doi:10. 3969/j.issn. 1672-4623.2023.02.003

基于卷积神经网络模型的照片精细化识别研究

张俊鑫[1]，郭海京[1,3*]，唐孝培[2]，金诗程[1]
（1. 广东省国土资源测绘院，广东 广州 510500；2. 中山大学地理科学与规划学院，广东 广州 510275；
3. 自然资源部华南热带亚热带自然资源监测重点实验室，广东 广州 510663）

摘　要：目前的照片精细化识别主要是基于传统机器学习，识别精度无法保证大范围应用，部分基于神经网络模型的方法在提升精度的同时难以兼顾效率。基于此，提出一种轻量化的卷积神经网络模型，减少了模型的参数量与计算量，从而提高速度。方法包含跨阶段融合以及多尺度融合策略，强化模型的特征学习能力，提升模型的预测精度。在利用三调照片对菠萝、柚子、香蕉等地表植被的识别中，通过与主流模型的对比实验发现，提出的模型在识别速度和精度方面均有较大提高。

关键词：国土三调；精准识别；地表覆盖；卷积神经网络
中图分类号：P208　　**文献标志码：**B　　**文章编号：**1672-4623（2023）02-0011-06

Research on Refinement Recognition of Pictures Based on Convolutional Neural Network Model

ZHANG Junxin[1], GUO Haijing[1,3], TANG Xiaopei[2], JIN Shicheng[1]
(1. Institute of Land and Resources Surveying and Mapping of Guangdong Province, Guangzhou 510500, China;
2. School of Geography and Planning, Sun Yat-Sen University, Guangzhou 510275, China;
3. Key Laboratory of Natural Resources Monitoring in Tropical and Subtropical area of South China, Ministry of Natural Resources, Guangzhou 510663, China)

Abstract: The current photo refined recognition is mainly based on traditional machine learning, and the recognition accuracy cannot guarantee a wide range of applications. Some methods based on neural network model are difficult to take into account the efficiency while improving the accuracy. In this paper, we proposed a lightweight convolutional neural network(CNN) model, which could reduce the amount of parameters and the calculation of model, thereby improving the speed. This method includes cross-stage fusion and multi-scale fusion strategies to strengthen the feature learning ability of model and improve the prediction accuracy of model. In the identification of pineapple, grapefruit, banana and other surface vegetation using pictures in the third territorial survey, we carried out the comparison experiment with the mainstream model. It is found that the model proposed in this paper has improved both in terms of recognition speed and accuracy.

Key words: the third territorial survey, precise identification, land coverage, CNN

深度学习是特殊的机器学习算法，其模型是深层的人工神经网络。深度学习最早由多伦多大学教授Hinton[1-2]提出，是指一种模拟人类大脑研究学习的人工神经网络算法。其主流算法包括受限波尔兹曼机（RBM）、深度信念网络（DBN）、卷积神经网络（CNN）以及自动编码器（AE）等，后3种算法广泛应用于遥感影像分类与识别中，以CNN模型最为典型[3]。然而，影像数据局限于单一视角，主要关注于目标的空间布局特征，忽略了剖面和层次特征，无法实现更细致的分类[4]。因实景照片多视角、高分辨率等特点，研究者开始关注基于照片的分类方法，通过深度学习CNN对照片进行识别[5-6]，实现更加精细化的分类。

本文提出一种轻量化的神经网络模型，基于三调照片实现地表覆盖类型的精细化识别。实验中，以广东省部分县区为对象，利用8 037张三调照片样本进行模型训练与分类实验研究，取得了较好的效果。研究结果表明，运用轻量化神经网络模型检测三调实地照片，可以识别地表上的柚子、香蕉、菠萝等作物，满足国土空间的精细化管理需求，同时为广东省地区的耕地恢复潜力分析提供数据支撑。

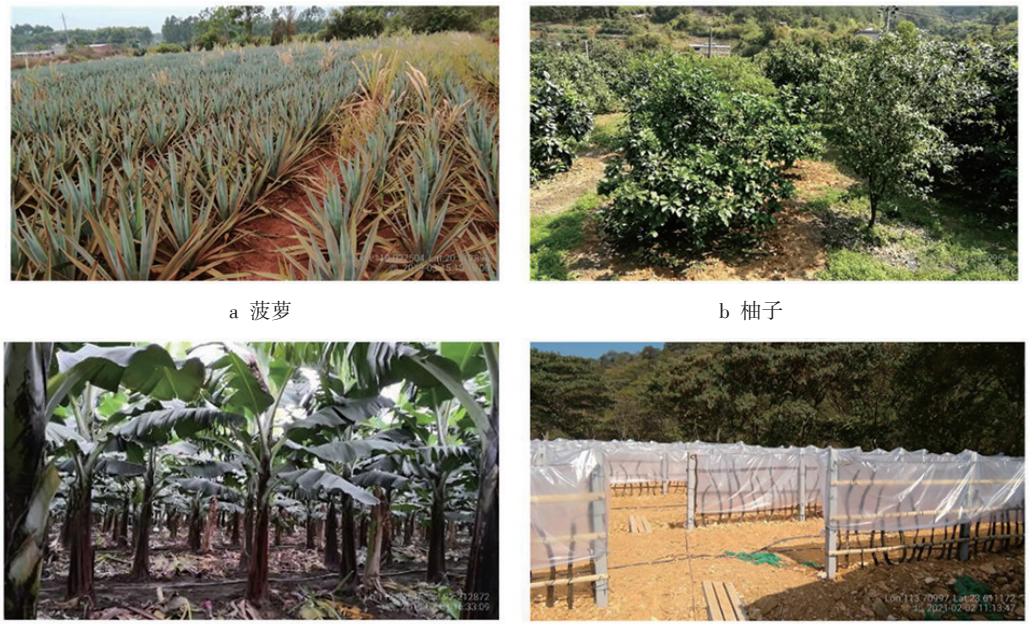
收稿日期：2022-02-05。
项目来源：广东省省级科技计划项目（2021B1212100003）。
（*为通信作者）

1 研究数据

本文使用广东省部分县区的三调照片作为模型的验证与分析数据，对野外实地拍摄的照片进行人工标注，照片像素为1 920×1 080，共标记了8 037张照片用于模型的训练与验证，其中菠萝照片4 073张、柚子照片1 798张、香蕉照片538张、其他照片1 628张（如表1所示），样本照片如图1所示。

表1 样本类别及数量

类别	编号	数量
菠萝	0	4 073
柚子	1	1 798
香蕉	2	538
其他	3	1 628
总计	–	8 037



a 菠萝 b 柚子 c 香蕉 d 其他

图1 样本照片

2 轻量化神经网络模型

卷积神经网络的参数量通常较多，在实际应用中受到计算能力的限制，为了使得本文的方法能够得到大范围推广应用，本文的核心目的是提出一个轻量化的神经网络模型，减少模型的计算量与参数量，具体的网络结构见图2。基础网络结构的灵感来自于EfficientNet[7]，主要包含CSPMBConv和CSPMBConvBlock等基础模块，各模块的详细结构见图3。同时考虑到三调数据的场景特征丰富，在网络中加入了多尺度特征融合，有效获取不同尺度的场景特征信息。基于此，本文提出了一个高效多尺度场景卷积神经网络（efficient multi-scale context network，EMCNet）。

神经网络通常是由许多卷积层构成，一个卷积网络层 i 可以定义为一个函数：$Y_i=F_i(X_i)$，其中 F_i 是算子，Y_i 为输出张量，X_i 为输入张量，张量形状为 $\langle H_i, W_i, C_i\rangle$，其中 H_i 和 W_i 为输入张量的空间分辨率大小，C_i 为通道数量。卷积神经网络 N 可以由一系列卷积层来表示：

$$N=F_k\odot\ldots\odot F_2\odot F_1(X_1)=\odot_{j=1\ldots k}F_k(X_1) \tag{1}$$

在实际应用中，卷积神经网络层通常被划分为多个阶段，每个阶段的所有层都具有相同的结构，例如ResNet[8]有5个阶段，除了第一层进行下采样外，其余每个阶段的卷积层都包含相同的卷积类型。因此，我们可以将卷积神经网络定义为：

$$N=\bigodot_{i=1\ldots s}F_i^{L_i}\left(X_{\langle H_i, W_i, C_i\rangle}\right) \tag{2}$$

式中，$F_i^{L_i}$ 为卷积层 F_i 在第 i 阶段重复了 L_i 次；$\langle H_i, W_i, C_i\rangle$ 为第 i 层的输入张量 X 的大小。

与常规的卷积神经网络设计主要关注于寻找最佳的网络层结构 F_i 不同，模型缩放试图扩大网络深度（L_i）、宽度（C_i）以及分辨率（H_i, W_i），而不改变基础网络中预定义的 F_i。通过修正 F_i，模型缩放简化了计算资源约束下的模型结构设计问题，但仍有很大的空间去探索网络每一层中不同的 L_i，C_i，H_i，W_i 组合。为此，谷歌大脑团队提出了EffcientNet[7]，可以

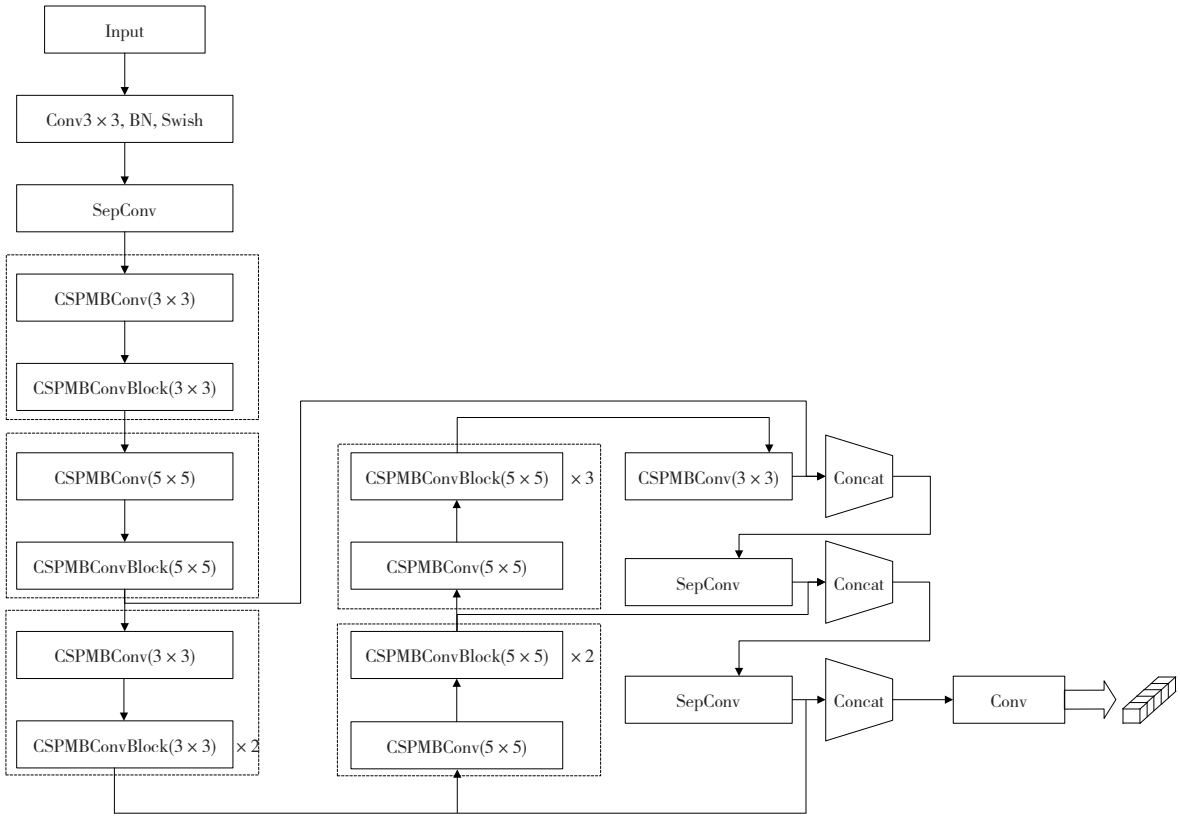


图2　网络结构图

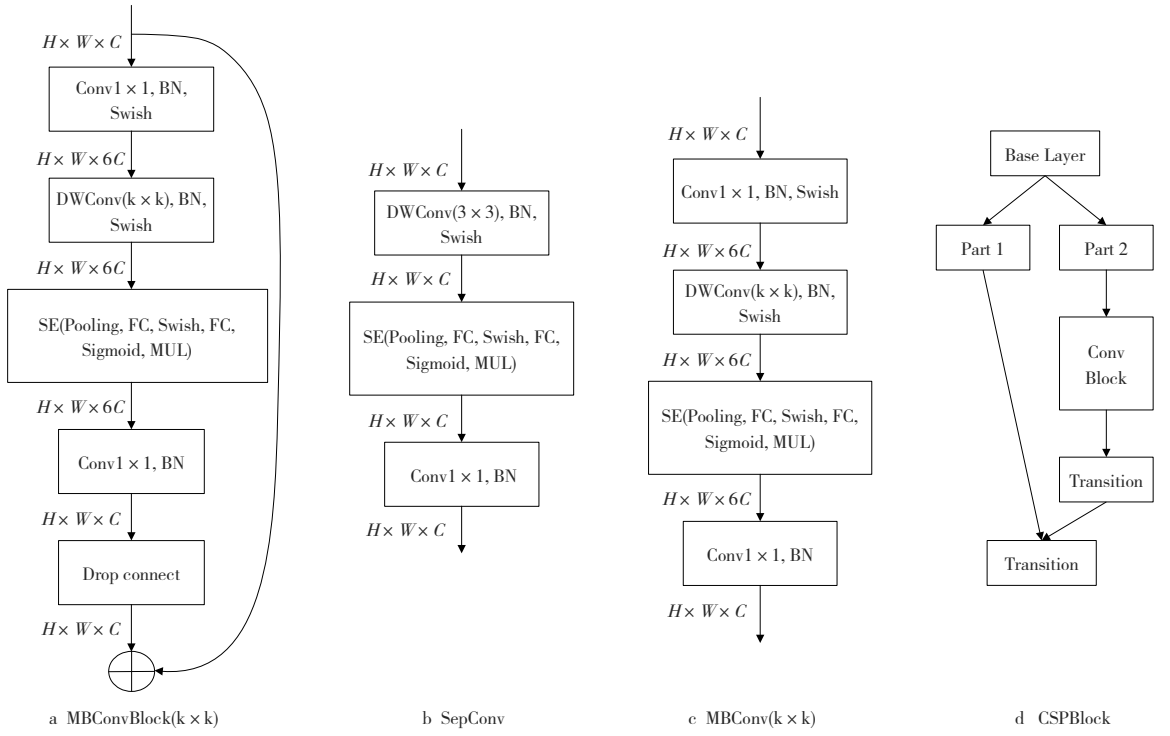


图3　网络模块细节图

将网络缩放表述为一个优化问题：

$$\max_{d,w,r} \text{Accuraccy}\left(N(d,w,r)\right) \tag{3}$$

$$\text{s.t. } N(d,w,r)=\bigodot_{i=1,2,\ldots,s}\hat{F}_i^{d\cdot\hat{L}_i}\left(X_{\langle r\cdot\hat{H}_i,\, r\cdot\hat{W}_i,\, w\cdot\hat{C}_i\rangle}\right) \tag{4}$$

$$\text{Memory}(N)\leqslant \text{targ}\,et_\text{memory} \tag{5}$$

$$\text{FLOPS}(N)\leqslant \text{targ}\,et_\text{flops} \tag{6}$$

式中，w,d,r 为网络缩放的宽度、深度和分辨率的系数；$\hat{F}_i$，$\hat{L}_i$，$\hat{C}_i$，$\hat{H}_i$，$\hat{W}_i$ 为基础网络中预定义的参数。

网络缩放的主要难点在于最优的 d,w,r 存在相互依赖，且在不同的资源约束下取值也会发生变化。正是由于这个难点的存在，传统方法一般只考虑其中一个参数的变化。

缩放网络深度是许多卷积网络最常用的方法[8, 10-11]。理想情况下，更深层的卷积神经网络可以捕捉到更丰富、更复杂的特征信息，并且同时可以保持良好的泛化能力。然而，梯度消失问题会导致更深的网络也更难以训练[12]。尽管近年来提出的跳跃连接（skip connections）[8]和批归一化（batch normalization）[13]缓解了训练问题，但更深的网络所带来的精度增益逐渐减小。例如 ResNet-1000 虽然有更多的层，但精度与 ResNet-101 相似。

缩放网络宽度通常用于小尺寸模型[14-16]。在 Zagoruyko[12]的研究中，更宽网络模型往往能够获取更细粒度的特征，也更容易训练。然而，极宽但较浅的网络往往难以捕捉更高层次的特征。

使用更高分辨率的输入图像，卷积神经网络可以获取更细粒度的特征。从早期 VGG[21]、AlexNet[22]等网络的 224 输入大小开始，现代卷积神经网络倾向于使用 299×299 [10]或 331×331[17]的输入大小来获得更好的精度。Huang[18]等通过输入 480×480 的大小实现了最优的网络分类精度，分辨率为 480×480。而更高的分辨率，如 600×600，也被广泛应用于目标检测网络中[19-20]。

平衡网络深度、宽度和分辨率 3 个维度获得更好的准确率和效率，EfficientNet 使用复合系数 ϕ 对 3 个维度进行缩放，具体如下：

$$d=\alpha^{\phi}, w=\beta^{\phi}, r=\gamma^{\phi} \tag{7}$$

$$\text{s.t. } \alpha\cdot\beta^2\cdot\gamma^2\approx 2 \tag{8}$$

$$\alpha\geqslant 1,\beta\geqslant 1,\gamma\geqslant 1 \tag{9}$$

式中，α,β,γ 为对应 d,w,r 运行维度的资源分配参数；ϕ 为资源控制系数；在满足约束条件的情况下，通过神经架构搜索对各参数进行优化调整；在参数量和运算量不增加的基础上，EfficientNet 模型通过不断调整网络深度、网络宽度和图像分辨率的系数达到最优的精度。

网络缩放实现了最优的输入分辨率，网络深度以及网络宽度的组合，能够在减少网络参数量的同时保证网络的分类精度。而有效的网络结构也能够减少网络的参数量[23]。最近提出的跨阶段局部网络（CSPNet）[24]在减少计算量的同时可以实现更丰富的梯度组合。CSPNet 是通过将基础层的特征图划分为两部分，然后通过提出的跨阶段层次结构将其合并来实现的。其主要思想是通过分割梯度流，使梯度流通过不同的网络路径传播。现有的 CNN 通过轻量化后，准确率大大下降，而 CSPNet 能够加强 CNN 的学习能力，使其在轻量化的同时保持足够的准确性。同时，过高的计算瓶颈会导致更多的循环来完成推理过程，或者一些算术单元会经常闲置。而 CSPNet 能够将计算量均匀的分布在 CNN 的每一层，从而有效的提升每个计算单元的利用率，减少不必要的计算消耗。

3 实 验

3.1 实验流程

本文所采用的实验流程如图 4 所示，从三调数据筛选出部分样例照片，并对照片进行人工标注，然后根据样例照片制作数据集，将数据集按照 7∶3 的比例

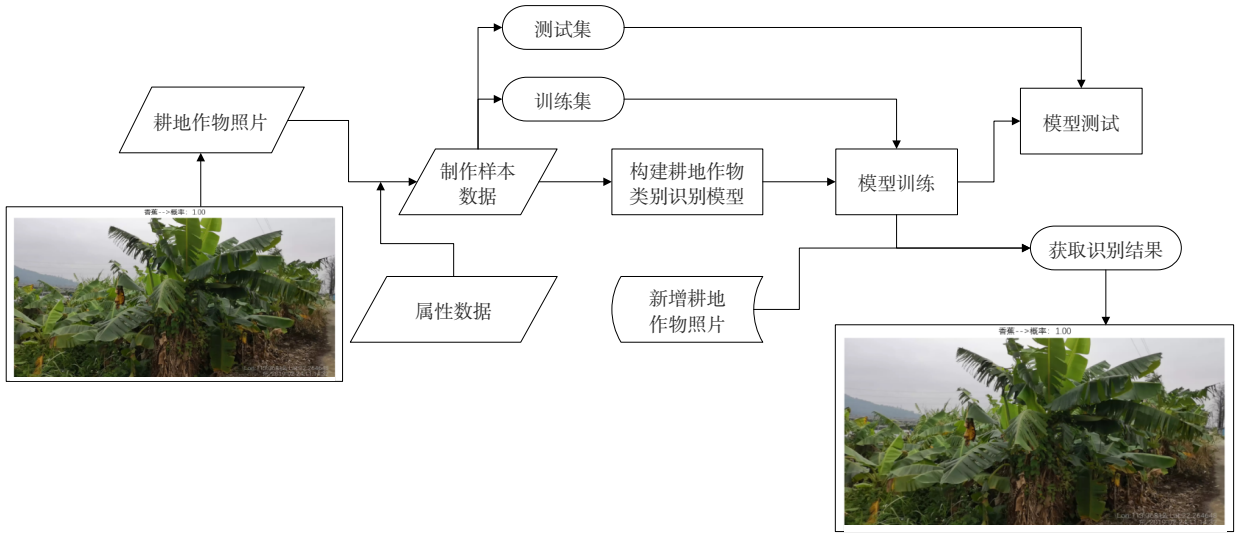


图 4 实验流程图

划分训练集与测试集。训练过程中对数据进行随机增强，包括水平翻转、镜像翻转、上下翻转、高斯噪声模糊、0~15°的随机旋转，然后将数据采样至512×512的大小并归一化输入进网络进行训练。训练结束后对模型进行测试，获取模型的测试精度，检验模型的鲁棒性。

3.2　参数设置

本文训练所使用的框架为pytorch，训练周期为100，优化器选择SGD，初始学习率为0.01，衰减系数为0.000 1，学习率采用WarmUp学习率调整策略，后续学习率每25个周期衰减10倍，训练过程中采用NVIDIA GTX 2080 ti GPU进行加速。

由于训练数据存在不均衡的现象，本文的损失函数并未使用普通的交叉熵损失函数，而是对交叉熵损失函数进行了加权，加权的交叉熵损失函数公式如下：

$$L(y,\hat{y})=-\sum_{i=1}^{N}w^{(i)}y^{(i)}\lg\hat{y}^{(i)}+\left(1-y^{(i)}\right)\lg\left(1-\hat{y}^{(i)}\right) \quad (10)$$

式中，y为样本真实标签；$\hat{y}$为样本通过网络模型得到的预测标签；N为样本数；$w^{(i)}$为样本权重；$y^{(i)}$为第i个真实样本；$\hat{y}^{(i)}$为第i个预测样本。

3.3　结果与分析

在对网络模型进行训练后，加载模型的训练权重，并利用预先划分的测试集对模型进行测试，测试的结果如图5展示的混淆矩阵所示。可以看到各个类别的分类精度均达到了非常高的精度，整体误分情况较少，测试集的F1-score达到了0.988 3，这说明本文提出的网络模型能够有效捕获三调数据的场景特征，提升最终的分类精度。

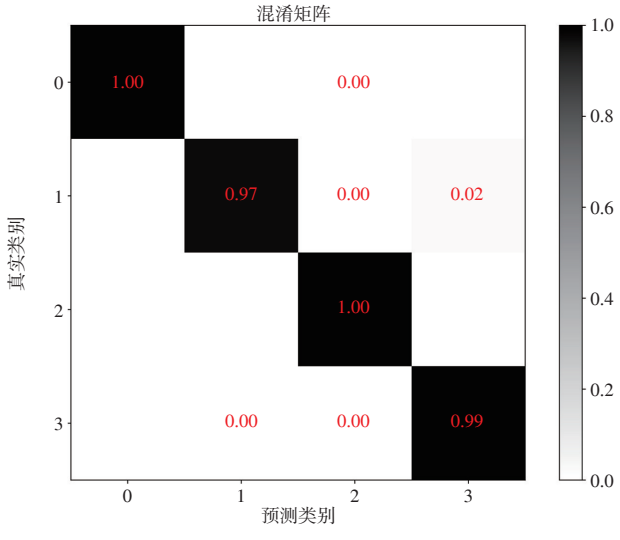


图5　测试集混淆矩阵

为验证本文提出的模型的有效性，本文对比了Inceptionv3[11]，ResNet101[7]，DenseNet169[9]，ResNeXt50[25]和WideResNet50[12]5种模型，在实验参数上保持一致，得到的结果如表2所示。根据结果可以看出，本文的网络模型在精度上均优于其余模型，并且参数量较少，其余模型参数量均在千万级别，本文的方法参数量在百万级别。在计算量方面，以输入512×512大小为基准，统计了各个模型的计算量，其中WideResNet50的计算量最高，达到了119.4 BFLOPs，其余模型也均超过30 BFLOPs，可以看到主流网络模型的计算量均较高，不适合大规模推理预测，耗费时间的同时也需要消耗大量的计算资源。而本文提出的模型计算量仅为15.3 BFLOPs，在减少计算量的同时并没有降低模型的精度，由此可见，本文提出的模型无论在速度和精度方面均有所提高。

表2　EMCNet与主流网络模型的对比结果

方法	召回率	精确度	F1-score	参数量/m	计算量/BFLOPs
Inceptionv3	0.939 0	0.942 5	0.940 6	21.8	35.1
ResNet101	0.919 4	0.939 1	0.928 5	42.5	81.8
DenseNet169	0.941 1	0.951 7	0.946 0	12.5	35.5
ResNeXt50	0.949 8	0.959 2	0.954 3	23.0	44.5
WideResNet50	0.949 2	0.936 9	0.942 3	66.8	119.4
EMCNet	0.991 4	0.985 4	0.988 3	6.9	15.3

4　结　语

本文提出了一种轻量化的神经网络模型，该模型利用了最新的网络缩放技巧，综合精度与效率，并对特征进行跨阶段融合，能够对特征进行有效利用。与目前主流的神经网络模型进行对比，本文提出的模型不仅在速度上具有较大的优势，且在精度上也表现更优，在实际的场景应用中能够达到较高的分类精度，可以实现利用该模型对三调数据进行自动筛选，有效提高筛选效率，减少人工参与，具有重要的推广意义。

参考文献

[1]　自然资源部. 国务院第三次全国国土调查领导小组办公室关于印发《第三次全国国土调查实施方案》的通知[EB/OL]. (2019-04-12) [2022-02-01]:http://www. gov. cn/xinwen/content_5381961. htm

[2]　Hinton G E, Osindero S, Teh Y W. A Fast Learning Algorithm for Deep Belief Nets[J]. Neural Computation, 2006, 18(7):1 527-1 554

[3] 王斌,范冬林.深度学习在遥感影像分类与识别中的研究进展综述[J].测绘通报,2019(2):99-102

[4] 徐世武,曾珏,张诗慧,等.一种深度卷积神经网络土地利用场景照片的分类方法[J].测绘通报,2020(2) :24- 28

[5] 周家奇.基于CNN的图像分类及在电商图片分类中的应用研究[D].武汉:武汉理工大学,2019

[6] 苏婷.基于照片识别分类的智能推荐研究[D].厦门:厦门大学,2017

[7] Tan M, Le Q. Efficientnet:Rethinking Model Scaling for Convolutional Neural Networks[C]//International Conference on Machine Learning,PMLR,2019:6 105-6 114

[8] 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

[9] Huang G, Liu Z, Vander M L, et al. Densely Connected Convolutional Networks[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition,2017

[10] Szegedy C,Liu W,Jia Y,et al. Going Deeper with Convolutions[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition,2015

[11] Szegedy C, Vanhoucke V, Ioffe S, et al. Rethinking the Inception Architecture for Computer Vision[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition,2016:2 818-2 826

[12] Zagoruyko S, Komodakis N. Wide Residual Networks[J]. ArXiv Preprint ArXiv,2016(12):1 605-1 607

[13] Ioffe S , Szegedy C. Batch Normalization:Accelerating Deep Network Training by Reducing Internal Covariate Shift[J]. JMLR. org,2015,37(8):448-456

[14] Howard A G, Zhu M, Chen B, et al. Mobilenets:Efficient Convolutional Neural Networks for Mobile Vision Applications[J]. ArXiv Preprint ArXiv,2017(10):1 704-1 708

[15] Sandler M, Howard A, Zhu M, et al. Mobilenetv2:Inverted Residuals and Linear Bottlenecks[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition,2018:4 510-4 520

[16] Tan M, Chen B, Pang R, et al. Mnasnet: Platform-aware Neural Architecture Search for Mobile[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition,2019:2 820-2 828

[17] Zoph B, Vasudevan V, Shlens J, et al. Learning Transferable Architectures for Scalable Image Recognition[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition,2018:8 697-8 710

[18] Huang Y, Cheng Y, Bapna A, et al. Gpipe:Efficient Training of Giant Neural Networks Using Pipeline Parallelism[J]. Advances in Neural Information Processing Systems,2019,32(9):425-429

[19] He K, Gkioxari G, Dollár P, et al. Maskr-cnn[C]//Proceedings of the IEEE International Conference on Computer Vision,2017:2 961-2 969

[20] 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, 2017:2 117-2 125

[21] Simonyan K, Zisserman A. Very Deep Convolutional Net-Works for Large-scale Image Recognition[J]. ArXiv Preprint,2014,35(11):1 409-1 415

[22] Krizhevsky A, Sutskever I, Hinton G E. Imagenet Classification with Deep Convolutional Neural Networks[J]. Advances in Neural Information Processing Systems, 2012, 58(12):32-39

[23] Hu J,Shen L,Sun G. Squeeze-and-excitation Networks[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition,2018:7 132-7 141

[24] Wang C Y,Liao H Y M,Wu Y H,et al. CSPNet: A New Backbone That Can Enhance Learning Capability of CNN[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops,2020:390-391

[25] Xie S,Girshick R,Dollár P,et al. Aggregated Residual Transformations for Deep Neural Networks[C]//Proceedings of the IEEE Conference on Computer Vision and Qattern Recognition,2017:1 492-1 500

第一作者简介：张俊鑫（1984—），高级工程师，主要从事测绘遥感方法、国土调查监测、多源地理空间数据处理等研究。