ARTICLE

Esophagus Achalasia Diagnosis from Esophagoscopy Based on A Serial Multi-scale Network

Kai Jiang[a], Masahiro Oda[b,a], Yuichiro Hayashi[a], Hironari Shiwaku[c], Masashi Misawa[d] and Kensaku Mori[a,e,f]

[a]Graduate School of Informatics, Nagoya University, Furou-cho, Chikusa-ku, Nagoya, 464-8601, Japan; [b]Information Strategy Office, Information and Communications, Nagoya University, Nagoya, Japan; [c]Department of Gastroenterological Surgery, Fukuoka University Faculty of Medicine, Fukuoka, Japan; [d]Digestive Disease Center, Showa University Northern Yokohama Hospital, Yokohama, Japan; [e]Information Technology Center, Nagoya University, Nagoya, Japan; [f]Research Center for Medical Bigdata, National Institute of Informatics, Tokyo, Japan

ARTICLE HISTORY
Compiled September 14, 2022

ABSTRACT
Esophageal achalasia is a primary esophageal motility disorder disease. To diagnose esophagus achalasia, physicians recommend endoscopic evaluation of the esophagus. However, a low sensitivity still accompanies esophagoscopy on esophagus achalasia diagnosis. Especially for early-stage achalasia, only less than half of patients can be correctly identified from esophagoscopy. Thus, a quantitative diagnosis system is needed to support physicians diagnose achalasia from the esophagoscopy video. This paper proposes a Serial Multi-scale Network for classifying achalasia images from the esophagoscopy video. The proposed method contains two main components, a Dense-pooling Net, and a Serial Multi-scale Dilated encoder. We construct the Dense-pooling Net using a convolution neural network with dense mixed-pooling connections to extract features from esophagoscopy images. We design the Serial Multi-scale Dilated encoder based on a dilated encoder composed of four residual-style dilated convolution blocks. We combine the dilated encoder and spatial attention modules to focus on extracting features we need from esophagoscopy images. We trained and evaluated our method with a dataset that was extracted from several esophagoscopy videos of achalasia patients. The evaluation results reveal a state-of-the-art accuracy of achalasia diagnosis. Furthermore, we developed a real-time computer-aided achalasia diagnosis system with the trained network. In the real-time test, the achalasia diagnosis system can stably output the diagnosis results in only 0.138 seconds. The extended experiments demonstrate that the constructed diagnosis system can diagnose achalasia from esophagoscopy videos.

KEYWORDS
Esophageal achalasia; esophagoscopy; computer-aided diagnosis; deep learning

1. Introduction

Esophagus Achalasia (achalasia) (Gennaro et al. 2011) is a chronic gastrointestinal disease. A standard definition of achalasia is the inability of the lower esophageal

CONTACT Kai Jiang Email: kjiang@mori.m.is.nagoya-u.ac.jp

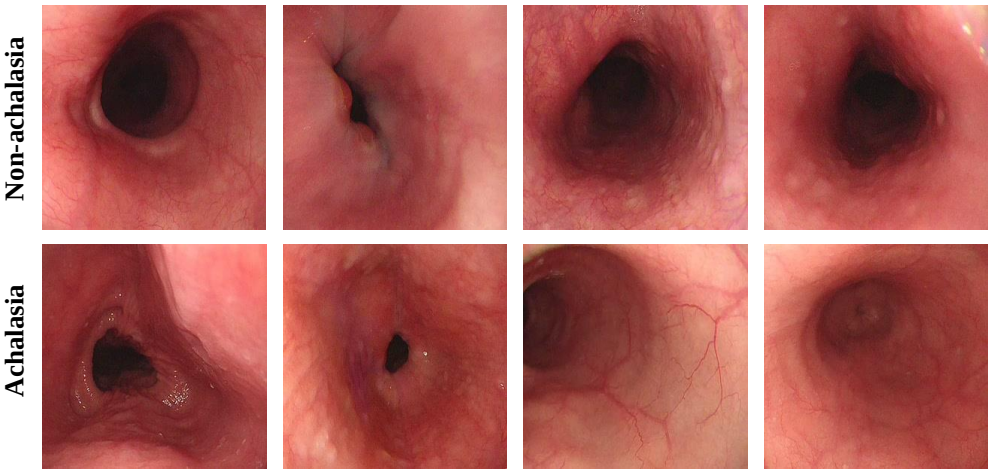

Figure 1. Examples of achalasia and non-achalasia in esophagoscopy images. The first row is esophagoscopy images collected from non-achalasia patients. The second row is esophagoscopy images collected from achalasia patients. Physicians can diagnose achalasia patients through images in the second row.

sphincter to relax without peristalsis (Boeckxstaens et al. 2014). The annual incidence of achalasia is approximately 1 in 100,000 people worldwide, with an overall prevalence of 9 to 10 in 100,000 people (Patel et al. 2017). Regardless of the stage at which achalasia is diagnosed, the treatment of it is the same as Peroral Endoscopic Myotomy (POEM) (Inoue et al. 2010). Thus, early diagnosis can not reduce the cost of the treatment. However, diagnosing achalasia earlier is very meaningful, achalasia carries a risk of complications, including aspiration pneumonia and oesophageal cancer (Torres-Aguilera and Troche 2018). Early diagnosis of achalasia can prevent esophageal cancer occurrence, and reduce the risk of POEM complications. About 65% to 90% of patients can be effectively treated with pneumatic dilation, Heller esophagotomy, or POEM once it be correctly diagnosised (Reynolds and Parkman 1989; Rakita et al. 2005; Barbieri et al. 2015). However, the etiology characterized by achalasia remains unknown, which causes physicians can not precisely identify achalasia (Vaezi et al. 1999). The general diagnosis methods of achalasia are through esophagus endoscopy (esophagoscopy), radiology, and manometry (Pohl and Tutuian 2007). Esophagoscopy is a necessary achalasia diagnosis method, which can rule out esophageal squamous cell carcinoma complicated with achalasia or secondary achalasia associated with malignancy (Pohl and Tutuian 2007).

The clinical manifestations of achalasia are similar to mechanical obstruction or inflammatory process. Thus, physicians can not diagnose achalasia with high accuracy through esophagoscopy, radiology, or manometry. Generally, as achalasia progresses, the esophagus dilates and eventually curves. The development of achalasia is accompanied by three stages, the straight type, the sigmoid type, and the advanced sigmoid type (Society 2017). Early-stage achalasia often refers to the straight type achalasia which has poor esophageal dilation and is difficult to detect by examination. Esophagoscopy and radiology diagnosis are especially difficult in dealing with patients of early-stage achalasia. Only about half or even fewer early-stage achalasia patients can be diagnosed precisely (Pohl and Tutuian 2007). Figure 1 shows esophagoscopy images of achalasia and non-achalasia. Physicians diagnose achalasia by observing esophageal body contraction or esophageal lumen dilation (Shiwaku et al.

2018). However, esophageal contraction or dilation is not very conspicuous in early-stage achalasia. Physicians may not accurately diagnose early-stage achalasia with inconspicuous contraction or dilation; it is common for the correct diagnosis to be delayed by 2 or 3 years from the onset of symptoms (Pohl and Tutuian 2007). Therefore, there is a demand for a videos-based Computer-Aided Diagnosis (CAD) system to support physicians to identify early-stage achalasia by classifying each frame into achalasia or non-achalasia image. We believe fast processing and accurate diagnosis are essential for the esophagoscopy videos-based CAD system. This paper proposes a method for constructing a CAD system that can diagnose achalasia fast and accurately from esophagoscopy videos.

Deep learning has been widely used in computer-aided diagnosis through medical images in recent years. Many deep learning methods, e.g., ConvMixer and Supervised Contrastive (Khosla et al. 2020; Trockman and Kolter 2022) have been proposed for cancer and bleeding diagnosis. However, achalasia does not have distinctive lesions, unlike cancer or bleeding. Physicians distinguish achalasia from esophagoscopy images by observing abnormal contraction and dilation of the esophageal body and lumen, respectively (Shiwaku et al. 2018). Mucosal thickening, liquid or food remnant, and whitish change or pinstripe pattern are also helpful in achalasia diagnosis (Shiwaku et al. 2018). Thus, a method that can capture multi-type and multi-scale features is necessary for achalasia diagnosis. Methods for cancer and bleeding diagnosis are designed for detecting typical lesions, which locate in part of esophagoscopy images. Those methods may not capture multi-type and multi-scale features which observe in the entire esophagoscopy images, leading to the wrong diagnosis.

In this work, we propose an automated classification method named Serial Multi-scale Network (SMN) for achalasia diagnosis from esophagoscopy videos. The proposed method contains two main components: a Dense-pooling Net and a Serial Multi-scale Dilated (SMD) encoder. We use the Dense-pooling Net, which is a Convolution Neural Network (CNN) with dense mixed-pooling connections (Playout et al. 2018) to extract feature maps from an esophagoscopy frame. The Dense-pooling Net aims to preserve the spatial resolution of features and more details of the esophageal. Since achalasia diagnosis requires multi-scale features detection, we propose a SMD encoder to generate features with multiple receptive fields outside the Dense-pooling Net. With this characteristic, the proposed method can classify esophagoscopy images with different scales of features. We train the proposed network with a private dataset that extracts from esophagoscopy videos collected from achalasia and non-achalasia patients. We quantitatively compared the proposed method and state-of-the-art image classification methods on our dataset. We experimentally test the diagnosis accuracy with 50 esophagoscopy videos. Furthermore, we build a CAD system using the proposed method for real-time processing from esophagoscopy videos. The CAD system has been used in clinical experiments. However, we can not provide the results of the experiments due to we have not received permission to publish clinical results.

In short, our contributions can be summarized three-fold:

(1) We propose an architecture for achalasia classification from esophagoscopy videos. We propose a Dense-pooling Net and a SMD encoder to extract different textures and scales of features from esophagoscopy images. The proposed method achieved state-of-the-art performance of achalasia diagnosis through both images and videos.
(2) We collected an image and a video achalasia datasets from several esophagoscopy videos for validating the proposed method. We implemented and compared mod-

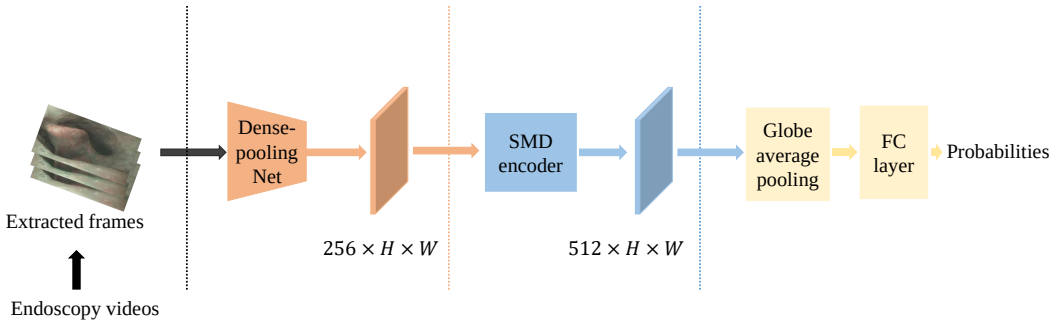

Figure 2. The sketch of the SMN, which consists of three parts: the Dense-pooling Net, the SMD encoder, and the classification part. The inputs of the Dense-pooling Net are esophagoscopy images. The SMD encoder receives the output feature maps from the Dense-pooling Net. Finally, the classification part performs classification results using the SMD encoder's output. In this figure $H \times W$ represents the height and width of the output feature maps.

ern data augmentation methods on the image dataset for achalasia classification.

(3) We construct a CAD system with the proposed method using the NVIDIA Jetson Xavier NX Developer Kit. The experiment reveals that the constructed CAD system can process esophagoscopy video in real-time.

2. Method

2.1. Overview

We propose a method called Serial Multi-scale Network for classifying esophagoscopy images. Figure 2 shows the illustration of the SMN. The proposed SMN consists of a Dense-pooling Net, a SMD encoder, and a classification part. The classification part comprises a global average pooling layer and a fully connected layer, which receives the output of the SMD encoder for calculating classification probabilities. The inputs of the SMN are esophagoscopy images extracted from esophagoscopy videos. The outputs of the SMN are two probabilities which stand for achalasia and non-achalasia of one input image, respectively.

2.2. Dense-pooling Net

We propose a CNN with dense-pooling connections, which is called Dense-pooling Net to get multi-type and multi-scale features from the input images. Figure 3 shows the architecture of the proposed Dense-pooling Net. In the proposed network, we used four serial connected residual block (He et al. 2016) as the backbone. We use dense pooling connection (Playout et al. 2018) based on the multi-scale spatial information in the network and the bottleneck layer for feature extraction. As shown in Fig 3, dense pooling connections connect four residual blocks with different filter sizes. We further use mixed pooling (Playout et al. 2018) instead of max-pooling or average-pooling to keep the spatial information in the dense pooling connections. The proposed CNN can capture features with less spatial information loss by using dense pooling connections and mixed pooling. The residual style in the proposed CNN can prevent overfitting in the training procedure. We believe the network can extract esophageal features with

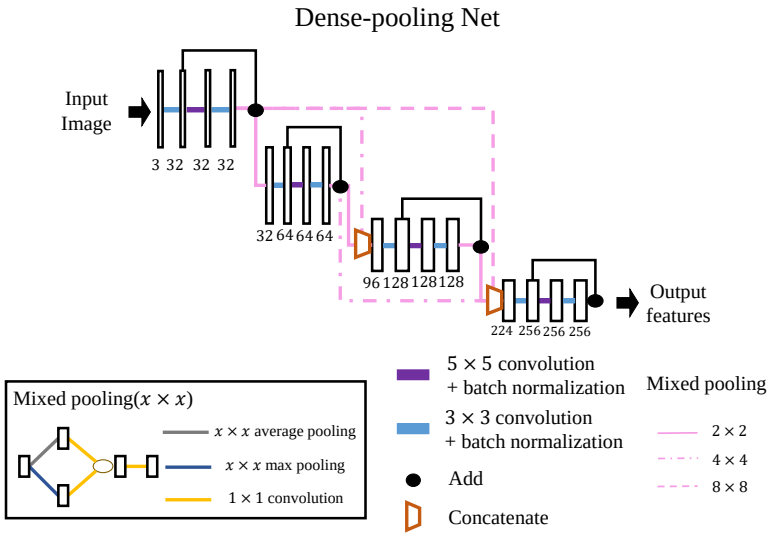

Figure 3. The architecture of the proposed Dense-pooling Net. In this figure, white boxes are feature maps or input images. The numbers below boxes are numbers of kernel or color channels. Dense pooling connections are represented as pink connections, which are implemented as a combination of mixed pooling. We describe the architecture of a mixed pooling connection on the bottom left, where $x \times x$ represents the size of the filter used in a mixed pooling.

residual style, dense pooling connection, and mixed-pooling. The Dense-pooling Net also preserves color and pattern details of mucosal in the feature maps. The resized esophagoscopy image is directly input into the Dense-pooling Net, and the output of this network is a 256 channel feature map.

2.3. SMD Encoder

We propose a SMD encoder to distribute representations for detecting multi-type and multi-scale features from the feature map. We design the SMD encoder based on a dilated encoder (Lin et al. 2017; Chen et al. 2021) that extract features for object detection and localization in the YOLOF. The dilated encoder consists of a Projector and four residual dilated blocks. The Projector, which is designed for channel dimension reduction, has the same structure in the FPN (Lin et al. 2016). As for the residual dilation block (Yu and Koltun 2015) with different dilation rates, it generates output features with multiple receptive fields in 3×3 convolutional layers, covering many object scales. The dilated encoder design enables it to detect objects on multiple-level instead of single-level features. The residual dilated block utilizes dilated convolution to increase the receptive field of input features. The YOLOF uses the residual style to ensure the encoder can get a multi-scale receptive field. Experiments have proved that the dilated encoder can detect multi-scale features from feature maps (Chen et al. 2021). However, for achalasia diagnosis, the network requires to detect inconspicuous features and pinstripe patterns from the background, which may miss by the dilated encoder. To solve this problem, we propose the SMD encoder.

Figure 4 illustrates the structure of the proposed SMD encoder. We enlarge kernels in the Projector from 1×1 and 3×3 to 3×3 and 5×5, respectively. Many researchers

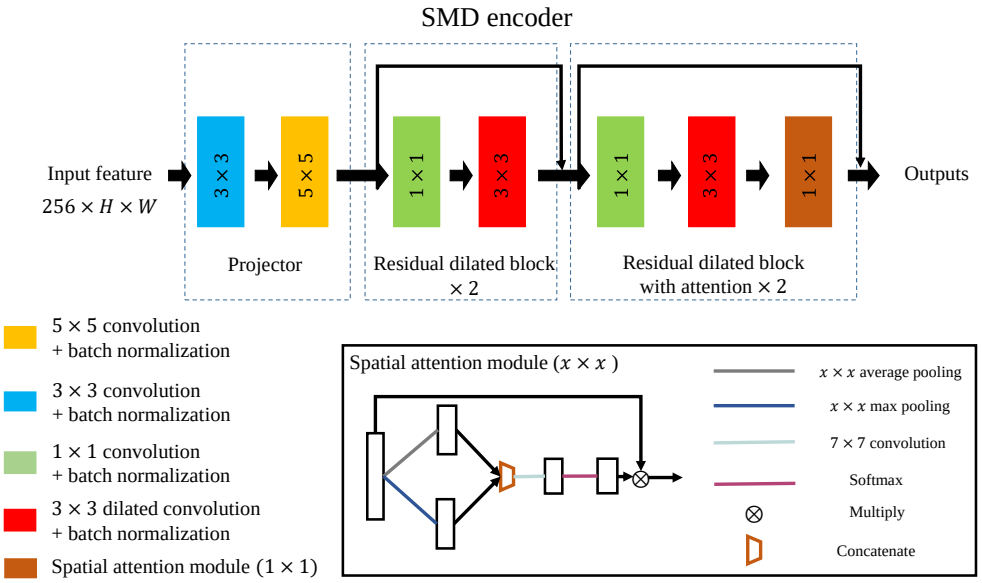

Figure 4. The structure of the SMD encoder. In this figure, white boxes are feature maps. The inputs of the SMD encoder are feature maps output from the Dense-pooling Net. In this figure, 1×1, 3×3 and 5×5 represents the filter size of the corresponding convolution layer. Block$\times 2$ stands for two same successive blocks. We describe the architecture of a spatial attention module on the bottom right. In the spatial attention module, $x \times x$ denotes $x \times x$ pooling. Furthermore, a batch normalization layer (Ioffe and Szegedy 2015) and a ReLU layer (Nair and Hinton 2010) are introduced after all convolution layers.

have demonstrated that a few convolution layers with large kernels have a better effective receptive field (Ding et al. 2022; Peng et al. 2017). We believe a large receptive field helps detect contraction or dilation of the esophageal. Besides, when the view of the esophagoscopy is tiny, large kernels can extract more features of the mucosal. Four residual dilated blocks serially connect with the Projector. We modified all residual dilated blocks by removing the last convolution layers. We set the dilation rates from the first to the last residual dilated blocks to $0, 2, 4, 8$ in order. Then, we introduce the spatial attention module (Woo et al. 2018; Vaswani et al. 2017) which helps the encoder focus on the meaningful features in the feature map. It makes the encoder distinguish the difference between inconspicuous features such as whitish change or pinstripe pattern from the normal mucosal. We add spatial attention modules in the last two residual dilated blocks. The input of the SMD encoder should be the feature map output from the Dense-pooling Net. The output of the SMD encoder is a 512 channel feature map.

3. Experiments And Results

3.1. Dataset

3.1.1. Image dataset

We collected esophagoscopy videos from patients in the Fukuoka University Faculty of Medicine with IRB approval for network training and test. In our dataset, all achalasia images and videos are collect from patients with straight type achalasia which con-

Table 1. Numbers of different types of images in training, validation, and test data in the image dataset.

Dataset	Train		Validation		Test	
Type	WLI	NBI	WLI	NBI	WLI	NBI
Number of non-achalasia images	122,600	67,167	25,824	13,456	35,863	21,499
Number of achalasia images	56,656	83,302	14,868	20,307	12,697	18,841

tains early-stage achalasia patients. Expert endoscopist annotated all achalasia frames in collected videos under the same standard, but not specifically labeled early-stage achalasia frames for the experiment. Based on these annotations, we extracted achalasia and non-achalasia frames from esophagoscopy videos by 30 fps. Images in our dataset contain images of Narrow Band Imaging (NBI) (Kuznetsov et al. 2006) and White Light Imaging (WLI) (Cummins et al. 2019). We resized all extracted images to 224×224 pixel with the Lanczos interpolation method (Fadnavis 2014). We manually removed all images with strong specular and serious blurred for a massive training set. We split all resized images into training, validation, and test datasets randomly without patient duplication. Table 1 shows the number of extracted WLI and NBI images in training, validation, and test datasets.

3.1.2. Video dataset

To evaluate whether the proposed method can apply to the clinical situation, we collected 50 esophagoscopy videos from different patients in the Fukuoka University Faculty of Medicine with IRB approval. All videos in the video dataset are collected from different patients in the image dataset. Expert physician annotated the class each video belongs to instead of annotating all achalasia frames in videos under the same standard of annotation in the image dataset. Among all videos, 25 videos have been annotated as achalasia videos; others are annotated as non-achalasia videos.

3.2. Implementation Details

For the training process, we set the minibatch size to be 64 to train our method for 200 epochs on NVIDIA Tesla V100 PCIe 32 GB with CUDA 10.0. We used binary-crossentropy as the loss function and an Adam as the optimizer function. The initial learning rate for training was set to be 1.0×10^{-3}. We implemented our method with Keras (Gulli and Pal 2017). To evaluate our method, we trained our method and other state-of-the-art methods using our image dataset in the same condition. For all training images, we implement resize, ZCA whitening (Kessy et al. 2018) for all images preprocess. We implement random flip horizontally, and random flip vertically to train all methods. Furthermore, we implement cutout (DeVries and Taylor 2017) for data augmentation for training the proposed method with our image dataset. We completed the CAD system with an NVIDIA Jetson Xavier NX developer kit which carries a trained SMN.

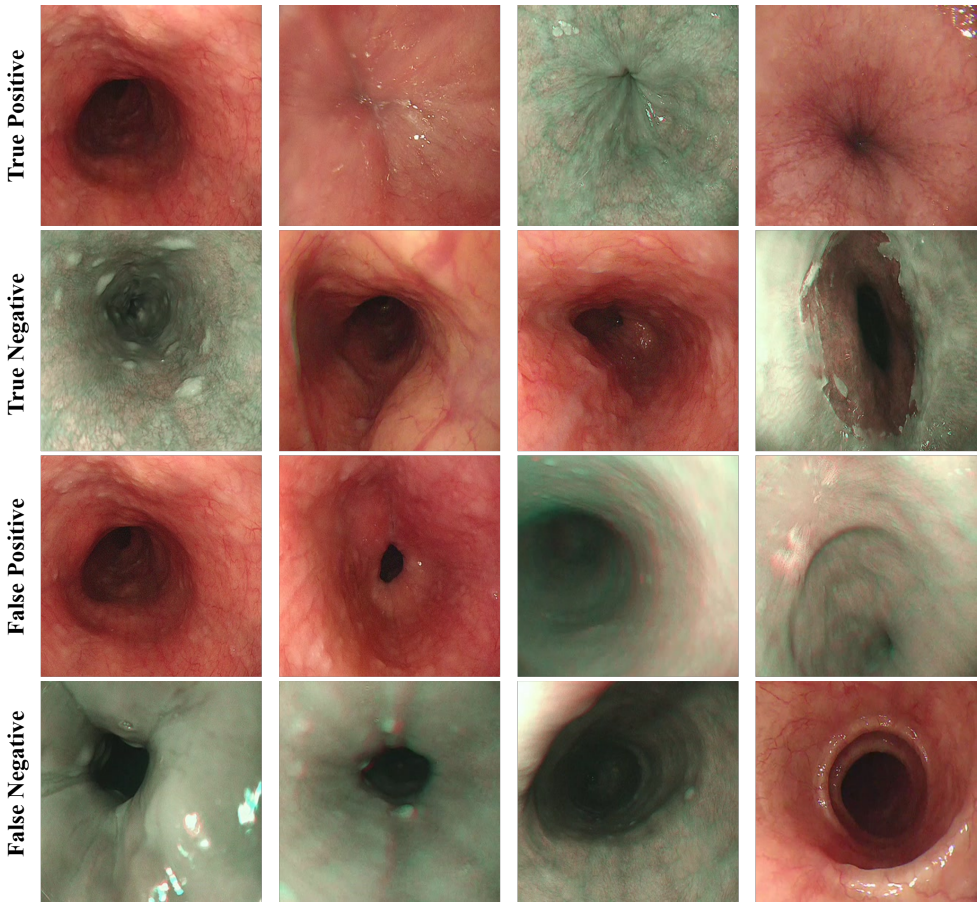

Figure 5. Examples of the test dataset that the SMN classifies. From the first row to the last row are true positive, true negative, false positive, and false negative images in order.

3.3. Results

3.3.1. Quantitative evaluation on image dataset

We introduced accuracy, precision, recall, specificity, and Area Under the Curve (AUC) of Receiver Operating Characteristic (ROC) to evaluate the classification accuracy of all trained models for evaluating their performance. We defined an image as a predicted positive image when the predicted achalasia probability of this image is greater than the threshold τ_p. On the contrary, an image was defined as a predicted negative image when the predicted achalasia probability is lower than or equal to the threshold. We defined a true positive/negative sample when an achalasia/non-achalasia image was correctly classified as a predicted positive/negative image. When an achalasia/non-achalasia image was not correctly classified, we defined a false positive/negative sample. For compare with other classification methods, we set $\tau_p = 0.5$ which is the common setting for binary image classification. Table 2 shows the quantitative evaluation results of all trained methods using our image dataset. Figure 5 shows examples that our method classified success and failure in the image test dataset.

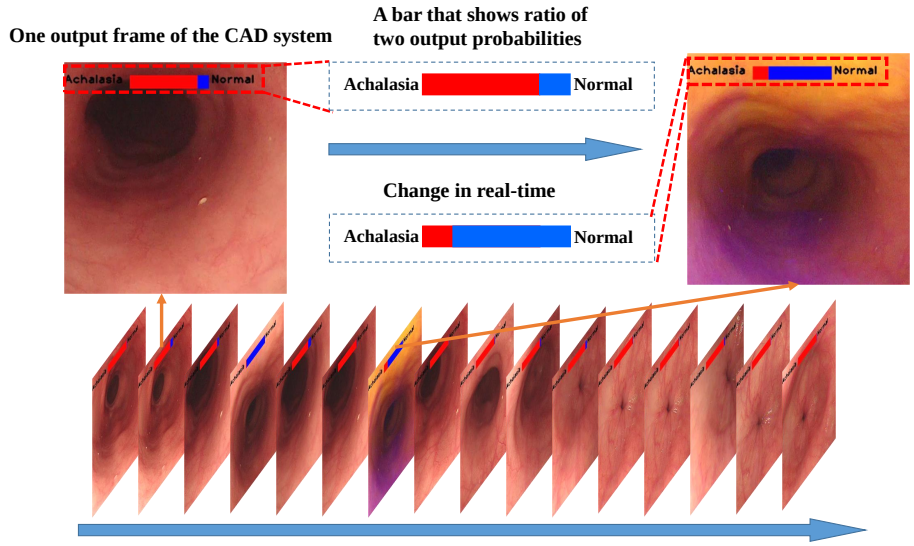

Figure 6. The output of the CAD system while the input is one video in the video dataset. In one output frame, a bar shows the ratio of achalasia and non-achalasia probabilities in the red and blue parts, respectively, in real-time.

Table 2. Quantitative evaluations in different methods on the image dataset.

Method	Accuracy	Precision	Recall	Specificity	AUC score
ResNet50 (He et al. 2016)	0.684	0.419	0.783	0.419	0.807
DenseNet121 (Huang et al. 2016)	0.691	0.423	0.671	0.521	0.799
U-Net Contracting path (Ronneberger et al. 2015)	0.669	0.529	0.861	0.894	0.899
ConvMixer (Trockman and Kolter 2022)	0.683	0.298	0.711	0.632	0.623
Supervised Contrastive (Khosla et al. 2020)	0.557	0.375	**0.891**	0.310	0.699
Gated-Attention (Ilse et al. 2018)	0.654	0.287	0.832	0.299	0.714
Our method (SMN + cutout)	**0.892**	**0.863**	0.826	**0.897**	**0.951**

3.3.2. Quantitative evaluation on video dataset

We used a SMN trained with cutout to diagnose all videos in the video dataset. The trained SMN classified every frame in one video. We set the threshold $\tau_p = 0.5$ for frame classification in the dataset. We introduce another threshold τ_f for video classification. When the proportion of predicted achalasia frames among all frames in a video is greater than the τ_f, we predicted the video as an achalasia video. We used accuracy, precision, recall, and specificity to evaluate the performance of the SMN on our video dataset. Table 3 shows the diagnosis results when we set different τ_f. Table 3 illustrates that the SMN performs well in achalasia diagnosis when the threshold τ_f is set to be 0.95.

Furthermore, to prove whether our method can be applied to the clinic, we test our CAD system's capture and process speed by connecting with an endoscopy instrument. The experiment results demonstrate that our CAD system can stably output diagnosis results in only 0.138 ± 0.04 seconds. Figure 6 shows one example in the video dataset diagnosed using the CAD system. However, diagnosis of achalasia videos by calculating the ratio of achalasia frames is not a stable method in the clinical. The CAD system temporarily uses a threshold $\tau_f = 0.95$ diagnosis achalasia videos in

Table 3. Performance of video classification in different τ_f by using the SMN.

τ_f	0.05	0.15	0.25	0.35	0.45	0.55	0.65	0.75	0.85	0.95
Accuracy	0.50	0.50	0.50	0.50	0.60	0.66	0.70	0.78	0.88	0.92
Precision	0.5	0.5	0.5	0.5	0.5	0.55	0.62	0.69	0.80	0.89
Recall	1	1	1	1	1	1	1	1	1	0.96
Specificity	0	0	0	0	0	0.2	0.4	0.56	0.85	0.88

Table 4. Performance comparison with different networks instead of Denes-pooling Net in the SMN.

Network	Accuracy	Precision	Recall	Specificity	AUC score
ResNet50 (He et al. 2016)	0.379	0.373	**0.991**	0.229	0.721
DenseNet121 (Huang et al. 2016)	0.690	0.456	0.943	0.891	0.890
U-Net (Ronneberger et al. 2015)	0.820	0.771	0.732	0.889	0.895
Denes-pooling Net w/o dense pooling	0.829	0.709	0.880	0.883	0.923
Denes-pooling Net	**0.874**	**0.865**	0.766	**0.899**	**0.947**

Table 5. Performance comparison with difference encoders in the SMN.

Encoder	Accuracy	Precision	Recall	Specificity	AUC score
Dilated encoder (Yu and Koltun 2015)	0.737	0.615	0.774	0.893	0.846
SMD encoder w/o attention module	0.871	**0.891**	0.721	0.876	0.946
SMD encoder	**0.874**	0.865	**0.766**	**0.899**	**0.947**

clinical experiments. We will evaluate the performance of CAD system's video classification when we select the threshold $\tau_f = 0.95$ in clinical experiments. Besides, the CAD system has another mode, which only classifies each video frame, then displays the classification result of the frame on the monitor to support physicians in deciding on achalasia diagnosis.

3.4. Ablation study

3.4.1. Effectiveness of Dense-pooling Net in the SMN

We compared the performances when we used different networks instead Dense-pooling Net in the proposed architecture. We used the ResNet-50, DenseNet, and encoder part of U-Net instead of Dense-pooling Net in the proposed method. We also used Dense-pooling Net without dense pooling in the SMN as a comparison to investigate if the dense pooling structure is helpful. Table 4 reports the quantitative evaluation results of the SMN using different networks instead of the Dense-pooling Net. Table 4 demonstrated that the SMN using the Dense-pooling Net performs best among all methods.

3.4.2. Effectiveness of SMD encoder in the SMN

To validate all improvements we made for the SMD encoder have a good influence. We used the original dilated encoder in the YOLOF, and SMD encoder without different improvements as the encoder part in the SMN. Table 5 compares the performances of SMNs using different encoders. Table 5 illustrated that all modified in the SMD encoder have a good influence on achalasia frame classification.

Table 6. Comparison of different data augmentation methods for training the SMN.

Method	Accuracy	Precision	Recall	Specificity	AUC score
SMN	0.874	**0.865**	0.766	**0.899**	0.947
SMN + cutmix	0.159	0.135	0.254	0.775	0.088
SMN + mixup	0.261	0.283	0.702	0.495	0.195
SMN + cutout	**0.892**	0.863	**0.826**	0.897	**0.951**

Table 7. Ablation study of different components in the SMN.

Method	Accuracy	Precision	Recall	Specificity	AUC score
Denes-pooling Net	0.803	0.834	0.584	0.877	0.892
SMD encoder	0.684	0.668	0.287	0.859	0.746
SMN	**0.874**	**0.865**	**0.766**	**0.899**	**0.947**

3.4.3. Effectiveness of cutout data augmentation

We investigated the effectiveness of different modern data augmentation methods on achalasia and non-achalasia esophagoscopy images. We experimentally compared the performance of different data augmentation methods such as cutmix (Yun et al. 2019), mixup (Zhang et al. 2017), and cutout (DeVries and Taylor 2017) on the image dataset for training the proposed SMN. Table 6 compares the performance of different data augmentation methods on the image dataset. Table 6 illustrates that cutout can large improve the accuracy, recall and AUC of the SMN with almost the same precision and specificity.

3.4.4. Comparison with different component in the SMN

We investigated the influence of two-component in SMN, which is proposed in this paper. Table 7 compares the performance of different components in the SMN on our test dataset. Table 7 illustrates that all components in the SMN provide a good influence on the achalasia image classification.

4. Discussion

Table 2 reports that our method produces the highest accuracy, precision, specificity, and AUC score, which illustrates that our method has the best ability to diagnose achalasia from still images among all methods evaluated in our experiments. Table 3 illustrates that the SMN is very sensitive to achalasia frames: the SMN classified nearly all achalasia frames, but misclassified many non-achalasia frames into achalasia frames. By selecting a high threshold τ_f, the SMN can provide high accuracy on esophagoscopy video diagnosis. This characteristic shows that the proposed method has the potential to provide high accuracy in the clinical.

Table 2 shows that the diagnosis recall of our method is lower than the Supervised Contrastive method. However, the Supervised Contrastive method has low specificity, which illustrates that this method can not classify true negative images precisely. Table 4 shows the SMN using ResNet50 instead the Dense-pooling Net also provides higher recall than our method, but with low precision and specificity. Physicians believe precision, recall, and specificity are equally crucial in disease diagnosis (Akobeng 2007). Table 2 illustrates the proposed method has both high precision, recall, and specificity. The proposed method provided the highest AUC score showing it is the most suitable method for achalasia diagnosis among all methods evaluated in this pa-

per, which are automatically done by computers. Another problem with the method is that the proposed method diagnoses achalasia by calculating the ratio of achalasia frames in a limited time. Inexperienced endoscopists may not be able to provide stable esophagoscopy videos for diagnosis. An esophagoscopy video with many noise frames may not be diagnosed by calculating the ratio of achalasia frames.

5. Conclusion and Future work

This paper proposed an automated achalasia diagnosis method SMN for achalasia diagnosis assistance. This proposed SMN utilizes a Dense-pooling Net to extract features from esophagoscopy frames and a SMD encoder for distribute representations to detect inconspicuous features. We collected an esophagoscopy image dataset and an esophagoscopy video dataset for model training and testing. We further constructed a CAD system using the proposed method. The proposed method achieved the best performance on the image test dataset among all methods in Table 2. The evaluation results on the video dataset demonstrate that the proposed method can diagnose achalasia from esophagoscopy video with high accuracy. For future work, we aim to propose a more robust method for achalasia diagnosis from classified frames. We plan to annotate all early-stage achalasia patients in our dataset. By extracting frames and videos from early-stage achalasia patients in our dataset, we can validate the performance of early-stage achalasia diagnosis with the proposed method. In order to perform better annotation and more justified verification results, we will ask more physicians to annotate our dataset and take average annotations among all physicians in the future. Furthermore, we plan to add a decoder to replace the global average pooling in the SMN.

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

Parts of this research was supported by the JST CREST (JPMJCR20D5), the MEXT&JSPS KAKENHI (26108006, 17H00867, 17K20099), the JSPS Bilateral International Collaboration Grants, the AMED (19hs0110006h0003) and the Hori Sciences & Arts Foundation.

References

Akobeng AK. 2007. Understanding diagnostic tests 1: sensitivity, specificity and predictive values. Acta paediatrica. 96(3):338–341.

Barbieri LA, Hassan C, Rosati R, Romario UF, Correale L, Repici A. 2015. Systematic review and meta-analysis: efficacy and safety of poem for achalasia. United European gastroenterology journal. 3(4):325–334.

Boeckxstaens GE, Zaninotto G, Richter JE. 2014. Achalasia. The Lancet. 383(9911):83–93.

Chen Q, Wang Y, Yang T, Zhang X, Cheng J, Sun J. 2021. You only look one-level feature. CoRR. abs/2103.09460. Available from: https://arxiv.org/abs/2103.09460.

Cummins G, Cox BF, Ciuti G, Anbarasan T, Desmulliez MP, Cochran S, Steele R, Plevris JN, Koulaouzidis A. 2019. Gastrointestinal diagnosis using non-white light imaging capsule endoscopy. Nature Reviews Gastroenterology & Hepatology. 16(7):429–447.

DeVries T, Taylor GW. 2017. Improved regularization of convolutional neural networks with cutout. arXiv preprint arXiv:170804552.

Ding X, Zhang X, Han J, Ding G. 2022. Scaling up your kernels to 31x31: Revisiting large kernel design in cnns. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. p. 11963–11975.

Fadnavis S. 2014. Image interpolation techniques in digital image processing: an overview. International Journal of Engineering Research and Applications. 4(10):70–73.

Gennaro N, Portale G, Gallo C, Rocchietto S, Caruso V, Costantini M, Salvador R, Ruol A, Zaninotto G. 2011. Esophageal achalasia in the veneto region: epidemiology and treatment. Journal of Gastrointestinal Surgery. 15(3):423–428.

Gulli A, Pal S. 2017. Deep learning with keras. Packt Publishing Ltd.

He K, Zhang X, Ren S, Sun J. 2016. Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. p. 770–778.

Huang G, Liu Z, Weinberger KQ. 2016. Densely connected convolutional networks. CoRR. abs/1608.06993. Available from: http://arxiv.org/abs/1608.06993.

Ilse M, Tomczak JM, Welling M. 2018. Attention-based deep multiple instance learning. CoRR. abs/1802.04712. Available from: http://arxiv.org/abs/1802.04712.

Inoue H, Minami H, Kobayashi Y, Sato Y, Kaga M, Suzuki M, Satodate H, Odaka N, Itoh H, Kudo S. 2010. Peroral endoscopic myotomy (poem) for esophageal achalasia. Endoscopy. 42(04):265–271.

Ioffe S, Szegedy C. 2015. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In: International conference on machine learning. PMLR. p. 448–456.

Kessy A, Lewin A, Strimmer K. 2018. Optimal whitening and decorrelation. The American Statistician. 72(4):309–314.

Khosla P, Teterwak P, Wang C, Sarna A, Tian Y, Isola P, Maschinot A, Liu C, Krishnan D. 2020. Supervised contrastive learning. Advances in Neural Information Processing Systems. 33:18661–18673.

Kuznetsov K, Lambert R, Rey JF. 2006. Narrow-band imaging: potential and limitations. Endoscopy. 38(01):76–81.

Lin T, Dollár P, Girshick RB, He K, Hariharan B, Belongie SJ. 2016. Feature pyramid networks for object detection. CoRR. abs/1612.03144. Available from: http://arxiv.org/abs/1612.03144.

Lin T, Goyal P, Girshick RB, He K, Dollár P. 2017. Focal loss for dense object detection. CoRR. abs/1708.02002. Available from: http://arxiv.org/abs/1708.02002.

Nair V, Hinton GE. 2010. Rectified linear units improve restricted boltzmann machines. In: Icml.

Patel DA, Lappas BM, Vaezi MF. 2017. An overview of achalasia and its subtypes. Gastroenterology & hepatology. 13(7):411.

Peng C, Zhang X, Yu G, Luo G, Sun J. 2017. Large kernel matters–improve semantic segmentation by global convolutional network. In: Proceedings of the IEEE conference on computer vision and pattern recognition. p. 4353–4361.

Playout C, Duval R, Cheriet F. 2018. A multitask learning architecture for simultaneous segmentation of bright and red lesions in fundus images. vol. 11071. Springer. p. 101–108.

Pohl D, Tutuian R. 2007. Achalasia: an overview of diagnosis and treatment. Journal of Gastrointestinal and liver diseases. 16(3):297.

Rakita S, Bloomston M, Villadolid D, Thometz D, Zervos E, Rosemurgy A. 2005. Esophagotomy during laparoscopic heller myotomy cannot be predicted by preoperative therapies and does not influence long-term outcome. Journal of gastrointestinal surgery. 9(2):159–164.

Reynolds JC, Parkman HP. 1989. Achalasia. Gastroenterology Clinics of North America. 18(2):223–255.

Ronneberger O, Fischer P, Brox T. 2015. U-Net: Convolutional networks for biomedical image

segmentation. vol. 9351. Springer. p. 234–241.
Shiwaku H, Yamashita K, Ohmiya T, Nimura S, Shiwaku Y, Inoue H, Hasegawa S. 2018. New endoscopic finding of esophageal achalasia with st hood short type: Corona appearance. Plos one. 13(7):e0199955.
Society JE. 2017. Descriptive rules for achalasia of the esophagus, june 2012. Esophagus. 14(4):275–289.
Torres-Aguilera M, Troche JMR. 2018. Achalasia and esophageal cancer: risks and links. Clinical and experimental gastroenterology. 11:309.
Trockman A, Kolter JZ. 2022. Patches are all you need? arXiv preprint arXiv:220109792.
Vaezi MF, Richter JE, of Gastroenterology Practice Parameter Committee AC. 1999. Diagnosis and management of achalasia. Official journal of the American College of Gastroenterology—ACG. 94(12):3406–3412.
Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser Ł, Polosukhin I. 2017. Attention is all you need. Advances in neural information processing systems. 30.
Woo S, Park J, Lee J, Kweon IS. 2018. CBAM: convolutional block attention module. CoRR. abs/1807.06521. Available from: http://arxiv.org/abs/1807.06521.
Yu F, Koltun V. 2015. Multi-scale context aggregation by dilated convolutions. arXiv preprint arXiv:151107122.
Yun S, Han D, Oh SJ, Chun S, Choe J, Yoo Y. 2019. Cutmix: Regularization strategy to train strong classifiers with localizable features. In: Proceedings of the IEEE/CVF international conference on computer vision. p. 6023–6032.
Zhang H, Cisse M, Dauphin YN, Lopez-Paz D. 2017. mixup: Beyond empirical risk minimization. arXiv preprint arXiv:171009412.