IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received September 10, 2019, accepted September 30, 2019, date of publication October 11, 2019,
date of current version October 24, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2946582

M-Net: A Novel U-Net With Multi-Stream Feature Fusion and Multi-Scale Dilated Convolutions for Bile Ducts and Hepatolith Segmentation

XIAORUI FU[1], NIAN CAI[1], KEMIN HUANG[1], HUIHENG WANG[1], PING WANG[2], CHENGCHENG LIU[2], AND HAN WANG[3]

[1]School of Information Engineering, Guangdong University of Technology, Guangzhou 510006, China
[2]Department of Hepatobiliary Surgery, The First Affiliated Hospital of Guangzhou Medical University, Guangzhou 510120, China
[3]School of Electromechanical Engineering, Guangdong University of Technology, Guangzhou 510006, China

Corresponding authors: Nian Cai (cainian@gdut.edu.cn) and Ping Wang (wangping1219@126.com)

This work was supported in part by the National Natural Science Foundation of China under Grant 61001179 and Grant 61471132, in part by the Science and Technology Project of Guangdong Province under Grant 2017B090911012, in part by the Project of Jihua Laboratory, China, under Grant X190071UZ190, in part by the Key Laboratory Construction Projects in Guangdong under Grant 2017B030314178, in part by the University Innovation and Entrepreneurship Education Major Project of Guangzhou City under Grant 201709P05, and in part by the Science and Technology Program of Guangzhou, China, under Grant 201803010065.

ABSTRACT Automatically segmenting bile ducts and hepatolith in abdominal CT scans is helpful to assist hepatobiliary surgeons for minimally invasive surgery. High-deformation characteristics of bile ducts and small-size characteristics of hepatolith make this segmentation task challenging. To the best of our knowledge, we make the first attempt to simultaneously segment bile ducts and hepatolith in this paper. Inspired by U-Net, a novel two-dimensional end-to-end fully convolutional network named M-Net is designed to implement this segmentation task. The M-Net is composed of four streams involving two encoder-decoder processes. Multi-scale dilated convolutions are designed to extract abundant semantic features and multi-scale context information at different scales. To make full advantages of multi-scale feature maps, a multi-stream feature fusion strategy is proposed to transfer the most abundant semantic features produced in the first stream to the other streams. To further improve the segmentation performance, a novel loss function is defined to focus the M-Net on hard pixels (difficultly distinguished) in the edges of bile ducts and hepatolith, which is based on the online bootstrapped method and cross entropy. By discarding pixels (easy to distinguish) with higher probability of class, the decline of loss is focused on hard pixels so that the training become more efficient and directional. Experimental results indicate that our proposed M-Net is superior to the state-of-the-art deep-learning methods for simultaneously segmenting bile ducts and hepatolith in the abdominal CT scans. The M-Net can simultaneously segment bile ducts and hepatolith in abdominal CT scans at a high performance with 98.678% Recall, 84.427% Precision, 89.831% DICE and 90.998% F1-score for bile ducts, and 99.894% Recall, 55.132% Precision, 71.248% DICE and 71.051% F1-score for hepatolith.

INDEX TERMS Segmentation of bile ducts and hepatolith, U-Net, multi-scale dilated convolution, multi-stream feature fusion, online bootstrapped loss function, cross entropy.

I. INTRODUCTION

Hepatobiliary stone disease is one of the most common surgical conditions in the world, especially in Asia [1]. At present, minimally invasive surgery for hepatolith removal is the dominate surgical method for the treatment of hepatolithiasis. Bile ducts and hepatolith should be well positioned in CT scans for preoperative plans so that hepatobiliary surgeons can make accurate surgical plans. This task should be cautiously done by the experienced hepatobiliary surgeons to achieve successful minimally invasive surgery. If an automatic segmentation method for bile ducts and hepatolith is designed, it will assist hepatobiliary surgeons to obtain accurate positions of bile ducts and hepatolith in CT scans so that they can achieve more

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

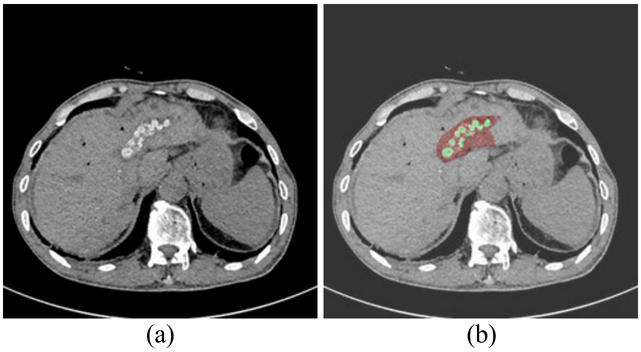


FIGURE 1. (a) An abdominal CT image; (b) Corresponding visualization of the lesion region: bile ducts are marked with the red region, and the hepatolith is marked with the green region.

intuitive judgments to improve the success rate of surgery. Fig. 1 illustrates an automatic segmentation example for bile ducts and hepatolith in abdominal CT scans. Bile ducts and hepatolith should be simultaneously and automatically segmented from the input original CT image. Here, bile ducts are marked with the red region, and the hepatolith is marked with the green region.

Classical image processing methods have successfully dealt with medical image segmentation, such as random forest classification [2], accurate model-based segmentation [3], level set [4] and sparse decomposition [5]. Recently, more and more researchers have introduced deep learning into medical image segmentation [6]–[20] due to its excellent ability of self-learning from a large amount of data through its special convolutional structures. Among them, U-Net [6], [7], [9], [10], [15]–[19], [22]–[24], an end-to-end full convolutional neural network, is a most promising network for medical image segmentation. Thanks to its skip-connection at different resolutions, more image details can be involved in the decoder process, resulting in better image segmentation [6]. A variety of researchers employed U-Net to segment tissues/organs or lesions such as brain tumor, left ventricle, prostate [7], [9], [10]. U-Net has an inherent architecture of pooling in the encoder process and interpolation in the decoder process, which will influence segmentation performance.

To improve segmentation performance, some researchers have modified the original U-Net by incorporating some modules or by adjusting some sub-structures [15]–[19], [22]. Zhang *et al.* [15] combined residual network (ResNet) and U-Net to propose a new end-to-end network named Res-U-Net, which was applied to ultrasound nerve segmentation. Md. Zahangir Alom [16] incorporated the ideas of recurrent convolutional neural network and residual network into U-Net to propose a novel network named R2U-Net. The R2U-Net makes full advantages of feature accumulation with recurrent residual convolutional layers and U-Net. Compared with the original U-Net and Res-U-net, it achieves better results in blood vessel segmentation in retina images, skin cancer segmentation, and lung lesion segmentation. Oktay *et al.* [17] proposed the Attention U-Net (Att U-Net) by introducing a novel attention gate model into the standard U-Net, which was used for multi-class medical image segmentation. The proposed attention gate model can suppress irrelevant regions and highlight salient features so that the prediction performance of the standard U-Net is improved. Jaeger *et al.* [18] proposed a Retina U-Net to detect lung nodules in CT scans, which fused the Retina Net one-stage detector with the U-Net architecture. Its architecture relearns the missing details by complementing object detection with an auxiliary task. Christ *et al.* [19] cascaded two standard U-Nets for a two-stage segmentation of the liver and its lesions, respectively. The first U-Net is trained to segment the liver from the CT image as an ROI input for the second U-Net. Then, the second U-Net is solely trained to segment lesions from the predicted liver ROIs. Explicitly, their network is not an end-to-end network. NasUnet [22] incorporates the Neural architecture search (NAS) strategy into the U-Net for medical image segmentation. Three types of primitive operation set on search space are designed to automatically find two cell architecture DownSC and UpSC. NAS can be seen as the subfield of AutoML (auto machine learning) and has significant overlap with hyper-parameter optimization and meta-learning. These improved U-Nets have succeeded in segmenting the tissues/organs or lesions from medical images.

To the best of our knowledge, no literature reports simultaneously segmenting bile ducts and hepatolith in abdominal CT scans. Bile ducts have high-deformation shapes and hepatolith in bile ducts is sometimes filled in bile ducts. Also, hepatolith stones are commonly of small sizes and separated individually. Thus, the multi-class segmentation for bile ducts and hepatolith is challenging. In this paper, we propose a novel U-Net to perform a multi-class task of bile ducts and hepatolith segmentation, which is named as M-Net since its structure looks like the character M. The M-Net cascades four streams including two encoder-decoder processes. Multi-scale dilated convolutions are designed to achieve abundant semantic features and different scale context information of bile ducts and hepatolith. Furthermore, the most abundant semantic features in the first stream is transferred to the other streams. Thus, multi-stream features are fused in the network to preserve edge details of segmented objects. To further refine the segmentation, an improved Bootstrapped loss function is defined by incorporating the idea of cross entropy into the Online Bootstrapped loss function [25], [26].

To summarize, this paper contains the following contributions.

(1) We propose an end-to-end network named M-Net to segment bile ducts and hepatolith in abdominal CT scans.

The M-Net combines two U-Nets via the strategy of multi-stream feature fusion. This end-to-end cascaded U-Net can avoid the problems of the reduction of the resolutions of feature maps and the loss of semantic features.

(2) The strategy of multi-stream feature fusion in the network can effectively fuse shallow feature information in different streams to preserve edge details with a high segmentation accuracy.

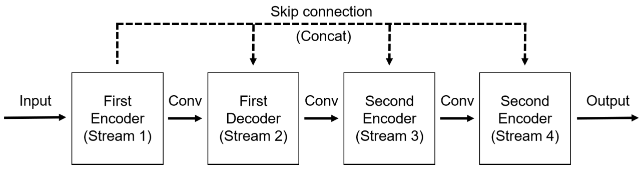


FIGURE 2. The sketch of the M-Net for bile ducts and hepatolith segmentation.

(3) The designed multi-scale dilated convolutions can effectively extract the context information at different scales. Thus, this strategy can simultaneously and excellently segment bile ducts and hepatolith with different shapes and sizes.

(4) The defined Bootstrapped cross entropy loss function makes full advantages of the Online Bootstrapped loss function and cross entropy. It can focus the network on hard pixels (difficultly distinguished) in the edges of bile ducts and hepatolith, which will promote the segmentation performance of the M-Net.

II. METHODS

Our method termed as M-Net is based on two-dimensional convolutional architecture, whose input and output are original abdominal CT images and the corresponding segmented images involving the marked lesion region, respectively. As sketched in Fig. 2, M-Net is composed of two cascaded encoder-decoders. Each coder corresponds to one stream, because different coders use different convolutional kernels to learn multi-scale features. The details of the M-Net architecture are described as following.

A. THE ARCHITECTURE OF THE M-NET

To improve the performance of a deep network, a common scheme for a fully convolutional network (FCN) is to deepen the network by adding convolutional layers of the network [27]. This scheme means a more significant increase of the number of network parameters compared with the accuracy improvement, which will result in more model memories for the network and more computation burden. It indicates that this deepening scheme should be implemented in an equipment with an excellent hardware. Also, more pooling operations will result in the reduction of the resolutions of feature maps. Therefore, this scheme is not suitable for bile ducts and hepatolith segmentation. To avoid the above problems, some attempts have been done, such as the cascaded U-Net [19]. The cascaded U-Net combines two U-Nets to implement the task of semantic liver and lesion segmentation. In this network, the first U-Net is trained to segment the liver as the ROI input for the second U-Net. The second U-Net is trained to segment lesions from the predicted liver ROIs via the first U-Net. It indicates that the cascaded U-Net is not an end-to-end network. Inspired by this network, we design a novel fully convolutional network named M-Net, which is an end-to-end network. The architecture of the M-Net is illustrated in Fig 3.

As illustrated in Fig. 3, the M-Net looks like the character M, which is divided into four streams. The four streams are composed of two encoders and two decoders. That is to say, Streams 1-4 imply the first encoder, the first decoder, the second encoder and the second decoder, correspondingly. More and more senior semantic feature information can be extracted stream by stream. Each stream includes four convolutional blocks with different resolutions. The shape information of bile ducts and hepatolith is the most abundant in Stream 1. With the flow of the information stream by stream, the shape information of bile ducts and hepatolith will be lost more and more. To involve more shape information with different resolutions, a multi-stream feature fusion strategy is proposed here. That is, each convolutional block in Stream 1 is skip-connected with the corresponding convolutional block at the same resolution in Streams 2, 3, 4. Also, the convolutions with different kernels are designed in different streams while all the convolutions are the same in one stream. This multi-scale dilated convolution strategy can achieve multi-scale context information of bile ducts and hepatolith. The inputs and outputs of the M-Net are original CT images and the corresponding predicted images in which bile ducts and hepatolith are segmented, respectively.

B. MULTI-SCALE DILATED CONVOLUTIONS AND MULTI-STREAM FEATURE FUSION

It is well-known that smaller convolutional kernels are more sensitive to small targets than larger ones [28]. Increasing the size of the receptive field can make use of the context information in a larger image region [29] and accelerate the convergence of the model [30]. Since the shape sizes of bile ducts and hepatolith are quite different in the CT images, the convolutional kernels with different sizes are designed for different streams in the proposed M-Net. Furthermore, the convolutional kernels with large sizes extract some redundant information, which will maybe influence the segmentation performance. Dilated convolutions can expand the receptive field without loss of resolution or coverage, which can aggregate multi-scale context information to improve the segmentation accuracy [31], [32]. Thus, a multi-scale dilated convolution strategy is designed in the M-Net. 1×1 convolutional kernels can achieve finer feature maps compared with the convolutional kernels with other sizes. It means that these maps can involve more details of bile ducts and hepatolith. Thus, 1×1 convolutional kernels are used in each block of Stream 1. In Streams 2-4, 3×3, 5×5, 7×7 dilated convolutional kernels, correspondingly. Then, the feature maps at different scales can be achieved by different streams. Stream 1 can produce the most abundant semantic features of bile ducts and hepatolith. Although the semantic features can be potentially transferred stream by stream, some semantic features may be lost during the information transmission. So, a multi-stream feature fusion strategy is implemented by transferring the semantic features achieved by Stream 1 to the other three streams. The semantic features achieved by each convolutional block of Stream 1 are transferred to the corresponding convolutional block with the same resolution

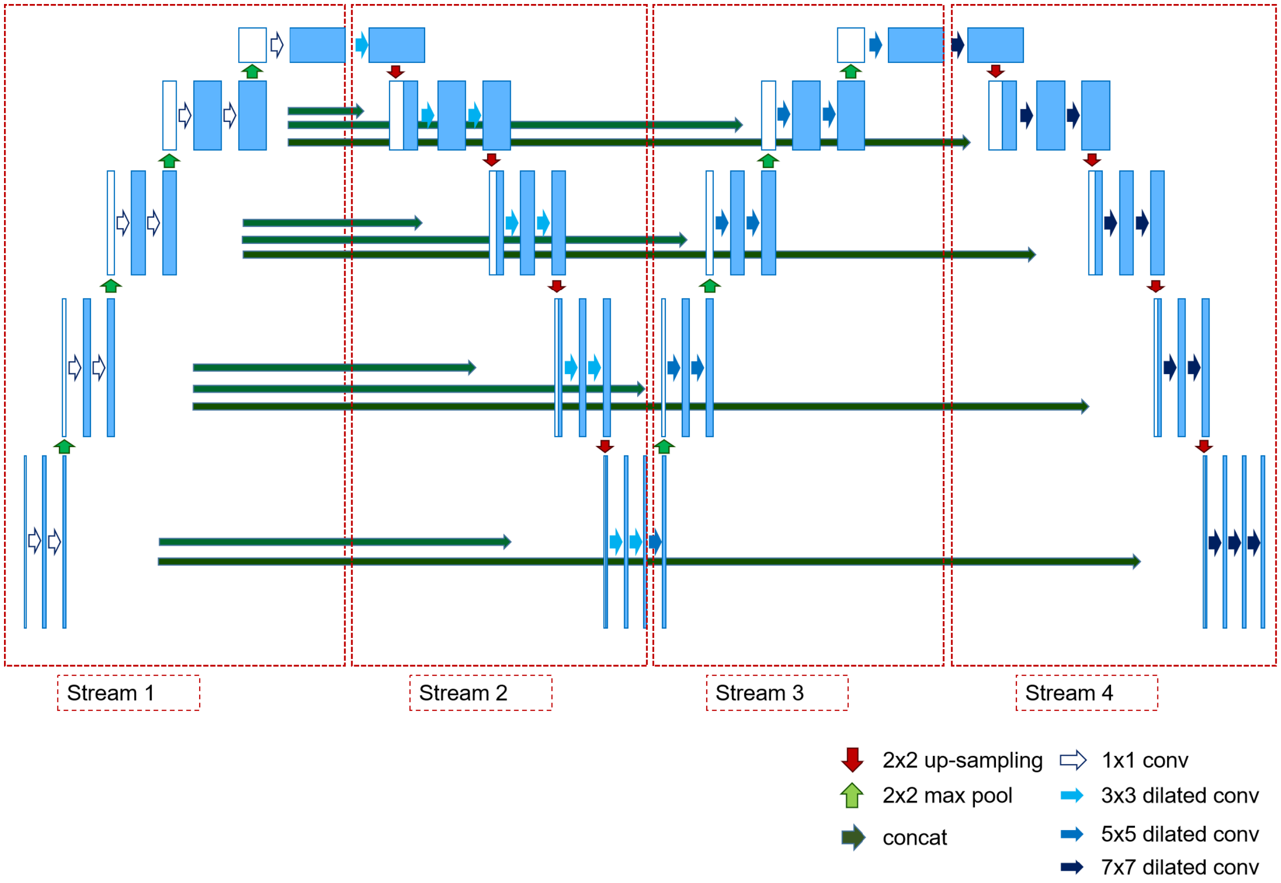


FIGURE 3. M-Net: the network consists of two encoder-decoder processes, corresponding to Streams 1, 2, 3, 4, respectively. Red arrows are 2×2 up-sampling; light green arrows are 2×2 max pooling; gray arrows are 1x1 convolutions; different blue arrows are dilated convolutions with the sizes of 3×3 in Stream 2, 5×5 in Stream 3, and 7×7 in Stream 4; and dark green arrows are concatenation.

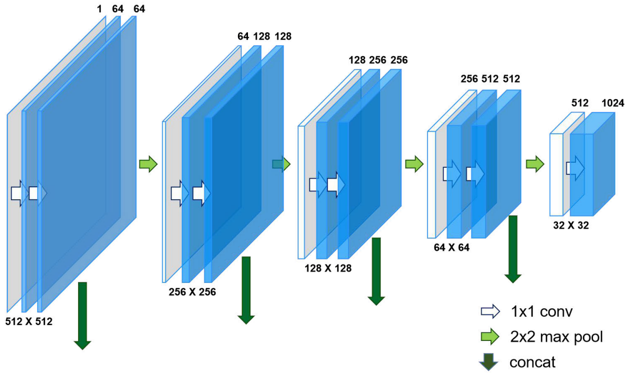


FIGURE 4. Stream 1: the first encoder process.

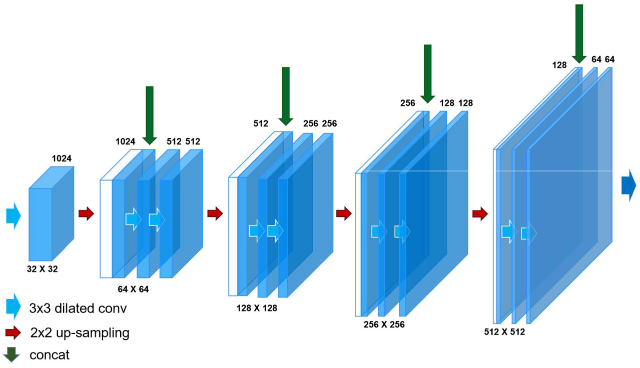


FIGURE 5. Stream 2: the first decoder process.

in Streams 2-4. This is indicated by the dark green arrows in Figs. 4-7.

Figs. 4-7 illustrate the implementation of each stream in detail. In each figure, the numbers above feature maps are the numbers of channels, and the numbers below are the resolutions. Streams 1 and 3 implement the tasks of the first and second encoder processes, in which a 2×2 pooling is employed between the two adjacent blocks. Thus, the resolution (512×512) of the input of the stream is reduced to a small resolution (32×32) in the output of the stream. Streams 2 and 4 implement the tasks of the first and second decoder processes, in which a 2×2 up-sampling is employed between the two adjacent blocks. Thus, the M-Net can output the segmented image with the same resolution as the original input CT image.

C. THE BOOTSTRAPPED CROSS ENTROPY LOSS FUNCTION

Cross entropy loss function (CELF) is a convex optimization function that effectively measures subtle changes [33]. When

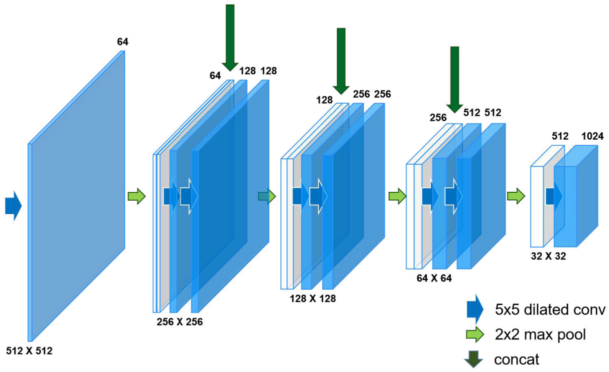


FIGURE 6. Stream 3: the second encoder process.

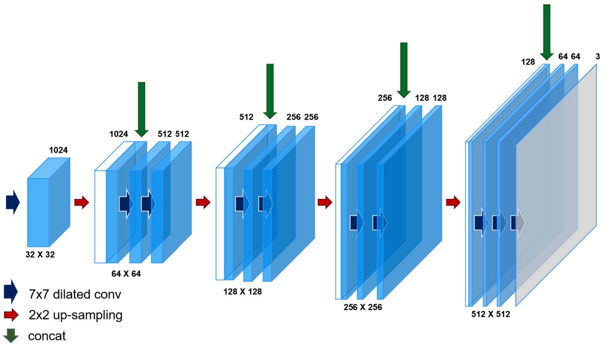


FIGURE 7. Stream 4: the second decoder process.

the loss between the prediction and the ground-truth is large, the gradient in the back propagation of training process also becomes large. So, the convergence speed is faster than the quadratic loss function. However, all the pixels in the image are learned equally by the network with the CELF since the CELF evaluates the class prediction for each pixel separately. This means that most of the pixels tend to be classified as the background pixels since bile ducts and hepatolith occupy small regions in abdominal CT scans. That is to say, some edge pixels of bile ducts and hepatolith are indistinguishable and even mis-classified as the background pixels, which can be considered as hard pixels defined in [25], [26]. The online bootstrap loss function (OBLF) defined by Wu *et al.* [26] can force the network to focus on hard pixels during training so that it can solve the above problem during segmentation. It is defined as

$$L = -\frac{1}{\sum_{\mathrm{i}=1}^{N}\sum_{\mathrm{j}=1}^{K} 1\left\{(y_i = j) \cap \left(p_{ij} \leq t\right)\right\}} \times \sum_{\mathrm{i}=1}^{N}\sum_{\mathrm{j}=1}^{K} 1\left\{(y_i = j) \cap \left(p_{ij} \leq t\right) \mathrm{log} p_{ij}\right\} \quad (1)$$

TABLE 1. The numbers of the labelled CT images in the GDPU-HS.

LABEL	NUMBER OF LABELLED CT IMAGES		
	TRAIN	TEST	TOTAL
Bile duct	250	80	330
Bile Duct & Hepatolith	350	106	456

where N and K denote the number of image pixels and the number of pixel categories, respectively. $y_i = j$ denotes that y_i belongs to the jth category, in which yi refers to the ground-truth label of the ith pixel. p_{ij} denotes the measured probability of the ith pixel belonging to the jth category. $t \in (0, 1]$ is a threshold. Here $1\{\bullet\}$ equals to 1 when the condition inside the brackets holds, and otherwise equals to 0.

The OBLF uses the idea of maximum likelihood estimation to punish the wrong classification by the logarithmic loss. However, the logarithmic loss is not sensitive to measure subtle changes. The subtle changes between the ground truths and the predictions can be effectively measured by the CELF. So, we incorporate the idea of CELF into the OBLF and define a novel loss function named bootstrapped cross entropy loss function (BCELF), which is formulated as

Simultaneously segmenting bile ducts and hepatolith in abdominal CT scans is a multi-class task. The model maybe does not achieve a good segmentation result if the threshold is a constant like t in (1). Thus, we use different thresholds to characterize different categories. As illustrated in (2), as shown at the bottom of this page., t_j refers to the threshold for the jth category.

III. EXPERIMENTAL RESULTS

A. DATASET, METRICS AND IMPLEMENTATION DETAILS

1) DATASET

The experimental data are clinical abdominal CT scans provided by the First Affiliated Hospital of Guangzhou Medical University. 9 cases with 9800 CT images are involved. The size of each CT image is 512×512. Dense segments at high quality pixel levels were manually labelled as ground truths by Ping Wang, an experienced hepatobiliary surgeon and one of the authors of this paper. He used the professional annotation tool named Labelme [34] to label hepatolith during the plain scan of CT, and to label bile ducts and hepatolith during the portal phase. Only 786 labelled CT images were selected to establish the dataset for this paper since most of CT images involved no bile duct or hepatolith. As shown in Table 1, 330 labelled CT images only involve bile ducts

$$LOSS = \frac{-\sum_{i=1}^{N}\sum_{j=1}^{K}[y_i \log y_i + (1-\hat{y}_i)\log(1-\hat{y}_i)] \cdot 1\{(y_i = j) \cap (p_{ij} \leq t_j)\}}{\sum_{i=1}^{N}\sum_{j=1}^{K} 1\{(y_i = j) \cap (p_{ij} \leq t_j)\}} \quad (2)$$

TABLE 2. Comparisons of M-Net models with different convolutional kernels.

Models	Bile Duct				Hepatolith			
	Recall (%)	Precision (%)	DICE (%)	F1-score （%）	Recall (%)	Precision (%)	DICE (%)	F1-score （%）
M-Net1 (3,3,3,3)	97.668	76.288	85.637	85.664	99.844	52.983	69.117	69.229
M-Net2 (3*,3*,3*,3*)	97.635	77.627	86.448	86.489	99.843	53.564	69.667	69.723
M-Net3 (1,3,5,7)	98.435	84.167	89.119	90.744	99.879	54.964	71.067	70.907
M-Net4 (1,3*,5*,7*)	98.678	84.427	89.831	90.998	99.894	55.132	71.248	71.051

and 456 labelled CT images involve bile ducts and hepatolith. 250 labelled CT images involving bile ducts and 350 ones involving bile ducts and hepatolith are randomly selected for training (GDPU-HS-train) from these labelled CT images. The rest of labelled CT images are for testing (GDPU-HS-test). A validation set is randomly split from the GDPU-HS-test set, in which there are about one half of the samples of the GDPU-HS-test. Finally, all the samples of the GDPU-HS-test are used for testing. Due to case privacy and confidentiality agreements, the GDPU-HS dataset is not open accessable.

2) EVALUATION METRICS

We use several common-used metrics to evaluate the segmentation performance, which are precision, recall, Dice similarity coefficient (DICE) and F1-score. For bile ducts, precision (positive prediction value) means the proportion of bile ducts in the correct predictions and those in the ground truths. Recall (sensitivity) is the ratio of bile ducts in the correct predictions and those in all the predictions. The larger their value, the better performance of segmentation. The precision and recall are calculated for each class, which are defined as

$$Precision = TP/(TP + FP) \tag{3}$$

$$Recall = TP/(TP + FN) \tag{4}$$

where TP (True positive) is defined as the number of correct segmented pixels in bile ducts or hepatolith. FP (False positive) and FN (False Negative) represent the wrongly segmented foreground (bile ducts or hepatolith) and background, respectively.

Dice similarity coefficient (DICE) and F1-score can more intuitively reflect the overall segmentation performance compared with precision and recall, which are defined as

$$DICE = 2TP/(FP + FN + 2TP) \tag{5}$$

$$F1\text{-score} = \frac{2 \cdot Precision \cdot Recall}{Precision + Recall} \tag{6}$$

As indicated in (5) and (6), Dice similarity coefficient and F1-score are both related with the components TP, FP and FN. Furthermore, the larger the value of Dice similarity coefficient/F1-score is, the better segmentation performance the model achieves.

3) IMPLEMENTATION DETAILS

We used Adam [35] to train the M-Net with alpha (learning rate) 10-3, beta 1 (first decay) 0.9, beta 2 (second decay) 0.999, epsilon (Prevent division by 0) 10-8, and 300 steps of each epoch. The model was trained until it converged and the number of training epochs was recorded. After cross-validation, the trained model was built for simultaneously segmenting bile ducts and hepatolith from abdominal CT scans. All the experiments were performed on a single machine with an Nvidia Quadro M2000M 4GB GPU. Our models were modelled on Keras [36] with TensorFlow backend.

B. MULTI-SCALE DILATED CONVOLUTIONS

Since bile ducts and hepatolith have different shapes and sizes, different convolutional kernels will definitely influence the segmentation results. We conducted an experiment to validate the proposed multi-scale dilated convolution strategy.

Four M-Net models are established with different convolutional kernels in the streams, which are simply named as M-Nets 1-4. The digits of corresponding positions in the parentheses correspond to the sizes of convolutional kernels in all the convolutional layers in the corresponding streams. For example, (1, 3*, 5*, 7*) means that Stream 1 uses 1×1 convolutional kernels, and Streams 2, 3, 4 use 3×3, 5×5 and 7×7 dilated convolutional kernels, respectively. The symbol * is used to represent the dilated convolution operation, such as 5* is a 5×5 dilated convolution.

As shown in Table 2, the models with multi-scale convolutions achieve better segmentation performance than those with convolutional kernels of the same scale in terms of recall, precision and DICE. This is because the convolutions with different scales can produce the feature maps involving the semantic information at different levels. Furthermore, 1×1 convolutional kernels can achieve finer feature maps to characterize more details of bile ducts and hepatolith. Compared with the models without dilated convolutions, the models with dilated convolutions can segment bile ducts and hepatolith more excellently. This is because dilated convolutions can expand the receptive field without loss of resolution or coverage. Since the multi-scale dilated convolution strategy combines the ideas of the multi-scale convolutions and the dilated convolutions, M-Net4(1,3*,5*,7*) is the best model to simultaneously segment bile ducts and hepatolith.

C. THE VALUES OF THE THRESHOLDS t_j

For our segmentation task, bile ducts and hepatolith have quite different shapes in terms of size and deformation. Compared with hepatolith, the shapes of bile ducts are of larger sizes with higher deformation. Thus, the thresholds $t_j (j = 1, 2)$ should be elaboratively selected for excellent segmentation. Here t_1 and t_2 correspond for bile ducts and hepatolith. Some interferences between those two thresholds will occur. We conducted an experiment to discuss the selection and the influence of the thresholds t_j. In this experiment, the M-Net4(1,3*,5*,7*) models with different values of the thresholds t_j are employed to simultaneously segment bile ducts and hepatolith. For clear illustration, the segmentation results of bile ducts and hepatolith are separately illustrated in Fig. 8. As indicated in Figs. 8(a) and 8(c), the models can achieve fair good segmentation results for bile ducts when $t_1 \in [0.6, 0.7]$ and $t_2 \in [0.5, 0.7]$. And hepatolith can be well segmented by the models with the thresholds $t_1 \in [0.6, 0.7]$ and $t_2 \in [0.5, 0.6]$, as shown by in Figs. 8(b) and 8(d). There should be one compromise selection of t_1 and t_2 at the same point in Figs. 8(c) and 8(d). Thus, the thresholds t_j of the proposed M-Net can be selected to simultaneously and well segment bile ducts and hepatolith in the range of $t_1 \in [0.6, 0.7]$ and $t_2 \in [0.5, 0.6]$. In this paper, the selection of $t_1 = 0.65$, $t_2 = 0.55$ is both illustrated by a black spot in Figs. 8(c) and 8(d).

D. BOOTSTRAPPED CROSS ENTROPY LOSS FUNCTION

We conducted a comparison experiment to discuss the validation of the proposed bootstrapped cross entropy loss function (BCELF). Three kinds of loss functions of cross entropy, online bootstrapped and bootstrapped cross entropy were employed to train the M-Net4(1,3*,5*,7*) models.

As shown in Table 3, the model with the online bootstrapped loss function (OBLF) achieves better segmentation performance than that with the cross entropy loss function (CELF) in terms of recall, precision and DICE. This is because the OBLF allows more learning opportunities for hard pixels difficult to be segmented accurately. Furthermore, the objective evaluations of the BCELF are the best among three loss functions. This benefits from the idea of the OBLF and the sensitivity of cross entropy for subtle changes.

A visualization example is illustrated in Fig. 9 to subjectively show the segmentation results achieved by the models with different loss functions. The red area marks bile ducts and the green area marks hepatolith. As shown in Fig. 9, the edges of bile ducts and hepatolith are segmented by the models with the bootstrapped-based loss function finer than that with the CELF. Especially, the model with the proposed BCELF can well separate the individual hepatolith, which approximates the ground truth. This is consistent to the objective results described above.

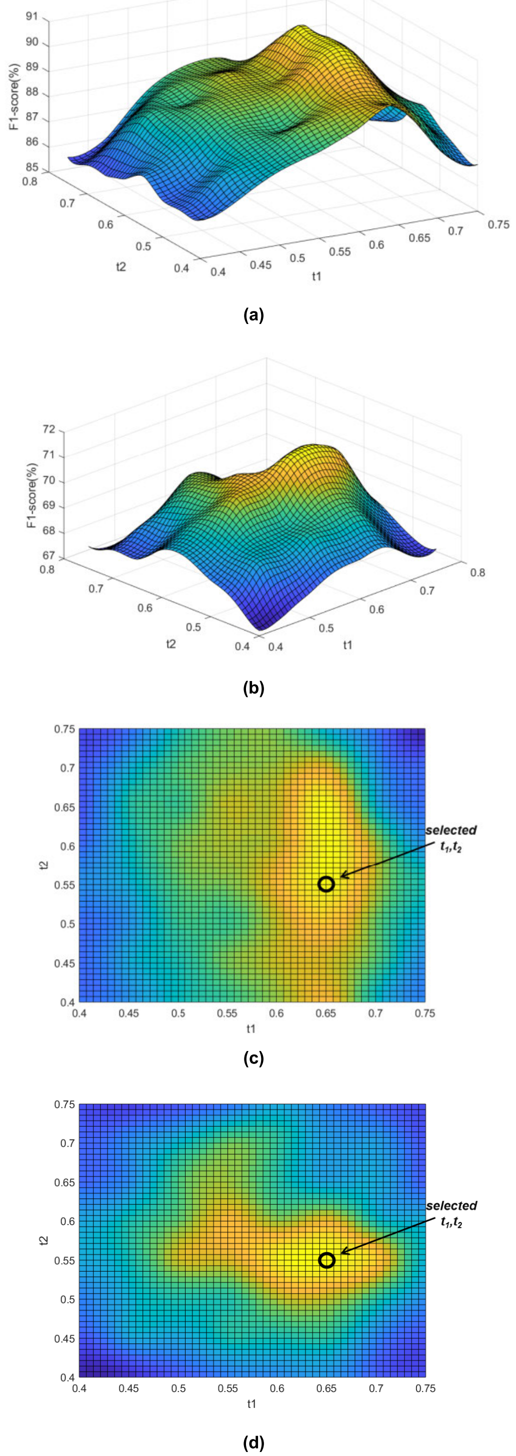


FIGURE 8. Segmentation results achieved by the M-Net models with different thresholds t_j. (a) and (b) are topographic maps for bile ducts and hepatolith, respectively; (c) and (d) are heatmaps for bile ducts and hepatolith, respectively. The black spot denotes the selected t_1, t_2.

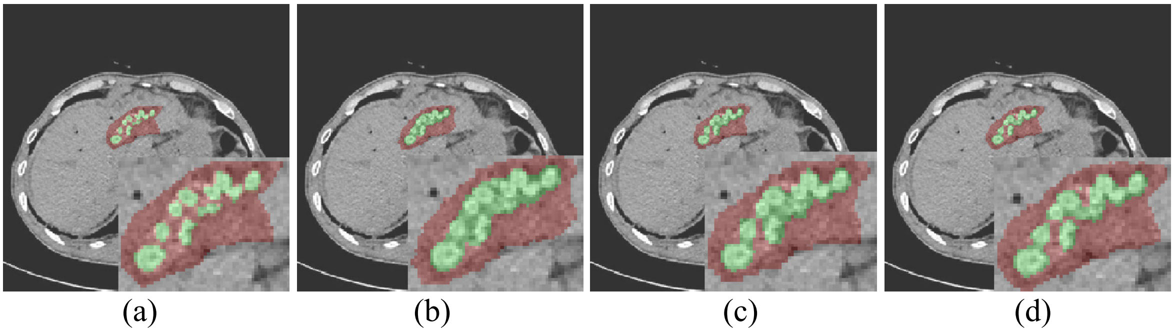


FIGURE 9. Segmentation results achieved by M-Net models with different loss functions, red and green region correspond to bile ducts and hepatolith. (a) Ground truth; (b) Cross entropy; (c) Online bootstrapped; (d) Bootstrapped cross entropy.

TABLE 3. Comparisons of M-Net models with different convolutional kernels.

LOSS FUNCTIONS	BILE DUCTS				HEPATOLITH			
	RECALL (%)	PRECISION (%)	DICE (%)	F1-SCORE （%）	RECALL (%)	PRECISION (%)	DICE (%)	F1-SCORE （%）
Cross entropy	88.092	61.313	71.031	72.303	99.789	44.345	61.153	61.403
Online Bootstrapped	95.142	74.125	82.754	83.329	99.884	50.856	67.258	67.397
Bootstrapped Cross entropy	98.678	84.427	89.831	90.998	99.894	55.132	71.248	71.051

TABLE 4. Comparisons of different deep-learning-based segmentation methods for bile ducts and hepatolith.

METHODS	BILE DUCT				HEPATOLITH			
	RECALL%	PRECISION %	DICE %	F1-SCORE （%）	RECALL %	PRECISION %	DICE %	F1-SCORE （%）
FCN-8s [27]	85.092	59.663	71.031	70.144	99.789	44.346	61.153	61.404
SegNet [37]	87.671	63.996	73.833	73.986	99.319	45.465	62.057	62.376
Deeplab V3 [28]	87.486	65.962	73.992	75.214	99.739	42.083	59.946	59.191
U-Net [6]	93.795	73.196	82.621	82.225	99.781	51.188	67.611	67.664
Retina U-Net [18]	89.954	67.309	78.904	77.001	99.659	47.812	65.604	64.621
R2U-Net [16]	92.150	69.779	79.931	79.419	99.693	49.174	66.873	65.862
Att U-Net [17]	95.610	72.896	82.711	82.722	99.850	54.606	69.934	70.601
Cascade U-Net [19]	95.981	74.372	83.796	83.806	99.682	53.136	69.199	69.320
NasUnet [22]	97.142	76.399	85.722	85.531	99.805	52.736	68.899	69.009
DEDN [20]	96.620	75.099	84.498	84.511	99.881	50.776	67.258	67.326
M-Net4 (1,3*,5*,7*)	98.678	84.427	89.831	90.998	99.894	55.132	71.248	71.051

E. COMPARISONS WITH THE STATE-OF-THE-ART METHODS

In this section, we compare the proposed M-Net with the BCELF with the state-of-the-art deep-learning-based medical image segmentation methods, which are FCN-8s [27], SegNet [37], Deeplab V3 [28], U-Net [6], Retina U-Net [18], R2U-Net [16], Att U-Net [17], Cascaded U-Net [19], NasUnet [22] and DEDN [20].

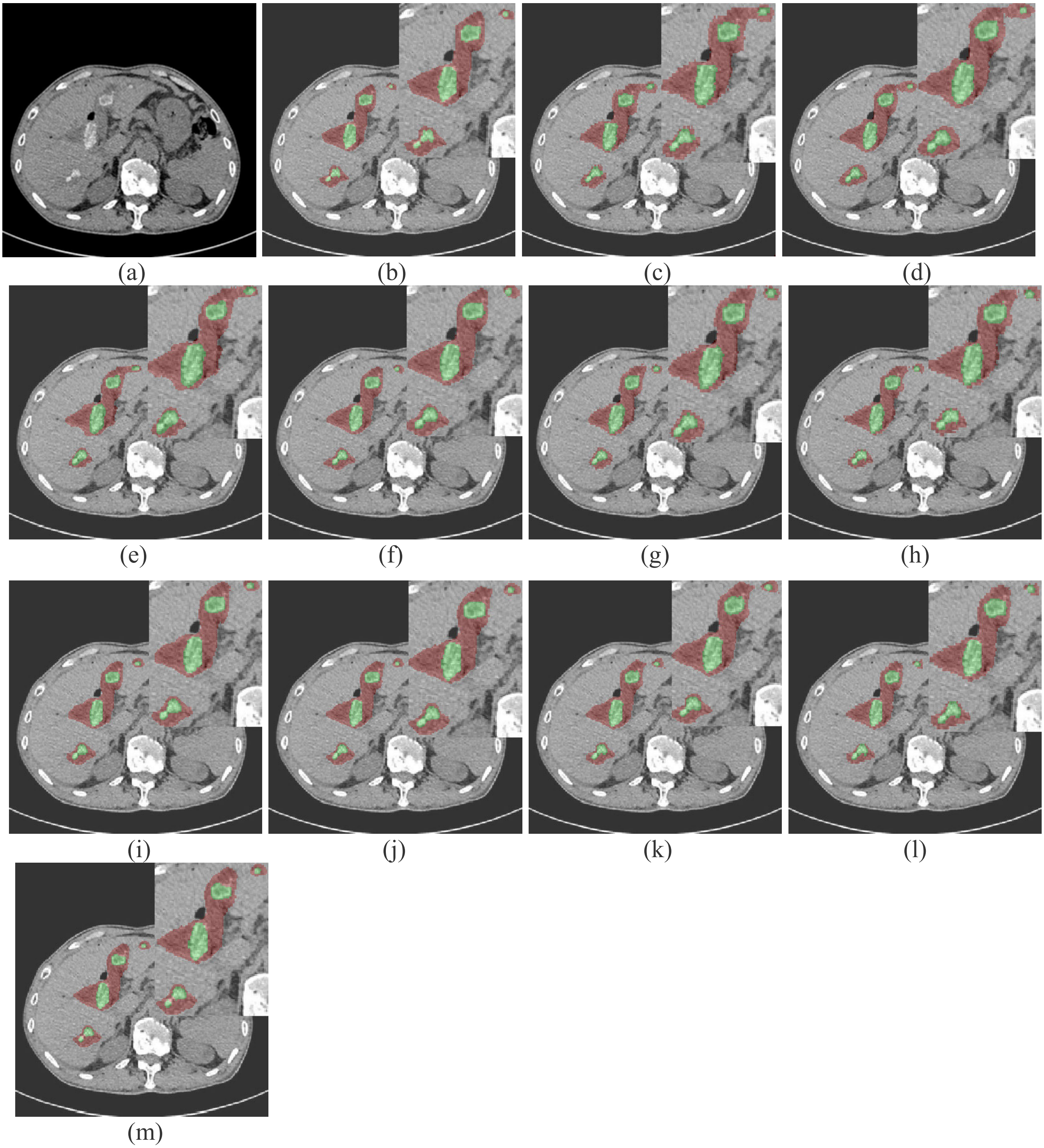


FIGURE 10. Segmentation results of various deep-learning-based methods, red and green regions correspond to bile ducts and hepatolith, respectively. (a) Original CT image; (b) ground truth; (c) FCN8s; (d) SegNet; (e) Deeplab V3; (f) U-Net; (g) Retina U-Net; (h) R2U-Net; (i) Att U-Net; (j) Cascaded U-Net; (k) NasUnet; (l) DEDN; (m) M-Net.

As shown in Table 4, the FCN-8s achieves the worst performance because a number of interpolation operations resulted from its recursive upsampling will maybe introduce many artifacts. Since pooling indices computed in the max-pooling layers are used for upsampling in the decoder of the SegNet, this kind of upsampling operation will neglect the context information especially for low-resolution feature maps. This results that the SegNet performs bad segmentation for bile ducts and hepatolith. The strategy of atrous spatial pyramid pooling (ASPP) in the DeepLab V3 promote the network to focus on more context information rather than detailed information. Bile ducts have the characteristics of

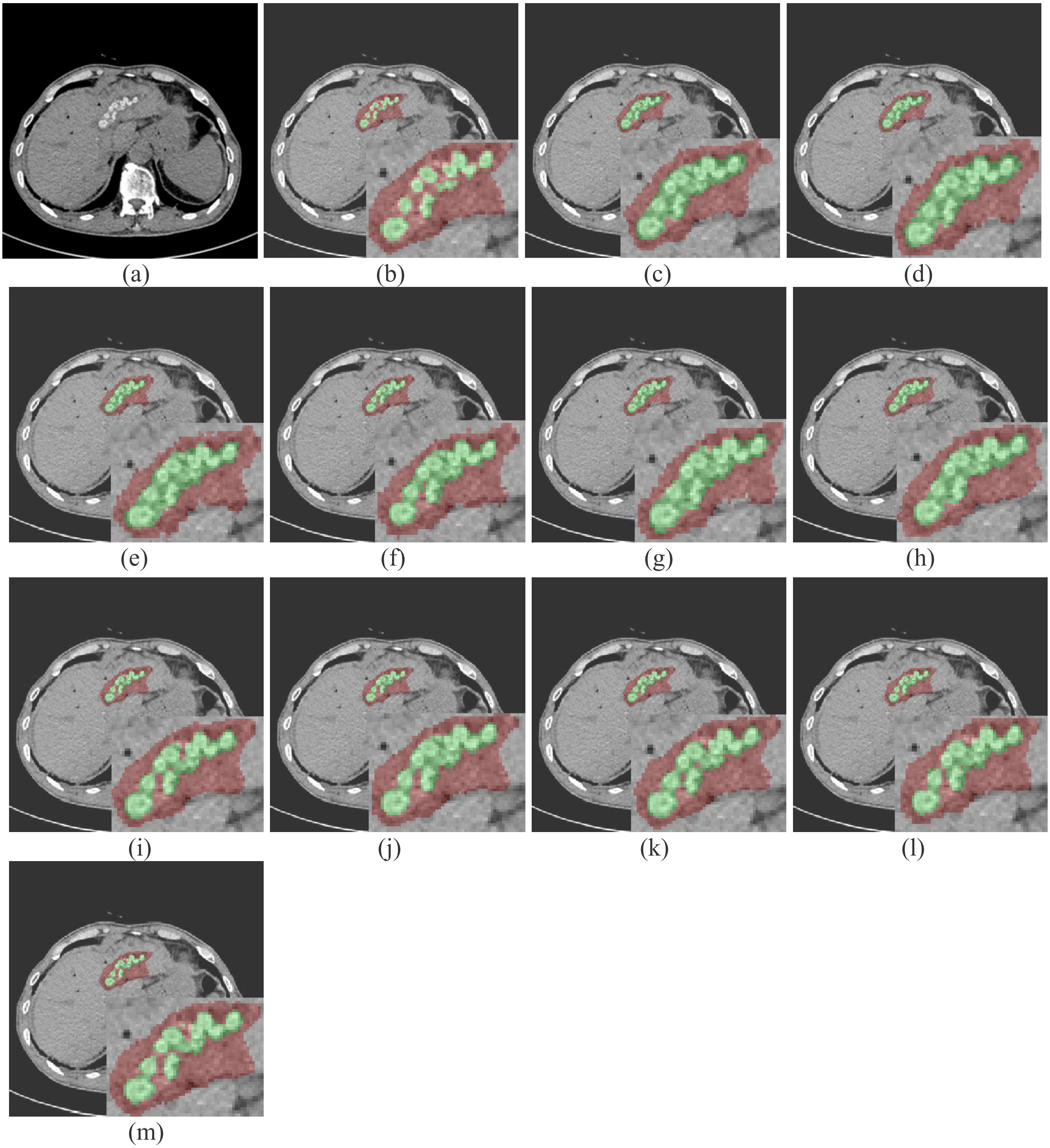


FIGURE 11. Segmentation results of various deep-learning-based methods, red and green regions correspond to bile ducts and hepatolith, respectively. (a) Original CT image; (b) ground truth; (c) FCN8s; (d) SegNet; (e) Deeplab V3; (f) U-Net; (g) Retina U-Net; (h) R2U-Net; (i) Att U-Net; (j) Cascaded U-Net; (k) NasUnet; (l) DEDN; (m) M-Net.

high deformation and hepatolith stones are of small sizes and densely appeared in bile ducts. Thus, the DeepLab V3 also does a bad job in segmenting bile ducts and individual hepatolith stones. Due to the principle of junior feature fusion, seven U-Net-based networks involved our proposed M-Net achieve better segmentation performance than the above three networks. The Retina U-Net is a one-stage detection model for detecting the tissues/organs or lesions in medical images, which makes full advantage of U-Net for semantic segmentation. However, our task is to accurately segment bile ducts and hepatolith. The strategy of pyramid dilated convolution may result that the detailed information of bile ducts and hepatolith will be lost in R2U-Net. Thus, Retina U-Net and R2U-Net achieve worse segmentation performance than U-Net. Att U-Net and Cascaded U-Net perform better in segmenting bile ducts and hepatolith since Att U-Net introduces the attention mechanism into U-Net and Cascaded U-Net cascades two U-Nets for two-stage segmentation. NasUnet [22] achieves a fairly good performance since its neural architecture search (NAS) strategy has significant overlap with hyper-parameter optimization and meta-learning. Our proposed M-Net achieves the best evaluation metrics compared with the above state-of-the-art deep-learning-based methods. Also, it performs better segmentation work than the latest deep learning network DCDN [20].

Figs. 10 and 11 illustrate two visualization examples of segmenting bile ducts and hepatolith in abdominal CT images (red and green regions correspond to bile ducts and hepatolith, respectively). Also, the segmented regions are zoomed in for clearly visual comparisons. As shown in Figs. 10(c)-10(e), some mis-segmentation occurs in bile ducts, which is achieved by FCN8s, SegNet and Deeplab V3. Seven U-Net-based methods and DCDN can well segment bile ducts in the CT images, as shown in Figs. 10(f)-10(m). To further evaluate these seven U-Net-based methods, another CT image with many individual hepatolith stones is employed for segmentation. As illustrated in Figs. 11(g) and 11(h), the hepatolith stones are segmented as a whole object and cannot be separated individually. Some individual hepatolith stones in bile ducts are well segmented by U-Net, Att U-Net, Cascaded U-Net, NasUnet and DCDN. And the proposed M-Net achieves the best segmentation for hepatolith stones, which even approximate the ground truth. Also, bile ducts are segmented by M-Net the most excellently among all the deep-learning-based methods. These subjective results are consistent to the above objective results.

IV. DISCUSSION

Automatic segmentation of bile ducts and hepatolith can assist hepatobiliary surgeons to accurately position bile ducts and hepatolith in abdominal CT scans, which is helpful for minimally invasive surgery. Although many deep-learning-based methods have proliferated for medical image segmentation, they may be not suitable to simultaneously segment bile ducts and hepatolith in abdominal CT scans due to the inherent characteristics of bile ducts and hepatolith. Hepatolith stones in bile ducts have small sizes and occupy very small areas in abdominal CT scans. Also, the high-deformation characteristics of bile ducts will result that the pixels (hard pixels) of their edges cannot be well distinguished. Thus, it is a challenging task for simultaneously segmenting bile ducts and hepatolith in abdominal CT scans.

We design the M-Net to simultaneously segmenting bile ducts and hepatolith, which is based on a U-Net as a backbone network. The M-Net involves the strategies of multi-scale dilated convolutions and multi-stream feature fusion and a novel loss function named bootstrapped cross entropy loss function (BCELF). Benefiting from multi-scale dilated convolutions, the most abundant semantic features can be effectively extracted in Stream 1 and multi-scale context information can be learned by the streams with different dilated convolutional kernels. Multi-stream feature fusion can effectively promote the transferring of multi-scale feature maps stream by stream by means of transferring the most abundant semantic features to the other streams. Furthermore, the defined BCELF combines the advantages of the online bootstrapped loss function (OBLF) that focuses on hard pixels and of the cross entropy loss function (CELF) that can characterize subtle changes. Therefore, as shown in Table 4 and Figs 10-11, the proposed M-Net achieves the best segmentation performance among the state-of-the-art deep-learning-based methods. It can simultaneous segment bile ducts and hepatolith in abdominal CT scans at a high performance with the DICEs of 89.831% and the F1-score of 90.998% for bile ducts, and the DICEs of 71.248% and the F1-score of 71.051% for hepatolith.

In the future work, it is recommended to extend the GDPU-HS dataset, which is helpful for the M-Net to further improve the segmentation accuracy and generalization. Moreover, the thresholds t_j will be adaptively determined according to the medical prior knowledge.

V. CONCLUSION

In this paper, we design a novel U-Net architecture named M-Net for simultaneously segmenting bile ducts and hepatolith. The M-Net depends on three strategies to effectively improve the segmentation performance, which are multi-scale dilated convolution, multi-stream feature fusion and bootstrapped cross entropy loss function. Comparison experiments indicate that the proposed M-Net can simultaneously segment bile ducts and hepatolith in abdominal CT scans at a high performance with 98.678% Recall, 84.427% Precision, 89.831% DICE and 90.998% F1-score for bile ducts, and 99.894% Recall, 55.132% Precision, 71.248% DICE and 71.051% F1-score for hepatolith, which is superior to some state-of-the-art deep-learning methods.

REFERENCES

[1] M. Y. Dar, S. Ali, A. H. Raina, M. A. Raina, O. J. Shah, M. A. Shah, and S. S. Mudassar, "Association of helicobacter pylori with hepatobiliary stone disease, a prospective case control study," *Indian J. Gastroenterol.*, vol. 35, no. 5, pp. 343–346, Sep. 2016.

[2] A. Mastmeyer, D. Fortmeier, and H. Handels, "Random forest classification of large volume structures for visuo-haptic rendering in CT images," *Proc. SPIE*, vol. 9784, Mar. 2016, Art. no. 97842H. doi: 10.1117/12.2216845.

[3] A. Mastmeyer, G. Pernelle, R. Ma, L. Barber, and T. Kapur, "Accurate model-based segmentation of gynecologic brachytherapy catheter collections in MRI-images," *Med. Image Anal.*, vol. 42, pp. 173–188, Dec. 2017.

[4] Z. Zhang and J. Song, "An adaptive fuzzy level set model with local spatial information for medical image segmentation and bias correction," *IEEE Access*, vol. 7, pp. 27322–27338, 2019.

[5] C. Lin, Y. Wang, T. Wang, and N. Dong, "Segmentation and recovery of pathological MR brain images using transformed low-rank and structured sparse decomposition," in *Proc. IEEE 16th Int. Symp. Biomed. Imag. (ISBI)*, Apr. 2019, pp. 1878–1881. doi: 10.1109/ISBI.2019.8759441.

[6] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention—MICCAI*. Cham, Switzerland: Springer, 2015, pp. 234–241.

[7] H. Dong, G. Yang, F. Liu, Y. Mo, and Y. Guo, "Automatic brain tumor detection and segmentation using U-net based fully convolutional networks," in *Proc. Annu. Conf. Med. Image Understand. Anal.*, Edinburgh, U.K., 2017, pp. 506–517.

[8] M. U. Dalmış, G. Litjens, K. Holland, A. Setio, R. Mann, N. Karssemeijer, and A. Gubern-Mérida, "Using deep learning to segment breast and fibroglandular tissue in MRI volumes," *Med. Phys.*, vol. 44, no. 2, pp. 533–546, Feb. 2017.

[9] S. Charmchi, K. Punithakumar, and P. Boulanger, "Optimizing U-net to segment left ventricle from magnetic resonance imaging," in *Proc. IEEE Int. Conf. Bioinf. Biomed. (BIBM)*, Madrid, Spain, Dec. 2018, pp. 327–332.

[10] W. Chen, Y. Zhang, J. He, Y. Qiao, Y. Chen, H. Shi, and X. Tang. "Prostate segmentation using 2D bridged U-net," 2018, *arXiv:1807.04459*. [Online]. Available: https://arxiv.org/abs/1807.04459

[11] M. Lai, "Deep learning for medical image segmentation," 2015, *arXiv:1807.04459*. [Online]. Available: https://arxiv.org/abs/1807.04459

[12] P. Moeskops, M. J. Wolterink, H. M. B. van der Velden, G. A. K. Gilhuijs, T. Leiner, A. M. Viergever, and I. Išgum, "Deep learning for multi-task medical image segmentation in multiple modalities," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.*, Athens, Greece, 2016, pp. 478–486.

[13] G. Wang, W. Li, A. Maria Zuluaga, R. Pratt, A. Premal Patel, M. Aertsen, T. Doel, L. Anna David, J. Deprest, S. Ourselin, and T. Vercauteren, "Interactive medical image segmentation using deep learning with image-specific fine tuning," *IEEE Trans. Med. Imag.*, vol. 37, no. 7, pp. 1562–1573, Jul. 2018.

[14] Z. Guo, X. Li, H. Huang, N. Guo, and Q. Li, "Deep learning-based image segmentation on multimodal medical imaging," *IEEE Trans. Radiat. Plasma Med. Sci.*, vol. 3, no. 2, pp. 162–169, Mar. 2019.

[15] Q. Zhang, Z. Cui, X. Niu, S. Geng, and Y. Qiao, "Image segmentation with pyramid dilated convolution based on resnet and U-net," in *Proc. Int. Conf. Neural Inf. Process. (ICONIP)*, Guangzhou, China, 2017, pp. 364–372.

[16] M. Z. Alom, M. Hasan, C. Yakopcic, M. T. Taha, and K. V. Asari, "Recurrent residual convolutional neural network based on U-Net (R2U-Net) for medical image segmentation," 2018, *arXiv:1802.06955*. [Online]. Available: https://arxiv.org/abs/1802.06955

[17] O. Oktay, J. Schlemper, L. Le Folgoc, M. Lee, M. Heinrich, K. Misawa, K. Mori, S. McDonagh, N. Y. Hammerla, B. Kainz, B. Glocker, and D. Rueckert, "Attention U-Net: Learning where to look for the pancreas," 2018, *arXiv:1804.03999*. [Online]. Available: https://arxiv.org/abs/1804.03999

[18] P. F. Jaeger, S. A. A. Kohl, S. Bickelhaupt, F. Isensee, T. A. Kuder, H.-P. Schlemmer, and K. H. Maier-Hein, "Retina U-net: Embarrassingly simple exploitation of segmentation supervision for medical object detection," 2018, *arXiv:1811.08661*. [Online]. Available: https://arxiv.org/abs/1811.08661

[19] P. F. Christ, A. M. E. Elshaer, F. Ettlinger, S. Tatavarty, M. Bickel, P. Bilic, M. Rempfler, M. Armbruster, F. Hofmann, M. D'Anastasi, W. H. Sommer, S.-A. Ahmadi, and B. H. Menze, "Automatic liver and lesion segmentation in CT using cascaded fully convolutional neural networks and 3D conditional random fields," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* Athens, Greece, 2016, pp. 415–423.

[20] N. Nguyen and S.-W. Lee, "Robust boundary segmentation in medical images using a consecutive deep encoder-decoder network," *IEEE Access*, vol. 7, pp. 33795–33808, 2019.

[21] P. Zaffino, G. Pernelle, A. Mastmeyer, A. Mehrtash, H. Zhang, R. Kikinis, T. Kapur, and M. F. Spadea, "Fully automatic catheter segmentation in MRI with 3D convolutional neural networks: Application to MRI-guided gynecologic brachytherapy," *Phys. Med. Biol.*, vol. 64, no. 16, Aug. 2019, Art. no. 165008. doi: 10.1088/1361-6560/ab2f47.

[22] Y. Weng, T. Zhou, Y. Li, and X. Qiu, "NAS-Unet: Neural architecture search for medical image segmentation," *IEEE Access*, vol. 7, pp. 44247–44257, 2019.

[23] Z. Zeng, W. Xie, Y. Zhang, and Y. Lu, "RIC-Unet: An improved neural network based on Unet for nuclei segmentation in histology images," *IEEE Access*, vol. 7, pp. 21420–21428, 2019.

[24] V. Zyuzin and T. Chumarnaya, "Comparison of Unet architectures for segmentation of the left ventricle endocardial border on two-dimensional ultrasound images," in *Proc. Ural Symp. Biomed. Eng., Radioelectron. Inf. Technol. (USBEREIT)*, Apr. 2019, pp. 110–113. doi: 10.1109/USBEREIT.2019.8736616.

[25] M. P. Shah, S. N. Merchant, and P. S. Awate, "MS-Net: Mixed-supervision fully-convolutional networks for full-resolution segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent. (MICCAI)*, Granada, Spain, 2018, pp. 379–387.

[26] Z. Wu, C. Shen, and A. van den Hengel, "Bridging category-level and instance-level semantic image segmentation," 2016, *arXiv:1605.06885*. [Online]. Available: https://arxiv.org/abs/1605.06885

[27] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Boston, MA, USA, Jun. 2015, pp. 3431–3440.

[28] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam, "Rethinking atrous convolution for semantic image segmentation," 2017, *arXiv:1706.05587*. [Online]. Available: https://arxiv.org/abs/1706.05587

[29] F. Zhang, N. Cai, G. Cen, F. Li, H. Wang, and X. Chen, "Image super-resolution via a novel cascaded convolutional neural network framework," *Signal Process., Image Commun.*, vol. 63, pp. 9–18, Apr. 2018.

[30] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Honolulu, HI, USA, Jul. 2017, pp. 2881–2890.

[31] F. Yu and V. Koltun, "Multi-scale context aggregation by dilated convolutions," 2015, *arXiv:1511.07122*. [Online]. Available: https://arxiv.org/abs/1511.07122

[32] R. Hamaguchi, A. Fujita, K. Nemoto, T. Imaizumi, and S. Hikosaka, "Effective use of dilated convolutions for segmenting small object instances in remote sensing imagery," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Lake Tahoe, NV, USA, Mar. 2018, pp. 1442–1450.

[33] L.-Y. Deng, *The Cross-Entropy Method: A Unified Approach to Combinatorial Optimization, Monte-Carlo Simulation and Machine Learning*, D. Apley, Ed. London, U.K.: Taylor & Francis, 2012, pp. 147–148. doi: 10.1198/tech.2006.s353.

[34] B. C. Russell, A. Torralba, K. P. Murphy, and W. T. Freeman, "LabelMe: A database and Web-based tool for image annotation," *Int. J. Comput. Vis.*, vol. 77, nos. 1–3, pp. 157–173, 2008.

[35] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2015, *arXiv:1412.6980*. [Online]. Available: https://arxiv.org/abs/1412.6980

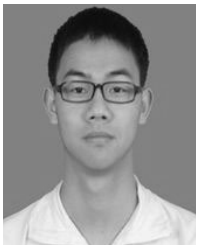
[36] *Keras: The Python Deep Learning Library*. Accessed: Jun. 5, 2014. [Online]. Available: https://keras.io/

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

XIAORUI FU received the B.E. degree from the College of Electronic Information, Guangzhou University, Guangzhou, China, in 2016. He is currently pursuing the M.E. degree in electronics and communication engineering with the Guangdong University of Technology, Guangzhou. His current research interests include medical segmentation and instance segmentation.

NIAN CAI received the B.Ed. degree in physics from Nanjing University, Nanjing, China, in 1999, and the Ph.D. Diploma degree in biophysics, a direct-promotion program, from the Institute of Biophysics, Chinese Academy of Sciences, Beijing, China, in 2004. He held a postdoctoral position in pattern recognition and intelligent system with Shanghai Jiao Tong University, Shanghai, China, in 2006. He is currently a Professor with the School of Information Engineering, Guangdong University of Technology, Guangzhou, China. His current research interests include machine learning, image processing, pattern recognition, computer vision, and signal processing and related areas.

KEMIN HUANG received the B.E. degree from the College of Automation, Guangxi University, Nanning, China, in 2017. He is currently pursuing the M.E. degree in electronics and communication engineering with the Guangdong University of Technology, Guangzhou, China. His current research interests include machine learning and computer vision.



HUIHENG WANG is currently pursuing the B.E. degree with the School of Information Engineering, Guangdong University of Technology, Guangzhou, China. His current research interests include deep learning and computer vision.

PING WANG is currently a Hepatobiliary Surgeon with The First Affiliated Hospital of Guangzhou Medical University, Guangzhou, China. He has long been engaged in the clinical and basic research of hepatobiliary and pancreatic surgery. He is good at surgical treatment of hepatobiliary and pancreatic tumors, bile duct stones and stenosis, bile duct injury, portal hypertension, and acute and chronic liver failure.

CHENGCHENG LIU received the B.E. degree from the College of Clinical Medicine, Yangtze University, Jingzhou, China, in 2015. He is currently pursuing the M.E. degree in surgery with Guangzhou Medical University, Guangzhou, China. His current research interests include diagnosis and treatment of common diseases in general surgery, and the operation of various minor operations. He is skilled in all kinds of general surgical laparoscopic surgery, especially percutaneous transhepatic hard choledochoscopy.

HAN WANG received the B.Ed. degree and the Ph.D. Diploma degree in mechanical engineering, a direct-promotion program, from the School of Mechanical and Electrical Engineering, Xiamen University, Xiamen, China, in 2003 and 2010, respectively. He is currently a Professor with the School of Mechanical and Electrical Engineering, Guangdong University of Technology, Guangzhou, China. His research interests include optical engineering, instrument engineering, and micro-nano processing technology and related areas.