激光与光电子学进展 先进成像

DeepLabV3_DHC:城市无人机遥感图像语义分割

孙国文[1]，罗小波[1*]，张坤强[2]

[1]重庆邮电大学计算机科学与技术学院重庆空间大数据智能技术工程研究中心，重庆 400065；

[2]昆明理工大学信息工程与自动化学院，云南 昆明 650500

摘要 高分辨率无人机遥感图像具有极为丰富的语义和地物特征，在语义分割中容易出现目标分割不全、边缘信息缺失、分割精度不足等问题。为了解决上述问题，基于DeepLabV3_plus模型提出改进的DeepLabV3_DHC。首先，利用多种主干网络进行下采样，采集图像的低级特征和高级特征。其次，将原模型的atrous spatial pyramid pooling(ASPP)全部替换成深度可分离混合空洞卷积，同时添加自适应系数，减弱网格效应。之后，抛弃传统上采样的双线性插值法，替换为可学习的密集上采样卷积。最后，在低级特征中串联注意机制。选用多种主干网络进行实验，数据集选用四川省隆昌市地区的部分图像，采用平均交并比和类别平均像素准确率作为评价指标。实验结果表明：所提方法不仅具有较高的分割精度，而且减少了计算量和参数量。

关键词 城市无人机遥感图像；语义分割；深度可分离混合空洞卷积；密集上采样；注意力机制；网格效应

中图分类号 TP751 **文献标志码** A **DOI**：10.3788/LOP230886

DeepLabV3_DHC: Semantic Segmentation of Urban Unmanned Aerial Vehicle Remote Sensing Image

Sun Guowen[1], Luo Xiaobo[1*], Zhang Kunqiang[2]

[1]*College of Computer Sciences and Technology, Chongqing Engineering Research Center for Spatial Big Data Intelligent Technology, Chongqing University of Posts and Telecommunications, Chongqing* 400065, *China*;

[2]*School of Information Engineering and Automation, Kunming University of Technology, Kunming* 650500, *Yunnan, China*

Abstract High-resolution unmanned aerial vehicle remote sensing images have extremely rich semantic and ground feature features, which are prone to problems such as incomplete target segmentation, missing edge information, and insufficient segmentation accuracy in semantic segmentation. To solve the above problems, based on DeepLabV3_plus model, an improved DeepLabV3_ DHC is proposed. First of all, multiple backbone networks are used for down-sampling to collect low-level and high-level features of the image. Second, the atrous spatial pyramid pooling (ASPP) of the original model is replaced by a depthwise separable hybrid dilated convolution, and an adaptive coefficient is added to weaken the mesh effect. After that, the traditional sampling bilinear interpolation method is abandoned and replaced by the learnable dense upsampling convolution. Finally, cascade attention mechanism in low-level features. In this paper, a variety of backbone networks are selected for the experiment, and some images of Longchang City, Sichuan Province are selected for the dataset. The evaluation index uses the average intersection and combination ratio and the average pixel accuracy of the category as the reference basis. The experimental results show that the method in this paper not only has higher segmentation accuracy, but also reduces the amount of computation and parameters.

Key words urban unmanned aerial vehicle remote sensing image; semantic segmentation; depthwise separable hybrid dilated convolution; dense upsampling convolution; attention mechanism; grid effect

1 引 言

无人机遥感图像分割能记录区域地表的综合特征和地物个体特征，提供细致、真实、可靠的数据，广泛应用于城市规划和地理信息系统的构建等[1]。但无人机图像由于自身具有数据量大、场景复杂多变和时空分

收稿日期：2023-03-17；**修回日期**：2023-05-17；**录用日期**：2023-06-01；**网络首发日期**：2023-06-11

基金项目：国家重点研发计划政府间国际科技创新合作项目(2021YFE0194700)、重庆市高技术产业重大产业技术研发项目(D2018-82)、重庆市教委重点合作项目(HZ2021008)

通信作者：*luoxb@cqupt.edu.cn

辨率高等特点，致使各类地物具有许多复杂的特性，如：尺度变化大、物体分布密集、小物体繁多等[2]。所以，研究适用于无人机图像的深度学习算法具有很重要的意义。传统分割算法根据图像形状、边界、纹理等人工交互的方法进行处理[3]。可该类算法处理效率低下，高语义特征难以全部表达，导致精度和鲁棒性都较差。基于区域耦合的图像分割方法[4]，在抗噪和速度方面有着不错的效果，但面对场景复杂的图像却难有成效。随着数理知识的发展，马尔可夫随机场、条件随机场以及利用图论和小波变换的分割算法也逐渐发展起来[5]。但这些人工设计的特征与高语义之间存在“语义鸿沟”，导致最终模型的泛化性比较差。

卷积神经网络（CNN）可避免传统算法中人工提取特征的不足，模型能学习更深层次、更深广度、更本质的特征。FCN[6]作为首个端到端像素级预测的全卷积网络，通过全卷积结构替换CNN中的全连接层，再利用反卷积上采样特征。然而，FCN分割没有考虑全局上下文信息，图像结果不精细，细节丢失较多。U-Net模型弥补了FCN的不足，利用U字形直观简单的结构，能连续提取图像的上下文信息和细节信息，从而使输出结果更精确。遥感图像语义信息丰富，且尺度复杂。U-Net[7]采用镜像对称的方式来对图像进行卷积，每次卷积都会损失部分特征，且只支持单尺度预测，对于复杂多变的图像并不适合。ResNet[8]提出残差结构和跳跃连接，解决了训练中梯度爆炸或梯度消失问题，本文借鉴了其中的跳跃连接，上采样前与前一个特征融合，完善特征细节。DeepLabV1[9]模型使用空洞卷积避免池化带来的影响，同时使用条件随机场优化分割精度，但多尺度预测较差，主干网络过于简单提取不到图像深层次语义。DeepLabV2[10]在V1的基础上使用atrous spatial pyramid pooling（ASPP）结构，解决了多尺度预测问题，再结合ResNet主干网络加深网络深度，提高了学习能力。DeepLabV3[11]去除了V2的条件随机场，引用更加通用的框架，提升网络泛化性。DeepLabV3_plus[12]引入编码器和解码器，能更好地捕捉空间信息，上采样的同时结合低语义特征，细节特征更丰富。DeepLab系列创造性地引入空洞卷积，虽然增大了感受野，但卷积过程并不连续且忽略了网格效应。而无人机遥感图像信息丰富，单一使用DeepLabV3_plus并不能很好完成分割任务。

深度学习虽然在特征提取方面有较大优势，但遥感图像的噪声仍会对提取的特征造成一定影响，因此近年来，越来越多的学者将传统预处理方法和深度学习方法结合。曹春林等[13]使用段落匹配算法结合循环卷积网络解决目标遮挡问题。王成龙等[14]将K-means++算法结合深度学习框架，增强了模型的鲁棒性并大大减少参数量和计算量。无人机遥感图像分割同卫星遥感图像分割具有一定的相似性，大多数学者也是借鉴通用的方法在研究。陈雨情等[15]利用改进的DeepLabV3+模型增强农田的边缘信息，以此提高检测精度。蒯宇等[16]提取无人机图像的多尺度特征，构建多尺度特征信息网络，在城市植被分割领域取得不错效果。申华磊等[17]通过通道注意力机制和多级特征融合的方法增强U2_Net的鲁棒性，提高了小麦倒伏面积识别的准确率。尽管大多数学者在各自领域内都能取得不错的效果，但遗憾的是，这些学者的研究方法大多适用于背景简单、地物不太复杂的场景。应用较多的农学领域，无人机的分割对象一般是小麦、农田、水体、病变的植株等，而这些地物在水平方向的凹凸感不强，因此产生的噪声也较小，分割相对简单。城市的分割大多集中于建筑和道路的研究，这些地物与背景之间是有明显差异的，因此对分割模型要求也不会很高。但本研究的目标是城市无人机图像中的建筑、树木、车辆和背景，这些地物彼此之间容易混淆，对此所提方法绝不能仅仅参考以上学者的研究。为应对以上所述的种种问题，首先对无人机图像进行一定的匀光匀色处理，避免光照和拍摄角度等的影响。接着提出基于DeepLabV3_plus的新模型DeepLabV3_DHC，深度可分离混合空洞卷积（DHC）的使用不仅避免了网格效应，同时降低了模型的复杂度。上采样时使用密集上采样卷积（DUC）能更好地恢复特征细节，避免精度丢失。最后联合注意力机制，密切关注主干网络提取的3个低级特征中的重点内容，达到最佳的分割效果。

2 DeepLabV3_DHC方法和原理

2.1 总体结构

DeepLabV3_plus模型在常规的卷积神经网络中引入编码器结构和解码器结构。在编码器端采用Xception主干网络对输入的图像进行16倍下采样，之后提取下采样的两个特征并分别命名为低级特征和高级特征。接着，对高级特征采用ASPP结构进行处理，具体措施是采用1个1×1卷积和3个空洞率为6、12和18的3×3的空洞卷积结合池化操作同时对高级特征进行处理，最后将处理结果融合至一起并调整最后的输出通道数为256。在解码器端，首先将低级特征的通道数调整至48，目的是减少非必要的通道数，避免参数浪费。接着，再将高级特征4倍上采样并与低级特征融合，随后将融合结果传入3×3卷积中调整通道数为256。最后，将结果再次4倍上采样并传入1×1卷积中，将特征恢复到原图大小，并将通道数调整为模型最终分割结果的数目。

对于城市无人机遥感图像，常见的深度学习模型难以提取其全部的语义信息，特别是对其进行下采样、上采样时，图像保留的细节会少于一般图像。对此，在DeepLabV3_plus模型的基础上进行改进，提出DeepLabV3_DHC，提高模型的分割能力。整体的网络模型如图1所示，输入图像的大小为512 pixel×512 pixel。

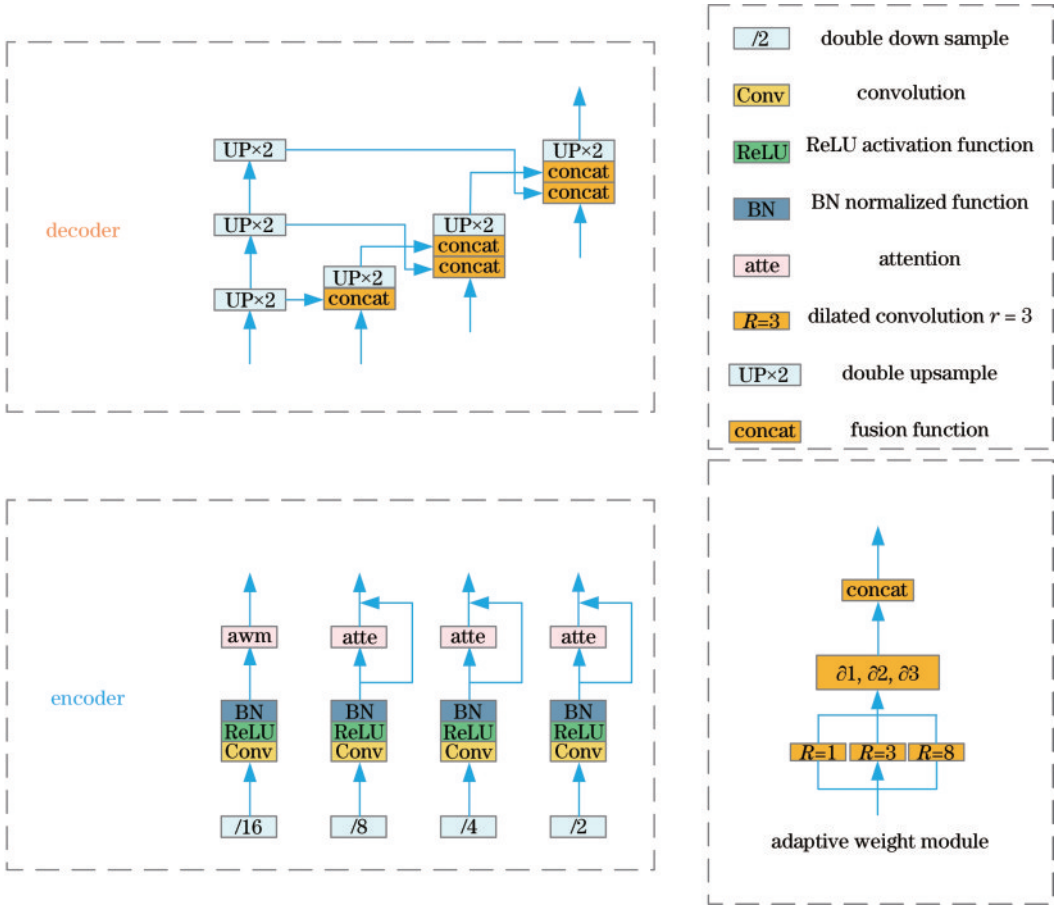

图 1　DeepLabV3_DHC 网络结构图

Fig. 1　DeepLabV3_DHC network structure diagram

在编码器端，由于无人机图像的语义信息丰富，利用主干网络进行 16 倍下采样，提取其 4 个特征代替原模型的 2 个特征。接着，由于 3 个低级特征的图形信息丰富，语义信息不明显，对其分别添加注意力机制，用以重点关注目标信息；相反，高级特征的语义信息丰富，但图形特征已然不清晰，这里只采用卷积进行处理。原模型中的 ASPP 结构虽然应用了空洞率 $r=[6,12,18]$ 的空洞卷积进行处理，但是仍然存在着网格效应[18]，并且池化层再次对高级特征下、上采样，尽管大小未变但是细节丢失得更多，不利于最终的分割。使用 DHC 方法，通过空洞率 $r=[1,3,8]$ 的深度可分离混合空洞卷积进行处理，并且对 3 个特征添加自适应权重，在避免网格效应的同时，提高特征提取能力并减少模型参数量和计算量。

解码器端，首先将经过 DHC 处理的高级特征分别进行 2 倍、4 倍和 8 倍密集上采样；接着，利用 1×1 卷积将经过注意力处理的 3 个特征调整至相同的通道数，便于之后的融合；随后，将 2 倍、4 倍和 8 倍密集上采样后的高级特征分别与前 3 个低级特征融合，增加低级特征的语义信息；最后，从高级特征开始，分别与前一个特征融合，再 2 倍密集上采样，直至最终的特征大小恢复到原来的 512 pixel× 512 pixel。

2.2　空洞卷积

主干网络处理图像时，通常会利用池化和卷积等操作，并且随着网络的深入图像分辨率会越来越低。这一过程会丢失一些像素值但图像的感受野会越来越大，直至获得一个高语义特征。普通卷积获取到的感受野有限，如 3×3 卷积、5×5 卷积，它们都只能获取来自上一层大小为 9 pixel 和 25 pixel 的感受野，再想扩大感受野就必须增大卷积核的大小，可这一操作会给模型带来大量的计算量和参数量。空洞卷积应运而生[19]。与普通卷积相比，空洞卷积提出了空洞率 r，普通卷积的空洞率 $r=1$，这意味着卷积核之间的参数距离为 0。而空洞卷积的 r 通常大于 1，即在卷积核各参数之间插入空洞 0，从而在不增加参数量的同时扩大感受野，提高特征提取效果。图 2 给出了 3 组 3×3 空洞卷积感受野示例。感受野的表达式为

$$K=k+(k-1)\times(r-1), \tag{1}$$

式中：K 为空洞卷积的感受野；k 为普通卷积的感受野。

2.3　深度可分离卷积

深度可分离卷积[20]又称为分组卷积，它将卷积分为逐通道卷积和逐点卷积。深度可分离卷积的结构如图 3 所示。

普通卷积对所有输入特征的通道数同时进行卷积

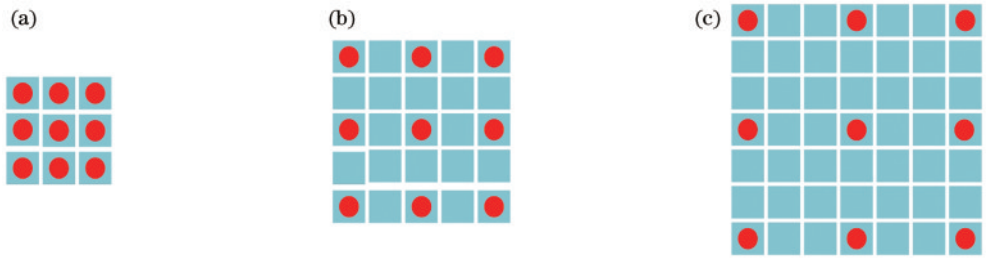

图 2　3组3×3空洞卷积感受野。(a) 空洞率为1;(b) 空洞率为2;(c) 空洞率为3

Fig. 2　Three groups of 3×3 dilated convolved receptive fields. (a) The dilated ratio is 1; (b) the dilated ratio is 2; (c) the dilated ratio is 3

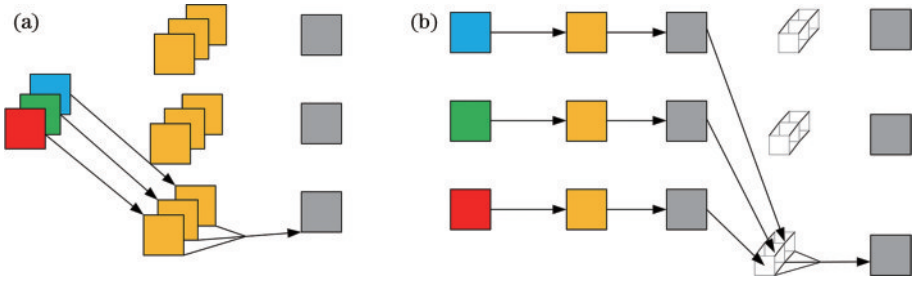

图 3　普通卷积和深度可分离卷积。(a) 普通卷积;(b) 深度可分离卷积

Fig. 3　Ordinary convolution and depth separable convolution. (a) Ordinary convolution; (b) depth separable convolution

操作,得到卷积特征。深度可分离卷积首先将输入特征的每个通道单独进行卷积处理,再将前一步结果分别利用多个1×1卷积调整通道数,最终得到卷积特征。得益于深度可分离卷积的这种特殊构造,其在保证卷积性能的同时可有效减少卷积过程产生的计算量和参数量。但其也有缺点,使用深度可分离卷积进行降通道处理时,一般不进行非线性处理,而是直接进行线性激活,否则会降低分割精度。

2.4　DHC方法

混合空洞卷积是一组带有不同空洞率r的特殊空洞卷积,它能有效避免网格效应,同时解决卷积过程不连续的问题。DHC函数表达式为

$$M_i=\max([M_{i+1}-2r_i,M_{i+1}-2(M_{i+1}-r_i),r_i])=\max([M_{i+1}-2r_i,-M_{i+1}+2r_i),r_i]),\quad(2)$$

式中:r_i为属于n组空洞率$[r_1,r_2,r_3,\cdots,r_{n-1},r_n]$中的第$i$位空洞率;$M_n=r_n$;卷积内核大小为$K$。当$M_2\leqslant K$时,表明该卷积符合混合空洞卷积的一个条件。除此之外,符合上述表达式,且n组空洞率之间的公因子为1,可最终确定为混合空洞卷积。DHC是在混合空洞卷积的基础上加上深度可分离卷积构建而成的,在减少模型复杂度的同时,保持相似或者更强的性能。DHC结构如图4所示。

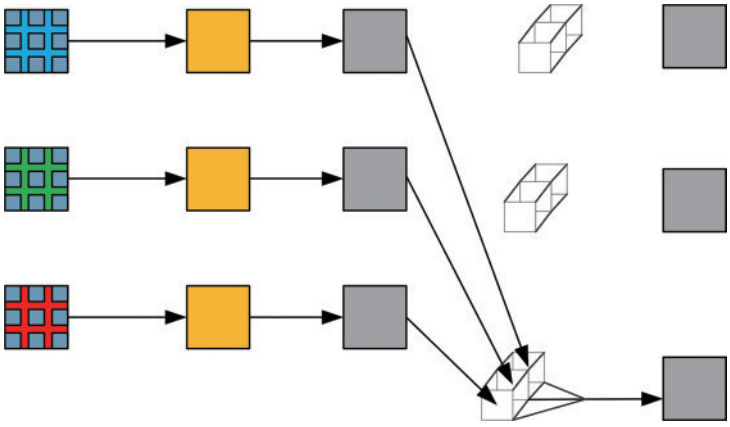

图 4　深度可分离混合空洞卷积

Fig. 4　Depthwise separable hybrid dilated convolution

选取的卷积核大小为3×3,空洞率$r=[1,3,8]$。$r=1$时,此时的DHC用于遍历整个特征图,避免采集的特征信息不全。$r=3$和$r=8$,是为了满足DHC表达式的前提下尽可能扩大感受野,同时限制感受野占整个特征大小的比重。尤其,当卷积的感受野接近或者超过待采集特征时,这时的卷积会退化成类似1×1

的卷积[21]，此时图像特征提取效果较差。

遥感无人机图像16倍下采样后，由于输入图像的大小为512 pixel×512 pixel，得到的高语义特征的大小为32 pixel×32 pixel。$r=3$时，此时卷积的感受野为7 pixel×7 pixel，同时r已然取到了最大值。例如：假设$r=4$，$M_2=\max([M_{i+1}-2r_i,-M_{i+1}+2r_i,4])$，得出$M_2>3$，不符合要求。$r=8$时，此时卷积的感受野为17 pixel×17 pixel，长、宽都约占高语义特征的一半，且经过大量实验对比验证，在此感受野下能尽最大可能平衡空洞对待采集特征的影响。

最后，尽管空洞卷积扩大了感受野，但由于卷积核各元素之间添加了过多的0元素，这仍可能会导致提取的信息不连续，影响最终的语义特征。为了尽量减少这类问题，设计了一种自适应的权重体系，在不影响结果的前提下抑制部分空洞卷积所提取的特征。根据卷积核各元素之间的距离，分别求取行列实际参与卷积计算的元素占总元素的比值：

$$K_i=\frac{k_{3\times3}^2}{\left[k_{3\times3}+(k_{3\times3}-1)\times(r-1)\right]^2},\tag{3}$$

式中：K_i为第$i(i=1,2,3)$个卷积行列的占比；$k_{3\times3}$为普通的3×3卷积的长或宽。

再对上述的结果进行一次Softmax得出最终各卷积的占比：

$$\partial_i=\mathrm{Softmax}(K_i)=\frac{\exp(K_i)}{\sum_{i=1}^{3}\exp(K_i)},\tag{4}$$

式中，∂_i为第i个的卷积的权重系数。

最后，将3个权重系数分别与对应的卷积相乘，再融合得到最终的高语义信息。

2.5　DUC方法

语义分割任务一般都有编码器和解码器，而大多数的解码器工作都会采用双线性插值法[22]进行上采样恢复特征。双线性插值法通过构建线性函数从而将上采样的空位补足，这种方法是不可学习的，会导致大量细节的丢失。DUC方法牺牲通道维度，可将上采样的长宽尺寸补足。它的好处在于，一切的操作都在整个特征中进行，没有引入额外的类似已知推导未知的方程来解决问题，特征没有被破坏。DUC的数学表达式为

$$\begin{cases} n,c,h,w \xleftrightarrow{\text{Conv}} n,r^2\times L,h,w \xleftrightarrow{\text{DUC}} \\ \qquad n,L,h\times r,w\times r \leftrightarrow \\ \qquad\quad n,L,H,W \end{cases},\tag{5}$$

式中：n是输入的banch_size大小；c为下采样后的通道数；h、w为下采样后长宽的大小且$h=H/r$，$w=W/r$，r为下采样的倍数；L是语义分割的种类数；H、W是输入原图的大小；Conv是普通卷积；DUC是密集上采样。DUC具体的操作和通道数的变化如图5所示。

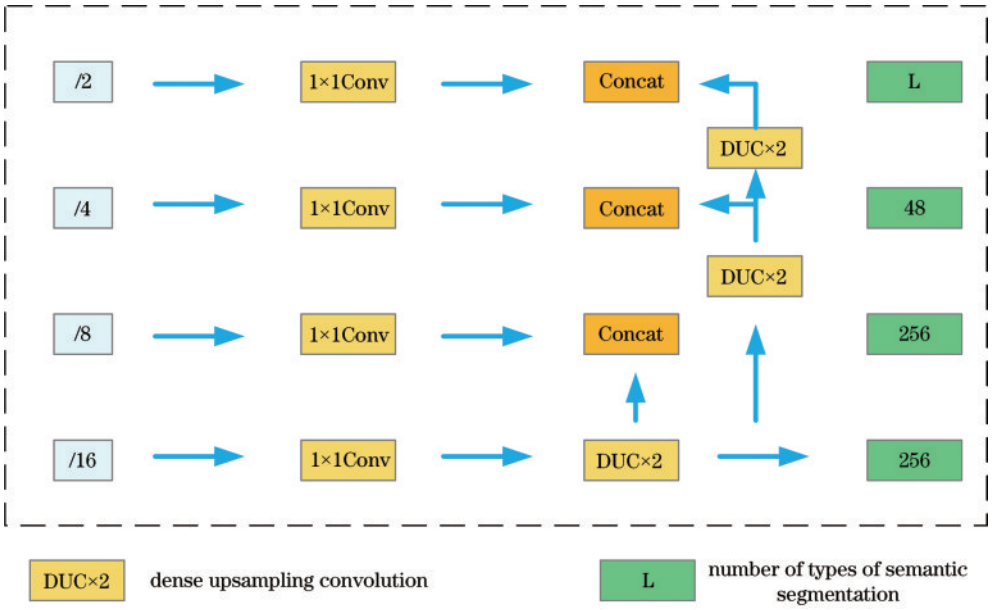

图5　密集上采样通道数变化

Fig. 5　Change in the number of dense upsampling channels

在编码器端，无论经过多少次下采样使用DUC皆可直接上采样为原图大小。由于提取4个特征，所以与传统的DUC不同，DUC方法的核心在于将通道维度reshape为空间维度。利用1×1卷积，首先将3个低级特征通道数调整为48，便于之后与高语义特征统一融合。再者将高语义特征通道数扩张为主干网络最后一层输出通道数（一般也是整个网络最大通道数），目的是增加更多的通道，以获得更多的特征。接着，将高语义特征2倍、4倍和8倍DUC之后的特征分别与8倍、4倍和2倍下采样的低级特征融合，增加低级特征的语义信息。最后，随着底层特征的不断融合，再利用1×1卷积结合DUC逐渐降低通道数，

减少参数损耗，调整的通道数为256、256、48，直到融合最后一个特征，将通道维度固定为语义分割的种类数，便于直接进行分割任务，避免多次使用1×1卷积增加内存访问量。

2.6　注意力机制

自从Transform[23]流行开来，注意力机制也逐渐在语义分割中成熟。注意力机制的本质是让网络自适应地关注到图像中的重点内容，避免学习到其他无用特征。借鉴CBAM[24]串联注意力机制的方式，将ECANet[25]通道注意力和CBAM中的空间注意力（SA）结合，组成ECA_SA串联注意力，如图6所示。

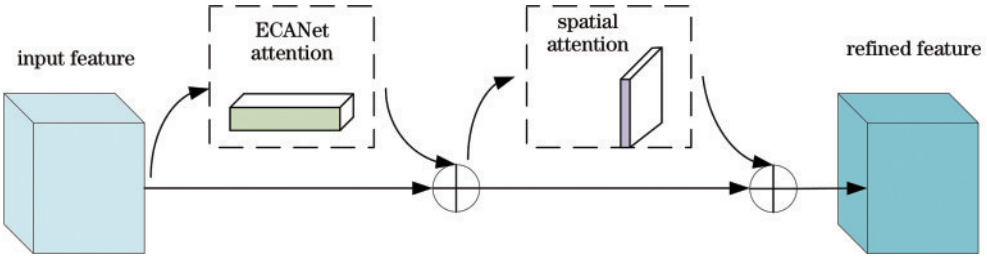

图6　注意力机制

Fig. 6　Mechanism of attention

原模型中通道注意力引入的平均池化层和最大池化层会降低特征表达，且产生不可忽略的计算量，导致模型臃肿。相比而言，ECANet抛弃了传统注意力机制常用的全连接层和最大池化层，仅使用全局平均池化获取特征，再利用1D卷积进行学习。1D卷积的大小会影响跨通道交互的覆盖率，所设计的1D卷积可根据输入图像通道数自适应调整大小，不仅减少参数量且提高了关注效率。接着，将1D卷积处理后的卷积特征传入Sigmiod函数，进行均一化处理，得到一组通道关注值。卷积核大小k表达式为

$$C=\phi(k)=2^{(\gamma*k-b)}, \tag{6}$$

$$k=\varphi(C)=\left|\frac{\log_2^{(C)}}{\gamma}+\frac{b}{\gamma}\right|, \tag{7}$$

式中：C为输入特征通道数；k为卷积核大小且取奇数；γ和b的值分别取2和1。

空间注意力对传入特征的每个特征点通道上取最大值和平均值，将二者结果融合后用1×1卷积调整通道数，再取Sigmiod函数，再次得到一组空间关注值。最终，将两组关注值分别与原特征相乘，获得带有注意力机制的关注特征。

将两种串联的注意力机制添加到主干网络提取的3个低级特征中，并借鉴ResNet的残差连接方法，在提高特征表达的同时不会降低原特征的信息。

3　分割结果与分析

3.1　数据集预处理

以于2022年3月使用大疆M300无人机P1相机拍摄的四川省隆昌市的部分城市图像作为数据集，图像的空间分辨率为0.03 m。由于原始图像的尺寸太大，系统运行效率过低，因此采用逐步叠加相切法对其进行处理。使用128步长对原始图像进行迭代相切，得到的目标图像尺寸为512 pixel×512 pixel。接着使用开源标注工具LabelMe[26]为每张图像绘制标签，标签的类别共计4类，分别是：背景、车、树木和建筑物。之后，得到2700张带有标签的图像，为了增加训练集数量，对切割后的2700张图像进行上下旋转和左右旋转，扩展得到5400张图像。最后，将处理完的图像按照2650∶1150∶1600分成3份，分别是训练集、验证集、测试集。图7为原始图像和对应的标签图示例。

图7　原始图像和对应的标签图

Fig. 7　Original image and corresponding label diagram

3.2　数据增强

为应对无人机拍摄过程中由于光线亮暗造成的亮度不均匀问题，对输入的图像进行局部自适应直方图匀光匀色处理。

利用Python库中的Opencv函数对输入图像的R、G、B三通道分别进行局部直方图均衡处理，为了防止均衡后的图像色彩失真，设定颜色对比度阈值为2，卷积核大小为10。接着，将处理后的三通道合并，得到如图8所示的图像。

除此之外，为应对模型过拟合的问题，对图像自适应添加高斯噪声，设定了自适应的左右、上下以及一定角度旋转等。

3.3　精度评价

由于研究对象是城市场景，因此在诸多的精度评价中应选择能代表整体分割精度的指标。平均交并比（MIOU）和类别平均像素准确率（MPA）作为整体分割的常用指标，可对分割结果提供重要的参考意义，因此选择MIOU和MPA对城市分割结果进行精度评价。MIOU和MPA的计算公式为

$$I = T_p/(T_p + F_p + F_n), \tag{8}$$

$$R_{MIOU} = \frac{1}{C}\sum_{i=1}^{C} I, \tag{9}$$

$$A = T_p/(T_p + F_n), \tag{10}$$

$$R_{MPA} = \frac{1}{C}\sum_{i=1}^{C} A, \tag{11}$$

式中：T_p表示预测值为某类且模型分类也为该类的像元数量；F_p表示预测值为其他类但模型分类为该类的像元数量；F_n表示预测值为该类但模型分类为其他类的像元数量；C为总的需要分类的个数；i代表的是第i类。

MIOU和MPA都可以反映模型整体的分割能力，且二者的取值范围皆在0～1之间，预测结果越接近1表示分割效果越好。

3.4　实验设计

在PyTorch环境下进行实验，服务器选择浪潮，GPU为单个Tesla-V100-PCIE-32GB，优化器策略选择ADAM[27]，损失函数选择常规交叉熵（CEL）损失函数。实验设定banch_size为4同时总计训练60个epoch。训练之前，为了统筹兼顾其他的数据集大小，均将图像大小调整至512 pixel×512 pixel，预测时会对图像大小进行复原，满足一致性要求。同时，对标签图像进行分类化处理，将像素值调整至0～3之间，提高运算速度。采用Xception、Resnet系列、MobilenetV2和MobilenetV3在ImageNet上的预训练模型，初始学习率设为0.00007，模型根据banch_size大小自动调整学习率，动量设为0.9。

3.5　实验结果及分析

分别在多个主干网络上进行实验用以检验模型的有效性，同时与DeepLabV3_plus模型进行对比。由表1可以看出，所提方法无论是在卷积结构较深的Xception和Resnet系列网络，还是在卷积结构较浅的Mobilenet系列网络上都有较大的提升。其中，前者效果最显著的是Xception主干网络，相比于原模型MIOU提升了2.77个百分点，MPA提升了3.61个百分点。而后者效果最佳的为MobilenetV2，对比原模型MIOU提升了5.15个百分点，MPA提升了4.02个百分点。相比于效果显著的Xception网络，尽管MobilenetV2网络提升较多，但无论是原模型还是改进的模型其分割效果都比较差。这主要因为无人机遥

表1　测试集数值评估

Table 1　Test set numerical evaluation

Model	Backbone	MIOU /%	MPA /%
DeepLabV3_plus	MobilenetV2	66.58	76.63
DeepLabV3_plus	MobilenetV3	63.51	74.13
DeepLabV3_plus	Resnet101	73.67	84.97
DeepLabV3_plus	Resnet152	75.58	86.37
DeepLabV3_plus	Resnext101	75.68	85.99
DeepLabV3_plus	Xception	78.22	86.48
DeepLabV3_DHC	MobilenetV2	**71.73**	**80.65**
DeepLabV3_DHC	MobilenetV3	**70.68**	**79.49**
DeepLabV3_DHC	Resnet101	**78.10**	**85.81**
DeepLabV3_DHC	Resnet152	**76.19**	**86.64**
DeepLabV3_DHC	Resnext101	**78.27**	**86.79**
DeepLabV3_DHC	Xception	**80.99**	**90.09**

图8　原始图像和局部直方图均衡图

Fig. 8　Original image and local histogram equalization

感图像语义信息丰富，特别在复杂城市中此点尤为突出，而较浅的网络难以充分提取其高语义特征，导致最终分割效果一般。不过DeepLabV3_DHC在多种主干网络中的分割效果仍旧大幅度超越原模型。

从表2可以看出，所提模型与DeepLabV3_plus模型相比，参数量（Parameters）和计算量（FLOPs）都有明显的减少，且由于其全部应用了深度可分离卷积，因此计算量减少得尤为明显。所提模型在精度和运算速度方面均超越了DeepLabV3_plus模型，具备实际应用于城市图像分割的能力。

表2　参数量和计算量对比

Table 2　Comparison of parameter number and calculation amount

Model	Backbone	Parameters /10^6	FLOPs /10^9
DeepLabV3_plus	MobilenetV2	5.831	26.433
DeepLabV3_DHC	MobilenetV2	**4.294**	**9.215**
DeepLabV3_plus	Xception	54.709	83.420
DeepLabV3_DHC	Xception	**51.513**	**65.195**
DeepLabV3_plus	MobilenetV3	11.725	30.535
DeepLabV3_DHC	MobilenetV3	**9.572**	**12.835**
DeepLabV3_plus	Resnet101	59.354	199.408
DeepLabV3_DHC	Resnet101	**56.181**	**181.863**

图9是所提模型和DeepLabV3_plus模型的可视化效果，其中，黑色、红色、绿色和黄色分别代表背景、建筑物、树木和车辆的分割结果。从图9可以看出，原模型分割结果误差较大，容易产生类似于椒盐噪声的零散点，建筑的分割容易产生裂痕或突增不存在的墙体，车辆的识别有残缺，树木的分割凌乱等。相对而言，由于所提方法加强了特征提取能力，且融合时统筹兼顾效果和效率，所以能够尽可能地避免以上的问题。所提处理方法能够很好地识别关键地物类别，不至于产生缺失或假目标，预测结果具有连贯性，极少中断，多类别预测不容易混淆。总之，所提方法在提升分割精度的同时可进一步减小误差。

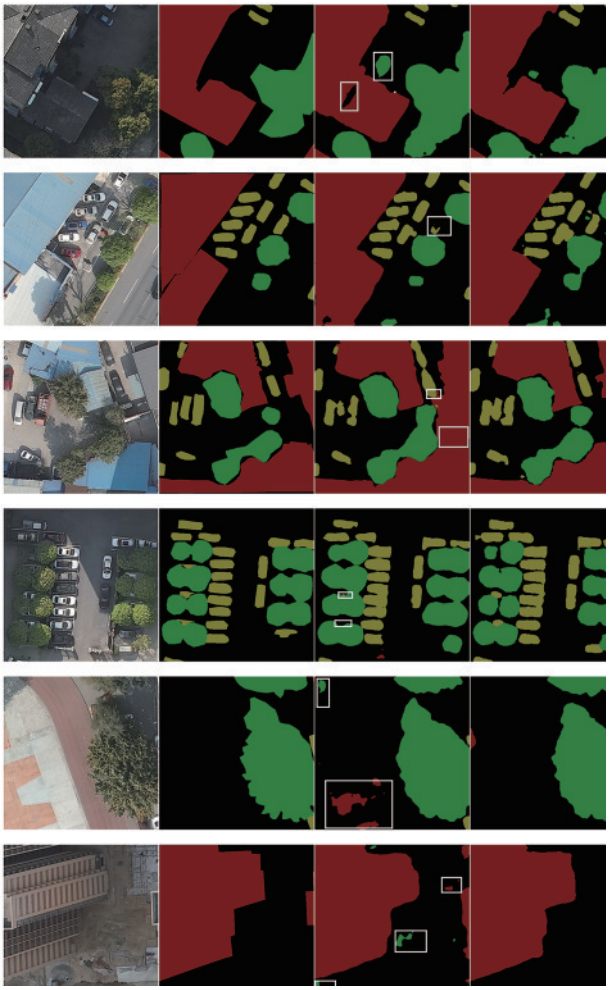

图9　可视化图结果图。(a)原始图像；(b)标签图像；(c) DeepLabV3_plus分割图像；(d) DeepLabV3_DHC分割图像

Fig. 9　Visualization result chart. (a) Original image; (b) Label image; (c) DeepLabV3_ plus image segmentation; (d) DeepLabV3_ DHC image segmentation

近年来，由于深度可分离卷积在卷积神经网络广泛应用，传统的3×3卷积不再优先作为提取特征的工具。相反，被抛弃的大卷积逐渐应用到各种模型中[28]。传统的大卷积具有大感受野的优点，能有效提升特征

提取效率且减少卷积层数量，但其高参数量和高计算量一直是学者避之不及的缺点。结合深度可分离卷积，在增加微量参数的基础上本研究探索大卷积能否应用于提取高语义特征。具体结果如表3所示，所设计的5×5卷积、7×7卷积、9×9卷积、11×11卷积、13×13卷积和15×15卷积对比3×3卷积可明显看出，大卷积的分割效果并没有比3×3卷积高出太多，总体效果在正常范围内波动。同时不可忽略的是，本研究控制了大卷积的空洞率，将其固定在较小的范围内，这主要是为了平衡参数量和计算量。由此可以得出结论，在控制参数量和计算量的前提下，大卷积的分割效果并不是特别突出，这也侧面印证了所提方法在提取特征方面有较为突出的优势。

表3 不同卷积在测试集的数值

Table 3 Different convolution values in the test set

Convolution size	Dilated rate	MIOU /%	MPA /%
3×3	1,3,8	80.99	**90.09**
5×5	1,3,5	**81.21**	89.41
7×7	1,2,3	80.58	89.07
9×9	1,2,3	80.43	89.68
11×11	1,2,3	80.63	88.15
13×13	1,2,3	**81.03**	88.97
15×15	1,2,3	80.73	88.92

3.6 消融实验

为了进一步验证所提方法的有效性，对网络有无DHC、DUC和ECA_SA注意力机制分别进行消融对比实验。通过表4可见，DeepLabV3_DHC模型的分割效果是最好的，而模型缺少DHC、DUC和ECA_SA注意力中的任一种，总体的分割效果皆不如完整模型。特别地，当模型去掉三种之中的任意一种，整体模型对车辆的分割效果较差，导致最终的效果很差。

表4 消融实验对比

Table 4 Comparison of ablation experiments

Model	MIOU /%	MPA /%
DeepLabV3_DHC (no DHC)	71.78	78.30
DeepLabV3_DHC (no DUC)	70.83	82.83
DeepLabV3_DHC (no attention)	77.11	83.78
DeepLabV3_DHC	**80.99**	**90.09**

4 结 论

提出DeepLabV3_plus的改进方法DeepLabV3_DHC。在编码器端引入的串联注意力增强了多级特征的表达，DHC方法提升了模型的分割能力。解码器端采用的DUC上采样方法能尽可能地避免特征的损失，多特征融合的方法将语义特征和图形特征结合进一步提升分割效果。对比所提模型和原模型的分割结果，在精度方面，所提模型识别精度高，能够改善分割中存在的识别混淆、分割不全、边缘信息缺失等问题。在效率方面，所提模型的参数量和计算量都较低，运行速度相较而言有所提高。因此，所提模型无论是在精度还是效率方面都优于原模型。但其也有不足之处，对于城市中的小物体车辆的分割效果较为一般，特别是车辆停靠的地方与屋檐重合时，这类缺点会放大。除此之外，网络结构较浅的主干网络提取特征较为困难，导致提取的高语义特征不明显，最终分割精度也较差。提出一种适用于所有主干网络的分割方法，是接下来的工作。

参 考 文 献

[1] 余帅, 汪西莉. 含多级通道注意力机制的CGAN遥感图像建筑物分割[J]. 中国图象图形学报, 2021, 26(3): 686-699.
Yu S, Wang X L. Remote sensing building segmentation by CGAN with multilevel channel attention mechanism [J]. Journal of Image and Graphics, 2021, 26(3): 686-699.

[2] 程擎, 范满, 李彦冬, 等. 无人机航拍图像语义分割研究综述[J]. 计算机工程与应用, 2021, 57(19): 57-69.
Cheng Q, Fan M, Li Y D, et al. Review on semantic segmentation of UAV aerial images[J]. Computer Engineering and Applications, 2021, 57(19): 57-69.

[3] 谌华, 郭伟, 闫敬文. 综合边界和纹理信息的合成孔径雷达图像目标分割[J]. 中国图象图形学报, 2019, 24(6): 882-889.
Chen H, Guo W, Yan J W. Synthetic aperture radar image target segmentation method based on boundary and texture information[J]. Journal of Image and Graphics, 2019, 24(6): 882-889.

[4] 李更生, 刘国军, 马文涛. 基于区域信息耦合的自适应图像分割[J]. 激光与光电子学进展, 2022, 59(2): 0210013.
Li G S, Liu G J, Ma W T. Adaptive image segmentation based on region information coupling[J]. Laser & Optoelectronics Progress, 2022, 59(2): 0210013.

[5] 肖春姣, 李宇, 张洪群, 等. 深度融合网结合条件随机场的遥感图像语义分割[J]. 遥感学报, 2020, 24(3): 254-264.
Xiao C J, Li Y, Zhang H Q, et al. Semantic segmentation of remote sensing image based on deep fusion networks and conditional random field[J]. Journal of Remote Sensing, 2020, 24(3): 254-264.

[6] 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.

[7] 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.

[8] 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.

[9] 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)[2023-02-03]. https://arxiv.org/abs/1412.7062.

[10] 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.

[11] Chen L C, Papandreou G, Schroff F, et al. Rethinking atrous convolution for semantic image segmentation[EB/OL]. (2017-06-17)[2023-02-06]. https://arxiv. org/abs/1706.05587.

[12] 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.

[13] 曹春林, 陶重犇, 李华一, 等. 实时实例分割的深度轮廓段落匹配算法[J]. 光电工程, 2021, 48(11): 22-33.
Cao C L, Tao C B, Li H Y, et al. Deep contour fragment matching algorithm for real-time instance segmentation[J]. Opto-Electronic Engineering, 2021, 48(11): 22-33.

[14] 王成龙, 赵倩, 赵琰, 等. 基于深度可分离卷积的实时遥感目标检测算法[J]. 电光与控制, 2022, 29(8): 45-49.
Wang C L, Zhao Q, Zhao Y, et al. A real-time remote sensing target detection algorithm based on depth separable convolution[J]. Electronics Optics & Control, 2022, 29(8): 45-49.

[15] 陈雨情, 王修信. 改进DeepLabv3+模型无人机图像农田信息提取[J/OL]. 计算机工程与应用:1-13[2023-02-03]. https://kns.cnki.net/kcms/detail/11.2127.tp.20220705.1859.018.html.
Chen Y Q, Wang X X. Improved DeepLabv3+ Model UAV image farmland information extraction[J/OL]. Computer Engineering and Applications: 1-13[2023-02-03]. https://kns.cnki.net/kcms/detail/11.2127.tp.20220705.1859.018.html.

[16] 蒯宇, 王彪, 吴艳兰, 等. 基于多尺度特征感知网络的城市植被无人机遥感分类[J]. 地球信息科学学报, 2022, 24(5): 962-980.
Kuai Y, Wang B, Wu Y L, et al. Urban vegetation classification based on multi-scale feature perception network for UAV images[J]. Journal of Geo-Information Science, 2022, 24(5): 962-980.

[17] 申华磊, 苏歆琪, 赵巧丽, 等. 基于深度学习的无人机遥感小麦倒伏面积提取方法[J]. 农业机械学报, 2022, 53(9): 252-260, 341.
Shen H L, Su X Q, Zhao Q L, et al. Extraction method of wheat lodging area by UAV remote sensing based on deep learning[J]. Transactions of the Chinese Society for Agricultural Machinery, 2022, 53(9): 252-260, 341.

[18] 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.

[19] Zhang J, Lin S F, Ding L, et al. Multi-scale context aggregation for semantic segmentation of remote sensing images[J]. Remote Sensing, 2020, 12(4): 701.

[20] Chollet F. Xception: deep learning with depthwise separable convolutions[C]//2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), July 21-26, 2017, Honolulu, HI, USA. New York: IEEE Press, 2017: 1800-1807.

[21] Yu F, Koltun V, Funkhouser T. Dilated residual networks[C]//2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), July 21-26, 2017, Honolulu, HI, USA. New York: IEEE Press, 2017: 636-644.

[22] 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.

[23] Vaswani A, Shazeer N, Parmar N, et al. Attention is all You need[C]//Proceedings of the 31st International Conference on Neural Information Processing Systems, December 4-9, 2017, Long Beach, California, USA. New York: ACM Press, 2017: 6000-6010.

[24] Woo S, Park J, Lee J Y, et al. CBAM: convolutional block attention module[M]//Ferrari V, Hebert M, Sminchisescu C, et al. Computer vision-ECCV 2018. Lecture notes in computer science. Cham: Springer, 2018, 11211: 3-19.

[25] Wang Q L, Wu B G, Zhu P F, et al. ECA-net: efficient channel attention for deep convolutional neural networks[C]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 13-19, 2020, Seattle, WA, USA. New York: IEEE Press, 2020: 11531-11539.

[26] Russell B C, Torralba A, Murphy K P, et al. LabelMe: a database and web-based tool for image annotation[J]. International Journal of Computer Vision, 2008, 77(1): 157-173.

[27] Kingma D P, Ba J. Adam: a method for stochastic optimization[EB/OL]. (2014-12-22)[2023-02-06]. https://arxiv.org/abs/1412.6980.

[28] Ding X H, Zhang X Y, Han J G, et al. Scaling up your kernels to 31 × 31: revisiting large kernel design in CNNs[C]//2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 18-24, 2022, New Orleans, LA, USA. New York: IEEE Press, 2022: 11953-11965.