Multi-Task Learning With Hierarchical Guidance for Locating and Stratifying Submucosal Tumors

Ruifei Zhang, Feng Zhang, Si Qin, Dejun Fan, Chaowei Fang, Jie Ma, Xiang Wan, Guanbin Li, *Member, IEEE*, and Xutao Lin

***Abstract*—Locating and stratifying the submucosal tumor of the digestive tract from endoscopy ultrasound (EUS) images are of vital significance to the preliminary diagnosis of tumors. However, the above problems are challenging, due to the poor appearance contrast between different layers of the digestive tract wall (DTW) and the narrowness of each layer. Few of existing deep-learning based diagnosis algorithms are devised to tackle this issue. In this article, we build a multi-task framework for simultaneously locating and stratifying the submucosal tumor. And considering the awareness of the DTW is critical to the localization and stratification of the tumor, we integrate the DTW segmentation task into the proposed multi-task framework. Except for sharing a common backbone model, the three tasks are explicitly directed with a hierarchical guidance module, in which the probability map of DTW itself is used to locally enhance the feature representation for tumor localization, and the probability maps of DTW and tumor are jointly employed to locally enhance the feature representation for tumor stratification. Moreover, by means of the dynamic class activation map, probability maps of DTW and tumor are reused to enforce the stratification inference process to pay more attention to DTW and tumor regions, contributing to a reliable and interpretable submucosal tumor stratification model. Additionally, considering the relation with respect to other structures is beneficial for stratifying tumors, we devise a graph reasoning module to replenish non-local relation knowledge for the stratification branch. Experiments on a Stomach-Esophagus and an Intestinal EUS dataset prove that our method achieves very appealing performance on both tumor localization and stratification, significantly outperforming state-of-the-art object detection approaches.**

***Index Terms*—Class activation map, endoscopy ultrasound, local feature enhancing, multi-task learning, tumor localization and stratification.**

I. Introduction

ENDOSCOPIC ultrasound (EUS) is widely applied for the diagnosis and treatment of diseases in the digestive tract [1], [2], [3]. It captures the imaging data by means of a miniature high-frequency ultrasound probe attached on the top of the endoscope. The morphological structures of the intracavitary lesions, stratification and cancer infiltration can be visualized in real time. Specifically, as shown in Fig. 1(a), we can observe five layers of the digestive tract wall (DTW) in an EUS image, including mucosa (M), mucosal muscle (MM), submucosa (SM), proper muscle (PM) and adventitia (A). Submucosal tumors refer to lesions originating from the layers below the mucosa of the digestive tract. This article concentrates on the localization and stratification of submucosal tumors. Automatic localization of submucosal tumors is very valuable since it can relieve the burden of radiologists and avoid misdiagnosis caused by all underlying factors that lead to human error such as tiredness. In practical diagnosis, the identification of the layer where the submucosal tumor originates is an essential step for tumor categorization, e.g. lipomyoma and mesenchymoma reside in the SM and PM layer respectively. We define this problem as tumor stratification which is paramount to in-depth diagnosis analysis. In experiments, we only consider three layers including MM, SM and PM since submucosal tumors originating from the adventitia layer are extremely rare.

With the vigorous development of deep learning techniques based on convolutional neural networks (CNNs), computer-aided diagnosis has already become an important role in clinical scenes, e.g. disease recognition in CT images [4], [5] and lesion

Manuscript received 30 September 2021; revised 12 September 2022 and 30 April 2023; accepted 22 June 2023. Date of publication 17 July 2023; date of current version 6 September 2023. This work was supported in part by the Chinese Key-Area Research and Development Program of Guangdong Province under Grant 2020B0101350001, in part by the Guangdong Basic and Applied Basic Research Foundation under Grant 2020B1515020048, in part by the National Natural Science Foundation of China under Grant 61976250, in part by the Shenzhen Science and Technology Program under Grants JCYJ20220530141211024 and JCYJ20220818103001002, in part by the National Key Clinical Discipline, and in part by the Guangdong Provincial Key Laboratory of Big Data Computing, The Chinese University of Hong Kong, Shenzhen. *(Corresponding author: Guanbin Li; Xutao Lin.)*

Ruifei Zhang, Jie Ma, and Guanbin Li are with the School of Computer Science and Engineering, Sun Yat-sen University, Guangzhou 510006, China (e-mail: zhangrf23@mail2.sysu.edu.cn; majie25@mail2.sysu.edu.cn; liguanbin@mail.sysu.edu.cn).

Feng Zhang is with the Department of Rheumatology, Sixth Affiliated Hospital, Sun Yat-sen University, Guangzhou 510655, China (e-mail: zhangf229@mail.sysu.edu.cn).

Si Qin is with the Department of Medical Ultrasonics, Sixth Affiliated Hospital, Sun Yat-sen University, Guangzhou 510655, China (e-mail: qins5@mail.sysu.edu.cn).

Dejun Fan and Xutao Lin are with the Department of Gastrointestinal Endoscopy, Sixth Affiliated Hospital, Sun Yat-sen University, Guangzhou 510655, China, and with the Guangdong Provincial Key Laboratory of Colorectal and Pelvic Floor Diseases, Sixth Affiliated Hospital, Sun Yat-sen University, Guangzhou 510655, China, and also with the Biomedical Innovation Center, Sixth Affiliated Hospital, Sun Yat-sen University, Guangzhou 510655, China (e-mail: fandj3@mail.sysu.edu.cn; linxt23@mail.sysu.edu.cn).

Chaowei Fang is with the School of Artificial Intelligence, Xidian University, Xi'an, Shaanxi 710071, China (e-mail: cwfang@xidian.edu.cn).

Xiang Wan is with the Shenzhen Research Institute of Big Data, Shenzhen 518000, China (e-mail: wanxiang@sribd.cn).

Digital Object Identifier 10.1109/JBHI.2023.3291433

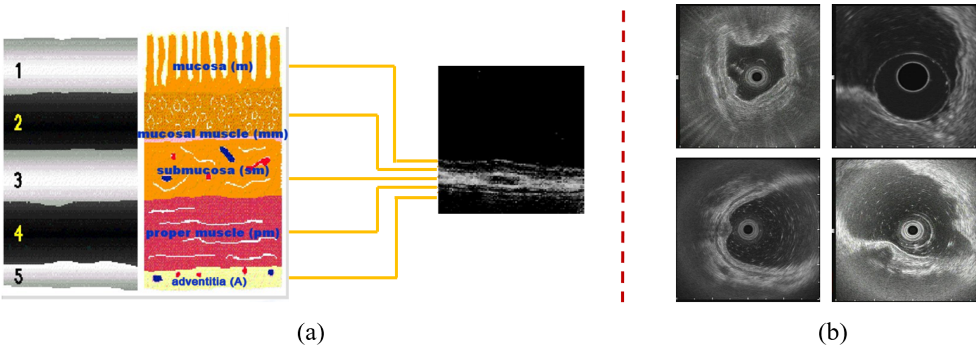

Fig. 1. (a) Hierarchical structure of the digestive tract wall (DTW). (b) Some examples of our endoscopic ultrasound (EUS) image dataset.

detection in chest X-ray images [6], [7]. However, CNN-based tumor localization and stratification in EUS images is still under explored. The work most related to ours is [8], which aims at analyzing gastric mesenchymal tumors in EUS images, differentiating gastrointestinal stromal tumors from benign mesenchymal tumors. Locating and stratifying submucosal tumors are challenging due to the low visual contrast between different structures in EUS images and the tiny thickness of DTW layers. A straightforward method to tackle the tasks is to regard the stratification as a three-way classification problem, and then mechanically leverage general object detection algorithms, e.g. Faster R-CNN [9] and YOLO [10]. However, such direct methods do not explicitly exploit the presence of the DTW which is a paramount analysis clue for the localization and stratification of the tumor regions. On the other hand, generalized object detection methods simply learn the image context features in a data-driven manner, treating each pixel equally. Due to the scarcity of data, the simple application of such methods cannot obtain effective results. At the same time, the lack of interpretability of the solution process also limits the applicability of the algorithm to the clinic. In practical diagnosis, it is critical for CNN-based methods to point out the grounds that the inferences depend on, instead of merely providing the final decisions.

Motivated by the above analysis, as shown in Fig. 2, we devise a multi-task learning framework with hierarchical guidance for tumor stratification, incorporating DTW segmentation, tumor localization and stratification tasks. Following clinical experiences, we believe that the segmentation of DTW is able to guide the tumor localization, and tumor location together with the DTW further play significant roles in tumor stratification. Thus, we propose an implicit attention mechanism based on hierarchical guidance, and an explicit online Class Activation Mapping (CAM) constraint, for exploring the clues of upstream tasks, e.g. DTW segmentation and tumor localization, to guide the implementation of downstream tasks, e.g. tumor stratification. Besides, a graph reasoning based Global Feature Perception (GFP) module is introduced to capture global context information for better tumor stratification.

The key contributions of this work are as follows:

- To our best knowledge, we are the first to tackle the joint localization and stratification task of submucosal tumors in endoscopy ultrasound images, which is an essential preliminary step for further tumor diagnosis.

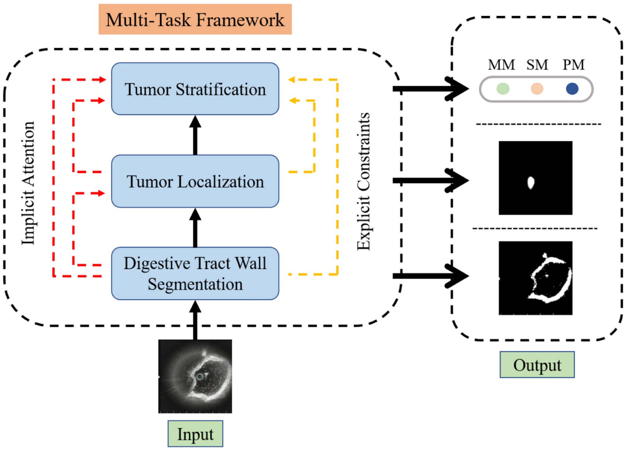

Fig. 2. Brief framework of our proposed multi-task learning for digestive tract wall segmentation, tumor localization and stratification. MM, SM and PM denote mucosal muscle, submucosa and proper muscle, respectively.

- We design a novel multi-task pipeline, incorporating DTW segmentation, tumor localization and stratification tasks. Upstream tasks are leveraged to guide the inference process of downstream tasks via implicit hierarchical guidance and explicit constraint on dynamic CAM. This helps to increase the interpretability of our approach, and also makes it coincide with the clinical experience. Moreover, a graph reasoning module is devised to capture global information for the tumor stratification task.
- Owing to the multi-task learning design and comprehensive feature exploration, our method achieves superior performance in locating and stratifying submucosal tumors on two EUS image datasets, compared to state-of-the-art object detection approaches.

II. Related Work

A. EUS Image Analysis

EUS plays an essential role in the diagnosis and treatment of submucosal tumors. Recently, some CNN-based computer-aided diagnosis (CNN-CAD) technologies have been applied to the EUS image analysis and made great progress [8], [11], [12], [13]. Specifically, these works mainly concentrate on the identification of gastrointestinal stromal tumors (GISTs), which are the most common submucosal tumors of the gastrointestinal tract. For instance, [8] develops a CNN-CAD system to differentiate GISTs from benign mesenchymal tumors such as leiomyomas and schwannomas. [11] aims to recognize GISTs of the higher-risk group from those of the lower-risk group on EUS images based on hand-craft feature extraction and a random forest classification model.

Different from the above mentioned methods which focus on the recognition of specific tumors, we define the general stratification task for submucosal tumors, and devote to provide more significant and comprehensive guidance for clinical diagnosis.

B. Lesion Detection

Since lesion detection relies heavily on object detection technologies in the computer vision community, we first make a brief review of object detection. As a traditional computer vision task, object detection aims to locate the object and predict its category simultaneously. Early deep learning based methods [9], [10], [14], [15], [16] widely adopt the anchor mechanism. Among them, two-stage approaches [9], [14], [15] introduce a region proposal network (RPN) to first generate the proposals, and then classify each proposal based on the aligned features. In contrast, one-stage methods [10], [16] jointly predict the object category and anchor box offsets, improving the inference speed. Recently, anchor-free methods abandon the anchor mechanism and predict key points to locate objects, such as corners [17], center points [18], [19] and hybrid extreme and center points [20]. Exploring the positive or negative sample selection strategy [21], [22] is an effective manner to boost the performance of object detection models.

With the prevalence of object detection approaches, plenty of works [23], [24], [25], [26] apply them to the field of ultrasound image analysis and lesion detection. However, as mentioned above, locating and stratifying the submucosal tumors rely heavily on global DTW guidance and contextual contrast reasoning. The general object detection methods may be incapable of tackling our proposed tasks since these approaches hardly learn significant clues for tumor localization and stratification due to the scarcity of data and the limitation of local region of interest (ROI) features.

C. Multi-Task Learning

Multi-task learning aims to simultaneously solve multiple related tasks, utilizing the mutual information among them to promote performance for each task [27]. It has been widely applied in medical image analysis. For instance, [28] proposes a multi-task UNet for gastrointestinal stromal tumor segmentation in EUS Images. [29] presents a multi-task attention based network for semi-supervised medical image segmentation, which incorporates supervised segmentation and unsupervised reconstruction tasks. [30] establishes a multi-task learning framework for segmentation and classification of tumors in 3D automated breast ultrasound images, consisting of an encoder-decoder network for segmentation and a lightweight multi-scale network for classification. A corpus of works [31], [32], [33] perform thoracic disease identification and localization on chest X-ray images, under limited supervision, namely only image-level annotations and a small amount of box-level annotations are available. Recently, the fast-spreading COVID-19 draws worldwide concerns. [34] devises a deep learning model to jointly identify the COVID-19 patient and segment the corresponding lesion region from CT images.

Compared to the above mentioned multi-task learning based methods, our approach employs the DTW segmentation, tumor localization and stratification sequentially, and integrates hierarchical guidance among tasks, including implicit attention and explicit constraints, to make full use of the inherited knowledge and also improve the interpretability of the model.

D. Attention Mechanism

Attention mechanism has proven its effectiveness in many computer vision and natural language processing tasks. According to the way obtaining the attention values, as in [33], we can roughly divide attention into activation-based attention and gradient-based attention.

In activation-based attention, the Sigmoid or Softmax activation function is usually employed to estimate attention values for re-weighting spatial positions or channel dimensions. [35] squeezes the feature map to a single vector and then obtains the channel-wise attention values through a fully connected layer followed by a Sigmoid operation. Non-local module [36] can be regarded as a special self-attention [37], which calculates the correlation between each pixel and all other pixels, and thus generates an attention map to enhance features with long-range dependency. The CBAM module proposed in [38] combines spatial and channel attention, capable of bringing benefit to various tasks. DANet [39] utilizes both position attention and channel attention to capture rich contextual information for scene segmentation. The other type is gradient-based attention, including CAM [40] and Grad-CAM [41], [42]. This line of attention technique can identify regions with significant responses to the inference result, thus they are widely used to explore the interpretability of deep models and implement weakly supervised object localization. Recently, some works [33], [43], [44], [45] further extend the CAM inference process as online trainable modules, which cooperates with the main classification task for improving classification performance and increasing interpretability simultaneously.

In this work, we combine activation-based attention and gradient-based attention. Different from previous works, our activation-based attention maps are obtained from the results of DTW segmentation and tumor localization tasks, and are employed to locally enhance the feature representations for the downstream task namely tumor stratification, and guide the dynamic CAM of tumor stratification.

III. Methodology

A. Problem Definition

This article targets at tackling the localization and stratification of tumors in EUS images. The goal is to locate the region of

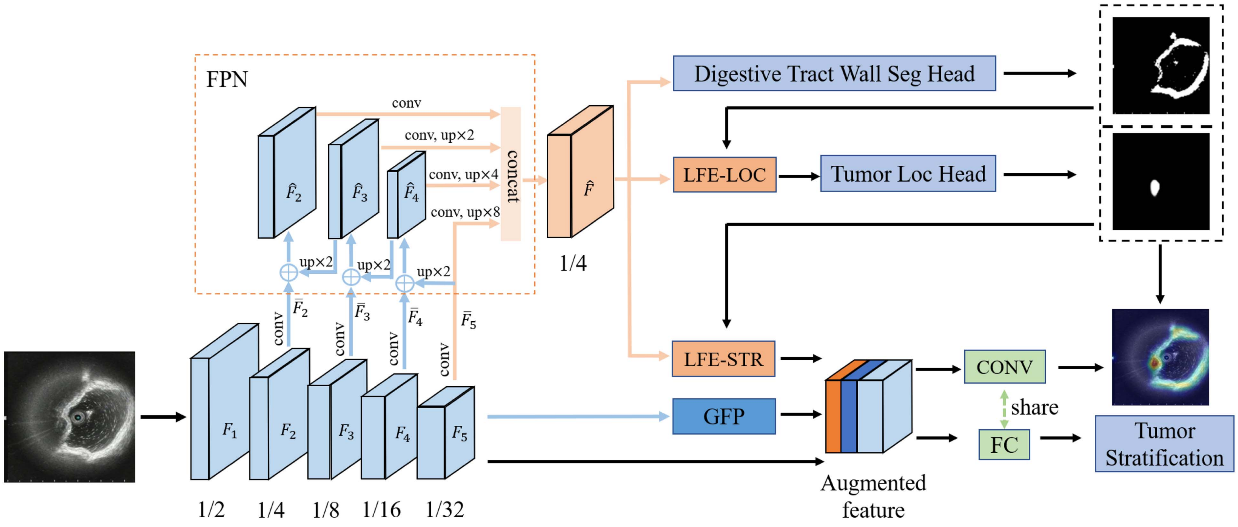

Fig. 3. Overview of our multi-task framework with hierarchical guidance. Feature pyramid network (FPN) is introduced to extract multi-scale features. The segmentation of digestive tract wall (DTW) is delivered to guide the tumor localization. And then tumor location together with the DTW segmentation further play significant roles in tumor stratification by implicit attention guidance and explicit class activation mapping (CAM) constraints. Moreover, the global feature perception (GFP) module is devised to capture global information to further facilitate the tumor stratification task.

the tumor inside the input image $\mathbf{X}$, and identify its stratification level namely predicting the probability vector $\mathbf{s} \in [0,1]^N$. Here, $N = 3$ denotes the number of stratification levels of tumors in EUS images.

B. Overview

The overview of our method is shown in Fig. 3. We propose a multi-task framework with hierarchical guidance for EUS image analysis, including DTW segmentation, tumor localization and stratification. Our network includes a shared encoder to obtain the feature representation of the input image. Due to the various shapes and sizes of DTW and tumors, the feature pyramid network (FPN) [46], built upon the backbone of the 34-layer residual networks (ResNet34) [47], is adopted to extract features for accurate DTW segmentation and tumor localization.

Given an image $\mathbf{X} \in \mathbb{R}^{H \times W \times 3}$, in which H, W, 3 are the height, width and channels of the image respectively, the i-th convolution block generates a feature map $\mathbf{F}_i \in \mathbb{R}^{H'_i \times W'_i \times C_i}$, where $H'_i = \frac{H}{2^i}, W'_i = \frac{W}{2^i}$, and C_i denotes the number of channels. $\{\mathbf{F}_i\}_{i=2}^5$ are fed into the FPN structure, resulting to the final feature map $\hat{\mathbf{F}}$ for segmenting DTW and locating the tumor. First, 1×1 convolutions are utilized to unify the dimensions of $\{\mathbf{F}_i\}_{i=2}^5$ to 64. We denote the features after dimensionality reduction as $\{\bar{\mathbf{F}}_i\}_{i=2}^5$, in which $\bar{\mathbf{F}}_i = Conv2D(\mathbf{F}_i)$. $Conv2D()$ denotes the 2D convolution operation. Then, we obtain the pyramid features $\hat{\mathbf{F}}_i$ by the following formulas:

$$\hat{\mathbf{F}}_i = \begin{cases} \overline{\mathbf{F}}_i + up_2(\hat{\mathbf{F}}_{i+1}), & i = 2,3 \\ \overline{\mathbf{F}}_i + up_2(\overline{\mathbf{F}}_{i+1}), & i = 4 \end{cases} \tag{1}$$

where up_2 denotes up-sampling the feature map by a factor of 2. Finally, the pyramid features $\hat{\mathbf{F}}_2$, $\hat{\mathbf{F}}_3$, $\hat{\mathbf{F}}_4$ and $\overline{\mathbf{F}}_5$ are post-processed by one convolution layer respectively. After up-sampled to the shape of $\frac{H}{4} \times \frac{W}{4}$, the resulted feature maps are concatenated into the final feature map $\hat{\mathbf{F}}$.

Referring to the clinical experience, the three tasks are implemented step by step. The DTW segmentation result is directly inferred from $\hat{\mathbf{F}}$. Then, after enhanced by the probability map of DTW, $\hat{\mathbf{F}}$ is exerted to predict the localization of the submucosal tumor. Finally, when predicting the stratification level, a comprehensive feature representation is acquired via merging a variant of $\hat{\mathbf{F}}$ enhanced by the hierarchical guidance module, another variant of $\mathbf{F}_5$ enhanced by the graph reasoning module [48], and the original $\mathbf{F}_5$.

In order to further enhance the dependency to DTW and tumor regions in the inference process of tumor stratification, we constrain the dynamic CAM with the DTW segmentation map and tumor localization map, increasing the model interpretability. In the subsequent sections, we will introduce each component of our proposed framework in detail.

C. Enhancing Features With Hierarchical Guidance

The three tasks, DTW segmentation, tumor localization, and tumor stratification, are implemented in a successive manner. We devise a hierarchical guidance module for the purpose of exploring the clues of upstream tasks to guide the inferring processes of downstream tasks.

a) DTW Segmentation: Prior knowledge of DTW is critical to the localization and stratification of tumors. Thus, we employ a lightweight segmentation head, which is composed of three convolution operations to infer the segmentation map $\mathbf{M}_{DTW}$ from $\hat{\mathbf{F}}$.

b) Tumor Localization: Locating tumors in ultrasound images is challenging because the visual contrast between different organizations is very low and tumors usually have diversified sizes and shapes. DTW can provide significant clues to the

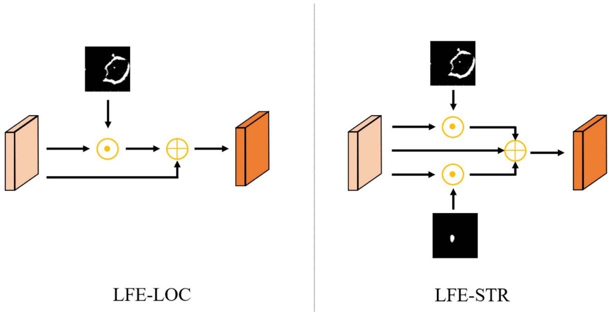

Fig. 4. Architectures of Local Feature Enhancement modules. Left: the module used to enhance features for tumor localization under the guidance of digestive tract wall (DTW) segmentation maps; right: the module employed to the feature for tumor stratification under the guidance of both DTW and tumor segmentation maps.

localization of tumors, e.g. tumors usually lie inside DTW and the contrast between abnormal and normal regions in DTW benefit the identification of tumors. Hence, we regard the inferred segmentation map of DTW as an attention map, and use it to enhance the feature representation for tumor localization. In practice, we design a local feature enhancement module (see LFE-LOC in Fig. 4) to transform $\hat{\mathbf{F}}$ into $\mathbf{F}^l$,

$$\mathbf{F}^l = \hat{\mathbf{F}} \odot \mathbf{M}_{DTW} + \hat{\mathbf{F}}. \tag{2}$$

'$\odot$' denotes the element-wise multiplication with broadcasting mechanism. $\mathbf{F}^l$ is fed into the tumor localization head which is also constituted by three convolution layers and derives a location map $\mathbf{M}_{Tumor} \in \mathbb{R}^{\frac{H}{4}\times\frac{W}{4}\times 1}$.

c) Multi-level Guidance for Tumor stratification: Based on the clinical experience, the task of tumor stratification requires to first locate the tumor region, and then observe the infiltration level in the DTW. Therefore, regions of the tumor and DTW are the key factors for tumor stratification. In order to incorporate the above two kinds of knowledge, we design two strategies: implicit attention guidance and explicit CAM constraints. In this section, we discuss the former in detail. Specifically, similar to the LFE-LOC module, we also design a local feature enhancement module for stratification (LFE-STR) to obtain the features $\mathbf{F}^s_l \in \mathbb{R}^{\frac{H}{4}\times\frac{W}{4}\times 256}$ through enhancing $\hat{\mathbf{F}}$ with $\mathbf{M}_{DTW}$ and $\mathbf{M}_{Tumor}$,

$$\mathbf{F}^s_l = \hat{\mathbf{F}} \odot \mathbf{M}_{DTW} + \hat{\mathbf{F}} \odot \mathbf{M}_{Tumor} + \hat{\mathbf{F}}. \tag{3}$$

$\mathbf{F}^s_l$ denotes the locally enhanced representation for tumor stratification. In addition, we also extract a global feature representation to strengthen the global understanding as introduced in the next subsection.

D. Stratifying Tumor With Global Feature Perception

To further capture the global structure dependencies for facilitating the tumor stratification, we design the global feature perception (GFP) module, which takes the $\mathbf{F}_5 \in \mathbb{R}^{\frac{H}{32}\times\frac{W}{32}\times 512}$ as input and generates a globally enhanced feature representation, $\mathbf{F}^s_g \in \mathbb{R}^{\frac{H}{32}\times\frac{W}{32}\times 256}$. As shown in Fig. 5, the core operation of the proposed GFP is based on graph reasoning, consisting of three steps.

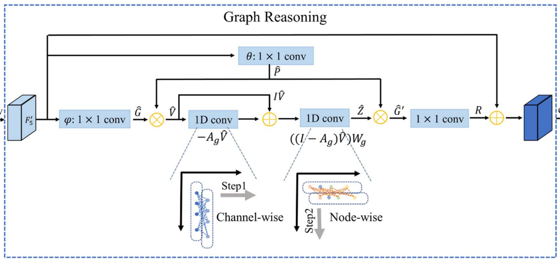

Fig. 5. Architecture of the Global Feature Perception module. The top one 1×1 convolution (θ) is used to produce the bi-projection between the coordinate and latent interaction spaces. The left (φ) and right 1×1 convolutions in the dashed box are utilized to reduce and expand dimensions. Two 1D convolutions are to perform graph reasoning.

1) Given the $\mathbf{F}'_5$, which has the same dimensions with the original $\mathbf{F}_5$ after the first convolution process, one 1×1 convolution is adopted to create a new embedding of $\mathbf{F}'_5$ to save the computation resources, resulting to $\mathbf{G} \in \mathbb{R}^{\frac{H}{32}\times\frac{W}{32}\times K}$, and the other 1×1 convolution is used to generate a tensor $\mathbf{P} \in \mathbb{R}^{\frac{H}{32}\times\frac{W}{32}\times Q}$ for node projection. Q denotes the number of feature nodes. The spatial dimensions of $\mathbf{G}$ and $\mathbf{P}$ are flattened, forming $\hat{\mathbf{G}} \in \mathbb{R}^{L\times K}$ and $\hat{\mathbf{P}} \in \mathbb{R}^{L\times Q}$ respectively ($L = \frac{H}{32} \times \frac{W}{32}$). The generated representation of nodes in the interaction space is $\hat{\mathbf{V}} = \hat{\mathbf{P}}^T\hat{\mathbf{G}}$.
2) Graph convolution [49], [50] is employed to explore the relationship among feature nodes. Specifically, we build a graph method based on a learnable adjacency matrix $\mathbf{A}_g$. Then, the graph convolution can be formulated as:

$$\hat{\mathbf{Z}} = ((\mathbf{I} - \mathbf{A}_g)\hat{\mathbf{V}})\mathbf{W}_g. \tag{4}$$

where $\mathbf{I}$ is the identity matrix, $\mathbf{A}_g$ indicates the adjacency matrix, and $\mathbf{W}_g$ denotes the state update function. As in [48], we implement this formula via two cascaded 1D convolutions along channel-wise and node-wise directions respectively:

$$\hat{\mathbf{Z}} = Conv1D((Conv1D(\hat{\mathbf{V}}) + \hat{\mathbf{V}})^{\mathrm{T}})^{\mathrm{T}}. \tag{5}$$

3) $\hat{\mathbf{Z}}$ are projected back into the original space, resulting to $\hat{\mathbf{G}}' = \hat{\mathbf{P}}\hat{\mathbf{Z}}$. Then $\hat{\mathbf{G}}'$ is reshaped to $\mathbf{G}' \in \mathbb{R}^{\frac{H}{32}\times\frac{W}{32}\times K}$ which is subsequently transformed into $\mathbf{R} \in \mathbb{R}^{\frac{H}{32}\times\frac{W}{32}\times 512}$ via a 1×1 2D convolution. The output of GFP $\mathbf{F}^s_g \in \mathbb{R}^{\frac{H}{32}\times\frac{W}{32}\times 256}$ is formed by applying another convolution to the addition of $\mathbf{F}'_5$ and $\mathbf{R}$.

The final feature used for stratification prediction is formed through fusing $\mathbf{F}^s_l$, $\mathbf{F}^s_g$ and original $\mathbf{F}_5$. First, we downsample the size of $\mathbf{F}^s_l$ into ($\frac{H}{32} \times \frac{W}{32}$) via max-pooling. Then, the downsampled variant of $\mathbf{F}^s_l$ is concatenated with $\mathbf{F}^s_g$ and $\mathbf{F}_5$ to construct the context-preserved fine-grained feature $\mathbf{F}^s \in \mathbb{R}^{\frac{H}{32}\times\frac{W}{32}\times 1024}$, which is aggregated into a 1024-dimensional vector via the global average pooling. Finally, one fully-connected layer attached with a softmax function is used to infer the stratification result $\mathbf{s}$.

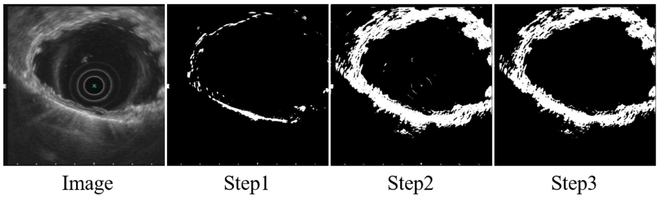

Fig. 6. Pseudo labels created by three steps.

E. Joint Training

We jointly train DTW segmentation, tumor localization and stratification to optimize our multi-task network. The specific loss item of each task is introduced in this section.

1) DTW Segmentation: Considering that annotating pixel-wise labels of the DTW region is very labor-intensive and cumbersome, we devise an interactive labeling algorithm to create pseudo labels for the DTW.

a) We normalize every ultrasound image into [0,1], and then predefine a threshold (0.6 in our experiment) to roughly segment the DTW.

b) A board-certificated ultrasound expert is required to select out samples whose segmentation results in the last step have acceptable quality. We train a U-Net segmentation model [51] with these selected images to re-annotate the remaining images. Practically, the trained U-Net is used to infer the DTW segmentation map of each remaining image, which is subsequently converted to a pseudo DTW label via cutting it with a threshold of 0.5. Only the largest connected component is preserved as the final annotation for every image.

An example of the annotation process is shown in Fig. 6. Every step in the interactive labeling algorithm benefits improving the region of DTW. The generated DTW masks are used to guide the learning of DTW segmentation in our proposed method. The loss function is formed by combining the binary cross entropy loss and the Dice loss:

$$\mathcal{L}^s = Bce(\mathbf{M}_{DTW}, \mathbf{G}_{DTW}) + Dice(\mathbf{M}_{DTW}, \mathbf{G}_{DTW}), \quad (6)$$

$$Bce(\mathbf{M}_{DTW}, \mathbf{G}_{DTW}) = -\frac{1}{HW}\sum_{i=1}^{H}\sum_{j=1}^{W}(g_{i,j}\log(m_{i,j}) + (1-g_{i,j})\log(1-m_{i,j})), \quad (7)$$

$$Dice(\mathbf{M}_{DTW}, \mathbf{G}_{DTW}) = 1 - \frac{2\sum_i^H\sum_j^W m_{i,j}g_{i,j} + \alpha}{\sum_i^H\sum_j^W(m_{i,j}+g_{i,j}) + \alpha}. \quad (8)$$

$\mathbf{M}_{DTW}$ is bilinearly upsampled to the size of the original image, and $\mathbf{G}_{DTW}$ is the pseudo label generated by the interactive labeling algorithm. $m_{i,j}$, $g_{i,j}$ denote the (i,j) pixel in $\mathbf{M}_{DTW}$ and $\mathbf{G}_{DTW}$, respectively. $\alpha(=1)$ refers to a smoothing factor for the Dice loss.

2) Tumor Localization: Only bounding box annotations are provided for the tumor localization task. For constraining the probability map of the tumor, we convert the bounding box into a mask, where pixels inside the bounding box are set to 1, and other pixels are set to 0. Again, the combination of the binary cross entropy loss and the Dice loss is used as the loss function for the tumor localization task:

$$\mathcal{L}^l = Bce(\mathbf{M}_{Tumor}, \mathbf{G}_{Tumor}) + Dice(\mathbf{M}_{Tumor}, \mathbf{G}_{Tumor}). \quad (9)$$

$\mathbf{M}_{Tumor}$ is upsampled to the size of the original image, and $\mathbf{G}_{Tumor}$ is the mask converted from the bounding box annotation.

3) Tumor Stratification: We adopt the cross entropy loss to constrain the stratification prediction result $\mathbf{s}$:

$$\mathcal{L}^c = -\sum_{n=1}^{N} y_n \log(s_n) \quad (10)$$

where s_n indicates the predicted probability of the tumor belonging to the n-th stratification level, and y_n is the corresponding ground truth label. N is the total number of levels.

Inspired from [43], we regularize the CAM of the tumor stratification with the probability maps of DTW and tumor, to further enhance the dependency to DTW and tumor regions in the inference process of tumor stratification. An extra branch consisting of one 1×1 convolution layer which shares the same parameters with the classification head for tumor stratification is introduced to dynamically estimate the CAM during the training stage. It generates one CAM for each stratification level. Since the identification of all levels depends on the same tumor and DTW regions, we average all CAMs to a $\frac{H}{32} \times \frac{W}{32} \times 1$ tensor $\mathbf{M}_{cam}$. Then, the following loss function is utilized to regularize $\mathbf{M}_{cam}$,

$$\mathcal{L}^a = Bce\left(\mathbf{M}_{cam}, \frac{\mathbf{M}_{DTW} + \mathbf{M}_{Tumor}}{2}\right). \quad (11)$$

The ground truth for CAM is obtained via averaging the predictions of DTW segmentation and tumor localization. $\mathbf{M}_{cam}$ is upsampled to the same size with the ground truth. The adoption of (11) enforces the stratification branch to focus on the DTW/tumor regions in the inference procedure, similar to the behaviour of clinicians. This helps to improve the interpretability of our method.

Overall Loss Function: The overall loss function for our multi-task framework is formed by summing up (6), (9), (10) and (11):

$$\mathcal{L} = \mathcal{L}^c + \lambda_1\mathcal{L}^s + \lambda_2\mathcal{L}^l + \lambda_3\mathcal{L}^a, \quad (12)$$

where λ_1, λ_2, and λ_3 are weights for balancing loss items $\mathcal{L}^s$, $\mathcal{L}^l$, and $\mathcal{L}^a$ respectively. We set $\lambda_1 = 10$, $\lambda_2 = 10$ and $\lambda_3 = 10$ in our experiments.

IV. Experiments

A. Datasets and Settings

1) Datasets: We evaluate our method on two EUS image datasets, which are collected from the sixth affiliated hospital of Sun Yat-sen University with approval from the local research ethics committee. The first is the Stomach-Esophagus EUS dataset which contains 737 images (618 patients) in total. The second is the Intestinal EUS dataset, containing 280 images (212 patients) in total. The tumors are categorized into 3 classes according to the invasion level, including mucosal muscle (MM),

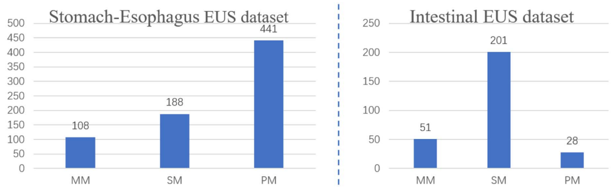

Fig. 7. Data distribution of the two endoscopic ultrasound (EUS) image datasets.

submucosa (SM) and proper muscle (PM). The number of each class in the two datasets is shown in Fig. 7. Each tumor is carefully annotated by experts for every image, in the format of the bounding box.

To fully evaluate the performance of our method, we adopt two experimental settings on both datasets, i.e., **vanilla-validation** and **cross-validation**. Specifically, for vanilla-validation, the Stomach-Esophagus EUS dataset is separated into a training set of 588 images (500 patients) and a testing set of 149 images (118 patients), and the Intestinal EUS dataset is split into 220 images (170 patients) for training and 60 images (42 patients) for testing. Note that in the two datasets, the patients have no overlap between the training set and test set. Besides, since the limited number of datasets, we also conduct 5-fold cross-validation on both datasets respectively to further evaluate our method. And the patients have no overlap among different folds of both datasets.

2) Experimental Details: Our model is implemented based on the PyTorch [52] framework. The K and Q in the graph reasoning process are set to 256 and 128 respectively. In the training phase, we use data augmentation methods such as random horizontal flipping and random rotation to enlarge the training set. All input images have a fixed size of 480×480. We set the batch size to 8, and use SGD optimizer with a momentum of 0.9 and a weight decay of 10^{-4}. The total number of epochs is set to 80. The learning rate is initially set to 0.001, and decayed at the 30th and 60th epochs by 0.05.

3) Evaluation Metrics: Following previous works [31], [32], [33], we adopt "the area under the ROC curve (AUC score)" and "Accuracy" to measure the tumor stratification and localization results. An inferred bounding box is identified as a correct prediction if the intersection over union (IoU) ratio between it and the ground-truth bounding box is larger than a specific threshold. And our method obtains the inferred bounding box based on the predicted tumor location map. For the DTW segmentation task, 30 testing samples are selected and annotated by experts to quantitatively evaluate the segmentation results with recall and precision metrics.

B. Ablation Study

We conduct extensive experiments on the Stomach-Esophagus EUS dataset under the vanilla-validation setting to verify the hierarchical guidance of three tasks and the effectiveness of each component.

1) Tumor Localization: The DTW segmentation provides fundamental prior knowledge for the tumor localization task. To verify our hypothesis, we compare the localization performance with and without the prior knowledge of DTW. In Table I, the experimental results show that under the guidance of the DTW segmentation, our localization accuracy has substantial improvement under different IoU thresholds, boosting the overall accuracy by 3.35%, 4.03%, 5.36% respectively.

TABLE I
ABLATION STUDY OF TUMOR LOCALIZATION PERFORMANCE (ACCURACY %). 'W/O LFE-LOC' MEANS THE LOCAL FEATURE ENHANCEMENT GUIDED BY DTW IS NOT USED

Threshold	Method	MM	SM	PM	ALL
0.1	Ours(w/o LFE-Loc)	68.18	76.32	75.28	74.50
	Ours	**77.27**	**78.95**	**77.53**	**77.85**
0.3	Ours(w/o LFE-Loc)	63.64	63.16	68.54	66.44
	Ours	**68.18**	**71.05**	**70.79**	**70.47**
0.5	Ours(w/o LFE-Loc)	54.55	52.63	58.43	56.38
	Ours	**63.64**	**63.16**	**60.67**	**61.74**

The best results are marked in bold.

TABLE II
ABLATION STUDY OF TUMOR STRATIFICATION PERFORMANCE (AUC SCORES %). ✓ INDICATES THE GFP MODULE OR CAM CONSTRAINT L^a IS ADOPTED. 'LFE' DENOTES THE GUIDANCE STRATEGY USED IN THE LOCAL FEATURE ENHANCEMENT

No.	GFP	LFE	L^a	MM	SM	PM	Mean
1				93.49	89.57	90.99	91.35
2	✓			94.27	94.19	92.83	93.76
3		DTW		94.24	93.76	91.65	93.22
4		Tumor		94.52	93.69	92.43	93.55
5		DTW+Tumor		95.17	93.43	93.28	93.96
6	✓	DTW+Tumor		95.45	95.19	92.90	94.51
7	✓	DTW+Tumor	✓	**96.06**	**95.33**	**93.30**	**94.90**

The best results are marked in bold.

2) Tumor Stratification: As shown in Table II, an elaborate ablation study is conducted on the tumor stratification task as well. No. 1 denotes the baseline model in which $\mathbf{F}_5$ is directly used to predict the stratification result. No. 2-6 incrementally include our designed components into the baseline model. In No. 2, the GFP module for capturing global relation information is incorporated into the baseline model, and the concatenation of $\mathbf{F}_5$ and global features $\mathbf{F}_g^s$ is utilized for tumor stratification. Comparing No. 2 to No. 1, the GFP module brings gains of 2.41% on the mean AUC metric. No. 3-5 introduce local feature enhancing strategies under different guidance maps into the baseline model. We can see that the awareness of either DTW or tumor region is beneficial for stratifying the tumor as they can provide valuable prior knowledge. The usage of both DTW and tumor generates better performance than using DTW or tumor only, which indicates that the knowledge of them is complementary to each other. Besides, the utilization of the tumor guidance is marginally better than that of the DTW guidance, since the stratification level is directly dependent on the location of the tumor. No. 6 indicates the variant of our method in which both GFP and LFE guided by both DTW and tumor are employed to improve the baseline model. No. 7 denotes the final version of our method. As we can observe from No. 6 and No. 7, the adoption of the CAM can benefit the stratification task quantitatively apart from increasing the interpretability.

To further explore the effect of each component and the stability of the whole framework, we record the stratification

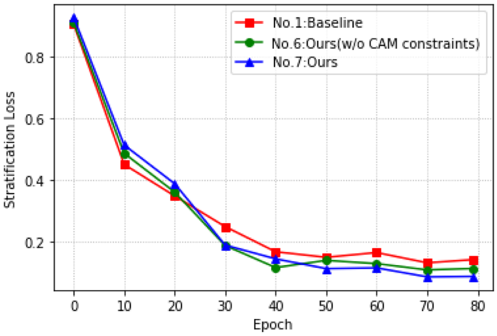

Fig. 8. Stratification loss in the training stage of different variants of our method.

TABLE III
ABLATION STUDY OF DTW SEGMENTATION PERFORMANCE

Threshold	Recall(%)	Precision(%)
0.1	65.94	40.27
0.2	60.24	42.12
0.3	56.31	43.19

loss in the training stage of three variants of our model and draw their loss curves, as shown in Fig. 8. Specifically, in the early stage of the training process, the baseline model exhibits a steeper descent pattern. We believe it's because the baseline model without well-designed feature guidance and constraints is easier to overfit the easy samples in the training set. However, as the training process going, thanks to the CAM constraints and other modules (i.e. GFP and LFE-STR), our method captures the essential stratification evidence and thus further decreases the stratification loss.

3) Digestive Tract Wall Segmentation: In Sections IV-B1 and IV-B2, we have validated the efficacy of the DTW segmentation to tumor localization and stratification. The quantitative performance of our method in segmenting DTW is presented in Table III. Although annotation data is not used, our method still achieves considerable precision and recall values under different thresholds.

C. Comparisons With Other State-of-the-Art Methods

To further demonstrate the superior performance of our method on the tumor localization and stratification tasks, we compare our approach with the state-of-the-art object detection methods, including FCOS [18], CornetNet [17], YOLOv3 [10], PAA [22], ATSS [21] and Faster-RCNN [9] on both datasets under two experimental settings. The object detection approaches may output multiple bounding boxes with different class predictions. The bounding box with the highest confidence score is regarded as the final result.

1) Tumor Localization: The detailed localization results of vanilla-validation are shown in Tables IV and V, and we also report the total accuracy under cross-validation setting in the format of (Mean ± Standard Deviation) in Tables VI and VII. With the guidance of DTW, our method achieves outstanding localization performance under different IoU thresholds on two

TABLE IV
TUMOR LOCALIZATION ACCURACY(%) ON STOMACH-ESOPHAGUS EUS DATASET (VANILLA-VALIDATION)

Threshold	Method	MM	SM	PM	ALL
0.1	FCOS [18]	72.73	76.32	73.03	73.83
	CornerNet [17]	72.73	**78.95**	75.28	75.84
	YOLOv3 [10]	68.18	73.32	73.03	73.15
	PAA [22]	72.73	73.68	75.28	74.50
	ATSS [21]	68.18	76.32	75.28	74.50
	Faster-RCNN [9]	72.73	73.68	74.16	73.83
	Ours	**77.27**	**78.95**	**77.53**	**77.85**
0.3	FCOS [18]	63.64	63.16	69.66	67.11
	CornerNet [17]	59.09	65.79	65.17	64.43
	YOLOv3 [10]	63.18	63.16	68.54	67.11
	PAA [22]	59.09	65.79	71.91	68.46
	ATSS [21]	63.64	63.16	68.54	66.44
	Faster-RCNN [9]	63.64	65.79	**70.79**	68.46
	Ours	**68.18**	**71.05**	**70.79**	**70.47**
0.5	FCOS [18]	40.91	52.63	60.67	55.70
	CornerNet [17]	45.45	52.63	55.06	53.02
	YOLOv3 [10]	59.09	47.37	64.04	59.06
	PAA [22]	54.55	55.26	62.92	59.73
	ATSS [21]	50.00	47.37	60.67	55.70
	Faster-RCNN [9]	54.55	57.89	**65.17**	**61.74**
	Ours	**63.64**	**63.16**	60.67	**61.74**

The best results are marked in bold.

TABLE V
TUMOR LOCALIZATION ACCURACY(%) ON INTESTINAL EUS DATASET (VANILLA-VALIDATION)

Threshold	Method	MM	SM	PM	ALL
0.1	FCOS [18]	81.82	80.95	85.71	81.67
	CornerNet [17]	54.55	76.19	42.86	68.33
	YOLOv3 [10]	54.55	83.33	71.43	76.67
	PAA [22]	72.73	78.57	85.71	78.33
	ATSS [21]	63.64	78.57	85.71	76.67
	Faster-RCNN [9]	**90.91**	85.71	57.14	83.33
	Ours	72.73	**92.86**	**100.00**	**90.00**
0.3	FCOS [18]	72.73	69.05	71.43	70.00
	CornerNet [17]	54.55	69.05	42.86	63.33
	YOLOv3 [10]	54.55	78.57	71.43	73.33
	PAA [22]	63.64	71.43	57.14	68.33
	ATSS [21]	45.45	64.29	71.43	61.67
	Faster-RCNN [9]	**81.82**	80.95	42.86	76.67
	Ours	72.73	**88.10**	**85.71**	**85.00**
0.5	FCOS [18]	**72.73**	59.52	42.86	60.00
	CornerNet [17]	36.36	57.14	28.57	50.00
	YOLOv3 [10]	54.55	61.90	42.86	58.33
	PAA [22]	45.45	54.76	14.29	48.33
	ATSS [21]	36.36	45.24	42.86	43.33
	Faster-RCNN [9]	63.64	61.90	42.86	60.00
	Ours	54.55	**66.67**	**71.43**	**65.00**

The best results are marked in bold.

TABLE VI
TUMOR LOCALIZATION ACCURACY(%) ON STOMACH-ESOPHAGUS EUS DATASET (CROSS-VALIDATION)

Threshold	0.1	0.3	0.5
Method			
FCOS [18]	74.77±0.92	67.71±2.27	56.85±1.96
CornerNet [17]	76.94±3.28	64.86±3.16	54.28±3.48
YOLOv3 [10]	74.09±1.75	67.71±2.68	58.89±3.46
PAA [22]	74.76±0.76	67.84±2.24	57.79±2.49
ATSS [21]	75.17±1.36	66.08±1.30	56.99±1.70
Faster-RCNN [9]	75.85±1.72	69.34±2.04	**59.01±2.96**
Ours	**78.70±2.58**	**69.88±1.77**	58.74±2.74

The best results are marked in bold.

TABLE VII
TUMOR LOCALIZATION ACCURACY(%) ON INTESTINAL EUS DATASET (CROSS-VALIDATION)

Threshold Method	0.1	0.3	0.5
FCOS [18]	83.97±4.53	78.73±5.50	61.45±5.80
CornerNet [17]	80.57±6.84	72.67±7.26	58.73±5.00
YOLOv3 [10]	81.88±5.92	77.21±7.44	68.03±7.98
PAA [22]	81.85±3.01	74.39±6.33	56.94±6.14
ATSS [21]	78.97±3.33	68.33±3.80	49.03±4.67
Faster-RCNN [9]	83.21±4.05	77.52±2.93	65.45±3.25
Ours	**86.36±4.07**	**81.36±4.58**	**70.82±4.98**

The best results are marked in bold.

TABLE VIII
TUMOR STRATIFICATION ACCURACY(%) ON STOMACH-ESOPHAGUS EUS DATASET (VANILLA-VALIDATION)

Method	MM	SM	PM	ALL
FCOS [18]	59.09	71.05	79.78	74.50
CornerNet [17]	68.18	78.95	89.89	83.89
YOLOv3 [10]	77.27	76.32	77.53	77.18
PAA [22]	50.00	71.05	82.02	74.50
ATSS [21]	54.55	68.42	80.90	73.83
Faster-RCNN [9]	59.09	65.79	83.15	75.17
Ours	**86.36**	**86.84**	**91.01**	**89.26**

The best results are marked in bold.

TABLE IX
TUMOR STRATIFICATION ACCURACY(%) ON INTESTINAL EUS DATASET (VANILLA-VALIDATION)

Method	MM	SM	PM	ALL
FCOS [18]	45.45	90.48	57.14	78.33
CornerNet [17]	54.55	92.86	42.86	80.00
YOLOv3 [10]	**72.73**	95.24	57.14	86.67
PAA [22]	36.36	97.62	28.57	78.33
ATSS [21]	36.36	88.10	28.57	71.67
Faster-RCNN [9]	36.36	95.24	42.86	78.33
Ours	54.55	**100.00**	**85.71**	**90.00**

The best results are marked in bold.

TABLE X
TUMOR STRATIFICATION AUC(%) ON STOMACH-ESOPHAGUS EUS DATASET (VANILLA-VALIDATION)

Method	MM	SM	PM	Mean
FCOS [18]	72.76	85.16	82.66	80.19
CornerNet [17]	76.25	86.90	89.60	84.25
YOLOv3 [10]	89.51	85.87	82.04	85.81
PAA [22]	73.98	63.16	54.87	64.00
ATSS [21]	63.10	77.03	88.93	76.35
Faster-RCNN [9]	82.50	88.93	83.46	84.96
Ours	**96.06**	**95.33**	**93.30**	**94.90**

The best results are marked in bold.

datasets. Compared with state-of-the-art object detection methods, our method shows competitive or superior performance. For example, under the IoU threshold value of 0.1 and 0.3, our method generates the best total accuracy of 78.70±2.58% and 69.88±1.77% on the Stomach-Esophagus EUS dataset. On the Intestinal EUS dataset, our method also achieves significantly better total accuracy than other methods under the cross-validation setting.

2) Tumor Stratification: Tables VIII, IX, X, and XI present the tumor stratification accuracy and AUC results on two datasets, respectively. And we also conduct 5-fold cross-validation and exhibit the total accuracy and mean AUC in Table XII. We can conclude that our method outperforms other object detection approaches by large margins. The experimental results demonstrate that the general object detection methods are not suitable for the tumor stratification task which relies heavily on the surrounding DTW knowledge and global relation information.

TABLE XI
TUMOR STRATIFICATION AUC(%) ON INTESTINAL EUS DATASET (VANILLA-VALIDATION)

Method	MM	SM	PM	Mean
FCOS [18]	72.54	72.09	60.65	68.43
CornerNet [17]	72.54	80.82	74.53	75.96
YOLOv3 [10]	79.59	83.33	90.57	84.50
PAA [22]	63.82	53.04	62.80	59.89
ATSS [21]	66.42	65.48	58.49	63.46
Faster-RCNN [9]	**88.68**	82.54	57.95	76.39
Ours	85.16	**88.62**	**99.73**	**91.17**

The best results are marked in bold.

TABLE XII
TUMOR STRATIFICATION PERFORMANCE(%) ON BOTH EUS DATASET (CROSS-VALIDATION)

Method	Stomach-Esophagus		Intestinal	
	Accuracy	AUC	Accuracy	AUC
FCOS [18]	76.12±1.83	82.54±1.87	78.57±1.78	75.58±4.64
CornerNet [17]	81.27±3.72	83.26±3.29	83.27±2.67	84.13±4.32
YOLOv3 [10]	77.88±0.97	84.84±1.12	83.88±2.53	87.69±3.10
PAA [22]	77.21±2.45	63.48±1.28	80.03±2.28	63.95±2.90
ATSS [21]	76.13±2.77	79.25±2.42	75.79±2.18	72.17±4.66
Faster-RCNN [9]	76.53±1.33	84.11±1.12	81.12±4.37	80.92±3.78
Ours	**86.56±2.25**	**94.25±1.42**	**86.00±3.88**	**90.71±3.52**

The best results are marked in bold.

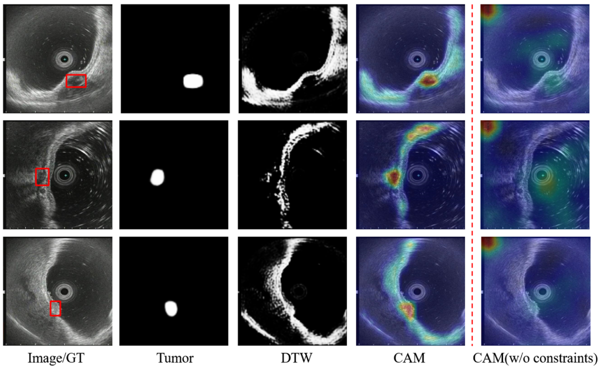

Fig. 9. Visualization results of our proposed multi-task network, including tumor localization, digestive tract wall (DTW) segmentation and class activation mapping (CAM) images, are shown in the second to fourth columns, respectively. The input image and corresponding ground truth (GT) annotation are shown in the first column. The last column is the CAM result obtained without explicit constraints, which is messy and lacks interpretability.

D. Visualization

A gallery of examples are provided in Figs. 9 and 10 to visualize the DTW segmentation and tumor localization results, together with CAMs of the tumor stratification task. As shown by the three examples in Fig. 9, our method can accurately identify the tumor and DTW regions. As shown by CAMs in Figs. 9 and 10, thanks to the implicit hierarchical guidance

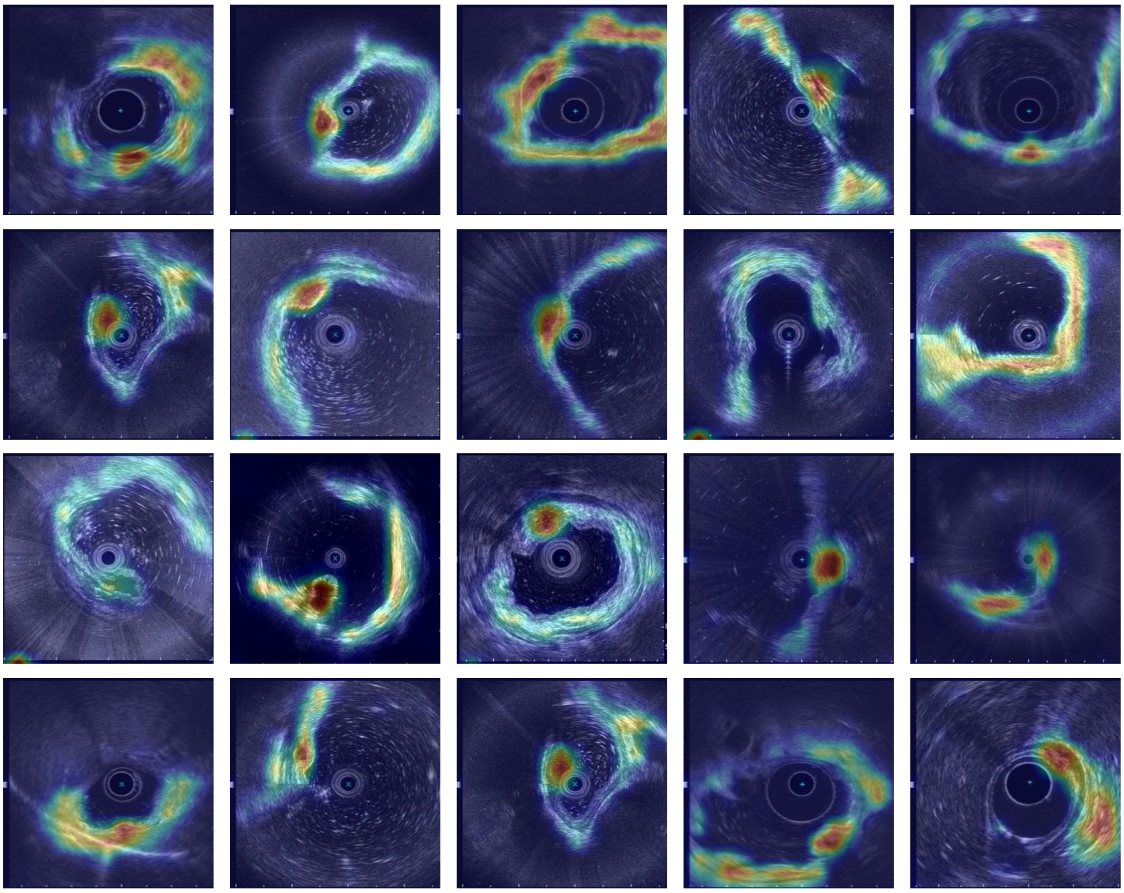

Fig. 10. Visualization of our class activation mapping (CAM) images.

module and the explicit CAM constraint, our model focuses on the tumor and the DTW regions when implementing the stratification prediction. This is consistent with the practical clinical experience of ultrasound experts. As shown in Fig. 9, the variant of our method without using constraint L^a in (11) generates CAMs lacking interpretability, while the CAMs of our method are apparently more explainable.

V. Conclusion

In this work, we aim at tackling the joint tumor localization and stratification problem in EUS images, which is of great significance for the preliminary diagnosis of submucosal tumors. Motivated by the experience of clinicians, we construct a multi-task framework which successively implements the DTW segmentation, tumor localization and tumor stratification tasks.

Quantitative experiments demonstrate that enhancing downstream tasks with the inference results of upstream tasks can improve the performance of downstream tasks, e.g. tumor localization and tumor stratification. The reason is that upstream tasks provide valuable clues for the implementation of downstream tasks. A global feature perception module based on graph reasoning is applied to enhance the high-level features with global relation information for tumor stratification. Performance improvement on the tumor stratification task is observed after the adoption of the global feature perception module. Constraining the dynamic class activation mapping with DTW and tumor probability maps is utilized to further enhance the dependency to the prior knowledge of DTW and tumor regions. The constraint helps our method derive class activation maps with high interpretability. Extensive experimental results on two EUS image datasets show that our method achieves tumor localization and stratification performance superior to state-of-the-art object detection methods.

References

[1] R. M. Kwee and T. C. Kwee, “The accuracy of endoscopic ultrasonography in differentiating mucosal from deeper gastric cancer,” *Official J. Amer. College Gastroenterol.*, vol. 103, no. 7, pp. 1801–1809, 2008.

[2] U. D. Siddiqui and M. J. Levy, “EUS-guided transluminal interventions,” *Gastroenterology*, vol. 154, no. 7, pp. 1911–1924, 2018.

[3] J. M. DeWitt et al., “Interventional endoscopic ultrasound: Current status and future directions,” *Clin. Gastroenterol. Hepatol.*, vol. 19, no. 1, pp. 24–40, 2021.

[4] Z. Wang, Q. Liu, and Q. Dou, “Contrastive cross-site learning with redesigned net for COVID-19 CT classification,” *IEEE J. Biomed. Health Inform.*, vol. 24, no. 10, pp. 2806–2813, Oct. 2020.

[5] Y. Xie et al., “Knowledge-based collaborative deep learning for benign-malignant lung nodule classification on chest ct,” *IEEE Trans. Med. Imag.*, vol. 38, no. 4, pp. 991–1004, Apr. 2019.

[6] G. Zhao,C. Fang, G. Li, L. Jiao, and Y. Yu, “Contralaterally enhanced networks for thoracic disease detection,” *IEEE Trans. Med. Imag.*, vol. 40, no. 9, pp. 2428–2438, Sep. 2021.

[7] J. Lian et al., “A structure-aware relation network for thoracic diseases detection and segmentation,” *IEEE Trans. Med. Imag.*, vol. 40, no. 8, pp. 2042–2052, Aug. 2021.

[8] Y. H. Kim et al., "Application of a convolutional neural network in the diagnosis of gastric mesenchymal tumors on endoscopic ultrasonography images," *J. Clin. Med.*, vol. 9, no. 10, 2020, Art. no. 3162.

[9] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2015, p. 28.

[10] J. Redmon and A. Farhadi, "YOLOv3: An incremental improvement," 2018, *arXiv:1804.02767*.

[11] X. Li, F. Jiang, Y. Guo, Z. Jin, and Y. Wang, "Computer-aided diagnosis of gastrointestinal stromal tumors: A radiomics method on endoscopic ultrasound image," *Int. J. Comput. Assist. Radiol. Surg.*, vol. 14, no. 10, pp. 1635–1645, 2019.

[12] G. Seven, G. Silahtaroglu, K. Kochan, A. T. Ince, D. S. Arici, and H. Senturk, "Use of artificial intelligence in the prediction of malignant potential of gastric gastrointestinal stromal tumors," *Dig. Dis. Sci.*, vol. 67, pp. 273–281, 2022.

[13] Y. Minoda et al., "Efficacy of endoscopic ultrasound with artificial intelligence for the diagnosis of gastrointestinal stromal tumors," *J. Gastroenterol.*, vol. 55, no. 12, pp. 1119–1126, 2020.

[14] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2014, pp. 580–587.

[15] R. Girshick, "Fast R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 1440–1448.

[16] W. Liu et al., "SSD: Single shot multibox detector," in *Proc. 14th Eur. Conf. Comput. Vis.*, 2016, pp. 21–37.

[17] H. Law and J. Deng, "CornerNet: Detecting objects as paired keypoints," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 734–750.

[18] Z. Tian, C. Shen, H. Chen, and T. He, "FCOS: Fully convolutional one-stage object detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 9627–9636.

[19] X. Zhou, D. Wang, and P. Krähenbühl, "Objects as points," 2019, *arXiv:1904.07850*.

[20] X. Zhou, J. Zhuo, and P. Krahenbuhl, "Bottom-up object detection by grouping extreme and center points," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 850–859.

[21] S. Zhang, C. Chi, Y. Yao, Z. Lei, and S. Z. Li, "Bridging the gap between anchor-based and anchor-free detection via adaptive training sample selection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 9759–9768.

[22] K. Kim and H. S. Lee, "Probabilistic anchor assignment with iou prediction for object detection," in *Proc. 16th Eur. Conf. Comput. Vis.*, 2020, pp. 355–371.

[23] Z. Cao et al., "Breast tumor detection in ultrasound images using deep learning," in *Proc. Int. Workshop Patch-Based Techn. Med. Imag.*, 2017, pp. 121–128.

[24] H. Li et al., "An improved deep learning approach for detection of thyroid papillary cancer in ultrasound images," *Sci. Rep.*, vol. 8, no. 1, pp. 1–12, 2018.

[25] M. H. Yap et al., "Breast ultrasound region of interest detection and lesion localisation," *Artif. Intell. Med.*, vol. 107, 2020, Art. no. 101880.

[26] S. Kulhare et al., "Ultrasound-based detection of lung abnormalities using single shot detection convolutional neural networks," in *Proc. Simul. Image Process. Ultrasound Syst. Assist. Diagnosis Navigation: Int. Workshops*, 2018, pp. 65–73.

[27] Y. Zhang and Q. Yang, "A survey on multi-task learning," *IEEE Trans. Knowl. Data Eng.*, vol. 34, no. 12, pp. 5586–5609, Dec. 2022.

[28] X. Li et al., "Multi-task refined boundary-supervision U-Net (MRBSU-Net) for gastrointestinal stromal tumor segmentation in endoscopic ultrasound (EUS) images," *IEEE Access*, vol. 8, pp. 5805–5816, 2020.

[29] S. Chen, G. Bortsova, A. G.-U. Juárez, G. van Tulder, and M. de Bruijne, "Multi-task attention-based semi-supervised learning for medical image segmentation," in *Proc. 22nd Int. Conf. Med. Image Comput. Comput.-Assist. Interv.*, 2019, pp. 457–465.

[30] Y. Zhou et al., "Multi-task learning for segmentation and classification of tumors in 3D automated breast ultrasound images," *Med. Image Anal.*, vol. 70, 2021, Art. no. 101918.

[31] Z. Li et al., "Thoracic disease identification and localization with limited supervision," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 8290–8299.

[32] J. Liu, G. Zhao, Y. Fei, M. Zhang, Y. Wang, and Y. Yu, "Align, attend and locate: Chest X-ray diagnosis via contrast induced attention network with limited supervision," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 10632–10641.

[33] X. Ouyang et al., "Learning hierarchical attention for weakly-supervised chest X-ray abnormality localization and diagnosis," *IEEE Trans. Med. Imag.*, vol. 40, no. 10, pp. 2698–2710, Oct. 2021.

[34] A. Amyar, R. Modzelewski, H. Li, and S. Ruan, "Multi-task deep learning based CT imaging analysis for Covid-19 pneumonia: Classification and segmentation," *Comput. Biol. Med.*, vol. 126, 2020, Art. no. 104037.

[35] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 7132–7141.

[36] X. Wang, R. Girshick, A. Gupta, and K. He, "Non-local neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 7794–7803.

[37] A. Vaswani et al., "Attention is all you need," in *Proc. 31st Int. Conf. Neural Inf. Process. Syst.*, 2017, pp. 6000–6010.

[38] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "CBAM: Convolutional block attention module," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 3–19.

[39] J. Fu et al., "Dual attention network for scene segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 3146–3154.

[40] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Learning deep features for discriminative localization," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 2921–2929.

[41] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 618–626.

[42] A. Chattopadhay, A. Sarkar, P. Howlader, and V. N. Balasubramanian, "Grad-CAM++: Generalized gradient-based visual explanations for deep convolutional networks," in *Proc. IEEE Winter Conf. Appl. Comput. Vis.*, 2018, pp. 839–847.

[43] K. Li, Z. Wu, K.-C. Peng, J. Ernst, and Y. Fu, "Tell me where to look: Guided attention inference network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 9215–9223.

[44] L. Wang et al., "Sharpen focus: Learning with attention separability and consistency," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 512–521.

[45] H. Fukui, T. Hirakawa, T. Yamashita, and H. Fujiyoshi, "Attention branch network: Learning of attention mechanism for visual explanation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 10705–10714.

[46] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 2117–2125.

[47] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.

[48] Y. Chen, M. Rohrbach, Z. Yan, Y. Shuicheng, J. Feng, and Y. Kalantidis, "Graph-based global reasoning networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 433–442.

[49] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," in *Proc. 5th Int. Conf. Learn. Representations*, 2017.

[50] Q. Li, Z. Han, and X.-M. Wu, "Deeper insights into graph convolutional networks for semi-supervised learning," in *Proc. AAAI Conf. Artif. Intell.*, 2018, pp. 3538–3545.

[51] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. 18th Int. Conf. Med. Image Comput. Comput. Assist. Interv.*, 2015, pp. 234–241.

[52] A. Paszke et al., "PyTorch: An imperative style, high-performance deep learning library," in *Proc. Adv. Neural Inf. Process. Syst.*, 2019, pp. 8026–8037.