sensors

MDPI

Article

Multi-Task Model for Esophageal Lesion Analysis Using Endoscopic Images: Classification with Image Retrieval and Segmentation with Attention

Xiaoyuan Yu [1], Suigu Tang [1], Chak Fong Cheang [1,*], Hon Ho Yu [2,*] and I Cheong Choi [2]

[1] Faculty of Information Technology, Macau University of Science and Technology, Taipa, Macau; 1709853eii30001@student.must.edu.mo (X.Y.); 2009853gii30001@student.must.edu.mo (S.T.)
[2] Kiang Wu Hospital, Santo António, Macau; alexchoi0@gmail.com
* Correspondence: cfcheang@must.edu.mo (C.F.C.); yuhonho@gmail.com (H.H.Y.)

Abstract: The automatic analysis of endoscopic images to assist endoscopists in accurately identifying the types and locations of esophageal lesions remains a challenge. In this paper, we propose a novel multi-task deep learning model for automatic diagnosis, which does not simply replace the role of endoscopists in decision making, because endoscopists are expected to correct the false results predicted by the diagnosis system if more supporting information is provided. In order to help endoscopists improve the diagnosis accuracy in identifying the types of lesions, an image retrieval module is added in the classification task to provide an additional confidence level of the predicted types of esophageal lesions. In addition, a mutual attention module is added in the segmentation task to improve its performance in determining the locations of esophageal lesions. The proposed model is evaluated and compared with other deep learning models using a dataset of 1003 endoscopic images, including 290 esophageal cancer, 473 esophagitis, and 240 normal. The experimental results show the promising performance of our model with a high accuracy of 96.76% for the classification and a Dice coefficient of 82.47% for the segmentation. Consequently, the proposed multi-task deep learning model can be an effective tool to help endoscopists in judging esophageal lesions.

Keywords: classification; image retrieval; segmentation; multi-task; esophageal endoscopic images

Citation: Yu, X.; Tang, S.; Cheang, C.F.; Yu, H.H.; Choi, I.C. Multi-Task Model for Esophageal Lesion Analysis Using Endoscopic Images: Classification with Image Retrieval and Segmentation with Attention. *Sensors* **2022**, *22*, 283. https://doi.org/10.3390/s22010283

Academic Editor: Sheryl Berlin Brahnam

Received: 29 November 2021
Accepted: 27 December 2021
Published: 31 December 2021

1. Introduction

The incidence of esophageal lesions is getting higher and higher due to the continuous growth of the population and the changes in the lifestyles of people. Especially, esophageal cancer is the tenth-most common cancer in the world and has the sixth-highest mortality rate [1]. Fortunately, the treatment of esophageal cancer benefits from early detection; that is, it has a 5-year relative survival rate of more than 90%, while the later survival rate is less than 20% [2]. It also brings great troubles to the health of people.

At present, the typical strategies currently used to detect esophageal lesions are gastrointestinal endoscopic screening such as white light imaging (WLI), narrow-band imaging (NBI), capsule endoscopy, and so on. Unfortunately, there are some deficiencies in using gastrointestinal endoscopic to diagnose esophageal diseases. It is difficult to accurately diagnose patients with esophageal cancer based on WLI because of its lower sensitivity and specificity [3]. NBI not only requires experienced endoscopists to perform operations but is also expensive for examinations [4]. Therefore, less-experienced endoscopists are more likely to be unable to distinguish similar esophageal lesions and there is a lack of NBI equipment in low-income countries or regions. Likewise, capsule endoscopy will produce a large number of esophageal images. The endoscopist needs to make a diagnosis within a limited time, find the diseased regions, mark them, and finally, determine the treatment schemes [5]. In other words, endoscopists are busy fighting esophageal diseases every day. As a result, the manual diagnosis process based on gastrointestinal endoscopic screening of

the esophagus is affected by many negative factors, such as the experience level and mental state of endoscopists, the limitation of the diagnosis time, the huge esophageal image base, the subjective differences of different endoscopists, and so on. This is why the diagnosis of clinical esophageal lesions still has a high rate of missed diagnosis and misdiagnosis [6,7]. Therefore, it is of great significance to develop a support tool based on deep learning to not only classify but segment the lesions in esophageal endoscopic images, so as to reduce the burden on endoscopists, thereby improving the diagnosis accuracy.

Recently, benefiting from the rapid development of deep learning, many advanced techniques based on deep learning have been applied in various medical fields, such as skin lesion segmentation [8], retinal blood vessels segmentation [9], prostate cancer analysis [10], etc. As for the esophageal lesions analysis, such deep learning approaches have shown success in object detection [11], image classification [12], and semantic segmentation [13].

Although deep learning methods have made great contributions, especially for those diseases with a high fatality rate or difficult to diagnose, most of them have a common problem that they only focus on a given task or a certain type of disease, such as esophageal squamous cell carcinoma [14]. Therefore, it is very necessary to use deep learning methods to do a comprehensive analysis instead of only targeting a certain disease. Moreover, the aid information they provide to endoscopists is also limited, usually to only one classification accuracy rate. Especially for difficult-to-diagnose or controversial samples, endoscopists should be provided with more effective aid information, not just a simple classification accuracy rate. In other words, a multi-task model based on deep learning can provide endoscopists with various aid information in clinical applications, thereby making a diagnosis more efficient and more accurate. For example, esophageal lesion classification first distinguishes the types of esophageal lesions, and then, esophageal lesion segmentation can further determine the lesion regions. If the deep learning-based classification and segmentation models are developed separately, this consumes a long training time and requires large storage to store all the subnetwork models.

To solve the above problems, developing a model to achieve multiple tasks is a good strategy by using the shared features between different single-task deep learning models. Multi-task learning is an important paradigm of deep learning. Its goal is to mine common features between different tasks to improve the performance of the model and its better generalization ability [15]. The basic idea of multi-tasking is that different tasks can share some common features, so they are jointly trained. There are two methods commonly used in multi-task learning based on convolutional neural networks: soft parameter sharing [16] and hard parameter sharing [17]. Soft parameter sharing designs a model for each task with its parameters and uses regularization as a constraint to realize parameter similarities. Hard parameter sharing is to share the same hidden layers of the model between multiple tasks but have different task layers of the model to implement different tasks. It is noted that hard parameter sharing is the most common method of multi-task learning in neural networks, which can be traced back to the literature [18].

Therefore, based on the hard parameter sharing of multi-task learning, we developed a multi-task deep learning model that achieves classification and segmentation for esophageal lesions using endoscopic images at the same time. The classification task no longer only focuses on the prediction of a certain type of disease but can also predict esophagitis, esophageal cancer, or normal images. Additionally, image retrieval in the classification task is used to provide more aid in diagnostic information. For each query image, it can find the five most similar images from the historical patient libraries. When the endoscopists encounter a controversial or difficult-to-diagnose sample, the retrieval can provide more aid information besides classification results. The segmentation task can locate the cancer lesion area. It is better than the methods based on detection, because it avoids the problem of inaccurate positioning but a high confidence level. To achieve a better segmentation performance, we designed a mutual attention module to capture more diverse features in the segmentation task.

In summary, our contributions are mainly the following four:

(1) We proposed a novel multi-task deep learning model for automatic esophageal lesion analysis. It can synchronously achieve multiple tasks, including classification and segmentation for esophageal lesions.
(2) To provide endoscopists with more supporting information in classification, we built a retrieval module on the classification branch to assign a confidence for each prediction result. Classification and retrieval can be optimized at the same time without affecting each other.
(3) To improve the performance of esophageal cancer segmentation, we designed a mutual attention module in the segmentation task that can generate weight matrices from different features and guide each other to obtain diversified features.
(4) The experiments show that the proposed model is better than other similar methods and can effectively help endoscopists improve the accuracy of a diagnosis.

2. Related Works

In this section, we discuss three types of works that are most related to our work, including esophagus classification, esophagus segmentation, and a multi-task medical image analysis.

2.1. Esophagus Classification

A lot of the traditional classification method was proposed to classify esophageal lesions based on color and texture information. For example, Munzenmayer et al. [19] proposed a method based on a color texture analysis in a content-based image retrieval framework for precancerous lesions classification. Riaz et al. [20] put forward the autocorrelation Gabor feature method to extract the texture features of gastroenterology imaging for gastroenterological classification (normal, precancerous, and cancerous) and achieve better performance. Additionally, some machine learning algorithms, such as support vector machines [21] and principal component analysis [22], were employed in esophageal lesion classification.

Instead of the traditional classification methods, the deep learning-based method has been used in the classification of esophageal disease using endoscopic images. Kumagai et al. [23] constructed a GoogLeNet-based artificial intelligence tool to distinguish malignant and nonmalignant esophageal squamous cell carcinoma. Liu et al. [24] brought forward a transfer learning framework by fine-tuning pretrained models, such as VGGNets, Inception, and ResNets, to successfully classify gastric images into chronic gastritis, low-grade neoplasia, and early gastric cancer. Du et al. [25] proposed an efficient channel attention deep dense convolutional neural network that can classify diseases into four categories with a higher area under the curve value. We can see that the above-mentioned deep learning models could achieve obvious success in esophagus classification. An overview comparison of the methods for esophageal lesion classification is shown in Table 1.

Table 1. Comparison of the methods for esophageal lesion classification.

Authors	Methods	Performance
Münzenmayer et al. [19]	content-based image retrieval	0.71 kappa
Riaz et al. [20]	autocorrelation Gabor features	82.39% accuracy
Yeh et al. [21]	color coherence vector	92.86% accuracy
Liu et al. [22]	support vector machines	90.75% accuracy
Nakagawa et al. [12]	SSMD	91.00% accuracy
Kumagai et al. [23]	GoogLeNet	90.90% accuracy
Liu et al. [23]	VGGNets, etc.	89.00% accuracy
Du et al. [25]	ECA-DDCNN	90.63% accuracy
Igarashi et al. [26]	AlexNet	96.50% accuracy

2.2. Esophagus Segmentation

Many efforts are devoted to addressing esophagus segmentation by developing effective methods. Before deep learning, most methods used shape or appearance models to guide esophagus segmentation [27,28]. However, this model is difficult to train and has poor robustness. Sommen et al. [29] then proposed the algorithm that computes local color and texture features based on the original and the Gabor-filtered image to annotate regions of early esophageal cancer. Furthermore, Yang et al. [30] proposed an online atlas selection approach to choose a subset of optimal atlases for automatic segmentation of the esophagus.

Inspiring by the successful application of deep learning-related methods in esophagus classification, a growing number of deep learning-based approaches have been used in esophagus segmentation. For instance, Mendel et al. [31] used convolutional neural networks based on pretrained ResNets by a transfer learning method to segment adenocarcinoma in Barrett's esophagus. With the great success of U-Net in medical image segmentation [32], many of its variants have been proposed to employ esophagus segmentation. Huang et al. [33] proposed channel attention U-Net to segment esophageal cancer with a higher Dice value. Tran et al. [34] proposed a novel U-Net with an attention mechanism combined and STA-PLEalgorithm to achieve esophagus segmentation using 3D images. An overview comparison of the methods for esophageal lesion segmentation is shown in Table 2.

Table 2. Comparison of the methods for esophageal lesion segmentation.

Authors	Methods	Performance
Sommen et al. [29]	local color and texture features	0.95 recall
Yang et al. [30]	online atlas selection	0.73 DSC
Mendel et al. [31]	transfer learning	0.94 sensitivity
Huang et al. [33]	channel-attention U-Net	0.725 DV
Tran et al. [34]	spatial attention network and STAPLE algorithm	0.869 Dice
Chen et al. [35]	U-Net Plus	0.79 DV
Diniz et al. [36]	Atlas-based Residual-U-Net	0.8215 Dice

2.3. Multi-Task in Medical Image Analysis

Although the above-mentioned deep learning-related methods have achieved significant results in esophagus classification or segmentation, they can only achieve a given task. There are few methods that simultaneously realize classification and segmentation for esophageal endoscopic images. Wu et al. [37] proposed an esophageal lesion network composed of a classification and segmentation network for the classification and segmentation of esophageal lesions, but the classification and segmentation networks were required to be trained separately. For other medical images, multi-learning methods have been used. Chakravarty et al. [38] presented a multi-task convolutional neural network combining appearance features and structural features to achieve the segmentation and classification of glaucoma. In order to analyze skin lesions, especially melanoma, the author in Reference [39] proposed a multi-task framework to achieve three tasks: detection, classification, and segmentation. Zhang et al. [40] proposed a 3D multi-attention guided multi-task learning network by visual attention and adaptive spatial attention for simultaneous gastric tumor segmentation and lymph node classification. There are not many multi-task models applied to esophageal lesions, and they also have poor performance. The information they provide to endoscopists about each task is only a simple accuracy or Dice. Therefore, it is necessary to design a new multi-task deep learning model that can provide more aid information for esophageal lesion analysis.

3. Proposed Methods

In this section, we will first introduce the multi-task deep learning model as a whole and then describe each subtask in detail separately.

3.1. Network Architecture of The Proposed Multi-Task Deep Learning Model

In order to provide richer and more effective diagnostic information, we used hard parameter sharing to develop a novel multi-task deep learning model that realizes two tasks. The first task is to distinguish whether this sample is cancer, esophagitis, or normal. Additionally, based on the features of classification, image retrieval was used to find a group of images that are the most similar to the input images. The second task is to determine the lesion area when the image is cancer. Figure 1 depicts the architecture of the proposed model.

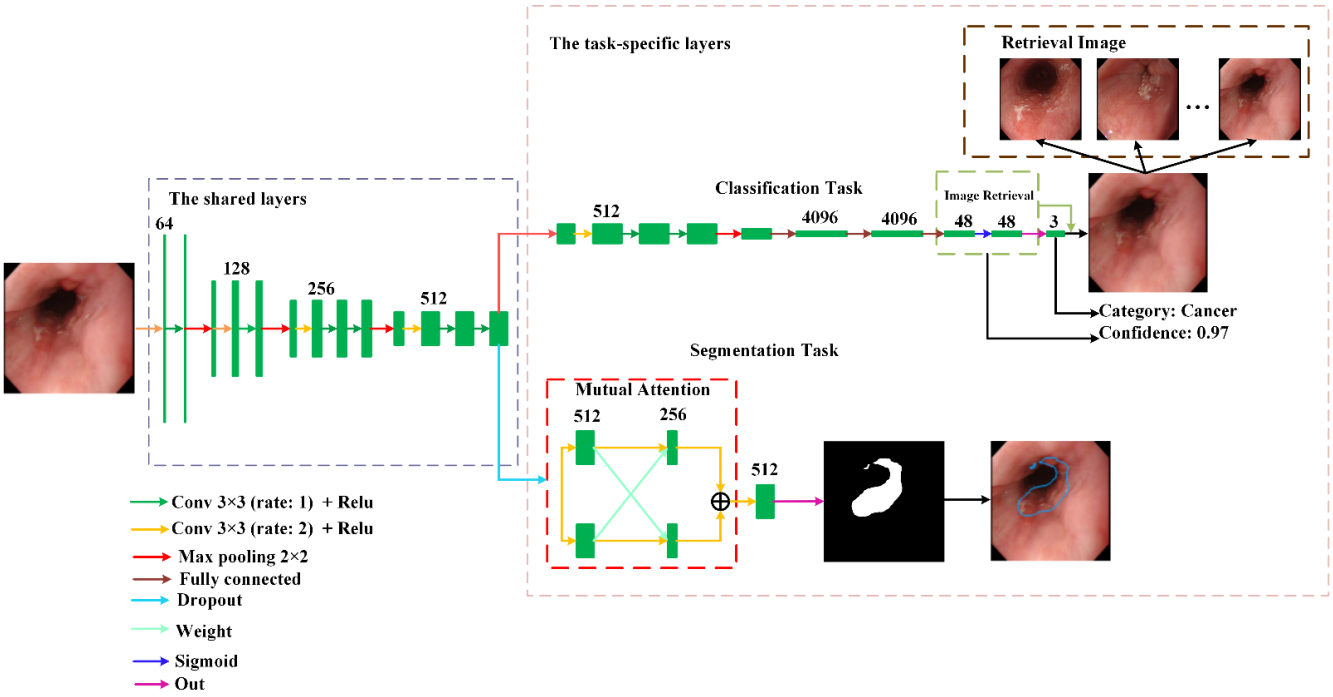

Figure 1. The network architecture of the proposed multi-task deep learning model.

It can be observed from Figure 1 that the proposed model is made up of the shared layers and task-specific layers. The shared layers located at the bottom of the model aim to extract common features between different tasks. The task-specific layers located at the upper region of the model consist of each task branch. Since the common features in the shared layers are not suitable for direct use in each task, convolution in the task-specific layers is used to extract the features suitable for each task to improve their performance.

At last, in order to reduce the negative impact caused by the imbalance of sample categories, the classification task adopts Focal loss [41] as the loss function. It is given by:

$$L_{cls} = -\alpha_{pred}\left(1 - p_{pred}\right)^{\gamma} log\left(p_{pred}\right) \quad (1)$$

where γ is the focusing parameter, and γ is 2, α is 0.25.

For the segmentation task, the cross-entropy loss function is used as the loss function. It is given by:

$$L_{seg} = -\frac{1}{K}\sum_{k=1}^{K}\left(g_n log(p_n) + (1 - g_n) log(1 - p_n)\right) \quad (2)$$

where K is the number of datasets, g is the truth label, and p is the output of the proposed model.

3.2. Classification and Segmentation Tasks

3.2.1. The Classification Task

The goal of the first task proposed is to determine the type of input image (cancer, esophagitis, or normal). This can be obtained through the classification branch. To help endoscopists make a more accurate diagnosis, we introduced a deep retrieval module [42] to provide more helpful information. This deep retrieval module consisted of a hash coding layer and a binary coding layer. The hash coding layer, which was a fully connected layer, was used to squeeze the feature into a fixed-length hash code. It was used to reduce the computational cost of image retrieval. The binary coding layer, which limits the characteristic parameters to 0 or 1, was aimed to binarize the hash code. For each image of the training set, the deep retrieval module outputted the corresponding binary hash code as its signature. We then used these signatures to build a feature library. When retrieving, every image query will get a signature from the deep retrieval module. By similarity calculations, we found a similar image ranking to the query from the feature library. Note that, since the feature is the binarized code, we adopted the appropriate Hamming distance as the similarity assessment. Next, we utilized image ranking to compute the confidence level of the predicted result.

In the training set, the number of samples for each category is different. When we obtain the predicted category and ranking of the query through the system, we take the top-n features from the ranking as candidates. n refers to the number of samples of the predicted category. Then, the missed candidates whose category is different from the predicted category of the query are removed from the candidates, leaving k hit candidates. Finally, we use the similarity of k hit candidates to average all n candidates as the confidence level of the prediction. The confidence level can be defined as follows:

$$C_q = \frac{\sum_{i=1}^{k} S_k}{n} \tag{3}$$

where C_q indicates the confidence level of the prediction. S_k means the similarity of the k_{th} hit candidate.

3.2.2. The Segmentation Task

When it was determined that the input image was a cancer lesion, the proposed model can mark where the lesion area was by the segmentation task. Compared with other detection-based methods that can only mark the approximate area of the lesion, Transformer [43] showed excellent performance in various visual fields. Inspired by SegFormer [44], we proposed a mutual attention module that can mark the lesion area more accurately, as shown in Figure 1. To enable our mutual attention module to capture different features, we used the dropout layer to generate differentiated feature maps. Then, the feature maps were fused using concatenation.

4. Experiments and Discussion

4.1. Dataset

The dataset used in this study contains 1003 upper gastrointestinal endoscopy images from Kiang Wu Hospital. All images can be categorized into three classes (240 normal, 473 esophagitis, and 290 esophagus cancer). Among them, the training set has 805 images (193 normal, 379 esophagitis, and 233 esophagus cancer), and the testing set has 198 images (47 normal, 94 esophagitis, and 57 esophagus cancer). All images have pathology reports, and the lesion areas are marked by experienced endoscopists. For data augmentation, we adopt random crop, random rotation between 45 and 135 degrees, horizontal flip, and vertical flip for the training set. This comprehensive data augmentation scheme makes the network converge better. The processes of training and testing the proposed model using the dataset are shown in Figure 2.

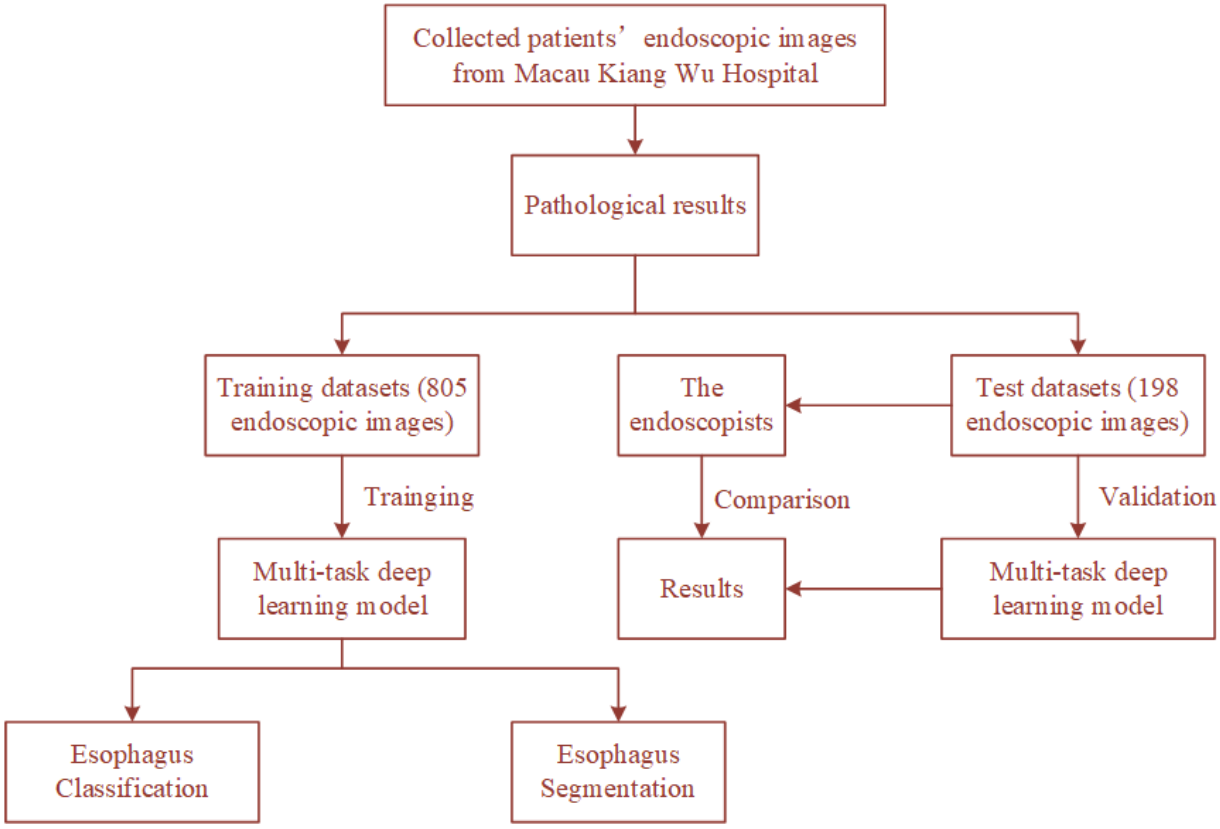

Figure 2. The processes of training and testing the proposed model using the dataset.

4.2. Evaluation Metric

To quantitatively analyze the performance of the proposed models, we employed the following three different metrics for two tasks.

For the classification task, we calculated the Accuracy Precision, Sensitivity, Specificity, Negative Predicted Value (NPV), and F1-score to evaluate the performance. They are defined as:

$$\text{Accuracy} = \frac{\sum_{c=1}^{C}(\text{TP}_c + \text{TN}_c)}{\sum_{c=1}^{C}(\text{TP}_c + \text{TN}_c + \text{FP}_c + \text{FN}_c)} \times 100\% \tag{4}$$

$$\text{Precision} = \frac{1}{C}\sum_{c=1}^{C}\frac{\text{TP}_c}{\text{TP}_c + \text{FP}_c} \times 100\% \tag{5}$$

$$\text{Sensitivity} = \frac{1}{C}\sum_{c=1}^{C}\frac{\text{TP}_c}{\text{TP}_c + \text{FN}_c} \times 100\% \tag{6}$$

$$\text{Specificity} = \frac{1}{C}\sum_{c=1}^{C}\frac{\text{TN}_c}{\text{TN}_c + \text{FP}_c} \times 100\% \tag{7}$$

$$\text{NPV} = \frac{1}{C}\sum_{c=1}^{C}\frac{\text{TN}_c}{\text{TN}_c + \text{FN}_c} \times 100\% \tag{8}$$

$$\text{F}_1 = 2 \times \frac{\text{Precision} \times \text{Sensitivity}}{\text{Precision} + \text{Sensitivity}} \tag{9}$$

where C is the number of types of esophageal lesions. TP (True Positives) means the number of positive samples is correctly classified. TN (True Negatives) means the number of negative samples is correctly classified. FP (False Positives) means the number of

negative samples is wrongly classified as positive. FN (False Negatives) means the number of positive samples is wrongly classified as negative.

To evaluate the image retrieval module, we adopted a ranking criterion to evaluate the retrieval performance. Given a query q, we obtained a ranking of each training set image using Hamming distance as the similarity measure. The precision of the query q in the top k rankings can be defined as:

$$precision@k = \frac{\sum_{i=1}^{k} hit(i)}{k} \tag{10}$$

where $hit(i)$ refers to whether the query q is consistent with the i_{th} image label in the ranking.

For the segmentation task, we adopted the most commonly used Dice coefficient and Intersection Over Union (IoU) as the evaluation metrics. They are defined as:

$$\text{Dice} = \frac{2|X \cap Y|}{|X|+|Y|} \times 100\% \tag{11}$$

$$\text{IoU} = \frac{|X \cap Y|}{|X \cup Y|} \times 100\% \tag{12}$$

where X represents the ground truth, which is masked by endoscopists, and Y is the segmentation region of the proposed model.

4.3. The Classification Results

The results of classification and retrieval are shown in Figure 3. For each image, it had a prediction given by the classification and the confidence level obtained by the retrieval. When the classification gives the correct diagnosis, the retrieval will also have a high confidence level, as shown in the first row of Figure 3. However, when the incorrect diagnosis is predicted, the confidence level will be low, as shown in the second row of Figure 3. This means that the proposed model can provide more effective supporting information when faced with difficult-to-diagnose or controversial samples. However, most of the current deep learning models just blindly improve the accuracy rate, ignoring that the main responsibility of the deep learning model is to provide effective diagnosis information.

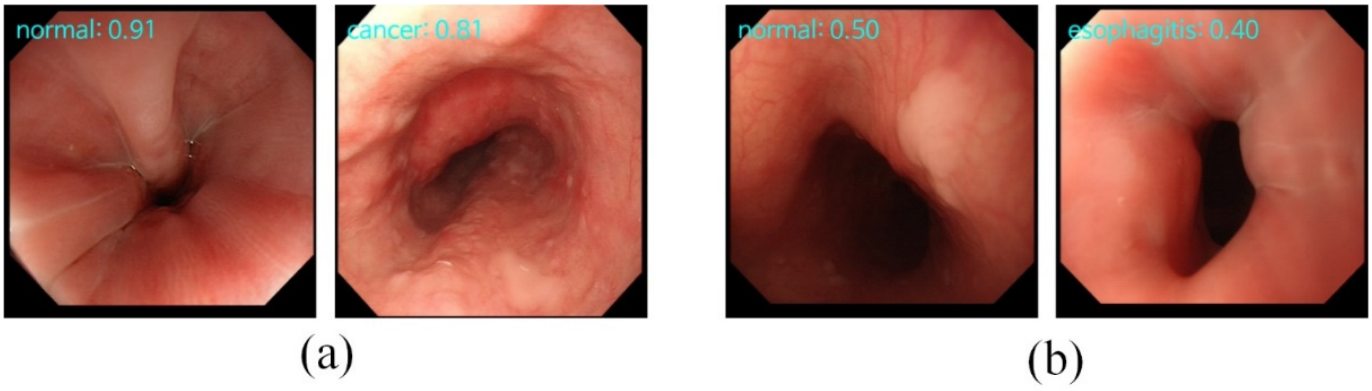

Figure 3. The results of classification and retrieval. "Type: 0.xx" on the top-left means the "predicted category: confidence level". (**a**) The input images with high confidence levels indicate the high possibility of a correct prediction made by the classification task. (**b**) The input images with low confidence levels indicate the high possibility of an incorrect prediction made by the classification task.

Additionally, in order to ensure the effectiveness and robustness of the proposed model, we compared five different CNN architectures, i.e., VGG-16 [45], ResNet-18 [46], ResNeXt-50 [47], Efficientnet-B0 [48], and RegNetY-400MF [49]. These architectures were trained and evaluated with the same protocol. We first evaluated the performance of the

classification task on the testing set separately. The compared results are shown in Table 3. It can be observed that the proposed model had a higher performance than the others in terms of the top-1 classification accuracy at 96.76 ± 0.22%. At the same time, we evaluated the performance of the retrieval module, and its accuracy was 91.67 ± 0.08%.

Table 3. Comparison of the classification results of our model and other models on the testing set.

Models	Top-1 Accuracy ± std	F1 Score ± std
VGG-16 [45]	92.68% ± 0.26	88.12% ± 0.26
ResNet-18 [46]	93.18% ± 0.25	88.36% ± 0.27
ResNeXt-50 [47]	94.34% ± 0.38	90.76% ± 0.33
Efficientnet-B0 [48]	95.15% ± 0.40	92.42% ± 0.39
RegNetY-400MF [49]	94.64% ± 0.52	91.57% ± 0.59
Ours	96.76% ± 0.22	94.22% ± 0.23

Next, to better verify the performance of the proposed model, we conducted confrontation ablation experiments on whether endoscopists refer to the results provided by the proposed model with the confidence of the predicted category. The endoscopists who participated in the testing included a senior (endoscopy experience > 10 years) and junior (endoscopy experience < 10 years), and the ratio was approximately 1:1. It is shown in Table 4.

Table 4. The diagnostic performance of the endoscopists without and with the proposed model.

Performance		Accuracy	Precision	Sensitivity	Specificity	NPV	F1-Score
Our model	cancer	98.48%	98.21%	96.49%	99.29%	99.59%	97.34%
	normal	96.46%	90.00%	95.74%	96.69%	98.65%	92.78%
	esophagitis	95.96%	96.74%	94.68%	97.12%	95.28%	95.70%
	all	96.96%	94.98%	95.64%	97.70%	97.84%	95.27%
Endoscopists only	cancer	91.41%	87.04%	82.46%	95.04%	93.06%	84.69%
	normal	83.84%	60.87%	89.36%	82.12%	96.12%	72.41%
	esophagitis	76.26%	81.33%	64.89%	86.54%	73.17%	72.19%
	all	83.84%	76.41%	78.90%	87.90%	87.45%	76.43%
Endoscopists (single classification)	cancer	93.43%	95.83%	78.90%	98.58%	92.67%	87.62%
	normal	87.04%	65.67%	93.62%	84.77%	97.71%	77.19%
	esophagitis	81.31%	84.34%	74.47%	87.5%	79.13%	79.10%
	all	87.26%	81.94%	82.33%	90.28%	89.84%	81.30%
Endoscopists (our model)	cancer	96.46%	93.10%	94.74%	97.16%	97.86%	93.91%
	normal	90.40%	73.33%	93.62%	89.4%	97.83%	82.24%
	esophagitis	89.9%	96.25	81.91%	97.12%	85.51%	88.50%
	all	92.25%	87.56%	90.09%	94.56%	97.73%	88.22%

In Table 4, we showed in detail the results of our model and the endoscopists. We can see that our model achieved the best performance on the test set. Its accuracy reached 96.96%. The precision, sensitivity, specificity, NPV, and F1 were 94.98%, 95.64%, 97.70%, 97.84%, and 95.27%. The endoscopists obtained 83.84% diagnosis accuracy without referring to the results provided by any deep learning models. Its precision, sensitivity, specificity, NPV, and F1-score were 76.41%, 78.90%, 87.90%, 87.45%, and 76.43%, respectively. To verify the effectiveness of the retrieval module, we conducted controlled experiments, as shown in Table 4. The endoscopists with a single classification indicated they only referred to the classification results. The endoscopists with our model indicated they referred to the classification and retrieval aid information provided by our model. Without the retrieval information, the average accuracy of the endoscopists was 87.26%. The average precision, sensitivity, specificity, NPV, and F1-score were 81.94%, 82.33%, 90.28%, 89.84%, and 81.30%. With the retrieval information, the average accuracy increased to 92.25%. The precision, sensitivity, specificity, NPV, and F1-score were 87.56%, 90.09%, 94.56%, 97.73%, and 82.22%,

respectively. In addition, we used Cohen's Kappa coefficient to evaluate the consistency of the diagnosis results between our model and the endoscopists. The Cohen's Kappa coefficients between our model and endoscopists, endoscopists with single classification, and endoscopists with our model were 0.5607, 0.7048, and 0.7258, respectively. We furthermore found that, after using the retrieval module, 27 of the 38 wrong diagnoses made by the endoscopists were corrected, as shown in Table 5.

Table 5. The diagnosis results of the endoscopists after referring to the results of the proposed model.

Counts		Endoscopists (Before)		Total
		Right	**Wrong**	
Endoscopists (after)	Right	152	**27**	175
	Wrong	8	11	23
Total		160	38	198

Figure 4 shows the output of the retrieval images selected for the endoscopists. For each input image, besides the confidence level of the predicted category, the top-five most similar labeled images retrieved from the training set are also provided to the endoscopists for making the diagnostic decision. This additional diagnostic information is helpful for endoscopists in dealing with difficult and controversial images.

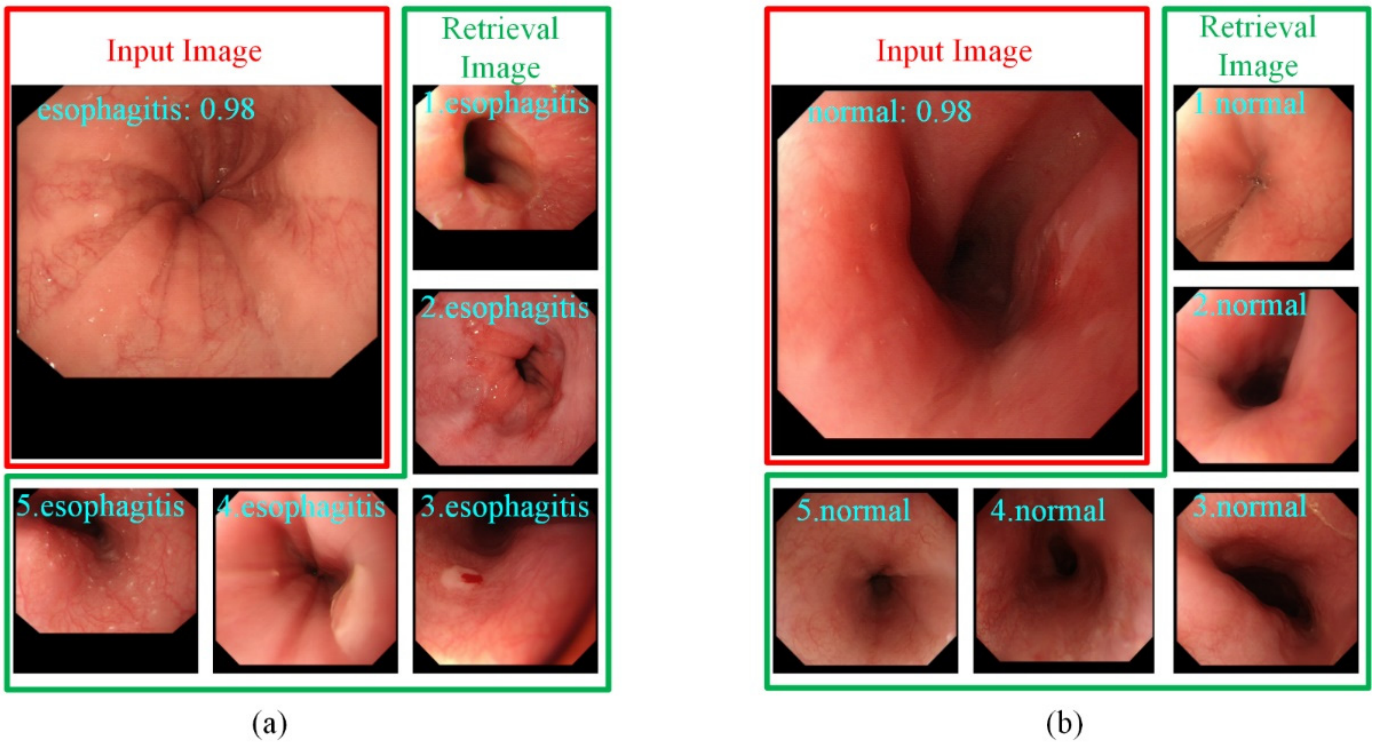

Figure 4. The top-5 most similar labeled samples are selected by images retrieval. (**a**) The input image is predicted as esophagitis. (**b**) The input image is predicted as normal.

Consequently, the proposed model can not only help endoscopists improve the accuracy of diagnosis, but the additional information provided by the retrieval module can further help endoscopists make a more accurate diagnosis. This demonstrates that the proposed model can be applied to daily clinical diagnoses.

4.4. The Segmentation Results

Six common segmentation models were compared, including U-Net [32], PSPNet [50], FCN [51], Deeplab V3+ [52], CCNet [53], OCRNet [54], and SegFormer [44]. The Dice and IoU are shown in Table 6. We could see that the IoU and Dice of the proposed model outperformed those of other models and were 71.27% and 82.47%. Additionally, we noticed that SegFormer, which is currently the best segmentation network, achieved the second-best results. This means the attention mechanism stimulated by the transformer enabled our model to accurately locate the cancerous area.

Table 6. Comparison of the segmentation results of our model and other models on the testing set.

Models	IoU	Dice
U-Net [32]	63.55%	75.12%
PSPNet [50]	62.28%	75.62%
FCN [51]	63.95%	76.72%
Deeplab V3+ [52]	66.24%	78.20%
CCNet [53]	62.52%	74.90%
OCRNet [54]	61.04%	73.63%
SegFormer [44]	67.25%	80.38%
Ours	71.27%	82.47%

Furthermore, the segmentation results of the proposed model and other models are shown in Figure 5. We observed that the cancer regions marked by our model were more accurate than the other models.

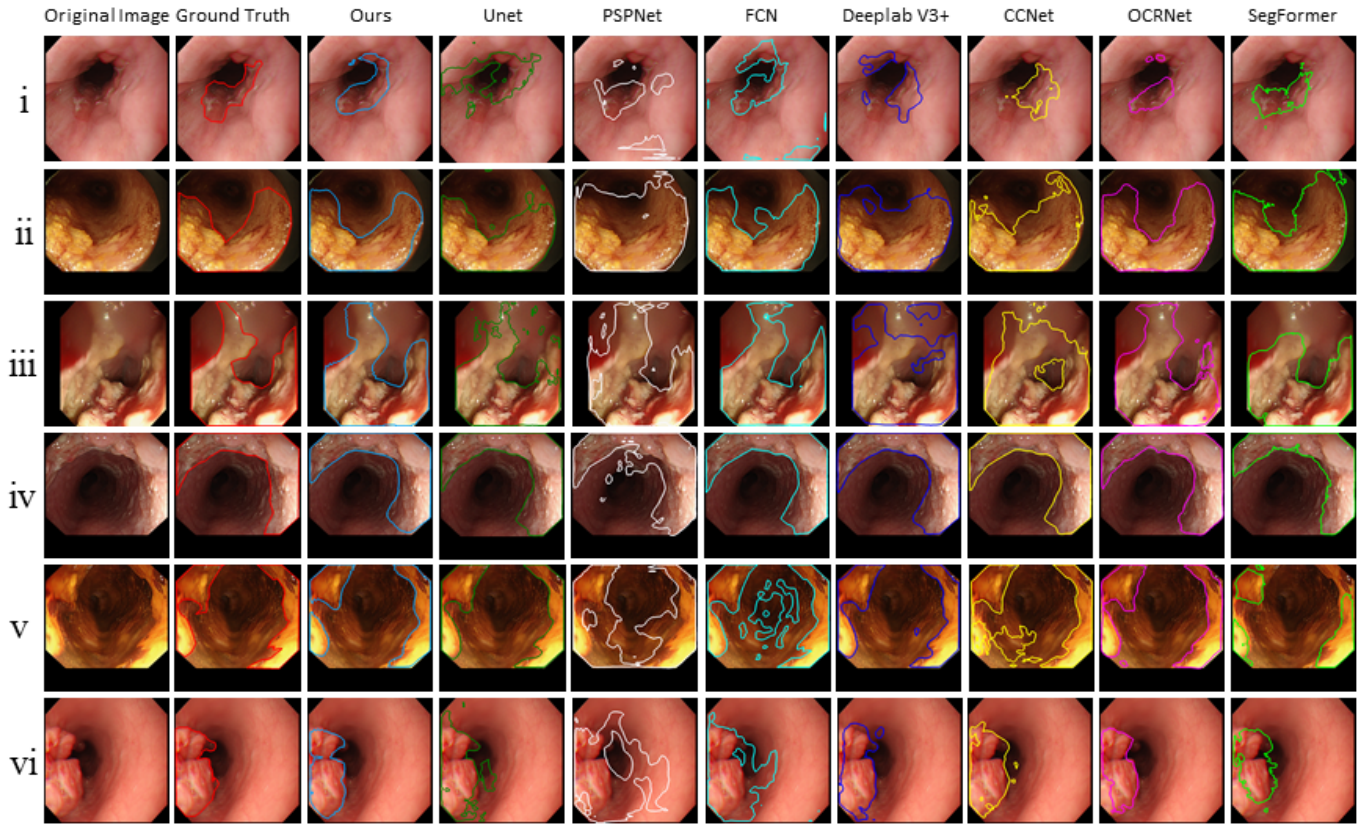

Figure 5. The segmentation results of the proposed model and other models.

5. Discussion

In this work, to solve the challenge of endoscopists in diagnosing esophageal lesions [55], we proposed a novel multi-task deep learning model to assist endoscopists in

improving the diagnosis accuracy of esophageal lesions. The proposed model showed a favorable performance for diagnosing esophageal diseases, with an accuracy of 96.76%. We can also intuitively see from the confusion matrix in Figure 6 that most of the images in the test set can be predicted correctly. Furthermore, to further verify the clinical application value of the proposed multi-task deep learning model, endoscopists were asked to review every image of the validation dataset with and without using the proposed model. By using this model, the average diagnostic accuracy was increased from 83.84% to 90.57%. The improvements in diagnostic ability confirmed the feasibility of the proposed model for helping endoscopists discover lesions ignored previously.

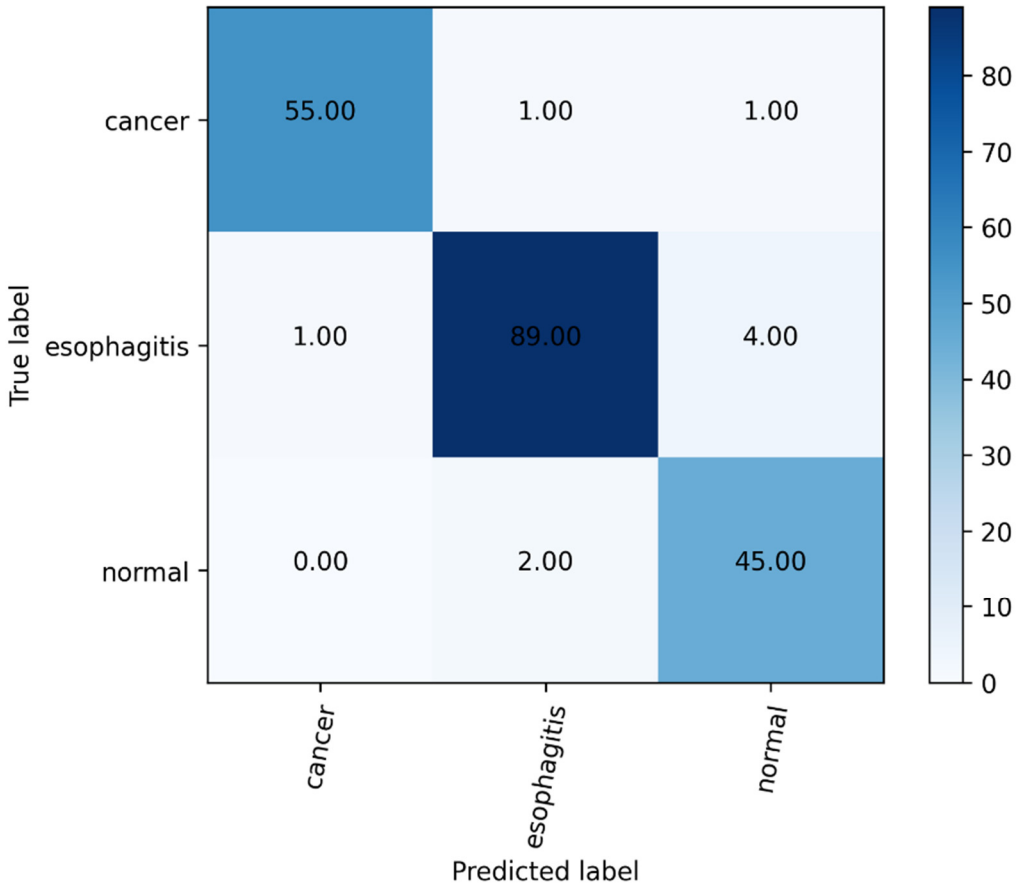

Figure 6. The confusion matrix of our model.

On the other hand, the proposed model can mask the esophageal cancer region with a high Dice coefficient (71.27%) and IoU (82.47%). Although previous studies have applied deep learning to classify or segment esophageal diseases, deep learning models can seldom classify and segment esophageal lesions at the same time. To the best of our knowledge, this is the first such multi-task deep learning model developed in Macau.

The proposed multi-task deep learning model not only achieved high accuracy in esophageal lesion classification but also output the mask of esophageal cancer, thereby reminding the endoscopists to pay attention to the location of the suspicious lesion. We hope that the proposed model can be used in the following situations: during the examination, it finds and masks a suspicious area under WLI; this will prompt the endoscopist to use the NBI mode and perform a biopsy. We are currently developing the multi-task deep learning model based on WLI and NBI images to establish a more subjective method that combines the current white light algorithm with the NBI algorithm.

Our work has several limitations. First, since our datasets only come from Macau Kiang Wu Hospital, the sample size (including images in the training and validation datasets) was small. Therefore, we plan to collect more images of different esophageal

types from different regions and invite more endoscopists to participate in our research. This will make up for the flaws of the imperfect data in this type of research. Second, our work only focused on cancer, esophagitis, and normal images and did not include other esophageal diseases such as esophageal polyps, esophageal leiomyoma, and ectopia of gastric mucosa. In the future, we will persistently collect these esophageal lesions and use them in the proposed model. Finally, we considered improving the robustness of the multi-task deep learning model for poor-quality images. The robustness of the multi-task deep learning model can be obtained by using poor-quality images during the training process. However, low-quality images will impair the convergence of the model and are not easily recognized by the model. Another feasible method is that we can use a model with a stronger learning ability to weaken the negative impact of low-quality images. We can also use a model with a stronger learning ability to weaken the negative impact of low-quality images. For example, we can consider the transformer [43] model that has recently shined in the field of deep learning or use NAS [56] to search for a specific model that deeply fits the endoscopic image of the esophagus. These methods will make the model more robust and able to cope with more complex situations.

6. Conclusions

In this paper, we constructed a multi-task deep learning model consisting of share layers and task-specific layers to achieve the classification and segmentation of esophageal lesions. The classification task determines the lesion type of the input image. Based on the classification task, image retrieval was used to provide more supporting information to endoscopists by finding a few samples that were the most similar in the input image. If the input image is cancer, the location of the cancer is further determined by cancerous area segmentation. To ensure the effectiveness and stability of the segmentation task, we developed an attention mechanism. The proposed model was evaluated on the testing set. The experimental results demonstrated that it was able to show a favorable diagnostic performance for classifying esophageal lesions with high accuracy and could achieve a high Dice coefficient and IoU for esophageal cancer segmentation. Furthermore, we invited endoscopists to compete with our model. The results showed our model achieved a classification accuracy of 96.76%. The accuracy of the endoscopists was 83.84%. Based on these promising results, the proposed multi-task deep learning model could become a potential assistant to help endoscopists in judging esophageal lesions.

Author Contributions: Conceptualization, X.Y., S.T. and C.F.C.; Data curation, S.T., C.F.C., H.H.Y. and I.C.C.; Formal analysis, X.Y.; Funding acquisition, C.F.C. and H.H.Y.; Investigation, S.T.; Methodology, X.Y.; Project administration, C.F.C. and H.H.Y.; Resources, C.F.C., H.H.Y. and I.C.C.; Software, X.Y.; Supervision, C.F.C.; Validation, X.Y., S.T. and C.F.C.; Visualization, X.Y.; Writing—original draft, X.Y. and S.T. and Writing—review and editing, C.F.C. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the Science and Technology Development Fund, Macau SAR (File no. 0023/2018/AFJ).

Institutional Review Board Statement: The ethical approval and inform consent were waived due to the retrospective design of this study, and due to the use of anonymized data that may not be connected to a real person.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Sung, H.; Ferlay, J.; Siegel, R.; Laversanne, M.; Soerjomataram, I.; Jemal, A.; Bray, F. Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J. Clin.* **2021**, *71*, 209–249. [CrossRef]
2. Rice, T.W.; Ishwaran, H.; Hofstetter, W.; Kelsen, D.; Apperson-Hansen, C.; Blackstone, E.H. Recommendations for pathologic staging (pTNM) of cancer of the esophagus and esophagogastric junction for the 8th edition AJCC/UICC staging manuals. *Dis. Esophagus* **2016**, *29*, 897–905. [CrossRef]

3. Ezoe, Y.; Muto, M.; Uedo, N.; Doyama, H.; Yao, K.; Oda, I.; Kaneko, K. Magnifying narrowband imaging is more accurate than conventional white-light imaging in diagnosis of gastric mucosal cancer. *Gastroenterology* **2011**, *141*, 2017–2025. [CrossRef]
4. Barbeiro, S.; Libanio, D.; Castro, R.; Dinis-Ribeiro, M.; Pimentel-Nunes, P. Narrow-band imaging: Clinical application in gastrointestinal endoscopy. *GE Port. J. Gastroenterol.* **2018**, *26*, 40–53. [CrossRef]
5. Pennazio, M. Capsule endoscopy: Where are we after 6 years of clinical use? *Dig. Liver Dis.* **2006**, *38*, 867–878. [CrossRef]
6. Mannath, J.; Ragunath, K. Role of endoscopy in early oesophageal cancer. *Nat. Rev. Gastroenterol. Hepatol.* **2016**, *13*, 720–730. [CrossRef]
7. Du, W.; Rao, N.; Liu, D.; Jiang, H.; Luo, C.; Li, Z.; Gan, T.; Zeng, B. Review on the applications of deep learning in the analysis of gastrointestinal endoscopy images. *IEEE Access* **2019**, *7*, 142053–142069. [CrossRef]
8. Ameri, A. A deep learning approach to skin cancer detection in dermoscopy images. *J. Biomed. Phys. Eng.* **2020**, *10*, 801–806. [CrossRef]
9. Liskowski, P.; Krawiec, K. Segmenting retinal blood vessels with deep neural networks. *IEEE Trans. Med. Imaging* **2016**, *35*, 2369–2380. [CrossRef]
10. Arif, M.; Schoots, I.; Tovar, J.; Bangma, C.; Krestin, G.; Roobol, M.; Nieesen, W.; Veenland, J. Clinically significant prostate cancer detection and segmentation in low-risk patients using a convolutional neural network on multi-parametric MRI. *Eur. Radiol.* **2020**, *30*, 6582–6592. [CrossRef]
11. Luo, H.; Xu, G.; Li, C.; He, L.; Luo, L.; Wang, Z.; Jing, B.; Deng, Y.; Jin, Y.; Li, B.; et al. Real-time artificial intelligence for detection of upper gastrointestinal cancer by endoscopy: A multicentre, case-control, diagnostic study. *Lancet* **2019**, *20*, 1645–1654. [CrossRef]
12. Nakagawa, K.; Ishihara, R.; Aoyama, K.; Ohmori, M.; Nakahira, H.; Matsuura, N.; Shichijio, S.; Nishida, T.; Yamada, T.; Yamaguchi, S.; et al. Classification for invasion depth of esophageal squamous cell carcinoma using a deep neural network compared with experienced endoscopists. *Gastrointest. Endosc.* **2019**, *90*, 407–414. [CrossRef]
13. Cao, R.; Pei, R.; Ge, N.; Zheng, C. Clinical target volume auto-segmentation of esophageal cancer for radiotherapy after radical surgery based on deep learning. *Technol. Cancer Res. Treat.* **2021**, *20*, 1–11. [CrossRef]
14. Guo, L.; Xiao, X.; Wu, C.; Zeng, X.; Zhang, Y.; Du, J.; Bai, S.; Zhang, Z.; Li, Y.; Wang, X. Real-time automated diagnosis of precancerous lesions and early esophageal squamous cell carcinoma using a deep learning model (with videos). *Gastrointest. Endosc.* **2020**, *91*, 41–51. [CrossRef]
15. Caruana, R. Multitask learning. *Mach. Learn.* **1998**, *27*, 95–133.
16. Misra, I.; Shrivastava, A.; Gupta, A.; Hebert, M. Cross-Stitch Networks for Multi-Task Learning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016.
17. Kokkinos, I. Ubernet: Training a Universal Convolutional Neural Network for Low-, Mid-, and High-Level Vision Using Diverse Datasets and Limited Memory. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017.
18. Caruana, R. Multitask Learning: A Knowledge Based Source of Inductive Bias. In Proceedings of the Tenth International Conference on Machine Learning, Amherst, MA, USA, 27–29 July 1993.
19. Münzenmayer, C.; Kage, A.; Wittenberg, T.; Mühldorfer, S. Computer-assisted diagnosis for precancerous lesions in the esophgus. *Methods Inf. Med.* **2009**, *48*, 324–330.
20. Riaz, F.; Silva, F.; Ribeiro, M.; Coimbra, M. Invariant gabor texture descriptors for classification of gastroenterology images. *IEEE Trans. Biomed. Eng.* **2012**, *59*, 2893–2904. [CrossRef]
21. Yeh, J.Y.; Wu, T.; Tsai, W.J. Bleeding and ulcer detection using wireless capsule endoscopy images. *J. Softw. Eng. Appl.* **2014**, *7*, 422–432. [CrossRef]
22. Liu, D.; Gan, T.; Rao, N.; Xing, Y.; Zheng, J.; Li, S.; Luo, C.; Zhou, Z.; Wan, Y. Identification of lesion images from gastrointestinal endoscope based on feature extraction of combinational methods with and without learning process. *Med. Image Anal.* **2016**, *32*, 281–294. [CrossRef]
23. Kumagai, Y.; Takubo, K.; Kawada, K.; Aoyama, K.; Endo, Y.; Ozawa, T.; Hirasawa, T.; Yoshio, T.; Ishihara, S.; Fujishiro, M.; et al. Diagnosis using deep-learning artificial intelligence based on the endocytoscopic observation of the esophagus. *Esophagus* **2019**, *16*, 180–187. [CrossRef]
24. Liu, X.; Wang, C.; Bai, J.; Liao, G. Fine-tuning pre-trained convolutional neural networks for gastric precancerous disease classification on magnification narrow-band imaging images. *Neurocomputing* **2020**, *392*, 253–267. [CrossRef]
25. Du, W.; Rao, N.; Dong, C.; Wang, Y.; Hu, D.; Zhu, L.; Zeng, B.; Gan, T. Automatic classification of esophageal disease in gastroscopic images using an efficient channel attention deep dense convolutional neural network. *Biomed. Opt. Express* **2021**, *12*, 3066–3081. [CrossRef]
26. Igarashi, S.; Sasaki, Y.; Mikami, T.; Sakuraba, H.; Fukuda, S. Anatomical classification of upper gastrointestinal organs under various image capture conditions using AlexNet. *Comput. Biol. Med.* **2020**, *124*, 103950. [CrossRef]
27. Fieselmann, A.; Lautenschläger, S.; Deinzer, F.; Matthias, J.; Poppe, B. Esophagus segmentation by spatially-constrained shape interpolation. In Proceedings of the Bildverarbeitung für die Medizin 2008: Algorithmen–Systeme–Anwendungen, Proceedings des Workshops, Berlin, Germany, 6–8 April 2008; pp. 247–251.
28. Feulner, J.; Zhou, S.; Cavallaro, A.; Seifert, S.; Hornegger, J.; Comaniciu, D. Fast automatic segmentation of the esophagus from 3D CT data using a probabilistic model. *Med. Image Comput. Comput. Assist. Interv.* **2009**, *12*, 255–262.

29. Sommen, F.; Zinger, S.; Schoon, E.; With, P. Supportive automatic annotation of early esophageal cancer using local gabor and color features. *Neurocomputing* **2014**, *144*, 92–106. [CrossRef]
30. Yang, J.; Haas, B.; Fang, R.; Beadle, B.M.; Garden, A.S.; Liao, Z.; Zhang, L.; Balter, P.; Court, L. Atlas ranking and selection for automatic segmentation of the esophagus from CT scans. *Phys. Med. Biol.* **2017**, *62*, 9140–9158. [CrossRef]
31. Mendel, R.; Ebigbo, A.; Probst, A.; Messmann, H.; Palm, C. Barrett's esophagus analysis using convolutional neural networks. In Proceedings of the Bildverarbeitung für die Medizin 2017: Algorithmen–Systeme–Anwendungen, Proceedings des Workshops, Heidelberg, Germany, 12–14 March 2017; pp. 80–85.
32. Ronneberger, O.; Fischer, P.; Brox, T. U-net: Convolutional networks for biomedical image segmentation. In Proceedings of the 18th International Conference on Medical Image Computing and Computer-Assisted Intervention, Munich, Germany, 5–9 October 2015.
33. Huang, G.; Zhu, J.; Li, J.; Wang, Z.; Zhou, J. Channel-attention U-Net: Channel attention mechanism for semantic segmentation of esophagus and esophageal cancer. *IEEE Access* **2020**, *8*, 122798–122810. [CrossRef]
34. Tran, M.; Kim, S.; Yang, H.; Lee, G.; Oh, I.; Kang, S. Esophagus segmentation in CT images via spatial attention network and STAPLE algorithm. *Sensors* **2021**, *21*, 4456. [CrossRef]
35. Chen, S.; Yang, H.; Fu, J.; Mei, W.; Ren, S.; Liu, Y.; Zhu, Z.; Liu, L.; Li, H.; Chen, H. U-Net Plus: Deep Semantic Segmentation for Esophagus and Esophageal Cancer in Computed Tomography Images. *IEEE Access* **2019**, *7*, 82867–82877. [CrossRef]
36. Diniz, J.; Ferreira, J.; Diniz, P.; Silva, A.; Paova, A. Esophagus segmentation from planning CT images using an atlas-based deep learning approach. *Comput. Methods Programs Biomed.* **2020**, *197*, 105685. [CrossRef]
37. Wu, Z.; Ge, R.; Wen, M.; Liu, G.; Chen, Y.; Zhang, P.; He, X.; Hua, J.; Luo, L.; Li, S. ELNet: Automatic classification and segmentation for esophageal lesions using convolutional neural network. *Med. Image Anal.* **2021**, *67*, 101838. [CrossRef]
38. Chakravarty, A.; Sivswamy, J. A deep learning based joint segmentation and classification framework for glaucoma assessment in retinal color fundus images. *arXiv* **2018**, arXiv:1808.01355. Available online: https://arxiv.org/abs/1808.01355 (accessed on 29 July 2018).
39. Song, L.; Lin, J.; Wang, Z.; Wang, H. An end-to-end multi-task deep learning framework for skin lesion analysis. *IEEE J. Biomed. Health Inform.* **2020**, *24*, 2912–2921. [CrossRef] [PubMed]
40. Zhang, Y.; Li, H.; Du, J.; Qin, J.; Wang, T.; Chen, Y.; Liu, B.; Gao, W.; Ma, G.; Lei, B. 3D multi-attention guided multi-task learning network for automatic gastric tumor segmentation and lymph node classification. *IEEE Trans. Med. Imaging* **2021**, *40*, 1618–1631. [CrossRef]
41. Lin, T.; Goyal, P.; Girshick, R.; He, K.; Dollar, P. Focal loss for dense object detection. *IEEE Trans. Pattern Anal. Mach. Intell.* **2020**, *42*, 318–327. [CrossRef]
42. Lin, K.; Yang, H.; Hsiao, J.; Chen, C. Deep Learning of Binary Hash Codes for Fast Image Retrieval. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, Boston, MA, USA, 11–12 June 2015; pp. 27–35.
43. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An image is worth 16 × 16 worlds: Transformers for image recognition at scale. *arXiv* **2020**, arXiv:2010.11929. Available online: https://arxiv.org/abs/2010.11929 (accessed on 22 October 2020).
44. Xie, E.; Wang, W.; Yu, Z.; Anndkumar, A.; Alvarez, J.; Luo, P. SegFormer: Simple and Efficient Design for Semantic Segmentation with Transformers. *arXiv* **2021**, arXiv:2105.15203. Available online: http://arxiv.org/abs/2105.15203 (accessed on 31 May 2021).
45. Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv* **2014**, arXiv:1409.1556. Available online: http://arxiv.org/abs/1409.1556 (accessed on 10 April 2015).
46. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016.
47. Xie, S.; Girshick, R.; Dollar, P.; Tu, Z.; He, K. Aggregated residual transformations for deep neural networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017.
48. Tan, M.; Le, Q. EfficientNet: Rethinking model scaling for convolutional neural networks. *arXiv* **2019**, arXiv:1905.11946v1. Available online: http://arxiv.org/abs/1905.11946v1 (accessed on 28 May 2019).
49. Radosavovic, I.; Kosaraju, R.; Girshick, R.; He, K.; Dollar, P. Designing Network Design Spaces. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 13–19 June 2020.
50. Zhao, H.; Shi, J.; Qi, X.; Wang, X.; Jia, J. Pyramid scene parsing network. *arXiv* **2017**, arXiv:1612.01105. Available online: http://arxiv.org/abs/1612.01105 (accessed on 4 December 2016).
51. Shelhamer, E.; Long, J.; Darrell, T. Fully convolutional networks for semantic segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 640–651. [CrossRef] [PubMed]
52. Chen, L.; Zhu, Y.; Papandreou, G.; Schroff, F.; Adam, H. Encoder-decoder with atrous separable convolution for semantic image segmentation. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 833–851.
53. Huang, Z.; Wang, X.; Wei, Y.; Huang, L.; Shi, H.; Liu, W.; Huang, T. CCNet: Criss-cross attention for semantic segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2020**, *14*, 1–14. [CrossRef] [PubMed]
54. Yuan, Y.; Chen, X.; Wang, J. Object-contextual representations for semantic segmentation. In Proceedings of the European Conference on Computer Vision (ECCV), Glasgow, UK, 23–28 August 2020; pp. 173–190.
55. Cotton, P.B. Quality endoscopists and quality endoscopy units. *J. Interv. Gastroenterol.* **2011**, *1*, 83–87. [CrossRef] [PubMed]
56. Elsken, T.; Metzen, J.H.; Hutter, F. Neural architecture search: A survey. *J. Mach. Learn. Res.* **2019**, *20*, 1–21.