Semi-Supervised Capsule cGAN for Speckle Noise Reduction in Retinal OCT Images

Meng Wang, Weifang Zhu, Kai Yu, Zhongyue Chen, Fei Shi, Yi Zhou, Yuhui Ma, Yuanyuan Peng, Dengsen Bao, Shuanglang Feng, Lei Ye, Dehui Xiang, *Member, IEEE*, and Xinjian Chen, *Senior Member, IEEE*

***Abstract*— Speckle noise is the main cause of poor optical coherence tomography (OCT) image quality. Convolutional neural networks (CNNs) have shown remarkable performances for speckle noise reduction. However, speckle noise denoising still meets great challenges because the deep learning-based methods need a large amount of labeled data whose acquisition is time-consuming or expensive. Besides, many CNNs-based methods design complex structure based networks with lots of parameters to improve the denoising performance, which consume hardware resources severely and are prone to overfitting. To solve these problems, we propose a novel semi-supervised learning based method for speckle noise denoising in retinal OCT images. First, to improve the model's ability to capture complex and sparse features in OCT images, and avoid the problem of a great increase of parameters, a novel capsule conditional generative adversarial network (Caps-cGAN) with small number of parameters is proposed to construct the semi-supervised learning system. Then, to tackle the problem of retinal structure information loss in OCT images caused by lack of detailed guidance during unsupervised learning, a novel joint semi-supervised loss function composed of unsupervised loss and supervised loss is proposed to train the model. Compared with other state-of-the-art methods, the proposed semi-supervised method is suitable for retinal OCT images collected from different OCT devices and can achieve better performance even only using half of the training data.**

***Index Terms*— Semi-supervision, capsule network, cGAN, optical coherence tomography, speckle noise.**

Manuscript received October 27, 2020; revised December 23, 2020; accepted December 27, 2020. Date of publication January 4, 2021; date of current version April 1, 2021. This work was supported in part by the National Key Research and Development Program of China under Grant 2018YFA0701700 and in part by the National Nature Science Foundation of China under Grant 61622114 and Grant 81401472. *(Meng Wang and Weifang Zhu contributed equally to this work.) (Corresponding author: Xinjian Chen.)*

Meng Wang, Weifang Zhu, Kai Yu, Zhongyue Chen, Fei Shi, Yi Zhou, Yuhui Ma, Yuanyuan Peng, Dengsen Bao, Shuanglang Feng, Lei Ye, and Dehui Xiang are with the School of Electronics and Information Engineering, Soochow University, Suzhou 215006, China (e-mail: wangmeng9218@126.com; wfzhu@suda.edu.cn; 578069383@qq.com; chenzy@suda.edu.cn; shifei@suda.edu.cn; zhouyi.zura@gmail.com; mayuhui@nimte.ac.cn; yypeng@stu.suda.edu.cn; baodengsen@126.com; 2470697802@qq.com; 1160776397@qq.com; xiangdehui@suda.edu.cn).

Xinjian Chen is with the School of Electronics and Information Engineering, Soochow University, Suzhou 215006, China, and also with the State Key Laboratory of Radiation Medicine and Protection, Soochow University, Suzhou 215006, China (e-mail: xjchen@suda.edu.cn).

Digital Object Identifier 10.1109/TMI.2020.3048975

I. Introduction

OPTICAL coherence tomography (OCT) is a non-invasive imaging technology proposed by Huang *et al.* [1], which can capture cross-sectional image of biological tissue and has been widely used. In the ophthalmology clinic for the diagnosis and monitoring of retinal diseases [2], [3]. Speckle noise is one of the most common noises generated in OCT imaging. Although the imaging technology and equipment have been continuously updated in recent years, the problem of speckle noise has not been solved very well, and it has seriously affected the performance of OCT image automatic analysis, such as retinal lesion region segmentation [4], [5], retinal layer information analysis [6]–[8] and registration [9]. Therefore, obtaining high-quality OCT images is essential to improve the performance of automatic analysis. Many hardware based methods, which depend on specially designed acquisition systems, have been proposed for speckle noise suppression during imaging. Iftimia *et al.* [10] proposed a high-speed method for implementing angular compounding by path length encoding (ACPE) for reducing speckle noise in OCT images. Kennedy *et al.* [11] presented a speckle reduction technique for OCT based on strain compounding. Based on angular compounding, Cheng *et al.* [12] proposed a dual-beam angular compounding method to reduce speckle noise and improve SNR of OCT images. However, these methods cannot be directly applied to commercial OCT scanners because they require specially designed acquisition systems. Recently, many algorithms have been proposed for speckle noise denoising in OCT images, which can be divided into two categories: traditional denoising algorithms and deep learning-based methods. In traditional speckle reduction methods, the partial differential equation (PDE) based methods such as anisotropic diffusion filtering are widely used in noise reduction [13], [14]. However, these methods have problems of overfitting and over-smoothing. Aum *et al.* [15] and Buades *et al.* [16] explored the non-local mean (NLM) based speckle noise denoising methods and achieved good performances both in visual effects and objective indicators. Bo and Zhu [17] proposed a wavelet modification based block matching and

3D filtering (BM3D) for speckle noise denoising in human finger skin OCT images. However, in the NLM-based method when the local regions can't be matched well, the edge information may be lost. The methods based on statistical model are also common for image denoising [18]. In addition, many sparse transform-based methods have achieved good performance in noise reduction tasks, such as adaptive wavelet thresholding [19], [20], curvelet transform [21], and dictionary learning based sparse representation [22], [23]. However, these methods still have some problems such as insufficient image feature representations, difficulties in choosing appropriate thresholds and time-consuming dictionary learning. Moreover, the low rank decomposition based methods are also effective for OCT image denoising [24], [25].

In last decade, deep learning based methods, especially convolutional neural networks(CNNs), have been widely used in image classification [26], [27], object detection [28], [29], and lesion region segmentation [30], [31]. Moreover, many CNNs based methods have achieved promising performance in image denoising as well. Mao *et al.* [32] proposed a very deep convolutional encoder-decoder network with symmetric skip connections for image restoration. Considering the long-term dependency problem of the model, Tai *et al.* [33] proposed a very deep persistent memory network (MemNet) and applied to image restoration. Zhang *et al.* [34] proposed a deep convolutional neural network based on residual learning (DnCNN) to suppress noise in natural images. Based on DnCNN, Cai *et al.* [35] further improved the denoising performance by introducing residual module and applied it to OCT image denoising. However, these methods are mainly used to reduce additive noise in images, which is quite different from the speckle noise in OCT images. To improve the performance of speckle noise reduction in OCT images, in our previous work, we [36] proposed a novel convolutional neural network named DeSpecNet, which combined residual learning and batch normalization to improve the network by using the shortcut connectivity blocks and leaky rectified linear units. In another our previous work, we [37] proposed an effectively method based on conditional generative adversarial networks (cGAN) to reduce the speckle noise in OCT images. Although these approaches have achieved impressive performance, there are still two major problems for CNNs-based method: (1) The fully-supervised method based on deep learning usually requires a large amount of labeled data, whose acquisition is usually time-consuming or expensive. (2) To improve the denoising performance, many CNNs-based methods have designed complex structure networks with a large number of parameters, which tends to cause the overfitting and poor generality of the networks.

In this article, to address these problems, we propose a novel semi-supervised method based on our newly proposed capsule conditional generative adversarial network called Caps-cGAN for speckle noise denoising in retinal OCT images, which can achieve outstanding denoising performance only using a small amount of labeled data and network parameters. The key contributions of this study are as follows:

(1) We propose a novel Caps-cGAN to develop a newly semi-supervised learning method for denoising the speckle noise, which can avoid the problem of a great increase in the amount of parameters.

(2) To alleviate the problem of retinal structure information loss in OCT images caused by the lack of detailed guidance during unsupervised learning, a novel joint semi-supervised loss function composed of unsupervised loss and supervised loss is designed to optimize the proposed network.

(3) We validate the effectiveness and generality of our proposed method by conducting comprehensive experiments on OCT images acquired from different types of OCT scanners. Results show that the proposed method outperforms the state-of-the-art methods in OCT image speckle noise reduction task.

II. Related Works

A. Semi-Supervised Learning (SSL)

The lack of labeled data has always been one of the biggest obstacles for applying CNNs-based methods to medical image processing tasks. SSL can use unlabeled data to improve the generalization performance of the supervised model [38]. To alleviate the dependence on labeled data, many SSL-based methods have been proposed and applied to medical image processing tasks, such as lesion region segmentation [38], disease detection [39] and registration [40]. However, to our best knowledge, there are no SSL-based methods for OCT image denoising.

B. Capsule Networks

Capsule network was first proposed by Sabour *et al.* [41]. Its biggest characteristic is to adopt vector to represent feature information, whose direction and amplitude indicate the attributes of the feature, such as posture, texture, etc. Therefore, the capsule network has better ability to capture the spatial relationship between features than standard CNNs. However, due to the high cost of memory and time consumption, the original capsule network was limited to image with small size. To overcome this shortcoming, Rodney and Bagci [42] extended the idea of convolutional capsules and rewrote the dynamic routing algorithm, and successfully applied the method to lung segmentation in CT image. Based on [42], Bass *et al.* [43] introduced the convolutional capsule network into the task of image synthesis. In addition, there were two previous studies which improved GAN by introducing non-convolutional capsule in discriminator [44], [45]. But both of them were only applied to image with small size 64×64.

III. Methods

A. Overview

In this article, we propose a novel semi-supervised method with small number of parameters for OCT image speckle noise reduction, which can outperform the state-of-the-art supervised approaches with less training data. Fig.1 shows our newly proposed semi-supervised learning system. Let's assume a

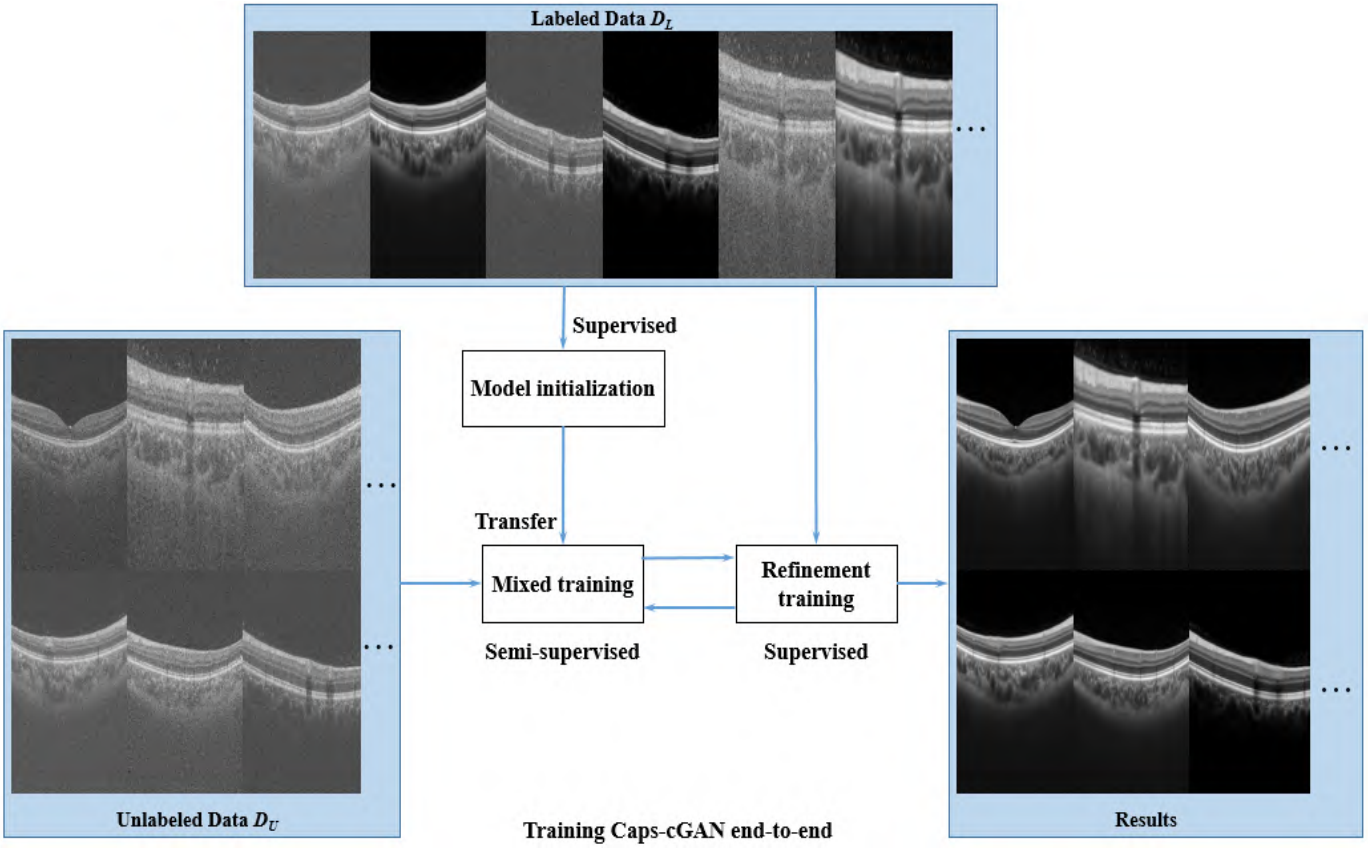

Fig. 1. Overview of the proposed semi-supervised system. D_L and D_U represent labeled data and unlabeled data, respectively.

total set of OCT images as $D_T = \{D_L, D_U\}$. $D_L = \{(x_i, y_i)\}$ and $D_U = \{x_i\}$ represent the labeled data and the unlabeled data respectively, in which x_i represents the original OCT image, while y_i is the corresponding ground truth of x_i. As can be seen from Fig.1, the proposed semi-supervised learning based model is optimized as follows:

(1) Model initialization: the model is first trained based on labeled data under the guidance of the fully-supervised objective function, which aims to guide the model to learn the distribution of labeled data and obtain the initial weights.

(2) Mixed training: the model is continuously trained with the initialized weights from step (1) using both a large amount of unlabeled data and the small amount of labeled data. The mixed training adopts the semi-supervised learning strategy, which composes of supervised learning based on small amount of labeled data and the un-supervised learning based on adversarial guidance optimization for large amount of unlabeled data.

(3) Refinement training: to avoid the mode crash caused by the diversity distribution of the unlabeled data and further improve the denosising performance, the model will be refinedly trained based on the small amount of labeled data again.

(4) Repeat the training process of (2) and (3) for several epochs. Finally, a model which can generate high-quality OCT images is obtained. Noting that the entire training process is end-to-end.

B. Capsule Conditional Generative Adversarial Network

GAN and its variants have been widely used due to its unique characteristics of adversarial game optimization, and have made promising achievements in many image processing tasks, such as image style transfer [46], target segmentation [30], [47], [48], and image manipulation [37], [49]. Different from the original GAN that generates image based on random noise, the cGAN generates the image conditioned on an observed image [37]. cGAN is mainly composed of two modules: the generator G that aims to generate the corresponding fake image based on the input image, and the discriminator D is used to distinguish whether the image is the real one or the generated one from G.

Generator: In the design of the cGAN generator, the most commonly used structure is the U-shape convolutional neural network [50]. The encoder of U-Net can gradually reduce the spatial dimension of feature maps and capture the feature information, while the decoder path can recover spatial dimension and features. Moreover, the skip connections are embedded between the encoder path and the decoder path to integrate downsampling feature with its corresponding upsampling feature. Despite the CNNs-based generator has shown remarkable flexibility and performance in many image processing tasks, there are still some inherit flaws: (1) In the standard CNNs, although a series of scalar values are used to represent the feature information of each neuron which make CNNs very good at detecting features, but the CNNs'

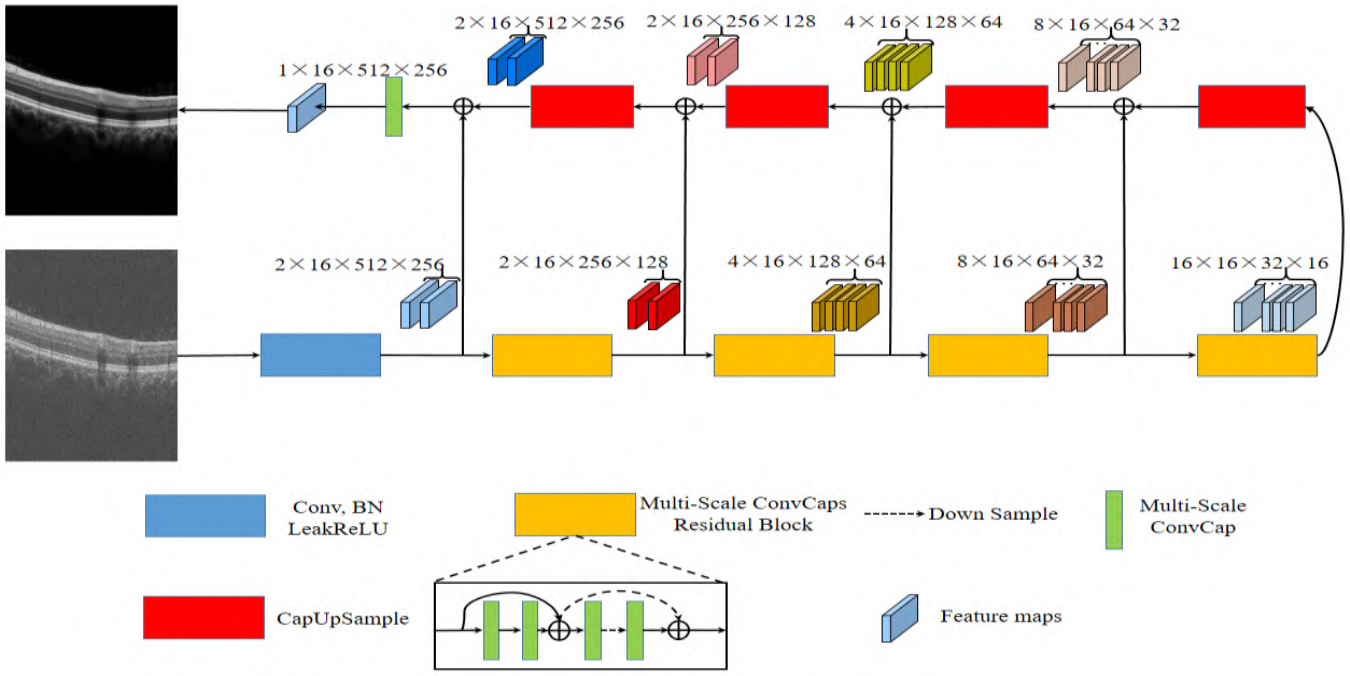

Fig. 2. The structure of the generator in the proposed capsule conditional generative adversarial network.

ability of capturing spatial feature relationship (viewing angle, size, direction) is insufficient.(2) Pooling operation in CNNs can make the network position-invariant, which also prompts CNNs to avoid overfitting. However, this invariance also results in feature loss. (3) Many CNNs-based generators use complex network structures with a large amount of parameters to improve the complex and sparse feature extraction ability. Although these approaches can improve the performance of the network, they also cause the great increase of parameters and increase the risk of overfitting.

To solve these problems, we propose a novel capsule conditional generative adversarial network, referred as Caps-cGAN, in which the generator is designed based on multi-scale dynamic routing convolutional capsule as shown in Fig.2. Different from standard CNNs-based methods, capsules output vectors, whose direction represents attributes (e,g. posture, texture, etc.) and makes the capsule represent the spatial relationship between the features well [41], [42]. Moreover, the other important component of the capsule network is an iterative algorithm called "dynamic routing", in which the output of the capsule is routed to the capsule in the upper layer based on the consistency of the prediction. As shown in Fig.2, the architecture of the proposed generator is still based on the U-shape encoder-decoder structure with skip connections. To further improve the feature representation capacity of the model, we introduce the shortcut connection and multi-scale residual mechanism like ResNet [51] in Caps-cGAN generator. As shown in Fig.2, the original OCT B-scans is first fed into a convolutional layer to get feature maps, then the feature maps are grouped into two capsules. The length of all capsule vectors in this article is set to 16, which is referred to [41] and [42]. The encoder path of the generator consists of four multi-scale ConvCaps residual blocks, where each block contains two ConvCaps residual layers. It has also been demonstrated that multi-scale feature information can improve the performance of feature extraction in [52]. Therefore, we propose a novel multi-scale dynamic convolutional routing to construct our convolutional capsule layer, which is shown as Algorithm 1.

Algorithm 1 Multi-Scale Convolutional Dynamic Routing for Multi-Scale ConvCap

Input: $\varphi \in R^{B,C,H,W}$
$\hat{\varphi} \leftarrow \text{Reshape}(\varphi) \in R^{B\times I,L,H,W}$
$\varphi^l_{k1} \leftarrow \text{Conv2d}(\hat{\varphi}) \in R^{B\times I,\text{O}\times L,H,W} k=1\times 1$
$\varphi^l_{k2} \leftarrow \text{Conv2d}(\hat{\varphi}) \in R^{B\times I,\text{O}\times L,H,W} k=3\times 3, d=1$
$\varphi^l_{k3} \leftarrow \text{Conv2d}(\hat{\varphi}) \in R^{B\times I,\text{O}\times L,H,W} k=3\times 3, d=3$
$\varphi^l_{k4} \leftarrow \text{Conv2d}(\hat{\varphi}) \in R^{B\times I,\text{O}\times L,H,W} k=3\times 3, d=5$
$\varphi^l = \varphi^l_{k1} + \varphi^l_{k2} + \varphi^l_{k3} + \varphi^l_{k4}$
$\hat{\varphi}^l \leftarrow \text{Reshape}(\varphi^l) \in R^{B,I,H,W,\text{O},L}$
$\mathbf{b} \leftarrow \mathbf{0} \in R^{B,I,H,W,O}$
for *iter* to r do:
  for all capsules i in layer l and capsule j in layer $(l+1)$: $c_{i,j} \leftarrow \text{Softmax}(b_{i,j})$
  for all capsules j in layer $(l+1)$: $s_j \leftarrow \sum_i \hat{\varphi}^l_{i,j} \cdot c_{i,j}$
  for all capsules j in layer $(l+1)$: $v_j \leftarrow \text{squash}(s_j)$
  for all capsules i in layer l and capsule j in layer $(l+1)$: $b_{i,j} \leftarrow b_{i,j} + \hat{\varphi}^l_{i,j} \cdot v_j$
end

Let's denote the feature maps from the first convolutional layer as $\varphi \in R^{B,C,H,W}$. In the multi-scale ConvCap, φ is first reshaped to $\hat{\varphi} \in R^{B\times I,L,H,W}$ and fed into a multi-scale 2D convolution module to get new feature map $\varphi^l \in R^{B\times I,\text{O}\times L,H,W}$ with multi-scale information, which is

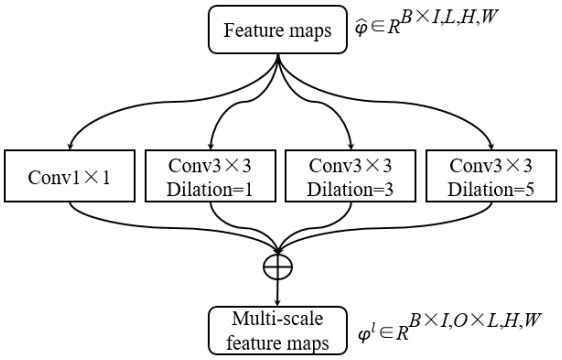

Fig. 3. Multi-scale 2D convolution module. It consists of a conv 1 × 1 branch and three conv 3 × 3 branches, and the dilation rates of conv 3 × 3 branches are 1, 3 and 5. Hence the respective fields of conv 3 × 3 branch are 3, 7 and 9. The conv 1 × 1 branch is adopted to perform feature compression and channel expansion.

shown in Fig.3. Then φ^l is reshaped to $\hat{\varphi}^l \in R^{B,I,H,W,O,L}$. B, I, L, O, H and W represent the batch size, input capsules, length of vectors, output capsules, feature map height and width, respectively. Finally, $\hat{\varphi}^l$ is fed into the iterative vote routing for r iterations to get the output capsules, where r is set as 3. Initialize weight coefficient vector $\mathbf{b} \in R^{B,I,H,W,O}$ as $\mathbf{0}$, which is used to iteratively calculate the weight $\boldsymbol{c}$ between the parent and child capsules. The weight $c_{i,j}$ corresponding to each children capsule and parent capsule is calculated by softmax normalization:

$$c_{i,j} = \frac{\exp\left(b_{i,j}\right)}{\sum_i b_{i,j}} \tag{1}$$

where i, j represents the index for child and parent capsules respectively. Softmax normalizes the weight coefficient $\boldsymbol{b}$ to $\boldsymbol{c}$ to increase the votes of capsules with similar characteristics and reduce the votes of capsules with dissimilar characteristics. $b_{i,j}$ is updated in every routing iteration. The intermediate vector v_j is obtained by squash operation defined as Eq. (2):

$$v_j = \text{squash}(s_j) = \frac{\|s_j\|^2}{1+\|s_j\|^2}\frac{s_j}{\|s_j\|} \tag{2}$$

where s_j denotes the output of capsule as follows:

$$s_j = \sum_i \hat{\varphi}^l_{i,j} \cdot c_{i,j} \tag{3}$$

where "·" denotes dot production.

To effectively fuse the high-resolution and low-semantic features in the lower layers of the network with the high-semantic and low-resolution features in the upper layers, we introduce capsule deconvolution module in the decoder path to upsample the feature maps. The upsampled feature vectors are added with the feature vectors in the lower layer to perform feature fusion. Feature fusion by vector addition can enhance the correlation of similar features and suppress unrelated features. As shown in Fig. 4, taking vector $\boldsymbol{a}$ and $\boldsymbol{b}$ as example, if $\boldsymbol{a}$ and $\boldsymbol{b}$ have similar features, the angle between them will be small, and the features will be enhanced by $\boldsymbol{a}+\boldsymbol{b}$ (Fig. 4(a)). Otherwise the features will be suppressed (Fig. 4(b)). The procedure of de-convolutional dynamic routing is shown in Algorithm 2.

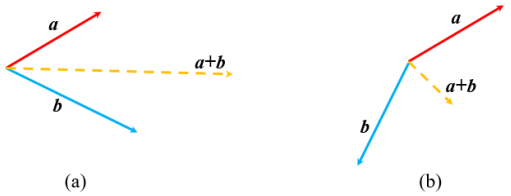

Fig. 4. Feature fusion by vector addition, where $\boldsymbol{a}$ and $\boldsymbol{b}$ indicate the feature vectors. (a) Two vectors with similar features, the angle between $\boldsymbol{a}$ and $\boldsymbol{b}$ will be small, and the features will be enhanced by $\boldsymbol{a}+\boldsymbol{b}$. (b) Two feature vectors with different features, the angle between $\boldsymbol{a}$ and $\boldsymbol{b}$ will be large, and the feature will be suppressed by $\boldsymbol{a}+\boldsymbol{b}$.

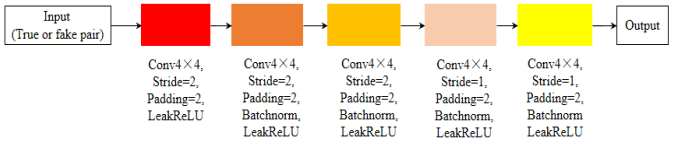

Fig. 5. The architecture of the discriminator.

Discriminator: The discriminator of PatchGAN [53] is adopted as our discriminator, which is shown in Fig.5. The ordinary discriminator in GAN maps the input to a probability value, that is to say, the probability of the input sample is a real sample. In contrast, patchGAN tries to map the input to an N× N matrix X, where X is the output feature map of the convolution layer. And the value of X_{ij} represents the probability whether each patch is a true sample. The mean value of X_{ij} is the final output of the discriminator. In our discriminator, the size of patches can be much smaller than the full size of the image, and its parameters are less than the original discriminator. Therefore, it can be applied to images of any size with higher computational efficiency.

Algorithm 2 De-Convolutional Dynamic Routing

Input: $\varphi \in R^{B,I,L,H,W}$
$\hat{\varphi} \leftarrow \text{Reshape}(\varphi) \in R^{B\times I,L,H,W}$
$\varphi^l \leftarrow \text{Transpose}(\hat{\varphi}) \in R^{B\times I,\text{O}\times L,2H,2W}$
$\hat{\varphi}^l \leftarrow \text{Reshape}(\varphi^l) \in R^{B,I,2H,2W,\text{O},L}$
$\mathbf{b} \leftarrow \mathbf{0} \in R^{B,I,2H,2W,O}$
for *iter* to r do:
  for all capsules i in layer l and capsule j in layer $(l+1)$:
$c_{i,j} \leftarrow \text{Softmax}(b_{i,j})$
  for all capsules j in layer $(l+1)$: $s_j \leftarrow \sum_i \hat{\varphi}^l_{i,j} \cdot c_{i,j}$
  for all capsules j in layer $(l+1)$: $v_j \leftarrow \text{squash}(s_j)$
  for all capsules i in layer l and capsule j in layer $(l+1)$:
$b_{i,j} \leftarrow b_{i,j} + \hat{\varphi}^l_{i,j} \cdot v_j$
end

C. Loss Function

Given an input image $\boldsymbol{X}$, the generator and discriminator are denoted as G and D, respectively. The output of G is represented as $G(\boldsymbol{X})$. In our proposed method, there are four kinds of possible inputs to discriminator network D: input image $\boldsymbol{X}$ concatenating ground truth, input image $\boldsymbol{X}$ concatenating generator prediction $G(\boldsymbol{X})$, ground truth concatenating ground

truth, and generator prediction $G(\boldsymbol{X})$ concatenating generator prediction $G(\boldsymbol{X})$.

Discriminator: The spatial binary cross entropy loss L_D as follows is adopted to optimize the discriminator,

$$L_D = -\sum_{h,w} (1-y) \log \left(1 - D\left(G\left(X\right)\right)^{h,w}\right) + y \log(D(y)^{h,w}) \tag{4}$$

where y = 0 if the sample is from generator prediction, and $y = 1$ if the sample is from the ground truth. $D\left(G\left(X\right)\right)^{h,w}$ denotes the probability map of $G(\boldsymbol{X})$ at location (h, w), and $D(y)^{h,w}$ is the probability map of y at location (h, w).

Generator: A multi-task loss L_G is employed to optimize the generator as follows:

$$L_G = L_{adv} + \lambda L_1 + \beta L_{SSIM} \tag{5}$$

where λ and β are weights for minimizing the proposed multi-task loss function and set as 100 and 10 respectively in this article. L_{adv}, as defined in Eq.(6), is the loss that discriminator recognizes the data from generator.

$$L_{adv} = -\sum_{h,w} \log(D(G(X))^{h,w}) \tag{6}$$

Previous studies have proven that it is beneficial to improving the performance of cGAN by mixing L_1 loss [34], which can represent errors that are sparsely distributed in space.

$$L_1 = \frac{1}{N} \sum_{i=1}^{N} \|y_i - G(X_i)\|_1 \tag{7}$$

where N is the sample size. In addition, the retinal layer is one of the most important structural information in retinal OCT image, which should be reserved as possible during the image denoising. To achieve this purpose, we introduce the structural similarity (SSIM loss) into the loss function. SSIM is an index that measures the similarity between two images from three aspects including brightness, contrast and structure, defined as follows:

$$L_{SSIM} = 1 - \frac{\left(2\mu_{G(X)}\mu_y + \mathrm{c}_1\right)\left(2\sigma_{G(X)}\sigma_y + \mathrm{c}_2\right)}{\left(\mu_{G(X)} + \mu_y + \mathrm{c}_1\right)\left(\sigma^2_{G(X)} + \sigma^2_y + \mathrm{c}_2\right)} \tag{8}$$

where c_1 and c_2 are constants to avoid system errors when the denominator is 0. $\mu_{G(X)}$, μ_y and $\sigma_{G(X)}$, σ_y are the means and standard deviations of $G(X)$ and y, respectively.

Objective Function for Few Labeled Data: In semi-supervised learning, few labeled data is mainly adopted to guide the model to learn the distribution of ground truth and avoid the mode crash may be caused by the distribution diversity of the large amount of unlabeled data. Therefore, the loss function for training model based on labeled data is defined as follows,

$$L_{supervised} = L_G + L_D \tag{9}$$

Objective Function for Unlabeled Data: Since there is no one-to-one ground truth for the unlabeled data, L_1 and L_{SSIM} can not be used to train the model, while L_{adv} is still applicable because it only requires the guidance of discriminator network, that is, $L_{unsupervised} = L_{adv}$.

The semi-supervised loss function is final defined as follows:

$$L_{semi} = L_{supervised} + L_{unsupervised} \tag{10}$$

D. Loss Function Application

As described in section of overview, our proposed semi-supervised learning method is optimized by three important steps: model initialization, mixed training and refinement training.

1) Model Initialization: the supervised loss function $L_{supervised}$ is used to optimize the model based on the labeled data, which aims to obtain the initial weights and initially learn the distribution of the labeled data.

2) Mixed Training: in mixed training process, the model is trained based on mixed data including labeled data and unlabeled data. Therefore, L_{semi} is adopted to optimize the model.

3) Refinement Training: Mode collapse is a common problem in generative adversarial models, especially in semi-supervised learning based tasks which lack label guidance. Therefore, to avoid the mode crash which may be caused by the distribution diversity of the unlabeled data in mixed training and further improve the denosising performance, the model is refinedly trained based on the small amount of labeled data using supervised loss function $L_{supervised}$ again.

E. Method Implementation

The implementation of our proposed denoising method is based on the public platform Pytorch and NVIDIA GeForce RTX 2080Ti GPU with 11GB memory. The Adam solver with momentum 0.5 is applied to optimize our models. Besides, we use the 'poly' learning rate policy, where learning rate $lr = base\_lr * (1 - \frac{iter}{total\_iter})^{power}$, the initial learning rate *base_lr* is set to 2e-4 and the power is set to 0.9.The batch size and total iteration epoch *total_iter* are set as 4 and 100, respectively. To be fair, all the supervised-based methods use the same loss function $L_{supervised}$ in our experiments. The code of Caps-cGAN will be released: https://github.com/wangmeng9218/Caps-cGAN.

IV. Dataset

The study is approved by the Institutional Review Board of Soochow University, and informed consent was obtained from all subjects. Based on the denoising ground truth acquisition method proposed in [36] and [37], we developed the labeled dataset D_L with 2 volumes (512 B-Scans), which were acquired from Topcon Atlantis DRI-1 SS-OCT scanner (Topcon, Tokyo, Japan) with center wavelength of 1050nm(256 B-scans) and the Topcon OCT-2000 SD-OCT scanner (Topcon, Tokyo, Japan) with center wavelength of 850nm (256 B-scans), respectively. The flowchart for obtaining the denoising ground truth of OCT B-scans is shown as Fig.6. First, repeat collecting M 3D OCT volumes from the same normal eye. Then one of the M volumes is randomly selected as the target volume, and its B-scans are also used as the target B-scans. The rest of NM-1 B-Scans surrounding the target B-scans are registered

TABLE I
DETAILS ABOUT THE DATASET

		Scanner	B-scan size (pixels)	B-scans	Center wavelength (nm)	Location	Normal/Pathological
D_L	Training Volume 1	Topcon DRI-1	512×992	256	1050	Macula	Normal
	Training Volume 2	Topcon 2000	512×885	256	840	Macula	Normal
D_U	Testing Volume 1	Topcon DRI-1	512×992	256	1050	Macula	Normal
	Testing Volume 2		512×992	256	1050	Macula+ONH	Normal
	Testing Volume 3		512×992	256	1050	Macula	Pathological (CSC)
	Testing Volume 4	Topcon 1000	512×480	64	840	Macula	Normal
	Testing Volume 5		512×480	64	840	Macula	Pathological (CSC)
	Testing Volume 6	Topcon2000	512×885	128	840	Macula	Normal
	Testing Volume 7		512×885	128	840	ONH	Normal
	Testing Volume 8	Zeiss Cirrus 4000	512×1024	128	840	Macula	Pathological (PM)
	Testing Volume 9		512×1024	128	840	Macula	Pathological (CSC)

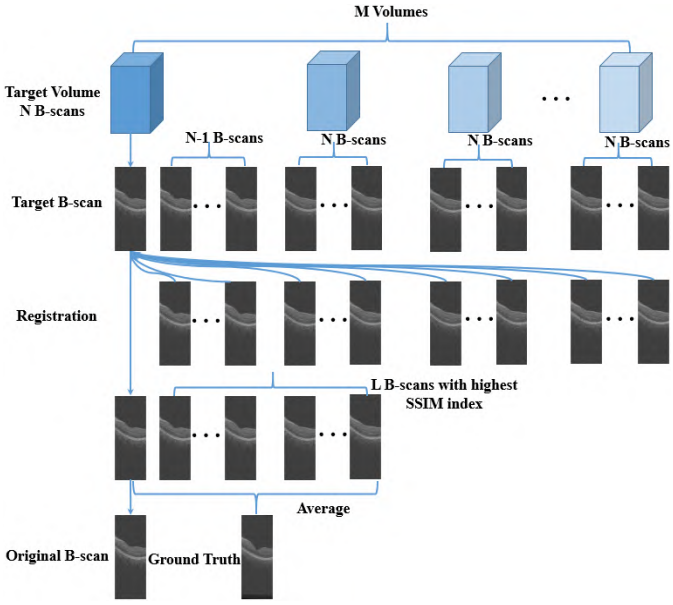

Fig. 6. The flowchart for obtaining the denoising ground truth of OCT B-scans.

with the target B-Scan. Finally, L B-scans with the highest SSIM scores from the NM-1 registered B-scans are selected and averaged with the target B-scan to obtain the ground truth corresponding to the target B-scan.

In additional, 9 retinal OCT volumes consisting of 1408 B-scans from four different types of OCT scanners were collected as the unlabeled dataset D_U, which were also adopted to assess the performance of our proposed method. Details about the data are listed in TABLE I.

V. EXPERIMENT

As shown in TABLE I, we have 512 B-scans with ground truth, which were acquired from two Topcon OCT scanners with different acquisition modes. In addition, we divide 9 test volumes without ground truth into 2 folds to perform cross-validation, one includes 768 B-Scans (Testing Volume 1-3) and the other includes 640B-Scans (Testing Volume 4-9). To prove the effectiveness and generality of our proposed semi-supervised method, we conduct experiments using five data strategies, as listed in TABLE II. The experimental results are qualitatively and quantitatively analyzed. Besides, we also conduct experiments to compare the performance of the proposed method with other state-of-the-art algorithms, in which all supervised CNNs-based methods are trained based on Strategy1. The same data preprocessing method is adopted in all experiments to ensure the fairness. Data augmentation is one of the effective strategies to increase the diversity of data distribution and alleviate the problem of overfitting. In this article, flipping and distortion are applied to augment the data to improve the algorithm performance with a factor of 2.

A. Evaluation Metrics

To evaluate the despeckle performance of different methods, four indicators including signal-to-noise ratio(SNR), contrast-to-noise ratio(CNR), equivalent number of looks(ENL) and edge preservation index(EPI) are adopted to quantitatively analyze the experimental results [20], [56], [57]. The region of interests (ROIs) including one background ROI and three signal ROIs are manually selected to calculate the index in each test image. As shown in Fig.7, the background ROI and signal ROIs are marked with green and red rectangles, respectively. Three signal ROIs are selected which are located in the retinal nerve fiber layer (RNFL), inner retina and the retinal pigment epithelium (RPE) complex respectively, with important structural information of retina. In addition, three boundaries represented by blue curves in Fig.7 including the upper boundary of RNFL, inner-outer retina boundary and the lower boundary of RPE are selected for calculating EPI. These four indicators are calculated as follows:

$$SNR = 10\log_{10}\left(\frac{\max(I)^2}{\sigma_b^2}\right) \tag{11}$$

TABLE II
EXPERIMENTAL DATA STRATEGIES

Strategy	Fold	Data	Data Distribution
Strategy 1		Training Data	512 labeled B-Scans from training volumes 1~2.
		Testing Data	1408 B-Scans from testing volumes 1~9.
Strategy 2	1st fold	Training Data	256 labeled B-Scans only from training volume 1, 256+768 unlabeled B-Scans from training volume 2 and testing volumes 1~3.
		Testing Data	640 B-Scans from testing volumes 4~9.
	2nd fold	Training Data	256 labeled B-Scans only from training volume 1, 256+640 unlabeled B-Scans from training volume 2 and testing volumes 4~9.
		Testing Data	768 B-Scans from testing volumes 1~3.
Strategy 3	1st fold	Training Data	256 labeled B-Scans only from training volume 2, 256+768 unlabeled B-Scans from training volume 1 and testing volumes 1~3.
		Testing Data	640 B-Scans from testing volumes 4~9.
	2nd fold	Training Data	256 labeled B-Scans only from training volume 2, 256+640 unlabeled B-Scans from training volume 2 and testing volumes 4~9.
		Testing Data	768 B-Scans from testing volumes 1~3.
Strategy 4	1st fold	Training Data	512 labeled B-Scans from training volumes 1~2, 768 unlabeled B-Scans from training volume 1 and testing volumes 1~3.
		Testing Data	640 B-Scans from testing volumes 4~9.
	2nd fold	Training Data	512 labeled B-Scans from training volumes 1~2, 640 unlabeled B-Scans from training volume 1 and testing volumes 4~9.
		Testing Data	768 B-Scans from testing volumes 1~3.
Strategy 5		Training Data	256 labeled B-Scans only from training volume 2.
		Testing Data	256 B-Scans with ground truth from training volume 1 and 1408 B-Scans from testing volumes 1~9.

$$CNR_i = 10\log_{10}\left(\frac{|\mu_i - \mu_b|}{\sqrt{\sigma_i^2 + \sigma_b^2}}\right) \tag{12}$$

$$ENL_i = \frac{\mu_i^2}{\sigma_i^2} \tag{13}$$

$$EPI = \frac{\sum_h \sum_w |I_d(h+1, w) - I_d(h, w)|}{\sum_h \sum_w |I_o(h+1, w) - I_o(h, w)|} \tag{14}$$

where μ_b and σ_b represent the mean and standard deviation of background region. μ_i and σ_i are the mean and standard deviation of i-th ($i = 1, 2, 3, \ldots$) signal region. I_o and I_d denote the original image and denoised image. $\max(I)$ is the maximum intensity value of the B-scan I. h and w are coordinates in height and width direction of image, respectively.

B. Qualitative Evaluation

Fig.7 shows the denoising results of B-scans corresponding to 9 testing volumes listed in Table I. These results are obtained by the proposed method based on strategy 4. It can be seen from Fig.7 that the proposed semi-supervised method performs well for all test volumes, the speckle noise in different areas is eliminated while the retinal layer structures and choroidal vessels are preserved and enhanced well. It can also be seen that although the training data is collected in the macula-centered mode without any lesions, the proposed method still works well in other collection modes (Fig.7(b) and Fig.7(d)) and the abnormal OCT B-scans (Fig.7(a), Fig.7(f), Fig.7(h) and Fig.7(i)). These results demonstrate the effectiveness and generality of our proposed method.

To further evaluate the performance of our proposed method, one example of denoising results with different methods are shown in Fig.8, in which the proposed method is based on strategy 4. It can be seen that BM3D, K-SVD and NLM do not remove noise completely, which cause artifacts inside the retinal layers and result in adhesion between layers. The layer edge of the denoising result with the MAP method is not smooth enough (Fig.8(e)), and there are still artifacts between the layers (Fig.8(e)). Although STROLLR-2D [54] can remove speckle noise well, the insufficient enhancement of the retina's layer structure results in the blurring of boundary between layers. DnCNN performs poorly on both test data (Fig.8(g)), which does not suppress speckle noise well and also causes the blurring of the retinal layer structures. The edges of retinal layers are distorted in the results of ResNet and cGAN and the external limiting membrane (ELM) is not enhanced well in the results of cGAN (Fig.8(i)). Compared with these methods, our proposed method removes speckle noise well and enhances the retinal layer information with clear layer boundaries, which proves the effectiveness of the proposed method (Fig.8(j)).

C. Quantitative Evaluation

To quantitatively evaluate the despeckling performance, four metrics including SNR, CNR, ENL and EPI of different methods listed in TABLE III.

The performances of some typical traditional methods are shown in upper part of TABLE III. The SNR of BM3D is low, which may be caused by its poor performance in speckle noise suppression. K-SVD has good indicators except the EPI, which may be caused by the blurred edges. Instead, NLM has the highest EPI and the lowest CNR and ENL, which

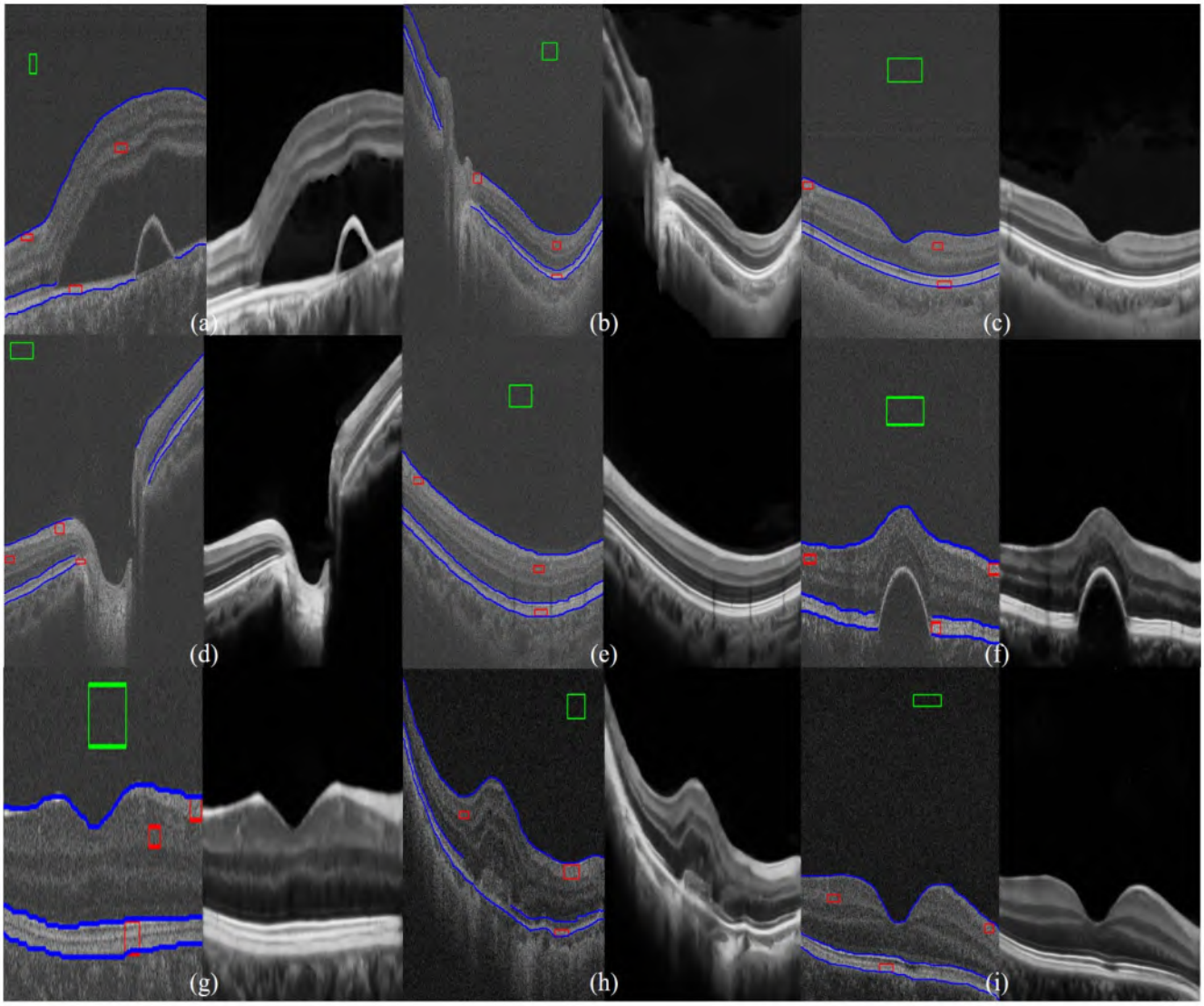

Fig. 7. The denoising results of B-scans from 9 testing volumes. For each panel, left: original B-scan, right: denoised B-scan. The green rectangle, red rectangles and blue curves represent the background ROI, signal ROIs and boundaries for calculating EPI, respectively.

may be caused by artifacts near the edges. All indicators of MAP except ENL are the low, especially SNR, which has great correlation with background noise. Similar to K-SVD, STROLLR has high CNR, SNR and ENL and low EPI. The middle part of TABLE III lists the quantitative performances of some state-of-art deep learning-based methods including DnCNN, ResNet, cGAN and Cycle-GAN. Compared with other deep learning-based methods, ResNet obtains the lowest indicators except for EPI. On the contrary, DnCNN obtains the lowest EPI. Due to the weakly supervised learning strategy, the indicators of Cycle-GAN [55] (except CNR) are lower than those of cGAN, especially EPI, which is mainly caused by the lack of one-to-one label guidance and in turn leads to the loss of retinal structural information. cGAN obtains quite balanced indices, which are still lower than those of the proposed semi-supervised Caps-cGAN. With the same training data, our proposed semi-supervised Caps-cGAN achieves the best SNR, CNR and ENL compared with other CNNs-based methods.

TABLE III also shows the indicators of the proposed supervised Caps-cGAN, which is trained using the same Strategy (Strategy 1 in TABLE II) and the same supervised loss function (Eq.(9)) with cGAN. It can be seen from TABLE III that compared with cGAN, the SNR, CNR, ENL and EPI of the proposed supervised Caps-cGAN have increased by 21.21%, 14.45%, 158.62% and 3.06%, respectively. These results show that the proposed Caps-cGAN can achieve better denoising performance than cGAN with standard CNNs. Due to its generator with 3 residual blocks, the parameter number of cGAN is 276.69M, while the proposed Caps-cGAN only has 5.31M. That is to say, the proposed Caps-cGAN can obtain better performances than cGAN with fewer parameters.

In addition, we also compare the denoising efficiency of different methods. Since the traditional denoising algorithms are executed on CPU, it results in the low efficiency. On the contrary, deep learning-based methods can be accelerated by GPU accelerator, which greatly improves the efficiency. It can be seen from TABLE III that our proposed method takes

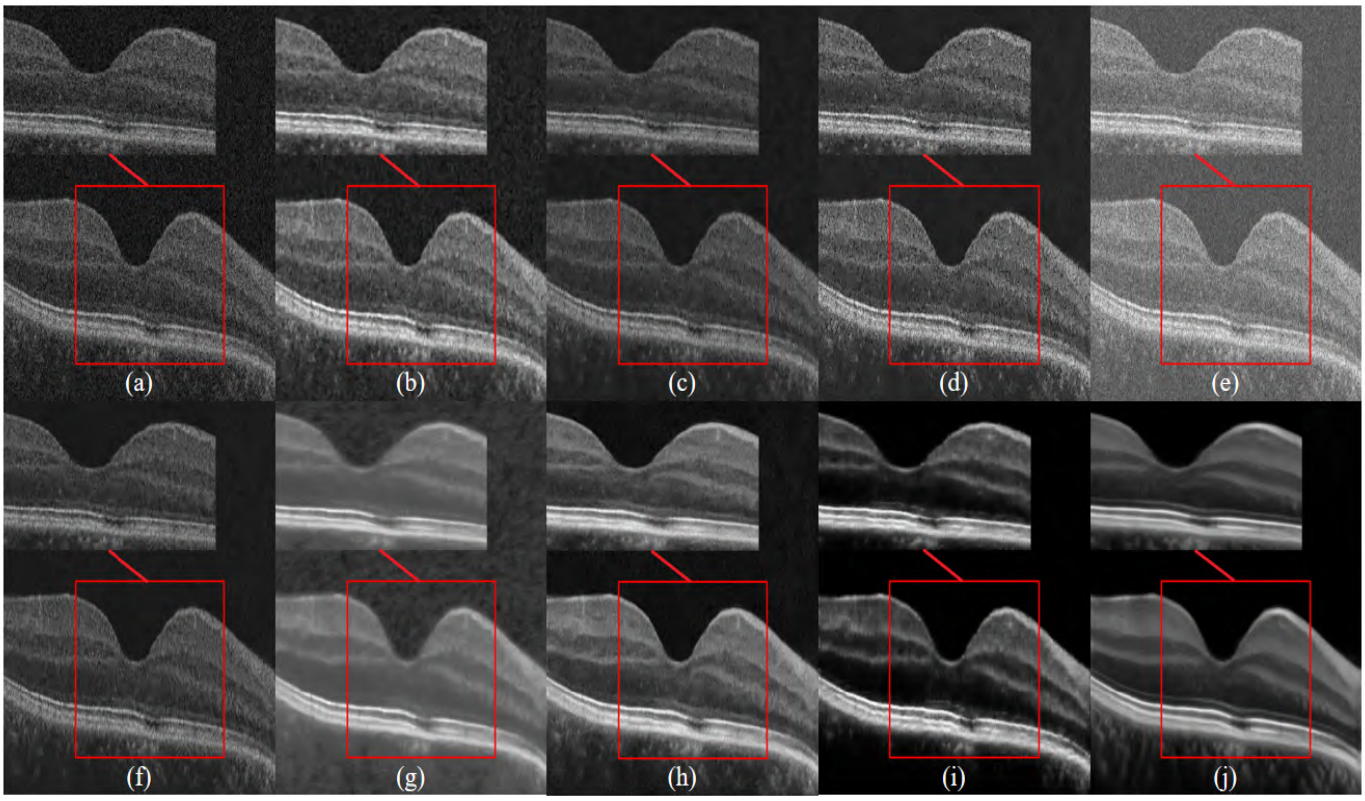

Fig. 8. One example of denoising results with different methods. (a) Original image (b) BM3D (c) K-SVD (d) NLM (e) MAP (f) STROLLR-2D (g) DnCNN (h) ResNet (i) cGAN (j) Proposed.

TABLE III
QUALITATIVE EVALUATION RESULTS OF DIFFERENT METHODS

Strategy	Learning method	Method	SNR	CNR	ENL	EPI	Times(s)
Strategy 1	Supervised	BM3D [17]	34.80±1.76	8.36±0.80	111.15±46.15	0.81±0.10	176.393
		K-SVD [20]	50.07±2.12	9.24±1.87	260.58±326.15	0.79±0.13	136.728
		NLM [16]	44.56±2.88	6.11±1.44	63.56±43.67	**1.04±0.09**	0.155
		MAP [18]	31.73±0.73	7.33±1.28	128.44±54.76	0.75±0.09	0.361
		STROLLR [54]	41.86±2.09	8.18±1.41	123.91±98.29	0.80±0.09	196.342
		DnCNN [34]	42.38±6.72	9.26±3.26	153.17±57.35	0.83±0.13	0.0710
		ResNet [35]	35.76±3.57	9.01±1.39	141.44±79.61	0.97±0.11	0.0919
		Cycle-GAN[55]	46.56±2.59	9.97±0.80	138.55±55.41	0.91±0.15	**0.0302**
		cGAN [51]	47.39±3.61	9.69±1.04	139.43±63.22	0.98±0.10	0.0464
		Caps-cGAN	57.44±9.94	11.09±1.20	360.59±317.40	1.01±0.18	0.0951
Strategy 2	Semi-Supervised	Caps-cGAN	56.02±9.55	10.72±1.05	264.69±179.33	1.03±0.19	
Strategy 3			56.21±7.47	11.07±1.06	304.46±200.76	1.00±0.17	
Strategy 4			**59.01±9.41**	**11.37±1.21**	**417.22±350.28**	1.03±0.17	

slightly longer time than other deep learning-based methods due to the introduction of vector and matrix operations in the capsule network. However, it can still meet the requirement of real-time processing. In summary, except that the EPI index is comparative to that of NLM, other indexes of the proposed Caps-cGAN have been greatly improved, which show the effectiveness and generality of the proposed Caps-cGAN.

In order to prove the advantage of the proposed semi-supervised learning strategy, we independently train the model using Strategy 2 and Strategy 3 with 256 labeled data and 256+768/640 unlabeled data. It can be seen from TABLE III that the indicators of the both semi-supervision trained models are higher than the other state-of-the-art approaches which are trained based on 512 labeled data. The result shows that the proposed semi-supervised Caps-cGAN can leverage less data to obtain comparable denoising performance.

It also can be seen from TABLE III that our proposed semi-supervised Caps-cGAN has achieved better performance than the supervision-based Caps-cGAN. The SNR, CNR, ENL and EPI of our proposed semi-supervised Caps-cGAN have been improved to 59.01, 11.37, 417.22 and 1.03 from 57.44, 11.09, 360.59 and 1.01, respectively. These results show that the unlabeled data is beneficial to improve the denoising performance and further demonstrate the effectiveness of the proposed semi-supervised method.

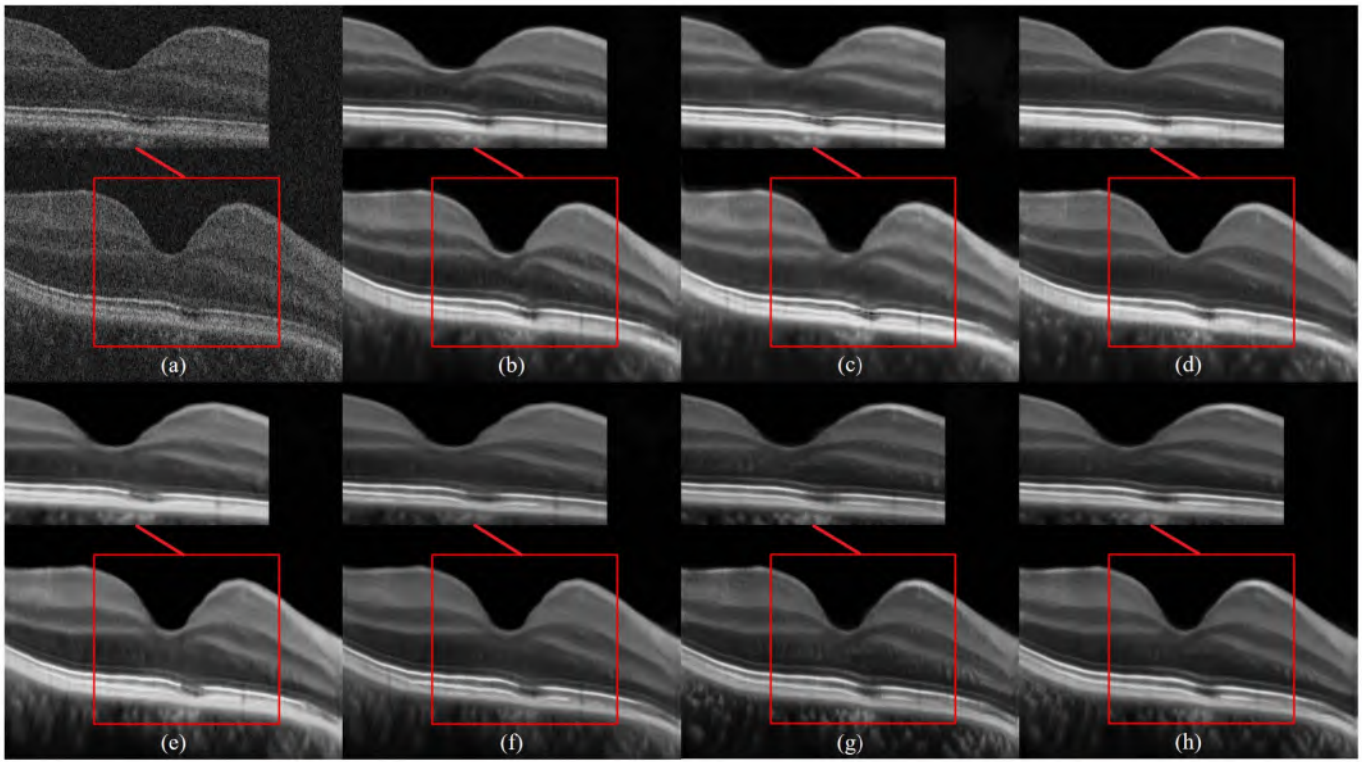

Fig. 9. Results of ablation experiments. (a) Original B-scan image (b) Caps-cGAN with original cGAN loss (c) Caps-cGAN+L1 (d) Caps-cGAN+SSIM (e) Single-Scale ConvCap (f) Multi-Scale ConvCap without residual (g) Proposed without refinement (h) Proposed.

TABLE IV
QUALITATIVE EVALUATION RESULTS OF DIFFERENT LOSS FUNCTIONS

Strategy	Methods	SNR	CNR	ENL	EPI
Strategy 4	Caps-cGAN	54.50±5.38	10.77±1.17	227.88±127.03	1.00±0.18
	Caps-cGAN+L1	51.18±5.23	11.16±0.83	293.01±124.04	0.97±0.19
	Caps-cGAN+SSIM	56.89±8.85	10.93±1.22	264.60±162.43	1.02±0.18
	Proposed	**59.01±9.41**	**11.37±1.21**	**417.22±350.28**	**1.03±0.17**

D. Ablation Experiment

To evaluate the contribution of the loss functions adopted in the proposed semi-supervised Caps-cGAN, four ablation experiments about original cGAN loss, original + L1 loss, original + SSIM loss and the proposed original + L1 +SSIM loss are performed with the same training Strategy 4. Fig.9(a)-(d) and (h) show the original B-scan and the corresponding denoising results with different loss functions and TABLE IV shows the corresponding quantitative evaluation results. It can be seen from TABLE IV that compared with the Caps-cGAN based on the original cGAN loss, the introduction of L1 loss can improve CNR and ENL and slightly reduce SNR and EPI, which may because that L1 loss improves the smoothness of the uniform area in the retinal layer and causes a slight blur simultaneously (as shown in the red rectangle area in Fig.9(b) and Fig.9(c)). On the contrary, all indicators of Caps-cGAN+SSIM (as shown in Fig.9(d)) have been improved because SSIM loss can introduce structural information of the retina during model training. Compared with the above two results, the result of Caps-cGAN trained using the proposed loss function (Eq. 10) not only improves the objective indexes (shown in TABLE IV), but also keeps the retinal layer structure clearer and smoother (as shown in the red rectangle area in Fig.9(h)).

The effectiveness of the proposed network architecture is also verified. Fig.9(e)-(g) show the denoising results with different network architectures and the corresponding quantitative evaluation results are shown in TABLE V. As can be seen from Fig.9(e) and Fig.9(f), the areas between the inner and outer boundaries of the retina and RPE suffers from blurred layer structure, but the proposed method obtains a clearer and smoother retinal layer structure in this region (Fig.9(h)). As shown in TABLE V, compared the Caps-cGAN based on single-scale ConvCaps that proposed in [42], the SNR, CNR, ENL and EPI of the proposed method have increased by 19.82%, 8.18%, 83.89% and 5.10%, respectively. Compared with the multi-scale Caps-cGAN without residual structure, the SNR, CNR, ENL and EPI of the proposed method have increased by 18.90%, 9.22%, 62.04% and 13.19%, respectively. It can be seen from Fig.9(g) and (h) and TABLE V that compared with the result of Caps-cGAN without refinement training, the proposed method gets slightly smoother

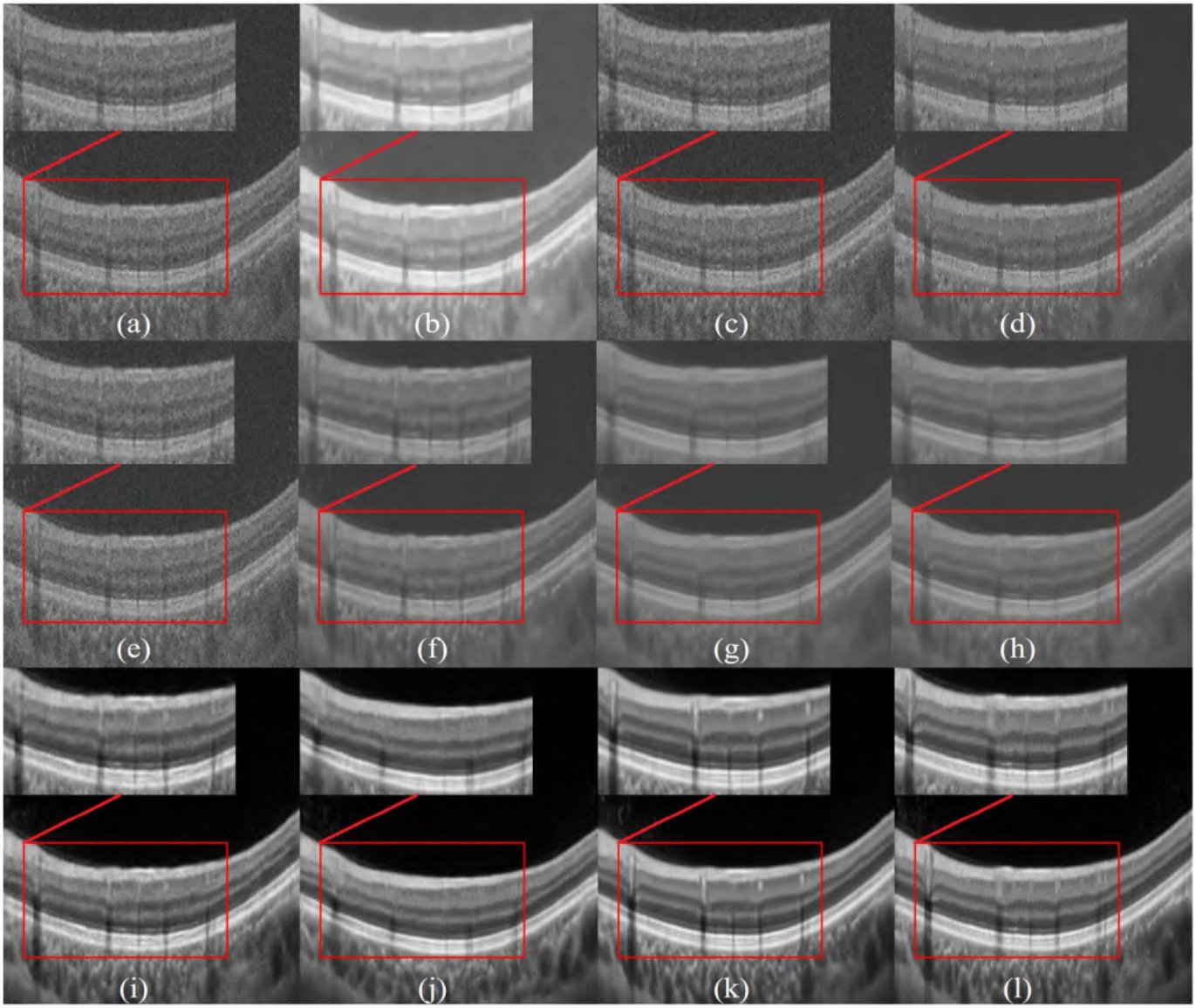

Fig. 10. Expanded experimental results from different methods. (a) Original image (b) BM3D (c) K-SVD (d) NLM (e) MAP (f) STROLLR-2D (g) DnCNN (h) ResNet (i) Cycle-GAN (j) cGAN (k) Caps-cGAN (l) Ground truth.

layer structure and better continuity, and the corresponding SNR, CNR, ENL and EPI have increased by 1.2%, 0.4%, 16.55% and 4.04%, respectively. These results demonstrate the rationality and effectiveness of the proposed network structure design.

E. Extended Experiment

Based on Strategy 5 listed in TABLE II, an extended experiment is performed to further demonstrate the effectiveness and generality of the proposed Caps-cGAN. The multi-scale-structural similarity index (Ms-SSIM) indicator is introduced to analyze the structural similarity between the result and the ground truth, which is the multi-scale score of SSIM.

$$Ms-SSIM=[L_M(X,Y)]^{\alpha_M} \times \prod_{J=1}^{M}[C_J(X,Y)]^{\beta_J}[S_J(X,Y)]^{\gamma_J} \quad (15)$$

$$L(X,Y)=\frac{2\mu_X\mu_Y+C_1}{\mu_X^2+\mu_Y^2+C_1} \quad (16)$$

$$C(X,Y)=\frac{2\sigma_X\sigma_Y+C_2}{\sigma_X^2+\sigma_Y^2+C_2} \quad (17)$$

$$S(X,Y)=\frac{\sigma_{XY}+C_3}{\sigma_X\sigma_Y+C_3} \quad (18)$$

where X and Y represent the denoised image and ground truth, respectively. $L(X,Y)$ is the brightness contrast factor, $C(X,Y)$ is the contrast factor, and $S(X,Y)$ is the structural contrast factor. μ_X, μ_Y and σ_X, σ_Y denote the mean and standard deviations of X and Y. The constants C_1, C_2 and C_3 are small values for numerical stability. M represents the number of scales. The ratio J indicates that the original image is down-sampled by a factor of 2^{J-1}. α_M, β_J and γ_J are used to adjust the relative importances of the components. In this article, referring to [58], M = 5, $\alpha_1 = \beta_1 = \gamma_1 = 0.0448$,

TABLE V
QUALITATIVE EVALUATION RESULTS OF DIFFERENT NETWORK ARCHITECTURES

Strategy	Methods	SNR	CNR	ENL	EPI
Strategy 4	Caps-cGAN (Single-Scale ConvCap[42])	49.25±2.48	10.51±1.15	226.89±157.08	0.98±0.17
	Caps-cGAN （Multi-Scale ConvCap Without Residual）	49.63±2.80	10.41±1.08	257.48±179.89	0.91±0.18
	Proposed without Refinement Training	58.30±9.92	11.32±1.06	357.98±279.99	0.99±0.17
	Proposed	**59.01±9.41**	**11.37±1.21**	**417.22±350.28**	**1.03±0.17**

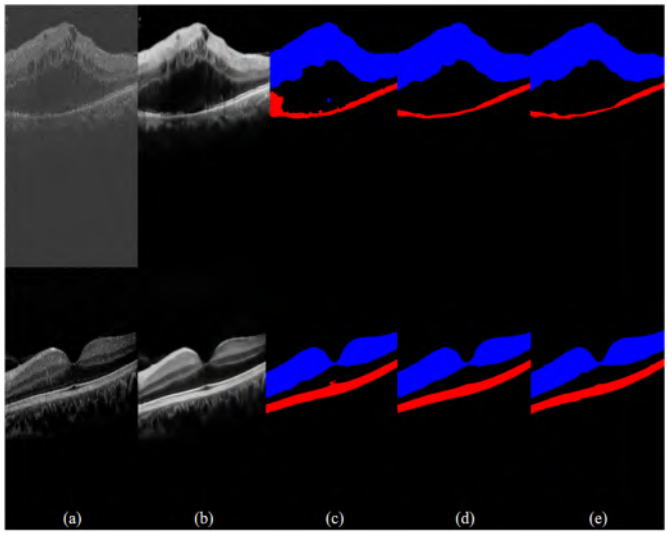

Fig. 11. Layer segmentation results. (a) Original image (b) Denoised image (c) The segmentation result of original image (d) The segmentation result of denoised image (e) Ground truth. Blue and red regions represent inner and outer retinal layer respectively.

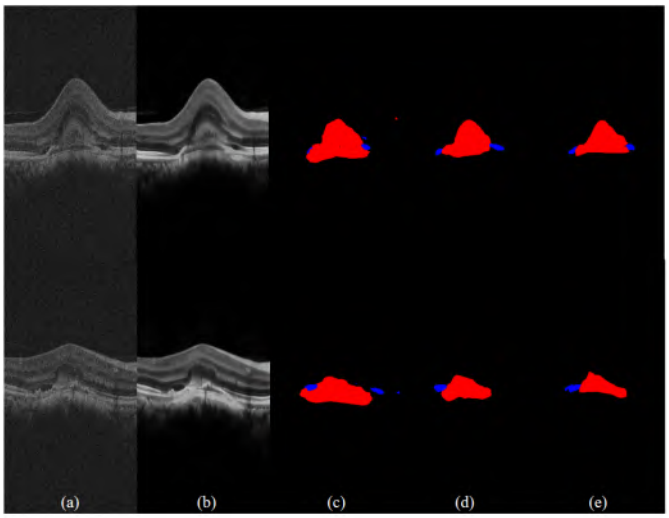

Fig. 12. SRF and CNV joint segmentation results. (a) Original image (b) Denoised image (c) The segmentation result of original image (d) The segmentation result of denoised image (e) Ground truth. Red and blue regions represent CNV and SRF respectively.

$\alpha_2 = \beta_2 = \gamma_2 = 0.2856$, $\alpha_3 = \beta_3 = \gamma_3 = 0.3001$, $\alpha_4 = \beta_4 = \gamma_4 = 0.2363$,$\alpha_5 = \beta_5 = \gamma_5 = 0.1333$. Fig.10 shows one example of denoising results of different methods. The quantitative indicators are listed in TABLE VI, where SNR, CNR, ENL and EPI are calculated based on 9 test sets without ground truth and training volume 1 with ground truth, while Ms-SSIM is calculated based on training volume1 as it requires ground truth.

It can be seen from Fig.10 and TABLE VI that compared with traditional typical denoising methods, such as BM3D, K-SVD, NLM, MAP, STROLLR, our proposed Caps-cGAN can suppress speckle noise more obviously, and the retina layer structure information has also been well enhanced. Although DnCNN has obtained the highest ENL index, the view of its denoised image (Fig.10(g)) is blurry and the layer information is not clear, which leads to lower EPI. The SNR, CNR and ENL of DnCNN associated with the speckle noise are improved because Strategy 5 adds training set 1 (from Topcon) into the test data, whose feature distribution is similar to the training set 2. However, its EPI and Ms-SSIM related to the retinal structure information are low, which may be caused by its poor retinal structure enhancement (shown as Fig.10(g)). Compared with DnCNN and ResNet, Cycle-GAN and cGAN can not only remove the speckle noise well, but also enhance the contrast and detail information of retina. As shown in Fig.10 and TABLE VI, compared with other denoising methods, our proposed Caps-cGAN has obtained better view of denoised image and higher SNR, CNR, EPI and Ms-SSIM indicators. It can be seen from Fig.10 that the retina structure information of denoised image with our proposed Caps-cGAN has been enhanced, especially for retina layer structure and choroidal vessels. The proposed Caps-cGAN has obtained the highest Ms-SSIM compared with other methods. These results further demonstrate the effectiveness and generality of our proposed Caps-cGAN in the task of removing speckle noise in retinal OCT images.

F. Application in Retinal Image Segmentation

To verify that the proposed denoising algorithm can facilitate image analysis, two experiments including retinal layer segmentation and joint segmentation of choroidal neovascularization (CNV) and sub-retinal fluid (SRF) are conducted. The corresponding segmentation results and Dice coefficient (DSC) are shown in Fig.11, Fig.12 and TABLE VII, respectively. The commonly used medical image segmentation network U-Net is adopted as the segmentation network. The training strategies and platform settings are consistent in all comparison experiments.

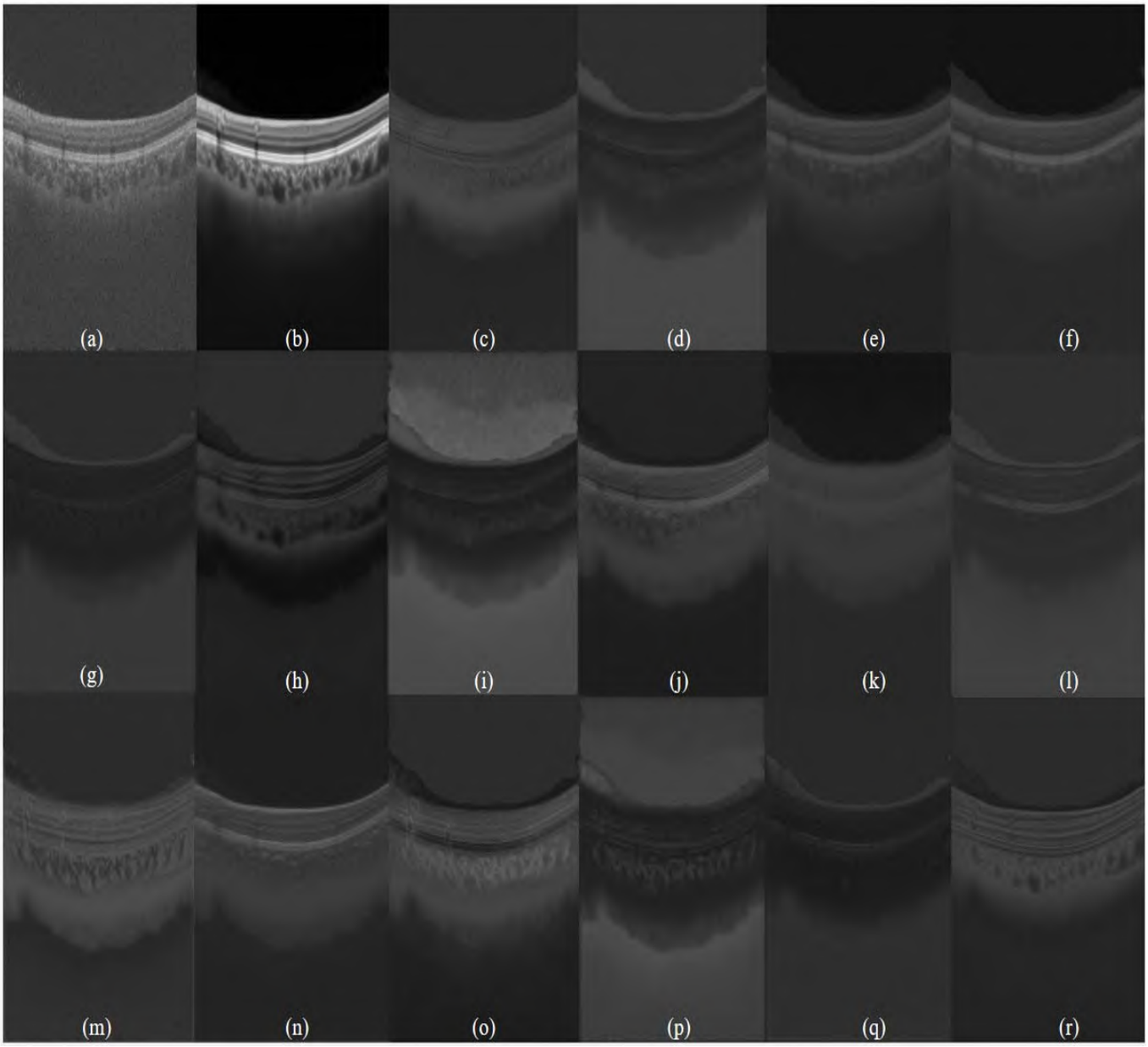

Fig. 13. Vector reconstruction results. (a) Original image (b) Denoised image (c-r) The reconstruction results of each dimension in vector.

1) Retinal Layer Segmentation: 2054 retinal OCT B-scan images were acquired from different OCT scanners and acquisition modes. The inner retinal layer and outer retinal layer were manually labeled as the ground truth under the supervision of the senior ophthalmologist. The data were randomly split into training set and test set according to the ratio of 4:1. As can be seen from Fig. 11 and TABLE VII, the layer structure information of the denoised image is clearer than the original image and the segmentation accuracy of the denoised image is higher than that of the original image, which proves the effectiveness of the proposed denoising method and is conducive to improve the layer segmentation performance.

2) Joint Segmentation of CNV and SRF: 1024 OCT B-cans from 6 eyes with CNV and SRF were included in this experiment. The CNV and SRF ground truth were manually labeled by two ophthalmologists independently. As shown in Fig.12 and TABLE VII, the segmentation performances with denoised data are better than those with original data, which indicates that the proposed denoising method is beneficial for disease region segmentation.

G. Vector Feature Analysis

To further explain why the proposed method works well in the OCT image denoising task, the feature responses of each dimension vector in the last capsule are reconstructed and shown in Fig.13, where Fig. 13(a) and Fig. 13(b) are the original noisy image and the denoised B-Scans, respectively. Fig. 13(c-r) denote the reconstructed results of each vector feature in the last capsule, which shows that the vector value of each dimension in the capsule focuses on different features in the OCT image. Fig. 13(c), Fig. 13(e), Fig. 13(f) and Fig. 13(j)

TABLE VI
THE EXPANDED EXPERIMENTS RESULTS OF DIFFERENT METHODS

Strategy	Methods	SNR	CNR	ENL	EPI	Ms-SSIM
Strategy 5	BM3D [17]	35.82±3.27	8.48±0.61	151.48±51.66	0.81±0.08	0.85±0.002
	K-SVD [20]	39.09±2.94	7.72±1.24	113.95±85.56	0.80±0.10	0.85±0.003
	NLM [16]	44.93±2.96	6.24±1.42	70.70±46.65	1.03±0.09	0.87±0.003
	MAP [18]	31.77±1.26	7.27±2.01	126.48±46.67	0.77±0.11	0.85±0.003
	STROLLR [54]	42.16±2.18	8.29±1.38	153.07±127.86	0.78±0.11	0.87±0.003
	DnCNN [34]	45.24±7.34	10.22±1.24	**332.86±185.67**	0.79±0.17	0.87±0.003
	ResNet [35]	36.40±4.66	8.92±1.39	166.04±120.69	0.94±0.15	0.87±0.003
	Cycle-GAN[55]	47.92±2.26	9.80±0.81	123.35±42.25	0.96±0.19	0.85±0.006
	cGAN [51]	46.90±4.75	10.52±0.95	174.40±66.89	1.01±0.16	0.91±0.011
	Caps-cGAN	**56.79±8.37**	**10.68±1.26**	233.07±141.34	**1.06±0.21**	**0.98±0.001**

TABLE VII
THE DICE COEFFICIENT OF SEGMENTATION (%)

Tasks	Target	Original	Denoised
Layer Segmentation	Inner retinal layer	96.75	**97.02**
	Outer retinal layer	93.35	**94.22**
	Mean	95.05	**95.62**
Joint Segmentation of CNV and SRF	SRF	76.78	**78.16**
	CNV	69.37	**70.00**
	Mean	73.08	**74.08**

characterize different main foreground features in OCT image, such as retinal layer edge, contrast and smoothness, etc. In contrast, Fig. 13(d), Fig. 13(g), Fig. 13(i), Fig. 13(p) and Fig. 13(q) learn diverse background features, especially in the Fig. 13(i) and Fig. 13(p), the background noise near the foreground edge of the retina is well represented. These results also explain why our results have high SNR, CNR and ENL. In addition, Fig. 13(h), Fig. 13(j), Fig. 13(m), Fig. 13(n), Fig. 13(o) and Fig. 13(r) capture different structural features of retina respectively, such as layer structure and choroidal vessels, etc. In Fig. 13(h), Fig. 13(m) and Fig. 13(r), the layer structure properties such as contrast, thickness, layer spacing and smoothness can be clearly observed, which result in the high EPI.

VI. CONCLUSION AND DISCUSSION

In this article, we propose a novel semi-supervision based method for speckle noise reduction in retinal OCT images. It is the first time to introduce the capsule network into the task of retinal OCT image denoising and achieve outperforming results. Unlike the previous CNNs-based methods, which improve the denoising performance via complex network structure and numerous parameters, our newly proposed Caps-cGAN can learn the feature information of the retinal OCT images via very few parameters. In addition, our proposed semi-supervision based network can get better performances with fewer training data than the fully supervision based networks. Comprehensive experiments are also conducted to evaluate the effectiveness and generality of the proposed method, which show that compared with other state-of-art algorithms, our proposed semi-supervised method obtains the best visual quality and higher objective indexes. The proposed semi-supervised method is suitable for retinal OCT images collected from different types of OCT devices and different scanning modes well.

There is still a limitation in this study that the model was trained only using the data from normal eyes, because the registration and average method for the ground truth acquisition is not applicable for the image with lesions. Although the proposed method has achieved promising generality on pathological data, we believe that if some data with lesions can be added into the training set, the performance of the proposed method will be further improved. Therefore, it is one of our future research tasks to explore the ground truth acquisition method that can be applied to pathological data. Besides, how to further improve the efficiency of matrix operations in the capsule network is another focus that will be continuously explored in future work.

REFERENCES

[1] D. Huang *et al.*, "Optical coherence tomography," *Science*, vol. 254, no. 5035, pp. 1178–1181, 1991.
[2] S. Zhu, F. Shi, D. Xiang, W. Zhu, H. Chen, and X. Chen, "Choroid neovascularization growth prediction with treatment based on reaction-diffusion model in 3-D OCT images," *IEEE J. Biomed. Health Informat.*, vol. 21, no. 6, pp. 1667–1674, Nov. 2017.
[3] W. Zhu *et al.*, "Automatic three-dimensional detection of photoreceptor ellipsoid zone disruption caused by trauma in the OCT," *Sci. Rep.*, vol. 6, no. 1, p. 25433, May 2016.
[4] Y. Rong *et al.*, "Deriving external forces via convolutional neural networks for biomedical image segmentation," *Biomed. Opt. Exp.*, vol. 10, no. 8, pp. 3800–3814, 2019.
[5] W. Zhu *et al.*, "An automated framework for intra-retinal cystoid macular edema segmentation in 3D-OCT images with macular hole," *J. Biomed. Opt.*, vol. 22, no. 7, 2017, Art. no. 076014.
[6] K. Yu, F. Shi, E. Gao, W. Zhu, H. Chen, and X. Chen, "Shared-hole graph search with adaptive constraints for 3D optic nerve head optical coherence tomography image segmentation," *Biomed. Opt. Exp.*, vol. 9, no. 3, pp. 962–983, 2018.
[7] D. Xiang *et al.*, "Automatic segmentation of retinal layer in OCT images with choroidal neovascularization," *IEEE Trans. Image Process.*, vol. 27, no. 12, pp. 5880–5891, Dec. 2018.
[8] D. Xiang *et al.*, "Automatic retinal layer segmentation of OCT images with central serous retinopathy," *IEEE J. Biomed. Health Informat.*, vol. 23, no. 1, pp. 283–295, Jan. 2019.
[9] L. Pan, L. Guan, and X. Chen, "Segmentation guided registration for 3D spectral-domain optical coherence tomography images," *IEEE Access*, vol. 7, pp. 138833–138845, 2019.
[10] N. Iftimia, B. E. Bouma, and G. J. Tearney, "Speckle reduction in optical coherence tomography by 'path length encoded' angular compounding," *J. Biomed. Opt.*, vol. 8, no. 2, p. 260, 2003.
[11] B. F. Kennedy *et al.*, "Speckle reduction in optical coherence tomography by strain compounding," *Opt. Lett.*, vol. 35, no. 14, pp. 2445–2447, 2010.

[12] W. Cheng, J. Qian, Z. Cao, X. Chen, and J. Mo, "Dual-beam angular compounding for speckle reduction in optical coherence tomography," *Proc. SPIE*, vol. 10053, Feb. 2017, Art. no. 100532Z.

[13] H. M. Salinas and D. C. Fernandez, "Comparison of PDE-based nonlinear diffusion approaches for image enhancement and denoising in optical coherence tomography," *IEEE Trans. Med. Imag.*, vol. 26, no. 6, pp. 761–771, Jun. 2007.

[14] P. Puvanathasan and K. Bizheva, "Interval type-II fuzzy anisotropic diffusion algorithm for speckle noise reduction in optical coherence tomography images," *Opt. Exp.*, vol. 17, no. 2, pp. 733–746, Jan. 2009.

[15] J. Aum, J. Kim, and J. Jeong, "Effective speckle noise suppression in optical coherence tomography images using nonlocal means denoising filter with double Gaussian anisotropic kernels," *Appl. Opt.*, vol. 54, no. 13, pp. 13–14, 2015.

[16] A. Buades, B. Coll, and J.-M. Morel, "A non-local algorithm for image denoising," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2005, pp. 60–65.

[17] B. Chong and Y.-K. Zhu, "Speckle reduction in optical coherence tomography images of human finger skin by wavelet modified BM3D filter," *Opt. Commun.*, vol. 291, pp. 461–469, Mar. 2013.

[18] M. Li, R. Idoughi, B. Choudhury, and W. Heidrich, "Statistical model for OCT image denoising," *Biomed. Opt. Exp.*, vol. 8, no. 9, pp. 3903–3917, 2017.

[19] F. Zaki, Y. Wang, H. Su, X. Yuan, and X. Liu, "Noise adaptive wavelet thresholding for speckle noise removal in optical coherence tomography," *Biomed. Opt. Exp.*, vol. 8, no. 5, pp. 2720–2731, 2017.

[20] R. Kafieh, H. Rabbani, and I. Selesnick, "Three dimensional data-driven multi scale atomic representation of optical coherence tomography," *IEEE Trans. Med. Imag.*, vol. 34, no. 5, pp. 1042–1062, May 2015.

[21] Z. Jian, L. Yu, B. Rao, B. J. Tromberg, and Z. Chen, "Three-dimensional speckle suppression in optical coherence tomography based on the curvelet transform," *Opt. Exp.*, vol. 18, no. 2, pp. 1024–1032, Jan. 2010.

[22] L. Fang, S. Li, Q. Nie, J. A. Izatt, C. A. Toth, and S. Farsiu, "Sparsity based denoising of spectral domain optical coherence tomography images," *Biomed. Opt. Exp.*, vol. 3, no. 5, pp. 927–942, May 2012.

[23] L. Fang, S. Li, D. Cunefare, and S. Farsiu, "Segmentation based sparse reconstruction of optical coherence tomography images," *IEEE Trans. Med. Imag.*, vol. 36, no. 2, pp. 407–421, Feb. 2017.

[24] I. Kopriva, F. Shi, and X. Chen, "Enhanced low-rank + sparsity decomposition for speckle reduction in optical coherence tomography," *J. Biomed. Opt.*, vol. 21, no. 7, Jul. 2016, Art. no. 076008.

[25] J. Cheng *et al.*, "Speckle reduction in 3D optical coherence tomography of retina by A-Scan reconstruction," *IEEE Trans. Med. Imag.*, vol. 35, no. 10, pp. 2270–2279, Oct. 2016.

[26] X. Guo and Y. Yuan, "Triple ANet: Adaptive abnormal-aware attention network for WCE image classification," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* Cham, Switzerland: Springer, 2019, pp. 293–301.

[27] Q. Wang, Y. Zheng, G. Yang, W. Jin, X. Chen, and Y. Yin, "Multiscale rotation-invariant convolutional neural networks for lung texture classification," *IEEE J. Biomed. Health Informat.*, vol. 22, no. 1, pp. 184–195, Jan. 2018.

[28] J. Hayashida and R. Bise, "Cell tracking with deep learning for cell detection and motion estimation in low-frame-rate," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* Cham, Switzerland: Springer, 2019, pp. 397–405.

[29] H. Tian, G. Chen, D. Xiang, and X. Chen, "Simultaneous and automatic two surface detection of renal cortex in 3D CT images by enhanced sparse shape composition," *Proc. SPIE*, vol. 10949, Mar. 2019, Art. no. 109492D.

[30] H. Jiang *et al.*, "Improved cGAN based linear lesion segmentation in high myopia ICGA images," *Biomed. Opt. Exp.*, vol. 10, no. 5, pp. 2355–2366, 2019.

[31] Z. Zhang, H. Fu, H. Dai, J. Shen, Y. Pang, and L. Shao, "ET-Net: A generic edge-attention guidance network for medical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* Cham, Switzerland: Springer, 2019, pp. 442–450.

[32] X. J. Mao, C. Shen, and Y. B. Yang, "Image restoration using very deep convolutional encoder-decoder networks with symmetric skip connections," in *Proc. Int. Conf. Neural Inf. Process. Syst. (NIPS)*, 2016, pp. 2802–2810.

[33] Y. Tai, J. Yang, X. Liu, and C. Xu, "MemNet: A persistent memory network for image restoration," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 4549–4557.

[34] K. Zhang, W. Zuo, Y. Chen, D. Meng, and L. Zhang, "Beyond a Gaussian denoiser: Residual learning of deep CNN for image denoising," *IEEE Trans. Image Process.*, vol. 26, no. 7, pp. 3142–3155, Jul. 2017.

[35] N. Cai, F. Shi, Y. Gu, D. Hu, Y. Chen, and X. Chen, "A resnet-based universal method for speckle reduction in optical coherence tomography images," 2019, *arXiv:1903.09330*. [Online]. Available: https://arxiv.org/abs/1903.09330

[36] F. Shi *et al.*, "DeSpecNet: A CNN-based method for speckle reduction in retinal optical coherence tomography images," *Phys. Med. Biol.*, vol. 64, no. 17, Sep. 2019, Art. no. 175010.

[37] Y. Ma, X. Chen, W. Zhu, X. Cheng, D. Xiang, and F. Shi, "Speckle noise reduction in optical coherence tomography images based on edge-sensitive cGAN," *Biomed. Opt. Exp.*, vol. 9, no. 11, pp. 5129–5146, Nov. 2018.

[38] S. Chen, G. Bortsova, A. G.-U. Juárez, G. Tulder, and M. Bruijne, "Multi-task attention-based semi-supervised learning for medical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.*, 2019, pp. 457–465.

[39] Y. Zhou *et al.*, "Collaborative learning of semi-supervised segmentation and classification for medical images," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 2079–2088.

[40] Z. Xu and N. M. Deepatlas, "DeepAtlas: Joint semi-supervised learning of image registration and segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* Cham, Switzerland: Springer, 2019, pp. 420–429.

[41] S. Sabour, N. Frosst, and G. E. Hinton, "Dynamic routing between capsules," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 3856–3866.

[42] R. LaLonde and U. Bagci, "Capsules for object segmentation," 2018, *arXiv:1804.04241*. [Online]. Available: http://arxiv.org/abs/1804.04241

[43] C. Bass *et al.*, "Image synthesis with a convolutional capsule generative adversarial network," *Med. Imag. Deep Learn.*, vol. 102, pp. 39–62, Mar. 2019.

[44] A. Jaiswal, W. AbdAlmageed, Y. Wu, and P. Natarajan, "CapsuleGAN: Generative adversarial capsule network," in *Proc. Eur. Conf. Comput. Vis. (ECCV) Workshops*. Cham, Switzerland: Springer, 2018, pp. 0–10.

[45] Y. Upadhyay and P. Schrater, "Generative adversarial network architectures for image synthesis using capsule networks," 2018, *arXiv:1806.03796*. [Online]. Available: http://arxiv.org/abs/1806.03796

[46] C. Li and M. Wand, "Precomputed real-time texture synthesis with Markovian generative adversarial networks," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2016, pp. 702–716.

[47] X. Dong *et al.*, "Automatic multiorgan segmentation in thorax CT images using U-net-GAN," *Med. Phys.*, vol. 46, no. 5, pp. 2157–2168, 2019.

[48] Z. Han, B. Wei, A. Mercado, S. Leung, and S. Li, "Spine-GAN: Semantic segmentation of multiple spinal structures," *Med. Image Anal.*, vol. 50, pp. 23–35, Dec. 2018.

[49] J.-Y. Zhu, P. Krähenbühl, E. Shechtman, and A. A. Efros, "Generative visual manipulation on the natural image manifold," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2016, pp. 597–613.

[50] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* Cham, Switzerland: Springer, 2015, pp. 234–241.

[51] 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.

[52] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. A. Alemi, "Inception-v4, inception-resnet and the impact of residual connections on learning," in *Proc. AAAI*, vol. 4, 2017, p. 12.

[53] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 5967–5976.

[54] B. Wen, Y. Li, and Y. Bresler, "When sparsity meets low-rankness: Transform learning with non-local low-rank constraint for image restoration," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Mar. 2017, pp. 2297–2301.

[55] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired Image-to-Image translation using cycle-consistent adversarial networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2223–2232.

[56] P. Bao and L. Zhang, "Noise reduction for magnetic resonance images via adaptive multiscale products thresholding," *IEEE Trans. Med. Imag.*, vol. 22, no. 9, pp. 1089–1099, Sep. 2003.

[57] Z. Amini and H. Rabbani, "Statistical modeling of retinal optical coherence tomography," *IEEE Trans. Med. Imag.*, vol. 35, no. 6, pp. 1544–1554, Jun. 2016.

[58] Z. Wang, E. P. Simoncelli, and A. C. Bovik, "Multiscale structural similarity for image quality assessment," in *Proc. 37th Asilomar Conf. Signals, Syst. Comput.*, 2003, pp. 1398–1402.