IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 27 December 2022, accepted 23 January 2023, date of publication 31 January 2023, date of current version 3 February 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3240999

RESEARCH ARTICLE

FIAS3: Frame Importance-Assisted Sparse Subset Selection to Summarize Wireless Capsule Endoscopy Videos

WEIJIE XIE[1,2], ZEFEIYUN CHEN[1,2], QINGYUAN LI[3], QINGFEI MA[4], YUSI WANG[3,4], TIANBAO LIU[1,2], YUXIN FANG[3], ZHANPENG ZHAO[4], SIDE LIU[3,4,5], AND WEI YANG[1,2,5]

[1]School of Biomedical Engineering, Southern Medical University, Guangzhou 510515, China
[2]Guangdong Provincial Key Laboratory of Medical Image Processing, Guangzhou 510515, China
[3]Department of Gastroenterology, Nanfang Hospital, Southern Medical University, Guangzhou, Guangdong 510515, China
[4]Guangzhou SiDe MedTech Company Ltd., Guangzhou, Guangdong 510530, China
[5]Pazhou Laboratory, Guangzhou, Guangdong 510330, China

Corresponding authors: Wei Yang (weiyanggm@gmail.com) and Side Liu (liuside2011@163.com)

This work was supported in part by the National Natural Science Foundation of China under Grant 12026605, in part by the Guangdong Provincial Key Laboratory of Medical Image Processing under Grant 2020B1212060039, and in part by the Key-Area Research and Development Program of Guangdong Province under Grant 2022B0303020003.

This work involved human subjects in its research. Approval of all ethical and experimental procedures and protocols was granted by the Institutional Review Board, the need for informed patient consent for inclusion was waived.

ABSTRACT Wireless capsule endoscopy (WCE) is a recently developed tool that allows for the painless and non-invasive examination of the entire gastrointestinal (GI) tract. The microcamera captures a large number of redundant frames for each WCE examination such that a video summarization technique is needed to assist in diagnosis. However, prevalent methods of summarizing WCE videos focus only on the representativeness of the frames owing to a lack of high-level information on their importance. This paper develops a Frame Importance-Assisted Sparse Subset Selection model, called FIAS3, to integrate the high-level frame importance from networks into a sparse subset selection model. The FIAS3 is optimized under three constraints: 1) a frame importance matrix to help pay more attention to important frames, 2) a sparsity constraint to make video summaries more compact, and 3) a similarity-inhibiting constraint to reduce redundancy. The results of experiments on a public dataset demonstrated that our FIAS3 outperforms other methods of summarizing WCE videos. Specifically, its coverage and video reconstruction error were 92% and 0.143, respectively, at a 90% compression ratio, recording respective at least 16.9% and 0.031 improvements over other methods. The results of generalization experiments showed that FIAS3 also achieves competitive results on private datasets.

INDEX TERMS Computer-aided diagnosis, deep learning, keyframe extraction, video summarization, wireless capsule endoscopy (WCE).

I. INTRODUCTION

Wireless capsule endoscopy (WCE) is a recently developed tool for gastrointestinal (GI) examination that uses a microcamera and wireless transmission technology. It is expected that patients will soon be able to use it on their own in the comfort of their home. During WCE, patients need to only swallow a pill-sized endoscope equipped with a camera that then visualizes the GI tract, as shown in Fig. 1(a). Although WCE is painless, non-invasive, and available for the entire GI tract [1], [2], reviewing the entire video is time consuming and tedious for gastroenterologists. Each WCE video on average contains more than 50,000 frames, as shown in Fig. 1(b), most of which capture the normal mucosa that do not have clinical diagnostic value. Some frames cannot provide useful information owing to their poor quality

The associate editor coordinating the review of this manuscript and approving it for publication was Byung-Gyu Kim.

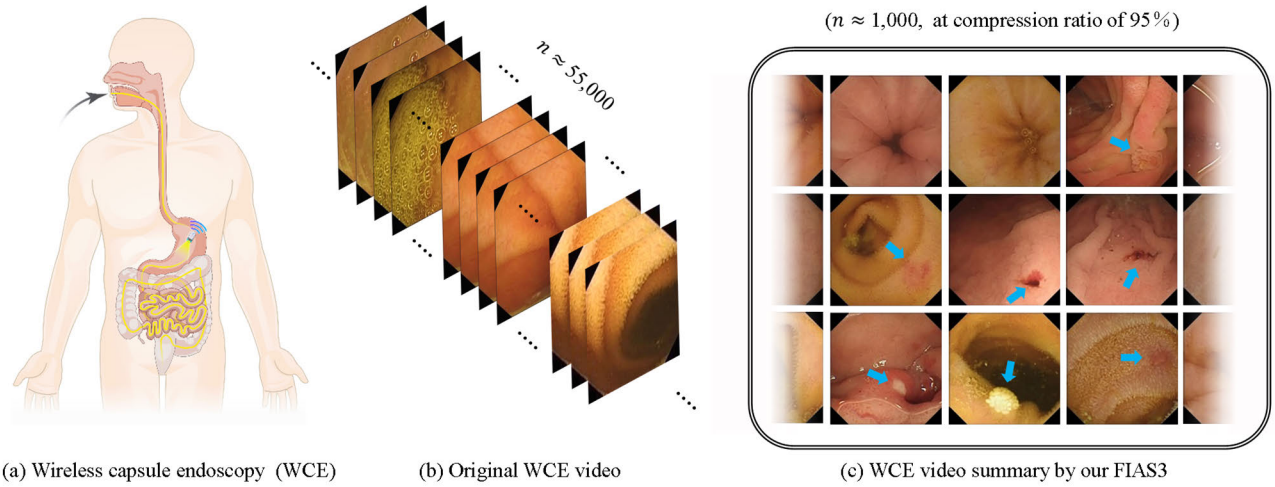


FIGURE 1. An illustration of WCE and WCE video summary. During (a) WCE, a wireless capsule is swallowed and moves along the entire GI tract; the captured images are transmitted to the receiving device to generate (b) an original WCE video. The proposed FIAS3 is able to summarize (b) into (c) a WCE summary while preserving the important findings, such as GI lesions (marked by blue arrows) and anatomical landmarks.

(e.g., due to reflections of light, darkness, bubbles, and undigested residues). The movement of the wireless capsule in the GI tract is uncontrollable, and it sometimes becomes stagnant in a specific location for a long time to yield a large number of redundant frames. One effective way to reduce the time taken by gastroenterologists is WCE video summarization.

Video summarization is widely used [3] in today's era of massive amounts of video data, with the aim of selecting a sequence of still keyframes or shots to represent the original video. Keyframe selection is usually applied in the task of WCE video summarization because keyframes offer more flexibility for browsing and navigating videos. Therefore, video summarization and keyframe extraction have the same meaning in the context of WCE videos, with the goal of choosing the most representative and important frames. Representative frames refer to a sequence of diverse frames whereas important frames contain such helpful information as images of GI lesions and anatomical landmarks, as shown in Fig. 1(c). However, frame importance has never been explicitly defined in the context of WCE videos, and is difficult to accurately quantify.

Most methods of WCE video summarization are shot-based approaches [4], [5], [6], [7], [8], [9], [10] that segment the original video into shots and then extract keyframes from each shot by relying on inter-frame relationships, such as those of similarity [4], [5], [6] and motion [7], [8]. These shot-based methods can eliminate local redundancy within individual shots, but similar scenes may recur in multiple shots such that they fail to eliminate global redundancy. In addition, the boundary of a shot in such methods is vague and cannot be clearly defined because WCE video is physically taken in a single shot. More critically, shot-based approaches may eliminate some frames containing important information owing to the lack of a mechanism for assessing frame importance. To avoid shot segmentation and introduce frame importance, a recent study [11] trained a sequence-to-sequence network to directly learn the underlying frame importance from human-labeled frame-wise importance scores. This network architecture can receive the long-range spatiotemporal information of the video and generate summaries of it that are consistent with human perception. However, these supervised methods require expensive frame-wise annotations for WCE video summarization, and this is impractical owing to the long duration of the video and the inconsistency of annotating among gastroenterologists, or even among repetitions of the exercise by one gastroenterologist. WCE video summarization thus remains a challenging task.

In this paper, we define frame importance for WCE videos and propose **F**rame **I**mportance-**A**ssisted **S**parse **S**ubset **S**election (FIAS3), to formulate WCE video summarization as a problem of subset selection, and estimate frame importance based on cost-effective labels for GI lesions and anatomical landmarks. Our FIAS3 reconstructs the WCE video by optimizing a coefficient matrix under three constraints: 1) a frame importance matrix for selecting important frames, 2) a sparsity constraint to make the information as compact as possible to generate short video summaries, and 3) a similarity-inhibiting constraint to eliminate global redundancy within the video summaries. By minimizing the weighted reconstruction loss under these constraints, FIAS3 can produce WCE video summaries with a high coverage of GI lesions and anatomical landmarks at different ratios of compression. Quantitative and qualitative experiments were conducted on the public dataset Kvasir-Capsule to verify the proposed method [12], and its capability for generalization was evaluated on a private dataset constructed by the authors.

TABLE 1. Comparative analysis of related works.

Year	Method	Feature	Shot-based	Frame importance	Demand for human annotation
2012	CCTS-MRFE [4]	color, texture	√		
2015	WCE-RIE [5]	color, texture	√		
2016	SNN-SVM [6]	SNN	√		
2020	Adaptive-SVD [10]	pretrained CNN, color	√		
2021	Adv-Ptr-Der-SUM [11]	end-to-end CNN		√	heavy
	FIAS3	SNN		√	light

The experimental results showed that our method outperforms prevalent methods of WCE video summarization on the public dataset and achieves competitive performance with them on the private dataset.

II. RELATED WORK

A. WCE VIDEO SUMMARIZATION

Many early studies in the area proposed shot-based approaches to WCE video summarization that involve first segmenting the entire video into shots and then separately selecting keyframes from each shot. Such methods are efficient for processing long videos. As shown in Table 1, WCE-RIE [5] and WCE-VS [6], proposed by Chen et al., segment shots based on similarities between adjacent shots, and then select keyframes by using an adaptive K-means clustering algorithm. The difference between them is that WCE-RIE uses color-related and textural features whereas WCE-VS extracts high-level semantic features by using a Siamese neural network (SNN) [13]. The relational rank matrix [4], motion analysis [7], [8], and factorization analysis [9], [10] have also been used to select keyframes in WCE videos from shots. These shot-based methods may separate numerous repeating and brief shots, leading to the selection of identical keyframes from various shots. Recently, Lan and Ye [11] collected an annotated dataset of frame-wise WCE videos from seven patients, similar to that in TV-Sum [14], and used it to propose a sequence-to-sequence network called Adv-Ptr-Der-SUM for WCE video summarization. Adv-Ptr-Der-SUM models spatiotemporal information by using long short-term memory (LSTM) [15] and predicts frame-wise scores that are similar to the ground truth. Although Lan et al. were thus able to avoid shot segmentation and select important frames that were aligned with human perception, annotating frame-wise importance is a subjective and expensive task for WCE videos. In this paper, we estimate frame importance without resorting to such expensive frame-wise annotations, and propose an optimization model that incorporates high-level frame importance while avoiding the problems of vagueness of shot segmentation and the annotation of frame-wise importance. A matrix of high-level frame importance is estimated by using GI lesion and anatomical landmark classification networks, thus making full use of the advantages of deep learning techniques as well as accurate and cost-effective image labels.

B. SPARSE SUBSET SELECTION

Sparse subset selection is a method of global modeling that has been verified on tasks of video summarization. Cong et al. [16] introduced an efficient global optimization algorithm to solve a row-sparse dictionary selection problem in consumer videos. Fei et al. [17] combined sparse subset selection with hierarchical clustering to improve the efficiency of keyframe extraction. In addition, sparse subset selection exhibits strong extendability. Ma et al. [18] extended the conventional linear sparse formulation into a block kernel sparse coding problem and introduced global inter-frame relationships by simply applying a transformation matrix. Wang et al. [19] merged changes in the gaze and content into prior cues to help a model of sparse dictionary selection choose important frames from a gastroscopic video. As in the work in Reference [19], we formulate WCE video summarization as a sparse subset selection problem, and estimate a frame importance matrix by using networks for classifying GI lesions and anatomical landmarks. The estimated frame importance is considered to be a weight that helps the sparse subset selection model pay more attention to important frames and thus improve the coverage of GI lesions and anatomical landmarks. To the best of our knowledge, this is the first study to define frame importance for WCE videos and use it to assist a sparse subset selection model.

C. WCE ABNORMALITY DETECTION

WCE abnormality detection is another way to assist gastroenterologists in quickly reviewing WCE videos. The emergence of open datasets is conducive to development in this direction. Many studies have sought to identify only one or several types [20], [21], [22], [23], [24] of abnormalities, such as bleeding [25], [26], tumors [26], polyps [27], ulcers [28], and hookworms [29]. In addition to the abnormalities, several GI anatomical landmarks and low-quality categories were also classified to help diagnose WCE. For example, Zhao et al. [20] produced a study synopsis by obtaining multiple labels for each frame of a WCE video, including those showing bile, air bubbles, extraneous matter, lesions, normal lumen, and polyps. They implemented frame-based classification by using support vector machines (SVMs) and sequence-based classification through hidden Markov models (HMMs) [30]. However, their synopsis provided multiple labels for every original video frame, including redundant and uninformative frames. By contrast, we aim to produce brief video summaries that provide the gist of the entire WCE video. To improve the coverage of GI lesions and anatomical landmarks in the summary, we train GI lesion and anatomical landmark classification networks to estimate a frame importance matrix to assist in WCE video summarization.

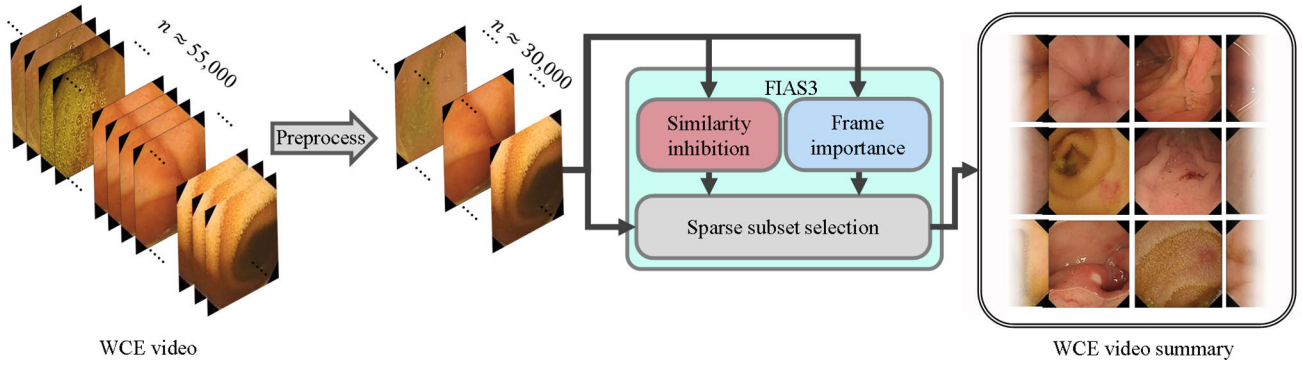


FIGURE 2. Process flow of WCE video summarization. "FIAS3" represents our frame importance-assisted sparse subset selection model.

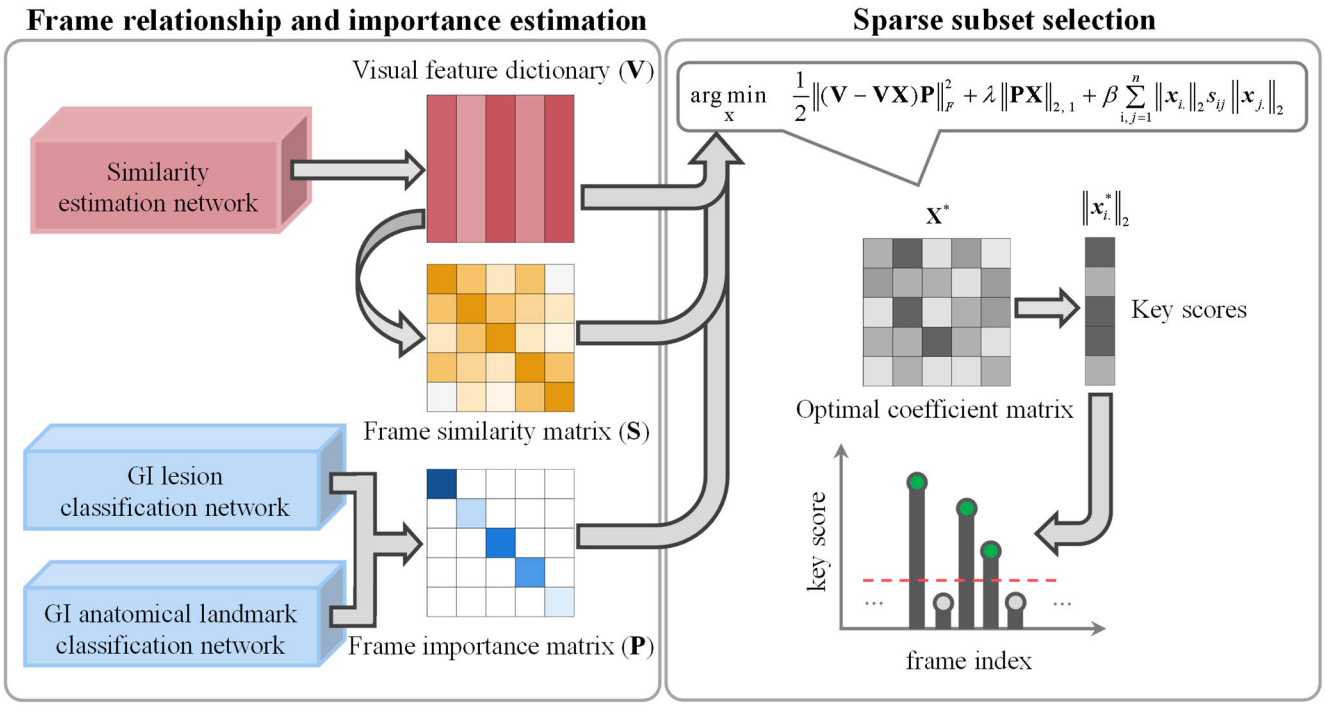


FIGURE 3. Overview of FIAS3.

III. METHODS

The process flow of our methods of WCE video summarization is illustrated in Fig. 2. Before the application of FIAS3, preprocessing is performed to reduce the amount of requisite computation. Uninformative and redundant frames are removed from the original video in this step. FIAS3 then extracts keyframes from the preprocessed WCE video by a sparse subset selection constrained by the similarity-inhibition and frame importance.

A. FIAS3

FIAS3 selects keyframes from a preprocessed WCE video. As illustrated in Fig. 3, a similarity estimation network first estimates the visual feature dictionary $\mathbf{V}$ and frame similarity matrix $\mathbf{S}$. The frame importance matrix $\mathbf{P}$ is obtained by using the GI lesion and anatomical landmark classification networks. The obtained matrices $\mathbf{V}$, $\mathbf{S}$, and $\mathbf{P}$ for each preprocessed WCE video are then introduced to the sparse subset selection model to solve for the optimal coefficient matrix $\mathbf{X}^*$. Finally, we select keyframes by thresholding the video with respect to the frame-wise key scores calculated from $\mathbf{X}^*$.

1) SIMILARITY ESTIMATION NETWORK

As the first step of image analysis, feature extraction is critical to the subsequent modeling. The most commonly used features are low-level hand-crafted features, such as color and

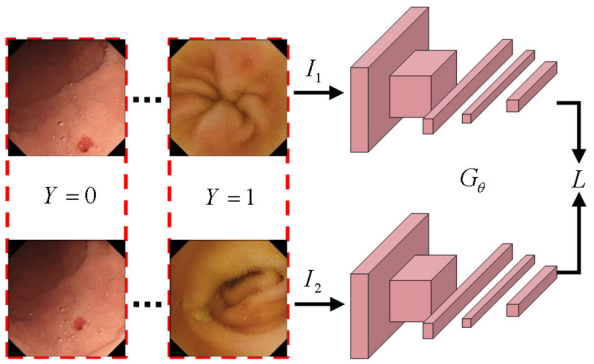


FIGURE 4. Similarity estimation network. Y, I_1, and I_2 are labels and frame pairs. G_θ represents CNNs that share structures and parameters (θ). L stands for contrastive loss.

texture. Although deep features extracted from a pretrained CNN are high dimensional, they still lack high-level semantic information. As illustrated in Fig. 4, we use an SNN [13] as our similarity estimation network to extract high-level semantic features. The network was trained on human-labeled similar and dissimilar frame pairs. The branches of the CNN G_θ consist of two convolutional layers and three linear layers each, and share the parameters θ. In the training phase, pairs of frames (I_1 and I_2) were fed into G_θ to obtain their features $G_\theta(I_1)$ and $G_\theta(I_2)$, respectively. The distance of the frame pair can be calculated by:

$$D_\theta(I_1, I_2) = \|G_\theta(I_1) - G_\theta(I_2)\|_2 \tag{1}$$

If I_1 is similar to I_2, the label Y is zero, and they are expected to be close, and vice versa. Hence, the parameters θ are optimized by using contrastive loss as:

$$L(\theta, (Y, I_1, I_2)) = \frac{1}{2}(1 - Y)(D_\theta(I_1, I_2))^2 + \frac{1}{2}Y\{\max(0, m - D_\theta(I_1, I_2))\}^2 \tag{2}$$

where m is a margin. In the inference phase, the similarity between frames I_1 and I_2 can be obtained by:

$$S_\theta(I_1, I_2) = \frac{1}{m}\max(0,\ m - D_\theta(I_1, I_2)) \tag{3}$$

Our similarity estimation network can extract high-level semantic features such that frames of the WCE video can be better distinguished than they can be based on low-level features, such as color, texture, and motion.

2) GI LESION AND ANATOMICAL LANDMARK CLASSIFICATION NETWORKS

To define frame importance, it is necessary to understand how gastroenterologists diagnose based on the results of WCE. During WCE diagnosis, gastroenterologists focus on the appearance and anatomical location of GI lesions. GI lesions can appear significantly different but the normal mucosae in the GI tract a highly similar such that localization is a challenge. Fortunately, several identifiable anatomical landmarks can help localize the GI tract. Therefore, we define the frame importance in WCE video as the possibility that a frame contains GI lesions or anatomical landmarks.

To extract high-level frame importance, we used massive amounts of public data to train GI lesion and anatomical landmark classification networks; it is benefiting from the rapid development of labeled data, optimization algorithms, GPU devices for efficient parallel computing, and deep CNN architecture. The classification networks, supervised by the human-labeled ground truths, can predict the frame-wise possibility of each class of GI lesions and anatomical landmarks. Their generated predictions are expressed as (c_{nor}, $\boldsymbol{c}_{\text{les}}$) and ($c_{\text{ord}}$, $\boldsymbol{c}_{\text{lan}}$), respectively. c_{nor} and c_{ord} are scalars representing the probabilities of predictions of the normal mucosa and the ordinary location, respectively. By contrast, $\boldsymbol{c}_{\text{les}}$ and $\boldsymbol{c}_{\text{lan}}$ are vectors representing the probabilities of predictions of multiple classes of GI lesions and anatomical landmarks, respectively.

3) FRAME RELATIONSHIP AND IMPORTANCE ESTIMATION

After training the similarity estimation network as well as the GI lesion and anatomical landmark classification networks, we construct a visual feature dictionary $\mathbf{V}$, a frame similarity matrix $\mathbf{S}$, and a frame importance matrix $\mathbf{P}$ for subsequent model optimization, as shown in the left half of Fig. 3. We use G_θ to extract the d-dimensional high-level visual features ($\boldsymbol{v}_i$) of the i-th frame and then concatenate them into the visual feature dictionary $\mathbf{V} = (\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_n)$. Further, we estimate the frame similarity matrix $\mathbf{S}$ by (1) and (3) based on $\mathbf{V}$, the element s_{ij} of which denotes the similarity between the i-th and the j-th frames.

By using predictions of the GI lesion and anatomical landmark classification networks, the i-th element of the diagonal frame importance matrix $\mathbf{P}$, representing the frame importance of the i-th frame, can be formulated as:

$$p_{ii} = \frac{c_{\text{nor}}}{\max(c_{\text{les}})} + \frac{c_{\text{ord}}}{\max(c_{\text{lan}})} \tag{4}$$

where c_{nor}, $\boldsymbol{c}_{\text{les}}$, c_{ord}, and $\boldsymbol{c}_{\text{lan}}$ have been introduced in Section III-B. We aim to obtain a small p_{ii} if the i-th frame contains GI lesions or anatomical landmarks. Therefore, c_{nor} and c_{ord} should be positively related to p_{ii} whereas $\boldsymbol{c}_{\text{les}}$ and $\boldsymbol{c}_{\text{lan}}$ should be inversely related to p_{ii}. We consider c_{nor} and c_{ord} to be molecules, and use the maximum values of $\boldsymbol{c}_{\text{les}}$ and $\boldsymbol{c}_{\text{lan}}$ as denominators for a more robust estimation. The element of the frame importance matrix is the sum of estimates of the two classification networks.

4) SPARSE SUBSET SELECTION

As in References [16], [19], we first formulate the task of WCE video summarization as a vanilla sparse subset selection problem without the similarity-inhibiting constraint and high-level frame importance information:

$$\arg\min_{\mathbf{X}} \frac{1}{2}\|\mathbf{V} - \mathbf{V}\mathbf{X}\|_F^2 + \lambda\|\mathbf{X}\|_{2,1} \tag{5}$$

where $\mathbf{V}$ and $\mathbf{X}$ are the visual feature dictionary and the coefficient matrix, respectively. The model attempts to reconstruct $\mathbf{V}$ through linear combinations of $\mathbf{V}$ and $\mathbf{X}$ under the row sparsity constraint $\|\mathbf{X}\|_{2,1}$. The j-th column of $\mathbf{X}$ contains the coefficients to reconstruct the j-th frame, and the i-th row of $\mathbf{X}$ is the coefficient of the i-th frame used to reconstruct $\mathbf{V}$. Therefore, the L2-norm of the i-th row vector $\|\boldsymbol{x}_{i.}\|_2$ can be used to estimate the key score of the i-th frame. With the row sparsity constraint, the model in (5) attempts to reconstruct $\mathbf{V}$ by using fewer frames. We can use the optimal coefficient matrix $\mathbf{X}^*$ to estimate the frame-wise key scores used for keyframe selection.

To impose a penalty for scenarios in which both similar frames obtain high key scores, we add a similarity-inhibiting constraint by using the frame similarity matrix $\mathbf{S}$ as:

$$\arg\min_{\mathbf{X}} \tfrac{1}{2}\|\mathbf{V}-\mathbf{V}\mathbf{X}\|_F^2 + \lambda\|\mathbf{X}\|_{2,1} + \beta\sum_{\mathrm{i},j=1}^{n}\|x_{i.}\|_2 s_{ij}\left\|x_{j.}\right\|_2 \tag{6}$$

where λ and β are the balance factors, and the element s_{ij} with the range of values of 0 to 1 represents the similarity between the i-th and the j-th frames. When two frames are similar (s_{ij} is large), their key scores $\|\boldsymbol{x}_{i.}\|_2$ and $\|\boldsymbol{x}_{j.}\|_2$ are not high at the same time. The model in (6) can use the similarity-inhibiting constraint to extract keyframes with lower redundancy than otherwise.

However, the model in (6) cannot adequately cover GI lesions or anatomical landmarks because it lacks high-level information on frame importance. We use elements of the frame importance matrix $\mathbf{P}$ as columnar weights for to reconstruct the visual feature ($\mathbf{V}$ - $\mathbf{V}\mathbf{X}$) and weights of the rows for the row sparsity constraint to obtain our final FIAS3:

$$\arg\min_{\mathbf{X}} \frac{1}{2}\|(\mathbf{V}-\mathbf{V}\mathbf{X})\mathbf{P}\|_F^2 + \lambda\|\mathbf{P}\mathbf{X}\|_{2,1} + \beta\sum_{\mathrm{i},j=1}^{n}\|x_{i.}\|_2 s_{ij}\left\|x_{j.}\right\|_2 \tag{7}$$

The smaller the element p_{ii} is, the more likely is a GI lesion or anatomical landmark to be present in the i-th frame. Therefore, we use $\mathbf{P}$ to reduce the reconstruction loss and the row sparsity constraint with regard to frames containing GI lesions or anatomical landmarks. This helps improve the coverage of GI lesions and anatomical landmarks.

5) MODEL OPTIMIZATION

To solve the non-convex function (7), we transform it into an L2, 1-norm problem [31] as follows:

$$\arg\min_{\mathbf{X}} \frac{1}{2}\|(\mathbf{V}-\mathbf{V}\mathbf{X})\mathbf{P}\|_F^2 + \lambda Tr(\mathbf{X}^{\mathbf{T}}\mathbf{P}\Psi\mathbf{P}\mathbf{X}) + \beta Tr(\mathbf{X}^{\mathbf{T}}\mathbf{W}\Phi\mathbf{W}\mathbf{X}) \tag{8}$$

where Tr stands for the trace of a square matrix, and the three latent variables $\boldsymbol{\Psi}$, $\mathbf{W}$, and $\boldsymbol{\Phi}$ are diagonal matrices whose elements are defined as follows:

$$\begin{aligned}\psi_{ii} &= 1/(2\|p_{i.}\mathbf{X}\|_2)\\ w_{ii} &= \sum_{j=1}^{n} s_{ij}\left\|x_{j.}\right\|_2\\ \varphi_{ii} &= 1/(2\|w_{i.}\mathbf{X}\|_2)\end{aligned} \tag{9}$$

The optimal coefficient matrix $\mathbf{X}^*$ can be obtained by:

$$\mathbf{X}^* = \left(\mathbf{V}^T\mathbf{V}\mathbf{P} + \lambda\mathbf{P}\boldsymbol{\Psi}\mathbf{P} + \beta\mathbf{W}\boldsymbol{\Phi}\mathbf{W}\right)^{-1}\mathbf{V}^T\mathbf{V}\mathbf{P} \tag{10}$$

Because $\boldsymbol{\Psi}$, $\mathbf{W}$, and $\boldsymbol{\Phi}$ depend on $\mathbf{X}$, a one-step optimization is not convex, and an iterative update strategy is thus adopted. The coefficient matrix $\mathbf{X}$ is initialized to a random matrix. For each iteration, $\boldsymbol{\Psi}$, $\mathbf{W}$, and $\boldsymbol{\Phi}$ are obtained by (9), and $\mathbf{X}$ is then updated by (10) by using the current values of $\boldsymbol{\Psi}$, $\mathbf{W}$, and $\boldsymbol{\Phi}$. When the change ($\|\Delta\mathbf{X}\|_F$) in the coefficient matrix is smaller than a predetermined threshold or the number of iterations reaches a predetermined number, we use $\mathbf{X}^*$ to estimate the frame-wise key scores used for keyframe selection.

B. PREPROCESSING

Our FIAS3 can be used on entire original WCE videos. Moreover, many prevalent techniques can be used to simplify our task and reduce the amount of computation required. These techniques include the removal of uninformative and redundant frames.

1) REMOVAL OF UNINFORMATIVE FRAMES

WCE videos contain a variety of low-quality frames, such as ones that are too dark, out of focus, contain blurred motion, bubbled frames, frames with highly reflective surfaces, and those containing residues. The complexity of and requisite computation for subset selection can be reduced by removing these low-quality frames. However, GI lesions or anatomical landmarks may be faintly visible in some low-quality frames, as shown in Fig. 5(a). Therefore, we remove only uninformative frames that do not contain any useful information.

The removal of uninformative frames is formulated as a traditional problem of supervised image binary classification. The color histogram is a typical visual feature used to remove uninformative frames from WCE videos [20], [32], [33], [34]. The luminance component usually does not contain useful information owing to complex shooting conditions. Therefore, we use HS histograms in which both the hue and the saturation components are quantified into 24 uniform bins. A simple but effective K-nearest neighbor (KNN) classifier [35] is used to classify uninformative frames by using 48-dimensional HS histograms.

2) REMOVAL OF REDUNDANT FRAMES

We can further reduce the complexity of subset selection and the amount of computation required by removing redundant frames. Some shot-based methods first segment a video into shots based on low-level similarities between adjacent frames and then select the ones closest to the cluster centers as representative frames. However, small GI lesions, shown in

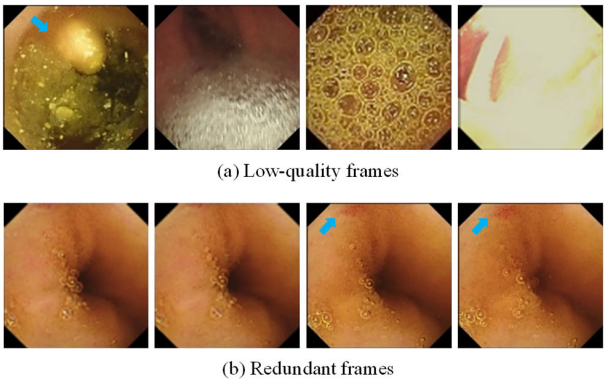


FIGURE 5. Examples of low-quality and redundant frames. (a) The first two low-quality frames contain a lymphoid hyperplasia (indicated by a blue arrow) or a pylorus, whereas the last two do not contain useful information (called uninformative frames). (b) The first two redundant frames contain no GI lesion but the last two contain a small angiectasia (indicated by the blue arrows).

Fig. 5(b), may be missed owing to the high similarity of their adjacent frames. Instead, we use a similarity estimation network to extract high-level semantic features for each frame. Furthermore, we use adaptive K-means clustering [5] to automatically select representative frames.

Rather than determining clusters without prior knowledge, adaptive K-means clustering generates them by limiting the maximum distance between samples within each cluster. The algorithm starts by selecting a frame as the first cluster. The remaining frames are then gradually grouped into the cluster nearest to them if their distance to the nearest cluster center is less than the predetermined maximum distance. Otherwise, a new cluster is created. The cluster centers need to be updated in every iteration. Finally, the frames closest to each cluster center are selected as representative ones to remove redundant frames.

C. EVALUATION METRICS

We quantitatively evaluated the video summaries generated by our method along three dimensions: the coverage [9], [36] of GI lesions and anatomical landmarks, the video reconstruction error (VRE) [36], [37], [38], and the compression ratio [6], [36]. We expected our FIAS3 to preserve as many diagnostically valuable frames as possible, and it is less subjective than general methods of video summarization. We evaluated its coverage of GI lesions and anatomical landmarks as follows:

$$\text{Coverage} = 1 - \frac{N_{\text{mis}}}{N_{\text{lab}}} \tag{11}$$

where N_{lab} denotes the number of labeled findings regarding GI lesions and anatomical landmarks and N_{mis} denotes the number of missed findings. A coverage of 100% indicates perfect performance. The VRE is defined as follows:

$$\text{VRE} = \frac{1}{N_{\text{ori}}} \sum_{k=0}^{N_{\text{key}}} \sum_{t=\text{T}_k}^{\text{T}_{k+1}} \min \left(\begin{matrix} \|f(t) - f(T_k)\|_2^2, \\ \|f(t) - f(T_{k+1})\|_2^2 \end{matrix} \right) \tag{12}$$

TABLE 2. Labeled data for the uninformative frame classifier and the similarity estimation network.

		Uninformative	Informative	Similar	Dissimilar
Public	Train	840	882	5818	6674
	Test	397	451	6352	6131
Private	Train	676	4336	11118	6455
	Test	311	2157	11276	6659

where N_{ori} denotes the total number of frames in the original video, N_{key} denotes the number of keyframes, $f(t)$ denotes the HS histogram of the t-th frame, and T_k denotes the frame index of the k-th keyframe. $k = 0$ and $k = N_{\text{key}} + 1$ correspond to the indices of the first and last frames, respectively, in the original video. A small VRE indicates good performance. The compression ratio is defined as follows:

$$\text{Compression ratio} = 1 - \frac{N_{\text{key}}}{N_{\text{ori}}} \tag{13}$$

A higher compression ratio indicates that more uninformative and redundant frames have been removed.

D. IMPLEMENTATION DETAILS

We validated our proposed method on the public dataset Kvasir-Capsule [12] and tested its capability for generalization on a private dataset. The public dataset was captured by using the Olympus EC-S10 endocapsule, in a resolution of 336 × 336, and the private dataset was generated by using the OMOM JS-ME-II capsule at Nanfang Hospital in Guangzhou, China, with pixel-resolutions of 256 × 240.

Table 2 lists the labeled data for our classifier of uninformative frames and the similarity estimation network on the public and private datasets. We collected the training and test sets from three videos. The videos for the public dataset were selected from the unlabeled videos. We evaluated the KNN-based uninformative frame classifier on the public and private datasets by setting K to seven. The color histogram of each frame was collected at the original image size. Its accuracy was higher than 98.4% and false positive rate was lower than 1.3% on the test sets of both the public and the private datasets. Similarly, we evaluated the similarity estimation network by setting the similarity threshold to 0.5 to classify pairs of frames. The input image was resized to 96 × 96. Its accuracy was higher than 96.6% and false positive rate was lower than 3.5% on the test sets of both the public and the private datasets.

Table 3 lists the labeled data for each category of the GI lesion and the anatomical landmark classification networks on the public dataset. Images of ordinary locations were collected from the classes of GI lesions. More details of the categories have been provided in Reference [12]. Both networks used the architecture of ResNet-152 [39]. The input images were resized to 224 × 224. We also used data augmentation techniques, including random rotation, flipping, and color jitter. Both networks were optimized by using weighted

TABLE 3. Labeled data for GI lesion and anatomical landmark classification networks.

	GI lesion classification network							GI anatomical landmark classification network		
	Normal mucosa	Angiectasia	Blood	Erosion	Erythematous	Lymphoid hyperplasia	Ulcer	Ordinary location	Pylorus	Ileo-cecal valve
Train	19946	760	22	305	127	224	727	23601	765	719
Test	14660	106	424	133	111	368	127	17614	755	698

TABLE 4. Labeled data for keyframe evaluation.

	Findings	Frames	Videos
Public	156	4217	19
Private	112	1471	5

TABLE 5. The performance of each process.

Process	Coverage	Compression ratio	Inference time (s)
Uninformative frame removal	100.0%	5.6%	49.02
Redundant frame removal	100.0%	35.1%	21.40
FIAS3	92.0%	90.0%	255.51

cross-entropy loss. Their sensitivity and specificity were 96% and 70% for normal mucosa, and 87% and 63% for ordinary locations on the test set.

Table 4 shows the labeled data used for the evaluation and comparison of keyframes. Each labeled finding, such as a GI lesion or an anatomical landmark, was visible in more than one frame. The 19 videos used from the public dataset were labeled videos in the test set.

To implement FIAS3, we set the maximum distance of adaptive K-means clustering to 0.03 in preprocessing. The preprocessed sequence of frames was uniformly sampled as successive non-overlapping segments due to limitations of the hardware. We used FIAS3 to extract keyframes from each segment and collected them as a final video summary. The Adam optimizer was used to optimize the similarity estimation network with an initial learning rate of 0.01, and this was reduced by a factor of 0.8 if the validation loss stopped decreasing after 10 epochs. The margin of contrastive loss was set to one, the total number of iterations of training was set to 200, and the batch size was set to 256. Both classification networks were also optimized by using Adam with a learning rate of 0.0001 and a batch size of 128 over 50 epochs. The threshold in the model optimization was set as 10^{-8}, and the maximum number of iterations was 200.

Our methods were implemented in Python 3.9.1 under with the deep learning framework of PyTorch 1.7.1. The results were mainly obtained on a PC equipped with an Intel Core i7-9700 with a 3.0 GHz CPU and 8 GB of RAM. The training and inference of CNNs, including the similarity estimation network as well as the GI lesion and the anatomical landmark classification networks, were achieved on a workstation equipped with a GPU (NVIDIA GeForce RTX 2080 Ti).

IV. EXPERIMENTS

We first validated the removal of uninformative and redundant frames through preprocessing. We then conducted ablation studies on the effects of the SNN feature extractor, the similarity-inhibiting constraint, and the weights assigned according to frame importance on the quality of the keyframes. We then compared our FIAS3 both quantitatively and qualitatively with other methods of WCE video summarization. Finally, we tested the generalization of the proposed method on the private dataset.

A. PREPROCESSING PERFORMANCE

Table 5 shows the performance of FIAS3 at each step in the process of removing uninformative frames, removing redundant frames, and selecting final keyframes. The coverages were 100% after the removal of uninformative and redundant frames, indicating that neither process had missed any GI lesion or anatomical landmark. The compression ratio showed that 5.6% of uninformative frames and 29.5% of redundant frames had been eliminated through preprocessing. In addition, the inference time of the preprocessing step was much shorter than that of keyframe extraction. These results suggest that the removal of uninformative and redundant frames can reduce the complexity of the system and the amount of required computation.

B. ABLATION STUDY

Fig. 6 shows the results of the ablation studies on the effects of our proposed SNN feature extractor, similarity-inhibiting constraint, and the importance weights assigned to frames on the quality of keyframes on the public dataset. We first fixed $\mathbf{S}$ as a zero matrix $\mathbf{0}_n$ and $\mathbf{P}$ as an identity matrix $\mathbf{I}_n$ in the model in (7), and then used high-level semantic visual features as our baseline (SNN). To validate the effectiveness of our SNN feature extractor, we varied the feature dictionary $\mathbf{V}$ by using different features to solve the models of sparse subset selection (color–texture, deep–color). The color–texture model combined a 255-dimensional color histogram with a four-dimensional texture feature that was extracted by using a grey-level co-occurrence matrix [5]. The deep–color model combined a 26-dimensional color histogram with 2048-dimensional deep features extracted by a pretrained ResNet-50 [10]. In another experiment to verify the similarity-inhibiting constraint and the weights

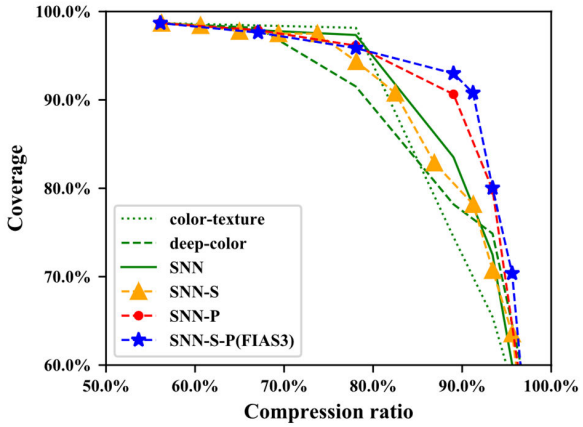


FIGURE 6. Coverage and compression ratio of the sparse subset selection model with different settings for the feature extractor, similarity-inhibiting constraint, and matrix of frame importance.

assigned to frames according to importance, we fixed the feature dictionary of the SNN and then replaced S with our estimated frame similarity matrix (SNN-S), replaced **P** with our estimated frame importance matrix (SNN-P), and replaced both (SNN-S-P). The performance curves show the coverage at different compression ratios to compare different methods. As expected, our SNN-based high-level semantic visual features yielded better performance than low-level features. The highest coverage was obtained by using both the similarity-inhibiting constraint and the matrix of frame importance. Specifically, the color–texture, deep–color, SNN, SNN-S, SNN-P, and SNN-S-P models yielded gains of 72.5%, 77.5%,81.1%, 79.5%, 88.2%, and 92.0% in coverage, respectively, when the compression ratio was 90.0%.

Table 6 listed the chosen factors and their effect on overall improvement at 90% compression ratio. By grid search, the highest performances were achieved when the λ and β were set to 0.01 and 0.0001 on the public dataset, respectively, and 0.005 and 5 on the private dataset. The length of each segment was 100 on both datasets. The best λ of the private dataset was smaller, and β was bigger than that of the public dataset. This may be because the accuracy of the estimated frame importance on the private dataset is lower than that on the public dataset, so it is necessary to reduce the weight of the frame importance constraint item in FIAS3.

C. COMPARISON WITH OTHER METHODS

We quantitatively and qualitatively compared our FIAS3 against four methods of WCE video summarization, denoted by CCTS-MRFE [4], WCE-RIE [5], SNN-SVM [6], and Adaptive-SVD [10], on the public dataset. As shown in Fig. 7, our FIAS3 generally obtained the best performance in terms of coverage and the VRE. The reproduced CCTS-MRFE method using seven-dimensional features could not attain a high compression ratio. The coverages of WCE-RIE, SNN-SVM, Adaptive-SVD, and our FIAS3 were 73.2%, 67.7%, 75.1%, and 92.0%, and their VREs were 0.174, 0.235, 0.177, and 0.143, respectively, when the compression ratio was 90%. Our FIAS3 thus obtained the best performance, with coverage that was 16.9% higher and VRE that was 0.031 lower than those of the other methods of WCE video summarization.

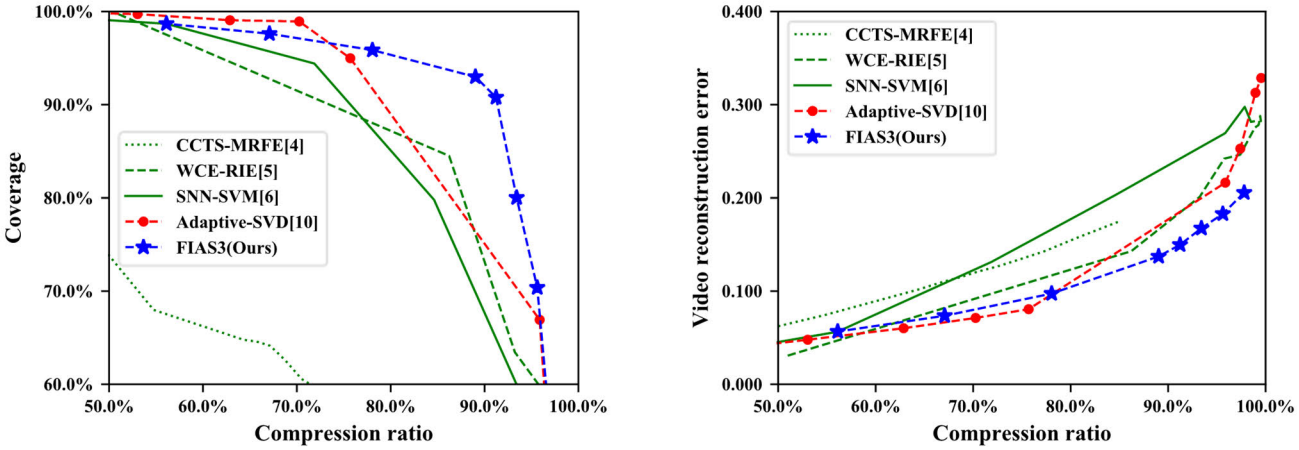
TABLE 6. Coverage with different factors at 90% compression ratio.

Dataset	λ	β	Segment length	Coverage
Public	0.005	0	100	82.9%
	0.01	0	100	88.2%
	0.05	0	100	75.1%
	0.1	0	100	73.4%
	0.01	0.00001	100	78.1%
	0.01	0.00005	100	69.9%
	0.01	**0.0001**	**100**	**92.0%**
	0.01	0.0005	100	67.0%
Private	0.0001	0	100	43.3%
	0.001	0	100	43.9%
	0.005	0	100	45.9%
	0.01	0	100	44.7%
	0.005	0.01	100	53.5%
	0.005	0.1	100	55.7%
	0.005	**5**	**100**	**59.6%**
	0.005	10	100	58.2%
	0.005	5	50	48.7%
	0.005	5	200	56.8%
	0.005	5	500	44.8%

We chose two WCE videos from the public test set for qualitative evaluation. Two example intervals with abnormal findings were defined in these videos. Five methods were first applied to these videos at a compression ratio of 95%. The extracted keyframes within the example intervals are shown in Fig. 8. Specifically, the low-quality interval in Fig. 8(a) consists of frames with indices ranging from 27474 to 27534, where lymphoid hyperplasia frequently masked by food debris was captured. The redundant interval in Fig. 8(b) consisted of frames with indices ranging from 21087 to 21129, where minor angiectasia was identified. The CCTS-MRFE, WCE-RIE, SNN-SVM, and Adaptive-SVD successfully covered the lymphoid hyperplasia but selected more redundant frames more than FIAS3. Of the methods considered, only Adaptive-SVD and our FIAS3 successfully covered minor angiectasia. Furthermore, the keyframes of our FIAS3 provided more diverse views on it. The quantitative and qualitative results suggest that our FIAS3 can generate keyframes with higher coverage and lower redundancy than the other methods.

D. GENERALIZATION CAPABILITY

We retrained the KNN-based uninformative frame classifier and the similarity estimation network on the private dataset. The curves of performance in Fig. 9 show that FIAS3 still



(a) Comparison of coverage

(b) Comparison of video reconstruction error

FIGURE 7. Quantitative comparison of FIAS3 with other methods.

achieved competitive results with the other methods. However, the improvement due to it was less remarkable than on the public dataset owing to its incorrect estimation of the matrix of frame importance. Specifically, the coverages of WCE-RIE, SNN-SVM, Adaptive-SVD, and our FIAS3 were 75.5%, 69.6%, 78.3%, and 80.5%, and their VREs were 0.104, 0.128, 0.093, and 0.106, respectively, when the compression ratio was 80%. The cross-dataset results demonstrate that our FIAS3 could adapt to the private dataset and obtain competitive results with the other methods.

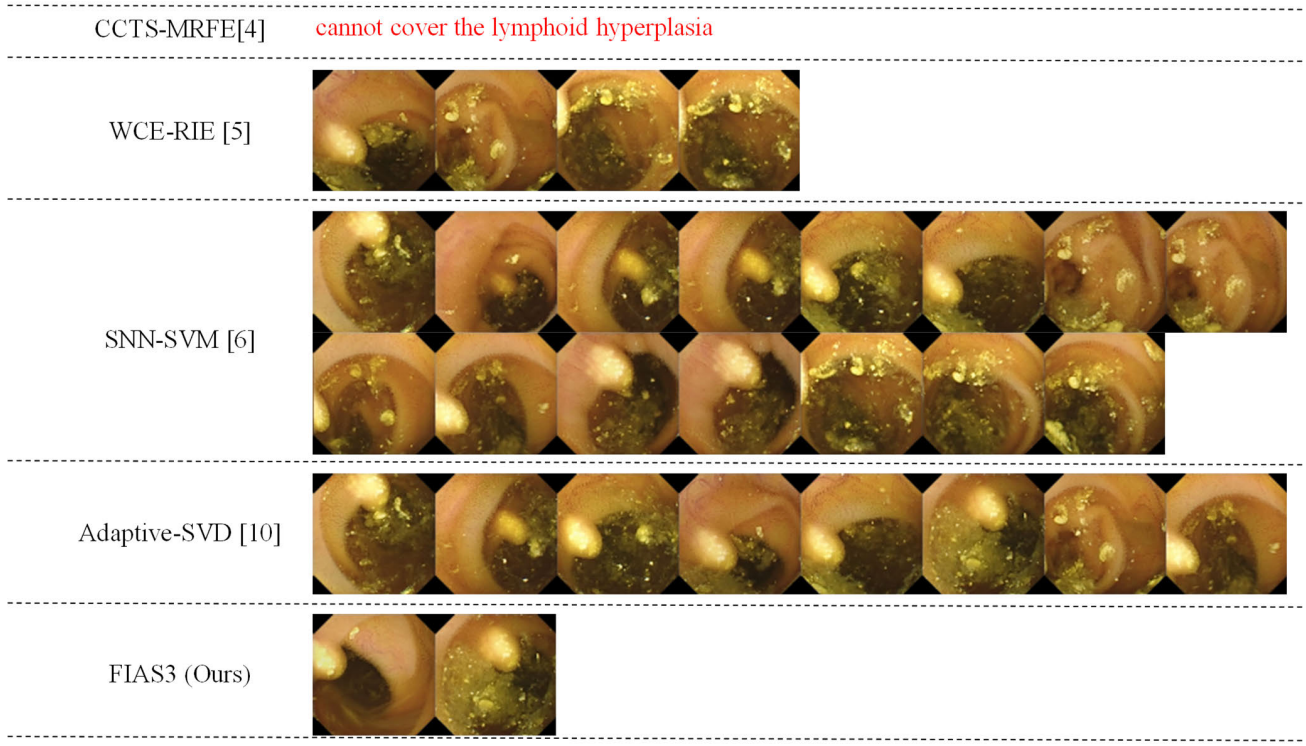
V. DISCUSSION

The results of both the quantitative and the qualitative experiments demonstrated the superiority of FIAS3, especially in terms of the coverage of GI lesions and anatomical landmarks. As illustrated by our ablation studies, the SNN feature extractor, similarity-inhibiting constraint, and matrix of frame importance contributed to its superiority over prevalent methods of WCE video summarization. The SNN feature extractor extracted high-level semantic features that could better distinguish between WCE frames than low-level features. The frame importance matrix estimated by the GI lesion and the anatomical landmark classification networks contained a large amount of high-level information that helped the model pay more attention to frames containing GI lesions and anatomical landmarks. Rather than introducing high-level information, the similarity-inhibiting constraint helped improve the coverage of the model by eliminating global redundancy. It is worth mentioning that the proposed summarization framework is not only suitable for WCE videos, but can be used on videos in different areas as long as frame importance is clearly defined. The similarity estimation network as well as the GI lesion and the anatomical landmark classification networks are plug-n-play, which means that they can be improved by leveraging advanced techniques in these domains.

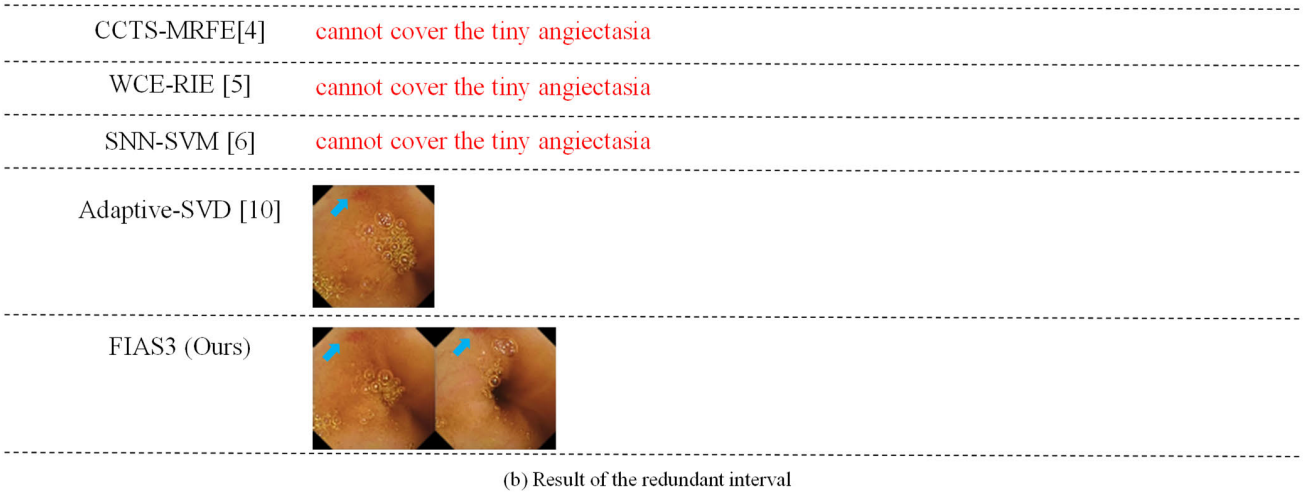
The uninformative frame classifier and similarity estimation network performed well on both public and private datasets with different resolution, providing $\mathbf{V}$ and $\mathbf{S}$ estimates of comparable quality. However, the results of the generalization experiment showed that the advantage of high coverage of FIAS3 became less prominent on the private dataset. Rather than the different resolution, this occurred is probably due to a gap in the distribution of features, such as different colors, lighting, and shooting angle, caused by the different wireless capsules on the public and the private datasets. The gap in distribution will lead to inaccurate $\mathbf{P}$ estimates and can be minimized by applying domain adaptation [40]. For instance, Dong et al. [41] trained a model of endoscopic lesion segmentation on gastroscope annotations by using domain adaptation. Hence, we believe that the summarization performance of the proposed model on wireless capsules from various manufacturers can be improved by integrating of our FIAS3 with methods of domain adaptation.

Compared with shot-based approaches, our FIAS3 can perceive information over a longer range to obtain better keyframes. It first transforms a long sequence of frames into a visual feature space and then selects keyframes by solving the problem of global optimization of the model of sparse subset selection. Although we increased the range of perception further through preprocessing, we still could not feed the entire WCE video into a single optimization model due to limitations of hardware. Such architectures as the LSTM [15] and the video transformer [42] can accommodate the entire video to avoid the bottleneck imposed by the hardware, but a method to introduce frame importance to these architectures has yet to be designed and verified.

Furthermore, our FIAS3 is more flexible than shot-based approaches in terms of the length of the video summaries. It can derive frame-wise key scores from the optimal coefficient matrix to generate video summaries at different rates



(a) Result of the low-quality interval

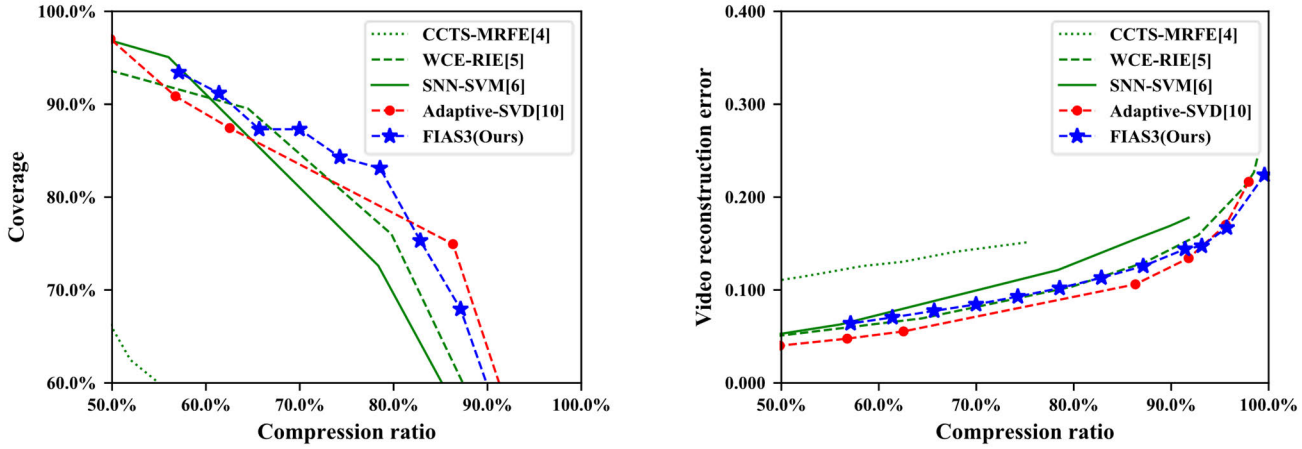


(b) Result of the redundant interval

FIGURE 8. Qualitative comparison of FIAS3 with other methods of WCE video summarization. (a) Keyframes within a 60-frame low-quality interval containing lymphoid hyperplasia. (b) Keyframes within a 42-frame redundant interval containing a small angiectasia (indicated by blue arrows). The red text highlights methods that could not cover the abnormalities.

of compression without resolving the model. Such flexibility may be useful for gastroenterologists as it allows them to adjust the compression ratio based on their experience.

In addition to the abovementioned possible improvements in performance, methods to assess WCE video summarization should be further advanced. Video summarization is still a subjective task in general, although it is more objective for WCE videos because there is consensus among gastroenterologists on which frame is important. Such quantitative metrics as coverage and compression ratio can be used to roughly evaluate summarization performance, but they might be sub-optimal. To resolve this problem, we can have one group of gastroenterologists offer a diagnosis based on the original WCE video and another group provide it based on the corresponding video summaries generated. We can then compare their performance on detection and reviewing times to reflect the clinical practicability of the video summarization.



(a) Comparison of coverage

(b) Comparison of video reconstruction error

FIGURE 9. Generalization-related performance of FIAS3 and the other methods on the private dataset.

VI. CONCLUSION

In this study, we developed a method of WCE video summarization called FIAS3 to generate video summaries by using both the relationships between frames and frame importance. Our frame importance matrix, estimated from the GI lesion and the anatomical landmark classification networks, significantly improves the coverage of GI lesions and anatomical landmarks. The similarity-inhibiting constraint used together with the frame importance matrix further improves coverage. The similarity estimation network can extract high-level semantic features from the video that provide better performance than low-level features. In addition, the proposed steps of preprocessing can reduce the amount of computation required. In general, the proposed FIAS3 outperformed prevalent methods on public and private datasets.

ACKNOWLEDGMENT

(Weijie Xie and Zefeiyun Chen contributed equally to this work.)

REFERENCES

[1] G. Iddan, G. Meron, A. Glukhovsky, and P. Swain, "Wireless capsule endoscopy," *Nature*, vol. 405, no. 6785, p. 417, May 2000.

[2] Y.-C. Wang, J. Pan, Y.-W. Liu, F.-Y. Sun, Y.-Y. Qian, X. Jiang, W.-B. Zou, J. Xia, B. Jiang, N. Ru, J.-H. Zhu, E.-Q. Linghu, Z.-S. Li, and Z. Liao, "Adverse events of video capsule endoscopy over the past two decades: A systematic review and proportion meta-analysis," *BMC Gastroenterol.*, vol. 20, no. 1, pp. 1–11, Dec. 2020.

[3] H. B. U. Haq, M. Asif, and M. B. Ahmad, "Video summarization techniques: A review," *Int. J. Sci. Technol. Res.*, vol. 9, pp. 146–153, Nov. 2020.

[4] J. S. Huo, Y. X. Zou, and L. Li, "An advanced WCE video summary using relation matrix rank," in *Proc. IEEE-EMBS Int. Conf. Biomed. Health Inform.*, Jan. 2012, pp. 675–678.

[5] J. Chen, Y. Wang, and Y. X. Zou, "An adaptive redundant image elimination for wireless capsule endoscopy review based on temporal correlation and color-texture feature similarity," in *Proc. IEEE Int. Conf. Digit. Signal Process. (DSP)*, Jul. 2015, pp. 735–739.

[6] J. Chen, Y. Zou, and Y. Wang, "Wireless capsule endoscopy video summarization: A learning approach based on Siamese neural network and support vector machine," in *Proc. 23rd Int. Conf. Pattern Recognit. (ICPR)*, Dec. 2016, pp. 1303–1308.

[7] B. Sushma and P. Aparna, "Summarization of wireless capsule endoscopy video using deep feature matching and motion analysis," *IEEE Access*, vol. 9, pp. 13691–13703, 2021.

[8] H.-G. Lee, M.-K. Choi, B.-S. Shin, and S.-C. Lee, "Reducing redundancy in wireless capsule endoscopy videos," *Comput. Biol. Med.*, vol. 43, no. 6, pp. 670–682, 2013.

[9] D. K. Iakovidis, S. Tsevas, and A. Polydorou, "Reduction of capsule endoscopy reading times by unsupervised image mining," *Computerized Med. Imag. Graph.*, vol. 34, no. 6, pp. 471–478, Sep. 2010.

[10] A. Biniaz, R. A. Zoroofi, and M. R. Sohrabi, "Automatic reduction of wireless capsule endoscopy reviewing time based on factorization analysis," *Biomed. Signal Process. Control*, vol. 59, May 2020, Art. no. 101897.

[11] L. Lan and C. Ye, "Recurrent generative adversarial networks for unsupervised WCE video summarization," *Knowl.-Based Syst.*, vol. 222, Jun. 2021, Art. no. 106971.

[12] P. H. Smedsrud, V. Thambawita, S. A. Hicks, and H. Gjestang, "Kvasir-Capsule, a video capsule endoscopy dataset," *Sci. Data*, vol. 8, no. 1, p. 142, 2021.

[13] G. Koch, R. Zemel, and R. Salakhutdinov, "Siamese neural networks for one-shot image recognition," in *Proc. ICML Deep Learn. Workshop*, 2015, pp. 1–8.

[14] Y. Song, J. Vallmitjana, A. Stent, and A. Jaimes, "TVSum: Summarizing web videos using titles," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 5179–5187.

[15] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997.

[16] Y. Cong, J. Yuan, and J. Luo, "Towards scalable summarization of consumer videos via sparse dictionary selection," *IEEE Trans. Multimedia*, vol. 14, no. 1, pp. 66–75, Feb. 2011.

[17] M. Fei, W. Jiang, W. Mao, and Z. Song, "New fusional framework combining sparse selection and clustering for key frame extraction," *IET Comput. Vis.*, vol. 10, no. 4, pp. 280–288, Jun. 2016.

[18] M. Ma, S. Mei, S. Wan, Z. Wang, D. D. Feng, and M. Bennamoun, "Similarity based block sparse subset selection for video summarization," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 31, no. 10, pp. 3967–3980, Oct. 2021.

[19] S. Wang, Y. Cong, J. Cao, Y. Yang, Y. Tang, H. Zhao, and H. Yu, "Scalable gastroscopic video summarization via similar-inhibition dictionary selection," *Artif. Intell. Med.*, vol. 66, pp. 1–13, Jan. 2016.

[20] Q. Zhao, G. E. Mullin, M. Q.-H. Meng, T. Dassopoulos, and R. Kumar, "A general framework for wireless capsule endoscopy study synopsis," *Comput. Med. Imag. Graph.*, vol. 41, pp. 108–116, Apr. 2015.

[21] M. Zhu, Z. Chen, and Y. Yuan, "DSI-Net: Deep synergistic interaction network for joint classification and segmentation with endoscope images," *IEEE Trans. Med. Imag.*, vol. 40, no. 12, pp. 3315–3325, Dec. 2021.

[22] X. Xing, Y. Yuan, and M. Q.-H. Meng, "Zoom in lesions for better diagnosis: Attention guided deformation network for WCE image classification," *IEEE Trans. Med. Imag.*, vol. 39, no. 12, pp. 4047–4059, Dec. 2020.
[23] Z. Ding, H. Shi, H. Zhang, L. Meng, M. Fan, C. Han, K. Zhang, F. Ming, X. Xie, H. Liu, J. Liu, R. Lin, and X. Hou, "Gastroenterologist-level identification of small-bowel diseases and normal variants by capsule endoscopy using a deep-learning model," *Gastroenterology*, vol. 157, no. 4, pp. 1044–1054, 2019.
[24] X. Guo and Y. Yuan, "Semi-supervised WCE image classification with adaptive aggregated attention," *Med. Image Anal.*, vol. 64, Aug. 2020, Art. no. 101733.
[25] F. Rustam, M. A. Siddique, H. U. R. Siddiqui, S. Ullah, A. Mehmood, I. Ashraf, and G. S. Choi, "Wireless capsule endoscopy bleeding images classification using CNN based model," *IEEE Access*, vol. 9, pp. 33675–33688, 2021.
[26] O. H. Maghsoudi, M. Alizadeh, and M. Mirmomen, "A computer aided method to detect bleeding, tumor, and disease regions in wireless capsule endoscopy," in *Proc. IEEE Signal Process. Med. Biol. Symp. (SPMB)*, Dec. 2016, pp. 1–6.
[27] X. Mo, K. Tao, Q. Wang, and G. Wang, "An efficient approach for polyps detection in endoscopic videos based on faster R-CNN," in *Proc. 24th Int. Conf. Pattern Recognit. (ICPR)*, Aug. 2018, pp. 3929–3934.
[28] J. A. O. Afonso, M. M. Saraiva, J. Ferreira, H. E. L. Cardoso, T. Ribeiro, P. I. C. Andrade, M. Parente, R. N. Jorge, and G. Macedo, "Automated detection of ulcers and erosions in capsule endoscopy images using a convolutional neural network," *Med. Biol. Eng. Comput.*, vol. 60, pp. 719–725, Jan. 2022.
[29] J.-Y. He, X. Wu, Y.-G. Jiang, Q. Peng, and R. Jain, "Hookworm detection in wireless capsule endoscopy images with deep learning," *IEEE Trans. Image Process.*, vol. 27, no. 5, pp. 2379–2392, May 2018.
[30] S. R. Eddy, "Hidden Markov models," *Current Opinion Struct. Biol.*, vol. 6, no. 3, pp. 361–365, 1996.
[31] F. Nie, H. Huang, X. Cai, and C. Ding, "Efficient and robust feature selection via joint $\ell_{2,1}$-norms minimization," in *Proc. Adv. Neural Inf. Process. Syst.*, no. 23, 2010, pp. 1–9.
[32] Z. Sun, B. Li, R. Zhou, H. Zheng, and M. Q.-H. Meng, "Removal of non-informative frames for wireless capsule endoscopy video segmentation," in *Proc. IEEE Int. Conf. Autom. Logistics*, Aug. 2012, pp. 294–299.
[33] M. K. Bashar, K. Mori, Y. Suenaga, T. Kitasaka, and Y. Mekada, "Detecting informative frames from wireless capsule endoscopic video using color and texture features," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent. (MICCAI)*, 2008, pp. 603–610.
[34] J. Oh, S. Hwang, J. Lee, W. Tavanapong, J. Wong, and P. C. de Groen, "Informative frame classification for endoscopy video," *Med. Image Anal.*, vol. 11, no. 2, pp. 110–127, Apr. 2007.
[35] L. E. Peterson, "K-nearest neighbor," *Scholarpedia*, vol. 4, no. 2, p. 1883, 2009.
[36] H. Bhaumik, S. Bhattacharyya, and S. Chakraborty, "Redundancy elimination in video summarization," in *Image Feature Detectors and Descriptors*. Cham, Switzerland: Springer, 2016, pp. 173–202.
[37] H.-C. Lee and S.-D. Kim, "Iterative key frame selection in the rate-constraint environment," *Signal Process., Image Commun.*, vol. 18, no. 1, pp. 1–15, Jan. 2003.
[38] T. Liu and J. R. Kender, "An efficient error-minimizing algorithm for variable-rate temporal video sampling," in *Proc. IEEE Int. Conf. Multimedia Expo (ICME)*, Aug. 2002, pp. 413–416.
[39] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.
[40] H. Guan and M. Liu, "Domain adaptation for medical image analysis: A survey," *IEEE Trans. Biomed. Eng.*, vol. 69, no. 3, pp. 1173–1185, Mar. 2022.
[41] J. Dong, Y. Cong, G. Sun, Y. Yang, X. Xu, and Z. Ding, "Weakly-supervised cross-domain adaptation for endoscopic lesions segmentation," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 31, no. 5, pp. 2020–2033, May 2021.
[42] S. Khan, M. Naseer, M. Hayat, S. W. Zamir, F. S. Khan, and M. Shah, "Transformers in vision: A survey," *ACM Comput. Surv.*, vol. 54, no. 10s, pp. 1–41, Jan. 2022.

WEIJIE XIE received the B.S. degree in biomedical engineering from the Southern Medical University, Guangzhou, China, in 2020. He is currently pursuing the Master of Engineering degree with the Department of Biomedical Engineering, Southern Medical University. His research interests include anatomical landmark classification, lesion detection, and video summarization in wireless capsule endoscopy.

ZEFEIYUN CHEN received the B.S. degree from the Department of Biomedical Engineering, Southern Medical University, Guangzhou, China, in 2019, where he is currently pursuing the Ph.D. degree. His research interests include medical image analysis and computerized-aid diagnosis.

QINGYUAN LI received the M.D. degree in clinical medicine from Southern Medical University, Guangzhou, China, in 2017. She is currently working as an Attending Doctor with the Department of Gastroenterology, Nanfang Hospital, Southern Medical University. Her current research interests include digestive system tumors and digestive endoscopy.

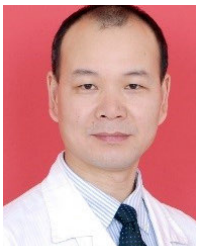
QINGFEI MA received the B.S. degree in engineering physics and the M.S. degree in electronic and communication engineering from Tsinghua University, Beijing, China, in 2003 and 2010, respectively. He is currently working as an Engineer with Guangzhou SiDe MedTech Company Ltd., Guangzhou, Guangdong, China. His current research interests include medical apparatus and instruments.

YUSI WANG received the bachelor's degree in clinical medicine from Chuanshan College, University of South China, Hengyang, China, in 2015, and the master's degree in internal medicine from Southern Medical University, Guangzhou, Guangdong, China, in 2021. She is currently working as a Medical Consultant with Guangzhou SiDe MedTech Company Ltd., Guangzhou. Her research interests include capsule endoscopy diagnosis, endoscopic diagnosis, and treatment of common diseases of the digestive tract and early cancers.

TIANBAO LIU is currently pursuing the master's degree with the Department of Biomedical Engineering, Southern Medical University, Guangzhou, China. His research interests include medical image analysis, machine learning, deep learning, and computerized-aid diagnosis.

YUXIN FANG received the M.D. degree in clinical medicine from Southern Medical University, Guangzhou, Guangdong, China, in 2018. He is currently working as an Attending Doctor with the Department of Gastroenterology, Nanfang Hospital, Southern Medical University. His current research interests include digestive system tumors and digestive endoscopy.

ZHANPENG ZHAO received the B.Sc. degree in communication engineering from Wuyi University, Jiangmen, China, in 2018. He is currently working as an Engineer with Guangzhou SiDe MedTech Company Ltd., Guangzhou, Guangdong, China. His research interests include medical image analysis, machine learning, deep learning, and computerized-aid diagnosis.



SIDE LIU received the bachelor's degree in military medicine from Third Military Medical University, Chongqing, China, in 1986, the master's degree in internal medicine from The First Affiliated Hospital, Third Military Medical University, in 1992, and the M.D. degree in internal medicine from First Military Medical University, China, in 1997. Since 2014, he has been with Nanfang Hospital, Southern Medical University, China, where he is currently a Professor and the Director of the Department of Gastroenterology. His research interests include early diagnosis and treatment strategies for gastrointestinal tumors, new capsule endoscopy systems, and new biological cell material for artificial liver.

WEI YANG received the B.Sc. degree in automation from the Wuhan University of Science and Technology, Wuhan, China, in 2001, the M.Sc. degree in control theory and control engineering from Xiamen University, Xiamen, China, in 2005, and the Ph.D. degree in biomedical engineering from Shanghai Jiao Tong University, Shanghai, China, in 2009. He is currently a Professor with the School of Biomedical Engineering, Southern Medical University, Guangzhou, China. His research interests include medical image analysis, machine learning, and computerized-aid diagnosis.