基于 GLNet 和 HRNet 的高分辨率遥感影像语义分割

赵紫旋[1,2]，吴 谨[1,2]，朱 磊[1,3]
（1. 武汉科技大学 信息科学与工程学院，湖北 武汉 430081；2. 冶金自动化与检测技术教育部工程中心，湖北 武汉 430000；
3. 中冶南方连铸技术工程有限责任公司，湖北 武汉 430223）

摘要：在 GLNet（Global-Local Network）中，全局分支采用 ResNet（Residual Network）作为主干网络，其侧边输出的特征图分辨率较低，而且表征能力不足，局部分支融合全局分支中未充分学习的特征图，造成分割准确率欠佳。针对上述问题，提出了一种基于 GLNet 和 HRNet（High-Resolution Network）的改进网络用于高分辨率遥感影像语义分割。首先，利用 HRNet 取代全局分支中原有的 ResNet 主干，获取表征能力更强，分辨率更高的特征图。然后，采用多级损失函数对网络进行优化，使输出结果与人工标记更为相似。最后，独立训练局部分支，以消除全局分支中特征图所带来的混淆。在高分辨率遥感影像数据集上，对所提出的改进网络进行训练和测试，实验结果表明，改进网络在全局分支和局部分支上的平均绝对误差（Mean Absolute Error，MAE）分别为 0.0630 和 0.0479，在分割准确率和平均绝对误差方面均优于 GLNet。

关键词：高分辨率遥感影像；语义分割；全局分支；局部分支；独立训练
中图分类号：TP751.1 **文献标识码**：A **文章编号**：1001-8891(2021)05-0437-06

High-resolution Remote Sensing Image Semantic Segmentation Based on GLNet and HRNet

ZHAO Zixuan[1,2]，WU Jin[1,2]，ZHU Lei[1,3]
(1. *School of Information and Engineering, Wuhan University of Science and Technology, Wuhan* 430081, *China*;
2. *Engineering Research Center of Metallurgical Automation and Measurement Technology, Ministry of Education, Wuhan* 430000, *China*;
3. *WISDRI CCTEC Engineering Co. Ltd, Wuhan* 430223, *China*)

Abstract: The backbone of a convolutional neural network global branch, a residual network (ResNet), obtains low-resolution feature maps at side outputs that lack feature representation. The local branch aggregates the feature maps in the global branch, which are not fully learned, resulting in a negative impact on image segmentation. To solve these problems in GLNet (Global-Local Network), a new semantic segmentation network based on GLNet and High-Resolution Network (HRNet) is proposed. First, we replaced the original backbone of the global branch with HRNet to obtain high-level feature maps with stronger representation. Second, the loss calculation method was modified using a multi-loss function, causing the outputs of the global branch to become more similar to the ground truth. Finally, the local branch was trained independently to eliminate the confusion produced by the global branch. The improved network was trained and tested on the remote sensing image dataset. The results show that the mean absolute errors of the global and local branches are 0.0630 and 0.0479, respectively, and the improved network outperforms GLNet in terms of segmentation accuracy and mean absolute errors.

Key words: high-resolution remote sensing image, semantic segmentation, global branch, local branch, trained independently

0 引言

图像的语义分割将属于相同目标类别的图像子区域聚合起来，是高分辨率遥感影像信息提取和场景理

收稿日期：2020-04-08；**修订日期**：2020-06-23.
作者简介：赵紫旋（1997-），女，湖北武汉人，硕士研究生，研究方向为图像处理、深度学习。E-mail：zhaozixuan19970708@163.com。
通信作者：吴谨（1967-），女，安徽芜湖人，博士，教授，研究方向为图像处理与模式识别、信号处理与多媒体通信等。E-mail：wujin@wust.edu.cn。
基金项目：国家自然科学基金青年基金项目资助（61502358，61702384）。

解的基础，也是实现从数据到信息对象化提取的关键步骤，具有重要的意义。

在对高分辨率遥感影像进行语义分割时，传统方法的抗噪性能较差，难以获得较好的分割准确率和分割速度。随着大规模数据集和硬件计算能力的发展，深度学习的方法在图像处理任务中取得了较好的成绩，基于深度学习的图像语义分割方法也可以更好地应用于实际任务。

目前，基于深度学习的图像语义分割可以分为两类：基于区域分类的图像语义分割和基于像素分类的图像语义分割。其中，基于像素分类的图像语义分割方法增加了模型的整体契合度，而且可以有效提升分割准确率和分割速度。在实际应用中，多采用全监督学习的像素分类方式进行训练[1]。

全卷积网络[2]（Fully Convolutional Network, FCN）是最早实现基于像素分类的图像语义分割网络之一，在牛津大学计算机视觉组（Visual Geometry Group, VGG）所提出的 VGG-16[3]网络的基础上进行改进，将全连接层替换为卷积层以实现逐像素的密集预测，该方法可以分割出图像的大致轮廓，但是结果较为粗糙。在 FCN[2]的基础上，U-Net[4]采用编码器-解码器的对称网络结构，并通过跳层连接（skip-connection）的方式将低级特征融合至高级特征，DeconvNet[5]和 SegNet[6]也采用了类似的结构。Deeplab[7]采用空洞卷积的方式来扩大卷积的感受野，大感受野下所获取的特征能够有效地编码图像中的上下文信息。然而，这些网络应用于高分辨率遥感影像时，随着计算量增加，需要占用更多的 GPU 内存，造成运行速度减慢等问题。

随着语义分割在许多实时应用中变得越来越重要，高效和快速的分割网络得到了更多的关注。ENet[8]在进行语义分割时，采用分解滤波器策略，通过低阶近似（low-rank approximation）的方法简化卷积操作，以减少运算。图像级联网络（Image Cascade Network, ICNet）[9]将不同尺寸的低分辨率图像输入主干网络得到粗糙分割图，然后通过级联特征融合单元来融合高分辨率特征图，提高了分割速度。尽管实时性能得到提升，但这些网络对于高分辨率遥感影像的分割准确率不佳。

GLNet[10]由全局分支（global branch）和局部分支（local branch）构成，分别以降采样的全局图像和全分辨率的裁剪图像作为输入，有效保留了细节信息和全局上下文信息，并能减少 GPU 内存的使用量，该网络以 ResNet[11]和特征金字塔网络（Feature Pyramid Network，FPN）[12]作为两个分支的主干。GLNet 在保证高分辨率图像分割准确率的前提下，提高了内存的使用效率，但是其全局分支主干网络侧边输出的特征图分辨率较低，而且表征能力不足，局部分支的学习也存在被混淆的问题。

HRNet[13]能够全程保持高分辨率的特征图，得到更为精准的空间信息，其多尺度融合策略也可以得到更为丰富的高分辨率表征，使预测的热点图（heatmap）更为准确。采用 HRNet 替代 GLNet 中的全局分支 ResNet，可以得到分辨率更高，而且特征信息更为丰富的侧边输出特征图，可以提高分割的准确率。

本文基于 HRNet 和 GLNet 提出了一种改进网络用于高分辨率遥感影像的语义分割。在全局分支中，将 GLNet 中的 ResNet 调整为 HRNet，以 HRNet 和 FPN 作为主干网络，在保持高分辨率特征图的同时，融合丰富的多尺度信息，得到更具代表性的表征。将辅助损失函数修改为多级损失结构优化网络[14]，使分割结果更为准确。在局部分支中，采用 ResNet 和 FPN 作为主干网络，独立训练该部分网络，不采用原有的特征共享策略，以消除全局分支中，未充分学习的特征图所带来的混淆。GLNet 和改进网络的流程，分别如图 1(a)和图 1(b)所示，在高分辨率遥感影像数据集上对两者进行比较，实验结果表明，该网络在分割准确率上优于 GLNet，得到了更好的结果。

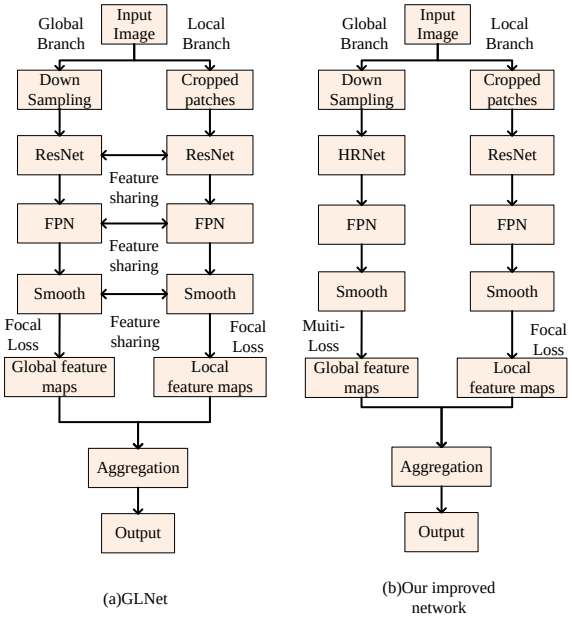


图 1 GLNet 和本文改进网络的流程对比

Fig.1 The comparison of GLNet and network structure proposed in this paper

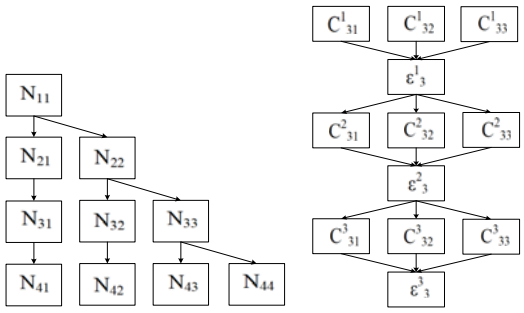
1 HRNet

HRNet 并行连接由高到低的子网络，在不采用中间热点图监督的条件下，重复融合子网络产生的表征，得到可靠的高分辨率表征。与通过由低到高的上采样

进程，聚合低层和高层表征的大多数网络相比，HRNet 具有较好的计算复杂度和参数效率。

1.1 并行多分辨率子网

并行多分辨率子网通过并行连接由高分辨率到低分辨率的子网构建而成，每个子网包含多个卷积序列，临近的子网间存在降采样层，以将特征图分辨率减半。以高分辨率子网作为第一个阶段，逐步增加由高到低分辨率的子网，组成新的阶段，然后并行连接多个分辨率子网。并行子网后一阶段的分辨率由前一阶段的分辨率和下一阶段的分辨率组成。通过 4 个并行子网组成的网络结构，如图 2(a)所示。图中：N_{sr} 表示在第 s 阶段的子网，其分辨率为初始阶段图像的 $1/(2^{r-1})$。



(a) 并行网络结构 (b) 尺度融合网络结构

(a) Paralled network structure (b) Scale fusion network structure

图 2 HRNet 的基本网络结构

Fig.2 Basic Network structure of HRNet

1.2 重复的尺度融合

HRNet 引入跨并行子网的交换单元，使每个子网多次接收来自其它子网的信息。信息交换单元的示例如图 2(b)所示，将第 3 阶段分隔为多个交换块，而每个交换块由 3 个并行卷积单元和 1 个交换单元组成。图中：$C_{sr}^{\ b}$ 表示第 s 阶段的第 b 个交换块，其交换单元的分辨率为初始阶段的 $1/(2^{r-1})$，而 ε_s^b 表示相应的交换单元。HRNet 中交换单元聚合不同分辨率特征信息的具体实现，如图 3 所示。

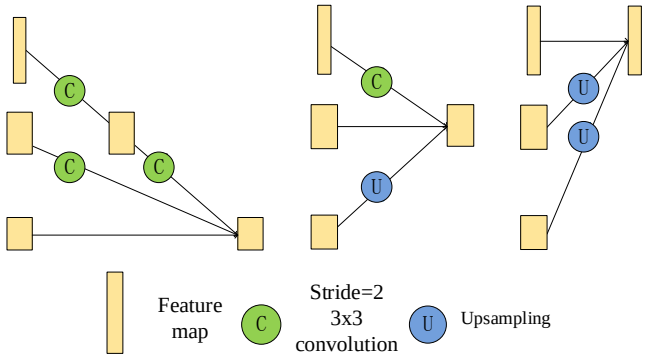


图 3 HRNet 的交换单元

Fig.3 Exchange unit of HRNet

交换单元以 s 个响应图$\{X_1, X_2, \cdots, X_s\}$作为输入，每个输出均由输入的响应图聚合而得，相应的输出用$\{Y_1, Y_2, \cdots, Y_s\}$表示，其中，$Y_i$ 与 X_i 的分辨率及维度一致。由输入到输出的表示为：$Y_k=\sum_{i=1}^{s}a(X_i,k)$，每个跨阶段的交换单元有一个额外的输出 Y_{s+1}，且 $Y_{s+1}=a(Y_s, s+1)$。

$a(X_i, k)$表示将输入 X_i 的分辨率由 i 变换到 k 的过程，通过降采样或上采样的方式实现。HRNet 的交换单元采用步长为 2 的 3×3 卷积进行降采样，而上采样则利用双线性插值的方式实现。值得注意的是，如果 $i=k$，则 $a(x_i, k)$表示恒等映射，即 $a(x_i, k)=x_i$。

2 基于 HRNet 和 GLNet 的网络

基于 HRNet 和 GLNet 的网络由全局和局部两个分支构成。全局分支以降采样后的整体图像作为输入，保留了图像的全局上下文信息，但缺少了部分的细节信息；局部分支以全分辨率的裁剪图像作为输入，高分辨率图像保留了细节信息，但缺少了空间信息和邻近区域依赖信息。本文改进的网络在全局分支采用了 HRNet 的结构提高了特征图的分辨率，在局部分支采用独立训练的方式保证网络学习效率，并有效整合两个分支，实现更好的分割性能。

2.1 全局分支和局部分支

基于 HRNet 和 GLNet 的网络结构如图 1(b)所示，将高分辨率图像数据集 D 中的 N 张原始图像和分割图作为网络的输入，$D=\{(I_i,S_i)\}_{i=1}^{N}$，其中 I_i 和 S_i 分别表示第 i 幅的原始图像和分割图，$I_i,S_i\in R^{H*W}$，R^{H*W} 表示图像的尺寸为 $H*W$。全局分支以降采样后的低分辨率图像数据集 D^G 作为输入，$D^G=\{(I_i^G,S_i^G)\}_{i=1}^{N}$，$I_i^G$ 和 S_i^G 分别表示第 i 幅的低分辨率原始图像和分割图；而局部分支对 D 进行裁剪，并以全分辨率的裁剪图像组 D^{L} 作为输入，$D^{\mathrm{L}}=\left\{\{I_{ij}^{\mathrm{L}},S_{ij}^{\mathrm{L}}\}_{j=1}^{n_i}\right\}_{i=1}^{N}$，将 D 中的 I_i,S_i 分别裁剪至 n_i 块子图像，I_{ij}^{L} 和 S_{ij}^{L} 分别表示对第 i 幅图像进行裁剪后的第 j 块子图及其分割图，I_i 和 S_i 并非随机裁剪，而是有序的完全裁剪，以便于训练和测试。其中，$I_i^G,S_i^G\in R^{h_1*w_1}$，$R^{h_1*w_1}$ 表示全局分支输入图像的尺寸为 h_1*w_1，$I_{ij}^{\mathrm{L}},S_{ij}^{\mathrm{L}}\in R^{h_2*w_2}$，$R^{h_2*w_2}$ 表示局部分支输入图像的尺寸为 h_2*w_2，$h_1,h_2\ll H$，$w_1,w_2\ll W$。

在全局分支中，采用 HRNet 作为主干网络。其并

行连接由高分辨率到低分辨率的子网，重复融合多分辨率特征，生成了可靠的高分辨率表征。与原有 ResNet 网络相比，提高了特征图的分辨率，在丰富了全局上下文信息的同时，保留了更多的细节信息，提高了分割效率，其网络结构如图 4 所示。

在局部分支中，仍采用 ResNet 作为主干网络。与原有网络不同，本文所提出的改进网络并未与全局分支深度共享特征图。全局分支在对局部分支缺少的上下文信息进行补充的同时，也对其特征图的学习造成混淆。因此，采用独立训练的方式，以提高局部分支的分割效果，局部分支的网络结构如图 5 所示。

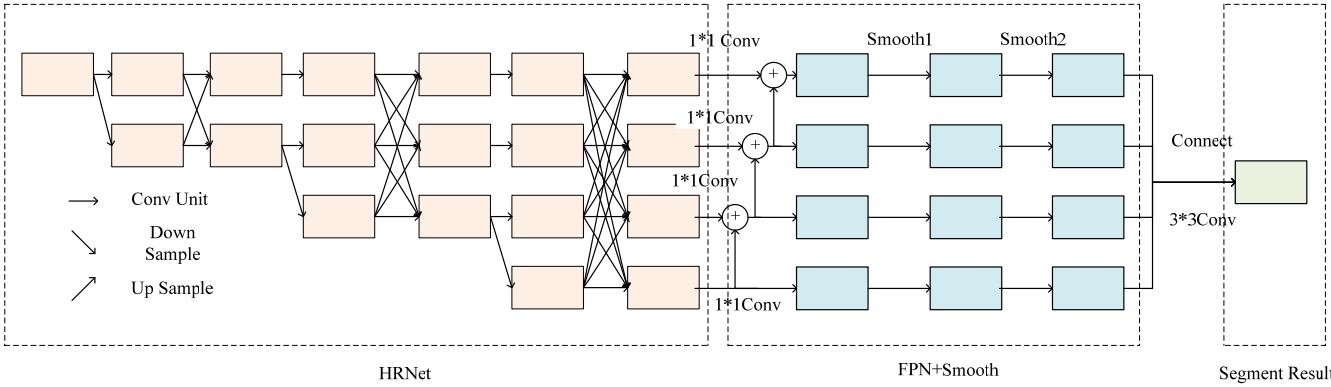


图 4 GLNet 的全局分支结构

Fig.4 The structure of global branch

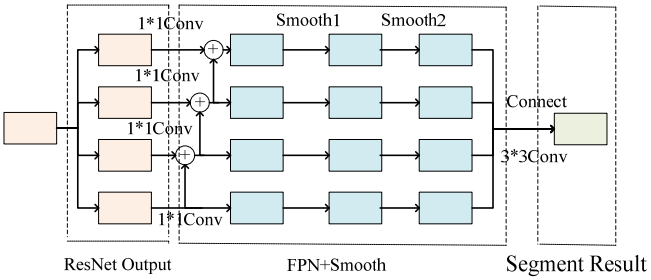


图 5 GLNet 的局部分支网络结构

Fig.5 The structure of local branch of GLNet

2.2 分支聚合

如图 6 所示，令两个分支间的聚合层为 f_{AGG}，该层由 3×3 卷积构成，实现了两个分支特征图之间的聚合（ensemble）。从局部分支和全局分支中提取的特征图可以分为 L 层，分别用 $X_{L,i}$ 和 $X_{G,i}$ 表示，其中 $i \in L$，$L=4$，将最后一层特征图沿着维度相连，通过聚合层 f_{AGG} 得到最后的分割图，令其为 S_{AGG}。除了针对于 S_{AGG} 的主损失函数，本文还采用了 2 个辅助损失函数，分别使全局分支的分割图为 $S_{G,L}$、局部分支的分割输出为 $S_{L,L}$，与相对应的人工标记结果（Ground Truth，GT）更为接近，该操作也使得训练过程更为稳定。

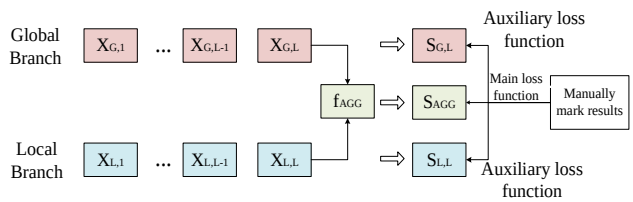


图 6 聚合过程

Fig.6 The process of aggregation

3 实验结果

3.1 实现细节

在全局分支，采用 HRNet 和 FPN 作为主干网络，其中，HRNet 包含 4 个并行子网，每个子网分辨率递减一半，而通道数增加至上一阶段的 2 倍。HRNet 的第一阶段由 4 个与 ResNet-50 结构一致的残差单元构成；第二，三，四阶段分别包含有 1，4，3 个交换块，每个交换块在相同分辨率之间包含有 2 个 3×3 卷积，而在不同分辨率之间包含 1 个交换单元。在进入第一阶段前，HRNet 需要经过 2 次降采样，本文实验，在输入降采样后的遥感图像前提下，仅通过 1 次降采样即可实现最优效果。

在局部分支，仍采用 ResNet 和 FPN 作为主干网络，提取 ResNet 在从第 2 个至第 5 个残差块的侧边输出进行学习。

全局分支输入的降采样图像与局部分支输入的裁剪图像均采用 500×500 的像素大小。局部分支中相邻的裁剪子图有 50 个像素的重叠，以避免卷积层的边界消失，并采用 $\gamma=6$ 的主损失函数和两个辅助损失函数来优化目标[15]，全局分支采用多级损失[14]的计算方式，每条分支的损失权重平均分配为 1，局部分支采用二元交叉熵损失函数。

实验是在 PyTorch 深度学习框架下进行，采用 Adam 优化器[16]（$\beta_1=0.9$，$\beta_2=0.999$），全局分支的学习率设置为 1×10^{-4}，局部分支的学习率设置为 2×10^{-5}，所有分支训练时批量处理的数量均为 6。

3.2 实验数据及评估标准

在遥感数据集上进行训练和测试，数据集中包含有 759 幅高分辨率的遥感影像（分辨率为 2248×2248），以及对应的人工标记结果。将数据集按 8:2 的比例划分成训练集和验证集，607 幅遥感影像用于训练，152 幅遥感影像用于测试。

在实验过程中，采用平均绝对误差作为评估标准。MAE 是单个观测值与标准值的偏差的绝对值平均，所有个体差异在平均值上的权重都相等，可以更好地反映预测值误差的实际情况，其表达式如下所示：

$$\mathrm{MAE}(x,h)=\frac{1}{m}\sum_{i=1}^{m}\left|h(x_i)-y_i\right|$$

式中：x 表示输入数据集；m 表示 x 中数据的总量；h 表示预测过程；$h(x_i)$和 y_i 分别表示第 i 个数据的预测值和标准值。

在 Linux 实验环境下，通过 Nvidia-smi 命令调取并记录 GPU 的内存占用情况。

3.3 实验结果

分别采用 GLNet 和改进网络对两幅相同的高分辨率遥感影像进行分割，结果对比如图 7 所示。

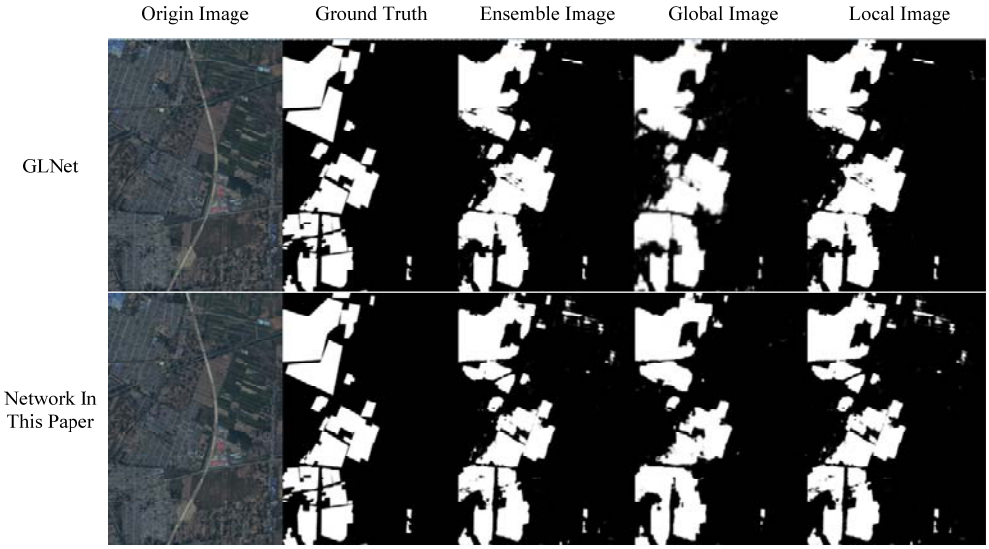


图 7 高分辨率遥感图像分割结果

Fig.7 The segmentation results of high resolution remote sensing images

在全局分支上，将本文所提出的改进网络与 GLNet 进行对比。平均绝对误差和内存占用量的对比数据如表 1 所示。

表 1 Global 分支实验对比

Table 1 Comparison of global branch experiments

	MAE	GPU Memory/M
GLNet	0.0730	1980
Ours	0.0630	2715

在局部分支上，将本文所提出的改进网络与 GLNet 中特征图共享的网络结构进行对比。平均绝对误差和内存占用量的对比数据如表 2 所示。

表 2 Local 分支实验对比

Table 2 Comparison of local branch experiments

	MAE	GPU Memory/M
GLNet	0.0572	1900
Ours	0.0479	1869

精度召回率（Precision Recall，PR）曲线是衡量学习器优劣的标准之一，其曲线下的面积（Area Under Curve，AUC）用以定量的评估分割效果。本文所提出的改进网络和 GLNet 的 PR 曲线如图 8 所示。

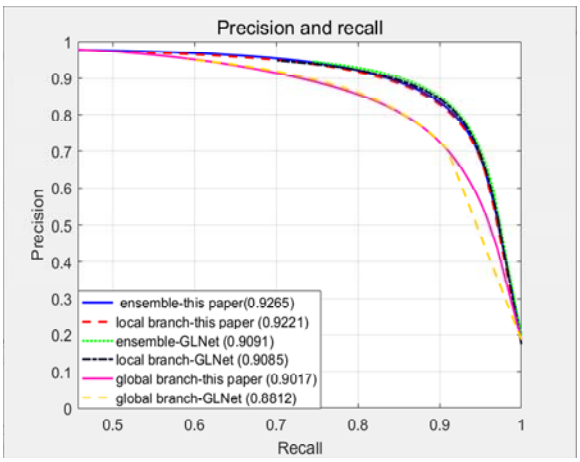


图 8 高分辨率遥感图像 PR 曲线图

Fig.8 The PR curves of high-resolution remote sensing image

在全局分支，本文基于 HRNet 主干结构的 MAE＝0.0630，平均绝对误差相比于 GLNet 降低了 0.01，由图 7 所见，改进网络全局分支的分割结果图也明显优于 GLNet，证明本文所提出的结构在全局分支可以更好地进行高分辨率遥感影像语义分割；在局部分支，未融入全局分支特征图的方法，MAE＝0.0479，平均绝对误差更低，可以推断全局分支对局部分支的学习造成了混淆，从分割图也可看出，不采用特征图共享的结构分割效果更好。由图 8 中的 PR 曲线定量分析可知，改进网络在融合、局部和全局模块的 AUC 均高于 GLNet 的对应模块，分别为 0.9265、0.9221 和 0.9017，具有更好的分割效果。综上所述，本文基于 HRNet 和 GLNet 的方法，在分割准确度和平均绝对误差方面性能更优。值得注意的是，该方法在内存使用效率方面，稍弱于 GLNet，全局分支的 GPU 内存使用仅比 GLNet 多了 735M，优于多数主流分割网络（U-Net：5507M，FCN-8s：5227M，PSPNet[17]：6289M，SegNet[6]：10339M）。

4 结论

高分辨率遥感影像分割是当前非常重要的计算机视觉技术之一。GLNet 在保证分割效率的同时，优化 GPU 内存的使用效率。本文在 GLNet 的基础上，对分割准确率进一步提升。在全局分支，采用 HRNet 作为主干网络，并采用多级损失函数优化网络；在局部分支，独立训练该部分网络，证明了全局分支中的特征图对局部分支特征图的学习造成了混淆。实验结果表明，本文所提出改进网络的平均绝对误差更低，且分割准确率更优，该方法可有效用于高分辨率遥感影像的语义分割。

参考文献：

[1] 田萱，王亮，丁琪. 基于深度学习的图像语义分割方法综述[J]. **软件学报**, 2019, **30**(2): 250-278.
TIAN X, WANG L, DING Q. Review of Image Semantic Segmentation Based on Deep Learning[J]. *Journal of Software*, 2019, **30**(2): 440-468.

[2] LONG J, Shelhamer E, Darrell T. Fully convolutional networks for semantic segmentation[C]//*Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015: 3431-3440.

[3] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition[C]// *Proceedings of International Conference on Learning Representation*, 2015: 1-13.

[4] Ronneberger O, Fischer P, Brox T. U-net: Convolutional networks for biomedical image segmentation[C]//*Proceedings of International Conference on Medical Image Computing and Computer-assisted Intervention*, 2015: 234-241.

[5] Noh H, Hong S, Han B. Learning deconvolution network for semantic segmentation[C]//*Proceedings of the IEEE International Conference on Computer Vision*, 2015: 1520-1528.

[6] 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.

[7] CHEN L C, Papandreou G, Kokkinos I, et al. Deep Lab: 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.

[8] Paszke A, Chaurasia A, Kim S, et al. Enet: A deep neural network architecture for real-time semantic segmentation[J/OL]. *arXiv preprint*, 2016, https://arxiv.org/abs/1606.02147.

[9] ZHAO H, QI X, SHEN X, et al. Icnet for real-time semantic segmentation on high-resolution images[C]//*Proceedings of the European Conference on Computer Vision*, 2018: 405-420.

[10] CHEN W, JIANG Z, WANG Z, et al. Collaborative global-local networks for memory-efficient segmentation of ultra-high resolution images[C]// *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019: 8924-8933.

[11] HE K, ZHANG X, REN S, et al. Deep residual learning for image recognition[C]//*Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016: 770-778.

[12] LIN T Y, Dollár P, Girshick R, et al. Feature pyramid networks for object detection[C]//*Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017: 2117-2125.

[13] SUN K, XIAO B, LIU D, et al. Deep high-resolution representation learning for human pose estimation[C]//*Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019: 5693-5703.

[14] HOU Q, CHENG M M, HU X, et al. Deeply supervised salient object detection with short connections[C]//*Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017: 3203-3212.

[15] LIN T Y, Goyal P, Girshick R, et al. Focal loss for dense object detection[C]//*Proceedings of the IEEE International Conference on Computer Vision*, 2017: 2980-2988.

[16] Kingma D P, Ba J. Adam: A method for stochastic optimization[J/OL]. arXiv preprint arXiv:1412.6980, 2014.

[17] ZHAO H, SHI J, QI X, et al. Pyramid scene parsing network[C]//*Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017: 2881-2890.