Highlights

Query^2: Query over Queries for Improving Gastrointestinal Stromal Tumour Detection in an Endoscopic Ultrasound

Qi He,Sophia Bano,Jing Liu,Wentian Liu,Danail Stoyanov,Siyang Zuo

- A novel endoscopic ultrasound dataset, GIST514-DB, which includes detailed tumour locations, tumour types from biopsies, and additional anatomical locations from a gastroendoscopy.
- A novel framework, i.e., Query^2, exploits anatomical locations to improve recognition accuracy.
- A detailed evaluation and comparison of Query^2 with the related deep learning-based image recognition methods using the GIST514-DB dataset.

Query^2: Query over Queries for Improving Gastrointestinal Stromal Tumour Detection in an Endoscopic Ultrasound

Qi He[a], Sophia Bano[b], Jing Liu[c], Wentian Liu[c], Danail Stoyanov[b] and Siyang Zuo[a,*]

[a]The Key Laboratory of Mechanism Theory and Equipment Design of Ministry of Education, Tianjin University, Tianjin, China
[b]The Welcome/EPSRC Center for Interventional and Surgical Sciences (WEISS), University College London, London, UK
[c]Department of Gastroenterology, Tianjin Medical University General Hospital, Tianjin, China

ARTICLE INFO

Keywords:
Gastrointestinal stromal tumours
Endoscopic ultrasound
Object detection
Anatomical location

ABSTRACT

Gastrointestinal stromal tumour (GIST) lesions are mesenchymal neoplasms commonly found in the upper gastrointestinal tract, but non-invasive GIST detection during an endoscopy remains challenging because their ultrasonic images resemble several benign lesions. Techniques for automatic GIST detection and other lesions from endoscopic ultrasound (EUS) images offer great potential to advance the precision and automation of traditional endoscopy and treatment procedures. However, GIST recognition faces several intrinsic challenges, including the input restriction of a single image modality and the mismatch between tasks and models. To address these challenges, we propose the Query^2 (Query over Queries) framework to identify GISTs from ultrasound images. The proposed Query^2 framework applies an anatomical location embedding layer to break the single image modality. A cross-attention module is then applied to query the queries generated from the basic detection head. Moreover, a single-object restricted detection head is applied to infer the lesion categories. Meanwhile, to drive this network, we present GIST514-DB, a GIST dataset that will be made publicly available, which includes the ultrasound images, bounding boxes, categories and anatomical locations from 514 cases. Extensive experiments on the GIST514-DB demonstrate that the proposed Query^2 outperforms most of the state-of-the-art methods.

1. Introduction

Gastrointestinal stromal tumours (GISTs) are the most common mesenchymal neoplasms of the gastrointestinal (GI) tract and are derived from the Canjal intestinal cells [1]. Similar to other subepithelial lesions (SELs), such as leiomyomas and schwannomas, GISTs are regularly encountered as incidental findings during an upper GI endoscopy and show very few clinical symptoms and complications [2]. Many SELs are estimated to be benign, such as leiomyomas, schwannomas or duplication cysts. However, up to 13% of upper GI tract lesions are malignant, and an additional 8% at least have malignant potential [2]. In addition, retrospective studies show that GISTs dominate potentially malignant SELs in the upper GI tract [2]. Approximately 20-30% of GISTs are malignant, and the rest reveal an indeterminate risk of aggressive behaviour that may have the capability to become malignant and then metastasise [2]. Therefore, it is important to recognise and manage potentially malignant GISTs.

Diagnosis and follow-up treatment of GISTs rely on numerous imaging modalities, which can be roughly divided into three parts based on a *3-step algorithmic approach* (see Fig. 1), such as endoscopic assessment, endoscopic ultrasound (EUS) criteria and classification, and tissue acquisition [3]. As the first step, the gastric endoscope is sent into the stomach through the mouth to capture standard gastroendoscopic images. The gastroendoscopic image is used to describe the appearance of a lesion and its location, but it usually cannot distinguish between different types of SELs. Fortunately, EUS has become a helpful tool for evaluating SELs. Description of an ultrasonic image, e.g., borders, layer of origin, size, contour, echogenicity, echopattern and vascularity can effectively narrow down the differential diagnosis list, but a definite diagnosis can hardly be established on the basis of a EUS image because the GIST is apparently similar to its benign counterpart, such as leiomyomas. Different from most benign hyperechoic lesions, the EUS image of a typical GIST is a hypoechoic, solid mass originating from the proper muscle layers (4^{th} EUS layer) of the GI wall [3]. However, gastric leiomyomas share similar characteristics with GISTs (as shown in Fig. 2), which may explain why distinguishing benign from

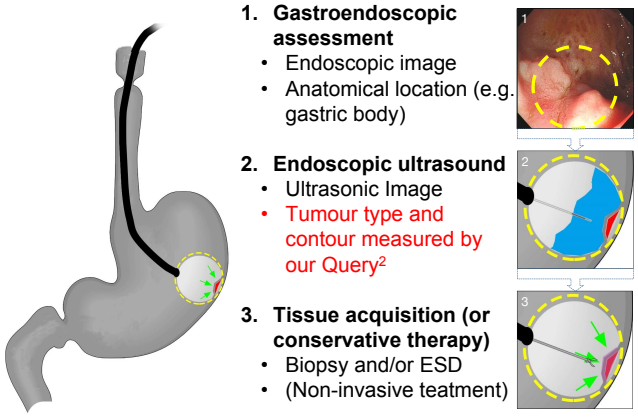

Figure 1: A 3-step algorithmic approach for the diagnosis and treatment of GISTs that is composed of an endoscopic assessment, endoscopic ultrasound, and tissue acquisition.

*Corresponding author
siyang_zuo@tju.edu.cn (S. Zuo)

potentially malignant lesions is suboptimal [4]. In addition, based on these measurements, the surgeon can choose the optimal means of tissue acquisition, such as ultrasound-guided biopsy and tumour excision, or accordingly plan the best strategy to further conservative therapy if non-invasive treatment is required [2, 3]. Our Query^2 is designed to offer helpful descriptions for EUS imaging, allowing more precise surgery and more justified conservative therapy.

Deep learning (DL)-based methods [5–8] have achieved great success in real-world image recognition while further attracting the endoscopic ultrasound community to develop computer-assisted diagnosis (CAD) applications [9–12]. However, most applications [9–12] still have several limitations of EUS, such as similar appearance of different SELs, insufficient training data, and imbalanced categories in the dataset. However, existing artificial neural networks for real-world image recognition [5–7] or object recognition [8] are not suitable for modelling EUS-based SEL recognition because the EUS image dataset has different object distributions than real-world datasets, such as ImageNet [13] and Stanford Cars [14] for classification tasks and COCO [15] for object detection tasks. For example, unlike most objects in ImageNet and Stanford Cars, which are large and centre-located, objects in a typical EUS image are small and indeterminately positioned. Additionally, in contrast to COCO, which has an indeterminate number of objects per image, there is only one object per image in the EUS dataset. In addition, information outside the image can also be helpful for image recognition. For example, even fine-grained visual classification algorithms [16] can incorrectly distinguish between two species, such as the American crow and common raven, with extremely similar appearance characteristics but different habitats. In other words, in this case, the tags of the habitats can help with image recognition. Therefore, in this study, we mainly focus on solving the problem of task and model mismatch and introduce labels of anatomical locations to help image recognition.

In this paper, we propose Query^2, a novel query-based single-object detection network for GIST detection, by embedding additional annotation of tumour anatomical region classification from gastroendoscopy (referred to as an anatomical location for short) to strengthen the features of EUS images with both end-to-end network training and inference. We first extract sparse queries that represent image features of bounding boxes through a query-based detection pipeline. We then adaptively learn the anatomical location feature from image features through a multi-head cross-attention module and memorise the feature for each anatomical location through an anatomical location embedding layer. We make a simple but important assumption that the number of SELs in EUS images is at most one for a fair comparison between classification and detection models. Thus, we can infer image classes in an end-to-end style or based on the bounding box with maximum probability. In contrast to previous works, our Query^2 requires the anatomical location of the tumour as additional tags to improve image recognition. Thus, we build a EUS dataset i.e., GIST514-DB and extensively evaluate our method in three typical tasks, including classification, detection, and instance segmentation, on GIST514-DB. Although the average horizontal diameter of SELs in GIST514-DB is smaller than 11 mm, our Query^2 still achieves a high accuracy of 95.1%. Our method outperforms existing state-of-the-art methods and is of high clinical relevance, as it offers helpful information, including tumour type and contour for potential downstream applications. Our main contributions can be summarised as follows:

(i) We build a novel EUS dataset, i.e., GIST514-DB (see Tables 2 and 3), which will be made publicly available. The dataset includes detailed tumour locations from EUS, tumour types from biopsies, and *additional anatomical locations from gastroendoscopy*. The training and validation dataset contains 251 GIST cases and 263 leiomyoma cases to avoid data imbalance.

(ii) We propose a novel framework, i.e., Query^2 (see Fig. 4), to accurately recognise GISTs and leiomyomas from EUS images. Superior to previous modelling and struggling with single image modality and mismatch between task and model, our Query^2 can leverage additional annotations from gastroendoscopy, i.e., anatomical locations, and the additional assumption, i.e., single-object restriction, for improving recognition accuracy.

(iii) We compare Query^2 with existing models in classification, object detection and instance segmentation on GIST514-DB. Through 5-fold cross-validation, we show that our method outperforms the most related methods (see Table 5) in GIST recognition and the state-of-the-art approach (see Tables 6, 7 and 9) from the real-world dataset by a large margin. The code is available at `https://github.com/howardchina/query2`.

2. Related works

In this section, we discuss the three most related works, including EUS features differentiating GISTs from leiomyomas, CAD applications and datasets to recognise GISTs, and relationships with real-world object recognition.

2.1. EUS features differentiating GISTs from leiomyomas

EUS is a popular technique for diagnosing different gastric SELs [2–4, 18]. Abnormal tumour size (larger than 30 or 40 mm) and irregular tumour margins are reported to be the most useful EUS features to predict high-risk GISTs and very high-risk GISTs, but low-risk GISTs cannot be differentiated from benign SELs simply based on tumour size and EUS appearance [2]. Fortunately, Kim et al. [18] demonstrated that 4 EUS features (inhomogeneity, hyperechogenic spots, marginal halo, and higher echogenicity compared with the surrounding muscle layer appearing more frequently) are of strong relevance to differentiating gastric GISTs from gastric leiomyomas with a sensitivity and specificity of 89%

Table 1
Comparison of EUS features and distribution between GISTs and leiomyomas

Subepithelial lesion	EUS layer	Size	Echogenicity	Border	Distribution in the GI tract	Malignant potential
Leiomyoma	4^{th} or 2^{nd}	Differing sizes	Hypoechoic and nearly similar to the muscle layer, homogeneous	Regular	Mostly oesophagus and stomach, but can occur anywhere in the GI tract	None
Schwannoma	3^{rd} or 4^{nd}	Differing sizes	Hypoechoic, well-demarcated, homogeneous	Regular	Stomach 70%, colon and rectum 15%	Extremely rare
GIST (very low risk, low risk)	4^{th} or 2^{nd}	Small (≤2 cm)	Hypoechoic but relatively hyperechoic compared to muscle layer, homogeneous	Regular	Stomach 60%, small bowel 35%, esophagus 5%, rectum 5%	10-30% clinically malignant
GIST (intermediate, high and very high risk)		Large (>3-5 cm)	Hypoechoic, heterogeneous, cystic spaces, hypervascularity, marginal halo, hyperechoic spots/echogenic foci	Irregular		

EUS: endoscopic ultrasound; GIST: gastrointestinal stromal; GI: gastrointestinal. Correspondence between EUS and histological layers: superficial mucosa (1^{st}), deep mucosa and muscularis mucosa (2^{nd}), submucosa (3^{rd}), muscularis propria (4^{th}), Serosa/adventitia (5^{th}) [17]. Complete comparisons see [3, 18]

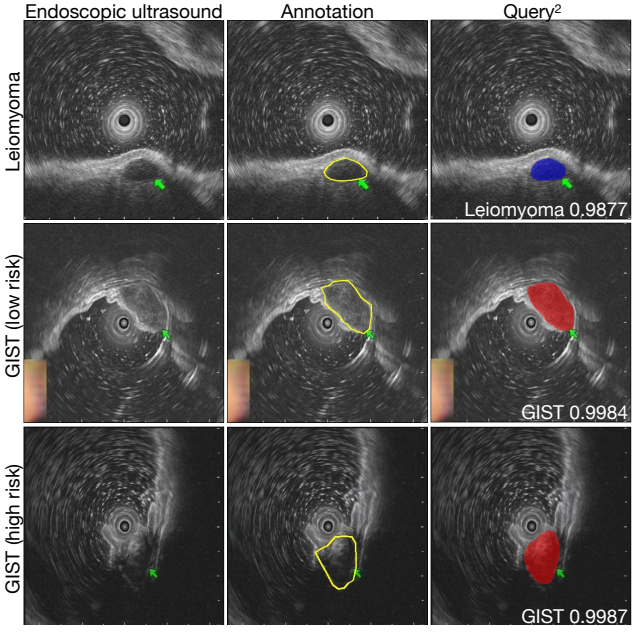

Figure 2: Examples of EUS images in GIST514-DB. The first row shows a typical image of leiomyoma (8.3 mm x 5.3 mm) originating from the 4^{th} layer in the gastric fundus. The second row shows a typical image of a GIST (13.2 mm x 6.6 mm) with malignant potential originating from the 4^{th} layer in the gastric fundus. The third row shows a typical image of malignant GIST (11.0 mm x 8.5 mm) originating from the 4^{th} layer in the gastric fundus. We present the result of Query^2 on segmentation and classification in the third column.

and 86%, respectively. Karaca et al. [4] further supported that EUS can contribute to the diagnosis of SELs in visualising tumour size and extent, but found that it had a relatively low accuracy at 45% for presumptive diagnosis. In 2015, Eckardt et al. [3] proposed a EUS-guided approach to the evaluation and management of SEL and offered a detailed analysis of EUS features and anatomical sites of common SELs, such as leiomyoma, schwannoma and GIST. We summarise and compare the different EUS features and distributions that distinguish GIST from leiomyoma and schwannoma in Table 1, which shows that there are substantial differences between high-risk GISTs and non-GISTs in size, echogenicity and border. Unfortunately, it also states that it is challenging to distinguish very low- and low-risk GISTs from non-GISTs simply on the basis of EUS layer, size, echogenicity, and border since their distributions are highly overlapped on these four dimensions.

2.2. CAD applications and datasets to recognize GISTs

Related methods to recognise GISTs can be roughly divided into two categories based on their tasks, e.g., classification [9, 10, 12] and detection [11] (see Table 5). Minoda et al. [9] proposed a complete workflow including data collection and construction of the EUS-AI leveraging Xception and showed the model's capability to differentiate GISTs from non-GIST SELs larger than or equal to 20 mm with 93.3% accuracy. However, for SELs smaller than 20 mm, the recognition performance is far from satisfactory. Similarly, Kim et al. [10] proposed a convolutional neural network (CNN)-based CAD system to classify GISTs and non-GISTs on a small dataset using a 6-layer CNN. Hirai et al. [12] also collected a larger dataset of 631 valid cases from 12 hospitals and experimentally demonstrated the efficiency of EfficientNet for discriminating GISTs from non-GISTs, such as leiomyomas, schwannomas, neuroendocrine tumours and ectopic pancreases on EUS images. In contrast to previously mentioned methods using object classification models, Oh et al. [11] applied an object detection model, i.e., EfficientDet, to discriminate GISTs from leiomyomas, which was trained

on images of 114 patients and validated on 54 patients, reaching promising sensitivity, specificity, and accuracy of 95.6%, 82.1%, and 91.2%, respectively. Different from using dense detectors, such as EfficientDet, which requires complex label assignment and post-processing, we build our model on the basis of a sparse detector, allowing us to focus on the design of location features. Moreover, unlike [11], which predicts the GIST class by the top-1 scoring bounding box from detectors, we end-to-end aggregate features from multiple bounding boxes to a single prediction through a multi-head cross-attention module.

Most existing datasets for GIST recognition are not publicly available. We summarise the main characteristics of the datasets in Table 3 from the aforementioned works [9–12]. The mean tumour size for these datasets ranged from 20.0 mm to 34.9 mm, but the mean tumour size of GIST514-DB was considerably smaller at 10 mm, indicating that the main difference between GIST514-DB and other datasets is that there are more very-low risk and low-risk GISTs in it. In addition to the commonly used annotations, such as categories of images [9, 10, 12], [11] also annotated bounding boxes for object detection. In our study, we further annotate contours/masks for instance segmentation and anatomical locations for Query[2]. For the EUS dataset collection, we follow the dataset setting of patient biographies and tumour type in [9] and only collect GISTs and leiomyomas because they look similar in the EUS. Compared with previous datasets [9–12], the category distribution in our GIST514-DB dataset is more uniform, which intuitively brings a more balanced recognisability over different categories. We provide a detailed description of our GIST514-DB in Sec. 3.

2.3. Relationship with Real World Object Recognition

Real-world object recognition mainly consists of three tasks: i.e., classification, object detection, and segmentation. Classification models [5, 7, 19–22] are widely used to solve most image recognition tasks and further extract features for downstream tasks such as object detection. Object detection models [23–29] usually generate proposals before recognition when the location of an object is required or the number of objects is uncertain. Among them, query-based methods [28, 29] that have emerged in recent years have a more elegant and efficient architecture than anchor-based [23, 24, 26] and anchor-free [25, 27] methods because query-based methods are free from notorious post-processing, such as non-maximum suppression. Instance segmentation [23, 26, 29] has been extensively combined with object detection models as a complement to provide object contours. For example, QueryInst [29] is demonstrated as capable of improving the detectability for query-based models while generating instance segmentation results. Thus, we adapt QueryInst as our baseline. Unlike images captured by a common camera, each EUS image in our study only contains at most one object due to the narrow scope of the ultrasound probe. Since at most one object occurs, we are able to improve the model inference by the single-object restriction.

Table 2
Baseline Characteristics of our GIST514-DB

	Leiomyomas (n = 251)	GISTs (n = 263)	P-value
Gender			0.281
Male	112	105	
Female	139	158	
Age (yr, mean±SD)	54.5 ± 10.3	59.9 ± 8.7	0.001
Tumour location			0.001[a]
Esophagus	128	7	
Cardia	18	0	
Fundus	76	202	
Fundus/body[b]	10	0	
Body	15	41	
Angle	0	4	
Antrum	4	9	
Size (mm, mean ± SD)			
Horizontal diameter	10.1 ± 6.0	10.9 ± 5.8	0.125
Longitudinal diameter	6.2 ± 3.6	7.5 ± 4.5	0.001
The tumour risk			
Very low-risk		218	
Low-risk		30	
Intermediate-risk		2	
High-risk		2	
Undetermined[c]		11	

SD: standard deviation; [a]: we map tumour locations to integers (e.g., esophagus: 0, cardia: 1, ..., antrum: 6) when calculating p-value; [b]: ambiguous location; tumour risk: AFIP risk; [c]: some cases did not provide data on tumour risks.

Table 3
Comparison of GIST Datasets in the EUS

Ref	GIST	N.G.	Mean size	Cls.	Bbox.	Seg.	Ana.
[10]	157	91	34.9 mm	✓			
[9]	184	89	20.0 mm	✓			
[12]	435	196	25.6 mm	✓			
[11]	125	43	25.0 mm	✓	✓		
Ours	263	251	10.5 mm	✓	✓	✓	✓

Ours: GIST514-DB; N.G: non-GISTs; Cls.: label annotations for classification; Bbox.: bounding box annotations for object detection; Seg.: contour annotations for instance segmentation; Ana.: anatomical location annotations for Query[2]. For [9], the median size is taken as the mean size since the mean size is not given.

Consequently, we build our model on the basis of an object detector with an instance segmentation module (see Sec. 4.2.4) and in addition, predict the class of the whole image under the single-object restriction (see Sec. 4.2.3).

3. GIST514-DB: Dataset

3.1. Data Collection and Annotation

The data collection was approved by the local institutional review board. A total of 514 fully anonymised cases in the endoscopy centre of the General Hospital of Tianjin Medical University from June 2016 to October 2021 were retrospectively collected for this study. These GIST cases and leiomyoma cases were randomly selected without any specific selection and exclusion criteria. Our data collection is designed for the current workflows to make better use of existing data. Before tissue acquisition, such as endoscopic submucosal dissection or biopsy, the EUS assessment procedure was performed using a GI endoscopy system (CV-260SL, Olympus, Tokyo, Japan) and ultrasonic microprobes (UM-DP20-25R, frequency 20 MHz). The immunohistochemistry analysis was conducted with CD117, CD34, S-100, and DOG-1 [30] on the acquired sample and offered ground truth of pathological classification and risk level.

Considering data scalability, additional information such as gender, age, anatomical location classification obtained from the gastroendoscopy, originating layer, lesion size, pathological classification, risk level and the EUS images were recorded for each case (see Table 2). The EUS images saved by the operator were considered to be the optimal image for that scan, given the limitation of patient position and probe workspace. As the virtual callipers were applied by the EUS operator in most cases to measure the SEL size, images with virtual callipers were saved as a reference to facilitate further annotations on images without callipers. The most relevant image without callipers was extracted from each case to form our dataset, i.e., GIST514-DB. Each image was classified as GIST or leiomyoma, which was acquired from biopsy results. The Labelme annotation tool [1] was then used to manually annotate the data for object detection and segmentation. Two non-clinicians first observed the lesion in paired images having callipers on it and then annotated the lesion contour on the calliper-free image. Two expert clinicians then verified the annotations. The same number of GIST and leiomyoma images were collected, which avoids the class imbalance problem.

3.2. Data splitting

The data on GIST514-DB are equally divided into $K = 5$ folds simply using the following strategy to further balance the number of samples across different anatomical locations and tumour sizes:

$$D^k = \{\bigcup_{c=0}^{C-1}\bigcup_{h=0}^{H-1} D_{c,h}^{(k)} \mid \left|D_{c,h}^{(k)}\right| = \lfloor\frac{|D_{c,h}|}{K}\rfloor, \\ \mathrm{hd}(D_{c,h}) \in [hI,(h+1)I)\} \quad (1)$$

where dataset D is naturally divided into $C = 2$ subsets, leiomyoma D_0 and GISTs D_1. Since the horizontal diameter $\mathrm{hd}(\cdot)$ of GIST in GIST514-DB ranges from 0 to 40 millimetres, each subset D_c is further divided into $H = 4$ smaller subsets $D_{c,h}$ by an interval of $I = 10$ millimetres; then, each subset $D_{c,h}$ is approximately uniformly distributed to $K = 5$ smaller subsets $D_{c,h}^{(k)}$ that collectively form the five final folds $D^{(k)}$. Given two sets of data D_{train} and D_{test}, where $D_{\text{train}} \bigcap D_{\text{test}} = \varnothing$ and $D_{\text{train}} \bigcup D_{\text{test}} = D$, let $D_{\text{test}} = D^{(k)}$ for the evaluation of the k^{th} split. In this way, the data in each fold are unseen to the other folds, of which the distributions are similar to each other, as illustrated in Fig. 3.

[1] https://github.com/wkentaro/labelme

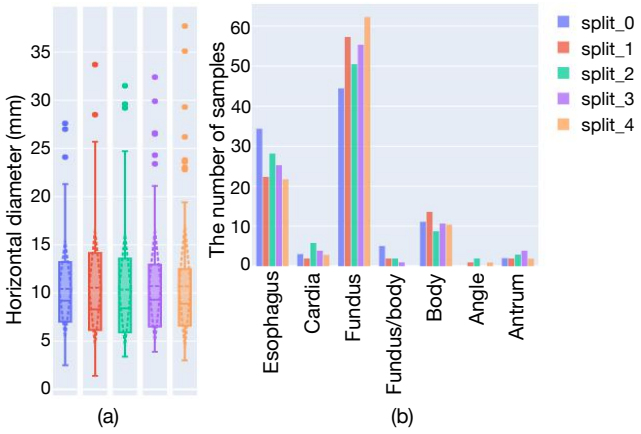

Figure 3: Data spitting into 5 folds based on the tumour size and anatomical location. (a) Distribution of the horizontal diameter, and (b) the anatomical site location containing SEL.

4. Proposed Method

This paper tackles several challenges towards developing automatic approaches for supporting high-accuracy classification in GIST diagnosis. In particular, following clinical practice [3], we present a novel single-object detection architecture (shown in Fig. 4) that exploits the anatomical location of lesions in the upper GI tract. Our main idea is to aggregate multiple queries q_t^{box} on image features x^{FPN} into one query q_t^{img} conditioned on the anatomical location x^{add} and use this to improve accuracy by iterative refinement.

4.1. Problem Formulation

Assume the category of the image represents the category of SEL because there is only one lesion in each EUS image. Given an EUS image x^{img} and an additional input of anatomical location x^{add}, the task designs and trains a model such that it outputs the category of image $\hat{c}$, bounding box $\hat{b}$, and mask $\hat{m}$. The data for each batch are organised as follows:

$$\{x, y\} = \{(x^{\text{img}}, x^{\text{add}}), (c_{\text{gt}}, b_{\text{gt}}, m_{\text{gt}})\} \quad (2)$$

where $x = (x^{\text{img}}, x^{\text{add}})$ denotes a pair consisting of an EUS image $x^{\text{img}} \in \mathbb{R}^{H\times W\times 3}$ and an anatomical location $x^{\text{add}} \in \{0,\dots,(L-1)\}$. There are $L = 7$ anatomical landmarks in our settings, namely, oesophagus, cardia, gastric fundus, undetermined between gastric fundus and gastric body, gastric body, angle and antrum in ascending order (mentioned in Table 2 and Fig. 3). In equation (2), $y = (c_{\text{gt}}, b_{\text{gt}}, m_{\text{gt}})$

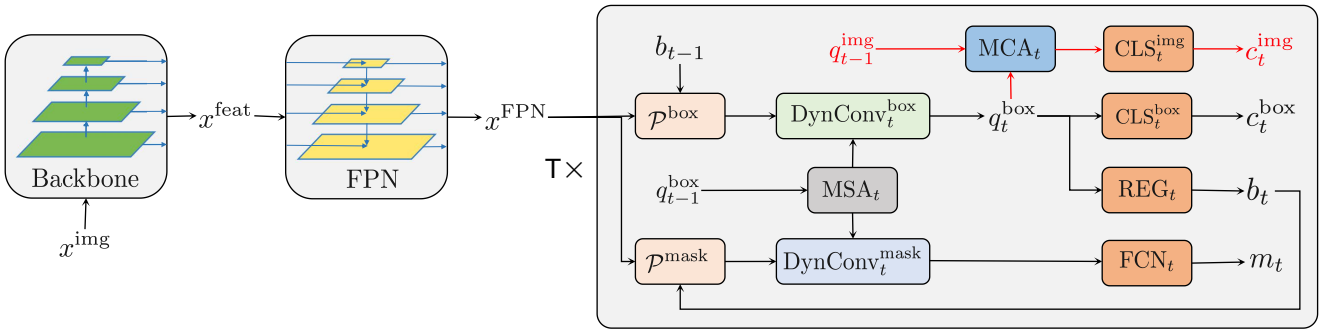

Figure 4: Overall of Query^2. Query^2 consists of T stages ($T = 6$ in our settings). The image feature is first extracted by the backbone and FPN module and then fed to T stages. The submodules for processing anatomical location are highlighted by red arrows.

denotes the ground truth. $c_{\text{gt}} \in \{0, 1\}$ denotes the ground truth of the image category, where 0 indicates benign lesions (leiomyoma) and 1 indicates malignant lesions (GIST), and is used to calculate classification metrics, such as accuracy, sensitivity and specificity. Additionally, $b_{\text{gt}} \in \mathbb{R}^4$ and $m_{\text{gt}} \in \mathbb{R}^{H\times W}$ denote the bounding box and mask, respectively, both of which are further used to evaluate the performance of object detection and instance segmentation.

4.2. Query^2 Architecture

We propose Query^2, a high-performance single-object detector that consists of a query-based backbone for extracting image features, a novel bounding box head driven by single-object restriction for GIST detection, and an additional mask head for GIST segmentation. Our strategy is also suitable for query-based object detectors, such as Sparse-RCNN [28] and QueryInst [29], *where we choose QueryInst as our baseline*. The overall architecture and experimental settings of Query^2 are introduced below.

4.2.1. Initial State with Anatomical Location Embedding

Query^2 consists of $T = 6$ cascade stages, which means that the output of the current stage is fed to the input of the next stage, e.g., the input of stage $t \in \{1, \dots, (T-1)\}$, such as the bounding box proposals b_{t-1}, bounding box query q_{t-1}^{box} and whole image query q_{t-1}^{img}, are generated from stage $t-1$. However, the input of the first stage is initialised by learnable parameters or embedding layers separately, which are formulated as follows:

$$\begin{aligned} b_0 &\leftarrow \theta^{\text{box}} \\ q_0^{\text{box}} &\leftarrow \theta^{\text{feat}} \\ q_0^{\text{img}} &\leftarrow \mathcal{E}_{\theta^{\text{img}}}(x^{\text{add}}) \end{aligned} \tag{3}$$

$b_0 \in \mathbb{R}^{N\times 4}$ is initialised by the encoded position of N bounding boxes $\theta^{\text{box}} \in \mathbb{R}^{N\times 4}$. Similarly, the bounding box query for the first stage $q_0^{\text{box}} \in \mathbb{R}^{N\times d}$ is initialised by N learnable parameters $\theta^{\text{feat}} \in \mathbb{R}^{N\times d}$ of length d representing bounding box features. $q_0^{\text{img}} \in \mathbb{R}^{1\times d}$ is initialised by *the anatomical location embedding layer* $\mathcal{E}$, where anatomical location x^{add} serves as the index of learnable parameters $\theta^{\text{img}} \in \mathbb{R}^{L\times d}$. In our experiments, θ^{box}, θ^{feat} and θ^{img} are implemented in PyTorch by the same function nn.Embeding.

4.2.2. Backbone and Neck

After we initialised the queries for the first stage above, we used the backbone and neck to extract image features from EUS images. The image feature extractor of the object detector can be formulated as follows:

$$\begin{aligned} x^{\text{feat}} &\leftarrow \text{Backbone}(x^{\text{img}}) \\ x^{\text{FPN}} &\leftarrow \text{FPN}(x^{\text{feat}}) \\ x_t^{\text{box}} &\leftarrow \mathcal{P}^{\text{box}}(x^{\text{FPN}}, b_{t-1}) \\ x_t^{\text{mask}} &\leftarrow \mathcal{P}^{\text{mask}}(x^{\text{FPN}}, b_t) \end{aligned} \tag{4}$$

Backbone denotes the main feature extractor, such as ResNet-50 or ResNet-101 [5], converting static images x^{img} into image features x^{feat} for the downstream tasks. FPN denotes the feature pyramid networks [31] allowing the pooling operators $\mathcal{P}^{\text{box}}$ and $\mathcal{P}^{\text{mask}}$ i.e., Region of Interest Align (RoIAlign) [23] to crop current bounding box features x_t^{box} and mask features x_t^{mask} from FPN features x^{FPN}. The pooling positions of this step are controlled by the bounding box prediction b_{t-1} from the previous stage and b_t from the current stage.

4.2.3. Bounding Box Head with Single-Object Restriction

We first build a vanilla bounding box prediction pipeline, which can be expressed as follows:

$$\begin{aligned} q_{t-1}^{\text{box}*} &\leftarrow \text{MSA}_t(q_{t-1}^{\text{box}}) \\ q_t^{\text{box}} &\leftarrow \text{DynConv}_t^{\text{box}}(x_t^{\text{box}}, q_{t-1}^{\text{box}*}) \\ c_t^{\text{box}} &\leftarrow \text{CLS}_t^{\text{box}}(q_t^{\text{box}}) \\ b_t &\leftarrow \text{REG}_t(q_t^{\text{box}}) \end{aligned} \tag{5}$$

where a multi-head self-attention (MSA) module MSA$_t$ [32] (see Fig. 5(a)) is applied to the bounding box query q_{t-1} from the last stage to obtain the enhanced query $q_{t-1}^{\text{box}*}$. DynConv$_t^{\text{box}}$ [28] (see Fig. 5(c)) denotes the box dynamic convolution module taking the bounding box features x_t^{box}

and the enhanced query $q_{t-1}^{\text{box}*}$ of the last stage as input while generating a bounding box query q_t^{box} for the next stage. The bounding box query q_t^{box} is then fed into the bounding box branch, consisting of a vanilla fully connected head of bounding box classification $\text{CLS}_t^{\text{box}}$ and a vanilla fully connected head of bounding box regression REG_t to generate the category prediction of bounding box $c_t^{\text{box}} \in \mathbb{R}^{N\times 2}$ and the position prediction of bounding box $b_t \in \mathbb{R}^{N\times 4}$. The vanilla fully connected head usually consists of several linear layers followed by layer normalisation and activation functions.

However, the vanilla bounding box prediction pipeline of query-based detectors, such as QueryInst [29] and Sparse R-CNN [28], and other nonquery-based detectors, such as EfficientDet [8, 11] and Cascade Mask R-CNN [26], cannot predict image category end-to-end because its prediction is generated by manual post-processing, leaving only the top scoring bounding box [11] and can be expressed as follows:

$$\begin{aligned} \hat{c} &\leftarrow \arg\max_j c_{T-1(i,j)}^{\text{box}} \\ \hat{b} &\leftarrow b_{T-1} \end{aligned} \tag{6}$$

The image category prediction is based on the top scoring bounding box from the category prediction of bounding box c_{T-1}^{box} in the last stage. The position predictions of bounding box b_{T-1} in the last stage are taken as the final output. In contrast, we propose a simple but end-to-end *multi-head cross-attention (MCA)* module to replace this manual post-processing and implement *single-object restriction (SOR)*, which can be expressed as follows:

$$\begin{aligned} q_t^{\text{img}} &\leftarrow \text{MCA}_t(q_{t-1}^{\text{img}}, q_t^{\text{box}}) \\ c_t^{\text{img}} &\leftarrow \text{CLS}_t^{\text{img}}(q_t^{\text{img}}) \\ \hat{c} &\leftarrow c_{T-1}^{\text{img}} \\ \hat{b} &\leftarrow b_{T-1} \end{aligned} \tag{7}$$

where a multi-head cross-attention module MCA_t [32] (see Fig. 5(b)) transforms the whole image query q_{t-1}^{img} by reading q_t^{box}. The whole image query q_t^{img} of the current stage is then fed into a vanilla fully connected layer head $\text{CLS}_t^{\text{img}}$ for image classification and converted to the image category prediction c_t^{img} of the current stage.

4.2.4. Mask Head

Following the settings of the mask head in QueryIns [29], the mask head is formulated as follows:

$$\begin{aligned} x_t^{\text{mask}*} &\leftarrow \text{DynConv}_t^{\text{mask}}(x_t^{\text{mask}}, q_{t-1}^{\text{box}*}) \\ m_t &\leftarrow \text{FCN}_t(x_t^{\text{mask}*}) \\ \hat{m} &\leftarrow m_{T-1}^{(i)} \end{aligned} \tag{8}$$

where $\text{DynConv}_t^{\text{mask}}$ [29] (see Fig. 5(d)) helps the bounding box head be partially driven by the mask head since it takes x_t^{mask} and $q_{t-1}^{\text{box}*}$ as input while generating enhanced mask features $x_t^{\text{mask}*}$ to the following mask head FCN_t. FCN_t consists of four sequentially connected convolution layers, an upsampling layer and a convolution layer, generating mask predictions m_t for instance segmentation tasks.

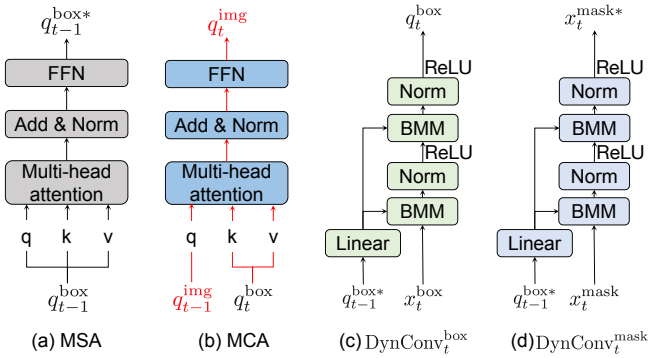

Figure 5: Illustrations of MSA, MCA, $\text{DynConv}_t^{\text{box}}$ and $\text{DynConv}_t^{\text{mask}}$ at stage t. (a) q_{t-1}^{box} is transformed by itself. (b) q_{t-1}^{img} is transformed by reading q_t^{box}. (c) x_t^{box} is enhanced by $q_{t-1}^{\text{box}*}$, while the output q_t^{box} serves as both new bounding box features for the current stage and new queries for the next stage. (d) x_t^{mask} is enhanced by $q_{t-1}^{\text{box}*}$. BMM: batch matrix multiplication.

5. Experimental Results

5.1. Implementation Details

Our implementation is in PyTorch and developed using MMDetection [33]. To additionally evaluate general classification models, we implemented the classification methods using MMClassification [34]. The training and inference of our models are supported by four Titan Xp GPUs, each having 12 GB memory.

5.1.1. Dataset Setup

Our models are evaluated on GIST514-DB by 5-fold cross validation based on the proposed data splitting setting described in Sec. 3.2. Input images are resized following the random scale settings in [29] during training and inference. To avoid overfitting, we introduce data augmentation adding random flipping and rotations at 90, 180 or 270 degrees since SELs can appear at any orientation. Note that brightness and texture, such as marginal halo and inhomogeneity of the tumour, show strong relevance with GIST classification [18]. Therefore, to retain these intensity features, we did not apply brightness augmentation and only applied geometric transformation.

5.1.2. Training Setup

Query^2 is first initialised with the weight of QueryInst [29] pretrained on the COCO dataset and then fine-tuned on GIST514-DB because the number of images in GIST514-DB is too small to support training from scratch. Without special mention, all detectors in our study are pretrained on COCO, and all classifiers are pretrained on ImageNet. Following query-based methods [28, 29], the training schedule is 36 epochs. The learning rate is warmed up from 2.5×10^{-8}

to 2.5×10^{-5} in the first 1000 iterations and then divided by 10 at the 27-th epoch and 33-th epoch. The weight decay of the AdamW optimiser is set to 1×10^{-4}. The 6 stages in our models are parallel driven by:

$$\begin{aligned} loss_t \leftarrow & w_0 \cdot \mathrm{L1}(b_t, b_{\mathrm{gt}}) + w_1 \cdot \mathrm{GIoU}(b_t, b_{\mathrm{gt}}) \\ & + w_2 \cdot \mathrm{FL}(c_t^{\mathrm{box}}, c_{\mathrm{gt}}) + w_3 \cdot \mathrm{CE}(c_t^{\mathrm{img}}, c_{\mathrm{gt}}) \\ & + w_4 \cdot \mathrm{DSC}(m_t, m_{\mathrm{gt}}) \end{aligned} \tag{9}$$

where $w_{(i)}$ denote the weights for the five loss functions. Following the setting of QueryInst[29], $w_{(i)}$ are set to 5, 2, 2, 1 and 8. Concretely, the encoded position of bounding box b_t is guided by the least absolute deviations (L1) loss function and generalised intersection over union (GIoU) loss function [35]. The category of bounding box c_t is supervised by the focal loss function [36]. The cross-entropy loss function CE is additionally applied to drive the classification head with single-object restriction. The dice coefficient (DSC) loss function is used to maximise the overlap between m_t and the ground truth. The number of proposals (N) referenced in Sec. 4.2.1 is set to 300.

5.1.3. Evaluation Metrics

Following the tasks of real world object detection and instance segmentation, mean average precision (mAP) is used to evaluate the performance of detectors. mAP indicates the average precision for IoU from 0.5 to 0.95 with a step size of 0.05, and AP_{50} denotes average precision acquired at IoU threshold 0.5. Importantly, we compared the sensitivity (Sen.), specificity (Spc.) and accuracy (Acc.) over previous methods by treating GIST recognition as a classification task. In this task, Sen. refers to GIST accuracy, Spc. refers to leiomyoma accuracy, and Acc. refers to the overall accuracy. In practice, we accumulate the confusion matrix from each split to calculate the sensitivity, specificity and accuracy for the entire dataset. Since the accumulation of mAP is complicated, we just average the mAP of each split to evaluate the entire dataset. We first repeated the ablation study 5 times with random seeds and reported the mean and standard deviation in Table 4. We then reported the accuracy with the best seed in subsequent sections.

5.2. Main Results

To validate the performance of the proposed Query^2 method, we performed an ablation study of anatomical location embedding layers and performed an extensive comparison with the existing CAD applications [9–12], classification models [5, 7, 19–22], object detectors [23–29], and instance segmentation [23, 26, 29].

5.2.1. Ablation study of the anatomical location embedding layer, MCA and SOR

MCA is essential because it is the only connection between anatomical location input and the rest of the model. We demonstrate that additional anatomical location input is crucial to classification performance. To evaluate the case of an embedding layer without the anatomical location input, we fix the index of the anatomical location embedding layer, i.e., $q_0^{\mathrm{img}} \leftarrow \mathcal{E}_{\theta^{\mathrm{img}}}(0)$, so that the MCA receives the same retrieval result from the embedding layer. As shown in Table 4, the removal of anatomical location input leads to a drop in classification accuracy, where the accuracy of the fifth model drops from 94.3% to 93.2%. Additionally, it is equally important to use SOR during inference. Models without SOR during inference are evaluated by computing the accuracy of the top scoring bounding box. As shown in Table 4, due to the existence of SOR during inference, the accuracy of the fifth model is 1.3% higher than that of the fourth model.

Table 4
Impacts of using the anatomical location embedding layer, MCA and SOR

$\mathcal{E}$	MCA	SOR	μ^{Sen}	σ^{Sen}	μ^{Spc}	σ^{Spc}	μ^{Acc}	σ^{Acc}
			92.4%	1.5%	93.0%	1.5%	92.7%	0.7%
$\mathcal{E}(0)$	✓		91.7%	1.1%	94.4%	1.0%	93.0%	0.5%
$\mathcal{E}(0)$	✓	✓	92.0%	0.9%	94.3%	1.5%	93.2%	1.0%
✓	✓		90.7%	0.6%	95.4%	0.8%	93.0%	0.5%
✓	✓	✓	92.2%	0.5%	96.5%	1.1%	94.3%	0.5%

$\mathcal{E}$: anatomical location of the embedding layer; MCA: multi-head cross-attention module; SOR: single-object restriction; μ: average for 5 repeated experiments; σ: standard deviation for 5 repeated experiments. The first model is baseline. The second and third models utilise $\mathcal{E}$, MCA and SOR without anatomical location input during training, but only the third model uses SOR during inference. The fourth and fifth model utilise $\mathcal{E}$, MCA and SOR with anatomical location input during training, but only the fifth model uses SOR during inference.

Table 5
Performances of CAD applications to recognize GISTs via EUS

Algorithm type	Dataset	Sen.	Spc.	Acc.↑
6-layer CNN[10]	[10]	83.0%	75.5%	79.2%
Xception[9]	[9][a]	77.3%	100%	83.3%
EfficientNetV2-L[12]	[12]	98.8%	67.6%	89.3%
EfficientDet[11]	[11]	95.6%	82.1%	91.2%
Xception[9]	[9][b]	91.7%	100%	93.3%
EfficientDet[11]	[11][c]	100%	85.1%	96.3%
6-layer CNN[10]⋆	GIST514-DB	61.2%	61.8%	61.5%
EfficientDet[11]⋆		81.4%	49.4%	65.8%
Xception[9]⋆		70.7%	76.9%	73.7%
EfficientNetV2-L[12]⋆		80.2%	81.3%	80.7%
Query^2 (Ours)	GIST514-DB	94.3%	96.0%	95.1%

CAD: computer assisted diagnosis; GIST: gastrointestinal stromal; EUS: endoscopic ultrasound; Sen.: sensitivity; Spc.: specificity; Acc.: accuracy; ⋆: reproduced methods; [a]: SELs < 20 mm; [b]: SELs ≥ 20 mm; [c]: evaluated on more than one image(s) for each case.

Table 6
Classification results on GIST514-DB.

Method	Resolution (H×W)	Epochs	#Params	#FLOPS	Sen.	Spc.	Acc.↑
SE-ResNet-101 [19]	224×224	100	47.03M	7.86G	82.5%	78.1%	80.4%
EfficientNet-b1 [7]	240×240	100	6.51M	0.03G	82.5%	78.5%	80.5%
Res2Net-101-26w-4s [20]	224×224	100	42.94M	8.13G	82.9%	79.7%	81.3%
ResNet-101 [5]	224×224	100	42.28M	7.85G	84.8%	80.5%	82.7%
ResNet-152 [5]	224×224	100	57.92M	11.58G	85.2%	80.6%	82.9%
Swin-B [21]	224×224	300	84.1M	15.14G	94.7%	71.4%	83.3%
EfficientNet-b3 [7]	300×300	100	10.7M	0.06G	92.0%	83.8%	87.9%
VGG-19-BN [22]	224×224	100	139.55M	19.69G	88.2%	88.0%	88.1%
Query^2(ours)	800×1280	36	200.36M	241.66G	94.3%	96.0%	95.1%

FLOPs are calculated under the input resolution illustrated on this table.

Table 7
Object detection results on GIST514-DB.

Method	Backbone	Epochs	#Params	#FLOPS	mAP^{box}↑	AP$^{box}_{50}$	FPS
Non-query based							
Mask R-CNN[23]	R-101-FPN	36	62.74M	334.24G	41.4	69.8	11.1
ATSS [24]	R-101-FPN	36	50.88M	277.53G	43.2	69.2	13.7
FCOS [25]	X-101-FPN	36	89.61M	434.75G	44.2	71.0	6.7
Cascade Mask R-CNN [26]	R-101-FPN	36	95.79M	465.04G	44.4	70.0	7.7
RepPoints [27]	X-101-FPN-DCN	36	57.81M	271.58G	46.9	75.3	6.7
Query based							
Sparse R-CNN [28]	R-101-FPN	36	124.99M	241.53G	53.9	86.5	9.1
QueryInst [29]	R-101-FPN	36	191.27M	241.53G	54.8	88.6	8.8
Query^2(ours)	R-101-FPN	36	200.36M	241.66G	55.8	88.8	8.4

FLOPs are calculated under the input resolution of 800×1280. R: ResNet; X: ResNext; DCN: deformable convolution.

5.2.2. Comparisons with the most relevant CAD applications

Table 5 shows the classification results of the most relevant CAD applications [9–12], where neither the code nor the datasets used for evaluation in these methods are publicly available. Therefore, we summarised the performance of the existing CAD methods on their private datasets as reported in their respective papers. For a fair comparison, we reproduce all of these methods in PyTorch and evaluate all of these methods on the GIST514-DB dataset. Following the resolution settings in these models, the images input to 6-layer CNN and Xception are cropped by the ground truth bounding boxes, while pixels outside the segmentation mask of images input to 6-layer CNN are removed.

5.2.3. Comparisons on GIST514-DB Classification

We compare Query^2 with the state-of-the-art classification methods of real world datasets (see Table 6) on GIST514-DB. The resolution, optimisers and learning schemes of each model are aligned with the default settings of their pretrained models. For fine-tuning, we froze the first stage of each model, divided the learning rate by 10, and applied label smoothing [6]. The accuracies of the SOTA classification models range from 80.4% to 88.1%. The experimental result suggests that VGG-19-BN outperforms other classification models with 88.1% accuracy, but Query^2 still considerably exceeds the accuracy of VGG-19-BN by 7%. We also provide the number of parameters and FLOPs to facilitate the selection of the optimal strategy for using the GIST514-DB dataset and the method on performance-constrained platforms. Classification models may perform worse than detectors because detectors are supervised by additional bounding box annotations. To validate this, we conducted extensive experiments on object detection models (Sec. 5.2.4).

5.2.4. Comparisons with GIST514-DB Object Detection

We first evaluate detection models by metrics commonly applied to the classification model, such as sensitive, specificity and accuracy, to fairly compare their performance with the aforementioned state-of-the-art classification models (see Table 8). Following [11], we also use the category of the top scoring bounding box to represent the classification results of these detectors, as shown in Eq. (6). The experimental results show that nonquery-based detectors,

Table 8
Performances of adapting object detectors to classification on GIST514-DB.

Method	Sen.	Spc.	Acc.↑
Non-query based			
FCOS [25]	81.4%	62.5%	72.2%
ATSS [24]	80.6%	67.7%	74.3%
Mask R-CNN[23]	81.7%	67.7%	74.9%
Cascade Mask R-CNN [26]	80.2%	69.7%	75.1%
RepPoints [27]	82.5%	72.9%	77.8%
Query based			
Sparse R-CNN [28]	91.2%	93.6%	92.4%
QueryInst [29]	92.0%	93.6%	92.8%
Query^2(ours)	94.3%	96.0%	95.1%

Table 9
Instance segmentation results on GIST514-DB.

Method	mAP$^{\text{mask}}$↑	AP$^{\text{mask}}_{50}$
Non-query based		
Mask R-CNN[23]	44.4	70.5
Cascade Mask R-CNN [26]	46.8	70.9
Query based		
QueryInst [29]	56.4	89.3
Query^2(ours)	57.4	89.0

such as FCOS, ATSS, Mask R-CNN, Cascade Mask R-CNN and Repoints, achieve accuracies between 72.2% and 77.8%, which are even lower than the worst state-of-the-art classification model illustrated in Table 6 (e.g., SE-ResNet-101 with 80.4% accuracy). In contrast, query-based detectors are experimentally shown to be at least 4.3% more accurate than the best state-of-the-art classification model, such as VGG-19-BN.

Table 7 shows the object detection results on GIST514-DB, evaluating bounding boxes and categories. The experimental results show that the query-based detectors show better localisation capabilities than the nonquery-based architectures. Sparse R-CNN, QueryInst and Query^2 surpass RepPoints by 7, 7.9 and 8.9 AP$^{\text{box}}$, respectively. Query^2 achieves 8.4 FPS on a single Titan Xp GPU during inference, which still surpasses several strong baselines. Overall, Query^2 achieves the best performance (55.8 AP$^{\text{box}}$), suggesting that it better models the task than the existing state-of-the-art methods.

5.2.5. Comparisons with GIST514-DB Instance Segmentation

Additionally, Table 9 shows the instance segmentation results on GIST514-DB, which shows that Query^2 improves instance segmentation performance by 1% mAP$^{\text{mask}}$ compared with the state-of-the-art QueryInst method. Moreover, Query^2 outperformed nonquery-based detectors (Mask R-CNN and Cascade Mask R-CNN).

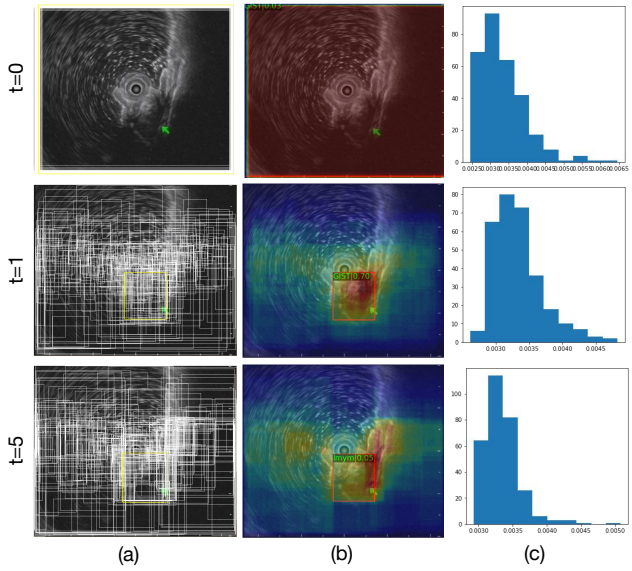

Figure 6: Visualising the distribution of bounding boxes in different stages. (a) Distribution of b_t, where the yellow box represents the assigned target box, (b) heatmap of b_t with the box with the highest MCA attention weight highlighted in red, where c_t^{box} is marked in green on the upper left corner of these boxes and (c) the histogram of the MCA attention weights.

5.2.6. Visualising the distribution of bounding boxes

The distribution of bounding boxes in each stage is illustrated in Fig. 6(a). To simplify the visualisation, the number of bounding boxes on each pixel is taken as the intensity of the heatmap, as shown in Fig. 6(b). The histogram of averaged MCA attention weights over different attention heads is shown in Fig. 6(c).

6. Discussion

SEL recognition using ultrasonic micro probes plays an indispensable role in gastroendoscopy, especially for the diagnosis of lesions with malignant potential, such as GISTs. Although several sonographic features have been shown to be relevant for high-risk GISTs, the existing sonographic features of low-risk GISTs remain limited [2–4, 18]; thus, the diagnosis of GISTs is a challenging problem. We conduct extensive experiments and prove that Query^2 outperforms existing methods with a large margin in classification, object detection and instance segmentation on the GIST514-DB dataset. Its superior performance is derived from two parts:

(i) spatial and semantic information are captured and aggregated by the end-to-end pipeline;
(ii) made full use of existing annotations to locate lesions and mimic the distribution of lesions using anatomical location.

Spatial information refers to the location of the lesion, which is coded with a bounding box in this study. In the pipeline of nonquery-based detectors, the positions of anchors that generate proposals are evenly distributed across different image regions. In fact, in the GIST514-DB dataset,

the target lesions are located more in a central region of the ultrasonic image than in a border region, which means that different regions have different importance. To define a box-wise importance score, Query^2 explicitly fits the distribution of the bounding box from training data by the bounding box proposals b_t and scores the importance of each generated bounding box query q_t^{box} through MCA aggregation. Concretely, the density of b_t is the approximation of the distribution of the bounding box, while the learned attention of MCA represents the box-wise importance score. As shown in Fig. 6(b), the distribution of b_t is concentrated in the lesion area, and the proposal with the top importance score is located in the most sampled areas. Notably, the bounding box proposal with the highest MCA attention weight is not necessarily the bounding box proposal with the highest prediction score c_t^{box}. The attention weights are also not concentrated in the boxes with the top MCA attention weight, as shown in Fig. 6(c), which implies that the contribution of each query q_t^{box} to c_t^{img} is relatively uniform. Therefore, the aggregation over all bounding boxes with MCA and SOR provides better performance than the top scoring bounding box, which explains why the fifth model in Table 4 performs better than the fourth model in Table 4. Moreover, utilising anatomical location input in a visual detection task is a key finding in this work, where we perform an ablation study on the effect of anatomical location input on GIST recognition. As shown in Table 4, the model with anatomical location input is compared fairly with the model without anatomical location input, showing that anatomical location input leads to better performance.

Equally important, as summarised in Table 3, the tumour size in GIST514-DB is considerably smaller than that in previously reported datasets, implying fewer high-risk GISTs in GIST514-DB. To evaluate the difficulty of GIST recognition on the GIST514-DB dataset, we reproduce the most relevant CAD applications. As shown in Table 5, although the comparison is made at the same resolution level, most methods achieve a relatively lower accuracy on GIST514-DB than on their own datasets, which implies that GIST514-DB is a more challenging dataset.

In Table 5 we can also observe that the classification model 6-layer CNN [10] has the lowest accuracy on the GIST514-DB dataset, and EfficientDet [11] has the second lowest accuracy on GIST514-DB. Due to differences in the number of parameters, FLOPs or image cropping, there is not enough evidence to suggest which classification model or detection model can perform better. However, it is worth noting that in Table 5, EfficientNetV2-L without image cropping outperforms the other classification models with image cropping, such as 6-layer CNN and Xception, which implies that image cropping is not a beneficial process for GIST recognition on the GIST514-DB dataset, and our experimental results also support this conclusion. We evaluate a wide range of classification models without image cropping, where most classification models without image cropping outperform the models with image cropping. As shown in Table 6, the classification accuracy of models without image cropping ranges from 80.4% to 88.1%, surpassing the 6-layer CNN with image cropping and Xception with image cropping. After excluding the effect of image cropping, the accuracy of classification models is almost positively correlated with the number of parameters or FLOPs, where EfficientNet is an exception due to its unique operator design.

Furthermore, more annotations, such as bounding boxes and segmentation masks, are expected to lead to better performance. In other words, detection models are expected to achieve better performance than most classification models. However, counter-intuitively, not all detection models outperform the classification models. As shown in Table 8, the accuracy of nonquery-based detectors is substantially lower than the above classification models in Table 6, while query-based detectors outperform most of the classification models. What classification models in Table 6 and query-based detectors in Table 8 have in common is that they consider pixels outside ground-truth bounding boxes for classification. As illustrated in Fig. 4, the query-based detector uses a self-attention module, such as the MCA, to interact between each bounding box query, which also includes bounding box queries outside the lesion region to generate weighted bounding box queries. In such a case, it is not surprising that further aggregation of bounding box queries, i.e., MCA and SOR, can improve the performance of query-based detectors step forward. As shown in Tables 7, 8 and 9, our Query^2 has state-of-the-art performance in detection, classification and instance segmentation on GIST514-DB.

The main limitation of the proposed method is related to interpretability. Since GISTs and Leiomyomas look very similar, it is difficult to make a qualitative comparison of our method. The main uncertainty comes from the noise signal from EUS probe, motion blur, and the low resolution of the EUS probe. From an imaging perspective, denoising methods and super-resolution methods can reduce the uncertainty. On the other hand, since deep learning is a data-driven approach, including more data can directly reduce uncertainty. Furthermore, how to extend the concept of anatomical location to similar applications, such as skin disease identification, is also an interesting question to explore for future work.

7. Conclusion

In this paper, we propose a novel GIST detection network named Query^2, which utilises the prior anatomical location and the prior single object to improve GIST identification. The proposed network is able to detect and segment even fine-grained lesions from the challenging GIST514-DB dataset that we collected. The GIST514-DB dataset is the first multimodal dataset of its kind, which contains detailed tumour locations from EUS, tumour types from biopsies and anatomical locations from endoscopy collected from the endoscopy centre of the General Hospital of Tianjin Medical University. Through an ablation study and extensive comparison with the existing classification, object detection, segmentation, and CAD methods, we show the robustness

and superiority of the proposed Query^2 method. In future work, we aim to construct a multicentre dataset and extend our architecture to more disease categories.

Declaration of competing interest

None declared.

Acknowledgement

This work was supported by National Natural Science Foundation of China (62133010), National Key R&D Program of China (2019YFB1311501), Wellcome/EPSRC Centre for Interventional and Surgical Sciences (WEISS) (203145/Z/16/Z), the Engineering and Physical Sciences Research Council (EPSRC) (EP/P027938/1), and the Major Special Project of Chronic Disease Prevention and Treatment of Tianjin Science and Technology Commission (17ZXMFSY00210).

References

[1] B. P. Rubin, M. C. Heinrich, C. L. Corless, Gastrointestinal stromal tumour, The Lancet 369 (2007) 1731–1741.

[2] M. Polkowski, Endoscopic Ultrasound and Endoscopic Ultrasound-Guided Fine-Needle Biopsy for the Diagnosis of Malignant Submucosal Tumors, Endoscopy 37 (2005) 635–645.

[3] A. J. Eckardt, C. Jenssen, Current endoscopic ultrasound-guided approach to incidental subepithelial lesions: Optimal or optional?, Annals of Gastroenterology (2015) 160–160.

[4] C. Karaca, B. G. Turner, S. Cizginer, D. Forcione, W. Brugge, Accuracy of EUS in the evaluation of small gastric subepithelial lesions, Gastrointestinal Endoscopy 71 (2010) 722–727.

[5] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 770–778.

[6] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, Z. Wojna, Rethinking the inception architecture for computer vision, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 2818–2826.

[7] M. Tan, Q. Le, Efficientnet: Rethinking model scaling for convolutional neural networks, in: International conference on machine learning, PMLR, 2019, pp. 6105–6114.

[8] M. Tan, R. Pang, Q. V. Le, EfficientDet: Scalable and Efficient Object Detection, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 10781–10790.

[9] Y. Minoda, E. Ihara, K. Komori, H. Ogino, Y. Otsuka, T. Chinen, Y. Tsuda, K. Ando, H. Yamamoto, Y. Ogawa, Efficacy of endoscopic ultrasound with artificial intelligence for the diagnosis of gastrointestinal stromal tumors, Journal of Gastroenterology 55 (2020) 1119–1126.

[10] Y. H. Kim, G. H. Kim, K. B. Kim, M. W. Lee, B. E. Lee, D. H. Baek, D. H. Kim, J. C. Park, Application of A Convolutional Neural Network in The Diagnosis of Gastric Mesenchymal Tumors on Endoscopic Ultrasonography Images, Journal of Clinical Medicine 9 (2020) 3162.

[11] C. K. Oh, T. Kim, Y. K. Cho, D. Y. Cheung, B.-I. Lee, Y.-S. Cho, J. I. Kim, M.-G. Choi, H. H. Lee, S. Lee, Convolutional neural network-based object detection model to identify gastrointestinal stromal tumors in endoscopic ultrasound images, Journal of Gastroenterology and Hepatology 36 (2021) 3387–3394.

[12] K. Hirai, T. Kuwahara, K. Furukawa, N. Kakushima, S. Furune, H. Yamamoto, T. Marukawa, H. Asai, K. Matsui, Y. Sasaki, D. Sakai, K. Yamada, T. Nishikawa, D. Hayashi, T. Obayashi, T. Komiyama, E. Ishikawa, T. Sawada, K. Maeda, T. Yamamura, T. Ishikawa, E. Ohno, M. Nakamura, H. Kawashima, M. Ishigami, M. Fujishiro, Artificial intelligence-based diagnosis of upper gastrointestinal subepithelial lesions on endoscopic ultrasonography images, Gastric Cancer 25 (2022) 382–391.

[13] J. Deng, W. Dong, R. Socher, L. Li, Kai Li, Li Fei-Fei, ImageNet: A large-scale hierarchical image database, in: IEEE Conference on Computer Vision and Pattern Recognition, 2009, pp. 248–255.

[14] J. Krause, M. Stark, J. Deng, L. Fei-Fei, 3d object representations for fine-grained categorization, in: 4th International IEEE Workshop on 3D Representation and Recognition (3dRR-13), Sydney, Australia, 2013, pp. 554–561.

[15] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, C. L. Zitnick, Microsoft coco: Common objects in context, in: European conference on computer vision, Springer, 2014, pp. 740–755.

[16] T.-Y. Lin, A. RoyChowdhury, S. Maji, Bilinear convolutional neural networks for fine-grained visual recognition, IEEE transactions on pattern analysis and machine intelligence 40 (2017) 1309–1322.

[17] T. Wiech, A. Walch, M. Werner, Histopathological Classification of Nonneoplastic and Neoplastic Gastrointestinal Submucosal Lesions, Endoscopy 37 (2005) 630–634.

[18] G. H. Kim, D. Y. Park, S. Kim, D. H. Kim, D. H. Kim, C. W. Choi, J. Heo, G. A. Song, Is it possible to differentiate gastric GISTs from gastric leiomyomas by EUS?, World Journal of Gastroenterology 15 (2009) 3376.

[19] J. Hu, L. Shen, G. Sun, Squeeze-and-excitation networks, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 7132–7141.

[20] S.-H. Gao, M.-M. Cheng, K. Zhao, X.-Y. Zhang, M.-H. Yang, P. Torr, Res2net: A new multi-scale backbone architecture, IEEE transactions on pattern analysis and machine intelligence 43 (2019) 652–662.

[21] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, B. Guo, Swin transformer: Hierarchical vision transformer using shifted windows, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 10012–10022.

[22] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, in: International Conference on Learning Representations, 2015, pp. 1–14.

[23] K. He, G. Gkioxari, P. Dollar, R. Girshick, Mask R-CNN, in: IEEE International Conference on Computer Vision, IEEE, 2017, pp. 2980–2988.

[24] S. Zhang, C. Chi, Y. Yao, Z. Lei, S. Z. Li, Bridging the gap between anchor-based and anchor-free detection via adaptive training sample selection, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2020, pp. 9759–9768.

[25] Z. Tian, C. Shen, H. Chen, T. He, Fcos: Fully convolutional one-stage object detection, in: Proceedings of the IEEE/CVF international conference on computer vision, 2019, pp. 9627–9636.

[26] Z. Cai, N. Vasconcelos, Cascade r-cnn: high quality object detection and instance segmentation, IEEE transactions on pattern analysis and machine intelligence 43 (2019) 1483–1498.

[27] Z. Yang, S. Liu, H. Hu, L. Wang, S. Lin, Reppoints: Point set representation for object detection, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 9657–9666.

[28] P. Sun, R. Zhang, Y. Jiang, T. Kong, C. Xu, W. Zhan, M. Tomizuka, L. Li, Z. Yuan, C. Wang, et al., Sparse r-cnn: End-to-end object detection with learnable proposals, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2021, pp. 14454–14463.

[29] Y. Fang, S. Yang, X. Wang, Y. Li, C. Fang, Y. Shan, B. Feng, W. Liu, Instances as queries, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 6910–6919.

[30] C. M. Kelly, L. Gutierrez Sainz, P. Chi, The management of metastatic GIST: Current standard and investigational therapeutics, Journal of Hematology & Oncology 14 (2021) 2.

[31] T.-Y. Lin, P. Dollar, R. Girshick, K. He, B. Hariharan, S. Belongie, Feature pyramid networks for object detection, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017.

[32] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, Advances in neural information processing systems 30 (2017).

[33] K. Chen, J. Wang, J. Pang, Y. Cao, Y. Xiong, X. Li, S. Sun, W. Feng, Z. Liu, J. Xu, Z. Zhang, D. Cheng, C. Zhu, T. Cheng, Q. Zhao, B. Li, X. Lu, R. Zhu, Y. Wu, J. Dai, J. Wang, J. Shi, W. Ouyang, C. C. Loy, D. Lin, MMDetection: Open mmlab detection toolbox and benchmark, arXiv preprint arXiv:1906.07155 (2019).

[34] M. Contributors, Openmmlab's image classification toolbox and benchmark, https://github.com/open-mmlab/mmclassification, 2020.

[35] H. Rezatofighi, N. Tsoi, J. Gwak, A. Sadeghian, I. Reid, S. Savarese, Generalized intersection over union: A metric and a loss for bounding box regression, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2019, pp. 658–666.

[36] T.-Y. Lin, P. Goyal, R. Girshick, K. He, P. Dollar, Focal Loss for Dense Object Detection, in: IEEE International Conference on Computer Vision, IEEE, 2017, pp. 2999–3007.