IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received January 31, 2020, accepted February 12, 2020, date of publication February 24, 2020, date of current version March 3, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.2975983

HRED-Net: High-Resolution Encoder-Decoder Network for Fine-Grained Image Segmentation

CHENGZHI LYU, GUOQING HU, AND DAN WANG
School of Mechanical and Automotive Engineering, South China University of Technology, Guangzhou 510641, China
Corresponding author: Guoqing Hu (gqhu@scut.edu.cn)

ABSTRACT Accurate segmentation of fine-grained information is an important step in medical image analysis applications. With the development of the encoder-decoder-based networks, various network structures and algorithms have made significant progress in semantic segmentation tasks. This work aims to present a novel high-resolution encoder-decoder network (HRED-Net) for fine-grained image segmentation that is highly accurate for small-scale targets. We design a multiscale context connection module to extract feature information without reducing the resolution, and propose a multiresolution fusion model to fine-tune the final results. In addition, these modules are trained together with a detail-oriented loss function to enhance the model's perception of fine-grained parts. Through experiments on the DRIVE dataset, we found a balance between these modules, and our comparison results show that in addition to the extraction multiscale features, the fusion of multiresolution prediction information is also beneficial for fine-grained segmentation. Our method yielded significant improvements in the accuracy and sensitivity in retinal vessel and lung segmentation tasks.

INDEX TERMS Fine-grained, multiscale, multiresolution, retinal vessel, semantic segmentation.

I. INTRODUCTION

Focusing on the details of image segmentation is an ongoing challenge, and accurate segmentation of medical images, including shapes, locations, and sizes, provides scientific assistance to doctors for making accurate diagnoses. Convolutional neural networks (CNN) based algorithms have made important contributions to the field of medical imaging, involving various aspects such as retinal blood vessel segmentation [1]–[4], pathological slice segmentation [5]–[7], organ segmentation [8]–[10], and tumor segmentation [11]–[13].

Due to the limitations of the standardization of clinical data collection programs and some manual interventions in the data collection process [14], fine-grained segmentation [15] of medical images is challenging. The first limitation is low tissue contrast: fine-grained targets tend to be similar to background pixel values, causing inconsistencies or disappearance at the extended end. The second limitation is noise interference: due to the similar physical properties at organizational junctions, and flowing tissue fluid, medical images are often accompanied by impurities and uncertainty shadows. The third limitation is that the ratio between foreground and background is unbalanced: fine-grained targets, such as tumors, blood vessels, and nerves, are more worthy of attention in medical images, and they are insignificant in images. For easy understanding, we select the retinal vessel segmentation task as an example, although similar situations exist in other tasks, such as organ and tumor segmentation. In Fig. 1, the left panel shows the collected retinal vessel images and the right panel shows the corresponding ground truth. We confirm the previous view in three aspects: 1) the selected parts of the box have low contrast and blurred blood vessel contours, and the targets are interrupted at the position indicated by the arrows, 2) irregular shadows are distributed throughout the image, and 3) the foreground occupies the image at a ratio of less 0.1.

To address this problem, many supervised and unsupervised segmentation methods have been proposed [16]–[21], including threshold processing, level-set, maximum entropy partition, and manual marking method. These methods, however, have a large dependence on the pixels in the region, and it is difficult to distinguish some fuzzy regions. Recently, many researchers have made various attempts with deep learning methods, and they have proposed new ideas for image segmentation. These studies can be grouped

The associate editor coordinating the review of this manuscript and approving it for publication was Jon Atli Benediktsson.

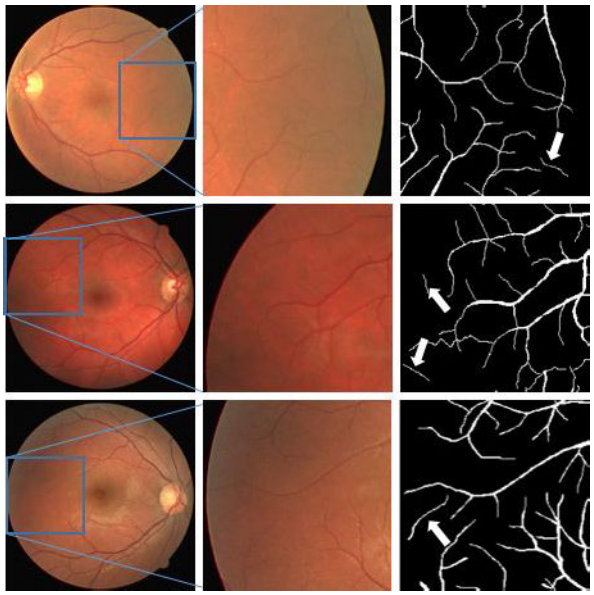


FIGURE 1. Illustration of the adjacent pixels in the foreground and background within retinal vessel images. First column: original images; second column: partially enlarged images; third column: ground truth of the identification area.

into three categories: 1) those proposing encoder-decoder feature extraction structures [5], [16] and implicitly using multiresolution features, 2) those proposing multiple resolution [22] combinations and explicit combinations of multiresolution information, and 3) those proposing multimodal feature extraction methods, including expanding the network width [23], increasing the network depth [24], and increasing the receptive fields of convolution [25].

We extract different scales of information at a resolution to extend the existing encoder network; additionally, we add an explicit fusion of multiresolution information to fine-tune the final results. During training, we also use a detail-oriented loss function to improve the sensitivity. We summarize the main contributions in the following five points:

1) Through careful analysis and experimental verification, an encoder module with residuals is used to extract semantic depth information and improve the semantic segmentation capability of the encoder-decoder structure.
2) A multiscale detail enhancement module is proposed to extract deep semantic information without reducing the resolution, and put it in the correct location through careful analysis and experiments to separate fine-grained targets.
3) We provide a shortcut between the low-resolution prediction maps to the final prediction and uses them to fine-tune the final results.
4) We propose a detail-oriented loss function that combines the weighted cross-entropy loss function and the Dice loss function to focus on the fine-grained parts.
5) We compare of the U-Net, SegNet, and context encoder network (CE-Net) with our proposed network under the same DRIVE dataset inputs, and implement extensive comparisons on other retinal vessel and lung segmentation tasks with the state-of-the-art methods.

II. RELATED WORKS

With the development of deep learning, CNNs have facilitated medical image segmentation tasks. To optimize the details of the segmentation targets, two strategies are often used in the literature: 1) improving the feature recognition and semantic reasoning capabilities of the network and inferring the attribution of pixel points by learning local and global information, and 2) improving the prediction ability of the network with multiresolution features and effectively integrating local and global context information.

To improve the logical reasoning and expression ability of the network, researchers have explored the use of patch-based CNNs in end-to-end learning to show the dawn in engineering applications. As a representative contribution, Long and Darrell. [26] made a major breakthrough when fully convolutional networks (FCNs) were introduced to address pixelwise prediction problems. FCNs define a skip layer that concatenates a deep, coarse layer with a global context and a shallow, fine layer with high-frequency details, leading to sharper boundaries between different classifications. Then, Ronneberger *et al.* [5] proposed U-Net with a 13-layer Visual Geometry Group (VGG13) framework, and Badrinarayanan *et al.* [27] proposed SegNet, which is topologically identical to the VGG16 framework; they both collected information on different resolutions for pixelwise segmentation, and these methods work well in medical segmentation tasks in small datasets. The skip connection provides a bridge for direct delivery of different resolution information between the encoder and decoder paths.

The deep convolution algorithm also has good performance in the medical field. To take full advantage of the different network levels of prediction results in one network, Guo *et al.* [28] and Liskowski and Krawiec [1] predicted the same resolution at different stages of the short-connection network. This short-connection approach passes low-level semantic information to a higher level to refine the high-level prediction and passes structural information to the lower-level to reduce the noise at the lower-level. Gu *et al.*[29] combined an inception module and dilated convolutions to form a context extraction module that links the encoder and decoder parts and captures more high-level information through the field of a different branch.

To improve the segmentation details, multiresolution fusion is widely used as an effective means of medical semantic segmentation. In reference [30], multiple U-Nets were connected into a chain, in which different resolution prediction maps were reused to improve the final accuracy. Feng *et al.* [31] proposed a more complicated method for connecting prediction graphs at different stages, in which the prediction in the primary path and the two branch paths are cross-connected, exhibiting strong robustness in image segmentation tasks.

Typically, better results can be obtained via hybrid approaches. In reference [9], a high-resolution pathway block was used as a skip-connection to fine-tune the final prediction map, and low-resolution prediction maps were also used to improve the top resolution. Both the dilated convolution kernels and low-resolution information were combined to obtain feature information with high recognition accuracy. To solve the problem of fuzzy boundary detection, Xie [32] proposed holistically-nested edge detection (HED), gradually reducing the resolution of the predicted images by FCNs, and then fusing them with weights. Lin *et al.* [22] exploited a multiscale pyramidal model and defined one pyramid for each feature map; the top-down convolutional pathway produces strong semantic information, and the bottom-up convolutional pathway yields accurate activation of the local information. These edge detection methods are important references for semantic segmentation.

In fine-grained segmentation tasks, Mavroudi *et al.* [33] used a temporal conditional random field module for fine-grained action segmentation. Zhao *et al.* [34] extracted fine-grained information with an improved pyramid neural network. To further improve the temporal convolutional encoder-decoder network, Nie and Shen [35] proposed a semantic-guided method to acquire accurate boundary information. In addition, many researchers have enhanced semantic segmentation by means of large receptive fields, Vo and Verma [36] combined two deep convolutions with multiple filter sizes for identifying fine-grained features. Zhou *et al.* [37] integrated fine-grained information from multiple scales with parallel multiresolution modules. Yang *et al.* [38] proposed a multiscale recurrent neural network to refine the details of boundary shapes.

The above literature has contributed to the semantic segmentation of medical images, and we can learn from these studies to further improve the segmentation performance. In the next sections, we describe a method to solve the problem and show the effectiveness of the proposed method through experimental comparisons.

III. METHOD

In this section, we describe our high-resolution encoder-decoder network (HRED-Net) in detail, and the entire process is shown in Fig. 2. The proposed network consists of four strategies, each of which is introduced in a separate section. First, an enhanced feature extraction module similar to U-Net is constructed, in which the feature information of different resolutions is fully extracted. Next, and most importantly, a multiscale pathway is used to improve the extraction of image details. Then, we elaborate on the multiresolution fusion module, sufficiently in the refined network for accurate segmentation targets. Finally, we force the network to focus more on the medical image features and design a hybrid loss function for fine-grained parts.

A. ENHANCED FEATURE EXTRACTION MODULE

Our feature extraction module learns from the efficient encoder-decoder structure [5], which provides local and global context information by extracting features at multiple resolutions.

To extract higher-resolution features, we use deeper convolution layers than U-Net. As shown in Fig. 2, each encoder module in our method is a three-layer convolution with a residual structure. The residual structure is easy to optimize [24]. Each convolution kernel uses batch normalization [39] and a rectified linear activation unit before the weight layer [40]. After that, max pooling is performed to achieve translation invariance at a low resolution.

B. MULTISCALE PATHWAY

Our second block is a multiscale pathway that connects the encoder and decoder. This module is designed to obtain high-resolution feature information and improve the perception of fine-grained parts.

Although the pooling operation has the characteristic of translation invariance, the resolution of the image is gradually reduced, which makes small targets disappear and hard to identify in the underlying layers. The current solution is the skip connection pathway which combines the coarse layers with the corresponding fine layers. However, regarding fine-grained segmentation targets, the deep layers cannot generate semantic information for the disappearing parts, which causes negative effects on the continuity of the segmentation target. To overcome this limitation, we propose a multiscale module. As shown in Fig. 1, we first perform a 1×1 convolution, a 3×3 convolution, a 7×7 convolution and 3×3 max pooling in a parallel, and in this way, we extract features of different scales at the same resolution. Then, we apply a 1×1 convolution to increase the depth of the network in every branch. Finally, we superimpose the features of all branches and use a 3×3 convolution for dimensionality reduction. Since our goal is to increase the sensitivity of the network to small isolated parts, no larger convolutions such as 7×7 convolutions or 9×9 convolutions, are used here.

Our module draws on the advantages of the Inception module [23], [41]. This operation has three advantages: 1) it expands the depth of the network and increases the nonlinearity of the network, 2) it increases the width of the network and improves the adaptability of the network to different scales, and 3) it operates at the same resolution with large receptive fields, to obtain a wider range of information without losing details.

C. MULTIRESOLUTION FUSION MODULE

Our third block is a weighted multiresolution fusion module for fine-tuning the final results. The multiresolution fusion module is designed to effectively fuse high-level global features and low-level local features to refine high-resolution output maps.

Fig. 4(a) and Fig. 4(b) represent two multiresolution network strategies; the difference between them is whether they explicitly utilize multiresolution prediction maps. By combining these two structures, we propose a multiresolution fusion module, which is shown in Fig. 4(c).

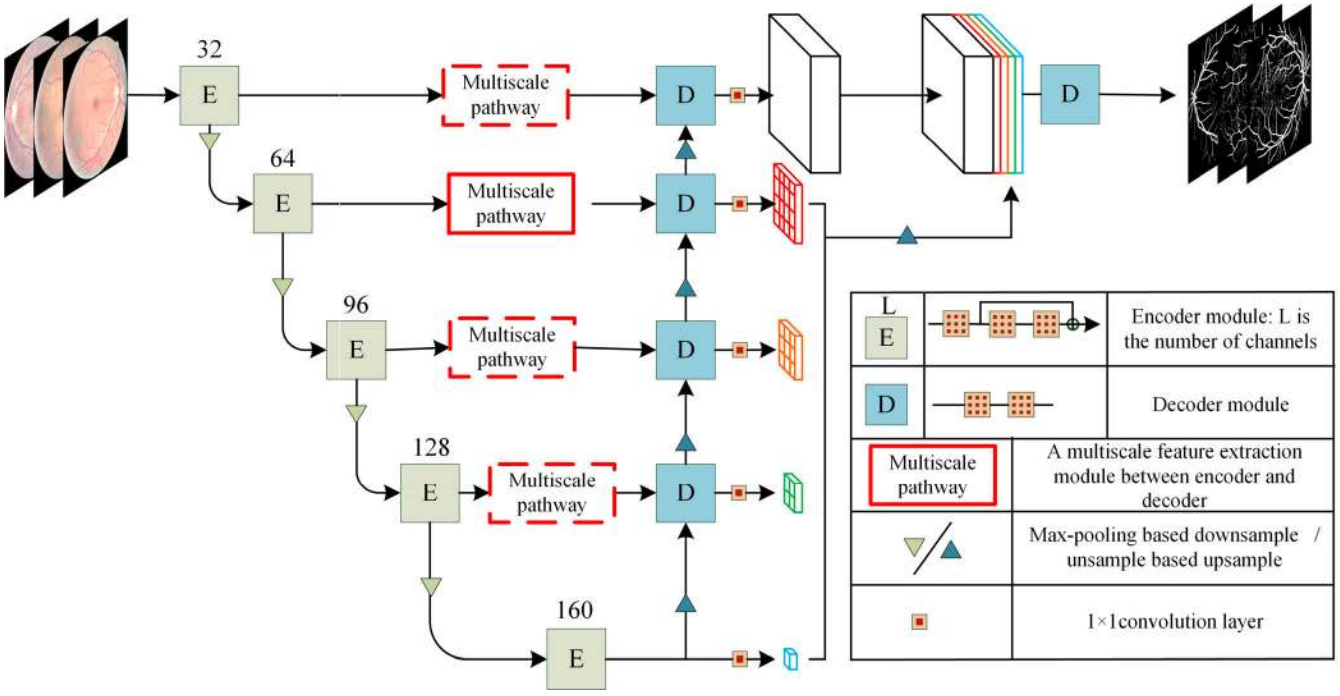


FIGURE 2. The proposed high-resolution encoder-decoder network (HRED -Net); our network consists of four modules: 1) an encoder module with residual E, 2) a multiscale feature extraction module, 3) a decoder module D, and 4) a multiresolution fusion module.

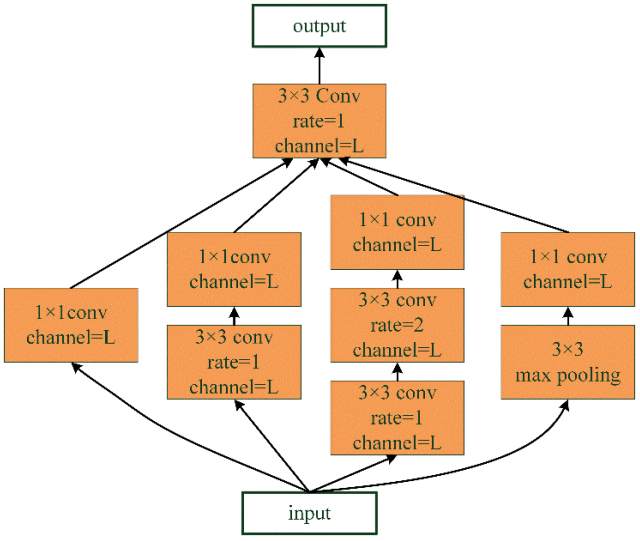


FIGURE 3. Illustration of the multiscale pathway block, which contains four branches with different receptive fields, and the input and output have the same number of channels.

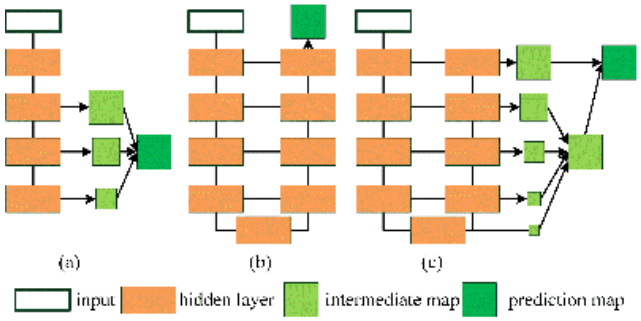


FIGURE 4. Illustration of the different multiresolution fusion structures: (a) holistically-nested network, (b) encoder-decoder feature extraction, and (c) our proposed multiresolution fine-tuning network.

In U-Net, the encoder provides a channel for extracting semantic information from different scales, the decoder implements a refinement process, and the skip connections combine coarse, deep features with fine shallow features [42]. In this manner, U-Net allows the depth semantic features to guide the subsequent fusion but does not explicitly produce multiresolution predictions. Comparatively, HED explicitly produces predictions for each level of features, and then a weighted fusion layer automatically combines outputs from multiple scales.

To inherit the advantages of the two networks, we followed the main outlines of the U-Net architecture and then fine-tuned the baseline with a multiresolution structure. Deep layers are helpful for instance detection. In U-Net, the transfer of deep features to the top is a long process, and the information is constantly adjusted during communication with the upper layers. Furthermore, semantic segmentation is a process of feature aggregation, which is assigned heuristically, and information discarded at other levels may contribute to the final prediction. Therefore, a shortcut is needed for the deep features to yield the final prediction.

Inspired by the structure of HED and the pyramid scene parsing networks [43], we provide a multiresolution fusion model for the deep layers. The difference from HED is the weight distribution for each resolution; we distinguish the top layer from the other layers; the top prediction map

is dominant, and the other prediction maps are secondary. As shown in Fig. 2, we add a 1 × 1 convolution before each prediction map, which not only automatically selects the scale, but also changes the number of channels. We convolve the number of channels from low-resolutions to a quarter of the original resolution [44] and superimpose them onto the original resolution prediction map; in this way, a shortcut to the final result is provided for results of different resolutions.

D. DETAIL-ORIENTED LOSS FUNCTION

As an end-to-end segmentation framework, our target is to train the proposed network to predict where each pixel belongs; i.e., this is a pixel-level classification problem. As a two-class task, the pixel values near the boundary are usually very similar, making the model easy to misclassify. Therefore, modeling the task as a regression problem is more accurate than a modeling it as a classification problem, which estimates the probability of each pixel belonging to the target.

The ratio of foreground to background is often unbalanced in medical images. Li *et al.* [45] showed that not all pixels are equal and that more power is given to the interesting pixels; to balance the pixel frequency between the region of interest and the background, we choose the weighted cross-entropy loss.

$$L_{wce}\,(\mathrm{p},\mathrm{y}) = -\sum\nolimits_{i\epsilon\Omega}(\alpha y_i log p_i + (1-y_i)\log(1-p_i)) \quad (1)$$

Here, Ω is the total number of pixels; p_i and y_i are the predicted probability of positive samples and the sample value of pixel i, and α denotes the weight of the positive pixels.

However, the Dice coefficient loss makes sense for clinical application, as more focus should be on the overlap between the prediction and fact in medical images. It is defined as follows:

$$L_{dice}(p, y) = 1-\frac{2\sum_{i\in\Omega} p_i.y_i}{\sum_{i\in\Omega} p_i^2 + \sum_{i\in\Omega} y_i^2} \quad (2)$$

Then, we define the joint morphological segmentation as follows:

$$L_{seg}(p, y) = \lambda L_{wce}(p, y) + L_{dice}(p, y) \quad (3)$$

Here, λ is the balanced weight. In our experiments, we set $\lambda = 1.5$ to obtain preferable results according to experience.

IV. EXPERIMENTS AND RESULTS

In this section, we evaluate the proposed method through experimental comparisons. First, we describe in detail the processing of the raw data, the implementation details, and the evaluation metrics. Then, we evaluate the modules we have proposed using the DRIVE dataset. Finally, we demonstrate the effectiveness by comparing our method with the state-of-the-art methods.

A. DATASET AND IMAGE PREPROCESSING

There are many publicly available benchmark datasets for image segmentation. We evaluate the proposed algorithm on three datasets: DRIVE [46], STARE [47], and LUNA. The DRIVE database contains 20 RGB training images and 20 testing images with the size of 565 × 584 (the retinal vessel occupies a radius of 540 pixels). We cropped the original images to 544 × 544 pixels with a reference mark, and then training and testing are performed for all of the images of this size. The STARE dataset contains 20 retinal fundus images with labels. We automatically generated marks based on the images. The original image size is 700 × 605 pixels, and we cropped the images to 656 × 544 pixels, and trained and tested images of this size. The LUNA dataset contains 534 2D samples with corresponding labels, and all images have a resolution of 512 × 512. The first two datasets are blood vessel segmentation datasets in which the segmentation targets are fine-grained and scattered, and the last dataset is an organ segmentation task, which is used to verify the generalization ability of our proposed network.

Due to the limitations in the number of training images, we augmented the dataset to enhance the expressive power of the training data. First, we adjusted the orientation of the images, including vertical flipping and horizontal flipping. Next, image preprocessing, which mainly involved random rotation, random shear, width shift and height shift, was implemented. Finally, we randomly processed the training images, including scaling from 0.9 to 1.1 and channel transformation. We also adopted noise reduction strategies on all the training and testing images. All the images underwent normalization, and then contrast-limited adaptive histogram equalization (CLAHE) and gamma correction were performed. The effect of noise reduction on the image is shown in Fig. 5. The segmentation targets become clearer, and the difference in the brightness decreases from the top row to the bottom row.

B. IMPLEMENTATION DETAILS

The dataset we use has the same standard training set and test set. All the pictures are subjected to the same image preprocessing and cropping processes. On this basis, the training data are subjected to additional amplification processing to compensate for the lack of data. We randomly divided the expanded training data into 80% for training and 20% for verification. All training ends with the early stop method. We also adopt the adaptive momentum (Adam) optimizer [48] with a learning rate of 0.0001, and step size hyperparameters $\beta_1 = 0.9$ and $\beta_2 = 0.999$. All the experiments are run on an NVIDIA Titan XP GPU.

C. EVALUATION METRICS

Subtle differences determine the quality of fine-grained segmentation, and how to select an evaluation indicator is very important to evaluate the segmentation results effectively. In this respect, we draw on the experience of other researchers. Many researchers have provided us with references for fine-grained segmentation, Angelova and Zhu [49] and Zhao *et al.* [50] used the accuracy as the metric of fine-grained segmentation, Zhang *et al.* [51] chose the precision and recall as metrics. We choose the accurate (Acc) as an

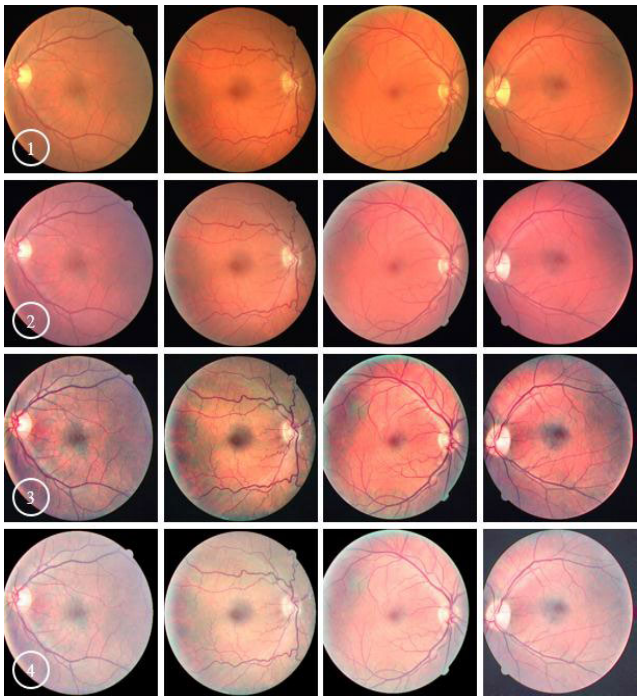


FIGURE 5. Image preprocessing details. Each line represents a different operation: 1) original images, 2) normalized images, 3) CLAHE corrected images, and 4) gamma corrected images.

indicator to evaluate fine-grained segmentation and use the sensitivity (Sen) and specificity (Spe) instead of the precision and recall for the medical images.

The accuracy is widely used to measure the percentage of correctly predicted pixels and is defined as:

$$\text{Acc} = \frac{TP + TN}{TP + TN + FP + FN} \tag{4}$$

Here, TP, TN, FP, and FN represent the numbers of true positives, true negatives, false positives, and false negatives, respectively.

The sensitivity and specificity often appear in pairs and are important metrics for binary segmentation. They are used to evaluate the correct prediction of foreground and background ratios; and are defined as:

$$\text{Sen} = \frac{TP}{TP + FN} \tag{5}$$

$$\text{Spe} = \frac{TN}{TN + FP} \tag{6}$$

In terms of model evaluation, we use the area under the curve (AUC) as a metric. The AUC is widely used as an essential indicator to confirm the effectiveness of machine learning algorithms and is obtained by integrating the area under the receiver operating characteristic (ROC) curve. The ROC curve, is the ratio of the true positive rate (TPR) to the false-positive rate (FPR).

$$\text{AUC} = \frac{1}{2} \sum_{i=0}^{m-1} (x_{i+1} - x_i)(y_i + y_{i+1}) \tag{7}$$

TABLE 1. Results comparison between U-Net and the multiscale pathway module in different locations.

Model	Sen	Spe	Acc	AUC	Parameter(M)
U-Net [5]	0.7924	0.9624	0.9565	0.9705	7.765
SED-Net 1	0.8673	0.9721	**0.9646**	0.9782	**2.396**
SED-Net 2	**0.8685**	**0.9735**	0.9633	**0.9786**	2.599
SED-Net 3	0.8550	0.9728	0.9651	0.9754	2.938
SED-Net 4	0.8643	0.9730	0.9645	0.9781	3.411

Here, x is the FPR and y is the TPR. We calculated the AUC by using the implementation provided in the scikit-learn Python library.

D. ABLATION ANALYSIS OF THE PROPOSED MODULE

To evaluate the performance of the proposed module, we designed a series of experiments for training and testing them on the DRIVE dataset. We first evaluate the effectiveness of the proposed modules and the loss function to determine the optimal combination, and then further confirm the performance with some classic networks under the same conditions.

1) EVALUATING THE MULTISCALE PATHWAY MODULE

To compare the effects of the multiscale pathway on different locations of the skip connection, we designed a comparative experiment. The experiment uses U-Net as a reference object, and the multiscale pathway module is placed in the four skip connections of the network, called SED-Net 1, SED-Net 2, SED-Net 3, and SED-Net 4. The difference between them is the number of channels convolved, which is the same as the number of channels in the corresponding encoder module.

We trained and tested all the networks under the same conditions as described in section A, and the test results are shown in Table 1. The table lists the parameters of the network and the corresponding AUC score, accuracy, and sensitivity for each network. From the experimental results, we can conclude that 1) our network outperform U-Net, and 2) the module performs best at the second encoder. There are three reasons for this phenomenon. The first reason is the resolution: in SED-Net 3 and SED-Net 4, the reduction in the image resolution affects the identification of fine-grained parts. The second reason is the receptive fields: operating at the same resolution with large receptive fields, our method obtains a wide range of information without losing details. The third reason is the width and depth: expanding the depth of the network increases the nonlinearity of the network, increases the width of the network and improves the adaptability of the network to different scales.

For a more intuitive understanding of the effects of the proposed module, we compared the prediction maps between our proposed networks and U-Net in detail (see Fig. 6). We show the images predicted by the two algorithms as well as the original image, comparing the segmentation effects on the fine-grained target. In those selected samples, our prediction

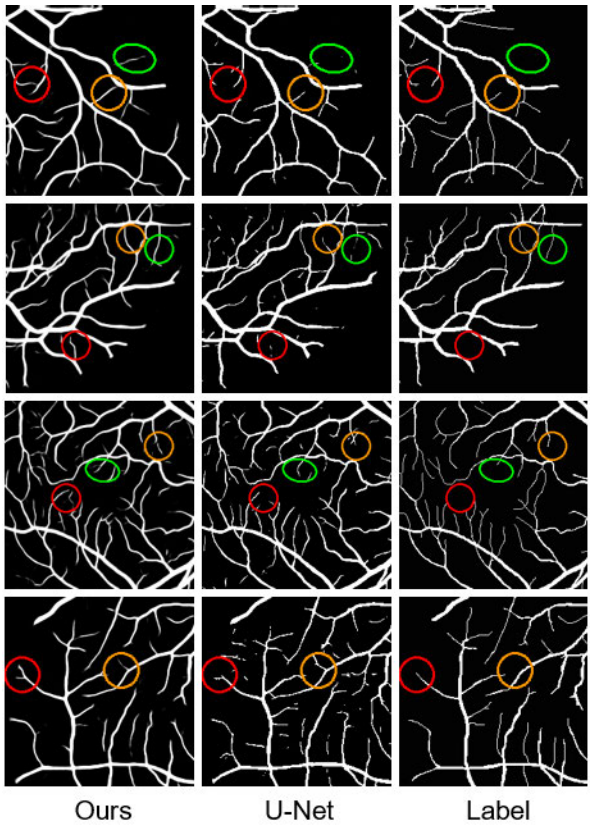


FIGURE 6. Examples of prediction maps showing the difference between our proposed network, U-Net and the ground truth. Each row shows a part of a prediction image, and the marked parts highlight the differences.

TABLE 2. Results comparison HRED-Net and SED-Net 2.

Model	Sen	Spe	Acc	AUC	Parameter(M)
SED-Net 2	0.8685	0.9735	0.9633	0.9786	**2.599**
HRED-Net	**0.8730**	**0.9742**	**0.9644**	**0.9796**	2.613

map performance was as follows: 1) high sensitivity to fine-grained parts (see the marked parts). We recognized the fine-grained contours and maintained the integrity. 2) Our method is better for low-contrast pasts than U-Net, referring to the parts marked by ellipses, where the contrast is low between the foreground and background, we segmented the blood vessel contours accurately.

2) EVALUATING THE MULTIRESOLUTION FUSION MODULE

To verify the performance of our multiresolution fusion module, we added a fusion part to the SED-Net 2 networks and called this network HRED-Net. We trained and tested the two networks under the same conditions as described in section A, and the results are shown in Table 2. The table lists the parameters of the network and the corresponding AUC score, accuracy, and sensitivity for each network.

From Table 2, we conclude that the fusion module 1) yields an approximate 0.5% improvement in the sensitivity with a 0.014 M parameter increase, and 2) increases both the AUC score and accuracy by 0.1%. This result proves the effectiveness of the multiresolution fusion module, indicating that there is a slight fine-tuning effect on the final prediction results by directly connecting the multiresolution prediction results.

TABLE 3. Comparison of the different loss function on the drive dataset.

Loss function	Sen	Spe	Acc	AUC
WCE	0.8615	0.9679	0.9597	0.9763
Dice	0.8692	0.9720	0.9602	0.9783
Detail-oriented	**0.8730**	**0.9742**	**0.9644**	**0.9796**

TABLE 4. Comparison of the segmentation results from different algorithms with the same inputs.

Model	Sen	Spe	Acc	AUC	Parameter(M)
U-Net [5]	0.7924	0.9624	0.9565	0.9705	7.765
SegNet [27]	0.8254	0.9707	0.9605	0.9734	10.963
CE-Net [29]	0.8537	0.9672	0.9583	**0.9815**	29.003
HRED-Net	**0.8730**	**0.9742**	**0.9644**	0.9796	**2.613**

3) EVALUATION OF THE DETAIL-ORIENTED LOSS FUNCTION

To evaluate the effectiveness of the detail-oriented loss function on the proposed network, we designed and tested a comparative experiment with different loss functions. From the numerical results shown in Table 3, all four metrics (sensitivity, specificity, accuracy, and AUC) showed a bright contrast in the DRIVE dataset, and the Dice coefficient loss performed better than the weighted cross-entropy loss. The detail-oriented loss function achieved the best segmentation scores. The sensitivity improved the most (0.4%), the specificity and accuracy improved by more than 0.2%, and the AUC score improved by 0.1%.

4) COMPARISON WITH THE STATE-OF-THE-ART METHODS

To further validate the effectiveness of our proposed module, we compare the proposed HRED-Net with the state-of-the-art algorithms on the same processed images described in section A. We run the code provided by the authors on the preprocessed images, and the results are shown in Table 4.

In Table 4, our network still has better realization than other methods under the same conditions with fewer parameters and shows the best accuracy and sensitivity, reflecting the advantages in fine-grained segmentation, mainly due to the following two points: 1) our loss function focuses more on the segmentation details, 2) our network handles prediction information in a multiscale and multiresolution way, enhancing the segmentation details.

Intuitive effects can be reflected by predictive comparisons, as shown in Fig. 7. All networks can segment the trunk of the blood vessel well and show differences in the small branches. U-Net and SegNet have visible breaks and discontinuities in the small parts, CE-Net's segmentation results are more complete than the results from the other methods but

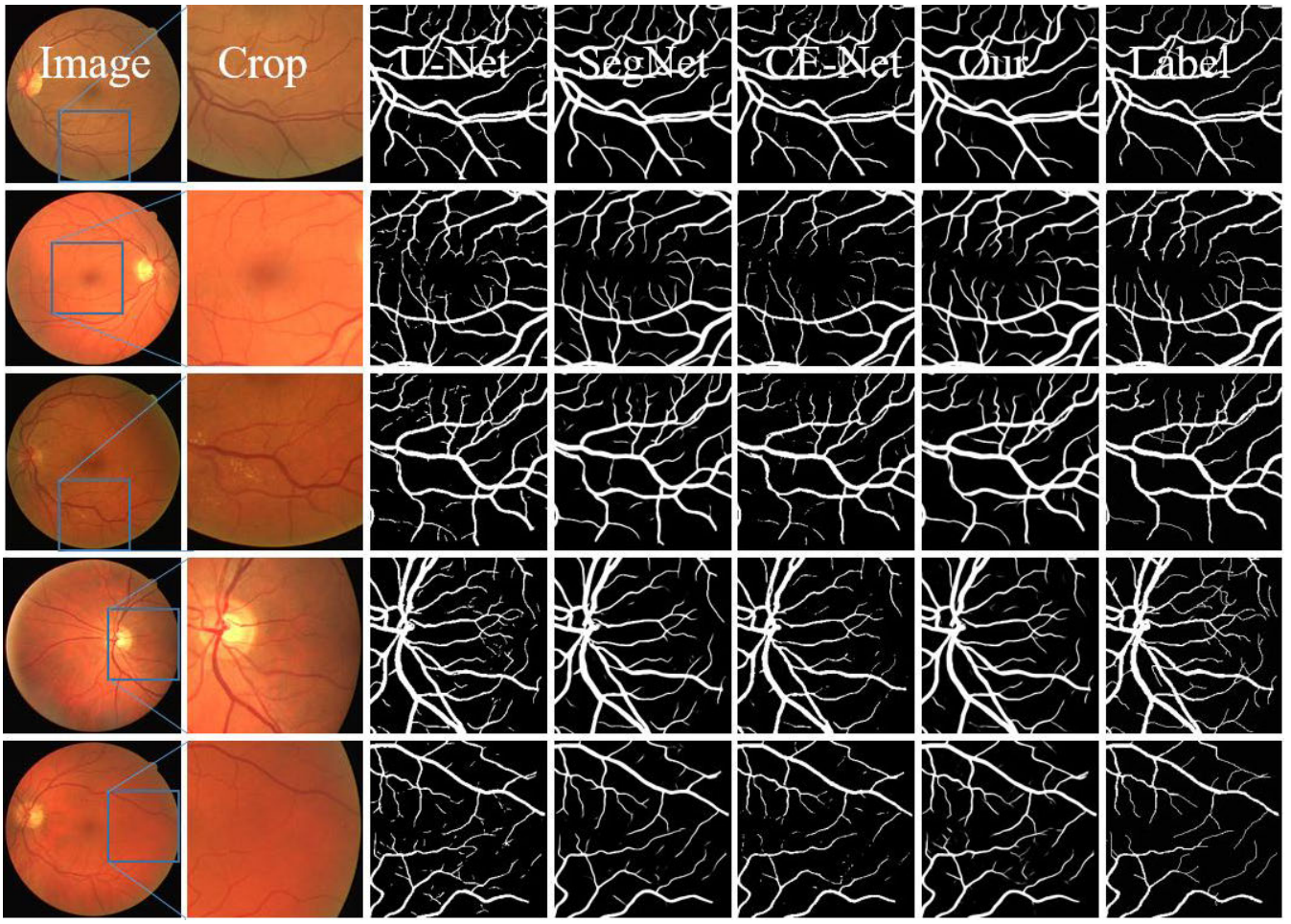


FIGURE 7. Examples of the prediction maps, from left to right: preprocessed images, state-of-the-art predictions obtained by U-Net, SegNet, CE-Net, HRED-Net and the ground-truth masks; all of the networks were trained with the same inputs.

lack the necessary details, and our predictions are closer to the label in the details than the other methods.

E. RESULTS

To further prove the validity of our proposed network, we experimented on different datasets and compared our results with other state-of-the-art approaches. All of the datasets were trained and tested under the same conditions as described in section B.

First, we compared our method with other machine learning methods on the DRIVE dataset and listed the results of the human observer. The results of the human observer come from people trained by an ophthalmologist [44], which can be used to measure of the effectiveness of machine learning methods. From Table 5, we can see that our proposed network increases the sensitivity from 0.8309 to 0.8730 by 4.2% and the accuracy decreases from 0.9576 to 0.9644, and we also achieve a sensitivity of 0.9742 and an AUC of 0.9796.

Then, we evaluated the accuracy and sensitivity of our method with the state-of-the-art algorithms on the STARE dataset. From the comparison results shown in Table 6, we can see that our proposed method achieves a sensitivity of 0.8044, a specificity of 0.9862, an accuracy of 0.9640, and an AUC of 0.9830.

TABLE 5. Comparison of segmentation results on the drive dataset.

Model	Year	Sen	Spe	Acc	AUC
Human Observer	-	0.7760	0.9725	0.9473	-
Zhang *et al.* [52]	2016	0.7861	0.9712	0.9466	0.9703
Orlando *et al.* [3]	2017	0.7897	0.9684	0.9454	0.9506
Oliveira *et al.* [2]	2018	0.8039	0.9804	0.9576	**0.9821**
Alom *et al.* [54]	2018	0.7792	**0.9813**	0.9556	0.9784
Jin *et al.* [55]	2018	0.7963	0.9800	0.9566	0.9802
Gu *et al.* [29]	2019	0.8309	-	0.9545	0.9779
HRED-Net	2019	**0.8730**	0.9742	**0.9644**	0.9796

Finally, we evaluated the performance on the LUNA database, which contains 534 2D samples with corresponding labels. All images have a resolution of 512 × 512. Different from the previous two datasets, the segmentation tasks are concentrated in this dataset. We compared the proposed method with U-Net [5], CE-Net [29] and the recurrent residual CNN based on U-Net (R2U-Net) [53]. From the

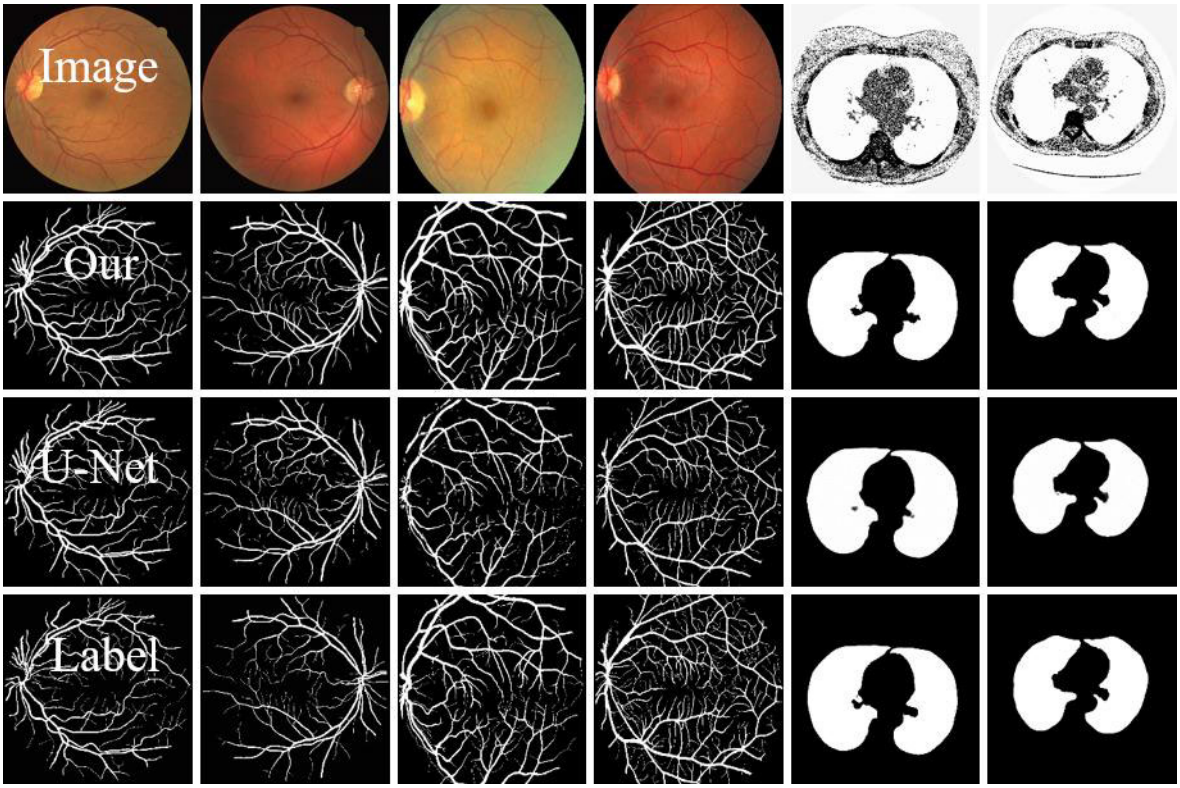


FIGURE 8. Sample results on vessel detection and lung segmentation dataset. Top to bottom are: original images, our proposed method, U-Net and ground truth.

TABLE 6. Comparison of the segmentation results on the stare dataset.

Model	Year	Sen	Spe	Acc	AUC
Human Observer	-	0.8956	0.9381	0.9346	-
Li *et al.* [55]	2016	0.7726	0.9844	0.9628	0.9879
Zhang *et al.* [52]	2017	0.7882	0.9729	0.9547	0.9740
Orlando *et al.* [3]	2017	0.7680	0.9738	0.9519	0. 9570
Oliveira *et al.* [2]	2018	**0.8315**	0.9858	0.9694	0.9905
Alom *et al.* [53]	2018	0.8298	0.9862	**0.9712**	**0.9914**
Jin *et al.* [54]	2018	0.7595	**0.9878**	0.9641	0.9832
HRED-Net	2019	0.8044	0.9862	0.9640	0.9830

TABLE 7. Comparison of the segmentation results on the LUNA dataset.

Model	Year	Sen	Spe	Acc	AUC
U-Net [5]	2015	0.9380	0.9872	0.9750	0.9784
Alom *et al.* [53]	2018	0.9832	**0.9944**	0.9918	**0.9889**
Gu *et al.* [29]	2019	0.9800	-	0.9900	-
HRED-Net	2019	**0.9917**	0.9935	**0.9923**	0.9879

comparisons shown in Table 7, our HRED-Net increases the sensitivity value from 0.9832 to 0.9917, the accuracy decreases from 0.9918 to 0.9923, and our network also achieves an AUC of 0.9879 and a specificity of 0.9935.

Tables 5, 6, and 7 above show the scores achieved by each model, which illustrate the effectiveness of our proposed network. These results further demonstrate that our proposed modules and the detail-oriented loss function are beneficial for all the target segmentation tasks. To obtain a more intuitive understanding of these scores, we show the same example. From the results shown in Fig. 8, we displayed the original image, U-Net segmentation results, our segmentation results, and the ground truth. We can conclude that our proposed algorithm achieves clear results in retinal vessel segmentation and lung organ segmentation.

V. CONCLUSION

Details are essential for medical image segmentation. In this paper, we proposed a multiscale connection encoder-decoder network that focuses on fine-grained parts. Our network draws on the encoder-decoder structure from U-Net, and it consists of an enhanced encoder module, a multiscale fine-grained extraction module, a decoder module and a multiresolution fusion module. We also added a detail-oriented loss function to the network. In the multiscale pathway, we extracted the location information of the small targets by increasing the width and the receptive field of the convolution. Through comparative experiments, we found that the multiscale module performed best in the second branch. Moreover, the multiresolution fusion path facilitates direct participation in the final prediction for low-resolution images, retaining complete semantic features. Our comparative experiments show that the proposed method improves fine-grained

part segmentation with fewer parameters, including retinal vessel segmentation and lung CT segmentation.

REFERENCES

[1] P. Liskowski and K. Krawiec, "Segmenting retinal blood vessels with deep neural networks," *IEEE Trans. Med. Imag.*, vol. 35, no. 11, pp. 2369–2380, Nov. 2016.

[2] A. Oliveira, S. Pereira, and C. A. Silva, "Retinal vessel segmentation based on fully convolutional neural networks," *Expert Syst. Appl.*, vol. 112, pp. 229–242, Dec. 2018.

[3] J. I. Orlando, E. Prokofyeva, and M. B. Blaschko, "A discriminatively trained fully connected conditional random field model for blood vessel segmentation in fundus images," *IEEE Trans. Biomed. Eng.*, vol. 64, no. 1, pp. 16–27, Jan. 2017.

[4] A. Diaz-Pinto, A. Colomer, V. Naranjo, S. Morales, Y. Xu, and A. F. Frangi, "Retinal image synthesis and semi-supervised learning for glaucoma assessment," *IEEE Trans. Med. Imag.*, vol. 38, no. 9, pp. 2211–2218, Sep. 2019.

[5] 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.

[6] T.-H. Song, V. Sanchez, H. EIDaly, and N. M. Rajpoot, "Dual-channel active contour model for megakaryocytic cell segmentation in bone marrow trephine histology images," *IEEE Trans. Biomed. Eng.*, vol. 64, no. 12, pp. 2913–2923, Dec. 2017.

[7] A. Veillard, M. S. Kulikova, and D. Racoceanu, "Cell nuclei extraction from breast cancer histopathologyimages using colour, texture, scale and shape information," *Diagnostic Pathol.*, vol. 8, no. 1, p. S5, Sep. 2013.

[8] X. Li, H. Chen, X. Qi, Q. Dou, C.-W. Fu, and P.-A. Heng, "H-DenseUNet: Hybrid densely connected UNet for liver and tumor segmentation from CT volumes," *IEEE Trans. Med. Imag.*, vol. 37, no. 12, pp. 2663–2674, Dec. 2018.

[9] S. Zhou, D. Nie, E. Adeli, J. Yin, J. Lian, and D. Shen, "High-resolution encoder-decoder networks for low-contrast medical image segmentation," *IEEE Trans. Image Process.*, vol. 29, pp. 461–475, Jun. 2020.

[10] E. Gibson, F. Giganti, Y. Hu, E. Bonmati, S. Bandula, K. Gurusamy, B. Davidson, S. P. Pereira, M. J. Clarkson, and D. C. Barratt, "Automatic multi-organ segmentation on abdominal CT with dense V-Networks," *IEEE Trans. Med. Imag.*, vol. 37, no. 8, pp. 1822–1834, Aug. 2018.

[11] A. Hafiane, F. Bunyak, and K. Palaniappan, "Clustering initiated multi-phase active contours and robust separation of nuclei groups for tissue segmentation," in *Proc. 19th Int. Conf. Pattern Recognit.*, Dec. 2008, pp. 1–4.

[12] S. Chen, C. Ding, and M. Liu, "Dual-force convolutional neural networks for accurate brain tumor segmentation," *Pattern Recognit.*, vol. 88, pp. 90–100, Apr. 2019.

[13] A. Zhao, G. Balakrishnan, F. Durand, J. V. Guttag, and A. V. Dalca, "Data augmentation using learned transformations for one-shot medical image segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 8543–8553.

[14] S. E. Petersen, "UK Biobank's cardiovascular magnetic resonance protocol," *J. Cardiovascular Magn. Reson.*, vol. 18, no. 1, p. 8, Dec. 2016.

[15] M. G. Harris and C. D. Giachritsis, "Coarse-grained information dominates fine-grained information in judgments of time-to-contact from retinal flow," *Vis. Res.*, vol. 40, no. 6, pp. 601–611, Mar. 2000.

[16] M. Ranzato, F. J. Huang, Y.-L. Boureau, and Y. LeCun, "Unsupervised learning of invariant feature hierarchies with applications to object recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2007, pp. 1–8.

[17] A. A. Qizhu Li and H. S. Philip Torr, "Weakly-and semi-supervised panoptic segmentation," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 102–118.

[18] A. Kolesnikov and C. H. Lampert, "Seed, expand and constrain: Three principles for weakly-supervised image segmentation," in *Computer Vision—ECCV*. Cham, Switzerland: Springer, 2016, pp. 695–711.

[19] H. Cholakkal, J. Johnson, and D. Rajan, "Backtracking ScSPM image classifier for weakly supervised top-down saliency," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 5278–5287.

[20] 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.

[21] S. Jegou, M. Drozdzal, D. Vazquez, A. Romero, and Y. Bengio, "The one hundred layers tiramisu: Fully convolutional DenseNets for semantic segmentation," presented at the IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW), Jul. 2017.

[22] T.-Y. Lin, P. Dollar, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," presented at the IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jul. 2017.

[23] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 1–9.

[24] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[25] F. Yu, V. Koltun, and T. Funkhouser, "Dilated residual networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 472–480.

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

[27] 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.

[28] S. Guo, K. Wang, H. Kang, Y. Zhang, Y. Gao, and T. Li, "BTS-DSN: Deeply supervised neural network with short connections for retinal vessel segmentation," *Int. J. Med. Informat.*, vol. 126, pp. 105–113, Jun. 2019.

[29] Z. Gu, J. Cheng, H. Fu, K. Zhou, H. Hao, Y. Zhao, T. Zhang, S. Gao, and J. Liu, "CE-net: Context encoder network for 2D medical image segmentation," *IEEE Trans. Med. Imag.*, vol. 38, no. 10, pp. 2281–2292, Oct. 2019.

[30] J. Zhuang, "LadderNet: Multi-path networks based on U-Net for medical image segmentation," 2018, *arXiv:1810.07810*. [Online]. Available: http://arxiv.org/abs/1810.07810

[31] S. Feng, Z. Zhuo, D. Pan, and Q. Tian, "CcNet: A cross-connected convolutional network for segmenting retinal vessels using multi-scale features," *Neurocomputing*, Apr. 2019.

[32] S. Xie and Z. Tu, "Holistically-nested edge detection," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, 2015, pp. 1395–1403.

[33] E. Mavroudi, D. Bhaskara, S. Sefati, H. Ali, and R. Vidal, "End-to-End fine-grained action segmentation and recognition using conditional random field models and discriminative sparse coding," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Mar. 2018, pp. 1558–1567.

[34] X. Zhao, W. Sun, W. Qian, S. Qi, J. Sun, B. Zhang, and Z. Yang, "Fine-grained lung nodule segmentation with pyramid deconvolutional neural network," *Proc. SPIE*, vol. 10950, Mar. 2019, Art. no. 109503S.

[35] D. Nie and D. Shen, "Semantic-guided encoder feature learning for blurry boundary delineation," 2019, *arXiv:1906.04306*. [Online]. Available: http://arxiv.org/abs/1906.04306

[36] H. H. Vo and A. Verma, "New deep neural nets for fine-grained diabetic retinopathy recognition on hybrid color space," in *Proc. IEEE Int. Symp. Multimedia (ISM)*, Dec. 2016, pp. 209–215.

[37] S. Zhou, D. Nie, E. Adeli, Y. Gao, L. Wang, J. Yin, and D. Shen, "Fine-grained segmentation using hierarchical dilated neural networks," in *Medical Image Computing and Computer Assisted Intervention—MICCAI*. Cham, Switzerland: Springer, 2018, pp. 488–496.

[38] X. Yang, L. Yu, L. wu, D. Ni, J. Qin, and P. Heng, "Fine-grained recurrent neural networks for automatic prostate segmentation in ultrasound images," presented at the 31st AAAI Conf. Artif. Intell., San Francisco, CA, USA, 2017.

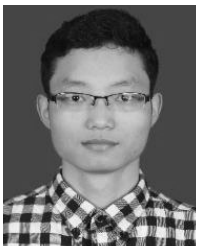
[39] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," presented at the ICML, 2015.

[40] K. He, X. Zhang, S. Ren, and J. Sun, "Identity mappings in deep residual networks," in *Computer Vision—ECCV*, Cham, Switzerland: Springer, 2016, pp. 630–645.

[41] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 2818–2826.

[42] C. Farabet, C. Couprie, L. Najman, and Y. LeCun, "Learning hierarchical features for scene labeling," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 8, pp. 1915–1929, Aug. 2013.

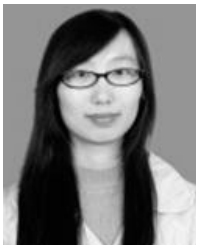
[43] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2881–2890.
[44] G. Lin, A. Milan, C. Shen, and I. Reid, "RefineNet: Multi-path refinement networks for high-resolution semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1925–1934.
[45] X. Li, Z. Liu, P. Luo, C. C. Loy, and X. Tang, "Not all pixels are equal: Difficulty-aware semantic segmentation via deep layer cascade," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 3193–3202.
[46] J. Staal, M. D. Abramoff, M. Niemeijer, M. A. Viergever, and B. van Ginneken, "Ridge-based vessel segmentation in color images of the retina," *IEEE Trans. Med. Imag.*, vol. 23, no. 4, pp. 501–509, Apr. 2004.
[47] A. D. Hoover, V. Kouznetsova, and M. Goldbaum, "Locating blood vessels in retinal images by piecewise threshold probing of a matched filter response," *IEEE Trans. Med. Imag.*, vol. 19, no. 3, pp. 203–210, Mar. 2000.
[48] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. ICLR*, 2015.
[49] A. Angelova and S. Zhu, "Efficient object detection and segmentation for fine-grained recognition," presented at the IEEE Conf. Comput. Vis. Pattern Recognit., Jun. 2013.
[50] B. Zhao, J. Feng, X. Wu, and S. Yan, "A survey on deep learning-based fine-grained object classification and semantic segmentation," *Int. J. Autom. Comput.*, vol. 14, no. 2, pp. 119–135, Jan. 2017.
[51] X. Zhang, H. Su, L. Yang, and S. Zhang, "Fine-grained histopathological image analysis via robust segmentation and large-scale retrieval," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 5361–5368.
[52] J. Zhang, Y. Chen, E. Bekkers, M. Wang, B. Dashtbozorg, and B. M. T. H. Romeny, "Retinal vessel delineation using a brain-inspired wavelet transform and random forest," *Pattern Recognit.*, vol. 69, pp. 107–123, Sep. 2017.
[53] M. Zahangir Alom, M. Hasan, C. Yakopcic, T. M. Taha, and V. K. Asari, "Recurrent residual convolutional neural network based on U-net (R2U-net) for medical image segmentation," 2018, *arXiv:1802.06955*. [Online]. Available: http://arxiv.org/abs/1802.06955
[54] Q. Jin, Z. Meng, T. D. Pham, Q. Chen, L. Wei, and R. Su, "DUNet: A deformable network for retinal vessel segmentation," *Knowl.-Based Syst.*, vol. 178, pp. 149–162, Aug. 2019.
[55] Q. Li, B. Feng, L. Xie, P. Liang, H. Zhang, and T. Wang, "A cross-modality learning approach for vessel segmentation in retinal images," *IEEE Trans. Med. Imag.*, vol. 35, no. 1, pp. 109–118, Jan. 2016.



CHENGZHI LYU is currently pursuing the Ph.D. degree with the South China University of Technology, China. His research interests are in computer vision based on image understanding and reconstruction and developing intelligent control systems based on machine learning.

GUOQING HU received the M.S. degree from Northwestern Polytechnical University, China, and the Ph.D. degree from Sichuan University, China.

He was a Professor with Xiamen University, China. He was an Advanced Visiting Scholar with The Chinese University of Hong Kong and the University of Nottingham. He is currently a Professor and a Ph.D. Student Supervisor with the School of Mechanical and Automotive Engineering, South China University of Technology, China. He was completed and participated in more than 90 projects, including the National 863 Project, the National Natural Science Foundation Project, the National Major Projects, the International Cooperation Projects, the Provincial Key Projects, the Province Fund Cooperation Projects. He has published 248 articles and two textbooks. He holds 22 patents. His research interests include amphibious flying machine, intelligent robot, industrial image processing, automation and industrial robot, electromechanical integration, and advanced sensor technology.



DAN WANG received the master's degree in mechanical engineering from Xiamen University, Xiamen. She is currently pursuing the Ph.D. degree with the South China University of Technology, Guangzhou, Guangdong, China. Her research interests are in computer vision and developing automated control systems based on image understanding and researching the machine vision-based mechanical automated processing and inspection systems.