



# A novel ResNet101 model based on dense dilated convolution for image classification

Qi Zhang<sup>1</sup>

Received: 28 October 2021 / Accepted: 22 November 2021

Published online: 07 December 2021

© The Author(s) 2021 [OPEN](#)

## Abstract

Image classification plays an important role in computer vision. The existing convolutional neural network methods have some problems during image classification process, such as low accuracy of tumor classification and poor ability of feature expression and feature extraction. Therefore, we propose a novel ResNet101 model based on dense dilated convolution for medical liver tumors classification. The multi-scale feature extraction module is used to extract multi-scale features of images, and the receptive field of the network is increased. The depth feature extraction module is used to reduce background noise information and focus on effective features of the focal region. To obtain broader and deeper semantic information, a dense dilated convolution module is deployed in the network. This module combines the advantages of Inception, residual structure, and multi-scale dilated convolution to obtain a deeper level of feature information without causing gradient explosion and gradient disappearance. To solve the common feature loss problems in the classification network, the up- down-sampling module in the network is improved, and multiple convolution kernels with different scales are cascaded to widen the network, which can effectively avoid feature loss. Finally, experiments are carried out on the proposed method. Compared with the existing mainstream classification networks, the proposed method can improve the classification performance, and finally achieve accurate classification of liver tumors. The effectiveness of the proposed method is further verified by ablation experiments.

## Highlights

- (a) The multi-scale feature extraction module is introduced to extract multi-scale features of images, it can extract deep context information of the lesion region and surrounding tissues to enhance the feature extraction ability of the network.
- (b) The depth feature extraction module is used to focus on the local features of the lesion region from both channel and space, weaken the influence of irrelevant information, and strengthen the recognition ability of the lesion region.
- (c) The feature extraction module is enhanced by the parallel structure of dense dilated convolution, and the deeper feature information is obtained without losing the image feature information to improve the classification accuracy.

**Keywords** Image classification · ResNet101 model · Dense dilated convolution · Semantic information

✉ Qi Zhang, 352720214@qq.com | <sup>1</sup>Department of Health Management, Nanyang Medical College, Nanyang 473000, People's Republic of China.



## 1 Introduction

Liver cancer is a malignant tumor of the liver, divided into primary and secondary. Hepatocellular carcinoma is one of the primary liver cancer with high incidence in China. According to the World Health Organization, liver cancer kills nearly 700,000 people in worldwide every year [1, 2]. Computed tomography (CT) is a common medical imaging method for the detection and diagnosis of malignant tumors and is widely used in clinical trials. Diagnosis is usually made by observing CT images of patients. However, due to the complexity of organs and the contrast of surrounding areas, misdiagnosis may occur when the workload is heavy. Therefore, how to achieve accurate classification of liver tumors with computer assistance is still a very challenging task [3].

Medical image classification methods can be divided into traditional machine learning-based classification methods and deep learning based classification methods according to the extraction method of image features. The classification methods based on traditional machine learning mainly use machine learning to classify medical images. Usually, the gray value attribute of the CT image is used to manually extract and study the statistical characteristics of the lesion area, and then build the classification model, finally, realize the image classification. Safia et al. [4] used three statistics-based and two model-based methods respectively to extract texture features from images, and used naive Bayes to classify extracted texture features after using each method individually and in pairs. Gatos et al., [5] combined the four types of image features calculated by traditional methods to obtain comprehensive features. Finally, Support Vector Machine (SVM) was used to classify and determine the optimal feature combination. Traditional machine learning methods generally have the following problems: (1) It takes a long time to extract features from texture, shape and other aspects of the image, and the selection of effective features should be carefully considered; (2) The calculation time of different dimensionality reduction methods is inconsistent with the obtained results, and improper dimensionality reduction methods may cause data redundancy; (3) There are little researches on the depth features of lesion regions, and due to the difference between lesions and regions, it is difficult to classify the lesion regions that are too small with traditional features [6, 7].

The methods based on deep learning solve the problem that traditional machine learning methods need to spend time manually extracting image features and selecting appropriate methods for dimensionality reduction. Meanwhile, advanced features can be obtained.

In recent years, deep learning has become the mainstream method to solve the problem of medical image classification. Zhang et al. [8] proposed a synergic deep learning (SDL) model and used multiple deep convolutional neural networks (DCNN) [9] to solve the problem of image difference within classes and similarity between classes. Firstly, the image features extracted from each pair of DCNNs were connected in parallel and used as the input of SDL, and whether the input image was the same category was predicted through full connection. When one medical image in a pair of DCNN was correctly classified [10], the other classification error resulted in a collaborative error and acted as an additional force to update the model. Using binary data sets and combining images with  $224 \times 224$  pixels, the classification accuracy was improved by 2.1% compared with the benchmark method. In order to distinguish between cysts and metastases, Romero et al. [11] proposed an end-to-end discriminant network framework for the classification of liver lesions, using InceptionV3 [12] to extract features from convolution with different sizes, migrated weights pre-trained on ImageNet. Finally, the pooling operations and auxiliary classifiers were used to improve the convergence of the model. Ghoneim et al. (2020) [13] input images into the convolutional neural network to extract deep-level image features, and then used the Extreme Learning Machine (ELM) classifier to classify the input images, and finally fine-tuned the network. Jiang et al. [14] designed an Attention Hybrid Connection Network architecture which combined soft and hard attention mechanism and long and short skip connections. And a cascade network was proposed based on the liver localization network, liver segmentation network, and tumor segmentation network to cope with this challenge. Seo et al. [15] showed a modified U-Net (mU-Net) with incorporation of object-dependent high level features for improved liver and liver-tumor segmentation in CT images. Bai et al. [16] proposed a liver tumor segmentation method on CT volumes using multi-scale candidate generation method (MCG), 3D fractal residual network (3D FRN), and active contour model (ACM) in a coarse-to-fine manner. Ahmed et al. [17] developed a real-time software-based respiration gating scheme that he implemented on a Verasonics ultrasound imaging system. The above methods are mainly classified by extracting the features of lesion regions in medical images, but have the following shortcomings:

- (1) The learned context information of lesion regions and the depth features of different lesion regions are not effectively used in the network training process;
- (2) Only the overall features of the image are extracted, without focusing on the local features of the focal region;

- (3) It is affected by irrelevant information, no attention is paid to the auxiliary judgment information in the image;
- (4) In the process of training, certain positions, details and other low-level feature information are lost, which reduces the classification accuracy.

The purpose of medical image classification task is to supplement and strengthen the features of lesions according to the contextual information around the lesion area and focus on the area itself. To extract global and local features of the focal area comprehensively and deeply. It needs to reduce the influence of background noise and avoid the loss of details and edge feature information of the lesion area. Therefore, a multi-scale deep feature extraction method based on ResNet101 for liver tumor classification is proposed in this paper. The main contributions are as follows.

- (a) The multi-scale feature extraction module is introduced to extract multi-scale features of images, it can extract deep context information of the lesion region and surrounding tissues to enhance the feature extraction ability of the network.
- (b) The depth feature extraction module is used to focus on the local features of the lesion region from both channel and space, weaken the influence of irrelevant information, and strengthen the recognition ability of the lesion region.
- (c) The feature extraction module is enhanced by the parallel structure of dense dilated convolution, and the deeper feature information is obtained without losing the image feature information to improve the classification accuracy.
- (d) The convolution substitution strategy is used to reduce the number of parameters and enhance the classification performance.
- (e) The up- down-sampling module in the network is improved, and multiple convolution kernels with different scales are cascaded to widen the network, effectively avoiding feature loss.

Section 2 of this paper describes the related work. Proposed image classification method is given in Sect. 3. Section 4 presents the experiments and results. Section 5 gives the conclusion.

## 2 Related works

### 2.1 Atrous convolution

In convolutional neural networks (CNN), with the size of the convolution kernel increasing, the corresponding

receptive field becomes larger. In addition, the number of learning parameters will also increase, resulting in overfitting in the training process. To address these problems, Yu et al. [18] proposed an Atrous convolution that could enlarge the receptive field of a feature graph without resolution loss. This is a special design for intensive forecasting tasks [19].

Atrous convolution is also called extended convolution. Different from conventional convolution, Atrous convolution is a convolution with a dilation rate. When the expansion rate is equal to 1, the atrous convolution can be regarded as the conventional convolution. However, when the expansion rate is greater than 1, the convolution kernel will conduct interval sampling on the feature graph by reducing the expansion rate by 1. The receptive field size  $F$  is calculated as:

$$F = (r - 1) \times (k - 1) + k \quad (1)$$

where  $r$  and  $k$  represent the expansion rate and convolution kernel size respectively. When the convolution kernel size is 3, the comparison of the atrous convolution with different expansion rates is shown in Fig. 1.

In the case that the size of convolution kernel remains unchanged, compared with conventional convolution, atrous convolution does not need to learn more parameters and does not produce information loss, so it can obtain a larger receptive field, which is conducive to enhancing the accuracy of image classification.

### 2.2 Inception

Szegedy et al. [20] applied the Inception module to the GoogLeNet network and achieved the best score in the classification and detection competition. Since then, Inception networks have been continuously improved and innovated for better performance from Inception-V2, V3 to V4 [21, 22]. Inception network is intended to solve the problem of convolutional layer stacking, avoid redundant computing, and make the network deeper and wider. The convolutional kernels of different scales can not only enhance the

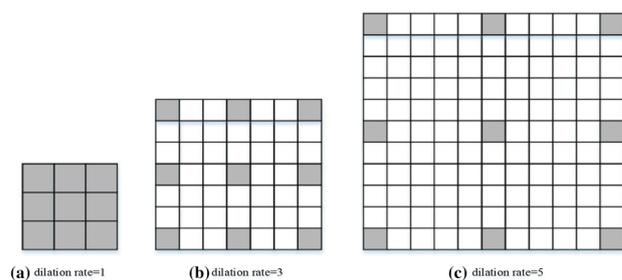


Fig. 1 Atrous convolution with different dilation rates

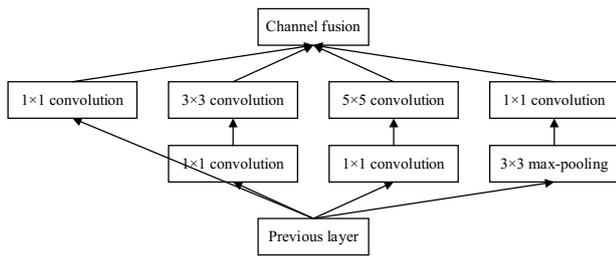


Fig. 2 Single Inception structure

generalization ability and structural expression ability of the network, but also add more nonlinearity to the network model, which greatly improves the feature learning ability of the convolutional neural network. Typically, an Inception network consists of a series of Inception modules.

As shown in Fig. 2, a single Inception module contains three convolutional kernels of different sizes and a maximum pooling layer. It uses a  $1 \times 1$  convolution at each layer for dimensionality reduction to improve computational efficiency. Splicing together the feature maps from these four branches and sending them to the next Inception module enables the network to acquire different receptive fields to increase the network width.

### 2.3 Residual network

For convolutional neural networks, if the network layer is deeper, the training is more difficult.

Because this will not only lead to network degradation, but also be easy to cause gradient disappearance and gradient explosion. In view of this problem, Wu et al. [23] proposed a residual network to enhance feature transmission by introducing shortcut connection into the convolutional neural network. Residual blocks are formed by adding a shortcut every two layers of conventional convolution. Several residual blocks are connected to form a residual network. As shown in Fig. 3,  $x$  is the input of the network.  $F(x)$  represents the output operated by two convolution layers. Before sending to the next layer, the original output will be superimposed with the mapping of quick connection  $F(x) + x$ . In this way, the difficulty of deep network training can be reduced and the network performance can be improved.

## 3 Proposed medical image classification method

In this paper, we use SE-ResNet101 [24] as the basic network architecture, and propose a multi-scale and deep feature extraction model for liver tumor CT image classification. The specific steps are as follows:

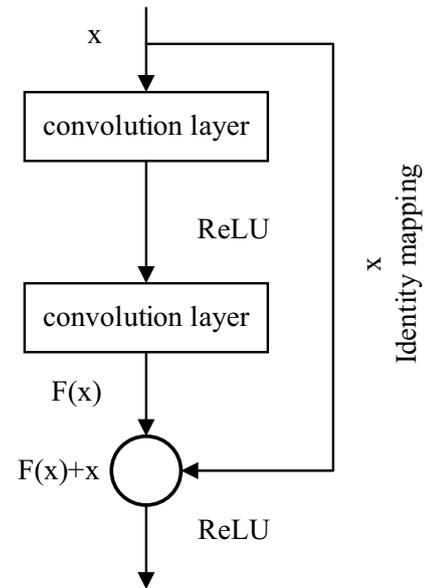


Fig. 3 The structure of the residual block

- (1) By improving the multi-scale expression ability of the network and increasing the receptive field of the network to strengthen the connection between the context of the lesion region.
- (2) By adding the attention mechanism module to extract the local features of the focal region and alleviate the influence of background noise.
- (3) Using the parallel dense cavity convolution to obtain a larger image receptive field, which can extract multi-scale features and retain more original details to improve classification accuracy.
- (4) The up- down-sampling module in the network is improved, and multiple convolution kernels with different scales are cascaded to widen the network, effectively avoiding feature loss.

### 3.1 Multi-scale feature extraction module

The focal regions in medical images appear with different sizes in a single image. Depending on the context information of the focal regions, it can better judge which extracted region of interest (ROI) belongs to. Therefore, Res2Net [25] is used to perceive information at different scales, improve the multi-scale expression ability of the network, and increase the receptive field of each network layer. The original bottleneck uses  $1 \times 1$ ,  $3 \times 3$  and  $1 \times 1$  convolution to map features. The proposed method replaces the original single  $3 \times 3$  convolution with multiple  $3 \times 3$  convolution groups. Specifically, the extracted coarse-grained features from the previous convolution layer are divided into  $s$  parts after a  $1 \times 1$  convolution. Each part represents a feature subset  $a_i$ . And

each feature subset  $a_i$  has the same space size [26]. Then, the original  $3 \times 3$  convolution is replaced by a smaller  $3 \times 3$  convolution group connected by residuals, which is represented by  $m_i(\cdot)$ . In order to reduce the number of parameters and increase the reuse of features, the  $3 \times 3$  convolution after  $a_i$  is omitted. Let  $b_i$  be the output of  $m_i(\cdot)$ . The output of  $m_{i-1}(\cdot)$  is added to the feature subset  $a_i$  and fed into  $m_i(\cdot)$ . So  $b_i$  is denoted as:

$$b_i = \begin{cases} a_i & i = 1 \\ m_i(b_{i-1} + a_i) & 1 < i \leq s \end{cases} \quad (2)$$

where  $i$  is an integer and  $i \in \{1, 2, \dots, s\}$ . Finally, the output  $b_i$  is concatenated through a  $1 \times 1$  convolution. In this paper,  $s=4$ , that is, the feature graph is divided into four parts on average. At this point, the best performance can be obtained, and more feature information can be obtained at different scales and processed efficiently.

### 3.2 Depth feature extraction module

In order to increase the expressiveness of features and suppress unimportant features while paying attention to important features, Woo et al. [27] introduced an attention module to generate attention charts from spatial and channel dimensions. The channel attention module and the spatial attention module are executed in series. The multiplication of the features of spatial attention and the features of the previous attention module through the channel comes from the refinement of adaptive features, which is specified as:

$$F' = A_S(A_C(F) \otimes F) \otimes A_C(F) \otimes F \quad (3)$$

$$F = f^{1 \times 1}[b_1; b_2; b_3; b_4] \quad (4)$$

where  $A_C \in R^{C \times 1 \times 1}$  represents channel attention module.  $A_S \in R^{1 \times H \times W}$  is spatial attention module.  $\otimes$  indicates the element-wise dot product operation between elements.  $F'$  represents the output features through the channel and spatial attention modules.

In the channel attention module,  $F \in R^{C \times H \times W}$  is first processed by global average pooling and global maximum pooling based on width and height, respectively, and also processed by a multi-layer perceptron (MLP). Then, the output features are added based on dot product, and finally the channel attention feature map is generated by activation function. The input  $F \in R^{C \times H \times W}$  of channel attention module is integrated into the spatial attention module. Namely,

$$A_C(F) = \sigma(M(P_{GA}(F)) + M(P_{GM}(F))) \quad (5)$$

where  $P_{GA}$  represents global average pooling.  $P_{GM}$  stands for global maximum pooling.  $M$  is multilayer perceptron.

$\sigma$  denotes sigmoid activation function. In the spatial attention module, feature maps are pooled through two  $1 \times 1$  convolutions respectively, and then activated by sigmoid activation function. In this way, the expression ability of the network can be improved and the problem of spatial information loss in input vector caused by the use of MLP can be solved, that is, Eq. (5) can be expressed as:

$$A_C(F) = \sigma(f^{1 \times 1}(f^{1 \times 1}(P_{GA}(F))) + f^{1 \times 1}(f^{1 \times 1}(P_{GM}(F)))) \quad (6)$$

The output of channel attention is dotted with its input as the input of the spatial attention module. After global average pooling and global maximum pooling based on the channel, concat is performed through the feature graphs generated by the two pooling operations, and then aggregation is performed through a  $7 \times 7$  convolution. Finally, the spatial attention feature map is generated by the activation function. Namely,

$$G = A_C(F) \otimes F \quad (7)$$

$$A_S(G) = \sigma(f^{7 \times 7}[P_{GA}(G); P_{GM}(G)]) \quad (8)$$

where  $f^{7 \times 7}$  represents the convolution with kernel size = 7.  $G$  represents the input of spatial attention module, that is, the obtained feature graph by element-wise dot product of the channel attention module output and its input. By using this module, the effective information of medical image can be further mined without the interference of background noise [28]. That is, it can focus on the in-depth features of detailed lesions needed for correct classification, and then send them to the enhanced feature extraction module.

### 3.3 Dense dilated convolution module

The existing classification task network structure usually outputs the classification prediction results through the convolutional pooling of the full connection layer. If the network structure can acquire receptive fields of different sizes in training, it can capture information of different scales, thus improving the classification accuracy.

U-net network has some common limitations. On the one hand, feature resolution is reduced through continuous pooling and convolutional operation, which often affects prediction tasks requiring very detailed spatial information, such as lung image classification in this paper. On the other hand, if the receptive field is increased by expanding the size of the convolution kernel, the model parameters will increase correspondingly, which is not conducive to model training. Considering the above situation, a dense dilated convolution (DDC) module is added in the middle of the network.

Although DDC looks like a dense connected block, the structure of this module aggregates Inception, residual networks, and dilated convolution. As shown in Fig. 4, the DDC module has four cascaded branches, and the number of dilated convolution in each branch increases one by one, and the corresponding receptive fields are 3, 7, 9 and 19 respectively. At the same time, each branch finally uses a  $1 \times 1$  convolution for ReLU activation. In addition, the module also introduces fast connections in the residual network to fuse original features with other features, avoiding the explosion and gradient disappearance. By combining dilated convolution with different expansion rates, DDC blocks can fully extract features at different scales.

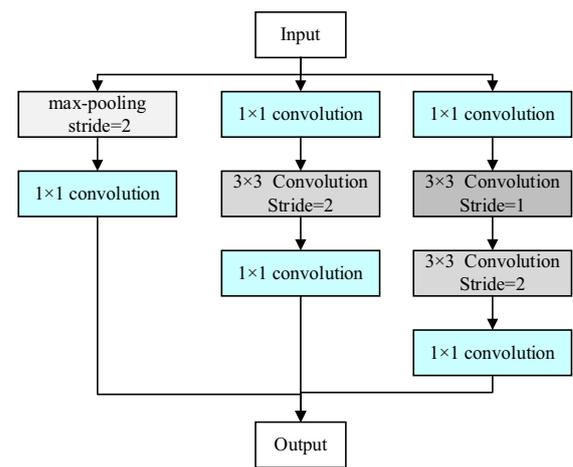
### 3.4 Improved up- down-sampling module

In the standard dense network framework, maximum pooling and up-sampling are used to reduce and increase the resolution of feature maps respectively, but feature loss and precision reduction may occur during training. Therefore, this paper improves the up- down-sampling module to avoid feature loss.

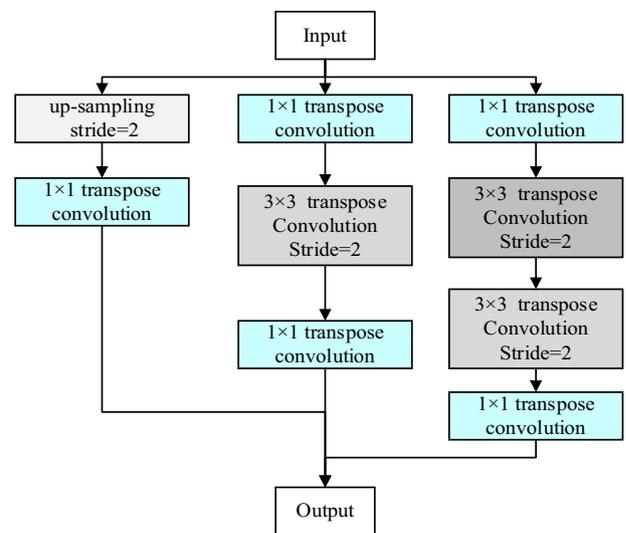
As shown in Fig. 5, the two modules have the same structures. The maximum pooling layer and the convolution layer in the down-sampling module are replaced by the up-sampling layer and the transpose convolution layer in the up-sampling module. This structure is composed of parallel cascades of multiple convolutional kernels with different scales, which can be regarded as a simple Inception structure, it can perceive local features of different scales and improve the learning ability of the network.

### 3.5 Loss function

The loss function can be used to assess the difference between the predicted result and the actual result. When the loss function is small, the robustness of the corresponding model is relatively strong. We replace the



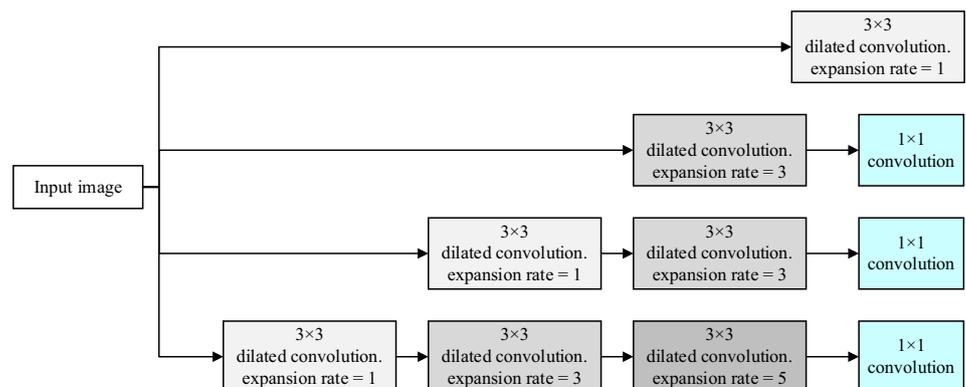
(a) down-sampling



(b) up-sampling

Fig. 5 Down-sample and up-sample block (a down-sample block; b up-sample block)

Fig. 4 DDC structure



commonly used binary cross entropy (BCE) [29] loss function with Dice loss function [30].

Dice similarity coefficient ( $C_{Dice}$ ) is a measurement method to calculate the overlap area between two samples. It ranges from 0 to 1, with 1 indicating complete overlap and 0 indicating no overlap area. Dice loss function is calculated as:

$$loss_{dice} = 1 - C_{Dice} = 1 - \frac{2 \sum_i^N p_i g_i}{\sum_i^N p_i^2 + \sum_i^N g_i^2} \quad (9)$$

where N represents the number of pixels.  $g_i \in \{0, 1\}$  represents whether pixel  $i$  is the label of the foreground, while  $p_i \in (0, 1)$  represents the prediction result of pixel  $i$  output by the Softmax layer.

### 3.6 Overall network architecture

The flowchart of the proposed method is shown in Fig. 6. The network consists of three  $3 \times 3$  convolutions, four blocks, DDC and Softmax. The input layer of the original SE\_ResNet101 is a  $7 \times 7$  convolution. In order to obtain the same receptive field with  $7 \times 7$  convolution and capture more features, it is replaced with three  $3 \times 3$  convolutions, which can reduce the number of parameters and learn more distinguishing features of the lesion edge region. In order to better extract the multi-scale and depth features of the lesion region, expand the convolution layer receptive field, avoid the loss of image features and alleviate the influence of background noise, this paper combines the multi-scale feature extraction module (MSFE), depth

feature extraction module (DFE) and DDC model to optimize the network bottleneck of SE\_ResNet (The detailed bottleneck is shown in Fig. 7).

$$B = \sigma(f_c(\delta(f_c(P_{GA}(F'')))))) \otimes F'' + X_t \quad (10)$$

In the formula, B represents the output through the bottleneck.  $f_c$  is the full connection layer.  $\delta$  represents ReLU activation function.  $X_t \in R^{C \times H \times W}$  represents the original feature map before feeding into the bottleneck.

The optimized bottleneck is composed of multi-scale feature extraction module, depth feature extraction module and DDC. In the training process, octave convolution is used to replace the ordinary convolution to reduce resource consumption and improve classification accuracy [31]. The network structure and parameter settings are shown in Table 1. C, AP, MP, and FC represent convolution layer, average pooling, maximum pooling and fully connected layer respectively. The following parameters are kernel size, step size and fill value in order.

## 4 Experiment and analysis

The experimental data set consists of 5200 abdominal CT images. The data comes from <https://competitions.codalab.org/competitions/17094>. In this paper, we only select 190 patients with portal vein CT scan (four lesions), including 35 metastasis (MET), 40 hemangiomas (HEM), 62 hepatocellular carcinoma (HCC) and 53 healthy tissues. The radiologist

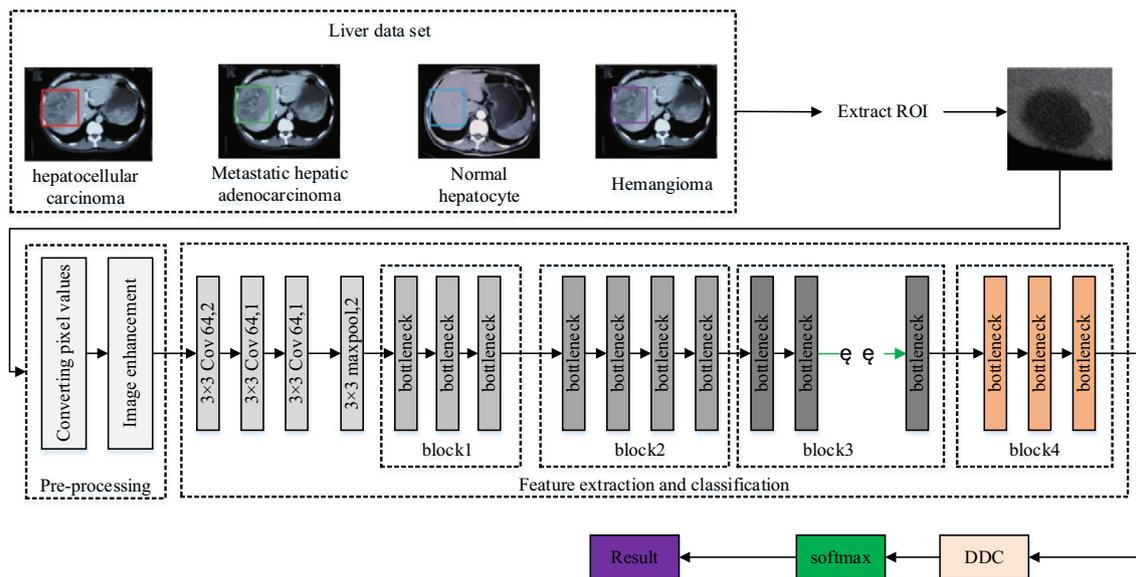


Fig. 6 Overview of the proposed method

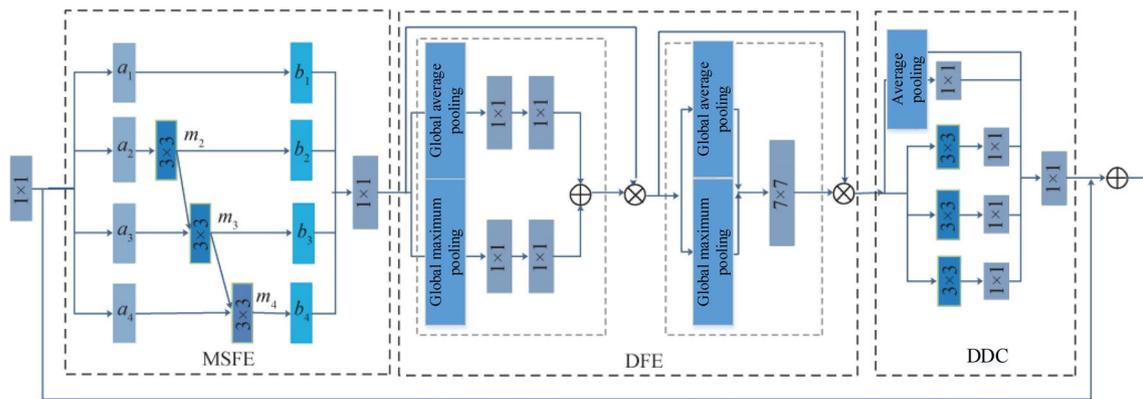


Fig. 7 Detailed bottleneck structure

Table 1 Network structure and parameter setting

Layer	Parameter			
Previous layer	C:3,S:2,P:1 BatchNorm	C:3,S:1,P:1 BatchNorm	C:3,S:1,P:1 BatchNorm ReLU	MP:3,S:2,P:1
Bottleneck	MSFE DFE DDC:256	MSFE DFE DDC:512	MSFE DFE DDC:1024	MSFE DFE DDC:2048
Latter layer	DDC:4	Softmax		

marks the edges of each lesion and confirms the corresponding diagnosis by biopsy or clinical follow-up. The data samples are shown in Fig. 8.

### 4.1 Experimental preprocessing

All data samples are preprocessed as follows:

- (1) ROI extraction. The required ROI is extracted from the liver tumor profile marked by an experienced radiologist, in order to increase the diversity of samples, the ROI of a healthy liver is delineated by the physician from healthy liver tissue.
- (2) Convert pixel values. The unit of CT value is Hounsfield (HU) and the range is (-1024,3071), which reflects the degree of X-ray absorption of tissues. Digital imaging and communications in medicine (DICOM) format, the image range is usually (0,4096). The RI (Rescale Intercept) and RS (Rescale Slope) need to be read from the DICOM header file during conversion. The conversion relationship between CT value and pixel value can be expressed as:

$$HU = PV \times RS + RI \tag{11}$$

where, PV represents pixel value. In this experiment, RS = 1 and RI = -1024 are taken in this paper [33].

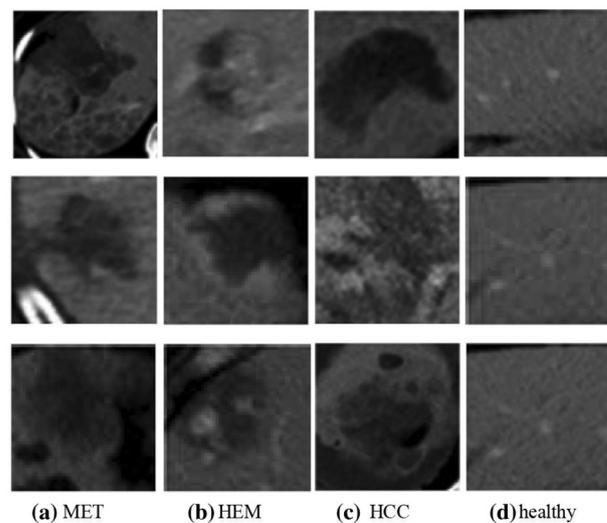


Fig. 8 Typical samples of liver dataset [32]

- (3) Image enhancement. In order to enhance the features of lesions and improve the generalization ability of the network, the extracted ROI is enhanced by random flipping, filling, clipping and affine transformation.

After pre-processing, the data set is randomly divided into the training set and testing set, the training set accounts for 80% of the total samples, and the testing set accounts for 20% of the total samples. The sample size is trimmed to  $64 \times 64$  pixels, and the data distribution is shown in Table 2. In the experiment, due to the limitation of graphics processing unit (GPU) graphics memory, the batch size is set as 16 and Adam optimization algorithm [34] is adopted in this paper. The initial learning rate is set as 0.002, the exponential decay rate is set as 0.98, and the decline period is set as 1. Experiments are carried out on

the PyTorch framework with NVIDIA GeForceGTX 1060 Ti GPU for all experiments to verify the validity of the proposed method.

## 4.2 Evaluation index

Accuracy (A), recall (R), precision (P), F1 score and receiver operating characteristic curve (ROC) are adopted to evaluate the classification results. They are defined as:

$$A = \frac{TP + TN}{TP + TN + FP + FN} \quad (12)$$

$$R = \frac{TP}{TP + FN} \quad (13)$$

$$P = \frac{TP}{TP + FP} \quad (14)$$

$$F1 = \frac{2TP}{2TP + FP + FN} \quad (15)$$

$$TPR = \frac{TP}{TP + FN} \quad (16)$$

$$FPR = \frac{FP}{TN + FP} \quad (17)$$

TP represents the number of correctly predicted tumor. TN represents the number of correctly predicted background. FP represents the sample number that the background is predicted as a tumor. FN represents the sample number that the tumor is predicted as background. The value is in the range [0,1], and the larger value denotes the better effect. FPR is the false positive rate, which refers to the probability of being correct in all other categories. TPR is the true positive rate, which refers to the probability that is actually predicted under this category. ROC is measured according to FPR and TPR. Area under the curve (AUC) is the area under the ROC curve [35]. If the value of AUC is close to 1, the authenticity of the testing method will be higher.

**Table 2** Distribution of the liver dataset

Type	MET	HEM	HCC	healthy
Training set	1300	662	1470	664
Testing set	325	165	366	165
Total	1625	827	1836	829

## 4.3 Comparison experiments

### 4.3.1 A. Comparison with the benchmark method.

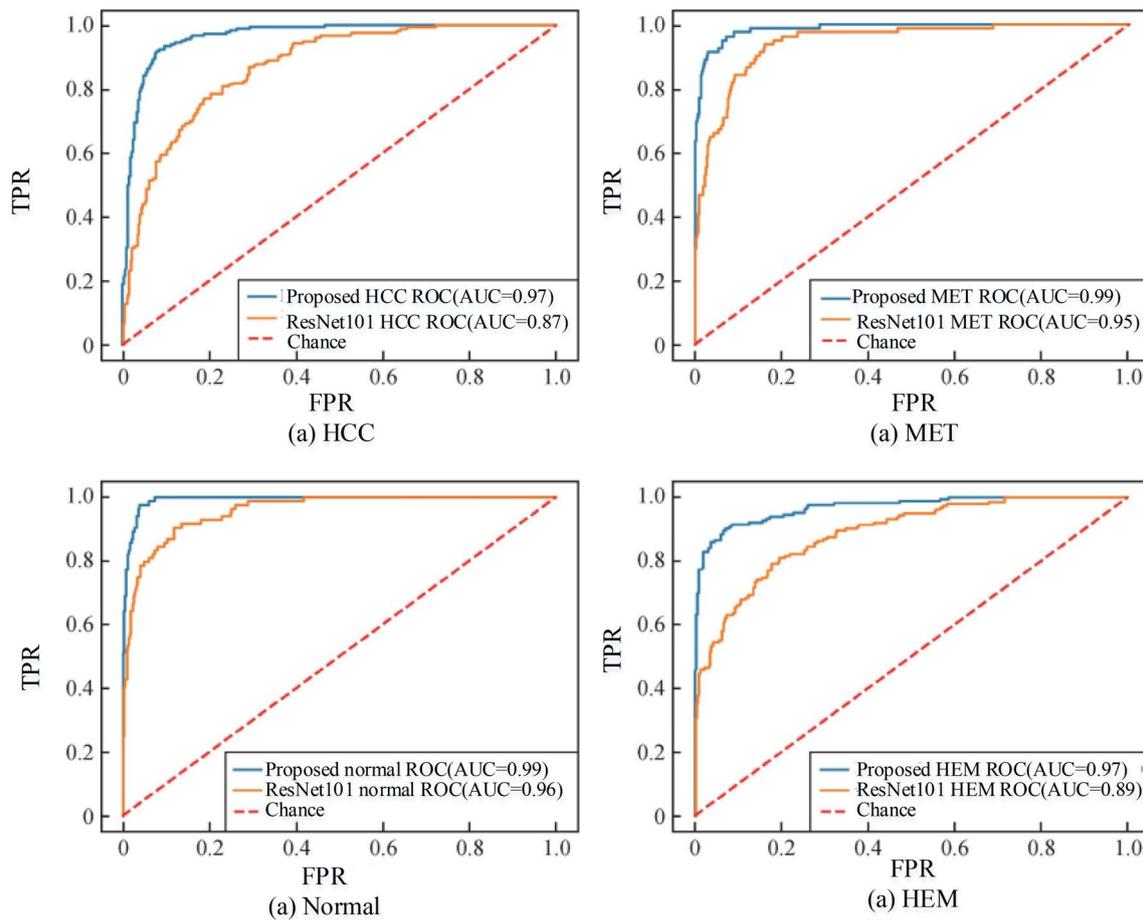
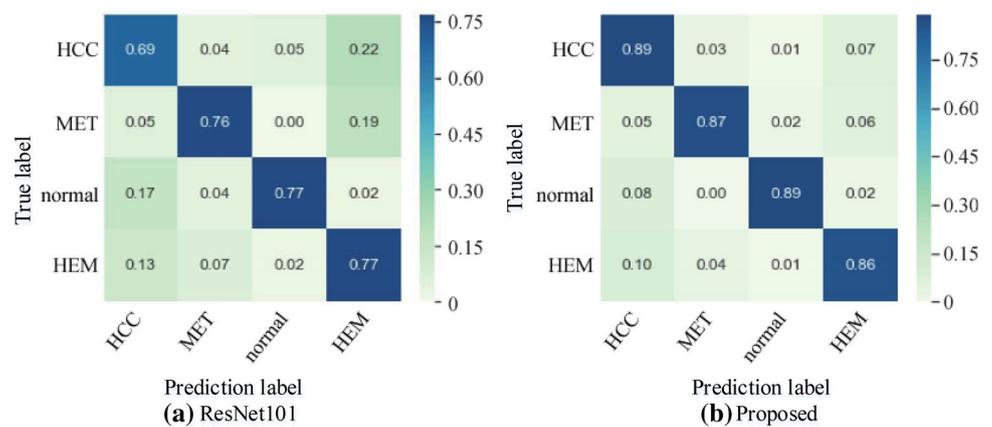
ResNet101 is used as the benchmark experiment. The multi-scale feature extraction module, depth feature extraction module and DDC module are added in ResNet101, and the octave convolution is used to replace ordinary convolution. Figure 9 is the confusion matrix comparison of HCC, MET, normal and HEM between the proposed method and the benchmark method. The x-axis represents the true label and the y-axis indicates the predicted label. Each row represents the probability of a correct prediction (recall). Each column represents the probability of a correct prediction in the case of the prediction label, that is, precision. As can be seen from Fig. 9, the recall and accuracy of the proposed method for each type of liver lesions are higher than or equal to that of the ResNet101, and each type of sample is balanced. It shows that the proposed algorithm can achieve better overall classification, enhance the feature extraction ability of the network, alleviate the influence caused by background noise, and strengthen the utilization of image features. However, due to the similarity of lesion areas being too close, there are still some misclassifications in the lesion types.

The ROC curves of different liver lesions and the ROC between the proposed method and benchmark model are shown in Figs. 10 and 11. It can be seen that the ROC curve obtained by the proposed method is more stable and smooth than the benchmark performance. It has a larger AUC, indicating that the proposed method in this paper has a better classification effect.

### 4.3.2 Ablation experiments

The proposed method mainly consists of multi-scale feature Extraction module (MSFE), deep feature extraction module (DFE), dense dilated convolution module (DDC) and convolution substitution strategy (CSS). Ablation experiments are performed on the same data set to verify the validity of each component. The results are shown in Table 3. It can be seen that by using MSFE, the classification accuracy is improved by 2.92%, indicating that strengthening the multi-scale extraction ability of the network and obtaining more context information can effectively improve the classification accuracy. Combined with MSFE and DFE, the accuracy is improved by 2.94% compared with the MSFE. It can be seen that the depth feature extraction module can strengthen the effective utilization of the features of the lesion region and reduce the impact of background noise on the classification task. The accuracy of the combination

**Fig. 9** Comparison of confusion matrix between baseline method and proposed method



**Fig. 10** ROC curves comparison of each type between ResNet101 model and proposed model

of MSFE, DFE and DDC is 2.7% higher than that of MSFE and DFE. DDC module enlarges the receptive field and increases the multi-scale features of the network during training, thus reducing the loss of accuracy. Combining

MSFE, DFE, DDC and CSS, the accuracy is improved by 1.36%, and the number of parameters is reduced by 3.67 M, which proves that the combination of CSS can effectively improve the classification accuracy while reducing redundancy

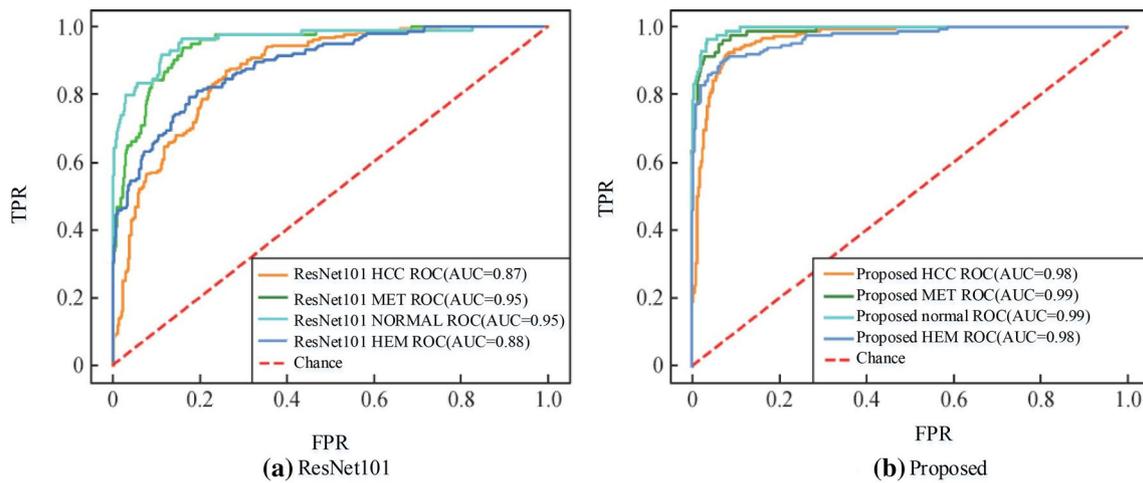


Fig. 11 ROC comparison between ResNet101 and proposed model

Table 3 Ablation experiments on liver dataset

Method	A	R	P	F1	AUC	Parameter size/M
ResNet101	77.93	77.89	78.86	78.86	91.11	49.23
MSFE	80.85	81.87	79.46	80.45	94.11	<b>34.64</b>
MSFE + DFE	83.79	83.11	84.47	83.96	95.11	56.25
MSFE + DFE + DDC	86.49	86.86	86.18	86.49	96.11	58.94
MSFE + DFE + DDC + CSS	<b>87.85</b>	<b>86.15</b>	<b>87.11</b>	<b>86.53</b>	<b>96.11</b>	55.27

The bold value denotes the best result

### 4.3.3 Comparison with classical networks

To further verify the classification performance, the proposed method is compared with DenseNet [36], ResNet101, MnasNet [37], MobileNet2 [38], ShuffleNetV2 [39], SK\_ResNet101 [40] and SE\_ResNet101 are compared under the same data sets. The results are shown in Table 4. It can be seen that by enhancing the multi-scale and depth features of the lesion region, the influence of background noise can be weakened, the utilization of useful information can be increased, and the classification performance can be improved without losing accuracy.

Figure 12 shows the overall ROC curve of the proposed method and the existing classical classification algorithms. It can be seen that, among all the classical classification models, the ROC curve of the proposed method has a smaller and smoother fluctuation range, and a larger AUC is obtained than that of other classical classification network models, which further proves the superiority of the proposed algorithm.

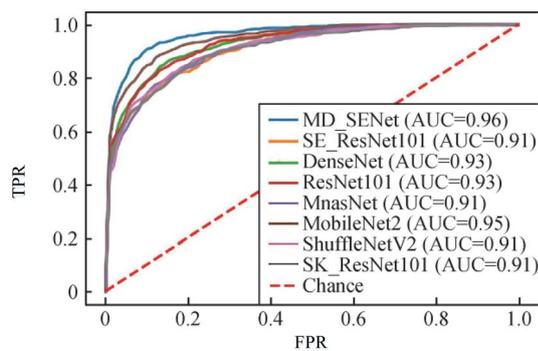
Table 4 Classification performance of the liver dataset

Method	A	R	P	F1	AUC
DenseNet	79.69	79.28	78.44	79.61	93.11
ResNet101	77.74	76.51	77.58	76.39	93.11
MnasNet	73.65	73.68	73.73	73.73	91.11
MobileNet2	81.24	81.16	80.42	80.43	95.11
ShuffleNetV2	76.18	76.67	77.00	77.01	91.11
SK_ResNet101	81.43	82.33	81.14	81.61	91.11
SE_ResNet101	77.93	77.89	78.86	78.86	91.11
Proposed	<b>87.85</b>	<b>86.15</b>	<b>87.12</b>	<b>86.54</b>	<b>96.21</b>

The bold values are the best results

## 5 Conclusion

To solve the problems of the existing classification methods, such as the insufficient mining of context information, the influence of background noise and the loss of image feature information, this paper proposes a novel ResNet101 model based on dense dilated convolution



**Fig. 12** Comparison of ROC curves with different methods

for medical liver tumors classification. The improvements are made in the following four aspects: (1) The multi-scale feature extraction module is used to strengthen the connection between the context information of the lesion region, and the semantic information can be fully mined while increasing the receptive field; (2) The depth feature extraction module is used to enhance the features of the lesion region and reduce the influence of background noise, so as to pay deep attention to the useful lesion information; (3) Dense dilated convolution is connected in parallel so that feature maps with different scales can be sampled to expand the receptive field of the network and improve feature utilization of the original image; (4) The up- down-sampling module in the network is improved, and multiple convolution kernels with different scales are cascaded to widen the network, effectively avoiding feature loss. Through the improvement of the above aspects, the accurate classification of liver tumors can be achieved. The superiority of the proposed method is verified on the liver data set, and the optimal performance is achieved under several evaluation indexes. Compared with the mainstream classical networks, the classification effect of the proposed method is better than that of the classical networks. In the future, it will be extended to more medical fields. As a method of early lesion detection, it will be of far-reaching significance to assist doctors in diagnosis and treatment. The proposed method in this paper still has some shortcomings. Since the collected organs CTs in different periods, such as venous phase, arterial phase, delayed phase and plain scanning, are different. The classification results will decline. Therefore, how to use CT values in different periods to accurately classify tumors remains to be studied.

**Data availability** The data that support the findings of this study are available from the corresponding author, Qi Zhang, upon reasonable request.

## Declarations

**Conflicts of interests** The author declares that there are no conflicts of interest.

**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

## References

- Xie DP, Gong YX, Jin YH et al (2020) Anti-tumor properties of Picrasma quassioides Extracts in H-Ras G12V Liver Cancer are mediated through ROS-dependent Mitochondrial Dysfunction[J]. *Anticancer Res* 40(7):3819–3830
- Zhu Y, Wang Siqi et al (2021) Integrative analysis of long extracellular RNAs reveals a detection panel of noncoding RNAs for liver cancer[J]. *Theranostics* 11(1):181–193
- Naeem S, Ali A, Qadri S et al (2020) Machine-learning based hybrid-feature analysis for liver cancer classification using fused (MR and CT) Images[J]. *Appl Sci* 10(9):3134
- Safia A, He DC (2013) New brodatz-based image databases for grayscale color and multiband texture analysis[J]. *Isrn Mach Vision* 2013:1–14
- Gatos I, Tsantis S, Karamesini M et al (2015) Development of a support vector machine - based image analysis system for focal liver lesions classification in magnetic resonance images[J]. *J Phys Conf* 633(1):012116
- Yin S, Bi J (2019) Medical image annotation based on deep transfer learning[J]. *J Appl Sci Eng* 22(2):385–390
- J Bi, A Shoulin, Yin N-S (2018) A new graph semi-supervised learning method for medical image automatic annotation[C]. In: *IEEE International Congress on Cybermatics-Things Halifax, NS, Canada, Canada* [https://doi.org/10.1109/Cybermatics\\_2018.2018.00041](https://doi.org/10.1109/Cybermatics_2018.2018.00041)
- Zhang J, Xie Y, Wu Q et al (2019) Medical image classification using synergic deep learning[J]. *Med Image Anal* 54:10–19
- Litjens G, Kooi T, Bejnordi BE et al (2017) A survey on deep learning in medical image analysis[J]. *Med Image Anal* 42(9):60–88
- Shin HC, Roth HR, Gao M et al (2016) Deep convolutional neural networks for computer-aided detection: CNN architectures, dataset characteristics and transfer Learning[J]. *IEEE Trans Med Imaging* 35(5):1285–1298
- Romero F P, Diler A, Bisson-Gregoire G, et al. (2019) End-To-End discriminative deep network for liver lesion classification[C]// 2019 IEEE 16th International symposium on biomedical imaging (ISBI). IEEE

12. Zhao Y, Xie K, Zou Z, He J-B (2020) Intelligent recognition of fatigue and sleepiness based on inceptionV3-LSTM via multi-feature fusion. *IEEE Access* 8:144205–144217. <https://doi.org/10.1109/ACCESS.2020.3014508>
13. Ghoneim A, Muhammad G, Hossain MS (2020) Cervical cancer classification using convolutional neural networks and extreme learning machines - ScienceDirect[J]. *Future Gen Comp Syst* 102:643–649
14. Jiang H, Shi T, Bai Z, Huang L (2019) AHCNet: An application of attention mechanism and hybrid connection for liver tumor segmentation in CT volumes. *IEEE Access* 7:24898–24909. <https://doi.org/10.1109/ACCESS.2019.2899608>
15. Seo H, Huang C, Bassenne M, Xiao R, Xing L (2020) Modified U-Net (mU-Net) with incorporation of object-dependent high level features for improved liver and liver-tumor segmentation in CT images. *IEEE Trans Med Imaging* 39(5):1316–1325. <https://doi.org/10.1109/TMI.2019.2948320>
16. Bai Z, Jiang H, Li S, Yao Y (2019) Liver tumor segmentation based on multi-scale candidate generation and fractal residual network. *IEEE Access* 7:82122–82133. <https://doi.org/10.1109/ACCESS.2019.2923218>
17. Ahmed R, Ye J, Gerber SA, Linehan DC, Doyley MM (2020) Preclinical imaging using single track location shear wave elastography: monitoring the progression of murine pancreatic tumor liver metastasis in vivo. *IEEE Trans Med Imaging* 39(7):2426–2439. <https://doi.org/10.1109/TMI.2020.2971422>
18. F Yu, Koltun V. 2016 Multi-scale context aggregation by dilated convolutions[C]// ICLR.
19. Chen Y, Zhao H, Hu Z et al (2021) Attention-based context aggregation network for monocular depth estimation[J]. *Int J Mach Learn Cybern* 11:1–14
20. Szegedy C et al (2015) Going deeper with convolutions. *IEEE Conf Comp Vision Pattern Recog (CVPR)* 2015:1–9. <https://doi.org/10.1109/CVPR.2015.7298594>
21. C Szegedy, V Vanhoucke, S Ioffe, J Shlens, Z Wojna, (2016) Rethinking the inception architecture for computer Vision. In: *Conference on computer vision and pattern recognition (CVPR)*, IEEE, pp 2818–2826, <https://doi.org/10.1109/CVPR.2016.308>
22. Szegedy C, Ioffe S, Vanhoucke V, et al. (2016) Inception-v4, Inception-ResNet and the impact of residual connections on learning[J]. [arXiv:1602.07261](https://arxiv.org/abs/1602.07261)
23. Wu S, Zhong S, Liu Y (2017) Deep residual learning for image steganalysis[J]. *Multimed Tools Appl* 77(9):10437–10453
24. Hu J, Shen L, Albanie S, Sun G, Wu E (2020) Squeeze-and-excitation networks. In *IEEE Trans Pattern Anal Mach Intell* 42(8):2011–2023. <https://doi.org/10.1109/TPAMI.2019.2913372>
25. Gao S-H, Cheng M-M, Zhao K, Zhang X-Y, Yang M-H, Torr P (2021) Res2Net: A new multi-scale backbone architecture. In *IEEE Trans Pattern Anal Mach Intell* 43(2):652–662. <https://doi.org/10.1109/TPAMI.2019.2938758>
26. Yin S, Zhang Y, Karim S (2019) Region search based on hybrid convolutional neural network in optical remote sensing images [J]. *Int J Distrib Sensor Netw* 15(5):155014771985203
27. Woo S., Park J., Lee JY., Kweon I.S. (2018) CBAM: Convolutional block attention module. In: *Computer Vision – ECCV 2018*. ECCV 2018. Lecture Notes in Computer Science. Springer, Cham, 11211, pp.3–19, [https://doi.org/10.1007/978-3-030-01234-2\\_1](https://doi.org/10.1007/978-3-030-01234-2_1)
28. Yin S, Meng L, Liu J (2019) A New apple segmentation and recognition method based on modified fuzzy C-means and hough transform[J]. *J Appl Sci Eng* 22(2):349–354
29. Teng L, Li H, Karim S (2019) DMCNN: A deep multiscale convolutional neural network model for medical image segmentation [J]. *J Health Eng.* 2019:1–10
30. Wang L, Wang C, Sun Z, Chen S (2020) An improved dice loss for pneumothorax segmentation by mining the information of negative areas. *IEEE Access* 8:167939–167949. <https://doi.org/10.1109/ACCESS.2020.3020475>
31. Y Chen et al (2019) Drop an octave: reducing spatial redundancy in convolutional neural networks with octave convolution. In: *IEEE/CVF International conference on computer vision (ICCV)*, pp 3434–3443, <https://doi.org/10.1109/ICCV.2019.00353>
32. Trivizakis E et al (2019) Extending 2-D convolutional neural networks to 3-D for advancing deep learning cancer classification with application to mri liver tumor differentiation. *IEEE J Biomed Health Inform* 23(3):923–930. <https://doi.org/10.1109/JBHI.2018.2886276>
33. Christ P.F. et al. (2016) automatic liver and lesion segmentation in CT Using cascaded fully convolutional neural networks and 3D Conditional Random Fields. In: *Medical Image Computing and Computer-Assisted Intervention-MICCAI 2016*. MICCAI 2016. Lecture Notes in Computer Science. Springer, Cham, vol 9901, pp. 415–423 [https://doi.org/10.1007/978-3-319-46723-8\\_48](https://doi.org/10.1007/978-3-319-46723-8_48)
34. Yin S, Li H, Teng L, Jiang M, Karim S (2020) An optimised multi-scale fusion method for airport detection in large-scale optical remote sensing images [J]. *Int J Image Data Fusion* 11(2):201–214. <https://doi.org/10.1080/19479832.2020.1727573>
35. Yin S, Li H (2020) Hot region selection based on selective search and modified fuzzy C-Means in remote sensing images[J]. *IEEE J Selected Topics in Appl Earth Observa Remote Sens* 13:5862–5871. <https://doi.org/10.1109/JSTARS.2020.3025582>
36. G Huang, Z Liu, L Maaten Van Der, KQ Weinberger, (2017) Densely connected convolutional networks. In: *Conference on computer vision and pattern recognition (CVPR)*, IEEE, pp 2261–2269, <https://doi.org/10.1109/CVPR.2017.243>
37. M Tan et al (2019) MnasNet: Platform-aware neural architecture search for mobile. In: *IEEE/CVF Conference on computer vision and pattern recognition (CVPR)*, pp 2815–2823, <https://doi.org/10.1109/CVPR.2019.00293>
38. M Sandler, A Howard, M Zhu, A Zhmoginov, L Chen (2018) MobileNetV2: Inverted residuals and linear bottlenecks. In: *IEEE/CVF Conference on computer vision and pattern recognition*, pp 4510–4520, <https://doi.org/10.1109/CVPR.2018.00474>
39. X Zhang, X Zhou, M Lin, J Sun, (2018) ShuffleNet: An extremely efficient convolutional neural network for mobile devices. In: *IEEE/CVF Conference on computer vision and pattern recognition*, pp 6848–6856, <https://doi.org/10.1109/CVPR.2018.00716>
40. X Li, W Wang, X Hu, J Yang, (2019) Selective Kernel Networks. In: *IEEE/CVF Conference on computer vision and pattern recognition (CVPR)*, pp 510–519, <https://doi.org/10.1109/CVPR.2019.00060>

**Publisher's Note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.