Intelligence & Robotics

Research Article

Open Access

A deep learning-based system for accurate detection of anatomical landmarks in colon environment

Chengwei Ye[1,#], Kaiwei Che[2,3,#], Yibing Yao[4,#], Nachuan Ma[5], Ruo Zhang[6], Yangxin Xu[1], Jiankun Wang[6], Max Q. H. Meng[1,6,7]

[1]Department of Electronic Engineering, The Chinese University of Hong Kong, Hong Kong 999077, China.
[2]School of Electronic and Computer Engineering, Peking University, Beijing 100871, China.
[3]Peng Cheng Laboratory, Shenzhen 518000, Guangdong, China.
[4]Department of Oncology, Air Force Medical Center, PLA, Beijing 100142, China.
[5]College of Electronic & Information Engineering, Tongji University, Shanghai 201804, China.
[6]Department of Electronic and Electrical Engineering, Southern University of Science and Technology, Shenzhen 518055, Guangdong, China.
[7]Shenzhen Research Institute of the Chinese University of Hong Kong, Shenzhen 518057, Guangdong, China.
[#]Authors contributed equally.

Correspondence to: Prof. Jiankun Wang, Prof. Max Q.-H. Meng, Department of Electronic and Electrical Engineering, Southern University of Science and Technology, 1088 Xueyuan Avenue, Shenzhen 518055, Guangdong, China. E-mail: wangjk@sustech.edu.cn; max.meng@ieee.org

How to cite this article: Ye C, Che K, Yao Y, Ma N, Zhang R, Xu Y, Wang J, Meng MQH. A deep learning-based system for accurate detection of anatomical landmarks in colon environment. *Intell Robot* 2024;4(2):164-78. http://dx.doi.org/10.20517/ir.2024.10

Received: 21 Dec 2023 **First Decision:** 15 Mar 2024 **Revised:** 9 Apr 2024 **Accepted:** 24 Apr 2024 **Published:** 28 Apr 2024

Academic Editor: Simon X. Yang, Zeng-Guang Hou **Copy Editor:** Dong-Li Li **Production Editor:** Dong-Li Li

Abstract

Colonoscopy is a standard imaging tool for examining the lower gastrointestinal tract of patients to capture lesion areas. However, if a lesion area is found during the colonoscopy process, it is difficult to record its location relative to the colon for subsequent therapy or recheck without any reference landmark. Thus, automatic detection of biological anatomical landmarks is highly demanded to improve clinical efficiency. In this article, we propose a novel deep learning-based approach to detect biological anatomical landmarks in colonoscopy videos. First, raw colonoscopy video sequences are pre-processed to reject interference frames. Second, a ResNet-101-based network is used to detect three biological anatomical landmarks separately to obtain the intermediate detection results. Third, to achieve more reliable localization, we propose to post-process the intermediate detection results by identifying the incorrectly predicted frames based on their temporal distribution and reassigning them back to the correct class. Finally, the average detection accuracy reaches 99.75%. Meanwhile, the average intersection over union of 0.91 shows a high degree of similarity between our predicted landmark periods and ground truth. The experimental results demonstrate that our proposed model can accurately detect and localize biological anatomical landmarks from colonoscopy videos.

Keywords: Biological anatomical landmark detection, colonoscopy videos, convolutional neural network (CNN), interference frame rejection, result cleaning

1. INTRODUCTION

Colorectal diseases, especially colorectal cancer (CRC), bring significant threats to public health worldwide. According to the statistical data from the American Cancer Society, the incidence rate of CRC ranged from 30 (per 100,000 persons) in Asians to 89 in Alaska Natives from 2012 through 2016 [1]. Fortunately, CRC can be prevented if adenomatous polyps and lesions within the colon can be detected and removed in a very early stage as they are essential precursors to CRC [2]. Colonoscopy is the standard screening method to detect polyps, with which the clinician can examine the entire rectum and colon of patients. Clinicians make diagnostic decisions by checking the collected colonoscopy videos. Once they detect a lesion area during manual inspection, they need to record its location relative to the colon for subsequent therapy or recheck if instant treatment is inappropriate or not required. However, given the absence of certain reference landmarks, it is difficult to obtain the relative location of lesion area with respect to the colon, since the endoscope can only capture the real-time view inside the colon but has no perception of location or distance. Biological anatomical landmarks can serve as a series of reference points that facilitate relative localization in colon environment. Hence, automatic detection of biological anatomical landmarks is highly demanded to improve clinical efficiency.

Before the advent of the machine learning and deep learning revolution, research in polyp detection was primarily dominated by traditional image processing-based methods [3–5]. Although these methods may show effectiveness in certain sample scenarios, they are susceptible to various environmental factors, such as illumination. After the emergence of machine learning, by exploiting low-level features such as color, texture, and shape information, researchers have developed some machine learning-based methods for polyp detection [6,7]. However, these techniques are hand-engineered; thus, they cannot achieve satisfactory performance due to poor characterization capability. Recently, various deep learning methods have been applied to polyp detection with the revolution of computational technologies [8–14]. Hierarchical feature learning and discrimination capabilities of neural networks can help significantly improve polyp detection accuracy. For instance, Jia *et al*. proposed a novel two-stage architecture called PLP-Net for automatic pixel-accurate polyp detection in colonoscopy images based on deep convolutional neural network (CNN) [8]. Although previous methods have managed to detect polyps robustly and efficiently, there is no work to localize polyps within the colon, which is crucial in clinical practice but challenging for clinicians to do manually.

According to the literature [15,16], biological anatomical landmark detection plays an essential role in various medical image analysis assignments, which can help achieve registration [17], segmentation [18], and localization tasks of medical images. Traditional landmark detection methods usually utilize classical machine learning algorithms and design specific image filters to extract invariant features [19–23]. For instance, Liu *et al*. leveraged the theory of submodular functions to search multiple human body landmarks including bone, organs, and vessels in 3D computed tomography (CT) images [20]. Lindner *et al*. proposed a novel landmark detection algorithm based on the supervised random forest regression-voting method for facial landmarks detection and the annotation of the joints of the hands in radiographs [21]. The stratified decision forests method was also utilized to detect anatomical landmarks in cardiac images [23].

Recently, researchers have proposed a large quantity of anatomical landmark detection algorithms [24–28] based on deep learning and reinforcement learning methods, and these algorithms showed more robust and accurate performance. Wester used a patch-based CNN to detect anatomical landmarks in 3D cardiovascular images, providing automatic registration between ultrasound and CT images of the same patient [24]. Song *et al*. proposed a two-step method to detect cephalometric landmarks automatically on skeletal X-ray images [25], utiliz-

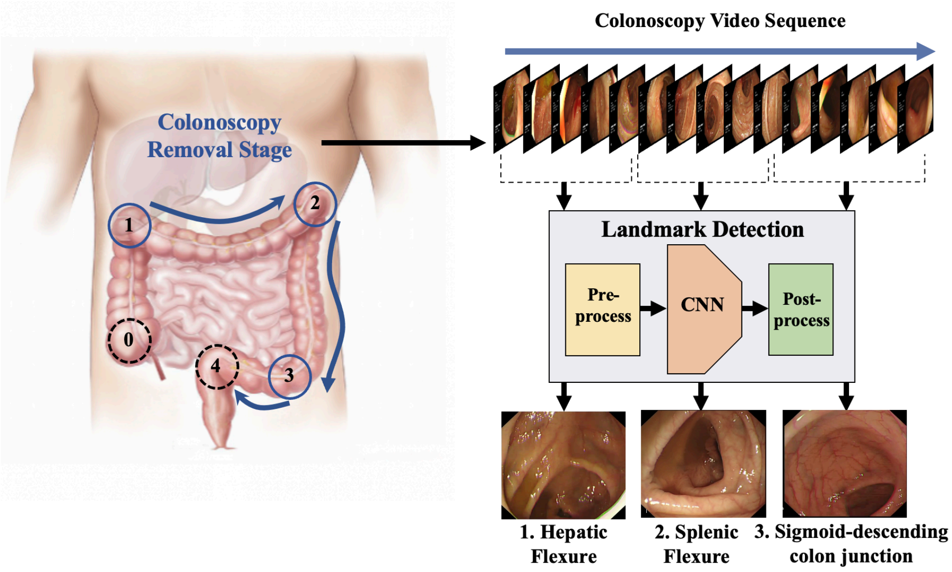
Colonoscopy Removal Stage
0
1
2
3
4
Colonoscopy Video Sequence
Landmark Detection
Pre-process
CNN
Post-process
1. Hepatic Flexure
2. Splenic Flexure
3. Sigmoid-descending colon junction

Figure 1. Biological anatomical landmarks (0: cecum; 1: hepatic flexure; 2: splenic flexure; 3: sigmoid-descending colon junction; 4: rectosigmoid junction) and the proposed threefold detection system. The left figure is quoted from Northern Care Alliance[32]. CNN: Convolutional neural network.

ing pre-trained networks with a backbone of ResNet-50[29]. Moreover, a novel communicative reinforcement learning agent system was presented for landmark detection in brain images and was evaluated on two datasets from adult magnetic resonance imaging (MRI) and fetal ultrasound scans[26]. Notably, detecting landmarks in colonoscopy videos is more challenging compared with other anatomical landmark detection tasks, since their locations in colonoscopy videos are dynamic and not directly descriptive.

Despite a few research studies[30,31] which worked on biological anatomical landmark detection in colonoscopy videos, their detection performance is not satisfactory. To fill this gap, we propose a novel deep learning-based algorithm to detect three biological anatomical landmarks in colonoscopy videos, providing a research basis for calculating the relative distances between the lesion areas (such as polyps and bleeding regions) and the landmarks. The proposed algorithm will help reduce human error and accelerate the diagnosis process significantly.

As shown in Figure 1, the three biological anatomical landmarks to be detected include hepatic flexure, splenic flexure, and sigmoid-descending colon junction. Cecum and rectosigmoid junction are the two end points of the colon. The colonoscopy video sequences are passed into the threefold system consisting of a pre-processing module, a detection network, and a post-processing module. The outputs of the system include prediction results for each frame and locations of the three landmark periods within the video sequence.

Our main contributions can be summarized in the following two aspects:

1. We collect a colonoscopy video dataset and finely label the time periods of three biological anatomical landmarks for each video.
2. We propose a novel three-fold biological anatomical landmark detection system for colonoscopy, consisting of a pre-processing module, a deep learning-based detection network, and a post-processing module.

The remainder of this article is organized as follows. We introduce the colonoscopy video dataset in Section 2. Section 3 outlines the proposed biological anatomical landmark detection system, while the experimental results are presented and analyzed in Section 4. Finally, we draw some conclusions and discuss the future work in Section 5.

2. DATASET

In a study approved by the local medical ethics committee, colonoscopy video data were obtained from 49 patients. All colonoscopies were performed on those who satisfied all of (1) at the age of 18-75; (2) cecal intubation; (3) adequate cleansing of colon. Patients were excluded if satisfied one or more of (1) reduced life expectancy; (2) history of CRC or adenomas; (3) history of serrated polyps with diameter above 10 mm; (4) history of surgical colon resection; (5) on-going chemotherapy or radiotherapy; (6) inflammatory disease. Each patient signed an informed consent form. The Olympus endoscopy system was used. The videos have a resolution of 560 × 720 and a frame rate of 50 fps. The duration of each video is approximately 10-20 min, with a maximal duration of 20.35 min (61,050 frames) and a minimal duration of 10.8 min (32,400 frames). Clinical information, including the serial number of the patient, testing date, and current timestamp, is also displayed in the videos.

During both the insertion stage and the removal stage of each colonoscopy video sequence, the timestamp and scope length when passing each biological anatomical landmark are recorded by clinicians. However, due to the resistance and disturbance, while inserting the scope, videos captured from the insertion stage contain a large number of interference frames such as turbid and camera shaking frames. Therefore, in view of the concern for data quality during the insertion stage, ground truth labels are generated from the timestamp and scope length data during the removal stage. As shown in Figure 1, for each video sequence, there are three time periods during which the anatomical landmarks are detected, each with a duration of 10 to 25 s. These periods are represented as positive periods in ground truth labels. The labels were manually annotated and verified by expert clinicians based on the bending features and biological characteristics of the landmark regions. Since the differences between landmark regions and normal regions are subtle and hardly perceptible, identifying the landmark periods is a challenging task involving significant difficulty, which can only be accomplished by clinicians with sufficient experience.

In terms of sampling images from the videos, the sampling period for each landmark begins at the timestamp of the previous landmark (or the timestamp when the removal stage starts) and ends at the timestamp of the next one (or the timestamp when the removal stage ends). As shown in Figure 2, we introduce a positive-and-negative gap (P-N gap) between the landmark and normal tissue periods to guarantee that the interclass difference is sufficiently large. Given the P-N gap, the labeled landmark period is taken as the positive period while 80% of the two remaining parts of the sampling period are taken as the negative periods. To tackle the sampling imbalance problem, we apply adaptive sampling frequencies for landmark and normal tissue periods. Details are presented in Section 5.1.

3. METHODS

In the proposed landmark detection system, the collected video sequences are first sampled into positive and negative frames separated by the P-N gap with adaptive sampling frequencies. Next, the pre-processing module is applied to reject interference frames. The pre-processed data are then shuffled, clipped, and divided into training, validation, and test sets. In total, 8,640 image frames are available, of which 5,529, 1,382, and 1,729 frames belong to the training, validation, and test sets, respectively. Details are shown in Table 1. The training and validation sets are applied to train and validate the detection model. After testing, the model outputs intermediate detection results. Finally, the post-processing module is applied to reassign the incorrectly predicted frames back to the correct class. The entire process of our proposed system is shown in Figure 2.

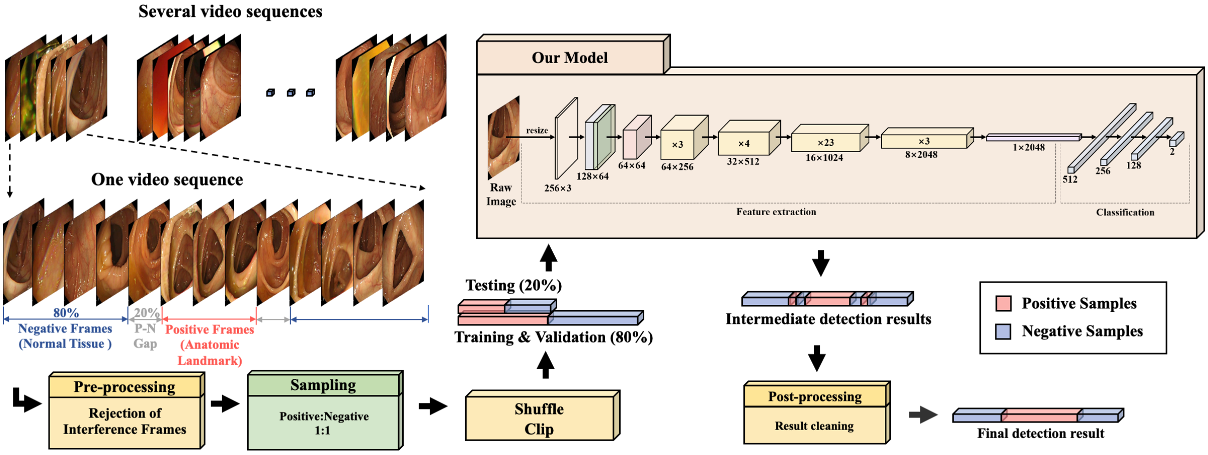

Figure 2. Workflow of the proposed landmark detection system. The pre-processing module rejects interference frames in input data. The pre-processed data are divided into training, validation, and test sets. The detection model based on ResNet-101 is trained and validated using the training and validation sets. After passing the test data into the trained model, the model outputs intermediate detection results indicating whether each frame should be classified as positive or negative. Finally, the post-processing module identifies the incorrectly predicted frames and reassigns them back to the correct class. P-N gap: Positive-and-negative gap.

In this section, we first introduce a pre-processing module to reject interference frames in the collected data. Then, we propose a novel network structure based on ResNet-101 to detect landmarks. Finally, we illustrate the post-processing module for cleaning the intermediate detection results, which enables localization of landmark periods within the video period and improves final detection performance.

3.1. Pre-process: rejection of interference frames

Colonoscopy video frames consist of biological anatomical landmark, normal tissue, and interference frames caused by feces, bubbles, and camera shaking. The crucial issue of the biological anatomical landmark detection task is distinguishing landmarks from normal tissue. However, during the sampling process, interference frames may appear in both normal tissue and anatomical landmark sampling periods, which will cause the detection model to take wrong samples as learning inputs and thereby reduce the detection accuracy. To facilitate the accurate detection of biological anatomical landmarks in the colon, it is necessary to pre-process all the video frames to reject those interference frames.

As shown in Figure 3, normal frames include biological anatomical landmarks and normal tissue, while interference frames include turbid, bubble, and blurry frames caused by camera shaking and black screen. It can be observed from Figure 3 that normal and interference frames exhibit clear distinctions, especially in terms of texture and color features. The textures of normal frames are finer and more complex than the textures of interference frames. Further, the color of normal frames involves frequent changes, whereas the interference frames are relatively monochromatic.

Based on the observation of the distinctions between normal and interference frames, the rejection of interference frames is achieved by applying an image processing tool named Canny edge detector[33]. The Canny edge detector has been widely applied in the field of computer vision to locate sudden changes in intensity and find object boundaries in images. In the direction of the maximum intensity change of the Canny edge detector, if the amplitude of gradient of one pixel is greater than the width of gradient of the pixels on both sides, the pixel is classified as belonging to an edge. The smallest value between the two thresholds is used for edge linking, while the largest value finds initial segments of strong edges. In our pre-processing module, a Canny edge detector with thresholds of 70 and 100 is applied. L_2 norm is used to calculate the image gradient magnitude.

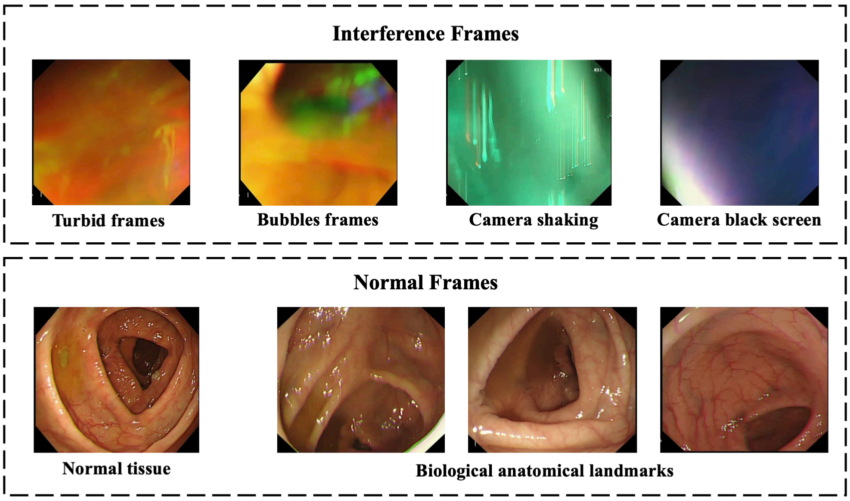

Figure 3. Interference frames and normal frames. Interference frames contain turbid frames, bubble frames, and blurry frames resulting from camera shaking and black screen. Normal frames contain biological anatomical landmark frames and normal tissue frames.

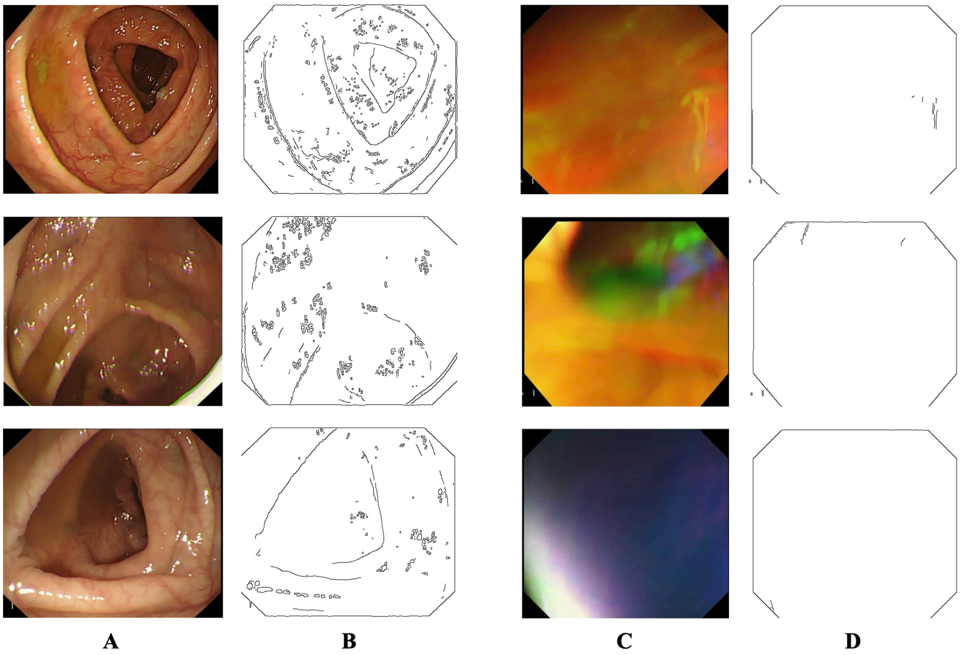

Figure 4. (A) Normal frames in colonoscopy video; (B) The edge features of normal frames extracted by Canny edge detector; (C) Interference frames in colonoscopy video; (D) The edge features of interference frames extracted by Canny edge detector.

Texture and color features can be regarded as edge indicators because the complexity of texture and the change of color can be represented by the amplitude of the gradient. Compared with interference frames, normal frames involve more complex textures and more sudden color changes. Therefore, the amplitude of the gradient of pixels in normal frames is greater than in interference frames.

As shown in Figure 4, column A shows three original normal frames and column B depicts their edge features, while column C shows three original interference frames and column D depicts their edge features, in which the black pixels belong to edges whereas the white pixels do not. It can be observed that the edge pixels in the normal frames are more numerous than the edge pixels in the interference frames. We calculate the number of black edge pixels for each frame passed through the Canny detector.

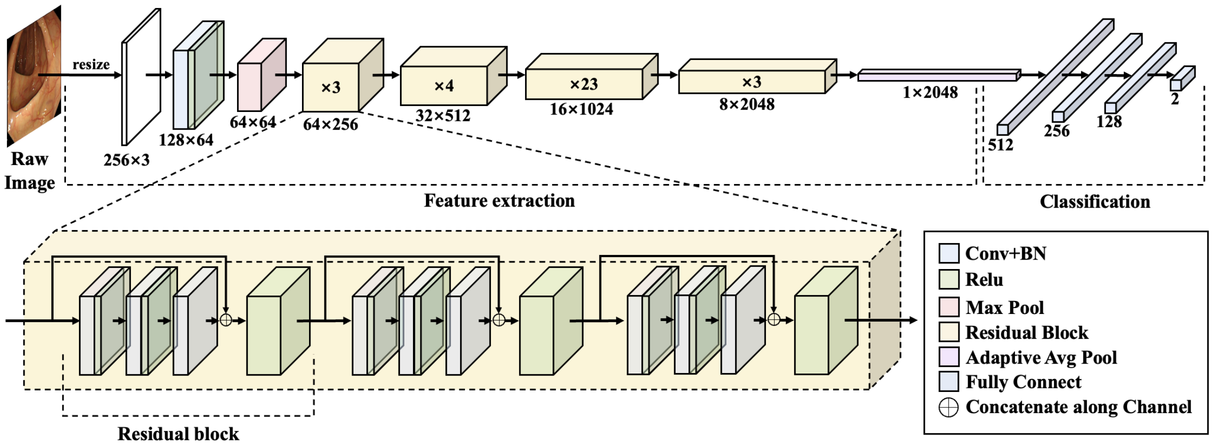

Figure 5. Our proposed landmark detection model based on ResNet-101. The single 1000-way fully connected layer is substituted by four consecutive fully connected layers. A dropout layer is inserted in between each pair of the four fully connected layers, which randomly drops neurons from visible and hidden layers to avoid overfitting.

To experimentally determine the threshold value of black edge pixels that distinguishes interference frames from normal frames, we classify 200 images into interference or normal classes using various threshold choices. It is observed that a threshold value above 2,000 will wrongly recognize some normal frames as interference frames, while a threshold value below 2,000 will fail to identify some interference frames. To balance eliminating interference frames and preserving normal frames, 2,000 is chosen as the optimal threshold, which means a frame should be classified as normal if it contains more than 2,000 black edge pixels. The threshold is then applied to classify outputs of the Canny detector into normal class or interference class and thereby reject interference frames in colonoscopy videos.

3.2. Landmark detection network based on ResNet-101

The network inputs are 256 × 256 colonoscopy images, while the network output is a two-dimensional vector indicating whether an input image should belong to the biological anatomical landmark (positive) or normal issue (negative) class. The network structure is developed based on the ResNet-101 model. We conduct comparative experiments with four other deep learning models, including Vgg16, Inception v3, ResNet-50, and ResNet-101, and the results demonstrate that our proposed model outperforms the others. Details are presented in Section 4.5.

As shown in Figure 5, the first layer is a 7 × 7 convolutional layer with a stride of 2. The second layer is a 3 × 3 max pooling layer with a stride of 2. Next, 4 different kinds of 3-layer building blocks are stacked 3, 4, 23, and 3 times, respectively. In the original model, the network ends with a global average pooling layer and a 1000-way fully connected layer with LogSoftMax. In the modified model for this task, the single 1,000-way fully connected layer is replaced by four consecutive fully connected layers. To prevent overfitting, we use a dropout layer between each pair of the four consecutive fully connected layers to randomly drop neurons from visible and hidden layers. For the three dropout layers, the probabilities of an element to be set as zero are 0.4, 0.3, and 0.3, respectively.

Due to the considerable depth of the network and the limited volume of training data, it is difficult to train the ResNet-101 model from scratch. Instead of randomly initializing the model weights, we load the weights pretrained on the ImageNet dataset[34] for all layers except the last four fully connected layers. The ImageNet is a public dataset containing over 14 million quality-controlled and human-annotated natural images belonging to 1000 categories.

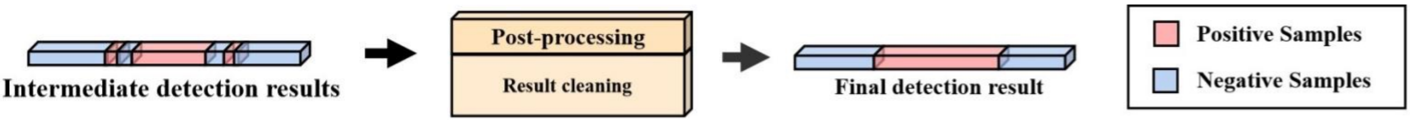

Figure 6. Post-processing module. In the intermediate detection results, some negative frames are discretely distributed in the presumably continuous landmark periods. Post-processing module aims to reassign the wrongly predicted frames into the correct class to obtain continuous landmark periods. It improves final detection performance and enables localization of landmark periods within the video period.

Output from the last fully connected layer could be activated by LogSoftmax activation, as shown in Equation (1). The LogSoftmax formulation can be expressed as:

$$LogSoftmax(x_i) = \log(\frac{\exp(x_i)}{\sum_j^N \exp(x_j)}) \tag{1}$$

where N is the number of classes, which is 2 in our case. x_i represents the specific predicted result of the landmark or normal tissue class.

Our loss function is the negative log likelihood loss (NLLLoss), which is useful for training a classification problem. Its input is a log probability vector and a target label. Our reduction is "mean". NLLLoss can be given as:

$$L(x, y) = \sum_{n=1}^{M} \frac{-x_n y_n}{M} \tag{2}$$

where M denotes the batch size, which is 32 in our case. x_n represents the predicted result of all classes. $x_n y_n$ denotes the dot product of vectors x_n and y_n.

Although freezing the model weights as pre-trained can reduce the amount of computation in the backward pass and decrease training time, freezing the model has the potential to hinder the learning of features from colonoscopy images, since the colonoscopy images exhibit distinctive features that vastly differ from images in the ImageNet dataset. Therefore, to obtain better performance, the weights of all layers are not frozen during the model training stage. They are updated through the backpropagation of gradients in every epoch.

In total, our proposed anatomical landmark detection model based on ResNet-101 contains 43,713,730 trainable parameters. It takes 0.0279 s on average for the model to detect a single video frame into the biological anatomical landmark or the normal tissue class. The statistics mentioned above are obtained from experiments conducted on an NVIDIA GeForce RTX 2080 Ti GPU.

3.3. Post-process: result cleaning

After passing the colonoscopy video frames through the network, we obtain the intermediate detection results indicating whether each frame should be classified as positive or negative. The next step toward the final results is to locate the landmark periods within the whole video period. However, it can be observed that some negative frames are discretely distributed in the presumably continuous landmark periods, while some positive frames are discretely distributed in the non-landmark normal periods. Therefore, it is necessary to post-process the intermediate detection results by identifying the incorrectly predicted frames and reassigning them back to the correct class.

As shown in Figure 6, in the intermediate detection results, within the landmark period, a large proportion of frames are predicted as positive while a small proportion are categorized as negative. In contrast, outside the

Table 1. Sizes of training, validation and test data

	Training	Validation	Total	Positive	Negative	Test
L1	2536	634	3170	1400	1770	793
L2	1686	421	2107	957	1150	527
L3	1307	327	1634	742	892	409

L1, L2, and L3 represent hepatic flexure, splenic flexure, and sigmoid-descending colon junction, respectively.

landmark period, a large proportion of frames are denoted as negative while a small proportion are labeled as positive. Since each landmark period should be continuous, the incorrectly predicted frames in intermediate detection results can be identified according to the temporal distribution.

The specific post-processing strategy is as follows. For each frame, we count the number of positive frames among eight neighboring frames, including four on the left and four on the right. If the count is greater than 3, the frame of interest should be considered as positive; otherwise, it should be classified as negative. For the first and the last four frames, we count the number of positive frames among ten neighboring frames. If the count is greater than 5, the first and the last four frames should be categorized as positive; otherwise, they should be denoted as negative. The detailed algorithm is shown in the Supplementary Materials.

4. EXPERIMENT RESULTS

4.1. Experimental data and setup

In our experiment, we introduce a P-N gap between positive and negative periods and sample them with adaptive sampling frequencies. The reason for leaving a P-N gap is that the labels of landmark periods may contain some errors, which blur the boundaries between positive and negative samples. Therefore, when parts of the positive and negative samples are highly similar, it is possible for the model to learn from the wrong samples.

Another crucial issue is avoiding sample imbalance, which would make our detection model suffer from serious bias. To keep the balance between positive and negative samples, we sample them from the videos with adaptive sampling frequencies to ensure that the ratio of positive and negative samples is approximately 1:1 [Figure 2]. We set the positive sampling frequency as 10 fps and set negative sampling frequency based on:

$$f_n = f_p \times \frac{t_p}{t_n} \tag{3}$$

where f_n and f_p denote negative and positive sampling frequencies, and t_n and t_p indicate negative and positive time periods in the video.

The training set contains 80% of the total samples, while the test set contains 20%. In order to guarantee no overlap between the training and test sets, all images sampled from the same video are assigned to either the training or test set. Images sampled from videos numbered 1-39 are assigned to the training set, while images sampled from videos numbered 40-49 are assigned to the test set. We also create a validation set from the training set to compare the performance of different models and to prevent the overfitting problem. Details are shown in Table 1.

4.2. Performance metrics

To quantitatively assess the classification and localization performance, five evaluation metrics applied in the experiments are accuracy, precision, recall, F1 score, and intersection over union (IoU), which are calculated by:

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \tag{4}$$

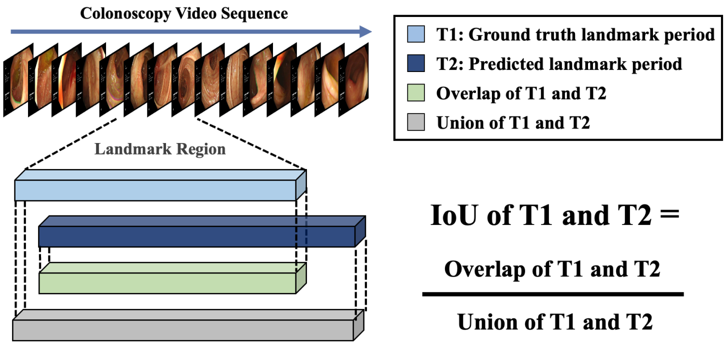

Figure 7. IoU of two time periods. IoU: Intersection over union.

$$Precision = \frac{TP}{TP + FP} \tag{5}$$

$$Recall = \frac{TP}{TP + FN} \tag{6}$$

$$F1 = 2 \times \frac{precision \times recall}{precision + recall} \tag{7}$$

where true positive (TP) cases are the biological anatomical landmark frames correctly predicted as landmarks while true negative (TN) cases represent the normal frames correctly predicted as normal. False positive (FP) cases are the normal frames incorrectly predicted as landmarks while false negative (FN) cases represent the biological anatomical landmarks frames incorrectly predicted as normal.

IoU is a metric frequently used to evaluate the extent of overlapping of two regions in applications related to object detection. It takes its value within the range between 0 and 1, with a greater value indicating a more considerable extent of overlapping. The traditional definition of IoU measures the similarity between the predicted and ground truth areas. Based on the IoU defined in a two-dimensional spatial context, we derive the IoU defined in a one-dimensional temporal context. As shown in Figure 7, the IoU of two time periods, T1 and T2, can be obtained by dividing the length of their overlapping period over the length of their union period.

4.3. Training and implementation details

We train three separate models to detect three biological anatomical landmarks respectively. The resolution of colonoscopy images is reduced from 521 × 478 to 256 × 256 to reach a trade-off between detection accuracy and computational cost. Adam optimizer with a learning rate of 0.001 is applied when training the model. The batch size is set to 32, and the number of epochs is set to 200.

During the testing process, the test samples are inputted into the trained detection model and classified into positive or negative classes to obtain the intermediate detection results. The intermediate detection results are then post-processed to retrieve the final detection results, which include the overlapping information of predicted landmark periods and ground truth labels. All the three biological anatomical landmarks are detected following the above pipeline.

4.4. Intermediate detection results

After implementing the experiments based on the above-mentioned procedure and parameter settings, the obtained intermediate detection results for the three biological anatomical landmarks in terms of various test indicators are shown in Table 2. In terms of intermediate detection accuracy, our proposed model based on ResNet-101 reaches 90.72%, 90.85%, and 92.03% for the three landmarks, respectively. Furthermore, the results

Table 2. Intermediate detection results in terms of multiple test indicators

	Accuracy	Precision	Recall	F1 score
L1	0.9072	0.9092	0.9077	0.9071
L2	0.9085	0.9170	0.8751	0.8916
L3	0.9203	0.9234	0.9101	0.9157

L1, L2, and L3 represent hepatic flexure, splenic flexure, and sigmoid-descending colon junction, respectively.

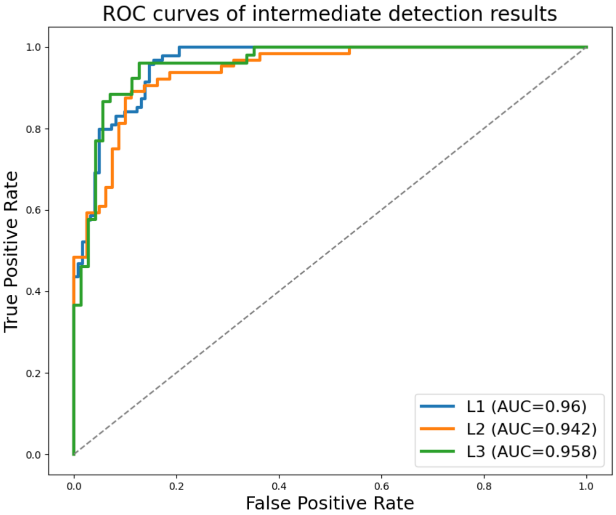

Figure 8. ROC curves of intermediate detection results. AUC represents area under the ROC curve. L1, L2, and L3 represent hepatic flexure, splenic flexure, and sigmoid-descending colon junction, respectively. ROC: Receiver operating characteristic; AUC: area under the ROC curve.

of precision, recall, and F1 score illustrate that our proposed model guarantees the balance between precision and recall, which demonstrates that our detection model is both precise and robust.

The Receiver operating characteristic (ROC) curves of intermediate detection results of three landmarks are plotted in Figure 8. They measure the classification performance by plotting the TP rate (TPR) against the FP rate (FPR) at various threshold settings. A trajectory ascending towards the upper-left corner indicates better performance in discriminating between positive and negative samples. Therefore, our proposed detection model shows satisfactory performance for the three biological anatomical landmarks. The area under the ROC curve (AUC) quantifies the model's capability of distinguishing between two classes. A higher AUC value indicates a more successful separation of the two classes. The AUC values of detecting hepatic flexure, splenic flexure, and sigmoid-descending colon junction are 0.96, 0.942, and 0.958, respectively, which quantitatively prove the discriminating capability of our detection model.

4.5. Comparison with other models

To further evaluate the performance of our proposed model based on ResNet-101, we conduct comparative experiments with four other frequently applied deep learning models: Vgg16, Inception v3, ResNet-50, and ResNet-101. To guarantee the fairness of comparison, the parameter settings and training process of each model are consistent with our model. We apply Adam optimizer with a learning rate of 0.001. The batch size is set to 32 and the number of epochs is set to 200. We summarize the corresponding results of accuracy, precision, recall, and F1 score in Table 3. As shown in Table 3 and Figure 9, in terms of intermediate detection accuracy, our proposed model outperforms the Vgg16 model by 40.14%, 23.95%, and 31.87% for the

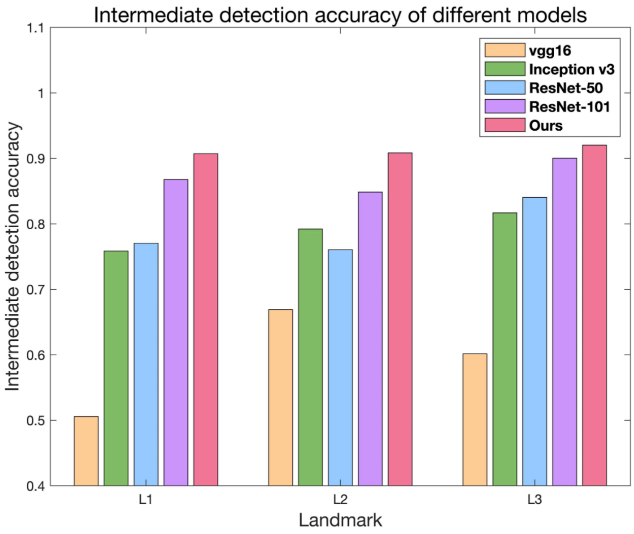

Figure 9. Intermediate detection accuracy of Vgg16, Inception v3, ResNet-50, ResNet-101, and our model. L1, L2, and L3 represent hepatic flexure, splenic flexure, and sigmoid-descending colon junction, respectively.

Table 3. Comparison with other models in intermediate detection results

		Accuracy	Precision	Recall	F1 score
L1	vgg16	0.5058	0.2529	0.5000	0.3359
	Inception v3	0.7587	0.7735	0.7573	0.7547
	ResNet-50	0.7703	0.7952	0.7720	0.7661
	ResNet-101	0.8677	0.8904	0.8682	0.8676
	Ours	0.9072	0.9092	0.9077	0.9071
L2	vgg16	0.6690	0.3345	0.5000	0.4008
	Inception v3	0.7923	0.8179	0.7050	0.7243
	ResNet-50	0.7606	0.7647	0.6679	0.6810
	ResNet-101	0.8486	0.8287	0.8304	0.8295
	Ours	0.9085	0.9170	0.8751	0.8916
L3	vgg16	0.6016	0.3008	0.5000	0.3756
	Inception v3	0.8167	0.8513	0.7784	0.7917
	ResNet-50	0.8406	0.8527	0.8152	0.8259
	ResNet-101	0.9004	0.9010	0.8902	0.8948
	Ours	0.9203	0.9234	0.9101	0.9157

L1, L2, and L3 represent hepatic flexure, splenic flexure, and sigmoid-descending colon junction, respectively.

three anatomical landmarks, respectively. It also surpasses the Inception v3 model by 14.85%, 11.62%, and 10.36%, while exceeding the ResNet-50 model by 13.69%, 14.79%, and 7.97% for the three anatomical landmarks, respectively. As for the ResNet-101 model, our proposed model excels it by 3.95%, 5.99%, and 1.99%, respectively. In addition, the comparison results of precision, recall, and F1 score further demonstrate its superiority over the others in anatomical landmark detection. The comparison among Vgg16, Inception v3, and the proposed model indicates the importance of residual blocks in network architecture, while the comparison among ResNet-50, ResNet-101, and the proposed model demonstrates that increasing the network depth could contribute to performance improvement.

4.6. Final detection results

To better locate the landmark periods within the whole video period, we propose to post-process the intermediate detection results by identifying the incorrectly predicted frames based on their temporal distribution and reassigning them back to the correct class.

As shown in Table 4, the final detection accuracy for the three landmarks improves by 9.02%, 8.95%, and 7.68% compared with the intermediate detection accuracy. The results demonstrate that our result cleaning

Table 4. Final detection results in terms of accuracy and IoU

	Intermediate accuracy	Final accuracy	Final IoU
L1	0.9072	0.9974	0.8960
L2	0.9085	0.9980	0.9039
L3	0.9203	0.9971	0.9332

[a]L1, L2, and L3 represent hepatic flexure, splenic flexure, and sigmoid-descending colon junction, respectively. IoU: Intersection over union.

algorithm can distinguish the wrongly classified frames based on their distribution and then correct them. Meanwhile, to evaluate the localization performance of the landmarks, we apply the IoU metric to measure the overlap between our predicted landmark periods and ground truth periods. The IoU values for the three landmarks reach 0.90, 0.90, and 0.93, respectively, showing a considerable extent of overlap. The final detection results demonstrate that our proposed system can accurately detect and localize landmarks against neighboring normal tissue.

5. CONCLUSIONS

In this paper, we present a novel deep learning-based architecture to automatically detect biological anatomical landmarks in colonoscopy videos. Comprehensive experimental results demonstrate that our proposed ResNet-101-based model outperforms other deep learning-based models in terms of accuracy, precision, recall rate, and F1 score. Quantitative results indicate that our proposed architecture can correctly differentiate biological anatomical landmarks from neighboring normal regions with an average accuracy of 99.75%.

In the future, there exist many promising research directions. First, now we only select three intermediate landmarks inside the colon as our detection objects. To obtain a more complete marking of the colon, we should consider adding cecum and rectosigmoid junction as detection objects, since they are the two end points of the colon. At the current stage of our work, the binary classification model is applied to detect each landmark. It would be more convenient and reasonable if all landmarks could be detected using a single model. We will convert the binary classification problem into multi-class classification and seek strategies to fuse the models without impairing the performance. We only compare the performance of the proposed model with traditional deep learning models. We will investigate the possibility of applying other state-of-the-art models, including You Only Look Once (YOLO) networks, transformers, generative adversarial networks (GAN), and recurrent neural networks (RNN) to improve the detection performance. In terms of pre-processing approaches, the Canny edge detector suffers from slow processing speed. In order to improve the pre-processing efficiency, the edge detector should be optimized. We will explore alternative edge detectors such as Sobel in our future work. In terms of post-processing strategy, we will compare our method with morphological operators such as median filter. Our long-term plan is to develop a novel positioning algorithm based on combining visual and magnetic trajectory information to estimate the relative distances between lesion areas and detected biological anatomical landmarks. Furthermore, another challenging future extension is to establish a 3D space that restores the internal structure of the colon based on 3D reconstruction technologies, which has the potential to significantly improve the efficiency of diagnosing lesion areas.

DECLARATIONS

Authors' contributions

Made substantial contributions to the research process and wrote the original draft: Ye C, Che K, Ma N, Zhang R, Xu Y
Performed data acquisition: Yao Y
Provided guidance and support: Wang J, Meng MQH

Availability of data and materials

Not applicable.

Financial support and sponsorship

This work was supported by the National Key Research and Development Program of China under Grant 2019YFB1312400.

Conflicts of interest

All authors declared that there are no conflicts of interest.

Ethical approval and consent to participate

The Ethics Committees of the Fourth Medical Center of PLA General Hospital (approval number: 2023KY007-KS001) and Beijing Aerospace General Hospital (approval number: 2022-50-PJ01) approved the protocols utilized in our investigation. Written informed consent was obtained from each participant in this study.

Consent for publication

Not applicable.

Copyright

REFERENCES

1. Siegel RL, Miller KD, Sauer AG, et al. Colorectal cancer statistics, 2020. *CA Cancer J Clin* 2020;70:145-64. DOI
2. American Cancer Society. Colorectal cancer facts & figures 2017–2019. Available from: https://www.cancer.org/content/dam/cancer-org/research/cancer-facts-and-statistics/colorectal-cancer-facts-and-figures/colorectal-cancer-facts-and-figures-2017-2019.pdf. [Last accessed on 25 Apr 2024]
3. Zhao Q, Meng MQH. Polyp detection in wireless capsule endoscopy images using novel color texture features. In: 2011 9th World Congress on Intelligent Control and Automation; 2011 Jun 21-25; Taipei, Taiwan, China. IEEE; 2011. pp. 948-52. DOI
4. Li B, Meng MQH. Automatic polyp detection for wireless capsule endoscopy images. *Expert Syst Appl* 2012;39:10952-58. DOI
5. Mamonov AV, Figueiredo IN, Figueiredo PN, Tsai YHR. Automated polyp detection in colon capsule endoscopy. *IEEE Trans Med Imaging* 2014;33:1488-502. DOI
6. Silva J, Histace A, Romain O, Dray X, Granado B. Toward embedded detection of polyps in wce images for early diagnosis of colorectal cancer. *Int J Comput Assist Radiol Surg* 2014;9:283-93. DOI
7. Iwahori Y, Shinohara T, Hattori A, et al. Automatic polyp detection in endoscope images using a hessian filter. In: MVA2013 IAPR International Conference on Machine Vision Applications; 2013 May 20-23; Kyoto, Japan. 2013. pp. 21-4. Available from: https://www.mva-org.jp/Proceedings/2013USB/papers/03-01.pdf. [Last accessed on 25 Apr 2024]
8. Jia X, Mai X, Cui Y, et al. Automatic polyp recognition in colonoscopy images using deep learning and two-stage pyramidal feature prediction. *IEEE Trans Autom Sci Eng* 2020;17:1570-84. DOI
9. Yuan Y, Meng MQH. Deep learning for polyp recognition in wireless capsule endoscopy images. *Med Phys* 2017;44:1379-89. DOI
10. Yuan Y, Qin W, Ibragimov B, et al. Densely connected neural network with unbalanced discriminant and category sensitive constraints for polyp recognition. *IEEE Trans Autom Sci Eng* 2019;17:574-83. DOI
11. Jia X, Xing X, Yuan Y, Xing L, Meng MQH. Wireless capsule endoscopy: a new tool for cancer screening in the colon with deep-learning-based polyp recognition. *Proc IEEE* 2019;108:178-97. DOI
12. Wu H, Zhao Z, Wang Z. META-Unet: multi-scale efficient transformer attention Unet for fast and high-accuracy polyp segmentation. *IEEE Trans Autom Sci Eng* 2023:1-12. DOI
13. Liu G, Chen Z, Liu D, Chang B, Dou Z. FTMF-Net: a fourier transform-multiscale feature fusion network for segmentation of small polyp objects. *IEEE Trans Instrum Meas* 2023;72:1-15. DOI
14. Ta N, Chen H, Lyu Y, Wu T. Ble-net: boundary learning and enhancement network for polyp segmentation. *Multimed Syst* 2023;29:3041-54. DOI
15. Zhou SK, Rueckert D, Fichtinger G. Handbook of medical image computing and computer assisted intervention. Elsevier; 2020. DOI
16. Zhou SK, Greenspan H, Davatzikos C, et al. A review of deep learning in medical imaging: imaging traits, technology trends, case studies with progress highlights, and future promises. *Proc IEEE Inst Electr Electron Eng* 2021;109:820-38. DOI
17. Lange T, Papenberg N, Heldmann S, et al. 3D ultrasound-CT registration of the liver using combined landmark-intensity information. *Int J Comput Assist Radiol Surg* 2009;4:79-88. DOI
18. Ibragimov B, Korez R, Likar B, Pernuš F, Xing L, Vrtovec T. Segmentation of pathological structures by landmark-assisted deformable

models. *IEEE Trans Med Imaging* 2017;36:1457-69. DOI

19. Chiras J, Depriester C, Weill A, Sola-Martinez MT, Deramond H. [Percutaneous vertebral surgery. Technics and indications]. *J Neuroradiol* 1997;24:45-59. Available from: https://pubmed.ncbi.nlm.nih.gov/9303944/. [Last accessed on 25 Apr 2024]
20. Liu D, Zhou KS, Bernhardt D, Comaniciu D. Search strategies for multiple landmark detection by submodular maximization. In: 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition; 2010 Jun 13-18; San Francisco, CA, USA. IEEE; 2010. pp. 2831-38. DOI
21. Lindner C, Bromiley PA, Ionita MC, Cootes TF. Robust and accurate shape model matching using random forest regression-voting. *IEEE Trans Pattern Anal Mach Intell* 2014;37:1862-74. DOI
22. Ebner T, Stern D, Donner R, Bischof H, Urschler M. Towards automatic bone age estimation from MRI: localization of 3D anatomical landmarks. *Med Image Comput Comput Assist Interv* 2014;17:421-8. DOI
23. Oktay O, Bai W, Guerrero R, et al. Stratified decision forests for accurate anatomical landmark localization in cardiac images. *IEEE Trans Med Imaging* 2016;36:332-42. DOI
24. Wester BH. Detecting anatomical landmarks in 3D cardiovascular images using convolutional neural network. 2020. Available from: https://www.duo.uio.no/handle/10852/80720. [Last accessed on 25 Apr 2024]
25. Song Y, Qiao X, Iwamoto Y, Chen Y. Automatic cephalometric landmark detection on X-ray images using a deep-learning method. *Appl Sci* 2020;10:2547. DOI
26. Leroy G, Rueckert D, Alansary A. Communicative reinforcement learning agents for landmark detection in brain images. In: MLCN 2020, RNO-AI 2020: Machine Learning in Clinical Neuroimaging and Radiogenomics in Neuro-oncology. Cham: Springer; 2020. pp. 177-86. DOI
27. Lian C, Wang F, Deng HH, et al. Multi-task dynamic transformer network for concurrent bone segmentation and large-scale landmark localization with dental CBCT. *Med Image Comput Comput Assist Interv* 2020;12264:807-16. DOI
28. Zhu H, Yao Q, Xiao L, Zhou SK. You only learn once: universal anatomical landmark detection. In: MICCAI 2021: Medical Image Computing and Computer Assisted Intervention – MICCAI 2021. Cham: Springer; 2021. pp. 85-95. DOI
29. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR); 2016 Jun 27-30; Las Vegas, NV, USA. IEEE; 2016. pp. 770-78. DOI
30. Kohn N, Fuchtmann J, Morandell J, Ostler D, Wilhelm D, Feußner H. Markerless endoluminal navigation (Project BIOPASS) - Deep learning based detection of intestinal segments for colorectal endoscopic investigations. In: Conference Proceedings: 17th Annual Meeting of the German Society for Computer- and Robot-Assisted Surgery e.V. 2018. pp. 223-25. (in German) Available from: https://www.curac.org/ images/advportfoliopro/images/CURAC2018/CURAC%202018%20Tagungsband.pdf. [Last accessed on 25 Apr 2024]
31. Horsch A, Allescher HD. Automatische lokalisationserkennung in der endoskopie des gastrointestinaltrakts - eine Machbarkeitsstudie. In: Meiler M, Saupe D, Kruggel F, Handels H, Lehmann TM, editors. Bildverarbeitung für die Medizin 2002. Berlin: Springer; 2002. pp. 163-66. DOI
32. Northern Care Alliance NHS Foundation Trust. Gastroenterology. Available from: https://www.ncaresearch.org.uk/research/gastroenterology/. [Last accessed on 25 Apr 2024]
33. Ding L, Goshtasby A. On the Canny edge detector. *Pattern Recognit* 2001;34:721-25. DOI
34. Deng J, Dong W, Socher R, Li LJ, Li K, LI FF. ImageNet: a large-scale hierarchical image database. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition; 2009 Jun 20-25; Miami, FL, USA. IEEE; 2009. pp. 248-55. DOI