Received April 18, 2020, accepted May 18, 2020, date of publication May 22, 2020, date of current version June 5, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.2996631

Depth Information-Based Automatic Annotation of Early Esophageal Cancers in Gastroscopic Images Using Deep Learning Techniques

DINGYUN LIU[1,2], HONGXIU JIANG[1,2], NINI RAO[1,2], WENJU DU[1,2], CHENGSI LUO[1,2], ZHENGWEN LI[1,2], LINLIN ZHU[3], AND TAO GAN[3]

[1]Center for Informational Biology, University of Electronic Science and Technology of China, Chengdu 610054, China
[2]School of Life Science and Technology, University of Electronic Science and Technology of China, Chengdu 610054, China
[3]Digestive Endoscopic Center of West China Hospital, Sichuan University, Chengdu 610017, China

Corresponding author: Nini Rao (raonn@uestc.edu.cn)

This work was supported in part by the National Natural Science Foundation of China under Grant 61872405 and Grant 61720106004, in part by the Key Project of Natural Science Foundation of Guangdong province under Grant 2016A030311040, and in part by the Scientific Platform Improvement Project of UESTC.

ABSTRACT The early diagnoses of esophageal cancer are of great significance in the clinic because they are critical for reducing mortality. At present, the diagnoses are mainly performed by artificial detection and annotations based on gastroscopic images. However, these procedures are very challenging to clinicians due to the large variability in the appearance of early cancer lesions. To reduce the subjectivity and fatigue in manual annotations and to improve the efficiency of diagnoses, computer-aided annotation methods are highly required. In this work, we proposed a novel method that utilized deep learning (DL) techniques to realize the automatic annotation of early esophageal cancer (EEC) lesions in gastroscopic images. The depth map of gastroscopic images was initially extracted by a DL network. Then, this additional depth information was fused with the original RGB gastroscopic images, which were then sent to another DL network to obtain precise annotations of EEC regions. In total, 4231 gastroscopic images of 732 patients were used to build and validate the proposed method. A total of 3190 of those images were EEC images, and the remaining 1041 were non-EEC images. The experimental results show that the combination of depth information and RGB information improved the annotation performance. The final EEC detection rate and mean Dice Similarity Coefficient (DSC) of our method were 97.54% and 74.43%, respectively. Compared with other state-of-the-art DL-based methods, the proposed method showed better annotation performances and fewer false positive outputs. Therefore, our method offers a good prospect in aiding the clinical diagnoses of EEC.

INDEX TERMS Gastroscopic image, early esophageal cancer, lesion annotation, deep learning, depth map.

I. INTRODUCTION

Esophageal cancer (EC) is one of the most fatal cancer types; it has a quickly rising incidence throughout the world and accounts for more than 450,000 deaths each year [1]. Early-stage EC is not very lethal; its five-year survival rate is over 95%. However, if EC lesions are diagnosed in advanced stages, a poor prognosis may become inevitable; at this stage, the five-year survival rate drops to only 5% [2], [3]. Therefore, diagnoses of EC in the early stage are of great significance in the clinic. At present, conventional gastroscopy-based artificial examination is the most common method for the clinical diagnosis of EC, but this examination procedure is susceptible to many negative factors for clinicians, such as time limitations, fatigue and insufficient experience. Moreover, the appearances of some early esophageal cancer (EEC) lesions are similar to those of benign inflammatory lesions, which makes the diagnoses more difficult, even for experienced clinicians. As a result, misdiagnosis of EEC often occurs in the clinic.

Recently, researchers have developed many computer-aided diagnosis (CAD) methods to improve the accuracy and efficiency of clinical endoscopic diagnoses [4]–[7]. In addition, deep learning (DL) techniques, especially conventional neural network (CNN)-based techniques, have made remarkable progress in recent years and have achieved

The associate editor coordinating the review of this manuscript and approving it for publication was Chaker Larabi.

state-of-the-art performances of image classification and segmentation. Therefore, they have been applied in many image-based CAD methods that aimed to automatically detect and annotate cancer lesions [8]–[10]. For example, Hirasawa *et al.* [11] designed a CNN-based method to automatically detect and annotate gastric cancer regions in gastroscopic images. A total of 13,584 gastric cancer images were utilized to train a CNN named Single Shot MultiBox Detector (SSD), and 2296 test images were used to validate the performance of the fully trained network. The results showed that this method obtained a high sensitivity of 92.2%. Similarly, Horie *et al.* [12] utilized the same SSD network to realize the annotation of EC regions in gastroscopic images, and they finally obtained a high sensitivity of 98%. However, the outputs of SSD are always square regions, which means the annotation results of this network cannot provide accurate locations of the edges of cancer lesions. Groof *et al*. [13]–[15] proposed an EEC annotation method based on high-definition (HD) gastroscopic images in [15]. They initially applied a traditional CNN, AlexNet, to extract deep features from local image windows. Then, a support vector machine (SVM) classifier was applied to identify EEC windows according to the extracted features. Finally, EEC windows were fused together and smoothed by bicubic interpolation to obtain the final annotation results. Experimental results showed that the area under the curve (AUC) of this method was 0.92, which was higher than that of other comparison CNN-based networks. Du *et al*. [16] put forward a CNN-based method to annotate EEC lesions in chromoendoscopic images. The authors applied the semantic segmentation network Deeplabv3+ as an end-to-end system to directly realize the prediction of EEC regions. This method finally obtained a high accuracy of 97.31% and a Mean Intersection-over-Union (MIoU) value of 92.09%. However, the EEC lesions in the chromoendoscopic image data already had high visibility and definition, which ultimately means that the clinical value of this method is limited because it contributes little to the reduction of misdiagnoses.

The above methods suggest the feasibility and effectiveness of applications of DL techniques in lesion annotations on gastroscopic images. However, most existing annotation methods directly use a mature end-to-end DL network to perform prediction and annotation. Therefore, there is a room to improve annotation performance with specific modifications that are performed with respect to the clinical properties of the gastroscopic images and lesions, e.g., novel loss functions, changes in the internal structure of the DL networks, etc.

In recent years, SegNet [17], U-Net [18], and Deeplabv3+ [19] have become highly recognized in the research field of image segmentation. Different from traditional CNN, these DL networks were designed with a decoder module in which the image features are up-sampled to reconstruct the large-sized feature maps. This property makes the networks be competent to perform classifications in pixel-level. Therefore, they were frequently used in the works which focus on lesion detection and annotation in medical images. For example, SegNet was employed in [20] and [21] for the annotation of pulmonary nodule lesions in CT images and polyp lesions in colonoscopic images, respectively. U-Net was utilized in [9] to annotate breast tumors in MRI images. Deeplabv3+ was applied in [16] and [22] in the annotation of EEC lesions in chromoendoscopic images and of brain tumors in MRI images, respectively.

Motivated by the methods above, we designed a novel method on the basis of a semantic segmentation network: Deeplabv3+, to realize the accurate annotation of EEC lesions in gastroscopic images. In this method, a depth map of a gastroscopic image was initially calculated by a CNN named "High-Resolution Depth Estimation Network" (HRDEN). Then, the original RGB gastroscopic image and the corresponding depth map were sent to a modified Deeplabv3+ network to obtain the final prediction of EEC regions. We name the proposed method "Depth-Deeplabv3+-Based Annotation" (DD-BA). Several experiments were implemented on real clinical images to validate the correctness and performances of the proposed method, and we finally obtained satisfactory results compared with other related methods.

The rest of the paper is organized as follows. Section II introduces details of the used images and the proposed EEC annotations method. The experiments and corresponding results are reported in section III. We further analyze the results and make the summary in section IV and draw conclusion in section V.

II. MATERIALS AND METHODS

A. MATERIALS

In total, 4231 esophageal images of 732 patients were used in this work, collected from the Digestive Endoscopy Center of the West China Hospital in Sichuan, China. The gastroscopic image capture devices used were OLYMPUS GIF-Q260 and Q290 gastroscopes. The collected original images were obtained at three sizes: 1920 px × 1080 px, 768 px × 576 px, and 480 px × 360 px, as shown in Fig. 1. Among the 4231 images, 3190 were EEC images that recorded 4310 EEC regions in total. If the lesions were diffuse over a large area or there were several small lesions very close to each other, we treated them as one integrated lesion [23]. The other 1041 images were non-EEC images that recorded normal tissue or benign inflammatory mucosa in the esophagus.

Most of the 4310 EEC lesions were accompanied by visible coarse mucosa, and some were accompanied by bleeding, edema, protuberances, or festering. All of these EEC lesions were confirmed by biopsy, and the corresponding ground truth (GT) annotations were made by two clinicians. Before the experiments, we calculated the size distribution of the EEC lesions and divided these EEC regions into ten size-levels, as shown in Fig. 2. In this figure, AP_{E} denotes the area ratio of each EEC region to the whole image area, which reflects the size of each EEC lesion region. P_{E} denotes the quantity proportion of images with the corresponding-sized

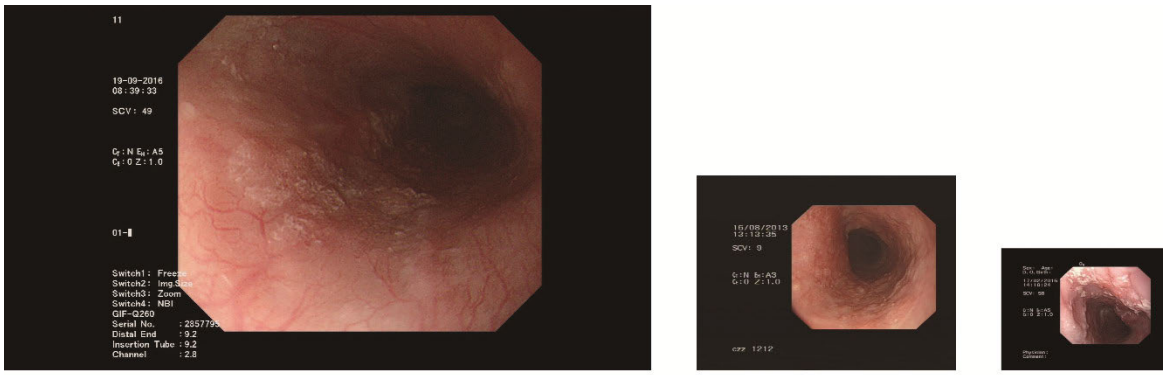

FIGURE 1. Original gastroscopic images at three different sizes. (a) 1920 px × 1080 px, (b) 768 px × 576 px, (c) 480 px × 360 px.

TABLE 1. Number of images, EEC lesion regions, and patients in the image groups used in this work.

	Training group			Validation group			Test group			All groups		
	EEC	Non-EEC	Total	EEC	Non-EEC	Total	EEC	Non-EEC	Total	EEC	Non-EEC	Total
Number of Images	2190	731	2921	500	155	655	500	155	655	3190	1041	4231
Number of EEC Regions	2947	0	2947	671	0	671	692	0	692	4310	0	4310
Number of Patients	525	173	525	106	42	106	101	45	101	732	260	732

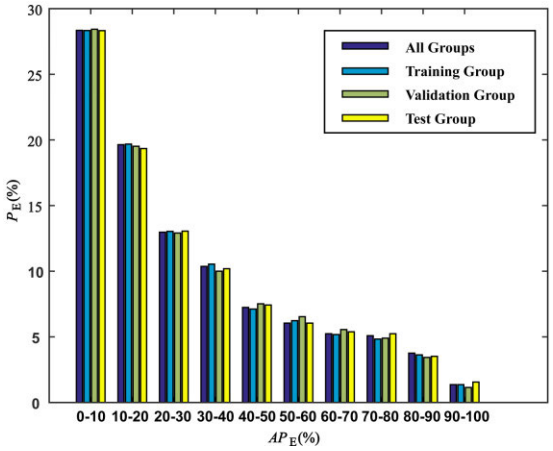

FIGURE 2. Size distribution of the EEC regions in the gastroscopic images.

EEC regions to the whole image group. The navy-blue parts of Fig. 2 show that, in 3190 EEC images, the higher the AP_E level, the lower the corresponding P_E value. Small EEC regions (i.e., EEC regions with AP_E values of 0%-10%) occur most frequently, whereas the EEC regions in the highest size-level rarely appear. Therefore, to enhance the generalizability and reliability of the DL model trained by gastroscopic images, we took this size distribution into account. In this work, three image groups were generated: a training group, a validation group, and a test group. The training group was used to train the parameters of the DL network. The validation group was utilized in the training procedure to adjust the hyper-parameters and reduce overfitting. The test group was used for the evaluation of performances. During the generation of the three groups, EEC images were randomly chosen at each size level according to the corresponding original P_E values for the three groups. Thus, a training EEC group of 2190 images, a validation EEC group of 500 images, and a test EEC group of 500 images were created, as summarized in Table 1. Using our selection method for the EEC images, the three image groups have similar EEC size distributions (as shown in Fig. 2), but there was no overlap in patients' names among them, that is, the images associated with one patients just appear in one group. Thus, the fully trained model could be more stable, and the results calculated by our test group could be more reliable.

The 1041 non-EEC images were also divided into training, validation, and test groups, which consisted of 731 images, 155 images, and 155 images, respectively. The three non-EEC groups were combined with the corresponding EEC groups; thus, the numbers of images in the combined training group, test group and validation group were 2921, 655, and 655 (approximately 70%, 15%, and 15% of all gastroscopic images), respectively. In this work, the 173 patients in the training non-EEC group were parts of the 525 patients in the training EEC group; therefore, the total number of patients in the whole training group remained 525. The validation group and test group also had this property, as described in the last row of Table 1.

At the start, the training group contained 2190 EEC images, which was insufficient for training a DL network. Therefore, the image augmentation procedure

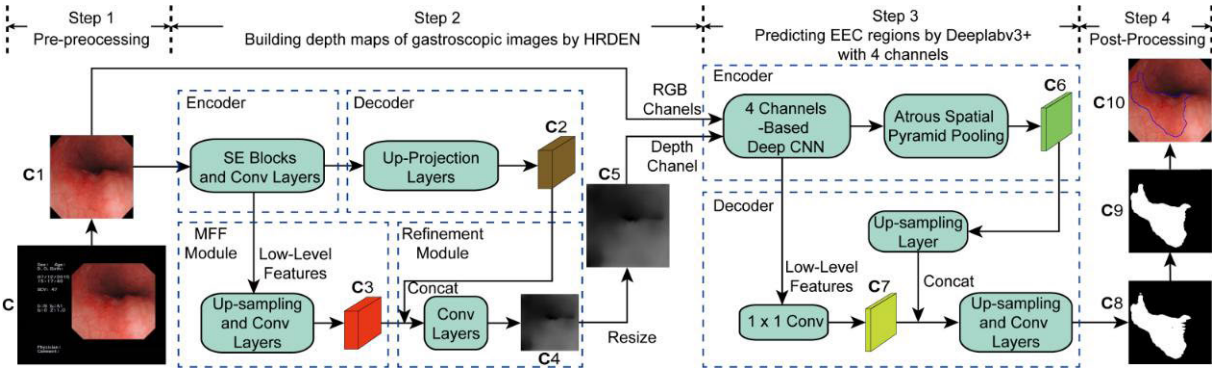

FIGURE 3. Main processes of our EEC annotation method.

became indispensable [24], [25]. In this work, the EEC images in the training group were augmented in the following ways: 1) clockwise rotation by 90 degrees; 2) clockwise rotation by 180 degrees; 3) clockwise rotation by 270 degrees; 4) horizontal flipping; 5) brightness increase by 25%; and 6) brightness decrease by 25%. We randomly chose four of the above six ways to perform the augmentation for each training EEC image. After that, the amount of training EEC images was expanded to $2190 \times 5 = 10{,}950$, which was adequate for the training procedures. Thus, the stability of the fully trained model was ameliorated, and the overfitting problem was prevented. It is worth indicating that the 731 non-EEC images in the training group were not augmented in our experiments because too many negative samples (i.e., non-EEC images) in the training group will damage the model's sensitivity to positive samples (i.e., EEC images); this could cause a sharp decrease in the EEC detection rate in the validation and testing groups, which cannot be tolerated in clinical use. Therefore, the number of training images was finally fixed as $10{,}950 + 731 = 11{,}681$.

B. METHODS

The proposed DD-BA consists of four steps, as shown in Fig. 3. Step 1 is pre-processing, which removes the black background regions and equalizes the sizes of the gastroscopic images. In Step 2, the pre-processed images are sent to HRDEN to realize the prediction of depth maps. Then, the RGB images and the corresponding depth maps are delivered to the 4-channel Deeplabv3+ network in Step 3 to obtain the predictions of the EEC regions. Finally, a post-processing step is performed on the EEC prediction results to complete the final annotations. The details of the four steps are systematically described as follows.

1) PRE-PROCESSING

The black background regions of gastroscopic image **C** record some auxiliary text information such as time, sex, and age. These regions do not contribute to the annotations, so they are firstly removed by fixed windows [26]. The original sizes among our gastroscopic images are different (see Fig. 1); therefore, we subsequently resize the remaining area of the gastroscopic images to 384 px × 384 px through bilinear interpolation. In this way, the sizes of the images are made identical, as shown in image **C**1 of Fig. 3.

2) BUILDING DEPTH MAPS OF GASTROSCOPIC IMAGES BY HRDEN

The depth map records the 3D-shape information of the objects. In recent years, depth information has been used in the detection of lesions in which shape features contribute a lot to the detection of lesions [27], [28], such as polyps. Inspired by these efforts, this work presents to utilize the depth information for improving the prediction performances of the EEC regions since it can help the computer to better recognize the internal structure of the esophagus. The usefulness of depth information will be further discussed and analyzed in section IV.

The DL-based network described in [29] was applied in this work. The network consists of 4 modules: an encoder, a decoder, a multi-scale feature fusion (MFF) module, and a refinement module, as shown in the step 2 of Fig. 3. The first two modules are the basic parts that realize the preliminary extraction of high-level depth feature (feature map **C**2), and the latter two modules are used for the extraction of multi-scale features (feature map **C**3) and high-resolution reconstruction of the depth maps (**C**4), respectively. For ease of description, the network in [29] is called "High Resolution Depth Estimation Network" (HRDEN). Compared with traditional depth estimation networks, HRDEN made two improvements. The first one is the extraction and fusion of multi-scale features, which reduced The loss of spatial resolution in the estimated depth maps; and the second one is the modified loss function which further enhanced the accuracy of reconstruction [29]. By virtue of the above improvements, HRDEN achieved state-of-the-art performance in depth prediction [29]. Therefore, HRDEN is utilized to calculate the depth maps of the gastroscopic images here.

HRDEN effectively reduced the loss of spatial resolution via the utilization and refinement of multi-scale features. However, to simplify the decoding of high-level features,

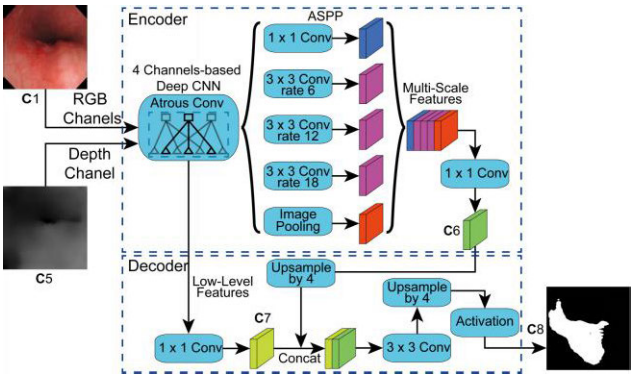

FIGURE 4. Main framework of the Deeplabv3+ network with 4 channels.

this DL network does not up-scale the feature maps to the size of the original input. As a result, the size of the estimated depth map is less than 1/2 of that of the input image [29]. In this work, similarly, the size of **C**4 is only 152 px × 114 px. Therefore, we performed bilinear interpolation on **C**4 to increase the size of the depth maps back to 384 px × 384 px, which is denoted by **C**5. The final depth map **C**5 is used as the input of the 4th channel in the subsequent DL network (Fig. 3).

3) PREDICTING EEC REGIONS BY DEEPLABV3+ WITH 4 CHANNELS

Deeplabv3+ [19] is a semantic segmentation network based on a traditional encoder-decoder structure. This DL network is famous for three unique points: 1) an improved Xception-based deep CNN module, which utilizes depth-wise separable atrous convolutions to replace the max-pooling layers; 2) the Atrous Spatial Pyramid Pooling (ASPP) module, which uses parallel atrous convolutions at different rates to capture the multi-scale information; and 3) the fusion of low-level features and multi-scale deep features, which leads to better segmentation results. Deeplabv3+ has shown satisfactory performance in similar lesion annotation works [16], [22], therefore, it is applied here to predict the EEC regions in the gastroscopic images. It is worth reminding that we made some modifications to the input and output layers of this network to achieve better EEC annotation performances. The main framework of the modified Deeplabv3+ is shown in Fig. 4.

The color images are initially processed by a deep CNN module in the encoder of Deeplabv3+. However, the original version of this module only supports 3 channels of inputs (i.e., red, green, and blue channels). To utilize the additional depth channel **C**5, the number of channels of the convolution kernels in the first layer is modified from 3 to 4. The subsequent layers of this deep CNN module remain unchanged to preserve its remarkable feature extraction performance. Next, ASPP is performed. Features extracted by the 4-channel–deep CNN are sent to parallel layers, the main contents of which are atrous convolutions with different atrous rates. After ASPP, the features from different scales are fused through a 1 × 1 convolution. A feature map, **C**6, that records the multi-scale information is obtained. In the decoder stage, the low-level features extracted by the 4-channel–deep CNN are simplified via a 1 × 1 convolution layer to obtain the low-level feature map **C**7. After that, skip connection is performed. The low-level features **C**7 and the high-level multi-scale features up-sampled from **C**6 are concatenated. Finally, the fused features are processed by convolution, up-sampling and activation layers. It should be mentioned that in the original Deeplabv3+, the function used in the last activation layer is Softmax because the original Deeplabv3+ was designed for the segmentation of 21 different kinds of objects in natural scenes [19]. However, our DD-BA only needs to distinguish two objects, namely, EEC regions and non-EEC regions. Therefore, we changed the activation function from Softmax to Sigmoid to allow the network to become more competent to the binary classification task of EEC annotation. After the classification is performed at the pixel level by the activation layer, a binary image **C**8 is obtained, the white regions (i.e., the ''1'' regions) of which record the location of EEC lesions.

4) POST-PROCESSING

The outputs of the 4-channel Deeplabv3+ are sufficient to predict the location of EEC lesions in most cases, but the visual accuracy of these predictions also deserves attention. In the clinic, one of the purposes of EEC annotation is to provide useful guidance for resections or dissections of lesion regions. Obviously, annotations with poor appearances (e.g., jagged edges, a large number of small holes) are not welcomed because it is difficult for clinicians to resect the lesions in jagged or perforated ways. However, the visual accuracy of the results predicted by Deeplabv3+ are sometimes unsatisfactory. For example, the upper-left part of the white region of image **C**8 shown in Fig. 4 is jagged. This is caused by the side-effect of atrous convolutions in Deeplabv3+. Generally, if the edges of the EEC lesions are not obvious enough in the gastroscopic images, the ''holes'' of the atrous convolution kernels may lead to unstable predictions in those edge areas. In most cases, these jagged edges do not affect the accuracy of annotation too much, but they reduce the visual accuracy of the annotation results. Therefore, to make the annotation results more acceptable to clinicians, we proposed a post-processing step to smooth jagged and perforated prediction results. The details are described as follows.

For the binary image **C**8, we firstly filled the small holes with ''0'' and ''1'' if their number of pixels are lower than the threshold n_1 and n_2, respectively. After that, morphological ''close'' and ''open'' operations are performed sequentially. For the operation of ''close'', the morphological ''erosion'' and ''dilation'' are implemented via a disk-shaped structuring elements of radius r_1. Similarly, the morphological ''open'' is realized via a disk-shaped elements of radius r_2. We repeat the above hole-filling and morphological operations until the ''1''

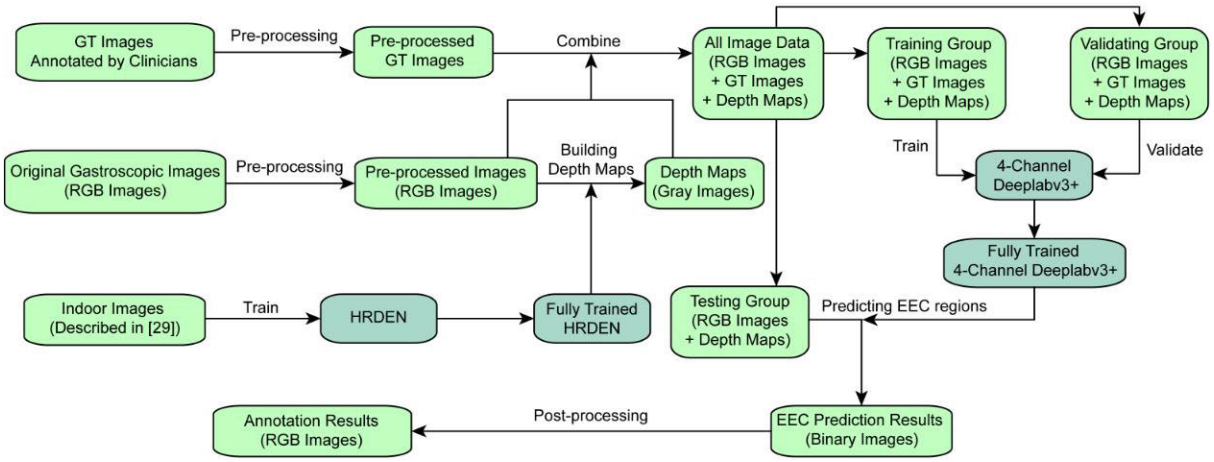

FIGURE 5. Implementation details of our EEC annotation method. The units marked by blue color represent DL models, and those marked by green color represent image data.

regions of **C**8 become constant. The constant binary image is denoted by **C**9. As shown in Fig. 3, the edges of **C**9 have become smooth, and the problem of the low visual accuracy is solved. Finally, we obtain the final annotated image **C**10 by mapping **C**9 to the original RGB image **C**1. A complete flow of DD-BA is finished.

It should be noted that in DD-BA, the two DL-networks, HRDEN and 4-channel Deeplabv3+, were separately trained because there was no GT for depth map available for our gastroscopic images. For this reason, we utilized the indoor image data described in [29] to train HRDEN. The details of this training will be described in the next section. The implementation processes of the proposed method are summarized in Fig. 5. This figure illustrates the details of ''how the image data is used'' and ''how the two DL networks are trained and applied''.

III. EXPERIMENTS AND RESULTS

Two experiments were conducted to verify the performances of DD-BA. The first experiment was the validation of the proposed idea, which utilized the depth maps as additional depth information of the DL network. The second experiment was the comparison between DD-BA and other state-of-the-art DL-based methods. The performances of the methods were evaluated in several ways. Recall (Rec), Precision (Pre), and Dice Similarity Coefficients (DSC) [30] were firstly used to evaluate annotation performances. Each index is expressed by the mean value and standard deviation (SD). In our experiment, positive pixels were defined as those in EEC regions, whereas negative pixels were those outside EEC regions. Thus, the Rec, Pre, and DSC were calculated using (1)-(3), respectively:

$$\mathrm{Pre} = TP_p/(TP_p + FP_p) \tag{1}$$

$$\mathrm{Rec} = TP_p/(TP_p + FN_p) \tag{2}$$

$$\begin{aligned}\mathrm{DSC} &= 2\times TP_p/(2\times TP_p + FN_p + FP_p)\\ &= 2\times Pre\times Rec/(Pre + Rec)\end{aligned} \tag{3}$$

where TP_p, FN_p, and FP_p denote the number of true positive pixels, false negative pixels, and false positive pixels in an EEC prediction, respectively. With these three indexes, DSC is the most crucial composite index. It can reach 100% only when an annotation region and the corresponding GT completely overlap with each other. Secondly, the EEC detection rate (DR) was used to evaluate the performance of detecting EEC lesions, and it was calculated by (4):

$$\mathrm{DR} = N_d/N_t \tag{4}$$

where N_d denotes the number of detected EEC regions and N_t denotes the total number of EEC regions (fixed at 692 in the test group). An EEC region was considered to be detected only when the DSC of the corresponding annotation was higher than 20% [23].

In addition, False Positive Annotations (FPAs, i.e., the annotation regions with a DSC value of 0%) also deserve attention. As we know, a qualified annotation method should not only annotate lesions accurately but also produce fewer and smaller FPAs [31]. Therefore, the mean values and SDs of two indexes, NP_FP and AP_FP [23], were used to measure FPA. NP_FP is the ratio of the number of FPAs to the number of gastroscopic images, which reflects the frequency of occurrence of FPAs. AP_FP is the ratio of the number of FPAs' pixels to the total number of image pixels and reflects the size and area ratios of the FPAs. We divided the test image group into two parts: the EEC test group and the non-EEC test group (as described in Table 1) to respectively evaluate the FPA outputs, the NP_FP and AP_FP of the EEC test group are denoted by NP_FP-E and AP_FP-E, respectively; and those of the non-EEC test group are denoted by NP_FP-N and AP_FP-N, respectively.

In our experiments, the basic settings and parameters were optimized and fixed as follows: In HRDEN, the training images were obtained from the NYU-Depth V2 database [29], [32], which consists of a large amount of indoor images and the corresponding GTs of depth. In the training process,

the activation function used in this network was ReLU, and the loss function L was the same as that of [29], as shown in (5)-(8):

$$l_{\text{depth}} = \frac{1}{n}\sum_{i=1}^{n} F(e_i) \tag{5}$$

$$l_{\text{grad}} = \frac{1}{n}\sum_{i=1}^{n} (F(\nabla_x(e_i)) + F(\nabla_y(e_i))) \tag{6}$$

$$l_{\text{normal}} = \frac{1}{n}\sum_{i=1}^{n} (1 - \frac{\sqrt{\langle \boldsymbol{n}_i^d, \boldsymbol{n}_i^g \rangle}}{\sqrt{\langle \boldsymbol{n}_i^d, \boldsymbol{n}_i^d \rangle}\sqrt{\langle \boldsymbol{n}_i^g, \boldsymbol{n}_i^g \rangle}}) \tag{7}$$

$$L = l_{\text{depth}} + l_{\text{grad}} + l_{\text{normal}} \tag{8}$$

In (5)-(8), l_{depth}, l_{grad}, and l_{normal} measures the errors in depth, gradients, and surface normals, respectively. $i = 1, 2, 3, \ldots, n$, where n is the total number of pixels of an output depth map. The local error $e_i = ||d_i - g_i||_1$, where d_i is the depth estimate at the i^{th} pixel and g_i is the corresponding GT value. Function $F(a) = \ln(a+0.5)$. $\nabla_x(e_i)$ and $\nabla_y(e_i)$ are the spatial derivative of e_i computed at the i^{th} pixel with respect to x direction and y direction, respectively. $\boldsymbol{n}_i^d$ and $\boldsymbol{n}_i^g$ are the surface normals of the depth estimate and GT, calculated by $\boldsymbol{n}_i^d \equiv [-\nabla_x(d_i), -\nabla_y(d_i), 1]^{\text{T}}$ and $\boldsymbol{n}_i^g \equiv [-\nabla_x(g_i), -\nabla_y(g_i), 1]^{\text{T}}$, respectively. In (8), the operation $\langle \boldsymbol{a}, \boldsymbol{b} \rangle$ represent the inner product of the vectors $\boldsymbol{a}$ and $\boldsymbol{b}$ [29].

For the other parameters and hyper-parameters of HRDEN, they are the same as those described in [29]. In 4-channel Deeplabv3+, the parameters were pre-trained in the PASCAL VOC2012 dataset [33]. However, the pre-trained parameters in the first convolution layer were not applied because the number of channels was changed in the first layer of DD-BA. Therefore, the parameters in this layer were randomly initialized. In the training procedure, the loss function used was the Dice coefficient, and the optimizer was AdaDelta with an initial learning rate of 1 and a mean decay rate of 0.95. The activation function used in the convolutional layers was ReLU. The batch size was 6, and the number of epochs was set to 50, which was sufficient for the network to converge in most cases. Other hyper-parameters were the same as those described in [19]. In post-processing, the thresholds n_1 and n_2 for the ''hole filling'' operation were both 100; the radius of the disk-shaped structuring elements were set to $r_1 = 7$ px and $r_2 = 4$ px.

In this work, the programming platforms were Python 3.6.4 (for HRDEN and Deeplabv3+) and MATLAB 2016a (for the other operations). Training procedures were implemented on double same GPUs of NVIDIA GeForce RTX 2080Ti.

A. EFFECTIVENESS VERIFICATION OF DEPTH INFORMATION USED IN DD-BA

To verify the utility of the 4$^{\text{th}}$ depth channel and evaluate the improvements brought by the combination of depth and RGB information, the first experiment was to compare the performances of our DD-BA and the Deeplabv3+-Based Annotation (D-BA). In D-BA, the pre-processing and post-processing steps were the same as those of DD-BA, but the step of building depth maps by HRDEN was removed. Therefore, the Deeplabv3+ network used in D-BA only employed 3 input channels (i.e., RGB). Other basic settings and hyper-parameters of Deeplabv3+ remained unchanged from those of DD-BA. The detection and annotation performances of the two methods are shown in Fig. 6 and Table 2.

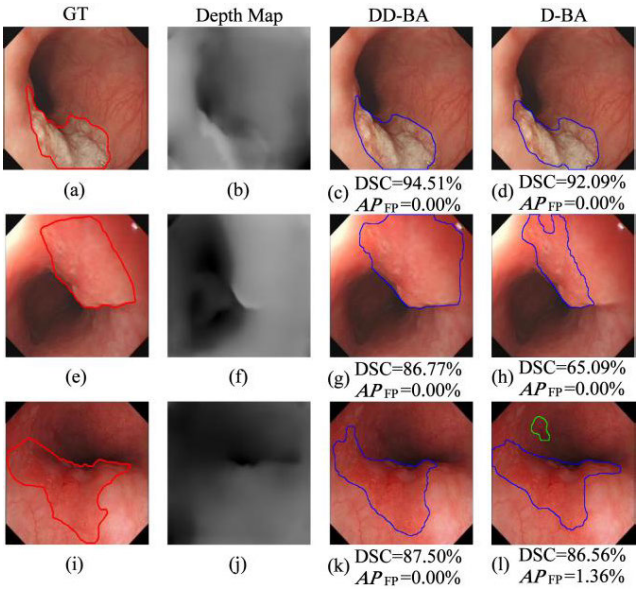

FIGURE 6. GTs, depth maps, and the corresponding annotation results of DD-BA and D-BA. The depth maps in the 2$^{\text{nd}}$ column are calculated by HRDEN. In the 3$^{\text{rd}}$ to 4$^{\text{th}}$ columns, the regions marked by blue circles are correct annotations, and those marked by green circles represent FPAs.

As seen in the 3$^{\text{rd}}$ column of Table 2, the DR of D-BA is 95.95%, which reveals that the features of the original RGB channels are a good foundation for EEC detection. However, DD-BA detected 11 more EEC regions, which led to a higher DR of 97.54%. This result reflects that the additional depth information improved the EEC detection performance. Focusing on the annotation performances, as displayed in the 4$^{\text{th}}$ to 6$^{\text{th}}$ columns of Table 2, D-BA achieved qualified annotation performances using RGB images; the mean DSC, Rec and Pre of this method are 72.95%, 74.06%, and 79.99%, respectively. Moreover, the 2$^{\text{nd}}$ row of Table 2 reflects that the addition of the depth channel does not affect the Pre substantially. The Pre values of DD-BA and D-BA differ by only 0.35%. However, the combination of depth and RGB channels resulted in a remarkable improvement of 4.07% in Rec. By virtue of this improvement, the mean DSC of DD-BA finally reached 74.43%, which is 1.48 percentage points greater than that of D-BA. Therefore, the addition of a depth channel also contributed to improving the annotation performance. Examples in Fig. 6 may be some good explanations for this result. As shown in Fig. 6 (a), there is a protruding EEC lesion which shows white and coarse surface. When annotating this EEC region, both of the two methods performed well. The DSC of them are both higher

TABLE 2. EEC detection and annotation performances, and FPA outputs of DD-BA and D-BA.

		DR	DSC	Rec	Pre	NP_{FP}-E	AP_{FP}-E	NP_{FP}-N	AP_{FP}-N
DD-BA	Mean	97.54% (675/692)	74.43%	78.13%	79.64%	0.1380 (69/500)	0.27%	0.1677 (26/155)	0.98%
	SD		0.1731	0.1998	0.2217		0.0171		0.0388
D-BA	Mean	95.95% (664/692)	72.95%	74.06%	79.99%	0.2120 (106/500)	0.44%	0.2129 (33/155)	0.99%
	SD		0.1782	0.2091	0.2102		0.0220		0.0362

than 90% and that of DD-BA is just 2.42% higher than that of D-BA. This result reveals that depth information contributes less to the enhancement of annotation performance when the abnormal color and texture of EEC regions are clearly visible. However, when the irregular color and texture of EEC regions are not salient enough, the addition of depth information may greatly improve the annotation performance. For example, the EEC region in Fig. 6 (e) is inflammatory tissue, but it is without visible bleeding or festering. As a result, this EEC tissue does not appear very coarse. In this case, D-BA only annotated the left part of this EEC region, and the flat part to the right was omitted. It is evident that this annotation is with low Rec value (50.11%), and the DSC becomes lower accordingly (merely 65.09%). By contrast, DD-BA utilized the depth map in Fig. 6 (f). Although this estimated depth map is not 100% accurate, it clearly indicates the integrity of the tissue in the upper-right part of the image. Therefore, the annotation result of DD-BA recalled the vast majority of this EEC region and achieved a satisfactory DSC of 86.77%.

For FPA outputs, as described in the 7th to 10th columns of Table 2, the NP_{FP}-E and AP_{FP}-E values of DD-BA are 0.1380 and 0.27%, respectively, which are only approximately 60% of those of D-BA. These results indicate that using depth and RGB information at the same time can help to reduce the FPAs in EEC images. The reason for this is that EEC lesions often appear together with benign inflammatory mucosa or lesions, which sometimes show irregular color and texture that are similar to those of EEC lesions. In these cases, the additional depth information can help the method to better recognize them. An example is shown in Figs. 6(i)-(l). In Fig. 6 (i), there is a large EEC region in the center, surrounded by benign coarse mucosa from the upper-left part. The coarse mucosa shows irregular texture and a dark red color, similar to some of the EEC lesions. However, the depth map in Fig. 6 (j) clarifies that this region of coarse mucosa is far from the camera. Therefore, its dark appearance is caused by a relative lack of illumination. As a result, the subsequent 4-channel Deeplabv3+ refuses to hastily annotate it as a EEC lesion and FPAs are avoided. By contrast, D-BA without using depth input channel produced the FPAs, as shown in the right part of Fig. 6 (l). In addition, the 3rd and 5th rows of Table 2 show that the SDs of the indexes of the two methods are similar, except for the SD of AP_{FP}-E. The SD of AP_{FP}-E of DD-BA is 0.0171, which is 22% lower than that of D-BA. This suggests that utilizing depth information also enhanced the stability of the method in constraining FPAs in EEC images. However, this improvement is not found in non-EEC test group. The AP_{FP}-N of the two methods are both close to 1%. This means the depth information does not contribute very much to the reduction of FPAs in non-EEC images. It could be explained by the fact that most of the non-EEC images used in this work were normal images which showed no coarse mucosa. The irregular texture of benign inflammatory mucosa or lesions did not appear very often. Thus, the improvement in FPA was restricted.

In summary, the above experimental results reveal the superiority brought by utilizing additional depth information. It can not only bring extra enhancements to the EEC detection and annotation performances but also limit the number and size of FPAs annotated in EEC images. Therefore, the combination of depth features and RGB information is demonstrated to be a feasible and effective strategy for EEC annotation.

B. COMPARISONS BETWEEN DD-BA AND OTHER STATE-OF-THE-ART DL-BASED METHODS

SegNet [17], U-Net [18], and Deeplabv3+ [19] are three state-of-the-art DL networks used for semantic segmentation. All three networks are based on the traditional encoder-decoder structure. SegNet is an improved fully convolutional network (FCN) that utilizes max-pooling indices to replace deconvolution operations in the decoder. This operation simplifies the up-sampling procedures and allows the network to retain more high frequency details. Thus, the segmentation results become more precise. U-Net is famous of its special structure, which applies multiple skip connections to combine the outputs in decoding layers and the high-resolution features in encoding layers. Compared with traditional DL-based segmentation networks, U-Net effectively reduced the loss of features in down-sampling layers. Therefore it achieved remarkable segmentation performance. Deeplabv3+ was already fully introduced in section II *B*, part 3). In recent years, the three networks have become highly recognized in the research field of image segmentation, and they are frequently used in lesion detection and annotation in medical images as well. Therefore, the three state-of-the-art networks deserve comparison with the proposed DD-BA. In the first experiment, we demonstrated

TABLE 3. EEC detection and annotation performances and FPA outputs of the three methods.

		DR	DSC	Rec	Pre	NP_{FP}-E	AP_{FP}-E	NP_{FP}-N	AP_{FP}-N
DD-BA	Mean	97.54% (675/692)	**74.43%**	**78.13%**	**79.64%**	**0.1380 (69/500)**	**0.27%**	**0.1677 (26/155)**	**0.98%**
	SD		**0.1731**	0.1998	0.2217		**0.0171**		**0.0388**
DS-BA	Mean	95.81% (663/692)	73.16%	74.51%	76.82%	0.2580 (129/500)	0.76%	0.3613 (56/155)	2.31%
	SD		0.1899	**0.1895**	**0.2132**		0.0236		0.0551
DU-BA	Mean	97.40% (674/692)	70.87%	75.36%	71.96%	0.3940 (197/500)	0.74%	(1.1806) 183/155	3.64%
	SD		0.1775	0.1996	0.2229		0.0271		0.0609

that DD-BA is superior to the traditional Deeplabv3+-based annotation method. In this experiment, we mainly analyzed the performances of SegNet-based annotation and U-Net-based annotation and compared the proposed DD-BA with these two methods. In addition, the experimental results recorded in Section III *A* have shown the utility of DD-BA's 4th depth channel. Therefore, we added the same depth channel to the first layer of SegNet and U-Net, to ensure the fairness of comparison. Thus, the two comparison methods were named Depth-SegNet-Based Annotation (DS-BA) and Depth-U-Net-Based Annotation (DU-BA), respectively.

For these two comparison methods, we only switched the DL step of DD-BA from Deeplabv3+ to SegNet (for DS-BA) and U-Net (for DU-BA). The remaining pre-processing and post-processing procedures remained unchanged from those of DD-BA to avoid differences in the performances caused by the changes in these steps. In the 4-channel SegNet of DS-BA and 4-channel U-net of DU-BA, the network structure and number of layers were the same as those described in [17] and [18], respectively (except the number of channels of the first layer). To obtain better binary prediction results, the activation function of the last layer was set to Sigmoid for both networks. When training the two networks, the basic settings, i.e., batch size, optimizer, learning rate, activation function of convolution layers, and loss function, were all the same as those of 4-channel Deeplabv3+ in DD-BA. The two networks were both trained with 50 epochs without using transfer-learning. The performances of DD-BA, DS-BA and DU-BA are shown in Table 3 and Fig. 7.

As illustrated in the 3rd column of Table 3, all three methods show qualified DRs higher than 95%, which means that the three methods performed well in the detection of EEC lesions. However, the best DR is achieved by DD-BA, with a value higher than 97.5%. Therefore, DD-BA has the best capability of EEC detection. The annotation performances are summarized in the 4th to 6th columns of Table 3. As seen in these columns, the Rec and Pre of DD-BA reached 78.13% and 79.64%, respectively, which are both higher than the respective values from DS-BA and DU-BA. Moreover, the DSC of DD-BA is 74.43%, which is 1.27% and 3.56% higher than that of DS-BA and DU-BA, respectively. In particular, the Pre of DD-BA is 7.68% higher than that of DU-BA, which is the largest margin observed in these

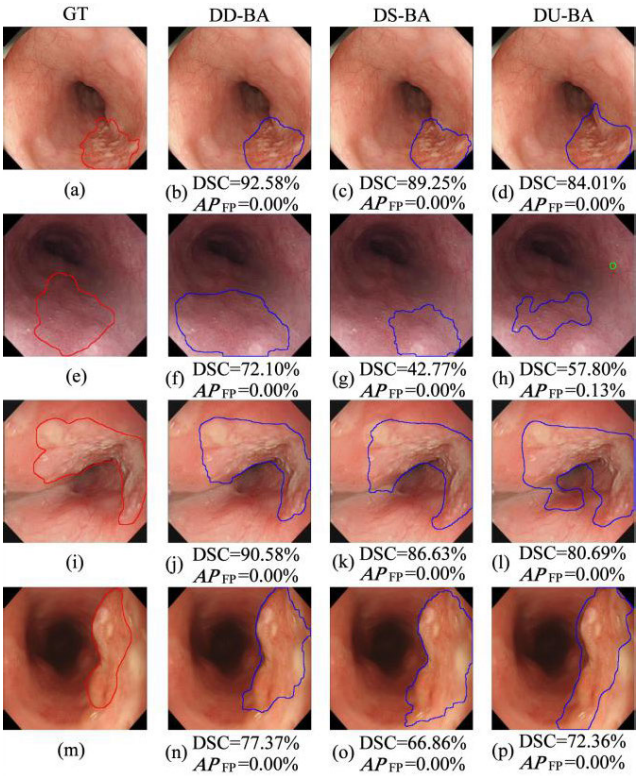

FIGURE 7. GTs and the corresponding annotation results for DD-BA, DS-BA and DU-BA. The regions marked by blue circles are correct annotations, and those marked by green circles represent FPAs.

performance indexes. These results fully prove that the proposed DD-BA comprehensively outperforms the two state-of-the-art, DL-based methods in terms of annotation performances. In addition, it also indicates that even depth and RGB information has both been utilized, the 4-channel SegNet and 4-channel U-Net cannot achieve the best performances for some special cases, which are illustrated in Fig. 7. In Fig. 7(a), there is a small EEC lesion with only a small amplitude of depth variation, and its irregular texture is very salient. In this case, all three methods performed well, and their DSCs are all higher than 80%. However, if the irregular color and texture of the EEC regions are not obvious, 4-channel SegNet and U-Net are

not sensitive enough to them. An examples of this is shown in Figs. 7(e)-(h). In Fig. 7(e), there is a large EEC lesion and its mucosa is not quite coarse. DS-BA and DU-BA just annotated a part of this lesion. However, DD-BA avoided this shortcoming and recalled the vast majority of this EEC region; thus, it obtained a high DSC of 72.10%. In Fig. 7(i), the coarse texture and white color of the EEC lesion are visible, but the edges of this EEC region are not clear. In this case, DD-BA successfully located those edges, whereas both DS-BA and DU-BA misjudged some of them. As a result, the DSCs of the two comparison methods were unsatisfactory. In addition, the cases shown in Figs. 7(d), (l), and (p) reflect that U-Net is more inclined to broadly annotate the EEC lesions, which is the reason why DU-BA obtained the lowest mean Pre value (71.96%).

In terms of FPAs, as depicted in the 7th to 10th columns of Table 3, the NP_{FP}-E, AP_{FP}-E, NP_{FP}-N, and AP_{FP}-N values of DD-BA are all lower than those of DS-BA and DU-BA, and the corresponding SDs of these indexes in DD-BA are also the lowest among those of the three methods. These results suggest that DD-BA has the best performance in constraining FPAs. Obviously, the Deeplabv3+ network contributed greatly to this performance, because the results in Table 2 and Table 3 have both shown that the additional depth information cannot effectively inhibit the output of FPAs in the non-EEC image group. Therefore, the advantage of using Deeplabv3+ as a segmentation network, with respect to constraining FPAs, is significant. In summary, whether for EEC annotation or constraining FPAs, DD-BA is superior to the two comparison methods.

IV. DISCUSSION

This study proposed an annotation method based on DL networks named DD-BA, which outperforms other state-of-the-art DL-based annotation methods. The most important innovation of DD-BA is the utilization of additional depth information. The result in section III *A* have shown that the use of depth information can help the computer better recognize the internal structure of the esophagus. In clinic, actually, the patterns of the tissues' depth variation also have impacts on the judgments of clinicians. An example is shown in Fig. 8. Fig. 8 (a) is an EEC image in which there is an EEC region annotated by the clinician. As seen in this figure, the EEC lesion shows visible coarse mucosa with irregular texture. In another image shown in Fig. 8 (c), the region annotated by the blue circle also shows coarse texture, which looks like that of the EEC region in Fig. 8(a). However, Fig. 8(c) is a non-EEC image which is without any lesion. Figs. 8(b) and (d) are the corresponding depth maps of Figs. 8(a) and (c), respectively. When we compare the two depth maps, we can find that the marked region in Fig. 8(d) has an amplitude of depth variation larger than that of Fig. 8(b). In other words, the depth information of the marked regions in Fig. 8 (a) and (c) are distinct. In most cases, the large amplitude of depth variation will make the object look more compact, and a reflection can be found in Fig. 9.

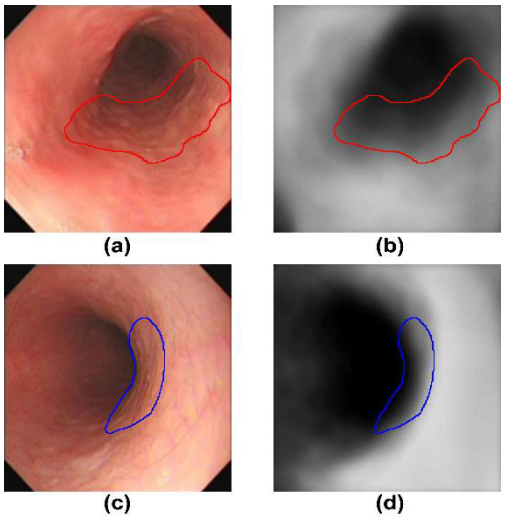

FIGURE 8. Gastroscopic images and their corresponding depth maps. (a) an EEC image, (b) the depth map of (a), (c) a non-EEC image, (d) the depth map of (c).

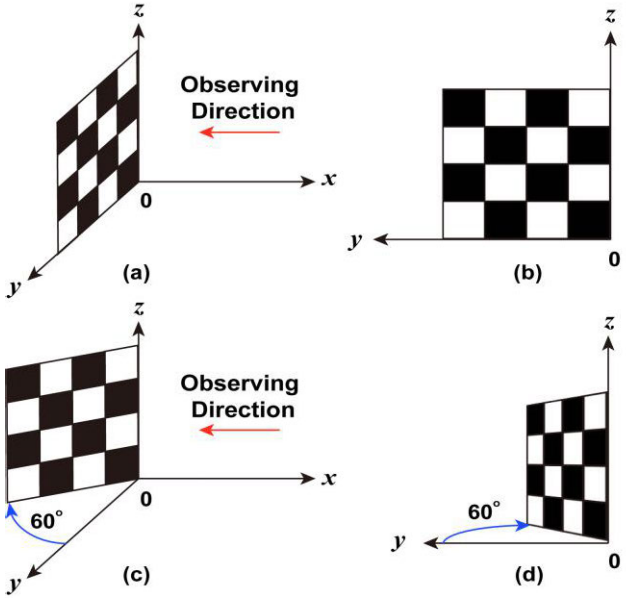

FIGURE 9. An example of the changes in the object's appearance caused by depth variation. (a) a mesh graph located in the y-z plane; (b) The appearance of the mesh graph to the observer's eyes; (c) rotated mesh graph; (d) The appearance of the rotated mesh graph to the observer's eyes.

Fig. 9 (a) shows a mesh graph located in y-z plane. If we assume that the x-axis represents depth, this mesh graph will be with no depth variation because $x = 0$ is hold for this graph. Here, we observe this mesh graph from the direction opposite to the x-axis, we can see the rectangular pattern in Fig. 9 (b). Next, we rotate the mesh graph by 60 degrees in the x-y plane, as shown in Fig. 9 (c). This operation brings a large depth variation (variation in the x-axis) to the mesh graph. After that, we observe this object from the direction same as that of Fig. 9 (a), we will see the

pattern shown in Fig. 9 (d). If we compare the appearances of the mesh graph in Figs. 9 (b) and (d), we can find that the grid pattern in Fig. 9 (d) is severely compressed on the y-axis. This example vividly shows that a large amplitude of depth variation will make an object's appearance denser in the observer's eyes. This is the reason why the phenomenon in Fig. 8 appears. In the gastroscopic images, the appearance of esophageal tissues will seem to be compressed in regions with a large amplitude of depth variation. Therefore, the irregular texture of the marked region in Fig. 8 (c) is caused by the large variation of depth, not by the coarse mucosa. In most cases, this kind of irregular texture will disappear from gastroscopic frames in few seconds if the camera keeps going ahead, but the textures of EEC lesion regions do not have this property. Even if the camera get closer to the lesion tissue, these irregular textures will not disappear. Taking the above analysis into account, the clinician finally made different judgements on the two coarse regions in Fig. 8. Inspired by this phenomenon, we proposed to combine the depth information and RGB information of gastroscopic images to improve the EEC annotation performance. The experimental results finally confirmed the effectiveness of this strategy and proved that it brought two advantages to the proposed method. First, it enhanced the method's adaptability to EEC regions which lack substantial irregularities in color and texture. When these EEC lesions appear, the utilization of the additional depth information enhanced the detection performances and preserved the integrity of the annotated results (e.g., Fig. 6 (g) and Fig. 7 (f)). Second, it reduced the number and size of FPAs in EEC images (e.g., Fig. 6(k)). Owing to these two advantages, DD-BA achieved the best performances compared with two comparison methods. Although the comparison methods are sufficiently sensitive to the coarse regions that appeared in most EEC cases, they sometimes show unsatisfactory annotation performances and FPA outputs because coarse regions are not always equivalent to EEC lesions. However, the combination of depth features and RGB information in DD-BA successfully alleviated the negative effects caused by this problem. In addition, the 4-channel Deeplabv3+ network used in DD-BA also made great contributions to its superior performance. The unique atrous convolution operations and ASPP module in Deeplabv3+ are effective ways to probe multi-scale features because they realize flexible control of the filters' receptive fields. This property further reduced the loss of multi-scale information and led to fewer FPAs in non-EEC images. Although the atrous convolutions caused jagged predictions in some cases, they were easily repaired by the brief post-processing procedure. Ultimately, DD-BA is a good example of a DL-based CAD method, and it is also a successful application of the depth-prediction algorithm.

However, there is still room for improvement of this work. The first potential area for improvement concerns the prediction of depth maps. In the first experiment, we confirmed the utility of the depth information, and the results showed that the depth maps produced by HRDEN are qualified to predict the depth variations in key positions of gastroscopic images. However, if we observe the predicted depth maps carefully, we can find that they are still not 100% accurate (e.g., the upper-left part in Fig 5 (f) and the lower-left corner in Fig. 9 (j)). The main reason for this is that the images used to train HRDEN are all indoor images. The objects in this kind of images are often neat and have clear edges and visible corners, but these properties cannot be obtained in gastroscopic images because the esophagus is a soft tubular organ. Indeed, it is almost impossible to obtain the GTs of depth for gastroscopic images at present because there is no depth camera available for clinical diagnosis. For this reason, we did not utilize gastroscopic image data to train HRDEN. However, if the GTs of depth become available for gastroscopic images in the future, we believe that the performances of DD-BA could be further enhanced. The second possible improvement could lie in the annotation accuracy. In this work, DD-BA utilized state-of-the-art DL techniques and showed better annotation performances than the comparison methods, but the mean DSC of DD-BA is still lower than 80%. There are two reasons for this imperfect performance. One is the diversity of the appearance of EEC lesions, and the other is the frequently appearing benign lesions in the EEC images that disrupted the annotations. If these factors could be considered in the segmentation networks, the annotation performances would be further improved. In future works, we will try to build synthetic image data that record esophagus-like tubular structures to further train the depth-estimation networks for improving the accuracy of the depth prediction results. At the same time, we will keep modifying the internal structure of the semantic segmentation networks based on the clinical properties of EEC to further improve the annotation performances and enhance the robustness of our method.

V. CONCLUSIONS

In this study, we designed a novel framework to annotate EEC lesions in gastroscopic images. Different from the existing DL-based annotation frameworks, the proposed framework utilized depth and RGB information simultaneously. In the depth prediction network HRDEN, the depth maps of gastroscopic images were calculated; in the semantic segmentation network Deeplabv3+, the first layer was modified into a 4-channel form, and the additional depth information was absorbed into this network to enhance EEC prediction performance. On this basis, we developed a novel EEC annotation method named DD-BA. Experimental results confirmed the correctness and necessity of the utilization of additional depth information and proved that the proposed DD-BA outperforms other state-of-the-art DL-based annotation methods. Whether for EEC detection, EEC annotation or FPA outputs, DD-BA showed the most satisfactory performances. In summary, our method has good potential for clinical use and could be helpful in enhancing the accuracy of EEC diagnoses.

ACKNOWLEDGMENT

Hongxiu Jiang is the joint first author.

APPENDIX

The abbreviations and the corresponding full names used in this paper are summarized in Table 4.

TABLE 4. Summary of abbreviations and the corresponding full names.

Full Name	Abbreviations
Atrous Spatial Pyramid Pooling	ASPP
Area Under the Curve	AUC
Computer Aided Dignosis	CAD
Convolutional Neural Network	CNN
Deeplabv3+-Based Annotation	D-BA
Depth-Deeplabv3+-Based Annotation	DD-BA
Depth-SegNet-Based Annotation	DS-BA
Depth-U-Net-Based Annotation	DU-BA
Deep Learning	DL
Detection Rate	DR
Dice Similarity Coefficient	DSC
Esophageal Cancer	EC
Early Esophageal Cancer	EEC
False Positive Annotation	FPA
Fully Convolutional Network	FCN
Ground Truth	GT
High Resolution Depth Estimation Network	HRDEN
Multi-scale Feature Fusion	MFF
Mean Intersection-over-Union	MIoU
Precision	Pre
Rectified Linear Unit	ReLU
Recall	Rec
Standard Deviation	SD
Single Shot multi-box Detection	SSD
Support Vector Machine	SVM

REFERENCES

[1] Globocan. (Mar. 1, 2019). *Estimated Cancer Incidence, Mortality and Prevalence Worldwide in 2012*. [Online]. Available: http://globocan.iarc.fr/Pages/online.aspx

[2] C. Ell, A. May, O. Pech, L. Gossner, E. Guenter, A. Behrens, L. Nachbar, J. Huijsmans, M. Vieth, and M. Stolte, "Curative endoscopic resection of early esophageal adenocarcinomas (Barrett's cancer)," *Gastrointestinal Endoscopy*, vol. 65, no. 1, pp. 3–10, Jan. 2007.

[3] D. C. Whiteman, "Esophageal cancer: Priorities for prevention," *Current Epidemiol. Rep.*, vol. 1, no. 3, pp. 138–148, Jul. 2014.

[4] Y. Chen and J. Lee, "A review of machine-vision-based analysis of wireless capsule endoscopy video," *Diagnostic Therapeutic Endoscopy*, vol. 2012, pp. 1–9, Nov. 2012.

[5] D. K. Iakovidis and A. Koulaouzidis, "Software for enhanced video capsule endoscopy: Challenges for essential progress," *Nature Rev. Gastroenterol. Hepatol.*, vol. 12, no. 3, pp. 172–186, Mar. 2015.

[6] Y. Mori, S. Kudo, H. E. N. Mohmed, M. Misawa, N. Ogata, H. Itoh, M. Oda, and K. Mori, "Artificial intelligence and upper gastrointestinal endoscopy: Current status and future perspective," *Digestive Endoscopy*, vol. 31, no. 4, pp. 378–388, Jul. 2019.

[7] M. Vasilakakis, A. Koulaouzidis, D. E. Yung, J. N. Plevris, E. Toth, and D. K. Iakovidis, "Follow-up on: Optimizing lesion detection in small bowel capsule endoscopy and beyond: From present problems to future solutions," *Expert Rev. Gastroenterol. Hepatol.*, vol. 13, no. 2, pp. 129–141, Feb. 2019.

[8] D. K. Iakovidis, S. V. Georgakopoulos, M. Vasilakakis, A. Koulaouzidis, and V. P. Plagianakos, "Detecting and locating gastrointestinal anomalies using deep learning and iterative cluster unification," *IEEE Trans. Med. Imag.*, vol. 37, no. 10, pp. 2196–2210, Oct. 2018.

[9] J. Zhang, A. Saha, Z. Zhu, and M. A. Mazurowski, "Hierarchical convolutional neural networks for segmentation of breast tumors in MRI with application to radiogenomics," *IEEE Trans. Med. Imag.*, vol. 38, no. 2, pp. 435–447, Feb. 2019.

[10] Y. Zhu, Q. C. Wang, M. D. Xu, Z. Zhang, J. Chen, and Y. S. Zhong, "Application of convolutional neural network in the diagnosis of the invasion depth of gastric cancer based on conventional endoscopy," *Gastrointestinal Endoscopy*, vol. 89, no. 4, pp. 806–817, Apr. 2019.

[11] T. Hirasawa, K. Aoyama, T. Tanimoto, S. Ishihara, S. Shichijo, T. Ozawa, T. Ohnishi, M. Fujishiro, K. Matsuo, J. Fujisaki, and T. Tada, "Application of artificial intelligence using a convolutional neural network for detecting gastric cancer in endoscopic images," *Gastric Cancer*, vol. 21, no. 4, pp. 653–660, Jul. 2018.

[12] Y. Horie, T. Yoshio, K. Aoyama, S. Yoshimizu, Y. Horiuchi, A. Ishiyama, T. Hirasawa, T. Tsuchida, T. Ozawa, S. Ishihara, Y. Kumagai, M. Fujishiro, I. Maetani, J. Fujisaki, and T. Tada, "Diagnostic outcomes of esophageal cancer by artificial intelligence using convolutional neural networks," *gastrointestinal Endoscopy*, vol. 89, no. 1, pp. 25–32, Jan. 2019.

[13] J. D. Groof, F. van der Sommen, J. van der Putten, M. R. Struyvenberg, S. Zinger, W. L. Curvers, O. Pech, A. Meining, H. Neuhaus, R. Bisschops, E. J. Schoon, P. H. de With, and J. J. Bergman, "The argos project: The development of a computer-aided detection system to improve detection of Barrett's neoplasia on white light endoscopy," *United Eur. Gastroenterol. J.*, vol. 7, no. 4, pp. 538–547, May 2019.

[14] F. van der Sommen, S. R. Klomp, A.-F. Swager, S. Zinger, W. L. Curvers, J. J. G. H. M. Bergman, E. J. Schoon, and P. H. N. de With, "Predictive features for early cancer detection in Barrett's esophagus using volumetric laser endomicroscopy," *Computerized Med. Imag. Graph.*, vol. 67, pp. 9–20, Jul. 2018.

[15] S. Van Riel, F. van der Sommen, S. Zinger, E. J. Schoon, and P. H. N. de With, "Automatic detection of early esophageal cancer with CNNS using transfer learning," in *Proc. 25th IEEE Int. Conf. Image Process. (ICIP)*, Athens, Greece, Oct. 2018, pp. 1383–1387.

[16] X. Du, Y. Li, J. Yao, B. Chen, J. Song, and X. Yang, "LoID-EEC: Localizing and identifying early esophageal cancer based on deep learning in screening chromoendoscopy," in *Proc. 2nd Int. Conf. Video Image Process. (ICVIP)*, Hong Kong, Dec. 2018, pp. 17–22.

[17] V. Badrinarayanan, A. Kendall, and R. Cipolla, "Segnet: A deep convolutional encoder-decoder architecture for image segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 12, pp. 2481–2495, Dec. 2017.

[18] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.*, Munich, Germany, 2016, pp. 234–241.

[19] L. C. Chen, Y. Zhu, G. Papandreou, H. Adam, and F. Schroff, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Munich, Germany, Sep. 2018, pp. 801–818.

[20] R. Roy, T. Chakraborti, and A. S. Chowdhury, "A deep learning-shape driven level set synergism for pulmonary nodule segmentation," *Pattern Recogn. Lett.*, vol. 123, pp. 31–38, May 2019.

[21] P. Wang, X. Xiao, J. R. G. Brown, T. M. Berzin, M. Tu, F. Xiong, X. Hu, P. Liu, Y. Song, D. Zhang, X. Yang, L. Li, J. He, X. Yi, J. Liu, and X. Liu, "Development and validation of a deep-learning algorithm for the detection of polyps during colonoscopy," *Nature Biomed. Eng.*, vol. 2, no. 10, pp. 741–748, Oct. 2018.

[22] A. R. Choudhury, R. Vanguri, S. R. Jambawalikar, and P. Kumar, "Segmentation of brain tumors using DeepLabv3+," in *Proc. Int. MICCAI Brainlesion Workshop*, Granada, Spain, Sep. 2018, pp. 154–167.

[23] D. Liu, N. Rao, X. Mei, H. Jiang, Q. Li, C. Luo, Q. Li, C. Zeng, B. Zeng, and T. Gan, "Annotating early esophageal cancers based on two saliency levels of gastroscopic images," *J. Med. Syst.*, vol. 42, no. 12, p. 237, Dec. 2018.

[24] J. G. Lee, S. Jun, Y. W. Cho, H. Lee, G. B. Kim, J. B. Seo, and N. Kim, "Deep learning in medical imaging: General overview," *Korean J. Radiol.*, vol. 18, no. 4, pp. 570–584, Jul. 2017.

[25] G. Litjens, T. Kooi, B. E. Bejnordi, A. A. A. Setio, F. Ciompi, M. Ghafoorian, J. A. van der Laak, B. van Ginneken, and C. I. Sánchez, "A survey on deep learning in medical image analysis," *Med. Image Anal.*, vol. 42, pp. 60–88, Dec. 2017.

[26] D.-Y. Liu, T. Gan, N.-N. Rao, Y.-W. Xing, J. Zheng, S. Li, C.-S. Luo, Z.-J. Zhou, and Y.-L. Wan, "Identification of lesion images from gastrointestinal endoscope based on feature extraction of combinational methods with and without learning process," *Med. Image Anal.*, vol. 32, pp. 281–294, Aug. 2016.

[27] C. I. Chen, D. Sargent, and Y. F. Wang, "Modeling tumor/polyp/lesion structure in 3D for computer-aided diagnosis in colonoscopy," *Proc. SPIE*, vol. 7625, Feb. 2010, Art. no. 76252F.

[28] P. Mesejo, D. Pizarro, A. Abergel, O. Rouquette, S. Beorchia, L. Poincloux, and A. Bartoli, "Computer-aided classification of gastrointestinal lesions in regular colonoscopy," *IEEE Trans. Med. Imag.*, vol. 35, no. 19, pp. 2051–2063, Sep. 2016.
[29] J. Hu, M. Ozay, Y. Zhang, and T. Okatani, "Revisiting single image depth estimation: Toward higher resolution maps with accurate object boundaries," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Waikoloa Village, HI, USA, Jan. 2019, pp. 1043–1051.
[30] L. R. Dice, "Measures of the amount of ecologic association between species," *Ecology*, vol. 26, no. 3, pp. 297–302, Jul. 1945.
[31] J. Chmelik, R. Jakubicek, P. Walek, J. Jan, P. Ourednicek, L. Lambert, E. Amadori, and G. Gavelli, "Deep convolutional neural network-based segmentation and classification of difficult to define metastatic spinal lesions in 3D CT data," *Med. Image Anal.*, vol. 49, pp. 76–88, Oct. 2018.
[32] N. Silberman, D. Hoiem, P. Kohli, and R. Fergus, "Indoor segmentation and support inference from RGBD images," in *Proc. Eur. Conf. Comput. Vis.*, Florence, Italy, Oct. 2012, pp. 746–760.
[33] M. Everingham, S. M. A. Eslami, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The PASCAL visual object classes challenge: A retrospective," *Int. J. Comput. Vis.*, vol. 111, no. 1, pp. 98–136, Jan. 2015.

DINGYUN LIU was born in Beijing, China, in 1990. He received the bachelor's degree in electronic engineering from the School of electronic engineering, UESTC, and the master's degree in biomedical engineering from the School of Life Science and Technology, UESTC, in 2012, and the combined master's and Ph.D. degree program, in 2014. He is currently pursuing the Ph.D. degree. His research interests include gastrointestinal endoscopic image processing, such as the detection and annotation of early gastric cancer, esophageal cancer, and the abnormal frame detection in wireless capsule endoscopic images, ECG signal processing, such as the detection of atrial fibrillation, bio-informatics of gastric cancer. He has published one patent and more than ten research articles on the above research fields.

HONGXIU JIANG received the bachelor's degree in biomedical engineering from the Southwest University of Science and Technology, in 2017. She is currently pursuing the master's degree from the School of Life Science and Technology, University of Electronic Science and Technology of China, Chengdu, China. Her research interest includes medical image/video processing.

NINI RAO received the B.S. and M.S. degrees in electronic engineering and the Ph.D. degree in biomedical engineering from the University of Electronic Science and Technology of China (UESTC), Chengdu, China, in 1983, 1989, and 2009, respectively. She had been a Visiting Scholar with the University of Georgia and with the Massachusetts General Hospital/Harvard Medical School, from April 2008 to October 2008 and from March 2016 to September 2016, respectively, and a Visiting Professor with the National University of Singapore, from July 2006 to August 2006. She is currently a Professor with the School of Life Science and Technology, UESTC. Her major research interests include biomedical signal and image processing, biomedical pattern recognition, and bioinformatics. She received more than 20 research grants and is the author or coauthor of more than 150 scientific articles. She was honored with the outstanding expert with outstanding contribution to Sichuan province, in 2005, and academic and technical leader in Sichuan province, in 2011, and acquired a third-class prize of progress of science and technology of Sichuan Province, in 2012.

WENJU DU was born in Shandong, China. She received the B.Eng. degree in biomedical engineering, in 2014, and the master's degree from the School of Life Science and Technology, University of Electronic Science and Technology of China (UESTC), Chengdu, China, in 2015. She is currently pursuing the combined master's and Ph.D. program in biomedical engineering. Her research interests include the application of deep learning on gastrointestinal image analysis and fuzzy control, and switched systems.

CHENGSI LUO received the B.Sc. and M.Sc. degrees from the School of Biomedical Engineering, University of Electronic Science and Technology of China, in 2015 and 2018, respectively, where he is currently pursuing the Ph.D. degree. His research interests include biomedical signal processing and deep convolutional neural networks.

ZHENGWEN LI was born in Shandong, China, in 1984. He received the master's degree in applied mathematics from Southwest Jiaotong University, in 2010. He is currently pursuing the Ph.D. degree with the College of Life Sciences, University of Electronic Science and Technology of China.

He worked as a Math Teacher with the Chengdu College of Electronic Science and Technology University, for four years. Since 2016, he has been mainly studied the application of convolutional neural networks in gastroscopic images. He has published several related articles.

LINLIN ZHU received the master's degree in medical in digestion from the West China College, Sichuan, China.

She is currently a Clinician with the West China Hospital, Sichuan University. She has been in charge of four provincial scientific projects. She has published four SCI papers. Her research interests include medical image processing and digestive endoscopy, including gastroscopy, colonsocpy, and wireless capsule endoscopy.

TAO GAN received the master's degree in medical in digestion from the West China College, Sichuan, China, in 2000, and the master's degree in computer science from the University of Electronic Science and Technology of China, Sichuan, in 2006.

He is currently an Associate Professor with the West China Hospital, Sichuan University. He has been in charge of provincial scientific project. He has published more than 20 articles and holds three Chinese patents. His research interests include medical image processing and digestive endoscopy.