基于二值语义分割网络的遥感建筑物检测

朱天佑[1,2,3]，董峰[1,2]，龚惠兴[1,2*]

[1]中国科学院红外探测与成像技术重点实验室，上海 200083；

[2]中国科学院上海技术物理研究所，上海 200083；

[3]中国科学院大学，北京 100049

摘要 针对遥感建筑物实时检测中深度卷积网络资源消耗大和硬件移植难的问题，提出一种基于二值与浮点数混用方法的语义分割网络 MBU-Net。通过对 FU-Net 网络全局权重进行二值化处理来压缩模型大小，并将占少量参数的网络输出层权重替换成浮点型（MBU-Net），解决了全局二值网络（GBU-Net）检测精度差、训练缓慢的问题。在 QuickBird 卫星遥感数据集上进行实验，MBU-Net 的像素准确率为 82.33%，F1 分数（召回率和精确率的调和平均数）为 73.15%；相比于 FU-Net，MBU-Net 在保证检测精度的前提下，模型大小大幅压缩，检测速度提升了 6.29 倍，功耗降为 37.78%，且优于其他同类方法（Deeplab、ENet），对遥感建筑物实时检测具有重要的实际工程价值。

关键词 遥感；卫星图像；建筑物检测；语义分割；二值神经网络

中图分类号 TP391.4 **文献标识码** A **doi**：10.3788/AOS201939.1228002

Remote Sensing Building Detection Based on Binarized Semantic Segmentation

Zhu Tianyou[1,2,3], Dong Feng[1,2], Gong Huixing[1,2*]

[1]*Key Laboratory of Infrared System Detection and Imaging, Chinese Academy of Sciences, Shanghai* 200083, *China*;

[2]*Shanghai Institute of Technical Physics, Chinese Academy of Sciences, Shanghai* 200083, *China*;

[3]*University of Chinese Academy of Sciences, Beijing* 100049, *China*

Abstract To address the problem of high resource consumption and difficulty of hardware transplantation involved in utilizing deep convolutional networks for real-time detection of remote sensing building, a semantic segmentation network based on the mixed method of binary and floating-point parameters, i.e., mixed binary U-shape network (MBU-Net), is proposed. To compress the model size, the weights of a float U-shape network (FU-Net) are binarized. The output layer weights that account for a small number of parameters are replaced by floating-point type parameters to resolve the poor detection accuracy and low training speed in a global binary network. Experiments using the QuickBird satellite remote sensing dataset show that the pixel accuracy of MBU-Net is 82.33% and the harmonic average of the recall rate and accuracy rate (F1 score) is 73.15%. Compared with the FU-Net, the MBU-Net can ensure the detection accuracy. The size of model is greatly compressed, the detection speed is increased by 6.29 times, and the power consumption is reduced to 37.78%, further demonstrating that the MBU-Net is superior to other similar methods (Deeplab and ENet). This finding has important practical engineering value for the real-time detection of remote sensing buildings.

Key words remote sensing; satellite images; building detection; semantic segmentation; binarized neural network

OCIS codes 120.0280; 150.1135; 110.2960; 100.3008

1 引 言

随着遥感成像技术的快速发展，现在可以获得数量可观的高分辨率遥感图像，从而使人们能够更详细地研究地表目标[1]。如今，利用卫星遥感数据进行人工建筑物检测是最具挑战性的任务之一[2]。语义分割，旨在对提取出的图像子区域中的每一个像素点进行分类。遥感图像中的建筑物具有粒度细、分布散的特点，采用语义分割方法能获取更精细的结果。遥感建筑物检测具有许多非常实际的应用，

收稿日期：2019-05-27；**修回日期**：2019-07-08；**录用日期**：2019-08-13

基金项目：十三五装备预研中科院联合基金(61XXA011XXXXX)

* **E-mail**：hxgong@mail.sitp.ac.cn

如土地资源管理、城市规划[3]和计算机制图等[4]。

在遥感图像的检测和分类[5]任务中，构造包含每一类特点的特征向量是非常重要的。在传统的研究工作中，通常通过人工构造包含形态滤波器[6]、纹理[7]、点描述符[8]、梯度方向[9]等信息的特征向量。随着深度学习的发展[10]，能够自动学习[11]并提取深层次[12]特征向量的卷积神经网络(CNN)开始被大规模应用于检测[13]任务。基于 CNN 的语义分割网络成为了研究热点，其中 FCN (Fully Convolutional Networks)[14]是第一种采用全卷积网络的语义分割方法。后来，基于 FCN 的方法在图像语义分割方面取得了重要进展。

在图像深度学习领域，神经网络方法的研究工作主要集中在两方面。一方面是采用越来越复杂且强大的模型网络来改善检测识别的效果：1)在分割方法中使用更深的网络层[15]、更丰富的网络结构，如 Deep Residual (深层残差网络)[16]结构、Inception (网中网)[17]结构，之后又涌现了一批代表性的新型网络，如 U-Net(U shape Network)[18]和 Deeplab (Deeper Network with Atrous Convolution)[19]等；2)将模拟人脑的注意力模型(self-attention 机制)应用到语义分割中，扩大感受野，如 DANet (Dual Attention Network for Scene Segmentation)[20]和 CCNet (Criss-Cross Attention for Semantic Segmentation)[21]等。

另一方面是针对复杂模型带来的存储空间过大以及计算资源消耗大等问题进行模型压缩研究，以提升其硬件的实用性，集中在加速网络结构设计、模型裁剪、权值量化三个方向。使用分组卷积(ShuffleNet[22]、MobileNet[23])、分解卷积(ErfNet[24]、ENet[25])等巧妙的网络结构可以减少参数量，但实现起来较为复杂；模型裁剪需要确定剪枝率和剪枝阈值，重新定义储存编码的数据结构[26]。相比而言，二值化[27]无需更改网络构架和存储结构，更为简便快捷，并且压缩效果明显。

因此，为提高遥感图像建筑物检测任务中语义分割网络的硬件实用性(降低存储和功耗)，本文提出一种混合类型权重的语义分割网络——MBU-Net。首先，将常用语义分割网络 FU-Net(浮点型 U 形框架)与图像分类领域中网络压缩方法——二值化结合起来，提出 GBU-Net (Global Binary U-shape Network)；然后，针对 GBU-Net 带来的精度损失过多、训练较慢的问题，将网络输出层权重替换成浮点型，改进成混合二值网络，即 MBU-Net；最后，在对快鸟(QuickBird)卫星获取的真实图片数据集进行三种网络对比实验的基础上，对结合性能较好的其他同类网络(Deeplab[19]、ENet[25])的分割结果进行分析。结果表明，本文方法可有效识别地上建筑物，且能保证检测精度，在低功耗硬件平台上具有一定的实际应用价值。

2　基本原理

2.1　二值神经网络系统抗过拟合性质分析

通常，神经网络可以模拟任何非线性函数，在训练数据不充分、不广泛、覆盖范围有限时，难免也会对噪声进行拟合，出现过拟合现象，导致训练出来的网络模型只在训练数据上表现良好，对训练集外的测试数据不起作用。为防止过拟合，提高网络的泛化能力，可在原始模型中引入额外信息，这种方法称为正则化方法。

常用的正则化方法有很多，如 L1 regularization、L2 regularization、Dropout (随机丢弃神经元)[28]、DropConnect (随机取消连接)[29]、数据集扩增。此外，在计算参数梯度时，将噪声添加到权重和激活中的方法可以看作是一种正则化方式。例如，对权重参数 w 来说，二值化后的权重 $\tilde{w}$ 可以写成

$$\tilde{w}=w+n_w, \tag{1}$$

式中：n_w 为二值化引入的权重噪声。权重的二值化带来的噪声比例 γ_w(WBNR)为

$$10\log_{10}(\gamma_w)=10\log_{10}\frac{E(w^2)}{E(n_w^2)}, \tag{2}$$

式中：$E(\cdot)$为能量。引入噪声后的激活 $\tilde{a}$ 同样可以表示为

$$\tilde{a}=a+n_a, \tag{3}$$

式中：a 为神经元的激活值；n_a 为激活的二值量化噪声。

在一个多层神经网络的前向传播中，$l+1$ 层的第 i 个激活值 $a_i^{(l+1)}$ 为

$$\begin{aligned} a_i^{(l+1)}&=\sum_{j=1}^{N} w_{i,j}^{(l+1)}a_j^{(l)}+b_i^{(l+1)}\\ a_i^{(l+1)}&=\sum_{j=1}^{N} w_{i,j}^{(l+1)}a_j^{(l)}+b_i^{(l+1)}, \end{aligned} \tag{4}$$

式中：j 为神经元序号；N 为神经元个数；$w_{i,j}^{(l+1)}$ 为第 l 层第 j 个激活向第 $l+1$ 层第 i 个激活传播时的权重系数；$a_j^{(l)}$ 为第 l 层的第 j 个激活；$b_i^{(l+1)}$ 为偏置。考虑到前向传播中的权重 w 要乘以前一层的激活 a，如(4)式所示，对较小的 n_w 和 n_a 来说，其乘积 $n_w\cdot n_a$ 可以忽略不计，则带噪声的传播可以

表示为

$$\tilde{w}\cdot\tilde{a}=(w+n_w)\cdot(a+n_a)=$$
$$w\cdot a+w\cdot n_a+n_w\cdot a+n_w\cdot n_a\cong \quad (5)$$
$$w\cdot a+w\cdot n_a+n_w\cdot a,$$

$$\frac{1}{\gamma_{w\cdot a}}=\frac{1}{\gamma_w}+\frac{1}{\gamma_a}, \quad (6)$$

$$\frac{1}{\gamma_{w_{i,j}^{(l+1)}a_j^{(l)}}}=\frac{1}{\gamma_{w_{i,j}^{(l+1)}}}+\frac{1}{\gamma_{a_j^{(l)}}}=\frac{1}{\gamma_{w^{(l+1)}}}+\frac{1}{\gamma_{a^{(l)}}}, \quad (7)$$

$$\frac{1}{\gamma_{\text{output}}}=\frac{1}{\gamma_{a^{(0)}}}+\frac{1}{\gamma_{w^{(1)}}}+\frac{1}{\gamma_{a^{(1)}}}+\cdots+\frac{1}{\gamma_{w^{(L)}}}+\frac{1}{\gamma_{a^{(L)}}}, \quad (8)$$

式中：γ_a 为单层激活的 WBNR；$\gamma_{w\cdot a}$ 为单层传播后的 WBNR；$\gamma_{w_{i,j}^{(l+1)}a_j^{(l)}}$ 为第 l 层向第 $l+1$ 层传播后的 WBNR；$\gamma_{w_{i,j}^{(l+1)}}$ 为第 l 层第 j 个激活向第 $l+1$ 层第 i 个激活传播权重的 WBNR；$\gamma_{a_j^{(l)}}$ 为第 l 层第 j 个激活的 WBNR；$\gamma_{w^{(l+1)}}$ 为第 $l+1$ 层权值的 WBNR；$\gamma_{a^{(l)}}$ 为第 l 层激活的 WBNR；γ_{output} 为整个系统输出的 WBNR；$\gamma_{w^{(L)}}$ 为第 L 层权值的 WBNR；$\gamma_{a^{(L)}}$ 为第 L 层激活的 WBNR；L 为网络层数。

$w\cdot a$ 的噪声比例 $\gamma_{w\cdot a}$ 满足(6)式，这是线性系统的特点。对二值化噪声单独引入权重和激活，等同于系统总噪声直接相加。因此在网络层的前向传播中，结合(4)式，可将 $\gamma_{w_{i,j}^{(l+1)}a_j^{(l)}}$ 表示为(7)式。对整个系统而言，γ_{output} 可表示为(8)式。

从(8)式中可见，网络层与层之间的噪声相互独立、互不影响，并且被叠加到整体噪声之中，共同作用于整个系统。本文提出的 MBU-Net 网络系统混用二值与浮点数权重，其最后一层输出层未经过二值化，仍保持高精度浮点型，因此最后一层的 $\gamma_{w^{(L)}}$、$\gamma_{a^{(L)}}$ 接近于无穷大，取倒数后，其值可忽略不计。结合(8)式可知，最后的浮点输出层并不影响前面特征提取层二值化引入的噪声以及系统整体噪声，二值化引入额外的噪声足够克服过拟合。理论分析表明，本文提出的基于二值与浮点数混用的 MBU-Net 网络具有正则化作用，能够克服过拟合。

2.2　二值化神经网络的训练步骤

网络模型的前向推理使用过程主要包括卷积和矩阵乘法操作，主要运算是乘法累加。当权重被约束为+1或−1时，很多乘法累加运算可由简单的加法(或减法)代替。这在硬件实现上有巨大的好处，因为定点加法器比乘法累加器的运算速度更快，所需资源更少[30]。

对传统的神经网络进行训练和参数更新一般分为三个步骤，如图 1 所示，即前向传播、反向传播以及参数更新，如此循环直至达到终止条件后结束。

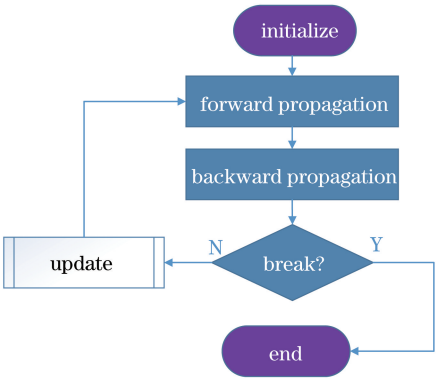

图 1　传统的神经网络训练流程图

Fig. 1　Flow chart for training of traditional neural network

二值化神经网络算法训练流程如图 2 所示，其中 w 为权重参数，w_t 为浮点型权重，t 为迭代次数，w_b 为二值化后的权重，η 为学习率，binarize(w)表示二值化，clip(w)表示权重裁剪，C 为二元交叉熵损失值。与传统网络相比，二值化神经网络在前向传播开始前，先将浮点型权重参数离散成二值，进行前向传播和反向传播，而参数更新仍然基于浮点型权重。因此使用二值化的关键是只在前向和后向传播期间对权重进行二值化，而不是在参数更新期间进行二值化。高精确权重始终需要保存，梯度仍以浮点型变量形式累积，在优化器更新参数期间使用。

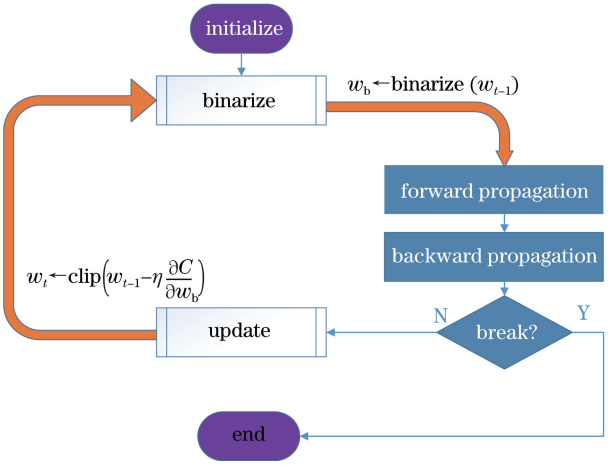

图 2　二值化神经网络的训练流程图

Fig. 2　Flow chart for training of binarized neural network

2.3　二值化方式

训练网络时，权重被约束为+1或−1，这两个值在硬件实现上非常有利。为了将实际变量转换为二值，需要使用二值化函数。常用的二值化方式包括确定式与随机式。确定式的二值化函数表达式为

$$x_B=\text{Sign}(x)=\begin{cases}+1, & x\geqslant 0\\ -1, & \text{otherwise}\end{cases}, \quad (9)$$

式中：Sign(•)为符号函数；x_B 为二值化后的变量；x 表示输入，范围为$(-\infty,+\infty)$。确定式二值化很容易直接实施，在实践中表现良好。还有一种更精细和更均衡的替代方案是随机二值化，表达式为

$$\sigma(x)=\mathrm{clip}\left(\frac{x+1}{2},0,1\right)=\max\left[0,\min\left(1,\frac{x+1}{2}\right)\right],\tag{10}$$

$$x_B=\begin{cases}+1, & \text{with probability } p=\sigma(x)\\ -1, & \text{with probability } 1-p\end{cases},\tag{11}$$

式中：clip(•)为截断函数；p 表示概率。在区间$[0,1]$上，$\sigma(x)$是 hard Sigmoid 函数，呈分段线性，并且有对应的上下边界约束。与 soft Sigmoid 函数相比，hard Sigmoid 的计算成本要低得多(在软件和专用硬件实现中)，并且在实际应用中取得了良好的效果。随机二值化比确定式二值化更具吸引力，但实现起来较为复杂，需要硬件在量化时生成随机位，资源开销较大。因此，本文主要使用确定性二值化函数，即符号函数。

2.4　损失函数选择

在网络学习训练过程中，为了在给定数据上评估预测结果与真实结果的偏离程度，需要选择误差函数(通常称为损失函数)。该函数可用于估计模型的损失，以便更新模型权重，减少下一次评估的损失。

神经网络模型学习从输入到输出的映射，而且损失函数的选择必须与特定建模任务的框架相匹配。因此，损失函数可以大致分为两类：分类损失和回归损失。输出层的激活函数必须适合于所选择的损失函数。对于二分类和多分类任务来说，对应神经网络输出层的激活函数常选用 Sigmoid 和 Softmax，匹配的损失函数是二元交叉熵(BCE)和多元交叉熵。

针对输入遥感卫星图片，本文需要输出像素级别的对应类别，这属于两类判别问题。因而本文的网络训练采用二元交叉熵作为损失函数，网络输出层采用 Sigmoid 激活函数。

$$f(x)=\frac{1}{1+\exp(-x)},\tag{12}$$

$$C=-[y\log_2(p)+(1-y)\log_2(1-p)],\tag{13}$$

式中：y 表示分类目标值，在集合$\{0,1\}$中取值；p 为网络输出的预测概率，其值在$[0,1]$区间；C 为计算得到的二元交叉熵损失值，用以衡量真实和预测概率分布之间的平均差异程度，差异越大，则计算的交叉熵值越大。

$f(x)$为 Sigmoid 函数，可将输入映射到$[0,1]$区间。(12)式是常见的 S 型函数，具有非线性和导数计算简单等特点[31]。二元交叉熵是适用于二分类问题的默认损失函数，可以计算得到一个损失分数值，如(13)式所示。

3　网络结构

为对比二值化前后检测精度的变化，本文网络层采用相同的框架结构，而参数分别采用二值型、浮点型以及二值加浮点的混合型，搭建了三种不同网络。网络特征提取层包括上采样层和下采样层，均采用 3×3 卷积核，而输出层采用 1×1 卷积核。三种网络特征提取层和输出层权重参数类型(即是否二值)对比示意图如图 3 所示。

第一个网络基于常用的语义分割网络结构 U-Net，所有卷积层的权重均采用 32 位浮点数，称为 FU-Net，如 4(a)所示；第二个网络结构如图 4(b)所示，它是在 FU-Net 基础上将所有卷积层的权重二值化，简称 GBU-Net；第三个网络如图 4(c)所示，

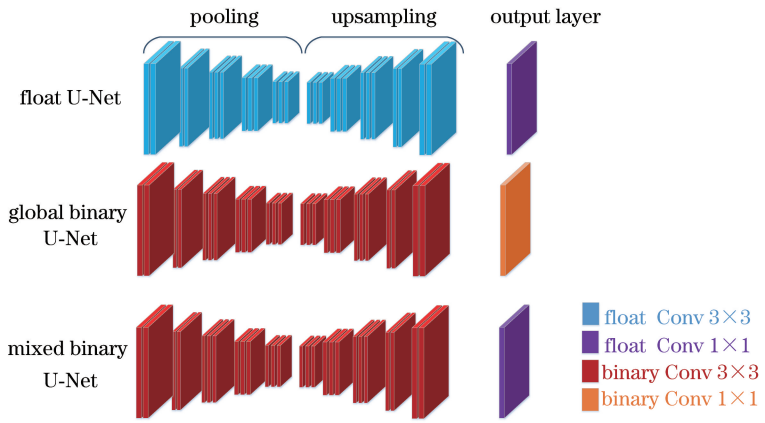

图 3　三种网络对比示意图

Fig. 3　Comparison of three neural networks

与 GBU-Net 相比，第三个网络将最后一层输出层的权重替换回 32 位浮点数，混合了两种不同类型的权值，简称 MBU-Net。三种网络中每层卷积核数、大小等详细参数设置如图 4 所示。

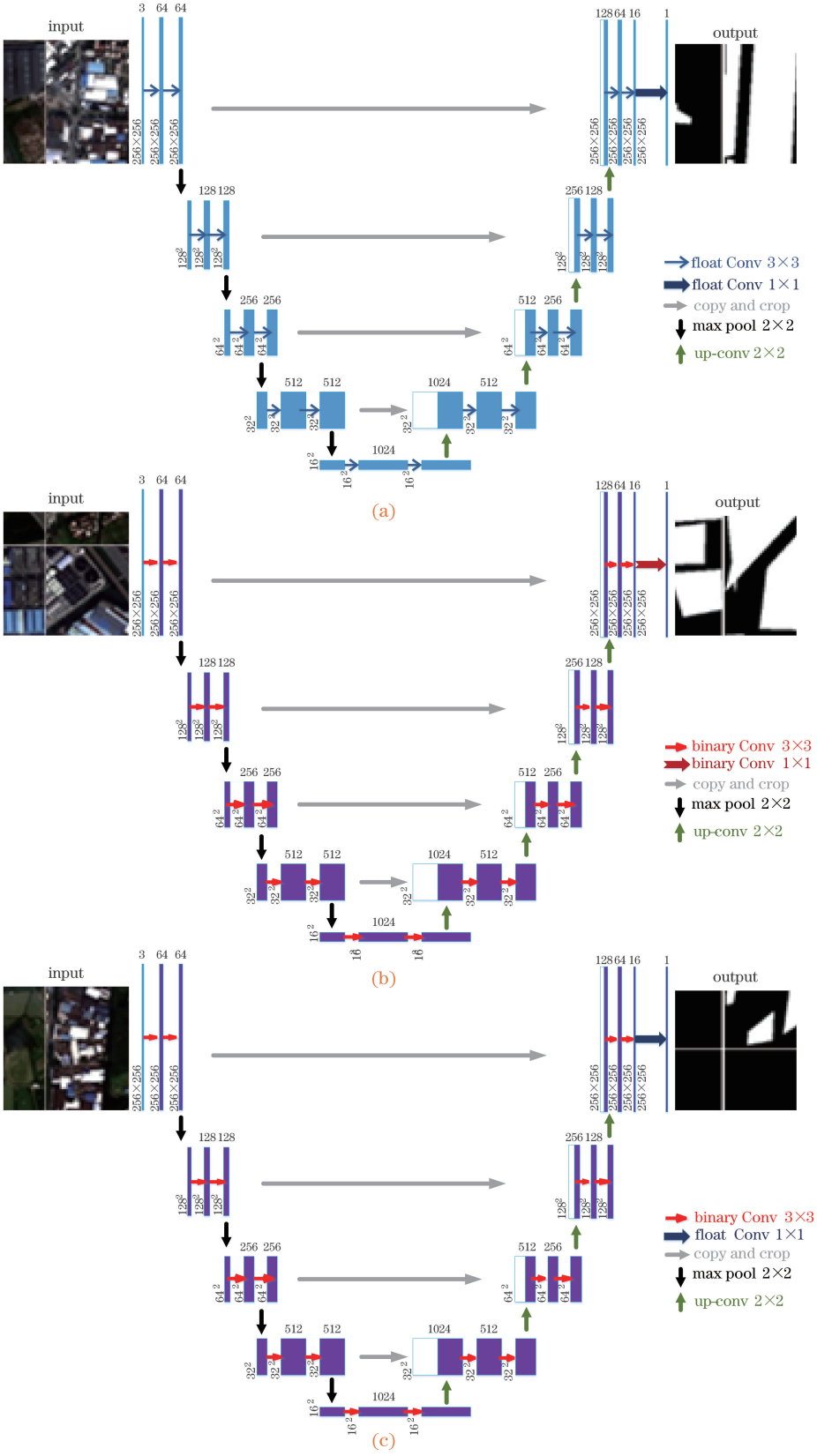

图 4　三种网络结构和详细参数。(a) FU-Net；(b) GBU-Net；(c) MBU-Net

Fig. 4　Details and architectures of three networks. (a) FU-Net; (b) GBU-Net; (c) MBU-Net

三种网络的结构框架和每层卷积核数、大小设置均一致，因此三种网络的总参数量相同。特征提取层的层数较多，且采用 3×3 大小的卷积核，根据深度计算每层的卷积参数，累加起来共 25317760 个参数，计算方式为 3×3×(3×64＋64×64＋64×128＋…＋128×64＋64×16)。输出层采用 1×1 大小的卷积核，单个卷积核的大小只有特征提取层的 1/9，且只有一层，参数量为 1×1×16×1＝16。可以看出，输出层参数远远少于特征提取层，因此网络总参数和所需存储空间主要由特征提取层决定。

4　实验过程及结果分析

4.1　数据集准备

数据集来自于 QuickBird 卫星从 450 km 处太空拍摄的地球表面卫星图片(覆盖广东省部分地区数百平方千米的土地)。QuickBird 卫星是世界上第一颗提供亚米级分辨率影像的商业卫星，它可在全球范围内实现大面积覆盖，更新频率快，93 min 即可环绕地球一周。

对获取的卫星图片进行手工标注，制作数据集。将图像像素级别的对应目标分为两类，一类是建筑物，一类是背景(非建筑物部分)。图 5(a)是 QuickBird 卫星获取的一张 R、G、B 三个通道的彩色图像，每个像素的对应类别如图 5(b)所示(白色代表建筑物，黑色代表背景)。通过旋转、翻转、拉伸、加噪、随机裁剪、平移变换等图像增强手段进行数据集扩增，最终获取 12000 张卫星图片。将其中的 10000 张作为训练集，用以训练不同的语义分割网络；将其余的 2000 张作为测试集，用以对比不同网络的效果差异。数据集部分样例如图 6 所示。

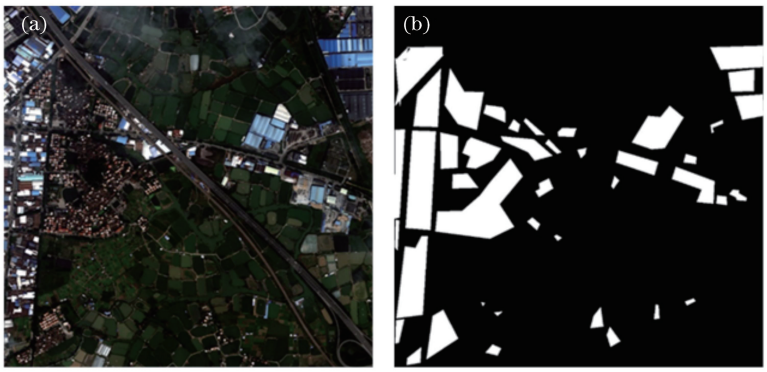

图 5　遥感图片上的建筑物及其对应标注。(a) 卫星遥感图片；(b) 标注图片

Fig. 5　Buildings in satellite remote sensing image and their corresponding labels. (a) Satellite remote sensing image; (b) labeled image

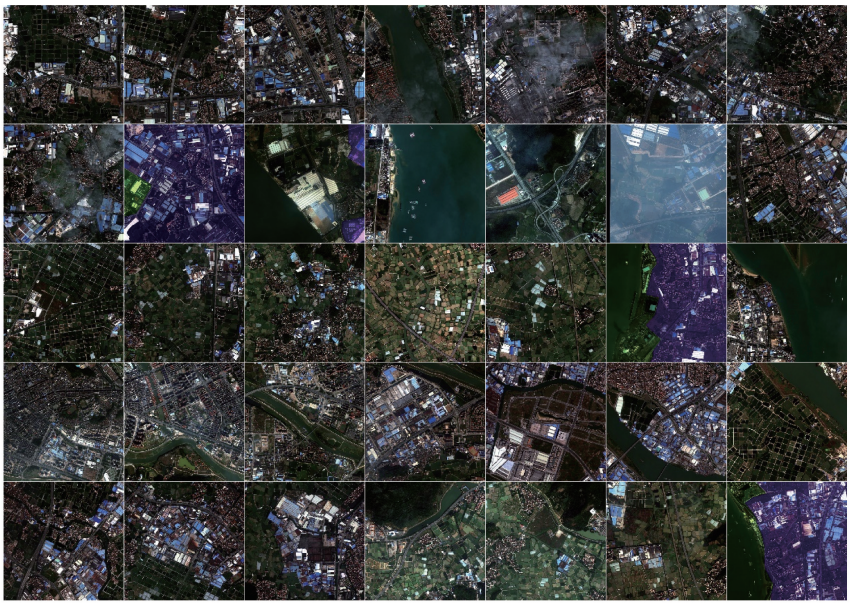

图 6　部分训练数据图片

Fig. 6　Part samples of satellite remote sensing data for training

4.2　评价指标

选取常用的语义分割性能评估指标(像素准确率 β_{PA}、精确率 P、召回率 R、F1 分数 $F_{1\text{-score}}$)进行评价,这些评估指标的计算公式分别为

$$P=\frac{N_{TP}}{N_{TP}+N_{FP}},\tag{14}$$

$$R=\frac{N_{TP}}{N_{TP}+N_{FN}},\tag{15}$$

$$\beta_{PA}=\frac{N_{TP}+N_{TN}}{N_{TP}+N_{TN}+N_{FP}+N_{FN}},\tag{16}$$

$$F_{1\text{-score}}=\left(\frac{R^{-1}+P^{-1}}{2}\right)^{-1}=2\cdot\frac{P\cdot R}{P+R},\tag{17}$$

式中:N_{TP}为正确识别出的建筑物像素的个数;N_{TN}为正确识别出的背景像素的个数;N_{FP}为将背景误识别为建筑物像素的个数;N_{FN}为将建筑物误识别成背景像素的个数。像素准确率 β_{PA}表示识别正确的像素(包括建筑物和背景)占所有像素的比例;精确率 P 表示所有被分类为建筑物像素中的真实建筑物像素所占的比例;召回率 R 表示在所有的真实建筑物像素中被正确分类的比例。通常希望同时获得较高的精确率和召回率,但这两者在某些情况下是互相矛盾的,因此需要综合考虑,常见的方法是取两者的调和平均数(称为 F1 分数),即 $F_{1\text{-score}}$。

4.3　中间结果的可视化

以 MBU-Net 为例,对网络权重和网络中间层激活的特征图进行可视化,展现不同卷积核的结果。MBU-Net 网络层详细的输入输出参数如表 1 所示。从表 1 中可以看出,除最后一层输出层采用浮点型卷积(Conv2D)和 1×1 大小的卷积核,其余特征提取层均采用二值卷积(BinaryConv2D)和 3×3 大小的卷积核。

表 1　MBU-Net 网络层详细的输入输出参数表

Table 1　Input and output parameters of MBU-Net layer

Convolution type	Input	Kernel size	Stride	Padding	Output
BinaryConv2D	256×256×3	3×3×64	1	Same	256×256×64
BinaryConv2D	256×256×64	3×3×64	1	Same	256×256×64
Pool	256×256×64	2×2	1		128×128×64
BinaryConv2D	128×128×64	3×3×128	1	Same	128×128×128
BinaryConv2D	128×128×128	3×3×128	1	Same	128×128×128
Pool	128×128×128	2×2	1		64×64×128
BinaryConv2D	64×64×128	3×3×256	1	Same	64×64×256
BinaryConv2D	64×64×256	3×3×256	1	Same	64×64×256
Pool	64×64×256	2×2			32×32×256
BinaryConv2D	32×32×256	3×3×512	1	Same	32×32×512
BinaryConv2D	32×32×512	3×3×512	1	Same	32×32×512
Pool	32×32×512	2×2			16×16×512
⋮	⋮	⋮	⋮	⋮	⋮
BinaryConv2D	256×256×128	3×3×64	1	Same	256×256×64
BinaryConv2D	256×256×64	3×3×16	1	Same	256×256×16
Conv2D	256×256×16	1×1×1	1	Same	256×256×1

以第三层卷积核组参数为例,维度为 3×3×64×128,将卷积核组平铺变换成二维,得到的结果如图 7 所示。因权重被约束成二值,故图 7 中只有黑白两种状态。

作为特征图可视化的输入如图 8 所示,经网络卷积核组前向传播后,选取第三个卷积层的 36 张特征图拼接到一起,得到的结果如图 9 所示。可以看出,不同卷积核组学到的特征图的侧重点有所不同,但大多对目标像素的响应明显有别于背景响应,最终所有卷积核组共同作用于最终的输出结果。

4.4　结果分析

实验所采用深度学习框架为 Keras,基于 Tensorflow 后端,系统为 64 位 Ubuntu16.04 LTS。网络训练与测试硬件平台的 GPU 采用英伟达 NVIDIA GeForce GTX 1080Ti,内存128G,图像读取保存及其他处理基于开源软件 OpenCV4.1.0 实现。除基于 U-Net 架构的三种网络(FU-Net、GBU-Net、MBU-Net)外,实验还在其他类别架构的网络

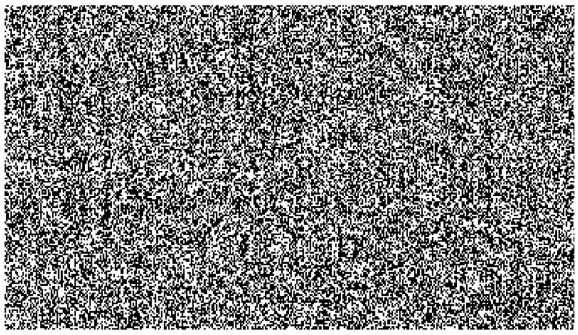

图 7　MBU-Net 卷积核组参数可视化结果

Fig. 7　Visualization result of convolution kernel group parameters of MBU-Net

图 8　MBU-Net 中间层可视化输入图

Fig. 8　Input image for visualizing of MBU-Net interlayer

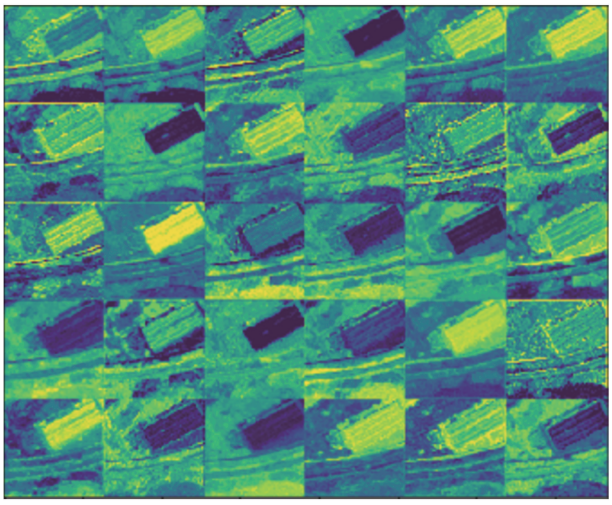

图 9　第三层 30 张特征图的可视化结果

Fig. 9　Visualization results of 30 feature maps in third layer

上进行，如检测效果较好的 Deeplab（基于空洞卷积和多分辨率结构）以及运行较快的 ENet（基于分解卷积结构）。

训练的损失函数为二元交叉熵（BCE），训练过程中随着迭代次数增加，五种网络的训练损失、训练准确率变化曲线分别如图 10 和图 11 所示。

五种网络训练迭代完成后，最终的训练损失和

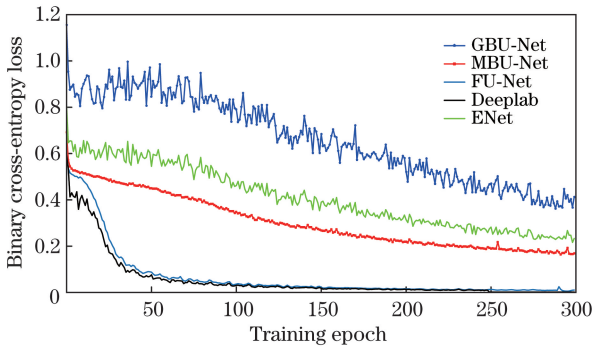

图 10　五种网络的训练损失变化曲线

Fig. 10　Trends of loss of five networks

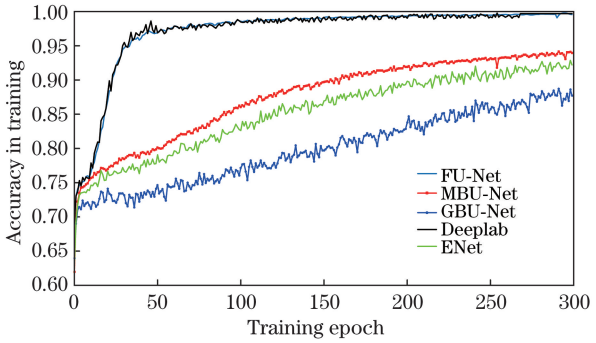

图 11　五种网络的训练准确率变化曲线

Fig. 11　Trends of accuracy of five networks

像素准确率如表 2 所示。结合训练曲线变化图和表 2 可以看出：FU-Net 和 Deeplab 趋势类似，训练较快，损失平稳下降，但容易过拟合；GBU-Net 损失下降缓慢，且振荡反复不稳定，处于欠拟合状态；MBU-Net 较 GBU-Net 损失下降平稳，相对 FU-Net 而言不易过拟合。这表明二值化带给权重的噪声具有正则化作用，可抗过拟合，与 2.1 节的理论分析相符合。

表 2　五种网络的训练损失值和准确率

Table 2　Results of training loss and accuracy of five networks

Algorithm	Training loss	Training accuracy /%
FU-Net	0.01	99.56
GBU-Net	0.41	87.73
MBU-Net	0.16	93.94
Deeplab	0.01	99.70
ENet	0.23	92.21

加载训练好的五种网络，分别在测试集上对测试图片进行像素级别的预测和分类判别，最终获得的各项指标如表 3 所示。

表 3　五种网络在测试集上的各项测试指标

Table 3　Test indexes of five networks on test set

Algorithm	β_{PA} /%	P /%	R /%	$F_{1\text{-score}}$ /%	Memory /MB	Time /ms
FU-Net	81.88	77.07	66.20	71.22	96.60	357
GBU-Net	74.74	84.41	54.22	66.03	3.02	47
MBU-Net	82.33	82.76	65.54	73.15	3.02	49
Deeplab	82.03	78.63	67.11	72.41	123.47	489
ENet	78.49	85.37	56.19	67.77	13.73	107

在测试集的遥感卫星图片上，对全局所有的卷积层权重都采取二值化的 GBU-Net。相比浮点型的 FU-Net，GBU-Net 的像素准确率下降了 7.14%，F1 分数下降了 5.19%，性能损失较多；而采用混合类型权重的 MBU-Net，F1 分数提高了 1.93%，检测识别效果与 Deeplab 网络相当，存储空间是基于卷积分解结构的 ENet 的 1/4，速度是 ENet 的 2 倍。

二值化将所有乘法操作转化成加法运算，对于普通的硬件平台来说，乘法远比加法操作费时，以 Intel Pentium CPU 为例，普通加法需要 1 个指令周期，而普通乘法则需要 10 或 11 个指令周期，乘法速度近乎是加减法的 1/10。由表 3 可知，MBU-Net 的速度相比二值化前提升了 6.29 倍。

大部分硬件平台浮点数的实现遵循 IEEE 754 标准，即浮点数存储为 32 位(bit)，相当于 4 字节(Byte，B)，而二值化后的参数只需要存储为 1 位，相当于 0.125 字节。因此将二值化引入特征提取层便可有效压缩存储空间。以 Xilinx Vertex-7 FPGA 为例，其最大片上内存为 8.5 MB 且不提供片外内存，不适合用于 FU-Net，但能够运行 MBU-Net。结合表 3 分析可知，本文提出的 MBU-Net 相比二值前所需空间减少了 96.87%，大大提升了硬件的实用性，且优于基于分解卷积加速的 ENet。

在过去的数十年中，能耗成为限制硬件性能的主要因素[32]，因此研究人员投入大量精力研究如何减少神经网络的能量消耗。45 nm 制造工艺技术在 0.9 V工作电压下各项运算操作能量消耗如表 4 所示。同样以 3×3 卷积核为例，32 位浮点数卷积一次需要 9 次乘法和 8 次加法，能量消耗为 40.50 pJ (3.7×9+0.9×8)；二值化后的权重卷积一次需要 17 次加法，只需消耗 15.30 pJ(0.9×17) 能量，相当于浮点型能耗的 37.78%。

使用 MBU-Net 网络对输入测试集的遥感卫星图片进行实验，获取的部分输出预测结果如图 12 所示。对比后可知，针对大部分建筑物，MBU-Net 都能很好地预测识别输出，满足实际使用要求。

表 4　不同类型参数加法和乘法运算操作的能耗对比[32]

Table 4　Comparison of energy consumption of addition and multiplication operations for different types of parameters[32]

Operation	Energy consumption /pJ	
	Multiplication	Addition
8 bit integer	0.20	0.03
32 bit integer	3.10	0.10
16 bit floating point	1.10	0.40
32 bit floating point	3.70	0.90

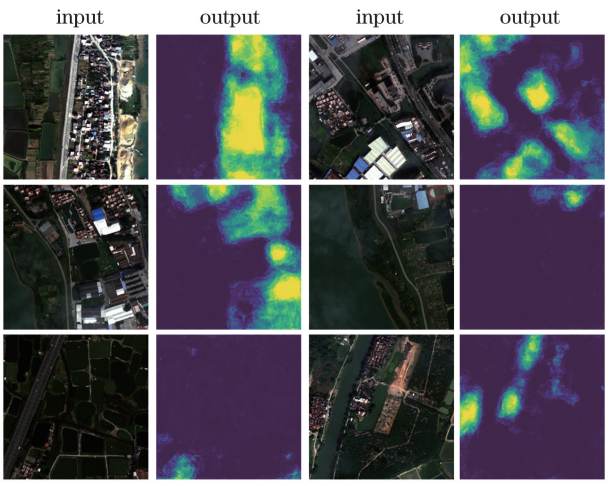

图 12　测试集中的部分输入图片及其对应的输出预测

Fig. 12　Partial input images and output prediction on test set

5　结　论

为解决目前语义分割网络模型太过庞大而无法在存储、功耗有限的硬件平台上稳定高效运行的问题，本文提出了一种只将特征提取层二值化的神经网络模型 MBU-Net。首先将图像分类领域中的模型压缩手段二值化引入到图像分割领域，结合语义分割网络 FU-Net，提出了一种全局二值化的语义

分割网络 GBU-Net，采用该网络进行遥感图片建筑物检测实验。针对全局权重二值化带来性能损失过多(像素准确率下降 7.14%，F1 分数下降 5.19%)、训练收敛较慢或者不收敛的问题，将 GBU-Net 改进成 MBU-Net，该网络采用混用二值和浮点型权重方法，即只将特征提取层二值化，而保留占少量参数和存储的输出层权重为浮点型。实验结果表明，本文方法有效解决了网络训练收敛慢或不收敛的问题，并且性能相比 FU-Net 并未损失(像素准确率为 82.33%，F1 分数为 73.15%)，在效果甚至超过 FU-Net 的情况下(F1 分数提高 1.93%)，存储空间减少了 96.87%，速度提升 6.29 倍，功耗降为 FU-Net 的 37.78%，且优于其他同类方法(Deeplab、ENet)。

本文提出了一种混合二值的语义分割网络 MBU-Net，该网络能对输入图像实时输出每个像素的类别，并可依据目标所占像素的多少估算出目标的面积，同时可解决目标定位和目标尺寸获取的问题。该网络除了可用于国土资源监察(地上建筑物的建、拆、改、扩)等民生领域，还有望在军事领域得到应用。

参　考　文　献

[1] Hu F, Xia G S, Hu J W, *et al*. Transferring deep convolutional neural networks for the scene classification of high-resolution remote sensing imagery[J]. Remote Sensing, 2015, 7(11): 14680-14707.

[2] Vakalopoulou M, Karantzalos K, Komodakis N, *et al*. Building detection in very high resolution multispectral data with deep learning features[C]// 2015 IEEE International Geoscience and Remote Sensing Symposium (IGARSS), July 26-31, 2015, Milan, Italy. New York: IEEE, 2015: 1873-1876.

[3] Zhang X N, Zhong X, Zhu R F, *et al*. Scene classification of remote sensing images based on integrated convolutional neural networks[J]. Acta Optica Sinica, 2018, 38(11): 1128001.
张晓男, 钟兴, 朱瑞飞, 等. 基于集成卷积神经网络的遥感影像场景分类[J]. 光学学报, 2018, 38(11): 1128001.

[4] Shao W, Yang W, Xia G S. Extreme value theory-based calibration for the fusion of multiple features in high-resolution satellite scene classification[J]. International Journal of Remote Sensing, 2013, 34(23): 8588-8602.

[5] Liu F, Lu L X, Huang G W, *et al*. Landform image classification based on discrete cosine transformation and deep network[J]. Acta Optica Sinica, 2018, 38(6): 0620001.
刘芳, 路丽霞, 黄光伟, 等. 基于离散余弦变换和深度网络的地貌图像分类[J]. 光学学报, 2018, 38(6): 0620001.

[6] Lefèvre S, Weber J, Sheeren D. Automatic building extraction in VHR images using advanced morphological operators[C]//2007 Urban Remote Sensing Joint Event, April 11-13, 2007, Paris, France. New York: IEEE, 2007: 9702691.

[7] Moser G, Serpico S B, Benediktsson J A. Land-cover mapping by Markov modeling of spatial-contextual information in very-high-resolution remote sensing images[J]. Proceedings of the IEEE, 2013, 101(3): 631-651.

[8] Wang M, Yuan S G, Pan J. Building detection in high resolution satellite urban image using segmentation, corner detection combined with adaptive windowed Hough Transform[C]//2013 IEEE International Geoscience and Remote Sensing Symposium-IGARSS, July 21-26, 2013, Melbourne, VIC, Australia. New York: IEEE, 2013: 508-511.

[9] Benedek C, Descombes X, Zerubia J. Building development monitoring in multitemporal remotely sensed image pairs with stochastic birth-death dynamics[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2012, 34(1): 33-50.

[10] Krizhevsky A, Sutskever I, Hinton G E. ImageNet classification with deep convolutional neural networks[C]//Advances in Neural Information Processing Systems, December 3-6, 2012, Lake Tahoe, Nevada, United States. USA: NIPS, 2012.

[11] Xi Z H, Hou C Y, Yuan K P, *et al*. Super-resolution reconstruction of accelerated image based on deep residual network[J]. Acta Optica Sinica, 2019, 39(2): 0210003.
席志红, 侯彩燕, 袁昆鹏, 等. 基于深层残差网络的加速图像超分辨率重建[J]. 光学学报, 2019, 39(2): 0210003.

[12] Ma H Q, Ma S P, Xu Y L, *et al*. Low-light image enhancement based on deep convolutional neural network[J]. Acta Optica Sinica, 2019, 39(2): 0210004.
马红强, 马时平, 许悦雷, 等. 基于深度卷积神经网络的低照度图像增强[J]. 光学学报, 2019, 39(2): 0210004.

[13] LeCun Y, Huang F J, Bottou L. Learning methods for generic object recognition with invariance to pose and lighting[C]//Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2004. CVPR 2004., June 27-July 2, 2004. Washington, DC, USA. New York: IEEE, 2004: 8168961.

[14] 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, 2015: 3431-3440.

[15] Szegedy C, Liu W, Jia Y Q, *et al*. Going deeper with convolutions[C]// 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 7-12, 2015, Boston, MA, USA. New York: IEEE, 2015: 15523970.

[16] 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, 2016: 770-778.

[17] Szegedy C, Vanhoucke V, Ioffe S, *et al*. Rethinking the inception architecture for computer vision[C]// 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 27-30, 2016, Las Vegas, NV, USA. New York: IEEE, 2016: 2818-2826.

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

[19] 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.

[20] Fu J, Liu J, Tian H J, *et al*. Dual attention network for scene segmentation[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, June 16-20, 2019, Long Beach, CA, USA. New York: IEEE, 2019: 3146-3154.

[21] Huang Z L, Wang X G, Huang L C, *et al*. CCNet: criss-cross attention for semantic segmentation[J/OL]. (2018-11-28)[2019-05-26]. https://arxiv.org/abs/1811.11721.

[22] Zhang X Y, Zhou X Y, Lin M X, *et al*. ShuffleNet: an extremely efficient convolutional neural network for mobile devices[C]// 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, June 18-23, 2018, Salt Lake City, UT, USA. New York: IEEE, 2018: 6848-6856.

[23] Howard A G, Zhu M L, Chen B, *et al*. MobileNets: efficient convolutional neural networks for mobile vision applications[J/OL]. (2017-04-17)[2019-05-26]. https://arxiv.org/abs/1704.04861.

[24] Romera E, Alvarez J M, Bergasa L M, *et al*. ERFNet: efficient residual factorized ConvNet for real-time semantic segmentation[J]. IEEE Transactions on Intelligent Transportation Systems, 2018, 19(1): 263-272.

[25] Paszke A, Chaurasia A, Kim S, *et al*. ENet: a deep neural network architecture for real-time semantic segmentation[J/OL]. (2016-06-07)[2019-05-26]. https://arxiv.org/abs/1606.02147arXiv.

[26] Ge S M, Luo Z, Zhao S W, *et al*. Compressing deep neural networks for efficient visual inference[C]// 2017 IEEE International Conference on Multimedia and Expo (ICME), July 10-14, 2017, Hong Kong, China. New York: IEEE, 2017: 667-672.

[27] Hubara I, Courbariaux M, Soudry D, *et al*. Binarized neural networks[C]// Advances in Neural Information Processing Systems, December 5-10, 2016, Barcelona, Spain. USA: NIPS, 2016.

[28] Srivastava N, Hinton G, Krizhevsky A, *et al*. Dropout: a simple way to prevent neural networks from overfitting[J]. Journal of Machine Learning Research, 2014, 15(1): 1929-1958.

[29] Wan L, Zeiler M, Zhang S X, *et al*. Regularization of neural networks using DropConnect[C]// International Conference on Machine Learning, June 16-21, 2013, Atlanta, GA, United States. USA: MIT Press, 2013, 28(3): 1058-1066.

[30] David J P, Kalach K, Tittley N. Hardware complexity of modular multiplication and exponentiation[J]. IEEE Transactions on Computers, 2007, 56(10): 1308-1319.

[31] Han J, Moraga C. The influence of the sigmoid function parameters on the speed of backpropagation learning[M]// Mira J, Sandoval F. From natural to artificial neural computation. IWANN 1995. Lecture notes in computer science. Berlin, Heidelberg: Springer, 1995, 930: 195-201.

[32] Horowitz M. 1.1 computing's energy problem (and what we can do about it)[C]// 2014 IEEE International Solid-State Circuits Conference Digest of Technical Papers (ISSCC), February 9-13, 2014, San Francisco, CA, USA. New York: IEEE, 2014: 10-14.