection process will be optimized in the future to avoid blurry images. Furthermore, we will collect and analyze the patients' data that cover a broader area to achieve real-time performance for clinical applications.

DATA AVAILABILITY STATEMENT

Due to the nature of this research, participants of this study did not agree for their data to be shared publicly, so supporting data is unavailable.

FUNDING

Funding this research was supported by Grants: Jinan healthcare science and technology plan (201907046), Shandong Province key research and development program (2019GSF108028), Youth Natural Science Foundation of Shandong Province (ZR2020QH040)

REFERENCES

1 Wang GQ, Jiao GG, Chang FB, et al. Long-term results of operation for 420 patients with early squamous cell esophageal carcinoma discovered by screening. *Ann Thorac Surg*. 2004;77(5):1740–1744.

2 Yang S, Wu S, Huang Y, et al.. Screening for oesophageal cancer. Cochrane Database Syst Rev.; 2012.

3 Zeng H, Zheng R, Guo Y, et al. Cancer survival in China, 2003-2005: a population-based study. *Int J Cancer*. 2003;136(8):25242378–25242378.

4 Riel SV, Sommen FVD, Zinger S, et al. Automatic Detection of Early Esophageal Cancer with CNNS Using Transfer Learning. *25th IEEE International Conference on Image Processing*. 2018;p. 1383–1387.

5 Lecleire S, Antonietti M, Iwanicki-Caron I, et al. Lugol chromo-endoscopy versus narrow band imaging for endoscopic screening of esophageal squamous-cell carcinoma in patients with a history of cured esophageal cancer: a feasibility study. *Dis Esophagus*. 2011;24(6):21309914–21309914.

6 Protano MA, Xu H, Wang G, et al. Low-Cost High-Resolution Microendoscopy for the Detection of Esophageal Squamous Cell Neoplasia: An International Trial. *Gastroenterology*. 2015;149(2):4547689–4547689.

7 Setio AAA, Sommen, Der FV, et al. Evaluation and comparison of textural feature representation for the detection of early stage cancer in endoscopy. In: Proceedings of the 8th International Conference on Computer Vision Theory and Applications (VISAPP'13). Spain INSTICC Press; 2013. p. 21–24.

8 Markus HA, Janse F, Sommen SVD, et al. Early esophageal cancer detection using RF classifiers. *Proc SPIE 9785*. 2016;p. 97851–97851.

9 Ghatwary N, Zolgharni M, Ye X. Early esophageal adenocarcinoma detection using deep learning methods. *Int J Comput Assist Radiol Surg*. 2019;14(4):6420905–6420905.

10 Ghatwary N, Ye X, Zolgharni M. Esophageal Abnormality Detection Using DenseNet Based Faster R-CNN With Gabor Features. *IEEE Access*. 2019;7:84374–84385.

11 Hashimoto R, Requa J, Dao T, et al. Artificial intelligence using convolutional neural networks for real-time detection of early esophageal neoplasia in Barrett's esophagus (with video). *Gastrointest Endosc*. 2020;91(6):31930967–31930967.

12 Wang YK, Syu HY, Chen YH, et al.. Endoscopic Images by a Single-Shot Multibox Detector for the Identification of Early Cancerous Lesions in the Esophagus: A Pilot Study. Cancers (Basel). PMCID; 2021.

13 Dai J. Deformable Convolutional Networks. *2017 IEEE International Conference on Computer Vision (ICCV*. 2017;p. 764–773.

14 Fang H, Sun J, Wang R, et al. InstaBoost: Boosting Instance Segmentation via Probability Map Guided Copy-Pasting. *IEEE/CVF International Conference on Computer Vision (ICCV)*. 2019;p. 682–691.

15 Cai Z, Vasconcelos N. Cascade R-CNN: Delving Into High Quality Object Detection. *IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2018;p. 6154–6162.

16 Chen K, Pang J, Wang J, et al. Hybrid Task Cascade for Instance Segmentation. *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2019;p. 4969–4978.

17 Huang Z, Huang L, Gong Y, et al. Mask Scoring R-CNN. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2019;p. 6402–6411.

18 Cao Y, Xu J, Lin S, et al. GCNet: Non-Local Networks Meet Squeeze-Excitation Networks and Beyond. *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*. 2019;p. 1971–1980.

19 Chen K, Wang J, Pang J, et al.. Open MMLab Detection Toolbox and Benchmark; 2019. Available from: arXiv: 1906.071552019.

20 He K, Gkioxari G, Dollár P, et al. Mask R-CNN. *2017 IEEE International Conference on Computer Vision (ICCV*. 2017;p. 2980–2988.

21 He K, Zhang X, Ren S, et al. Deep Residual Learning for Image Recognition. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2016;p. 770–778.

XR, ZZ, JJ and GL contributed equally.

LIST OF FIGURES

1 Illustration of deformable convolution 10
2 Illustration of deformable RoI pooling 11
3 Where “Input” as input image; “conv” as backbone convolutions; “pool” as region-wise feature extraction; “H” as network head; “B” as bounding box, and “C” classification. 12
4 A case from the cascade_mask_rcnn_r101_fpn_instaboost model on Real Boundary class with the original image (left), ground truth image (middle) and predicted image (right). 13
5 A case from the cascade_mask_rcnn_x101_32x4d_fpn_dconv_c3-c5 model on Distinct Boundary class with the original image (left), ground truth image (middle) and predicted image (right). 14
6 An Iodine Staining class case from the cascade_mask_rcnn_x101_32x4d_fpn_dconv_c3-c5 model with the original image (left), ground truth image (middle) and predicted image (right). 15
7 Distribution density of object pixel size 16
8 Example for first “FN” commonality 17
9 E [1–21]xample for second “FN” commonality 18

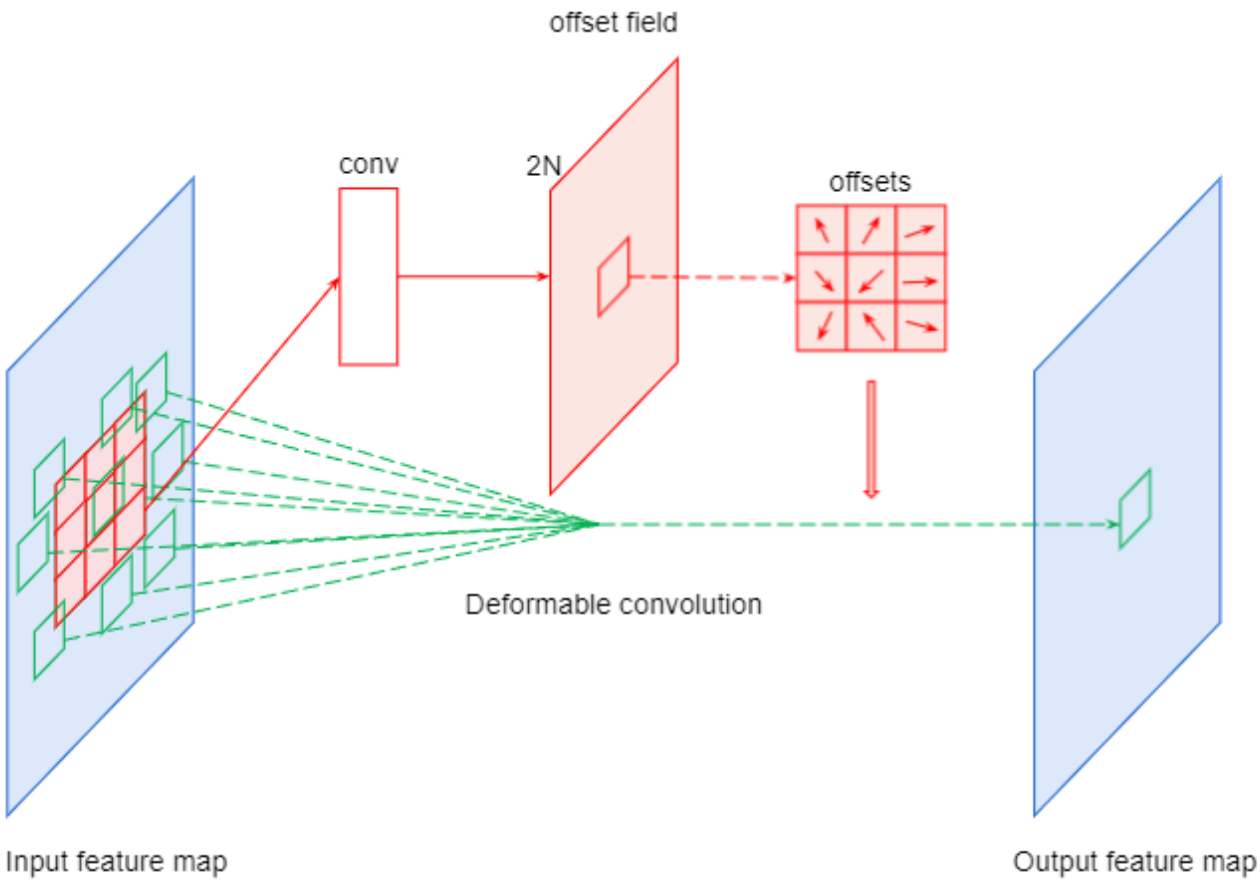

Figure 1 Illustration of deformable convolution

XR, ZZ, JJ and GL contributed equally.

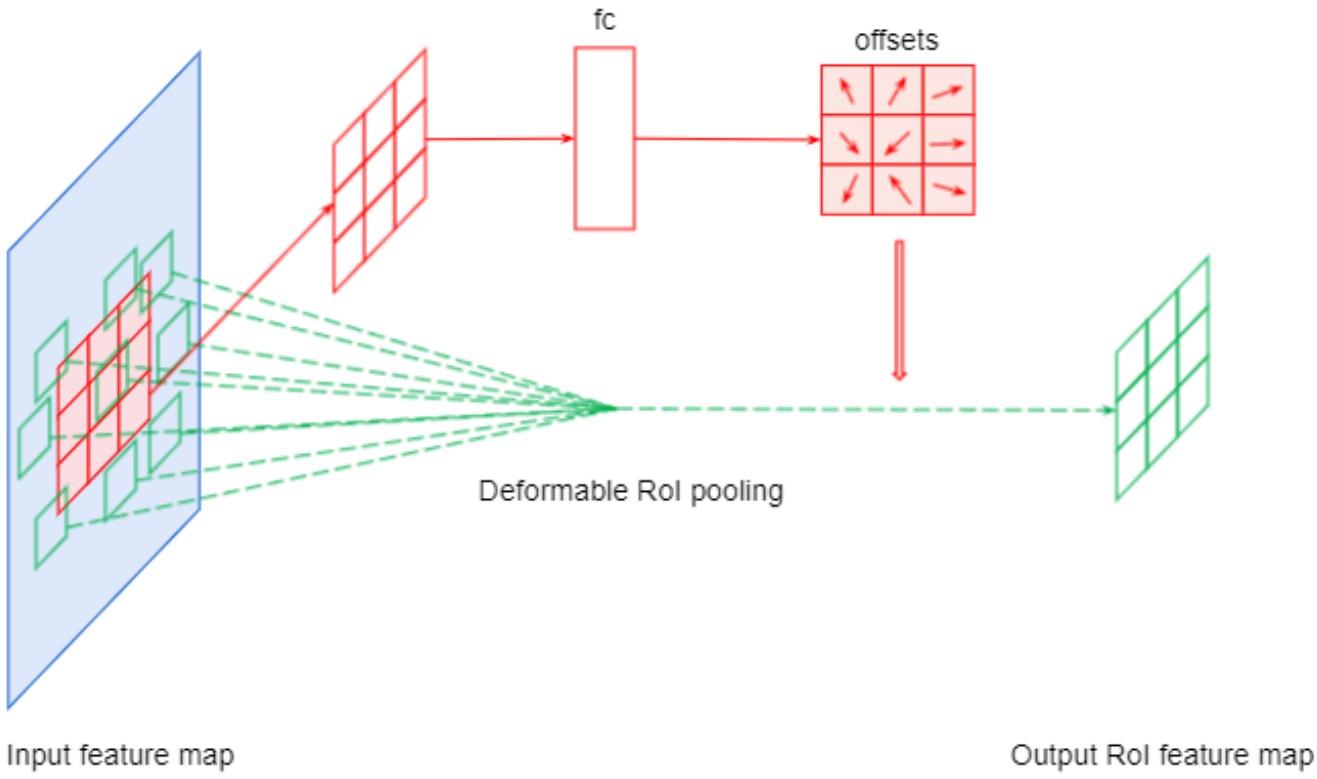

Figure 2 Illustration of deformable RoI pooling

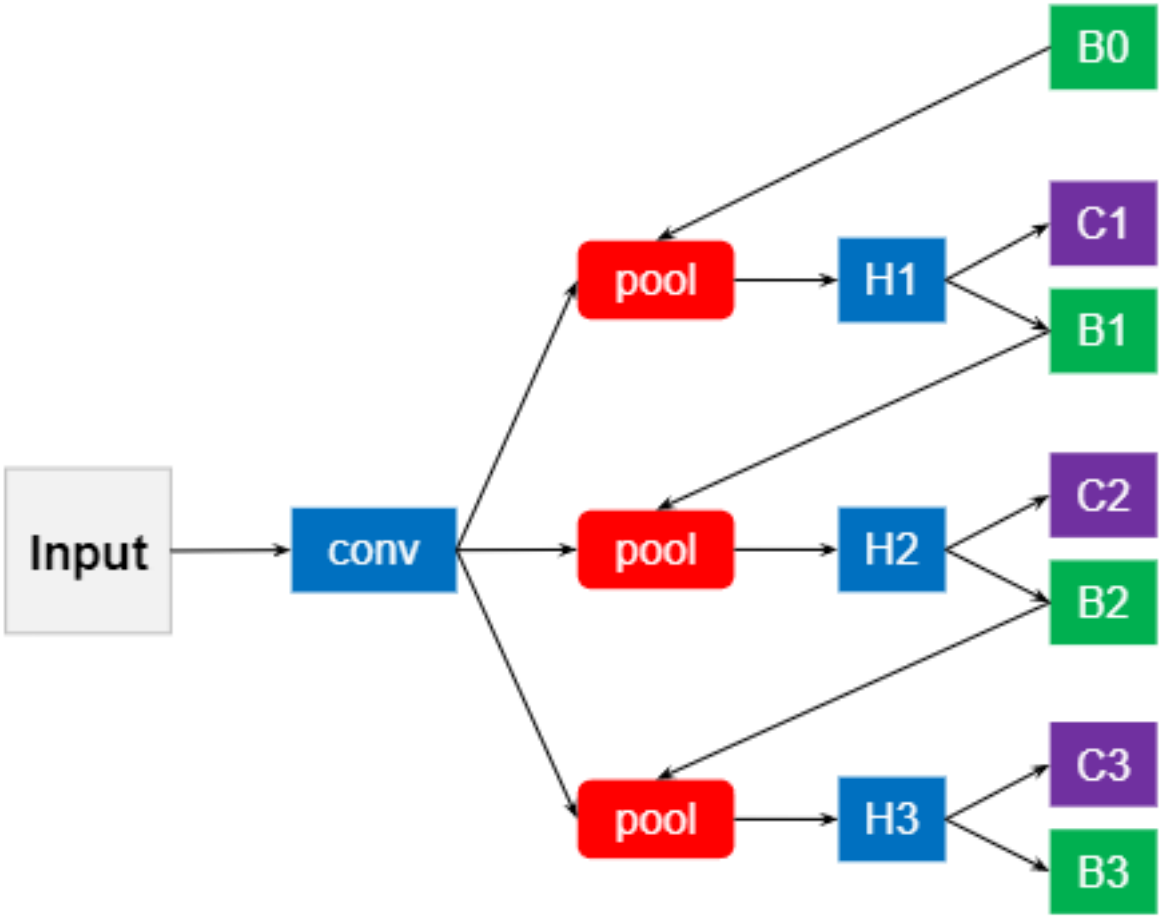

Figure 3 Where “Input” as input image; “conv” as backbone convolutions; “pool” as region-wise feature extraction; “H” as network head; “B” as bounding box, and “C” classification.

XR, ZZ, JJ and GL contributed equally.

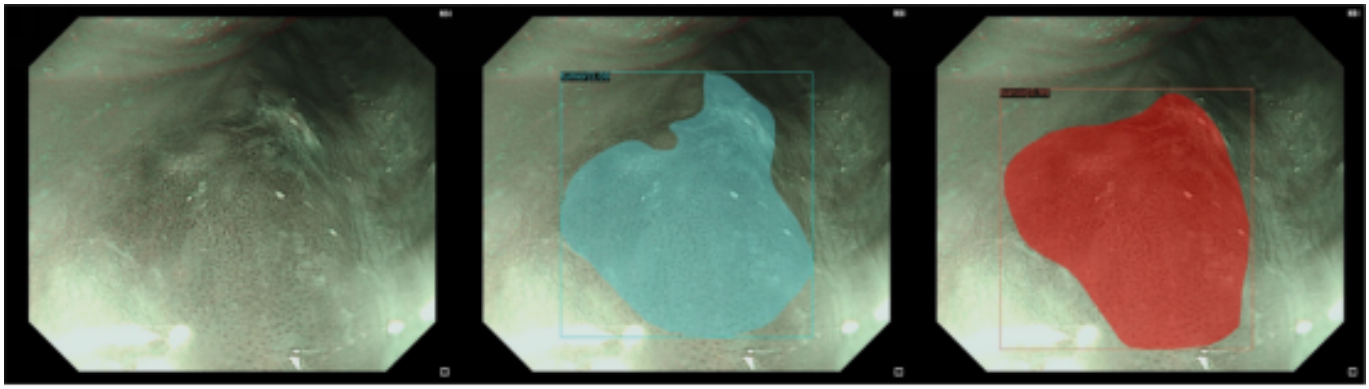

Figure 4 A case from the cascade_mask_rcnn_r101_fpn_instaboost model on Real Boundary class with the original image (left), ground truth image (middle) and predicted image (right).

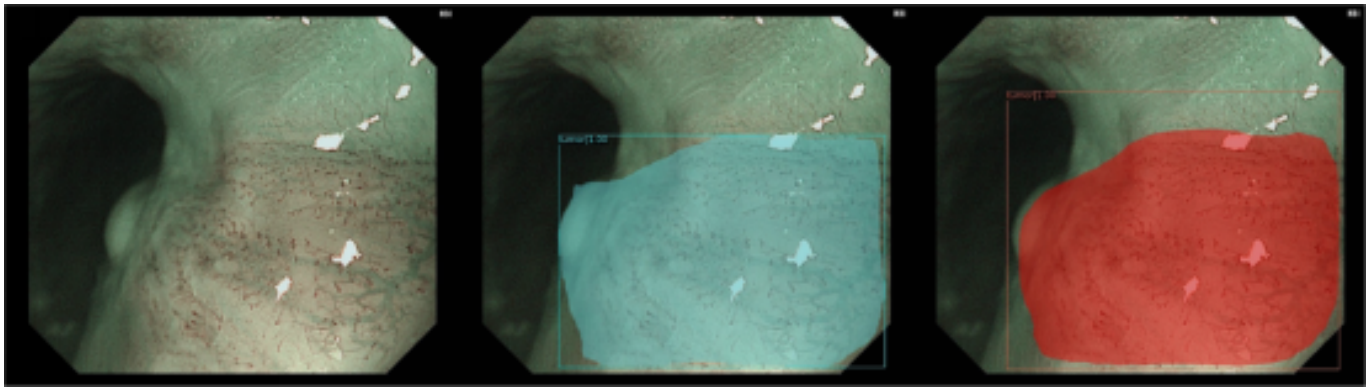

Figure 5 A case from the cascade_mask_rcnn_x101_32x4d_fpn_dconv_c3-c5 model on Distinct Boundary class with the original image (left), ground truth image (middle) and predicted image (right).

XR, ZZ, JJ and GL contributed equally.

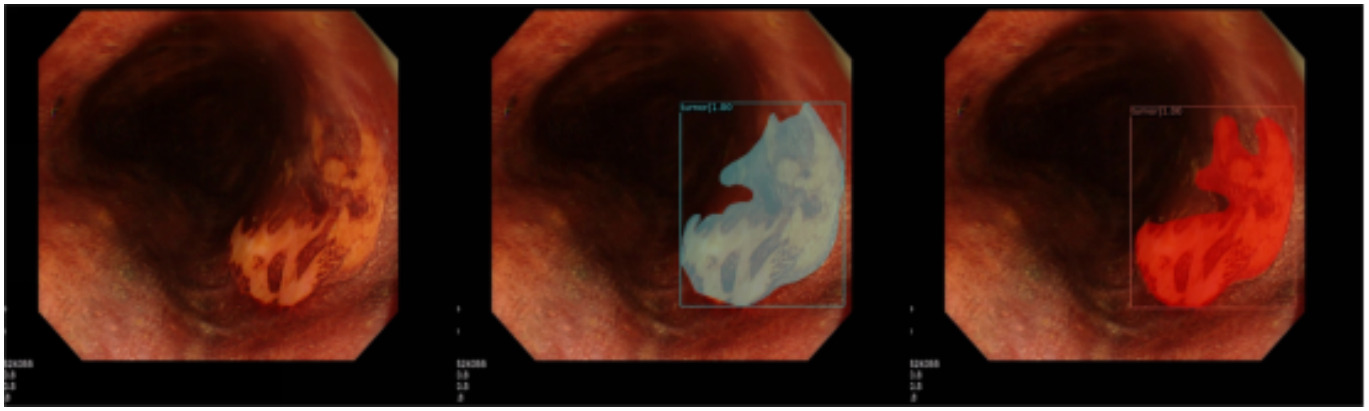

Figure 6 An Iodine Staining class case from the cascade_mask_rcnn_x101_32x4d_fpn_dconv_c3-c5 model with the original image (left), ground truth image (middle) and predicted image (right).

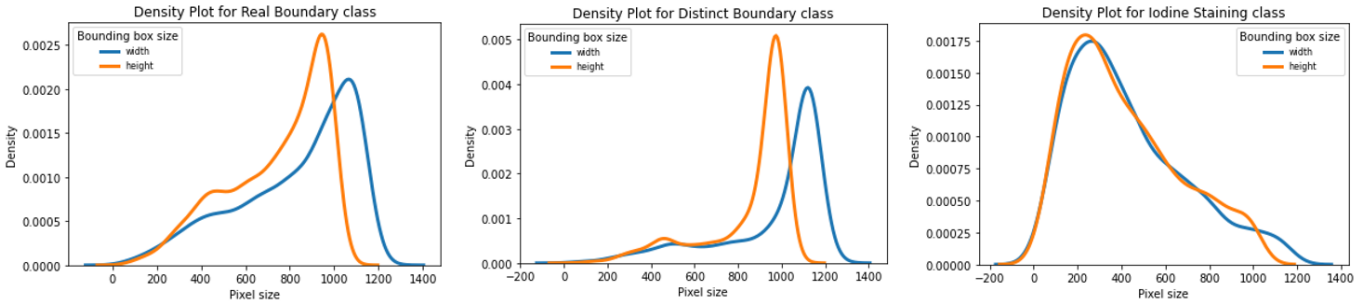

Figure 7 Distribution density of object pixel size

XR, ZZ, JJ and GL contributed equally.

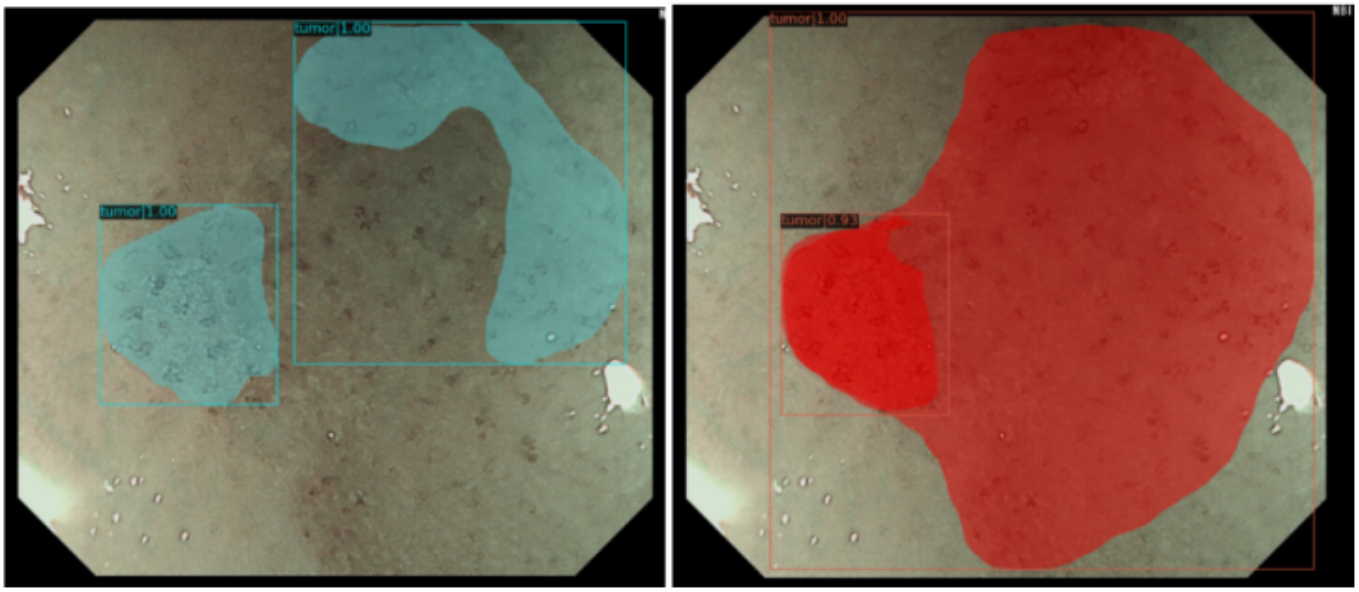

Figure 8 Example for first “FN” commonality

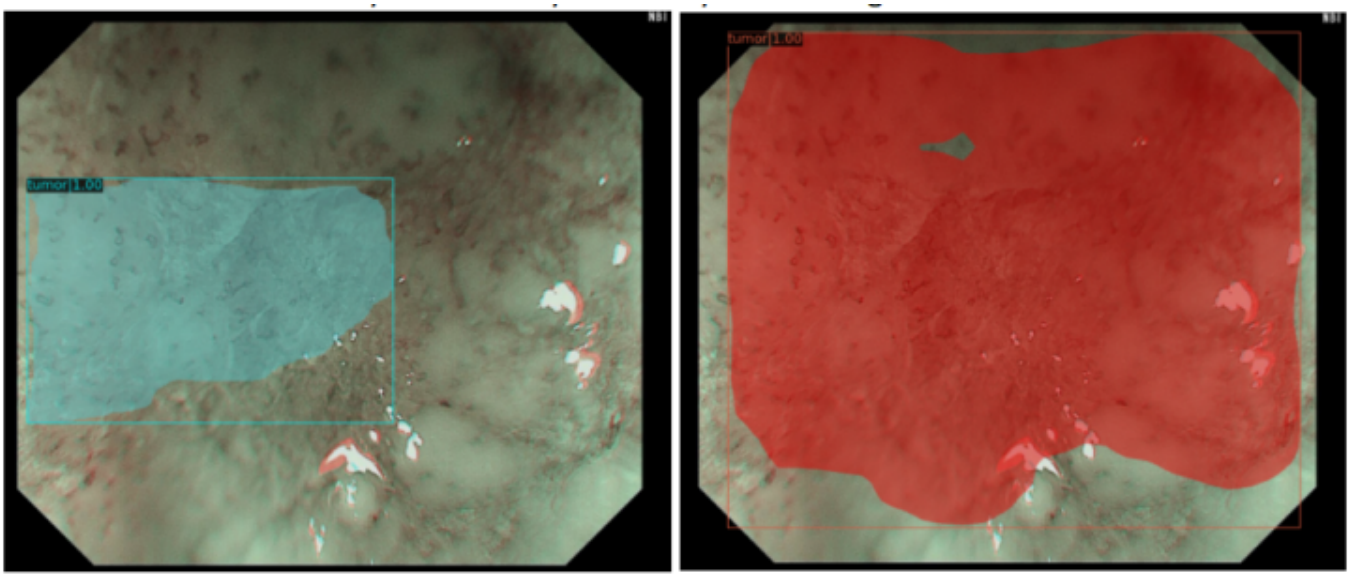

Figure 9 E [1–21]xample for second “FN” commonality

XR, ZZ, JJ and GL contributed equally.

LIST OF CHART

1 Explanation of the metrics 20
2 Abbreviation and the corresponding description used in the model name 21
3 Performance based on Bounding Box instance segmentation 22
4 Performance based on Polygon Masks instance segmentation 23
5 Object level performance based on Bounding Box instance segmentation 24
6 Object level performance based on Polygon Masks instance segmentation 25
7 Correlation between object size and IoU 26
8 Correlation between object size and IoU for object below 1000 pixel 27

N	total number of images
B^g	the ground truth area (bounding box/polygon) of the tumor region
B^p	the predicted area (bounding box/polygon) of the tumor region
mAP	the average of Average Precision (AP), which is finding the area under the precision-recall curve above
TP	count of True Positive tumor object, with a 0.5 IoU threshold
FP	count of False Positive tumor object, with a 0.5 IoU threshold
FN	count of False Negative tumor object, with a 0.5 IoU threshold

Chart 1: Explanation of the metrics

XR, ZZ, JJ and GL contributed equally.

Abbr	Full name
r101	ResNet 101
x101	ResNeXt 101
fpn	Feature Pyramid Network
dconv	Deformable Convolutional Networks
c3-c5	c3-c5 stages of backbone
htc	Hybrid Task Cascade
casacde	Cascade RoI Head
ms_rcnn	MaskScoring RoI Head

Chart 2: Abbreviation and the corresponding description used in the model name

Real Boundary on Bounding Box					
Model	avgIoU	avgDice	APR	ARR	mAP
cascade_mask_rcnn_r101_fpn_instaboost	0.8180	0.8945	0.8955	0.9119	0.8414
cascade_mask_rcnn_x101_32x4d_fpn_dconv_c3-c5	0.8087	0.8890	0.8933	0.9041	0.8356
cascade_mask_rcnn_x101_64x4d_fpn	0.8044	0.8858	0.8718	0.9207	0.7863
ms_rcnn_x101_64x4d_fpn	0.7926	0.8793	0.8780	0.9018	0.8273
htc_x101_64x4d_fpn	0.8141	0.8921	0.8874	0.9149	0.8127
htc_x101_64x4d_fpn_dconv_c3-c5	0.8233	0.8984	0.9060	0.9081	0.8511
mask_rcnn_x101_64x4d_fpn	0.7965	0.8820	0.8798	0.9020	0.8487
htc_x101_64x4d_fpn_dconv_c3-c5_instaboost	0.8202	0.8964	0.8987	0.9124	0.8455
Distinct Boundary on Bounding Box					
Model	avgIoU	avgDice	APR	ARR	mAP
cascade_mask_rcnn_r101_fpn_instaboost	0.8866	0.9354	0.9256	0.9570	0.9191
cascade_mask_rcnn_x101_32x4d_fpn_dconv_c3-c5	0.8773	0.9302	0.9285	0.9432	0.9158
cascade_mask_rcnn_x101_64x4d_fpn	0.8795	0.9315	0.9238	0.9518	0.9211
ms_rcnn_x101_64x4d_fpn	0.8684	0.9257	0.9276	0.9334	0.9117
htc_x101_64x4d_fpn	0.8835	0.9342	0.9270	0.9514	0.9059
htc_x101_64x4d_fpn_dconv_c3-c5	0.8865	0.9363	0.9386	0.9433	0.9236
mask_rcnn_x101_64x4d_fpn	0.8643	0.9224	0.9192	0.9385	0.9189
htc_x101_64x4d_fpn_dconv_c3-c5_instaboost	0.8877	0.9367	0.9371	0.9463	0.9364
Iodine Staining on Bounding Box					
Model	avgIoU	avgDice	APR	ARR	mAP
cascade_mask_rcnn_r101_fpn_instaboost	0.8417	0.9104	0.9083	0.9250	0.8315
cascade_mask_rcnn_x101_32x4d_fpn_dconv_c3-c5	0.8170	0.8958	0.9068	0.8988	0.8042
cascade_mask_rcnn_x101_64x4d_fpn	0.8390	0.9090	0.9144	0.9167	0.8332
ms_rcnn_x101_64x4d_fpn	0.8035	0.8881	0.9026	0.8883	0.8122
htc_x101_64x4d_fpn	0.8305	0.9037	0.9061	0.9155	0.8319
htc_x101_64x4d_fpn_dconv_c3-c5	0.8320	0.9050	0.9070	0.9173	0.8620
mask_rcnn_x101_64x4d_fpn	0.7997	0.8849	0.8938	0.8920	0.8434
htc_x101_64x4d_fpn_dconv_c3-c5_instaboost	0.8332	0.9055	0.8999	0.9237	0.8408

Chart 3: Performance based on Bounding Box instance segmentation

XR, ZZ, JJ and GL contributed equally.

Real Boundary on Polygon Mask					
Model	avgIoU	avgDice	APR	ARR	mAP
cascade_mask_rcnn_r101_fpn_instaboost	0.7704	0.8654	0.8679	0.8845	0.7947
cascade_mask_rcnn_x101_32x4d_fpn_dconv_c3-c5	0.7634	0.8608	0.8796	0.8639	0.7829
cascade_mask_rcnn_x101_64x4d_fpn	0.7429	0.8473	0.8544	0.8638	0.7165
ms_rcnn_x101_64x4d_fpn	0.7440	0.8483	0.8712	0.8489	0.7709
htc_x101_64x4d_fpn	0.7570	0.8570	0.8681	0.8676	0.7606
htc_x101_64x4d_fpn_dconv_c3-c5	0.7778	0.8705	0.8868	0.8747	0.8277
mask_rcnn_x101_64x4d_fpn	0.7666	0.8636	0.8778	0.8673	0.7871
htc_x101_64x4d_fpn_dconv_c3-c5_instaboost	0.7775	0.8703	0.8768	0.8849	0.7915
Distinct Boundary on Polygon Mask					
Model	avgIoU	avgDice	APR	ARR	mAP
cascade_mask_rcnn_r101_fpn_instaboost	0.8619	0.9215	0.9136	0.9406	0.8753
cascade_mask_rcnn_x101_32x4d_fpn_dconv_c3-c5	0.8483	0.9134	0.9177	0.9219	0.8926
cascade_mask_rcnn_x101_64x4d_fpn	0.8537	0.9168	0.9141	0.9319	0.8935
ms_rcnn_x101_64x4d_fpn	0.8445	0.9110	0.9208	0.9128	0.8792
htc_x101_64x4d_fpn	0.8559	0.9183	0.9174	0.9310	0.8795
htc_x101_64x4d_fpn_dconv_c3-c5	0.8511	0.9154	0.9245	0.9181	0.9079
mask_rcnn_x101_64x4d_fpn	0.8439	0.9106	0.9163	0.9172	0.8848
htc_x101_64x4d_fpn_dconv_c3-c5_instaboost	0.8562	0.9182	0.9246	0.9243	0.9087
Iodine Staining on Polygon Mask					
Model	avgIoU	avgDice	APR	ARR	mAP
cascade_mask_rcnn_r101_fpn_instaboost	0.7899	0.8791	0.8892	0.8800	0.8423
cascade_mask_rcnn_x101_32x4d_fpn_dconv_c3-c5	0.7753	0.8701	0.8866	0.8657	0.7960
cascade_mask_rcnn_x101_64x4d_fpn	0.7867	0.8771	0.8990	0.8679	0.8445
ms_rcnn_x101_64x4d_fpn	0.7753	0.8702	0.8933	0.8602	0.8325
htc_x101_64x4d_fpn	0.7897	0.8793	0.8967	0.8738	0.8403
htc_x101_64x4d_fpn_dconv_c3-c5	0.7933	0.8817	0.8984	0.8756	0.8501
mask_rcnn_x101_64x4d_fpn	0.7744	0.8695	0.8944	0.8571	0.8545
htc_x101_64x4d_fpn_dconv_c3-c5_instaboost	0.7914	0.8805	0.8889	0.8832	0.8295

Chart 4: Performance based on Polygon Masks instance segmentation

Real Boundary on Bounding Box						
Model	TP	FP	FN	Precision	Recall	F1
cascade_mask_rcnn_r101_fpn_instaboost	325	58	47	0.8486	0.8737	0.8609
cascade_mask_rcnn_x101_32x4d_fpn_dconv_c3-c5	328	59	44	0.8475	0.8817	0.8643
cascade_mask_rcnn_x101_64x4d_fpn	312	62	60	0.8342	0.8387	0.8365
ms_rcnn_x101_64x4d_fpn	323	68	49	0.8261	0.8683	0.8467
htc_x101_64x4d_fpn	320	54	52	0.8556	0.8602	0.8579
htc_x101_64x4d_fpn_dconv_c3-c5	331	48	41	0.8734	0.8898	0.8815
mask_rcnn_x101_64x4d_fpn	331	75	41	0.8153	0.8898	0.8509
htc_x101_64x4d_fpn_dconv_c3-c5_instaboost	330	68	42	0.8291	0.8871	0.8571
Distinct Boundary on Bounding Box						
Model	TP	FP	FN	Precision	Recall	F1
cascade_mask_rcnn_r101_fpn_instaboost	280	24	22	0.9211	0.9272	0.9241
cascade_mask_rcnn_x101_32x4d_fpn_dconv_c3-c5	279	24	23	0.9208	0.9238	0.9223
cascade_mask_rcnn_x101_64x4d_fpn	280	22	22	0.9272	0.9272	0.9272
ms_rcnn_x101_64x4d_fpn	278	34	24	0.8910	0.9205	0.9055
htc_x101_64x4d_fpn	276	25	26	0.9169	0.9139	0.9154
htc_x101_64x4d_fpn_dconv_c3-c5	281	18	21	0.9398	0.9305	0.9351
mask_rcnn_x101_64x4d_fpn	281	33	21	0.8949	0.9305	0.9123
htc_x101_64x4d_fpn_dconv_c3-c5_instaboost	286	19	16	0.9377	0.9470	0.9423
Iodine Staining on Bounding Box						
Model	TP	FP	FN	Precision	Recall	F1
cascade_mask_rcnn_r101_fpn_instaboost	181	39	33	0.8227	0.8458	0.8341
cascade_mask_rcnn_x101_32x4d_fpn_dconv_c3-c5	178	35	36	0.8357	0.8318	0.8337
cascade_mask_rcnn_x101_64x4d_fpn	183	29	31	0.8632	0.8551	0.8592
ms_rcnn_x101_64x4d_fpn	185	62	29	0.7490	0.8645	0.8026
htc_x101_64x4d_fpn	183	33	31	0.8472	0.8551	0.8512
htc_x101_64x4d_fpn_dconv_c3-c5	189	39	25	0.8289	0.8832	0.8552
mask_rcnn_x101_64x4d_fpn	190	69	24	0.7336	0.8879	0.8034
htc_x101_64x4d_fpn_dconv_c3-c5_instaboost	184	38	30	0.8288	0.8598	0.8440

Chart 5: Object level performance based on Bounding Box instance segmentation

XR, ZZ, JJ and GL contributed equally.

Real Boundary on Polygon Mask						
Model	TP	FP	FN	Precision	Recall	F1
cascade_mask_rcnn_r101_fpn_instaboost	310	73	62	0.8094	0.8333	0.8212
cascade_mask_rcnn_x101_32x4d_fpn_dconv_c3-c5	312	75	60	0.8062	0.8387	0.8221
cascade_mask_rcnn_x101_64x4d_fpn	290	83	82	0.7775	0.7796	0.7785
ms_rcnn_x101_64x4d_fpn	309	81	63	0.7923	0.8306	0.8110
htc_x101_64x4d_fpn	306	69	66	0.8160	0.8226	0.8193
htc_x101_64x4d_fpn_dconv_c3-c5	323	56	49	0.8522	0.8683	0.8602
mask_rcnn_x101_64x4d_fpn	315	85	57	0.7875	0.8468	0.8161
htc_x101_64x4d_fpn_dconv_c3-c5_instaboost	313	85	59	0.7864	0.8414	0.8130
Distinct Boundary on Polygon Mask						
Model	TP	FP	FN	Precision	Recall	F1
cascade_mask_rcnn_r101_fpn_instaboost	270	34	32	0.8882	0.8940	0.8911
cascade_mask_rcnn_x101_32x4d_fpn_dconv_c3-c5	275	28	27	0.9076	0.9106	0.9091
cascade_mask_rcnn_x101_64x4d_fpn	274	27	28	0.9103	0.9073	0.9088
ms_rcnn_x101_64x4d_fpn	271	37	31	0.8799	0.8974	0.8885
htc_x101_64x4d_fpn	270	31	32	0.8970	0.8940	0.8955
htc_x101_64x4d_fpn_dconv_c3-c5	278	22	24	0.9267	0.9205	0.9236
mask_rcnn_x101_64x4d_fpn	273	40	29	0.8722	0.9040	0.8878
htc_x101_64x4d_fpn_dconv_c3-c5_instaboost	281	25	21	0.9183	0.9305	0.9243
Iodine Staining on Polygon Mask						
Model	TP	FP	FN	Precision	Recall	F1
cascade_mask_rcnn_r101_fpn_instaboost	183	35	31	0.8394	0.8551	0.8472
cascade_mask_rcnn_x101_32x4d_fpn_dconv_c3-c5	177	36	37	0.8310	0.8271	0.8290
cascade_mask_rcnn_x101_64x4d_fpn	185	25	29	0.8810	0.8645	0.8726
ms_rcnn_x101_64x4d_fpn	189	50	25	0.7908	0.8832	0.8344
htc_x101_64x4d_fpn	184	31	30	0.8558	0.8598	0.8578
htc_x101_64x4d_fpn_dconv_c3-c5	187	41	27	0.8202	0.8738	0.8462
mask_rcnn_x101_64x4d_fpn	192	60	22	0.7619	0.8972	0.8240
htc_x101_64x4d_fpn_dconv_c3-c5_instaboost	182	34	32	0.8426	0.8505	0.8465

Chart 6: Object level performance based on Polygon Masks instance segmentation

Correlation between object's size and IoU in Real Boundary		
Model	corr(IoU,width)	corr(IoU,height)
cascade_mask_rcnn_r101_fpn_instaboost	0.1581	0.2096
cascade_mask_rcnn_x101_32x4d_fpn_dconv_c3-c5	0.1077	0.2271
cascade_mask_rcnn_x101_64x4d_fpn	0.1434	0.3282
ms_rcnn_x101_64x4d_fpn	0.0124	0.2367
htc_x101_64x4d_fpn	0.1424	0.2908
htc_x101_64x4d_fpn_dconv_c3-c5	0.0423	0.2159
mask_rcnn_x101_64x4d_fpn	0.1377	0.3619
htc_x101_64x4d_fpn_dconv_c3-c5_instaboost	0.0925	0.2311
Correlation between object's size and IoU in Distinct Boundary		
Model	corr(IoU,width)	corr(IoU,height)
cascade_mask_rcnn_r101_fpn_instaboost	0.5111	0.3759
cascade_mask_rcnn_x101_32x4d_fpn_dconv_c3-c5	0.4557	0.4157
cascade_mask_rcnn_x101_64x4d_fpn	0.4228	0.3167
ms_rcnn_x101_64x4d_fpn	0.3772	0.2387
htc_x101_64x4d_fpn	0.4103	0.3731
htc_x101_64x4d_fpn_dconv_c3-c5	0.5090	0.3778
mask_rcnn_x101_64x4d_fpn	0.5225	0.3946
htc_x101_64x4d_fpn_dconv_c3-c5_instaboost	0.4582	0.3927
Correlation between object's size and IoU in Iodine Staining		
Model	corr(IoU,width)	corr(IoU,height)
cascade_mask_rcnn_r101_fpn_instaboost	0.2074	0.2211
cascade_mask_rcnn_x101_32x4d_fpn_dconv_c3-c5	0.0575	0.0849
cascade_mask_rcnn_x101_64x4d_fpn	0.1724	0.1690
ms_rcnn_x101_64x4d_fpn	0.1495	-0.0279
htc_x101_64x4d_fpn	0.2810	0.2526
htc_x101_64x4d_fpn_dconv_c3-c5	0.3431	0.2774
mask_rcnn_x101_64x4d_fpn	0.0706	0.2130
htc_x101_64x4d_fpn_dconv_c3-c5_instaboost	0.2536	0.1961

Chart 7: Correlation between object size and IoU

XR, ZZ, JJ and GL contributed equally.

Correlation between below 1000 pixel object's size and IoU in Real Boundary		
Model	corr(IoU,width)	corr(IoU,height)
cascade_mask_rcnn_r101_fpn_instaboost	0.0798	0.1726
cascade_mask_rcnn_x101_32x4d_fpn_dconv_c3-c5	0.0421	0.2010
cascade_mask_rcnn_x101_64x4d_fpn	0.0816	0.2756
ms_rcnn_x101_64x4d_fpn	-0.1377	0.1590
htc_x101_64x4d_fpn	0.0176	0.3088
htc_x101_64x4d_fpn_dconv_c3-c5	-0.0046	0.2038
mask_rcnn_x101_64x4d_fpn	0.0263	0.3793
htc_x101_64x4d_fpn_dconv_c3-c5_instaboost	0.0441	0.2399
Correlation between below 1000 pixel object's size and IoU in Distinct Boundary		
Model	corr(IoU,width)	corr(IoU,height)
cascade_mask_rcnn_r101_fpn_instaboost	-0.0043	0.0559
cascade_mask_rcnn_x101_32x4d_fpn_dconv_c3-c5	0.0732	0.0837
cascade_mask_rcnn_x101_64x4d_fpn	-0.0843	-0.0747
ms_rcnn_x101_64x4d_fpn	-0.1554	-0.1547
htc_x101_64x4d_fpn	0.0905	0.1190
htc_x101_64x4d_fpn_dconv_c3-c5	0.0323	0.0659
mask_rcnn_x101_64x4d_fpn	0.1591	0.0755
htc_x101_64x4d_fpn_dconv_c3-c5_instaboost	0.0911	0.0952
Correlation between below 1000 pixel object's size and IoU in Iodine Staining		
Model	corr(IoU,width)	corr(IoU,height)
cascade_mask_rcnn_r101_fpn_instaboost	0.1328	0.1657
cascade_mask_rcnn_x101_32x4d_fpn_dconv_c3-c5	0.0363	0.0641
cascade_mask_rcnn_x101_64x4d_fpn	0.1468	0.1382
ms_rcnn_x101_64x4d_fpn	0.1495	-0.0279
htc_x101_64x4d_fpn	0.3496	0.2852
htc_x101_64x4d_fpn_dconv_c3-c5	0.1932	0.1446
mask_rcnn_x101_64x4d_fpn	0.0706	0.2131
htc_x101_64x4d_fpn_dconv_c3-c5_instaboost	0.1896	0.1343

Chart 8: Correlation between object size and IoU for object below 1000 pixel