第 48 卷 第 15 期/2021 年 8 月/中国激光 特邀论文

中国激光

用于视网膜 OCT 图像分层的多尺度补全卷积神经网络

袁坤，霍力*

清华大学电子工程系，北京 100084

摘要 本文提出了一种多尺度补全卷积神经网络(MsiNet)，用于光学相干层析(OCT)视网膜图像的分层。该网络充分利用了人眼视觉特性和视网膜层次特征。实验结果表明，与现有的视网膜分层网络相比，MsiNet 能够实现更高的分层正确率，同时具有网络规模小和参数量少的特性。

关键词 图像处理；人工神经网络；断层成像；光学相干层析

中图分类号 O436 **文献标志码** A **doi**：10.3788/CJL202148.1507004

1 引 言

光学相干层析(OCT)成像技术利用红外光、白光等照射生物样品，通过相干原理检测来自样品不同深度的背向散射光子，可以获得穿透深度为毫米量级、空间分辨率为微米量级的样品纵深剖面图(B-Scan)或三维结构图[1]。眼底视网膜成像是 OCT 应用最为广泛、技术最为成熟的领域，以谱域 OCT(SD-OCT)和扫频 OCT(SS-OCT)为代表的眼科 OCT 设备已经形成了一定规模的装机数量，日益成为眼底病图像学诊断的金标准。视网膜在解剖学上具有明显的层次状组织结构，这种层次结构也能清晰地体现在 OCT 的 B-Scan 中。

视网膜 OCT 图像是指利用 OCT 设备对视网膜进行三维重构得到的图像。视网膜图像的分层对于定量分析视网膜的结构变化以及病变特征的提取具有基础性支撑作用，甚至可以说，图像分层是所有眼底 OCT 数据处理中最为重要的算法单元，其优劣对于 OCT 设备的临床评价具有至关重要的意义。随着眼底 OCT 设备扫描速度的不断提升以及 OCT 血流成像(OCTA)技术的逐渐普及，OCT 数据规模迅速增加，因此，开发高效的具有鲁棒性的视网膜 OCT 自动分层算法变得越来越重要。大部分眼底疾病，例如糖尿病视网膜病变(DR)[2]，会改变眼底原本规则的层次结构，导致视网膜自动分层算法实现难度增加。

通常认为视网膜 OCT 图像的特性有以下三方面：1) 在视网膜的同一层(同一个语义类别)中，各像素的相似度极高，灰度值彼此差别不大，但由于入射光较强的相干性会形成类似噪声的散斑；2) 一般情况下，视网膜同一层中的各点连续，分割获得的视网膜的单一层次会形成一个连续的带状结构；3) 不同视网膜层次之间的边界较为模糊，层次间的灰度梯度不大。前两方面意味着视网膜图像的语义信息更加明确，可以在低分辨率对应的网络层级上采用更少的参数，而第三方面意味着视网膜边界处的分层比普通图像的分割难度更大。

近年来，深度卷积神经网络(DCNN)在图像识别领域获得了广泛应用，特别是在自动驾驶[3-5]、计算摄影[6]和生物医学图像分割[7-8]方面获得了较大成功。大多数情况下，语义分割网络[9-14]都具有庞大的规模和大量的参数，并且需要很多计算资源[15-19]，不适合用于医疗图像的分割。OCT 视网膜图像分层要求在桌面计算机上利用较为有限的计算资源，快速有效地完成大量视网膜 B-Scan 图像的处理，利用视网膜结构本身的独特性质，形成更有针对性的算法。

本课题组针对视网膜 OCT 图像的分层应用提

收稿日期：2021-04-12；**修回日期**：2021-06-03；**录用日期**：2021-07-13

基金项目：国家重点研发计划(2019YFB3800)、国家自然科学基金(61575107)

通信作者：lhuo@tsinghua.edu.cn

出了一种深度神经网络结构——MsiNet。总体上，该网络分为两步处理过程：第一步，继承编码-解码器结构，但在解码器上抛弃了常用的基于多个尺度的特征图的解码结构，只保留低分辨率特征图的部分，本文称之为单级解码器，该设计大幅减少了网络参数；第二步，将单级解码器的输出作为输入信号馈入一个图像补全网络（Inpainting Network）中，而将编码器输出的高分辨率特征图作为另一个输入送入图像补全网络，这一设计兼顾了图像细节特征。此外，本课题组设计了交织残差单元（IRU），用它来替代残差卷积单元（RCU），这种设计可使训练时的收敛速度更快，网络参数更少。此外，本课题组设计了联合权重方法以及多损失函数，这使得本文所提网络比现有的神经网络能更准确地划分层次边界。

2 相关工作回顾

早期有关于视网膜图像分割的工作主要是采用路径搜索方法[20]进行的，这种方法将视网膜图像中的像素看作是图中的节点，通过计算节点之间的某种特性（例如梯度信息）搜寻到视网膜层的交界面，这个过程将会进行若干次迭代，从而得到不同层的分界面。这一过程相对比较耗时。

由于 MsiNet 采用的网络结构是基于编码-解码器以及卷积神经网络（CNN）的，因此需要简要回顾一下相关工作。CNN[21-24]是一种基本的神经网络结构，实际中更常用的是 CNN 的变体结构，如 VGG[23]和 ResNet[24]，这类网络已被广泛应用于物体识别和图像分割任务中[25-27]。

在语义分割任务方面，Long 等[28]提出了全卷积网络（FCN），该网络去除了 CNN 网络中的全连接层[18,29]。FCN 能有效避免过拟合，但其中的解卷积过程会导致分割结果中丢失大量的细节。2015 年，Noh 等[30]提出了 DeconvNet 结构，FCN 中的解卷积由一个包含解卷积、去池化和 ReLu 的局部网络替代。Ronneberger 等[11]提出了 U-Net 结构并将其应用于生物医学图像的分割上。U-Net 建立在 FCN 的基础之上，也是一个编码-解码器结构。U-Net 的编码器部分是经典的 CNN 网络，而解码器则是对级联了多尺度和上卷积的特征图像进行处理。类似地，RefineNet[12]和 LW[31]也利用了不同分辨率下的特征，且都采用链式残差池化（CRP）在更大的图像区域中捕捉目标背景信息。DeepLab 系列[17,32-33]是非常有效的语义分割神经网络。DeepLab v3[33]总体上也是编码-解码器结构，但它采用空洞空间卷积池化金字塔（ASPP）而不是卷积-池化来获得更大的感受野，从而保留了分割图像中的细节；该结构中的空洞卷积在维度较高、特征图像规模较大时所需的计算资源较大。最近，人们又提出了 DFN[34]结构，该结构基于平滑网络和边界网络通过注意力机制进行语义分割。DeepRetina[35]以 DeepLab 为基础并以 Xception65 为骨干网络进行特征的提取，然后通过 ASPP 不同尺度的卷积来获取多维度特征并进行特征的组合，最后利用上采样恢复特征尺度，取得了良好的预测效果。对于视网膜分层，除了基于像素的语义分割外，另一个方法是利用神经网络直接寻找不同层次的交界线[8]，但这种方法得到的交界线的连续性还不够理想。

所有上述深度神经网络结构对于语义分割都有效，但计算量普遍较大，对于视网膜分层任务显得过于臃肿，不适合用于边缘设备或桌面计算机的高效计算。此外，可用于视网膜分割任务的精细像素级标注的视网膜图像数据集通常不够大，视网膜层数也局限在 10 类以内；因此，本课题组采用规模较小的 MsiNet 进行视网膜分层。

3 MsiNet 的基本原理

MsiNet 的结构如图 1 所示。输入的图像经过一系列卷积过程，形成一个特征图金字塔，金字塔的不同层级包含不同尺度上的图像信息。保留每一个

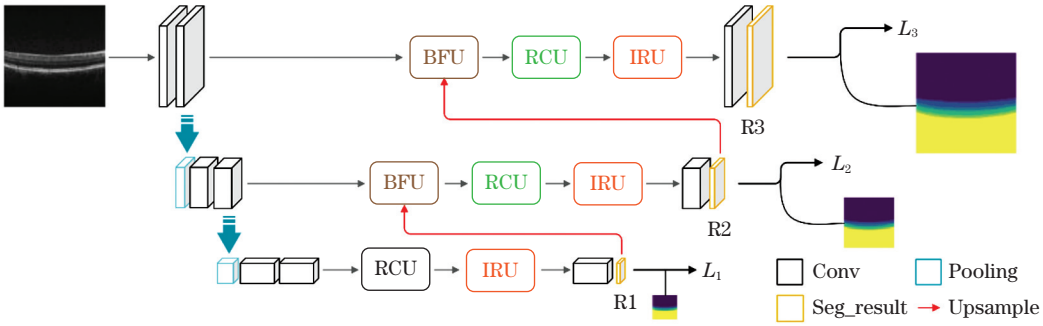

图 1 MsiNet 网络结构

Fig. 1 MsiNet structure

层级的特征，用于后续不同尺度上的图像信息的提取。所形成的特征金字塔与“编码-解码”结构中的编码器功能类似。特征金字塔最抽象层级（图 1 中最下方所示层）的特征图经过一个单级解码器直接输出分层结果 R1。R1 经过上采样后作为输入反馈到图像补全网络中，该网络的另一个输入是特征金字塔的中间层特征。图像补全网络包括一个融合单元（BFU）、一个残差单元（RCU）、一个交错残差单元（IRU）以及一个卷积层（Conv），补全网络的输出为结果 R2。R2 经过进一步的上采样后与编码器底层特征一起输入到最后一个图像补全网络中，输出分层结果 R3。可以认为本课题组设计的解码器是一个级联的多输出结构。由于获得 R1 输出只采用了单级解码器结构，因此大幅缩减了网络参数的数量。采用多输出结构能够在网络训练时提供不同尺度的监督信号，同时能够保证网络在不同尺度上的专注性。

对于卷积神经网络，当网络深度增加时会滋生梯度衰减问题，而 ResNet[21] 采用 RCU 作为基本构造模块，有效解决了梯度衰减问题。RCU 结构如图 2(a)所示。通过多个 RCU 的级联可以实现高层次的语义分割功能[12,31]。Nekrasov 等[31] 在 LW-RefineNet 中提出了轻量级残差卷积单元（LW-RCU），该单元利用了 1×1 的卷积核，因而极大地减少了参数量和计算量。然而这种方式需要大量的 LW-RCU 单元，这对于视网膜分层应用而言开销过大。为了降低参数量，本课题组提出了将两个 RCU 单元合并为一个交错残差单元（IRU）的方法，IRU 的结构如图 2(b)所示，IRU 可以看作是两个交织在一起的 RCU 单元。本课题组在实验中构造了两个神经网络，将第一个神经网络中的 IRU 单元全部替换为两个级联的 RCU 单元，其余部分保持不变；然后利用同样的数据集训练 10 次，IRU 单元和两个 RCU 单元的损失函数随训练次数的变化如图 2(c)所示，显然，IRU 的收敛速度更快，并且参数量减少了 25%。

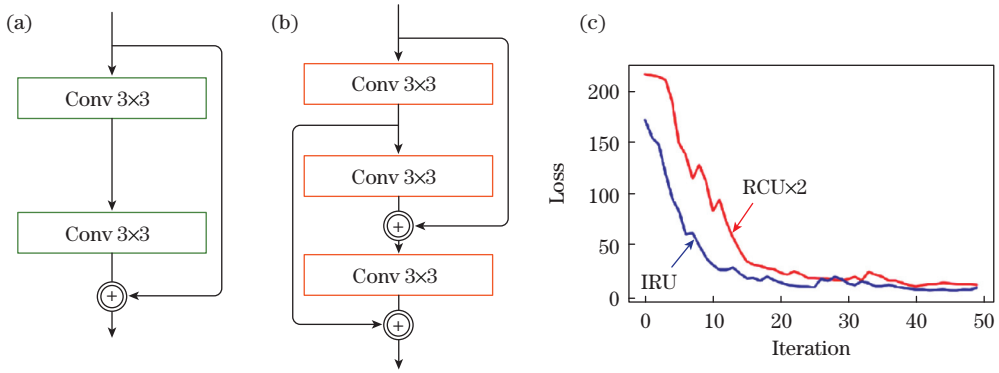

图 2 RCU 和 IRU 结构以及损失函数对比。(a) ResNet 中采用的 RCU 通用结构；(b) MsiNet 中采用的 IRU 单元结构；(c)训练过程中损失函数的趋势比较

Fig. 2 Structures of RCU and IRU and loss function comparison. (a) General network structure of residual convolution unit (RCU) in ResNet; (b) architecture of interlaced residual unit (IRU) in MsiNet; (c) loss comparison chart during training

本课题组借鉴 RefineNet 中的融合单元以及 ICNet 中的级联特征融合单元（CFF），提出 BFU 这一融合形式。图 3 给出了各种融合单元的结构示意图。与 RefineNet 中融合单元在不同尺度上采用相同大小的卷积核不同，BFU 对高分辨率特征图采用 3×3的卷积核，对低分辨率特征图采用1×1的卷积

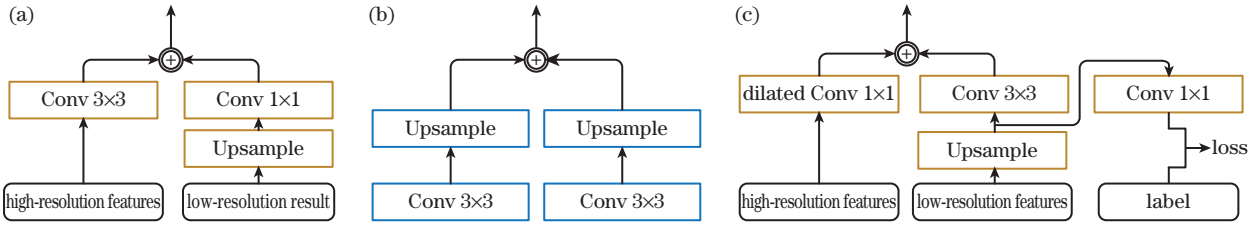

图 3 各融合单元的结构示意图。(a)MsiNet 中采用的 BFU 结构；(b)RefineNet 中的融合单元结构；(c)ICNet 中的 CFF 结构

Fig. 3 Structures of each fusion unit. (a) Structure of biased fusion unit (BFU) in MsiNet; (b) architecture of fusion unit in RefineNet; (c) CFF structure in ICNet

核，从而形成了不同的偏向性。上采样过程会保留绝大部分语义信息，但同时也会对图像边缘起到平滑作用。在 BFU 中，通过 3×3 卷积从高分辨率特征图中提取准确的边界信息对低分辨率结果进行修正。CFF 也采用偏融合的方式，但高低分辨率特征图所对应的卷积核大小与 BFU 恰好相反，其原因是通用的语义分割应用中物体的边缘非常清晰，而视网膜不同层次间的过渡比较平缓，需要引入较为模糊的处理方式才能避免分层被图像中的散斑噪声错误诱导。

MsiNet 中的图像补全网络由 BFU、RCU、IRU 和两个卷积层组成，其中：BFU 融合了高分辨率细节以及低分辨语义信息；RCU 由于其中 ReLu 函数的作用，处理效果与一个非线性判决器类似，对融合后的信息进行语义判决；IRU 和两个卷积层进一步提取特征后输出分层结果 R2 和 R3。

为了实现 MsiNet 端到端的像素级别分层训练，本课题组在不同层次（尺度）上定义了 3 个损失函数。对于低分辨率解码器输出结果 R1，利用交叉熵作为损失函数 L_1。对于图像补全网络，其作用是对低分辨输出结果进行修正，其重点应当放到“边界”信息上。本课题组提出了“广义语义边界”（generalized semantic boundary，GSE）的概念，如图 4 所示，GSE 是原始图像中的一个连续区域，包括已标注好的视网膜不同层次边界的各个像素，以及距离边界最短距离点小于等于三个像素的所有像素。GSE 能够有针对性地处理视网膜边界模糊的问题。

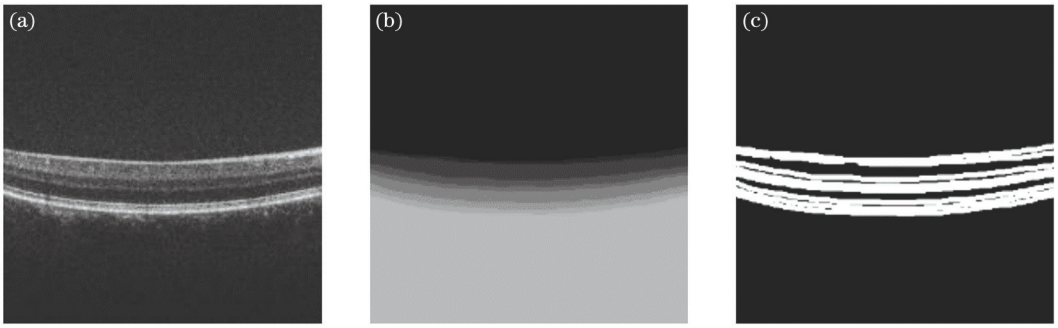

图 4 OCT 原始图像(a)、Ground Truth 分层图像(b)和 GSE(c)

Fig. 4 Original OCT image (a), Ground Truth image (b), and generalized semantic boundary(c)

L_2 和 L_3 是惩罚项修正分类结果，其定义为

$$L_{\mathrm{GSE}}(L_2,L_3)=-\sum_{i=1}^{N}(1+\theta_i)\lg\frac{\exp(\boldsymbol{x}^{(i)}[\boldsymbol{y}^{(i)}])}{\sum_{j=1}^{N}\exp(\boldsymbol{x}^{(i)}[\boldsymbol{y}^{(j)}])},\tag{1}$$

式中：$\boldsymbol{x}^{(i)}$ 是神经网络的输出向量；θ_i 是表征像素点是否在 GSE 上的数值，如果不在 GSE 上取 0，如果在 GSE 上则取一个常数 θ；θ 为超参数，需要在网络训练过程中尝试调整；$\boldsymbol{y}^{(i)}$ 是第 i 个样本的预测值输出向量；N 为训练样本的数量。

网络的总体损失函数定义为

$$L=\alpha_1L_1+\alpha_2L_2+\alpha_3L_3,\tag{2}$$

式中：α_1、α_2、α_3 表征最终 loss 的组成结构，且 $\alpha_1+\alpha_2+\alpha_3=3$。为了获得较好的边缘信息，通常 L_3 的权重要大一些，实际中 L_1、L_2 和 L_3 的取值分别为 0.8、0.8 和 1.4。

4 实　　验

采用 Duke 大学公开的 SD-OCT 眼底图像数据集，以及第三方（图湃（北京）医疗科技有限公司）提供的 SS-OCT 数据集进行训练和测试。前者包含 100 张分辨率为 512 pixel×512 pixel 的视网膜图像，通过数据增强将数据集规模扩大为 180 张，数据集中病变视网膜所占比例较大；后者共包括 953 张分辨率为 614 pixel×512 pixel 的已经过标注的视网膜 B-Scan 图像，这一数据集中健康眼底图像的比例较大。对于两个数据集，分别随机抽取其中的 30 张和 190 张图像作为测试集。

4.1 超参数 θ 的影响

超参数 θ 表征边界信息的相对重要性。在 MsiNet 的损失函数中选取 4 个不同的 θ 值（分别取 0、0.5、1.0、5.0）来评估其影响。图 5 给出了 θ 对于分层正确率的影响。图 6 中的第一列图像代表不同的具有代表性的眼底 B-Scan 图像，第二列为对应的分层 Ground Truth 标注结果。后四列代表所对应的视网膜分层输出图像。第一行和第二行图像是 SD-OCT 数据集的结果，可以看到，随着 θ 增大，视网膜水肿部分（图中异常凸起部分下面的黑色空洞）边界

的形状得到了越来越精确的修正，但当 θ 过大(θ=5.0)时，出现了一些过拟合特征，正常部位被识别为水肿。图 6 第三行和第四行图像是第三方数据集 SS-OCT 的结果，可以看出，所拍摄的视网膜边界较为简单和清晰，在第三行黄斑中央凹这一生理区域，通过增大 θ 的数值，分层的准确度会得到明显提升。总的来讲，无论是对于 SD-OCT 图像还是 SS-OCT 图像，适当增大 θ 值，都能起到增大分层准确度的目的，但过大的 θ 值反而会导致分层正确率下降。

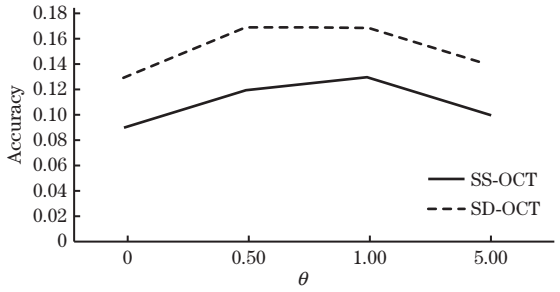

图 5　GSE 区域准确率随 θ 的变化

Fig. 5　Variation of accuracy with θ on GSE

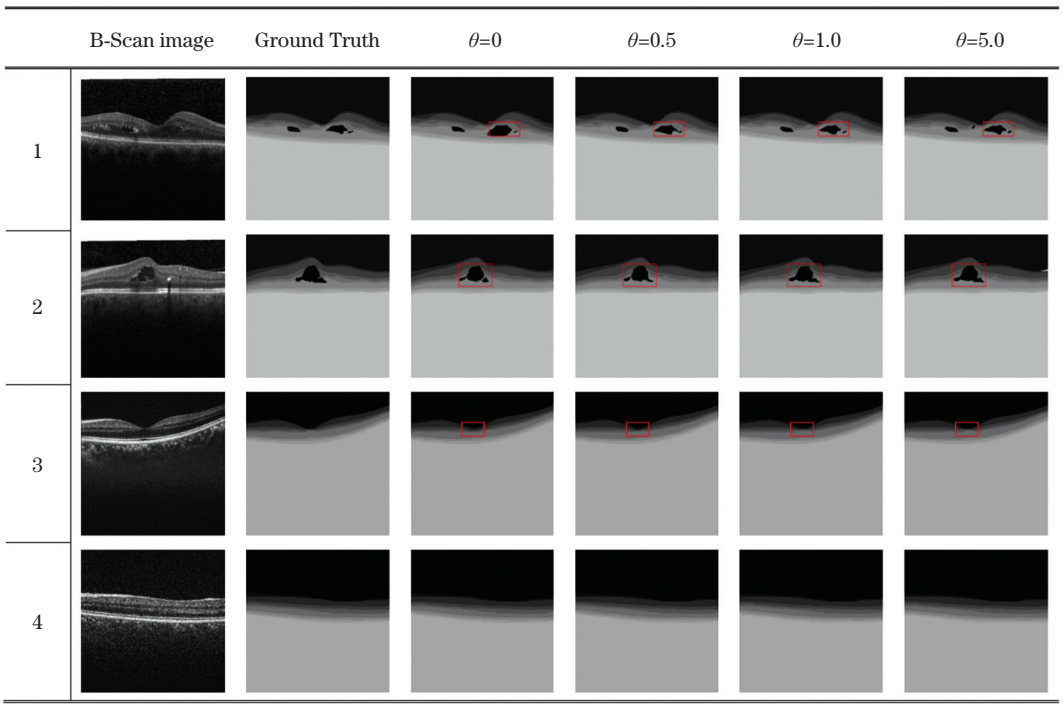

图 6　θ 取不同值时，代表性图像分层结果的定性比较

Fig. 6　Representative qualitative comparison for retina segmentation when different values of θ are applied

4.2　R1、R2 和 R3 的比较

如图 1 所示，MsiNet 有三个输出结果：R1、R2 和 R3。R1 和 R2 不仅是输出，同时也是更高分辨率图像补全网络的输入。由于下采样操作，R3、R2 和 R1 的图像尺寸逐级递减。为了清晰地看到 R1、R2 和 R3 视网膜分层结果的差异，本课题组通过线性差值的方法放大了 R1 和 R2 的尺寸，使得三者的最终尺寸与输入图像的大小相同。与多级解码器相比，R1 通过一个单级解码器获得，运算速度快，保留了大部分语义上的信息，但是分辨率低，包含了许多细节上的错误，如图 7 中的图像(1，R1)和(2，R1)所示。这种错误会在后续的图像补全网络中得到逐步修正。图 7 中的图像(1，R3)和(2，R3)由于融合了高分辨率下的细节信息和低分辨率下的语义信息，得到了正确的分层结果。在图 7 的第三列图像中，本文有意改变了输入图像的对比度，第四列图像中有部分睫毛遮挡的阴影，在这种情况下，低分辨率图像(3，R1)和(4，R1)包含了更多的错误结果，但高分辨率图像(3，R3)和(4，R3)仍能够正确地进行视网膜层的划分。

4.3　不同方法的比较

将 MsiNet 与两种其他的用于生物医学图像的神经网络进行比较性测试。第一种是 U-Net，这是一种用于医学图像处理的著名网络结构；第二种是 ReLayNet，它是专门用于视网膜分层的深度神经网络结构。本课题组搭建了三种网络，用同样的策略进行训练，三种网络的 Learning Rate 每经过 5 个 epoch 的训练后，设定到原值的 80%，Momentum 设为 0.9，训练一直持续到验证集准确度不再提升为止。对于 MsiNet，θ 值取 0.8。网络训练结束后，测试集分层结果的正确率如表 1 和表 2 所示，其中，第 2 列～第 9 列为视网膜不同层次分类的准确度，

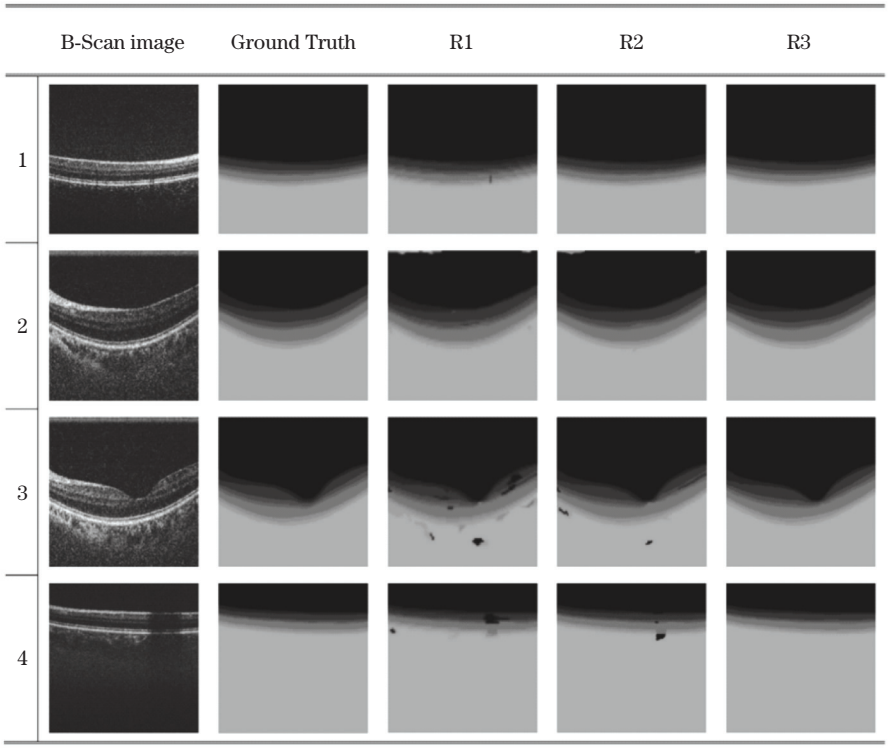

图 7 MsiNet 中不同输出分层的比较

Fig. 7 Comparison of three outputs in MsiNet

A 表示内界膜，B 表示神经纤维层到内丛状层，C 表示内侧核层，D 表示外丛状层，E 表示外侧核层到内侧节髓样体，F 表示内侧节椭球体，G 表示外侧节到视网膜色素上皮，H表示积液[7]。可以看到，无论是在 SD-OCT 数据集上还是在 SS-OCT 数据集上，MsiNet 的表现都优于其他两种网络，特别是在 SS-OCT 数据集上，各分层结果的正确性在三种网络中都呈现出最优结果。

表 1 不同方法在 SD-OCT 数据集上分层的定量结果

Table 1 Quantitative layer segmentation of each method on SD-OCT dataset

Method	Accuracy								mIOU	Classification accuracy of GSE
	A	B	C	D	E	F	G	H		
ReLayNet	0.90	0.94	0.87	0.84	0.93	0.92	0.90	0.76	0.69	0.08
U-Net	0.88	0.92	0.83	0.79	0.91	0.92	0.82	0.65	0.67	0.07
MsiNet(ours)	0.89	0.97	0.89	0.80	0.95	0.93	0.91	0.81	0.73	0.14

Notes: the 2nd to 9th columns are segmentation accuracy values of different layers. A: inter limiting membrane; B: nerve fiber layer to inner plexiform layer; C: inner nuclear layer; D: outer plexiform layer; E: outer nuclear layer to inner segment myeloid; F: inner segment ellipsoid; G: outer segment to retinal pigment epithelium; H: fluid. mIOU represents mean intersection over union.

表 2 不同方法在 SS-OCT 数据集上分层的定量结果

Table 2 Quantitative layer segmentation of each method on SS-OCT dataset

Method	Accuracy								mIOU	Classification accuracy of GSE
	A	B	C	D	E	F	G	H		
ReLayNet	0.93	0.94	0.88	0.89	0.90	0.93	0.93	—	0.71	0.12
U-Net	0.90	0.92	0.89	0.85	0.91	0.94	0.90	—	0.70	0.11
MsiNet(ours)	0.93	0.96	0.90	0.91	0.93	0.96	0.96	—	0.76	0.17

图 7 给出了三种网络具有代表性的一些分层结果对比，第一列为原始图像，第二、第三、第四列分别对应 U-Net，ReLayNet 和 MsiNet 的分层结果。U-Net 和 ReLayNet 在玻璃体、脉络膜中都引入了不正确的层次划分信息，而 MsiNet 则保持了良好的分层性能。通过表 1 可以看到，由于采用了多损失函数，MsiNet 方法将 GSE 区域的划分准确度提升到至少 5%，MsiNet 的 mIOU 明显优于其他两种网络。从参数量来看，MsiNet 的参数量大约是 4×10^6，U-Net 的参数量 3.1×10^7，而 ReLayNet 为 0.6×10^6。MsiNet 在计算量和分层准确度上取得了不错的平衡，并且适合桌面计算机进行处理。

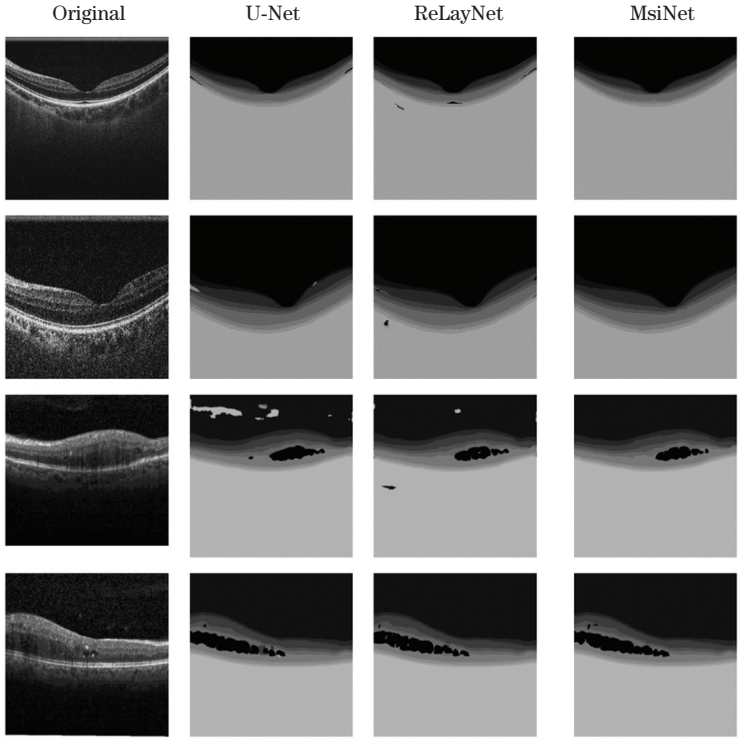

图 8 分层结果的定性比较

Fig. 8 Qualitative comparison of layer segmentation

5 结 论

结合视网膜 OCT 图像的特点，本课题组提出了一种新型的视网膜 OCT 图像分层卷积神经网络结构——MsiNet。该网络利用一个单级解码器提取语义信息，利用两个级联的图像补全网络，逐步校正分层细节。在损失函数中通过引入 GSE 提升了模糊分层边界的识别正确率。由于视网膜图像语义信息清晰，因此网络规模和参数量得以缩小。实验结果表明，MsiNet 比其他两种视网膜分层神经网络能更有效地划分出视网膜的不同层次。

参 考 文 献

[1] Huang D, Swanson E, Lin C, et al. Optical coherence tomography [J]. Science, 1991, 254 (5035): 1178-1181.

[2] Weng W, Liang Y J, Kimball E S, et al. Decreasing incidence of type 2 diabetes mellitus in the United States, 2007-2012: epidemiologic findings from a large US claims database[J]. Diabetes Research and Clinical Practice, 2016, 117: 111-118.

[3] Ess A, Mueller T, Grabner H, et al. Segmentation-based urban traffic scene understanding [C] // Procedings of the British Machine Vision Conference 2009, September, 2009, London. British: British Machine Vision Association, 2009: 84.1-84.11.

[4] Zhao H S, Qi X J, Shen X Y, et al. ICNet for real-time semantic segmentation on high-resolution images [M]//Ferrari V, Hebert M, Sminchisescu C, et al. Computer vision-ECCV 2018. Lecture notes in computer science. Cham: Springer, 2018, 11207: 418-434.

[5] Xu H Z, Gao Y, Yu F, et al. End-to-end learning of driving models from large-scale video datasets[C]// 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), July 21-26, 2017, Honolulu, HI, USA. New York: IEEE Press, 2017: 3530-3538.

[6] Yoon Y, Jeon H G, Yoo D, et al. Learning a deep convolutional network for light-field image super-resolution[C]//2015 IEEE International Conference on Computer Vision Workshop (ICCVW), December 7-13, 2015, Santiago, Chile. New York: IEEE Press, 2015: 57-65.

[7] Roy A G, Conjeti S, Karri S P K, et al. ReLayNet: retinal layer and fluid segmentation of macular optical coherence tomography using fully convolutional networks[J]. Biomedical Optics Express, 2017, 8(8): 3627-3642.

[8] Shah A, Zhou L X, Abrámoff M D, et al. Multiple surface segmentation using convolution neural nets: application to retinal layer segmentation in OCT images[J]. Biomedical Optics Express, 2018, 9(9): 4509-4526.

[9] Garcia-Garcia A, Orts-Escolano S, Oprea S, et al. A review on deep learning techniques applied to semantic segmentation [EB/OL]. (2017-04-22) [2021-04-10]. https://arxiv.org/abs/1704.06857.

[10] Wang P Q, Chen P F, Yuan Y, et al. Understanding convolution for semantic segmentation[C]//2018 IEEE Winter Conference on Applications of Computer Vision (WACV), March 12-15, 2018, Lake Tahoe, NV, USA. New York: IEEE Press, 2018: 1451-1460.

[11] Ronneberger O, Fischer P, Brox T. U-Net: convolutional networks for biomedical image segmentation[M]//Navab N, Hornegger J, Wells W M, et al. Medical image computing and computer-assisted intervention-MICCAI 2015. Lecture notes in computer science. Cham: Springer, 2015, 9351: 234-241.

[12] Lin G S, Milan A, Shen C H, et al. RefineNet: multi-path refinement networks for high-resolution semantic segmentation[C]//2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), July 21-26, 2017, Honolulu, HI USA. New York: IEEE Press, 2017: 5168-5177.

[13] Badrinarayanan V, Kendall A, Cipolla R. SegNet: a deep convolutional encoder-decoder architecture for image segmentation [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 39(12): 2481-2495.

[14] Peng C, Zhang X Y, Yu G, et al. Large kernel matters: improve semantic segmentation by global convolutional network[C]//2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), July 21-26, Honolulu, HI, USA. New York: IEEE Press, 2017: 1743-1751.

[15] Chen L C, Zhu Y K, Papandreou G, et al. Encoder-decoder with atrous separable convolution for semantic image segmentation[M]//Ferrari V, Hebert M, Sminchisescu C, et al. Computer vision-ECCV 2018. Lecture notes in computer science. Cham: Springer, 2018, 11211: 833-851.

[16] Zhao H S, Shi J P, Qi X J, et al. Pyramid scene parsing network [C]// 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), July 21-26, 2017, Honolulu, HI, USA. New York: IEEE Press, 2017: 6230-6239.

[17] Chen L C, Papandreou G, Kokkinos I, et al. DeepLab: semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2018, 40(4): 834-848.

[18] Farabet C, Couprie C, Najman L, et al. Learning hierarchical features for scene labeling [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2013, 35(8): 1915-1929.

[19] Pinheiro P H O, Collobert R. Recurrent convolutional neural networks for scene parsing[EB/OL]. (2012-06-12)[2021-04-10]. https://arxiv.org/abs/1306.2795.

[20] Chiu S J, Li X T, Nicholas P, et al. Automatic segmentation of seven retinal layers in SDOCT images congruent with expert manual segmentation [J]. Optics Express, 2010, 18(18): 19413-19428.

[21] He K M, Zhang X Y, Ren S Q, et al. Deep residual learning for image recognition [C]// 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 27-30, 2016, Las Vegas, NV, USA. New York: IEEE Press, 2016: 770-778.

[22] Szegedy C, Ioffe S, Vanhoucke V, et al. Inception-v4, inception-ResNet and the impact of residual connections on learning [EB/OL]. (2016-02-23) [2021-04-10]. https://arxiv.org/abs/1602.07261.

[23] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition[EB/OL]. (2014-09-04)[2021-04-10]. https://arxiv.org/abs/1409.1556.

[24] Krizhevsky A, Sutskever I, Hinton G E. ImageNet classification with deep convolutional neural networks [J]. Communications of the ACM, 2017, 60(6): 84-90.

[25] He K M, Gkioxari G, Dollár P, et al. Mask R-CNN [C]// 2017 IEEE International Conference on Computer Vision (ICCV), October 22-29, 2017, Venice, Italy. New York: IEEE Press, 2017: 2980-2988.

[26] Ren S Q, He K M, Girshick R, et al. Faster R-CNN: towards real-time object detection with region

proposal networks[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 39(6): 1137-1149.

[27] Liu W, Anguelov D, Erhan D, et al. SSD: single shot multibox detector[M]//Leibe B, Matas J, Sebe N, et al. Computer vision-ECCV 2016. Lecture notes in computer science. Cham: Springer, 2016, 9905: 21-37.

[28] Long J, Shelhamer E, Darrell T. Fully convolutional networks for semantic segmentation[C]//2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 7-12, 2015, Boston, MA, USA. New York: IEEE Press, 2015: 3431-3440.

[29] Ciresan D, Giusti A, Gambardella L M, et al. Deep neural networks segment neuronal membranes in electron microscopy images[J]. In Advances in Neural Information Processing Systems, 2012: 2843-2851.

[30] Noh H, Hong S, Han B. Learning deconvolution network for semantic segmentation[C]//2015 IEEE International Conference on Computer Vision (ICCV), December 7-13, 2015, Santiago, Chile. New York: IEEE Press, 2015: 1520-1528.

[31] Nekrasov V, Shen C H, Reid I. Light-weight refinenet for real-timesemantic segmentation[EB/OL]. (2018-10-08)[2021-04-10]. https://arxiv.org/abs/1810.03272.

[32] Chen L C, Papandreou G, Kokkinos I, et al. Semantic image segmentation with deep convolutional nets and fully connected CRFs[EB/OL]. (2014-12-22)[2021-04-10]. https://arxiv.org/abs/1412.7062.

[33] Chen L C, Papandreou G, Schroff F, et al. Rethinking atrous convolution for semantic image segmentation[EB/OL]. (2017-06-17)[2021-04-10]. https://arxiv.org/abs/1706.05587.

[34] Yu C Q, Wang J B, Peng C, et al. Learning a discriminative feature network for semantic segmentation[C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, June 18-23, 2018, Salt Lake City, UT, USA. New York: IEEE Press, 2018: 1857-1866.

[35] Li Q L, Li S Y, He Z Y, et al. DeepRetina: layer segmentation of retina in OCT images using deep learning[J]. Translational Vision Science & Technology, 2020, 9(2): 61.

Multiple-Scale Inpainting Convolutional Neural Network for Retinal OCT Image Segmentation

Yuan Kun, Huo Li*

Department of Electronic Engineering, Tsinghua University, Beijing 100084, *China*

Abstract

Objective Optical coherence tomography (OCT) has become the *de facto* gold standard of diagnosis in ophthalmology. In recent years, with the rapid improvement in imaging speed and the wide adoption of OCT angiography (OCTA), a large amount of retinal OCT B-Scan can be generated in one clinical scan. Automatic and effective retinal tissue segmentation is required to realize this trend. Conventional segmentation algorithms based on path searching are time consuming and error prone when dealing with morbid retinas. In such cases, neural-network (NN)-based methods such as U-Net and ReLayNet are promising approaches. These NN-based methods differ in complexity and performance. For the desktop computer in the current mainstream OCT equipment, an NN with moderate parameter sets and high performance is highly desirable. In this study, we demonstrate a novel end-to-end segmentation method for retina images, named multiple-scale inpainting convolutional NN (MsiNet) for retinal layer segmentation. MsiNet is based on human visual characteristics and can be implemented on a desktop computer with high performance. The framework was validated on two retina image datasets with comparisons against U-Net and ReLayNet, which are well established in retinal OCT image segmentation. MsiNet showed better performance than the other two methods in terms of both retinal layer segmentation accuracy and morbid tissue segmentation, with a moderate parameter set size suitable for desktop computers.

Methods MsiNet is based on semantic segmentation with encoder-decoder architecture and convolutional NNs (CNNs). We regarded retinal layers as different categories and predicted a pixel's probability to different categories. The human visual system usually detects objects in two steps: first, it tends to obtain semantic (outline, location, etc.) information; second, it is used to focusing on details. Inspired by this fact, we employed a small-scale network

as a decoder to refine semantic information and then used inpainting networks to extract spatial structures from high-resolution feature maps for inpainting low-resolution results. Thus, semantic and detailed information in different stages could be refined directly with less redundancy. To fit MsiNet into the limited computation resource, we reduced the number of parameters and floating-point operations using new structures: interlaced residual unit (IRU) and biased fusion unit (BFU). We also adopted a single-stage decoder instead of a traditional decoder and improved the segmentation results stage by stage. Further, we designed a joint weighting method for some special pixels to intensify punishment. Multiple losses were provided in different resolution stages for obtaining different resolution results. Compared with two well-established NN-based methods, the segmentation accuracy on edges was significantly improved.

Results and Discussions MsiNet was tested on two retinal OCT datasets: the SD-OCT dataset and a dataset provided by a third-party company (TOWARD π Medical Technology). We compared MsiNet with two well-established NN-based methods: U-Net and RaLayNet. First, using the GSE (generalized semantic boundary) weighting method, the accuracy of edge and disease tissue prediction of MsiNet was better than those of U-Net and RaLayNet (Table 1). Second, by comparing the outputs of different stages in Fig. 6, we confirmed that a high-level decoder significantly improves the accuracy of low-level outputs. Third, MsiNet outperformed U-Net and RaLayNet in terms of both retinal layer and morbid tissue segmentation accuracy, with a moderate parameter set size suitable for desktop computers. The results of the three methods are demonstrated in Fig. 7 and Table 1.

Conclusions Based on human visual characteristics, we propose MsiNet for retinal tissue segmentation. MsiNet replaces the traditional decoder that merges different-resolution feature maps with a single-stage decoder in low resolution, and an inpainting network is designed to rectify segmentation errors and add structural information to phased results. An extended GSE mask is applied to the loss function to adjust the weights of edge pixels. Because of the clear semantic information, the parameter set size is significantly reduced. Experiments show that MsiNet outperforms U-Net and ReLayNet in terms of both layer segmentation and morbid tissue segmentation, mainly due to the improvement in edge point classification.

Key words imaging process; artificial neural networks; tomographic; optical coherence tomography

OCIS codes 100.4996; 100.6950; 110.4500