Hindawi
Computational and Mathematical Methods in Medicine
Volume 2022, Article ID 1691075, 14 pages
https://doi.org/10.1155/2022/1691075

Research Article

Double-Balanced Loss for Imbalanced Colorectal Lesion Classification

Chang Yu,[1] **Wei Sun**,[1] **Qilin Xiong**,[1] **Junbo Gao**,[1] **and Guoqiang Qu**[2]

[1]*Information Engineering College, Shanghai Maritime University, Shanghai 201306, China*
[2]*Department of Gastroenterology, Eastern Hospital, Shanghai Sixth People's Hospital, Shanghai 201306, China*

Correspondence should be addressed to Wei Sun; weisun@shmtu.edu.cn

Received 23 April 2022; Revised 6 July 2022; Accepted 13 July 2022; Published 8 August 2022

Academic Editor: Lin Lu

Colorectal cancer has a high incidence rate in all countries around the world, and the survival rate of patients is improved by early detection. With the development of object detection technology based on deep learning, computer-aided diagnosis of colonoscopy medical images becomes a reality, which can effectively reduce the occurrence of missed diagnosis and misdiagnosis. In medical image recognition, the assumption that training samples follow independent identical distribution (IID) is the key to the high accuracy of deep learning. However, the classification of medical images is unbalanced in most cases. This paper proposes a new loss function named the double-balanced loss function for the deep learning model, to improve the impact of datasets on classification accuracy. It introduces the effects of sample size and sample difficulty to the loss calculation and deals with both sample size imbalance and sample difficulty imbalance. And it combines with deep learning to build the medical diagnosis model for colorectal cancer. Experimentally verified by three colorectal white-light endoscopic image datasets, the double-balanced loss function proposed in this paper has better performance on the imbalance classification problem of colorectal medical images.

1. Introduction

A survey shows that malignant tumors have become the first killer of Chinese residents' health by 2020. There is an evidence that the incidence rate and mortality of colorectal cancer are increasing at a particularly significant rate. In 2020, the number of new cases and deaths of colorectal cancer has increased to more than 500,000 and more than 300,000, respectively, which has seriously threatened the health of Chinese residents [1]. Colorectal white-light endoscopy is one of the most widely used of polyp detection and have been extensively used for early colorectal cancer screening. Previous research has shown that missed and misdiagnosed colorectal polyps will increase the possibility of colorectal cancer, and their survival rate is less than 10% [2]. Therefore, the proper classification of colorectal polyps by white-light endoscopic images can be used to assist physicians in the early screening of colorectal cancer.

With the application of artificial intelligence in the field of intelligent medicine, deep learning models are widely used in the lesion detection and classification of medical images. This paper reviews and compares the main research work of large intestine white-light endoscope image recognition from two aspects: deep learning models and lesion classification methods; the results are shown in Table 1. The existing body of research on colorectal polyps' classification suggests that classification standards are not uniform. Komeda et al. [3, 4] used convolutional neural networks to classify lesions into neoplastic and nonneoplastic and adenomatous and nonadenomatous. Gao et al. [5] used ResNet50 [6] to distinguish whether colonoscopic images contain lesions, then detected specific lesions and divided them into adenoma, cancer and polyp, which detected AP50 up to 0.903. Taş and Yilmaz [7] proposed to perform super-resolution reconstruction of colonoscopic images to obtain high-resolution images, then use Faster R-CNN [8] for detection, which improved the accuracy of model detection by 8% through super-resolution processing. Shin et al. [9] used image enhancement such as rotation and scaling to increase the number of training samples and then used Faster R-CNN

Table 1: Deep learning in the diagnosis of colorectal white-light endoscopy.

Study	Date	Model	Classes
Komeda et al. [3]	2016	CNNs	Neoplastic, nonneoplastic
Eduardo et al. [4]	2017	CNNs	Adenoma, nonadenoma
Gao et al. [5]	2020	Mask R-CNN	Cancer, adenoma, polyp
Taş and Yilmaz [7]	2021	Faster R-CNN	Polyp, nonpolyp
Shin et al. [9]	2018	Faster R-CNN	Polyp, nonpolyp
Shin et al. [10]	2019	Faster R-CNN	Polyp, nonpolyp
Nogueira-Rodríguez et al. [11]	2020	YOLO	Polyp, nonpolyp

for detection. Shin et al. [10] proposed a framework for generating synthetic polyp images to augment the training data to detect polyps using Faster R-CNN with a 20% improvement in accuracy. Among them, the literature [7, 9–11] did not classify the lesions but only detected the presence of polyps.

In early colorectal cancer screening, the classification of colorectal polyps is important. Clinically, physicians need to determine whether to perform resection surgery based on the specific category of polyps. At present, there is little research on the classification of colorectal polyps under deep learning, mainly due to the lack of large public datasets for polyp classification [11]. Based on the experience of doctors in the Shanghai Sixth People's Hospital (East Hospital) in the clinical diagnosis of colon white-light endoscopy, this paper adopts the classification criteria proposed in literature [4]; that is, colon lesions are divided into polyps (polyp), adenomatous polyps (adenoma), and cancerous structures (cancer). Furthermore, the collected colorectal images with lesions were labeled into three categories to carry out a classification study of colorectal white-light endoscopic images.

2. Problem Statement and Background

In recent years, deep learning has achieved good results in the detection and classification of medical image lesions, which can assist doctors in the diagnosis and treatment work to a certain extent. This mainly depends on the powerful learning ability of deep learning models in the image field and the recognition accuracy that can match or even exceed the human eyes. However, the deep learning model is based on the assumption that the samples are independently and equally distributed in the training set, but this premise does not hold in most medical image datasets. A survey showed that about 75% of colorectal polyps are adenomatous polyps [12], which leads to the fact that most of the colorectal white-light endoscopic images collected by doctors from clinical diagnosis are adenoma images. Unbalanced data makes it difficult for deep learning models to be applied in the field of medical-assisted diagnosis. There are two types of approaches designed to address imbalance classification. The first category is the data-level approach, which adjusts the adaptability of the model to the data by changing the data distribution of the training set, such as random oversampling and random undersampling. The second category is the algorithmic-level approach, which does not change the training set but adjusts the training strategy or model structure, such as reconfiguring the classifier, two-stage training, and improving the loss function. Moreover, methods that combine the two types are available.

2.1. Data-Level Approach. Random undersampling and random oversampling are the classical methods to solve data imbalance; both of them change the original distribution of the dataset. In 2015, Bae and Yoon [13] proposed an upsampling enhancement framework based on data sampling to use rebalanced datasets to learn comprehensive classifiers and use them to detect different types of polyps. The experimental results show that the performance is improved compared with other most advanced detectors. However, random undersampling may cause the samples to lose important feature information during undersampling, while random oversampling increases the risk of overfitting. SMOTE [14] is a more advanced sampling method to overcome this problem, in which new samples are generated by adding data points from the nearest neighbors by interpolation; the limitation of this method is that it cannot overcome the problem of data distribution in unbalanced datasets, which increases the difficulty of classification algorithms to classify them. Another sampling method is class-aware sampling for stochastic gradient descent optimization neural networks [15], whose main idea is to ensure that the sample classes of each batch are evenly distributed in training.

2.2. Algorithm-Level Approach. One algorithm-level approach to address data imbalance is to reconfigure the classifier, which can be implemented in various ways, such as one-class classifier, which uses small classes as outliers and transforms the classification problem into abnormal detection [16]. Two-stage training [17] is first performed on a balanced dataset, and then, the final output layer is fine-tuned on the unbalanced original dataset. Although these methods can alleviate the data imbalance problem to some extent, the improvement of loss function has more attractive features, such as ease of implementation. Cross-entropy loss is widely used in classification problems, but it cannot handle data imbalance. A simple improvement is to use weighted cross-entropy (WCE) according to the number of categories, which is often ineffective in practice.

In 2019, Cui et al. [18] proposed the class-balanced loss framework, using the effective number of samples per class to inversely weight the loss, and this method can effectively solve the class number imbalance problem. Kim et al. [19] proposed complement cross-entropy (CCE) to solve the problem of imbalanced classification and proved that suppressing the probability of the error class helps the deep learning model to learn discriminative information. By neutralizing the confidence of the error sample, the minority class samples get more learning opportunities. In addition, the sample difficulty imbalance can be handled by improving the loss function. In 2016, Shrivastava et al. [20] proposed the OHEM algorithm to screen out difficult samples based on the loss value of the input samples, and then, the screened samples are applied to training in stochastic gradient descent. This method achieves online hard example mining, but it completely discards simple samples, which leads to the model's inability to improve their detection accuracy. In 2017, Lin et al. [21] proposed focal loss to increase the model's focus on difficult samples by dynamically adjusting the loss contribution of difficult samples by introducing a $(1-p)^{\gamma}$ modulation factor in the cross-entropy, but this approach assigns high weights to outliers as difficult samples. The GHM [22] loss function can improve this problem, and its basic idea is to start from the gradient parity of the samples and to dynamically weight the samples according to the proportion of samples accounted for by the gradient parity so that easy samples with small gradients are downweighted, difficult samples with medium gradients are upweighted, and outlier samples with large gradients are downweighted.

3. Materials and Methods

3.1. The Proposed Methods. The most common loss function used in deep learning multiclassification tasks is cross-entropy (CE) loss. The cross-entropy loss gives equal importance to each data instance, which will lead to the network monitoring the classes with fewer number of observations. Therefore, CE loss is inappropriate in classification tasks under class imbalance. This paper proposes a new loss function named double-balanced (DB) loss. We derive it from the perspective of sample size and sample difficulty.

3.1.1. Imbalance of Sample Size. Usually, the method to deal with the unbalanced sample size is to assign a weight to the sample that is inversely proportionate to the class frequency. Since the weights are chosen with a fixed value to the number of samples of each class in the total sample, it does not work well for deep learning when using the batch gradient descent optimization method SGD. This is due to the overlap of features between different samples, as shown in Figure 1. As the number of samples increases, the features carried by new samples already exist in the original samples, and the model does not learn new features from the new samples, so this increase of sample size is ineffective for model training.

To address the problem of sample size imbalance, this paper rebalances the loss by effective samples size [18]. First,

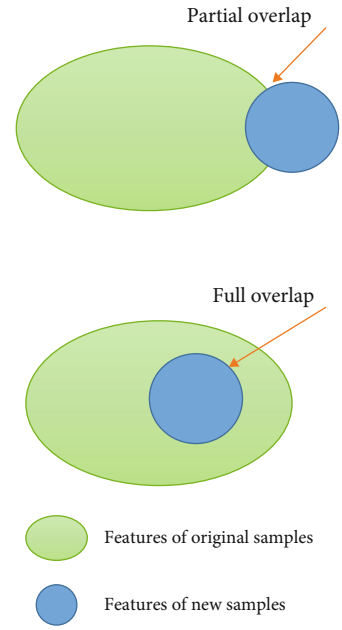


Figure 1: The features contained in the new samples may already be partially or fully contained in the original samples.

the effective sample size of each category is calculated using the following equation.

$$E_{n_i} = \frac{1-\beta^{n_i}}{1-\beta}, \tag{1}$$

where n_i denotes the true number of samples in each category, β is the hyperparameter that controls the growth rate of E_{n_i} with n_i, $\beta \in (0,1)$, and here, it is set at $[0.9,1)$. It can be seen from the above that a larger E_{n_i} indicates a larger effective sample size of category i in the training sample, and its loss proportion should be as small as possible. Therefore, the loss value is inversely related to the effective sample size as follows.

$$e_i = \frac{1}{E_{n_i}} = \frac{1-\beta}{1-\beta^{n_i}}. \tag{2}$$

Secondly, considering the large difference in the number of each category, the effective sample size is normalized to get the sample size balance factor α_i for each category:

$$\alpha_i = \frac{e_i}{\sum_{j=1}^{k} e_j} = \frac{(1-\beta)/(1-\beta^{n_i})}{\sum_{j=1}^{k}((1-\beta)/(1-\beta^{n_j}))}. \tag{3}$$

where k denotes the number of categories. The weight α_i based on the effective sample size of each category is obtained from (3), and in turn, this weight is added to the loss calculation.

$$\text{Loss} = L(p) + \alpha_i L(p) = (1+\alpha_i)L(p), \tag{4}$$

where $L(p)$ denotes the original loss and p denotes the predicted probability. Here, $\alpha_i L(p)$ is called the loss based on the effective sample size, and the total loss is the original loss plus the loss based on the effective sample size.

3.1.2. Imbalance of Sample Difficulty. The difficult samples refer to less clearly classified borders on the transition region between the foreground and background, while the easy samples refer to background samples that do not overlap with the real target or positive samples that have a high degree of overlap with the real target. We proposed a method to determine and calculate the difficulty of samples. The basic idea is to calculate the distance between the prior probability distribution and the predicted probability distribution of a category and then measure the difficulty of the sample based on the magnitude of the distance.

First, the empirical class frequencies of each category [17] were calculated as their prior probabilities:

$$p_p(c_i) = \frac{(1/n_i)^\rho}{\sum_{j=1}^{k}(1/n_j)^\rho}, \tag{5}$$

where n_i denotes the number of categories c_i and ρ, as a hyperparameter, denotes the degree of flexibility. After calculating the prior probability of each category, the prior distribution of the training samples is obtained. The cross-entropy function is then used to calculate the distance between the two distributions.

$$H\left(p_p, p_s\right) = \sum_{i=1}^{k} -p_p(c_i)\log p_s(c_i), \tag{6}$$

where p_p denotes the prior given matrix of the prior distribution and p_s denotes the matrix of each category of distribution obtained by the softmax function.

Considering the mutual exclusivity between categories in the classification problem, we only calculate the prior probabilities and predicted probabilities corresponding to the true categories in the two probability distribution matrices, where the true vector corresponding to a single category c_i is

$$\text{true}_{c_i} = [0, 0, \cdots 1, \cdots 0]^T. \tag{7}$$

true_{c_i} is a $k * 1$ matrix. In this matrix, the value of the category c_i is 1 at the corresponding position, and the rest are 0. Next, we do the following calculation.

$$p_p^T * \text{true}_{c_i} = \left[p_p(c_1), p_p(c_2), \cdots, p_p(c_k)\right] * \begin{bmatrix} 0 \\ 0 \\ \cdots \\ 1 \\ \cdots \\ 0 \end{bmatrix} = p_p(c_i),$$

$$p_s^T * \text{true}_{c_i} = [p_s(c_1), p_s(c_2), \cdots, p_s(c_k)] * \begin{bmatrix} 0 \\ 0 \\ \cdots \\ 1 \\ \cdots \\ 0 \end{bmatrix} = p_s(c_i). \tag{8}$$

In this way, we extract the individual category prior probability and the predicted probability, and this category must correspond to the true category, suppressing the interference of misinformation. In the following, a purer difficulty weight is obtained by calculating the distance between these two following the method in (6).

$$H\left(p_p^T * \text{true}_{c_i}, p_s^T * \text{true}_{c_i}\right) = -p_p(c_i)\log p_s(c_i), \tag{9}$$

where p_p denotes the prior probability for the category c_i and p_s denotes the predicted probability of the softmax output for the category c_i.

3.1.3. Double-Balanced Loss Function. The working principle of the double-balanced loss function is shown in Figure 2. In the classifier, the model obtains the prediction vector *Output* through the full connection layer, with the size of $[3 * 1]$, representing three kinds of lesions. Then, the model uses the softmax function to normalize this group of values, so that the probability vector *Prediction* of each model can be obtained. Before calculating the loss, the model calculates the quantity weight and the difficulty weight according to equations (4) and (9), respectively. Based on these two weights, the double-balanced loss calculates the distance between the predicted value (Prediction$[3 * 1]$) and the true value as the final loss value, and this loss value is used as a feedback signal passed to the optimizer, and then, the optimizer implements the neural network weight update through the back propagation algorithm.

In this paper, we use cross-entropy as the original loss, which is responsible for calculating the difference between the true value and the predicted value, and use the one-hot coding in the calculation to obtain a concise expression:

$$\text{CE}_{\text{loss}} = -\log p_s. \tag{10}$$

Bringing (10) into (4) yields the loss based on the effective sample.

$$\text{loss}' = -(1 + \alpha_i)\log p_s. \tag{11}$$

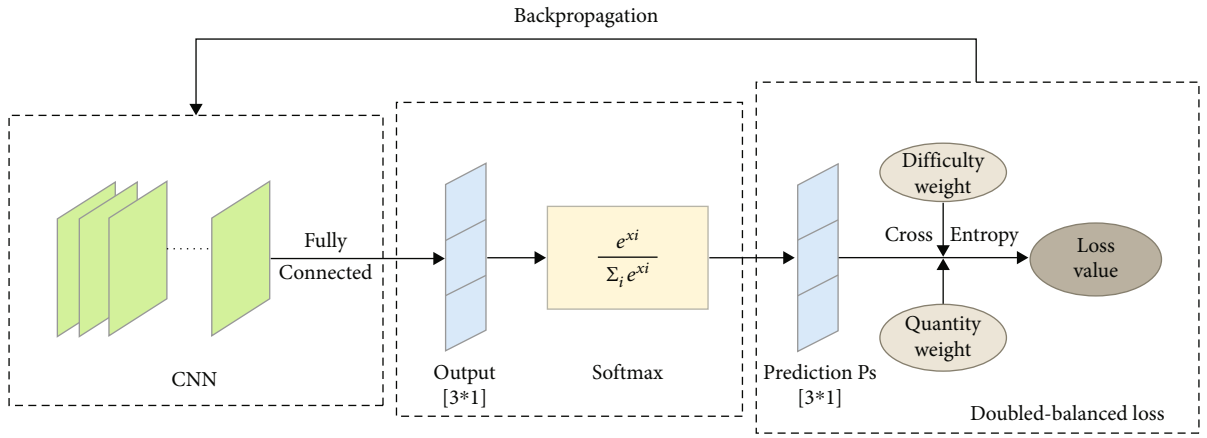


Figure 2: Double-balanced loss working principle, balancing loss in terms of both sample size and sample difficulty.

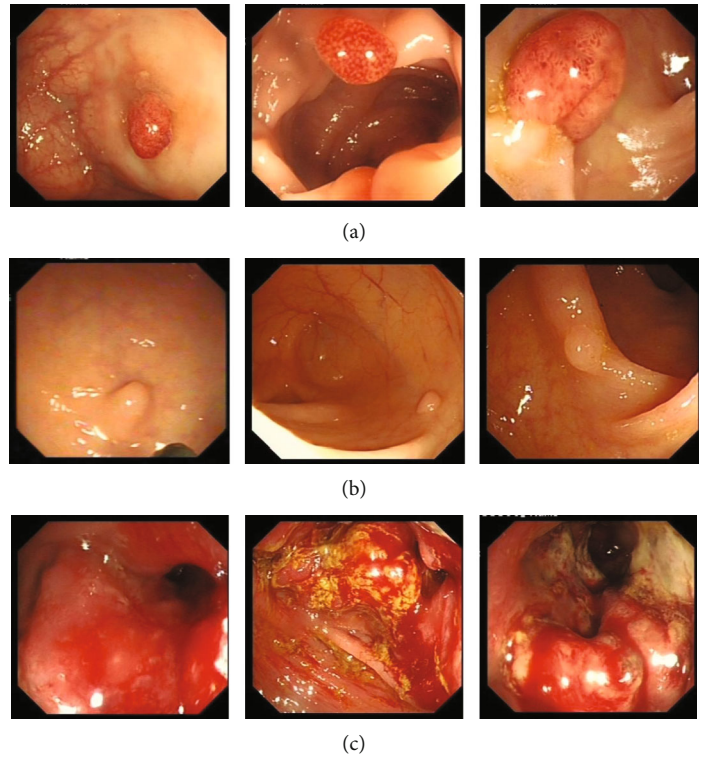


Figure 3: White-light endoscopy images of patients' colonoscopy: (a) adenoma, (b) polyp, and (c) cancer.

Then, the formula for calculating the difficulty weights in (9) is added to (11), and the double-balanced loss function can be obtained as follows.

$$\mathrm{DB}_{\mathrm{loss}} = \left(1 + \frac{(1-\beta)/(1-\beta^{n_i})}{\sum_{j=1}^{k}((1-\beta)/(1-\beta^{n_j}))}\right)\left(\frac{(1/n_i)^{\rho}}{\sum_{j=1}^{k}(1/n_j)^{\rho}}\right)(\log p_s)^2, \tag{12}$$

where β and ρ are hyperparameters, k is the number of categories, and n_i is the number of training samples corresponding to category i.

3.2. Materials. The dataset used in this paper was obtained from white-light endoscopy images of patients' colonoscopy, provided by the Gastrointestinal Endoscopy Center of the East Hospital of Shanghai Sixth People's Hospital. We named this dataset SSPH_WL. The SSPH_WL was collected by doctors under ethical approval. The colorectal lesions were classified into three categories (polyp, adenoma, and cancer) in combination with clinical diagnosis. Sample images of the three categories are shown in Figure 3.

The dataset was collected from June 2015 to September 2019, and a total of 1709 white-light endoscopic images containing lesions were collected and labeled by physicians with more than 5 years' clinical diagnostic experience in

Table 2: The datasets used in this experiment: SSPH_WL-I is used for training, SSPH_WL-II, Kvasir, and CVC is used for test.

Dataset	Used for	Description	Resolution ($w \times h$)	Public
SSPH_WL-I	Train	1367 images	375×347	No
SSPH_WL-II	Test	342 images	375×347	No
Kvasir	Test	428 images	Various resolutions	Yes
CVC	Test	95 images	388×284 574×500	Yes

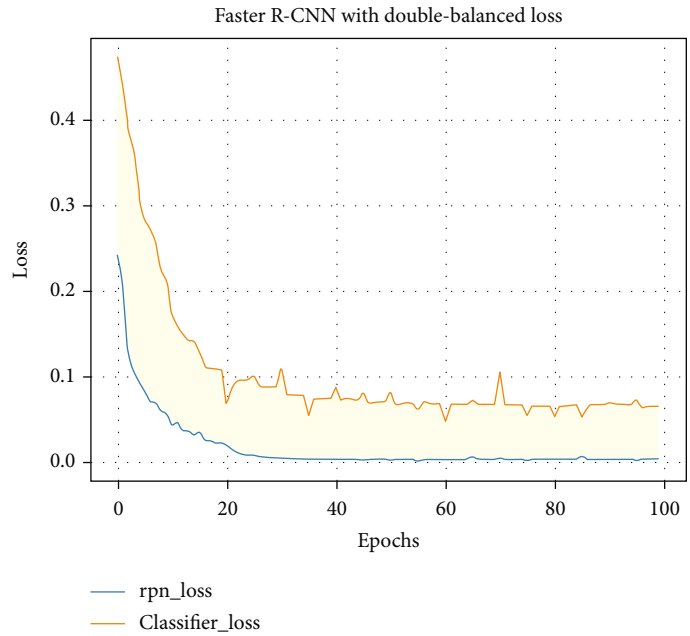


Figure 4: Loss variation curves for training: yellow curve represents RPN loss and blue curve represents classifier loss.

gastroenterology endoscopy. Among them, there were 1048 adenomatous cases, 381 polypoid cases, and 280 cancerous cases. The SSPH_WL was divided into the training set (SSPH_WL-I) and test set (SSPH_WL-II) according to 8 : 2.

We also use colorectal white-light endoscopic images from the public datasets CVC_ClinicDB [23], CVC_ColonDB [24], and Kvasir [25] as the test sets to further evaluate the generalization ability of the classification model. 428 colonoscopic images are selected from the Kvasir data, including 180 images of adenoma, 73 images of cancer, and 175 images of polyp. 95 images are selected from CVC_ClinicDB, CVC_ColonDB and our collection of videos, including 36 images of polyp and 40 images of adenoma, and 19 images of cancer from colorectal cancer videos, named CVC. All of the above datasets were annotated by experienced gastroenterology endoscopy clinicians. The detailed descriptions of these three datasets are shown in Table 2.

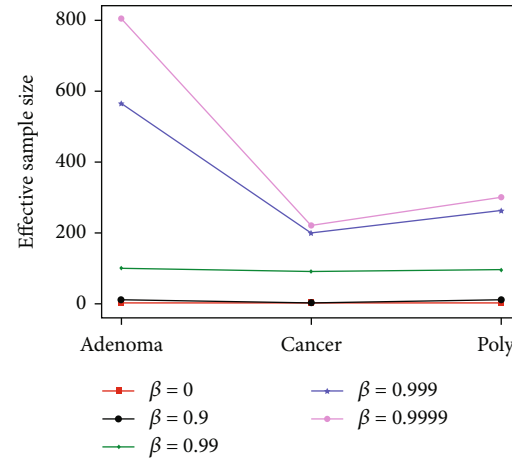


Figure 5: The effective sample size changes with the hyperparameter β.

4. Experimental Results and Discussion

4.1. Training Strategy and Evaluation Metrics. Deep learning-based object detection models can be divided into two categories: two stage and one stage. Due to the complex background of the gut, the two-stage algorithm performs background filtering first to overcome this problem, so we choose the two-stage object detection model Faster R-CNN [8] as the representative and focus on the classification effect

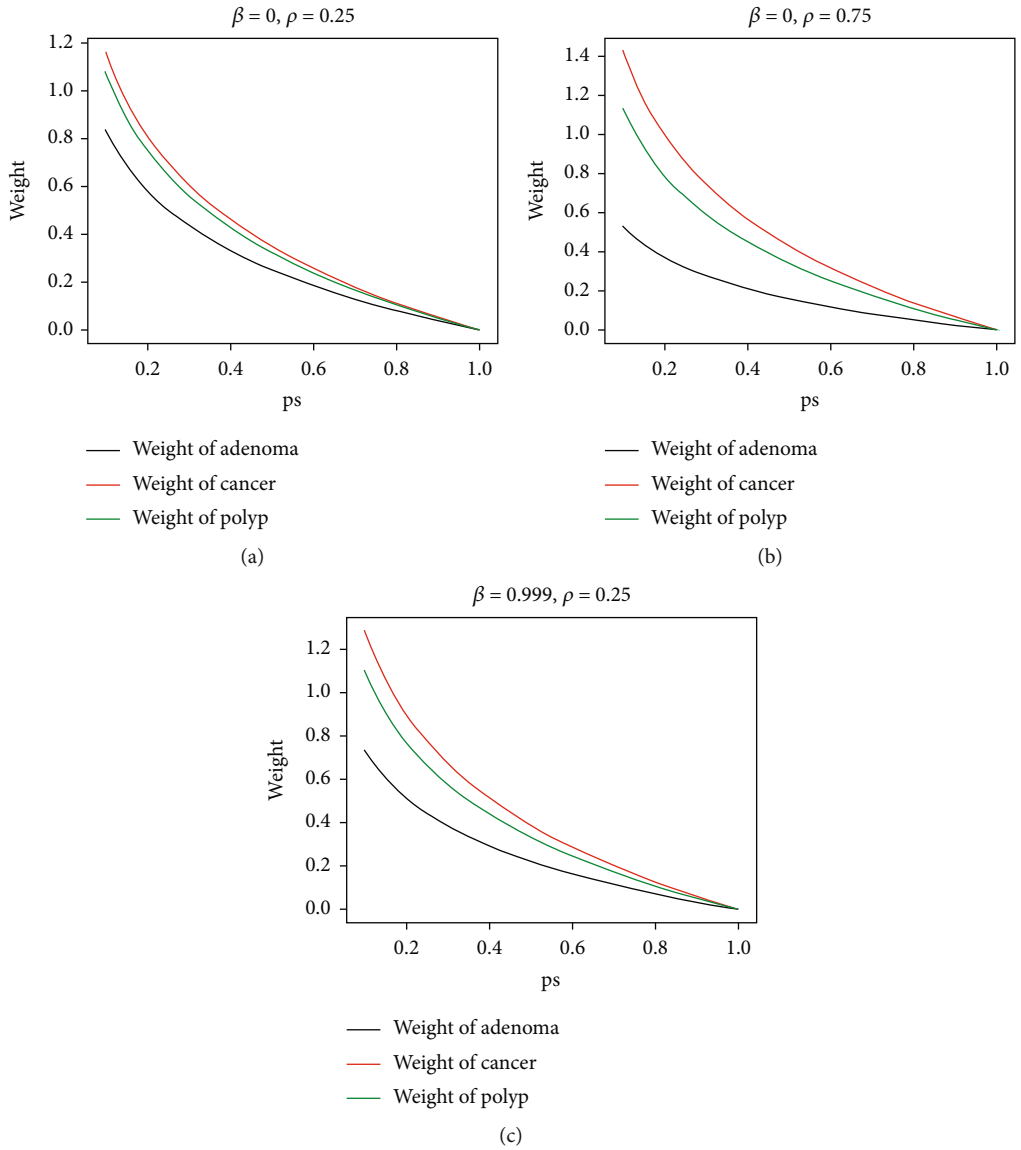


Figure 6: Analysis of β and ρ. Variation curve of weight.

Table 3: AP and AR of Faster R-CNN with different parameter values.

β	ρ	AP	AP_{50}	AP_{75}	AR	AR_{10}	AR_{100}
0.9	0.25	0.578	0.854	0.678	0.669	0.718	0.718
0.99	0.25	0.598	0.858	0.712	0.675	0.731	0.733
0.999	0.25	0.585	0.858	0.696	0.672	0.710	0.711
0.9999	0.25	0.57	0.833	0.694	0.656	0.707	0.707
0.9	0.5	0.587	0.853	0.695	0.662	0.717	0.717
0.99	0.5	0.582	0.838	0.7	0.671	0.726	0.726
0.999	0.5	0.582	0.843	0.675	0.653	0.715	0.716
0.9999	0.5	0.571	0.826	0.679	0.661	0.709	0.709
0.9	0.75	0.579	0.852	0.663	0.669	0.718	0.718
0.99	0.75	0.562	0.836	0.627	0.654	0.704	0.704
0.999	0.75	0.584	0.85	0.667	0.675	0.713	0.714
0.9999	0.75	0.597	0.848	0.717	0.674	0.719	0.719

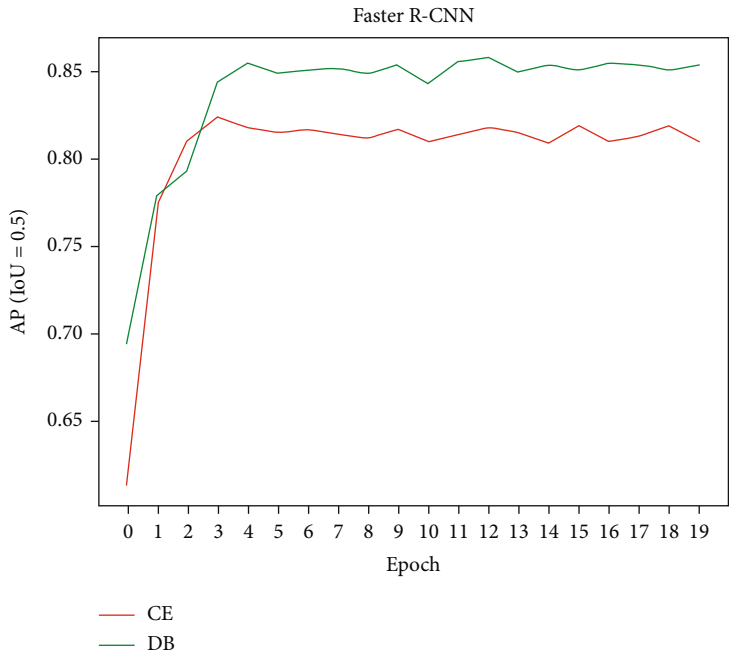


Figure 7: The AP50 (IoU = 0.5) changes with epoch on SSPH_WL-II by using CE and DB.

of the double-balanced loss in this model. The momentum is set to 0.9; the initial learning rate (lr) is 0.005 and automatically decreases by two-thirds every three epochs. For other parameters, we use the default values in the model. For the backbone feature extraction network, we choose ResNet50 based on natural image pretraining, which can make the initial performance of the model higher and converge faster by assisting us to train the model through transfer learning. As shown in Figure 4, Since the loss value no longer decreases, the model reaches the optimum at this time and stops training.

This study is a multilabel classification, and the accuracy is not suitable for evaluating the performance of a model for a single category due to the data imbalance. This paper uses AP and AR as model evaluation indicators. AP (Average Precision) is the area under the P-R curve and can be used to measure the performance of the model for a single category, and AR represents the average recall rate of the classification. These two indicators can be expanded as follows.

AP : AP at IoU = 0.50 : 0.05 : 0.95.

AP_{50} : AP at IoU = 0.50.

AP_{75} : AP at IoU = 0.75.

AR : AR given 1 detection per image.

AR_{10} : AR given 10 detections per image.

AR_{100} : AR given 100 detections per image.

IoU (Intersection-over-Union) represents the ratio of the intersection and the union of the predicted border and the ground truth bound. In addition, we add the missed detection rate (False Negative Rate, FNR) and the wrong detection rate (False Positive Rate, FPR), which are common evaluation indicators for computer-aided diagnosis. These two indicators are calculated as follows.

$$\mathrm{FNR} = \frac{\mathrm{FN}}{\mathrm{TP} + \mathrm{FN}}, \tag{13}$$

Table 4: Comparison of the improvement effects of Faster R-CNN and SSD on SSPH_WL-II.

Model	Loss	AP_{50}	AR_{100}
Faster R-CNN	DB	0.858	0.733
Faster R-CNN	CE	0.824	0.711
SSD	DB	0.835	0.701
SSD	CE	0.821	0.689

Table 5: Comparison of the improvement effects of Faster R-CNN on three test sets.

Dataset	Loss	AP	AP_{50}	AP_{75}	AR	AR_{10}	AR_{100}
SSPH_WL-II	DB	0.598	0.858	0.712	0.675	0.731	0.733
	CE	0.564	0.824	0.638	0.660	0.711	0.711
Kvasir	DB	0.528	0.833	0.552	0.601	0.661	0.661
	CE	0.475	0.750	0.505	0.594	0.652	0.652
CVC	DB	0.639	0.948	0.745	0.677	0.770	0.770
	CE	0.583	0.874	0.649	0.656	0.698	0.698

$$\mathrm{FPR} = \frac{\mathrm{FP}}{\mathrm{FP} + \mathrm{TN}}. \tag{14}$$

In equations (13) and (14), TP is true positive, TN is true negative, FP is false positive, and FN is false negative.

4.2. Experiment on Hyperparameter. The double-balanced loss function has two parameters, where β is responsible for regulating the effective sample size and ρ is responsible for regulating the prior probability of the category. Usually, $\beta \in [0, 1)$; in this paper, the number of training samples of

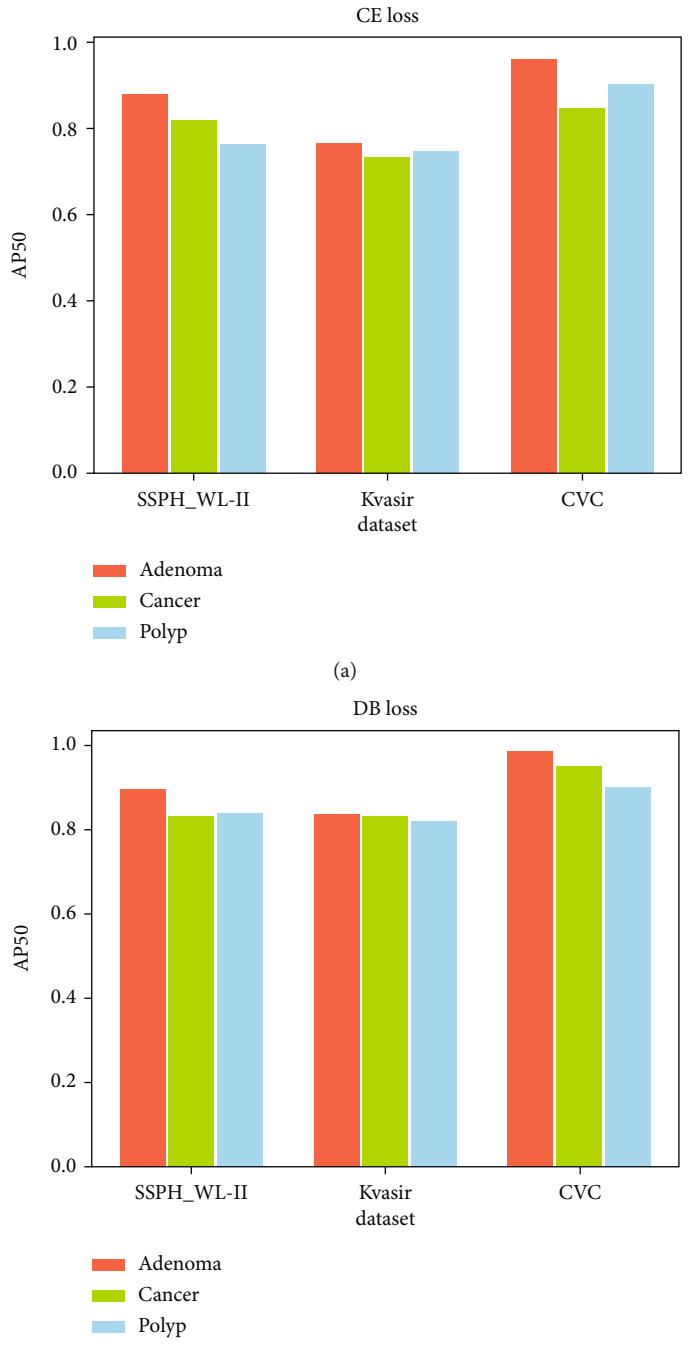


Figure 8: Comparison of classification effect by using CE and DB: (a) cross-entropy loss; (b) double-balanced loss.

each category is {adenoma = 838, cancer = 224, polyp = 305}, and when $\beta \in [0, 0.9)$, the three effective sample sizes calculated from the training samples are almost equal, and the difference of sample sizes cannot be reflected at this time, so $\beta \in [0.9, 1)$ is set. As shown in Figure 5, starting from $\beta = 0.99$, the difference of the effective sample size corresponding to the three categories starts to appear, and as β increases, the effective sample size is closer to the true number. For the parameter ρ, since the number of each category in the training data is fixed, when $\rho = 0$, it means that each category has the same prior distribution, and when $\rho = 1$, it is equivalent to each category having a prior distribution based on the inverse class frequency.

Table 6: Wrong detection rates and missed detection rates.

Loss	FPR	FNR
DB	3.24	3.01
CE	4.97	3.47

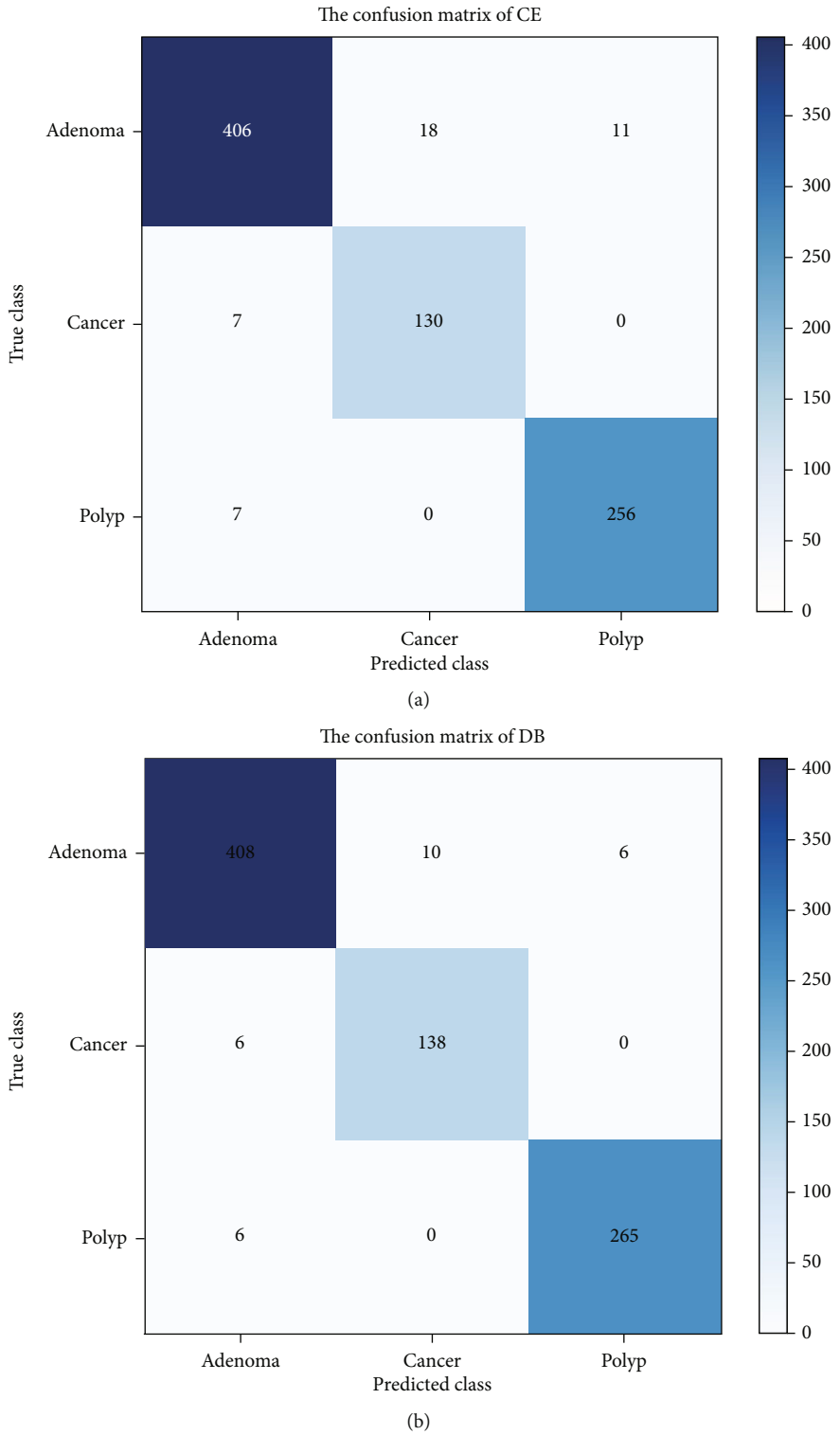
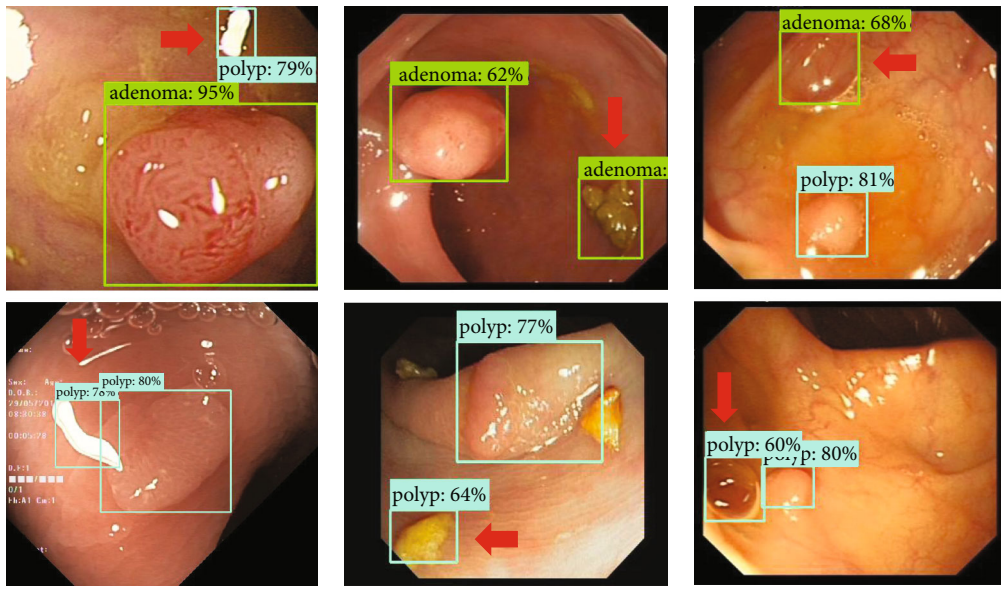


Figure 9: Confusion matrix: (a) cross-entropy loss function; (b)double-balanced loss function.

As shown in Figures 6(a)–6(c), the change curve of the weighted value of the loss function with the prediction probability, looking at each line alone, there is a trend: the higher the predicted probability, the smaller the weighted value, which means that the loss of simple samples is suppressed; looking at the three category curves, the weighted value of adenoma with the largest number of samples is always the smallest, which means that the loss of the multisample category is suppressed. In Figures 6(a) and 6(b), when β is constant, the larger ρ is and the larger the weighted gap between the three categories. In Figures 6(a) and 6(c), when ρ is constant, the larger β is, the larger the weighted gap between the



(a) (b) (c)

Figure 10: Three main categories of disturbances: (a) light spot, (b) foreign matter, and (c) bubble.

three categories will also be. By adjusting hyperparameter β and ρ, the model forms a dynamic weighting mechanism of loss function during training, which can make the model update parameters in a more reasonable way.

In this paper, the comprehensive experiments are conducted for the parameter values, and the corresponding results of different parameter values are shown in Table 3. From the overall performance, the average precision and average recall of detection are almost optimal for $\beta = 0.99$ and $\rho = 0.25$, so we mainly conduct experiments under this parameter value.

4.3. Improved Results on Faster R-CNN. The original loss function used for classification in Faster R-CNN is cross-entropy (CE) loss. Thus, comparing the CE loss, it is clear to see the improvement brought by the DB loss. The training set SSPH_WL-I and test set SSPH_WL-II are from the dataset SSPH_WL, so their data distribution are the same. Figure 7 shows the AP_{50} (IoU = 0.5) changes with epoch on SSPH_WL-II; DB loss can help the model steadily improve the AP_{50} value, where the AP_{50} using the DB loss in Faster R-CNN is stable at around 0.850 and up to 0.858, and the AP_{50} using the CE loss is stable at around 0.820 and up to 0.824. Table 4 compares the detection results of the one-stage model SSD [26]; for the colorectal polyp lesion detection, Faster R-CNN performs better than SSD, and the use of double-balanced loss function can significantly improve the detection performance of Faster R-CNN and SSD.

Table 5 shows a comparison of the results on SSPH_WL-II, Kvasir, and CVC. In these three test sets, DB loss achieves a great lead in all metrics. Kvasir and CVC are used as additional test sets to verify the generalization of DB loss, and the results show that Faster R-CNN trained based on DB loss has a good generalization ability.

In our study, we focus not only on the overall classification level of the model but also on the model's ability to distinguish the three categories. Figure 8 shows the comparison of the classification effect (AP_{50}) of the model on the three test sets. On SSPH_WL-II, the model has the worst classification effect on polyp, and after improvement, the classification precision of polyp is improved the most. On Kvasir and CVC, the model has the worst classification effect on cancer, and after improvement, the classification precision of cancer is improved the most. This indicates that the model can focus on the samples that are not good at classification after improvement, and the classification precision of the three categories is balanced in this way.

We use three test sets to further evaluate the classification ability of the model in terms of the wrong detection rate (FPR) and the missed detection rate (FNR). The wrong detection refers to the model that can locate the lesions but cannot classify them correctly. This is caused by the insufficient classification ability of the model. The missed detection refers to lesions that are not detected by the model, which are filtered out mainly because the model does not locate the lesion or the classification confidence is lower than the set threshold. As shown in Table 6, the FPR and FNR have decreased by using DB loss, indicating that the classification ability of the model has been improved.

The confusion matrices are shown in Figure 9. For the wrong detection, in the original model, 18 adenomas are wrongly classified as cancer, 7 cancers are wrongly classified as adenoma, 11 adenomas are wrongly classified as polyp, and 7 polyps are wrongly classified as adenoma; after using the double-balanced loss, 10 adenomas are wrongly classified as cancer, 6 cancers are wrongly classified as adenoma, 6 adenomas are wrongly classified as polyp, and 6 polyps are wrongly classified as adenoma. It is not difficult to find that the wrong detection never occurs between cancer and

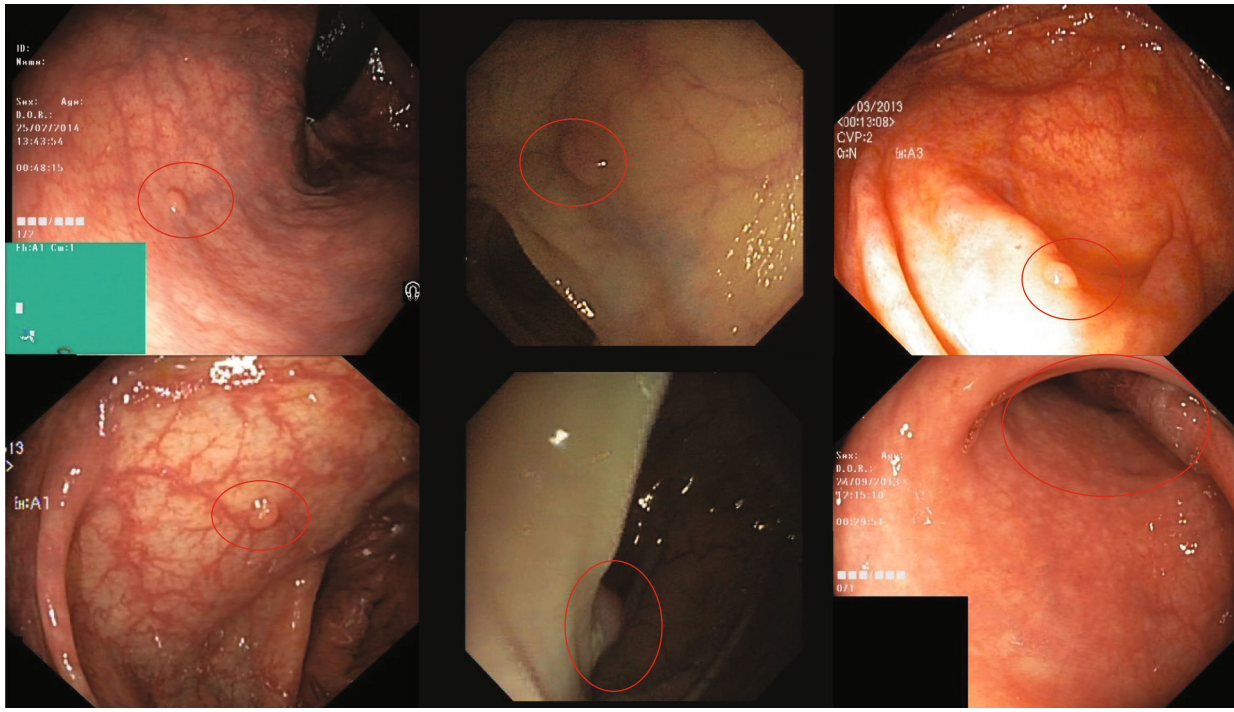


Figure 11: Missed detection images of three test datasets.

Table 7: Performance of different loss functions on the three test sets.

Dataset	Loss	AP	AP_{50}	AP_{75}	AR	AR_{10}	AR_{100}
SSPH_WL-II	DB	0.598	0.858	0.712	0.675	0.731	0.733
	WCE	0.577	0.842	0.669	0.658	0.724	0.724
	FL	0.571	0.813	0.662	0.667	0.719	0.719
	CB	0.574	0.834	0.681	0.661	0.728	0.726
Kvasir	DB	0.528	0.833	0.552	0.601	0.661	0.661
	WCE	0.458	0.764	0.459	0.581	0.637	0.637
	FL	0.453	0.739	0.459	0.581	0.639	0.639
	CB	0.500	0.788	0.570	0.605	0.661	0.661
CVC	DB	0.639	0.948	0.745	0.677	0.770	0.770
	WCE	0.593	0.885	0.668	0.662	0.705	0.705
	FL	0.597	0.864	0.668	0.663	0.711	0.712
	CB	0.619	0.905	0.676	0.672	0.735	0.735

polyp; this is due to the huge difference in characteristics between polyp and cancer.

In addition, some disturbances in the intestine may also lead to wrong detection of the model, as shown in Figure 10 There are three main categories of these disturbances: the first is the identification of the reflected light spots in the intestine as lesions, the second is the identification of foreign matter in the intestine as lesions, and the third is the identification of air bubbles in the intestine as lesions.

As for missed images, DB loss reduces the occurrence of missed detection, but there are still missed detections, as shown in Figure 11. Most of these images are characterized by small target areas, dim light, and occlusion, which may be the main reason for missing model detection.

4.4. Comparison Experiments of Different Loss Functions. We compare the classification effects of weight cross-entropy loss (WCE), multiclassification focal loss (FL), and class-balanced cross-entropy (CB) loss in Faster R-CNN, all of which can be used to solve the imbalance problem.

As shown in Table 7, the classification effects of different loss functions on different test sets are compared. On SSPH_WL-II, the double-balanced loss achieves the advantage in all metrics, and on Kvasir, the double-balanced loss and

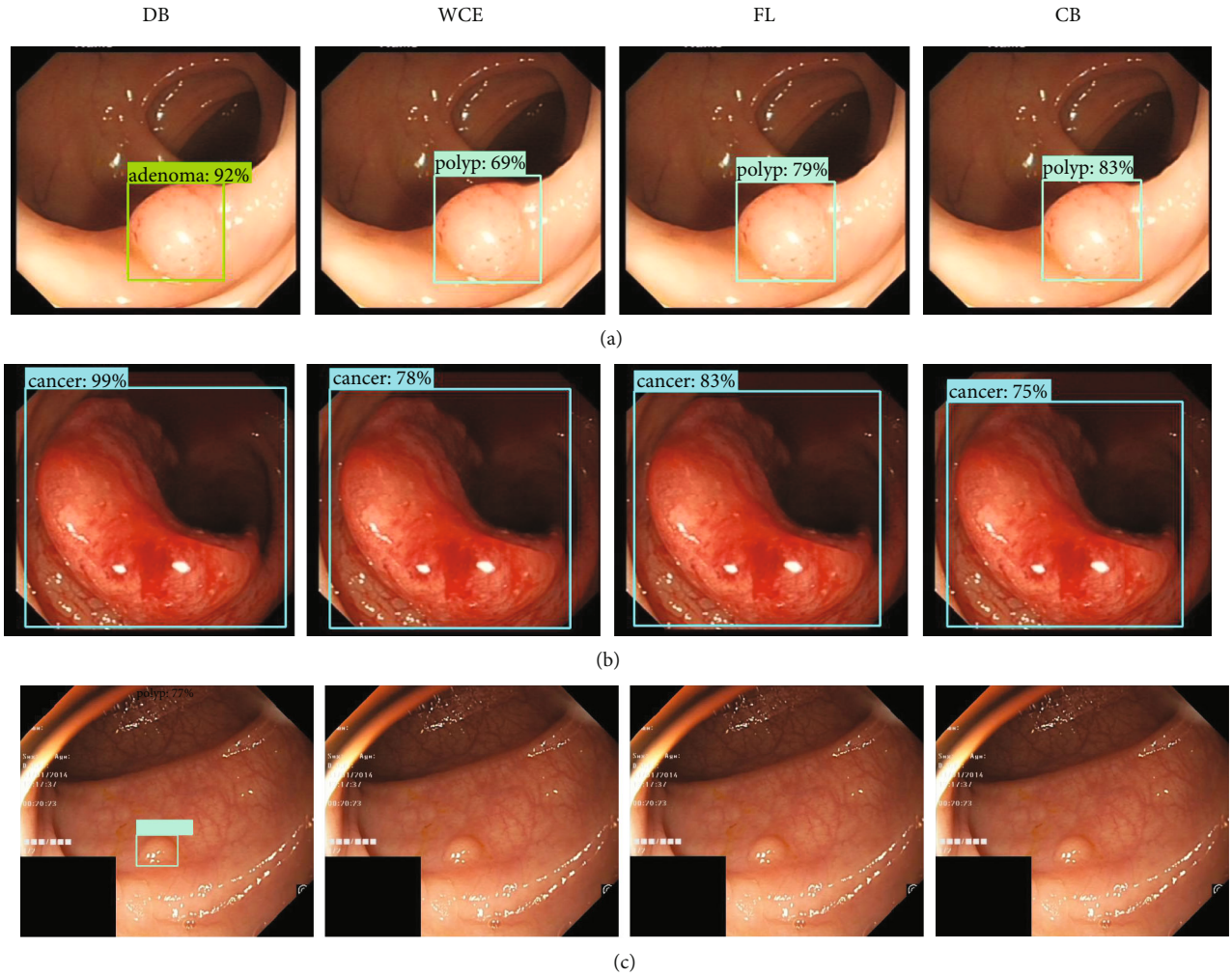


Figure 12: Comparison of different loss function detection results. True category is (a) adenoma, (b) cancer, and (c) polyp.

the class-balanced cross-entropy loss perform the best. On CVC, these four loss functions achieve better results, and the double-balanced loss function comes out ahead in all metrics. Experiments show that the double-balanced loss has good generalization ability on different test sets and is able to solve the data imbalance problem better than other loss functions.

Comparing the classification effects of weight cross-entropy (WCE), multiclassification focal loss (FL), and class-balanced cross-entropy (CB) loss functions in Faster R-CNN, as shown in Figure 12, we can see that the double-balanced loss function proposed in this paper has the following advantages: in example (a), the classification is more accurate; in example (b), the classification confidence is higher; and in example (c), the classification performance is better for small targets.

5. Conclusions

To address the imbalance problem in medical image classification, this paper proposes a new loss function, namely, the double-balanced loss. This loss function improves the classification ability of the model for this part of samples by increasing the focus on fewer sample categories and difficult samples during model training. In this paper, we mainly achieve the double-balanced loss function in Faster R-CNN, and after three test set validations, the model achieves the best detection effect at IoU = 0.5, when the AP values are improved by 3.4%, 8.3%, and 2.9%, indicating that the double-balanced loss function achieves the expected effect on the classification of colorectal white-light endoscopic images. However, there are various types of medical images, and we will verify the effectiveness of the double-balanced loss function on other imbalanced datasets and further promote the double-balanced loss function in the next work.

Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare no competing interests.

Acknowledgments

This research was funded by the Shanghai University of Medicine and Health Sciences Seed Foundation in China, grant number SFP-18-22-14-006.

References

[1] Y. Li, "Colorectal cancer should not be feared proactive screening can prevent and treat it," *Chinese Medical Information Herald*, vol. 36, no. 8, p. 1, 2021.

[2] D. H. Kim et al., "CT colonography versus colonoscopy for the detection of advanced neoplasia," *New England Journal of Medicine*, vol. 358, no. 1, pp. 88–90, 2008.

[3] Y. Komeda, H. Handa, T. Watanabe et al., "Computer-aided diagnosis based on convolutional neural network system for colorectal polyp classification: preliminary experience," *Oncology*, vol. 93, Suppl. 1, pp. 30–34, 2017.

[4] E. Ribeiro, A. Uhl, G. Wimmer, and M. Häfner, "Exploring deep learning and transfer learning for colonic polyp classification," *Computational and Mathematical Methods in Medicine*, vol. 2016, Article ID 6584725, 16 pages, 2016.

[5] J. Gao, Y. Guo, Y. Sun, and G. Qu, "Application of deep learning for early screening of colorectal precancerous lesions under white light endoscopy," *Computational and Mathematical Methods in Medicine*, vol. 2020, Article ID 8374317, 8 pages, 2020.

[6] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, Las Vegas, NV, USA, 2016.

[7] M. Taş and B. Yılmaz, "Super resolution convolutional neural network based pre-processing for automatic polyp detection in colonoscopy images," *Computers and Electrical Engineering*, vol. 90, no. 12, p. 106959, 2021.

[8] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: towards real-time object detection with region proposal networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 6, pp. 1137–1149, 2017.

[9] Y. Shin, H. A. Qadir, L. Aabakken, J. Bergsland, and I. Balasingham, "Automatic colon polyp detection using region based deep CNN and post learning approaches," *IEEE Access*, vol. 6, pp. 40950–40962, 2018.

[10] Y. Shin, H. A. Qadir, and I. Balasingham, "Abnormal colon polyp image synthesis using conditional adversarial networks for improved detection performance," *IEEE Access*, vol. 6, pp. 56007–56017, 2018.

[11] A. Nogueira-Rodríguez, R. Domínguez-Carbajales, H. López-Fernández et al., "Deep neural networks approaches for detecting and classifying colorectal polyps," *Neurocomputing*, vol. 423, pp. 721–734, 2021.

[12] S. A. Karkanis, D. K. Iakovidis, D. E. Maroulis, D. A. Karras, and M. Tzivras, "Computer-aided tumor detection in endoscopic video using color wavelet features," *IEEE Transactions on Information Technology in Biomedicine A Publication of the IEEE Engineering in Medicine & Biology Society*, vol. 7, no. 3, pp. 141–152, 2003.

[13] S. H. Bae and K. J. Yoon, "Polyp detection via imbalanced learning and discriminative feature learning," *IEEE Transactions on Medical Imaging*, vol. 34, no. 11, pp. 2379–2393, 2015.

[14] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: synthetic minority over-sampling technique," *The Journal of Artificial Intelligence Research*, vol. 16, pp. 321–357, 2002.

[15] L. Shen, Z. Lin, and Q. Huang, "Relay backpropagation for effective learning of deep convolutional neural networks," in *European Conference on Computer Vision*, Springer, Cham, 2016.

[16] H. J. Lee and S. Cho, "The novelty detection approach for different degrees of class imbalance," in *Neural Information Processing*, pp. 21–30, Springer, 2006.

[17] S. Zhang, Z. Li, S. Yan, X. He, and J. Sun, "Distribution alignment: a unified framework for long-tail visual recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2361–2370, 2021.

[18] Y. Cui, M. Jia, T. Y. Lin, Y. Song, and S. Belongie, "Class-balanced loss based on effective number of samples," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9268–9277, Long Beach, CA, USA, 2019.

[19] Y. Kim, Y. Lee, and M. Jeon, "Imbalanced image classification with complement cross entropy," *Pattern Recognition Letters*, vol. 151, article S016786552100266X, pp. 33–40, 2021.

[20] A. Shrivastava, A. Gupta, and R. Girshick, "Training region-based object detectors with online hard example mining," in *IEEE Conference on Computer Vision & Pattern Recognition IEEE Computer Society*, pp. 761–769, Las Vegas, NV, USA, 2016.

[21] T. Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 99, pp. 2999–3007, 2017.

[22] B. Li, Y. Liu, and X. Wang, "Gradient harmonized single-stage detector," *Proceedings of the AAAI conference on artificial intelligence*, vol. 33, no. 1, pp. 8577–8584, 2019.

[23] J. Bernal, F. J. Sánchez, G. Fernández-Esparrach, D. Gil, C. Rodríguez, and F. Vilariño, "WM-DOVA maps for accurate polyp highlighting in colonoscopy: validation vs. saliency maps from physicians," *Computerized Medical Imaging and Graphics*, vol. 43, pp. 99–111, 2015.

[24] D. Vázquez, J. Bernal, F. J. Sánchez et al., "A benchmark for endoluminal scene segmentation of colonoscopy images," *Journal of Healthcare Engineering*, vol. 2017, Article ID 4037190, 9 pages, 2017.

[25] D. Jha, P. H. Smedsrud, M. A. Riegler et al., "Kvasir-SEG: a segmented polyp dataset," in *International Conference on Multimedia Modeling*, pp. 451–462, Springer, Cham, 2020.

[26] W. Liu, D. Anguelov, D. Erhan et al., *SSD: single shot multi box detector*, Springer, Cham, 2016.