王威威, 罗杨, 李熹. 基于注意力机制的城市街景语义分割方法[J]. 智能计算机与应用,2026,16(7):212-217. DOI:10.20169/j.issn.2095-2163.24112103

基于注意力机制的城市街景语义分割方法

王威威, 罗 杨, 李 熹
(广西民族大学 人工智能学院, 南宁 530006)

摘 要: 现有的图像语义分割模型在应用于城市街景图像的分割任务时,由于城市街景环境中背景信息丰富,小目标较多,存在遮挡等问题,分割结果往往不佳。针对以上问题,提出一种基于 DeepLabv3+的改进算法。该算法基于编码解码架构,在编码端、输入图像经过主干网络提取到的特征,在进入空洞特征金字塔池化模块后,先通过一个 EMA 模块,强化模型对全局信息的提取。在解码端,将由主干网络提取到的低层次特征通过一个 ECA 模块,减少背景和无关小目标干扰后,再与编码器得到的高层次特征融合。该模型在 CityScapes 数据集上的实验,取得了平均交并比 74.5%的结果。

关键词: 语义分割; 城市街景; 注意力机制; 多尺度特征融合; 编码-解码架构
中图分类号: TP391.41 **文献标志码**: A **文章编号**: 2095-2163(2026)07-0212-06

Semantic segmentation of urban street scenes based on attention mechanism

WANG Weiwei, LUO Yang, LI Xi
(School of Artificial Intelligence, Guangxi Minzu University, Nanning 530006, China)

Abstract: When the existing image semantic segmentation models are applied to the segmentation task of urban street view images, the segmentation results are often poor due to the rich background information, many small objects, and occlusion in the urban street view environment. Aiming at the above problems, an improved algorithm based on DeepLabv3+ is proposed. The algorithm is based on the encoder-decoder architecture. At the encoder, the features of the input image extracted by the backbone network are inputted into the hole feature pyramid pooling module, and then through an EMA module to strengthen the extraction of global information of the model. At the decoder, the low-level features extracted by the backbone network are passed through an ECA module to reduce the interference of background and irrelevant small targets, and then fused with the high-level features obtained by the encoder. The experimental results on the CityScapes dataset show that the proposed model achieves an average Intersection over Union ratio of 74.5%.

Key words: semantic segmentation; urban street view; attention mechanism; multi-scale feature fusion; encoder-decoder architecture

0 引 言

语义分割是计算机视觉(CV)领域的一个核心研究课题,其研究目标是为图像中的每个像素点分配与之相对应的语义标签[1],从而实现像素级别的分类。语义分割主要应用于自动驾驶[2]、医学图像分析[3]、缺陷检测[4]等领域。

城市街景语义分割[5]任务是图像语义分割研究的一个重要课题,对于自动驾驶的研发和智慧城市[6]的建造具有至关重要的作用。有分析指出,使用经典的卷积神经网络模型进行城市街景语义分割研究时存在着构图复杂、小目标众多、遮挡和干扰较多等难题,从而导致分割精度较低。而现有的对经典卷积神经网络模型的改进模型也常常存在运算量大、目标边缘模糊等问题。

随着深度学习技术的快速发展,更多适用于城市街景图像语义分割任务的模型被提出。Chen 等学者[7]提出的 DeepLab 系列模型引入了空洞卷积,有效解决了模型对目标上下文信息的提取能力。Zhao 等学者[8]随后提出了 PSPNet,运用金字塔池化技术加强对上下文信息的利用。Wang 等学者[9]提出了 HRNet 模型,保留了特征图的高分辨率,解决

作者简介: 王威威(2000—),男,硕士。主要研究方向:机器学习,深度学习,计算机视觉; 罗 杨(1999—),男,硕士。主要研究方向:深度学习,图像处理。
通信作者: 李 熹(1982—),男,副教授。主要研究方向:软件设计,生物信息学。Email:18810997486@139.com。

收稿日期: 2024-11-21

了小目标分割问题。Google 推出的 EfficientNet[10] 和 EfficientDet[11]模型着眼于高效分割，兼顾了高效率和高精度。

本文提出一种基于改进 DeepLabv3+的语义分割算法。该算法对经典的 DeepLabv3+模型做了 4 项改进：

(1)将主干网络替换为 ResNet50 网络[12]，提高主干网络信息提取能力。

(2)在编码器端的 ASPP 模块之后添加一个 EMA 注意力模块，增强对长距离信息的捕捉。

(3)在解码器的前端加入 ECA 注意力模块，减少无关小目标的干扰。

(4)针对 CityScapes 数据集中存在的数据不平衡的问题，采用 focal loss[13] 损失函数，该函数在类别损失前加入权重系数以缓解由于数据不平衡引起的分割误差，提高分割精度。本文方法在 CityScapes 数据集上平均交并比达到了 74.5%。

1　网络结构

1.1　主干网络

本文的骨干网络选择了 ResNet50 模型，ResNet 系列模型首次在神经网络中使用了跳跃连接技术，提升了网络性能。ResNet50 是 ResNet 最常用的变种之一，包含 4 种残差层，残差层之间存在跳跃连接。4 种残差层分别重复 3、4、6、3 次，共计 50 层。为了保留更多来自原始图像的特征，本文将 ResNet101 模型第一层的 7×7 卷积替换为 1×1 卷积。同时，为加快收敛速度，将骨干网络在 ImageNet[14] 数据集上做了预训练。

1.2　注意力机制

1.2.1　EMA 注意力机制

捕获远程依赖关系在深度神经网络中至关重要。对于图像数据，远程依赖关系通常是由卷积操作的深度堆栈形成的大型接受场来建模的。由于一次卷积运算只处理一个局部邻域，捕捉像素点之间的远程依赖关系需要在不同位置重复多次卷积操作，导致计算效率低且优化困难。

Non-local 注意力[15] 机制被提出后，引发了计算机视觉领域的广泛关注。随后一系列文章证明了其在语义分割任务中的有效性。然而，基于自注意力机制的 non-local 注意力机制在计算时需要生成一个巨大的注意力图，导致其空间复杂度和时间复杂度都非常可观。

EMA 注意力机制[16]、也就是期望最大化注意力机制，是一种高效的多尺度空间特征学习注意力机制，摒弃了在全图上计算注意力图的流程，转而通过期望最大化(EM)算法迭代出一组紧凑的基，在这组基上运行注意力机制，从而大大降低了复杂度。具体采用分组策略而无需降维，避免了因降维导致的信息丢失，并通过跨维度交互作用有效保留每个通道的信息。又通过将通道维度划分为若干子集，实现空间语义特征在各个特征图中的均匀分布。

EMA 注意力机制具体实现方式如图 1 所示。给定任意输入 $X \in R^{C\times H\times W}$，EMA 按跨通道维度方向将输入特征图 X 划分为 G 个子特征 $X=[X_0,\cdots,X_{G-1}]$，$X_i \in R^{C//G\times H\times W}$ 用以学习不同的语义。其中，C 为特征图的通道数，G 为划分的子特征个数。模块由 3 条并行路线组成，并行多分支的网络结构设计能够捕获所有通道之间的依赖关系，同时降低计算预算。前 2 个分支为 1×1 分支，目的是将输入的特征图通过一维全局平均池化操作分别对水平、垂直 2 个方向通道进行编码；第 3 分支为 3×3 分支，通过 3×3 卷积来捕获多尺度特征表示。计算公式如下：

$$z_c^h=\frac{1}{H}\sum_{i=0}^{W}x_c(h,i) \tag{1}$$

$$z_c^w=\frac{1}{W}\sum_{j=0}^{H}x_c(j,w) \tag{2}$$

其中，z_c^h 和 z_c^w 分别表示输入特征图在垂直和水平 2 个方向的输出；h、w 分别表示通道 c 的高度和宽度；i、j 表示不同的空间方向。

在跨通道学习时，将分支中沿图像水平、垂直方向编码的 2 个特征图相连接，共享不进行降维的 1×1 卷积。然后再次分解为 2 个向量，采用 Sigmoid 函数进行线性拟合。公式如下：

$$f=\sigma(F_1([z^h,z^w])) \tag{3}$$

其中，σ 表示非线性函数，F_1 表示卷积变换函数。最后在每个特征组内通过乘法融合两方向通道注意力权重因子，从而实现 2 条并行路径之间的跨通道特征交互。

在跨空间学习阶段，对 3×3 分支和 2 个 1×1 分支融合后的输出使用二维全局平均池化技术进行全局空间信息编码，再通过 Softmax 函数转化为相应的维度形状。最后，将生成的 2 个空间注意力的权重值通过跨空间交互模块进行聚合，建立远程依赖关系，捕获像素级成对关系并突出所有像素的全局上下文信息。最终得到与输入 X 大小相同的输出。

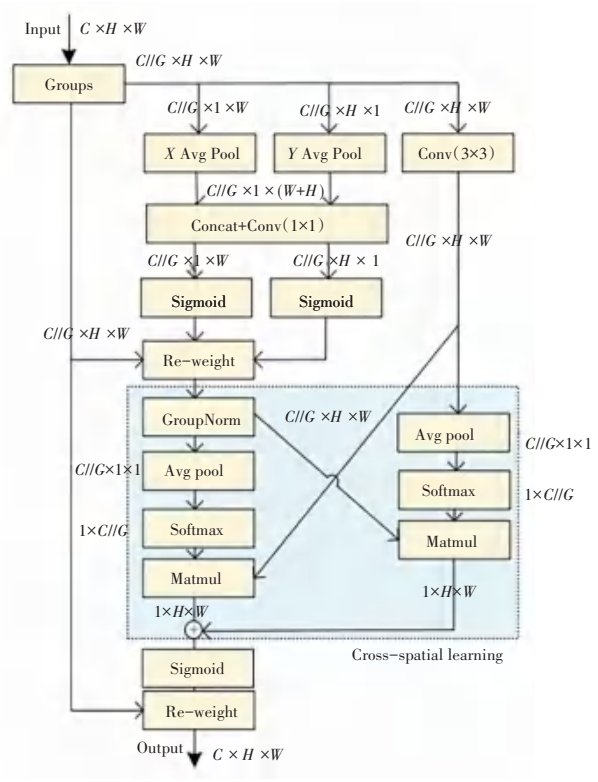


图 1 EMA 注意力模块

Fig. 1 EMA attention module

1.2.2 ECA 注意力机制

注意力机制使得模型更关注全局特征信息中重要特征,抑制学习无关特征,然而广泛使用的 SE[17]、CBAM[18]注意力,常常忽略上下文信息重要性。ECA 注意力机制[19]通过不降维的局部跨信道交互策略和自动学习实现空间和通道信息的有效融合,有效提高分类精确率;并选择一维卷积核代替传统的全连接层,增强了特征提取性能。ECA 模块架构如图 2 所示。

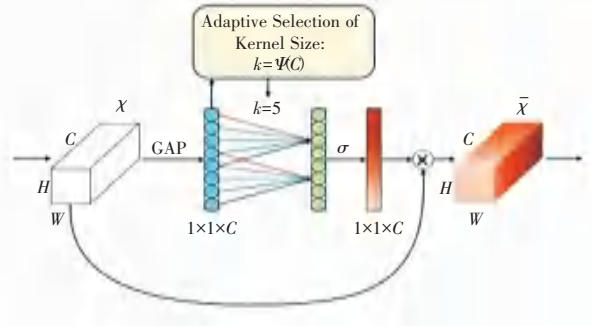


图 2 ECA 注意力模块

Fig. 2 ECA attention module

该模块通过使用全局平均池化技术来获得特征,通过一维卷积核获取通道权重。图 2 中,H、W 和 C 分别表示特征图的高度、宽度和通道数。

ECA 注意力的计算公式如下:

$$\tilde{x} = x \times \sigma \sum_{j=1}^{k} \omega^j y_i^j,\ y_i^j \in \Omega_i^k \tag{4}$$

$$y = \frac{1}{W \times H} \sum_{i=1}^{W} \sum_{j=1}^{H} x(i,j) \tag{5}$$

$$\omega = \sigma(C1D_k(y)) \tag{6}$$

其中,x 表示输入的特征图;y 表示全局平均池化操作的输出;$C1D$ 表示一维卷积;σ 表示 Sigmoid 激活函数;Ω_i^k 表示通道为 y 的 k 个相邻通道集合;y_i^j 表示通道集合中的一个通道。

自适应 Kernel Size 计算公式:

$$C = \varphi(k) = 2^{\gamma \times k - b} \tag{7}$$

$$K = \psi(C) = \left| \frac{\log_2 C}{\gamma} + \frac{b}{\gamma} \right|_{\text{odd}} \tag{8}$$

其中,$|x|_{\text{odd}}$ 为选择最近的奇数;γ 值取 2;b 值取 1;C 为通道维度;K 为核大小。

ECA 模块具体实现步骤如下:

(1)将输入的特征图进行全局平均池化操作,将其维度转换为[1,1, C]。

(2)根据特征图中通道的数量,确定自适应的一维卷积核 K 的大小。

(3)将一维卷积的输出再输入 Sigmoid 激活函数,输出结果为每个通道的权重系数。为输出与输入向量大小相同的权重向量,将权重系数控制在[0,1]区间内。

(4)将标准化后的权重与输入的特征图通道值相乘。

1.3 焦点损失函数

焦点损失(Focal Loss)函数是一种用于解决类别不平衡问题的损失函数,特别适用于目标检测[20]和图像分割[21]任务。当数据集存在类别不平衡问题时,常规的交叉熵损失函数对于大量的易分类样本会产生较大的损失,从而使模型过度关注这些样本。Focal Loss 通过引入一个调变因子,降低易分类样本的权重,从而更加关注难分类样本,提升整体模型性能。Focal Loss 的数学表达式如下:

$$\mathrm{FL}(p_t) = -(1-p_t)^{\gamma} \log(p_t) \tag{9}$$

其中,$(1-p_t)^{\gamma}$ 表示调变因子,且 $\gamma \geqslant 0$。

1.4 模型整体架构

本文模型基于 DeepLabv3+模型[22],改进后的模型架构如图 3 所示。本文模型属于经典的编码解码架构。在编码器部分,输入图像先经过 ResNet50 骨干网络的特征提取,在骨干网络中采用空洞卷积[7]以获取更大的感受野。从骨干网络获得 2 个不同层次的特征图,将获得的高水平特征输入 ASPP 空洞空间

金字塔池化[23]模块。在 ASPP 模块中，输入的特征图并行的进行 5 个操作，分别是 1 个 1×1 卷积操作、3 个空洞率分别为 6，12，18 的 3×3 空洞卷积操作和 1 次全局平均池化操作。5 个并行操作将获得 5 张不同的特征图，将获取的多个特征图拼接并使用 1×1 卷积改变通道数，然后将获取的特征图通过一个 EMA 注意力模块后再输入解码器。

在解码器部分，先将从编码器得到的高水平特征通过 4 倍双线性插值上采样操作扩大特征图。将从骨干网络提取到的低水平特征先输入 ECA 注意力模块，在 ECA 注意力模块中，目标的空间坐标信息被嵌入通道信息中，加强了对目标边缘信息和上下文信息的提取和利用，将获得的特征图通过 1×1 卷积改变通道数以确保能与从编码器获得的高水平特征图拼接。然后，将拼接后的特征图与从编码器获得的高水平特征图堆叠，接着再进行一次 3×3 卷积保证从 2 张特征图得到的特征充分融合，最后通过 4 倍双线性插值上采样操作得到与输入图像大小相同的输出结果。

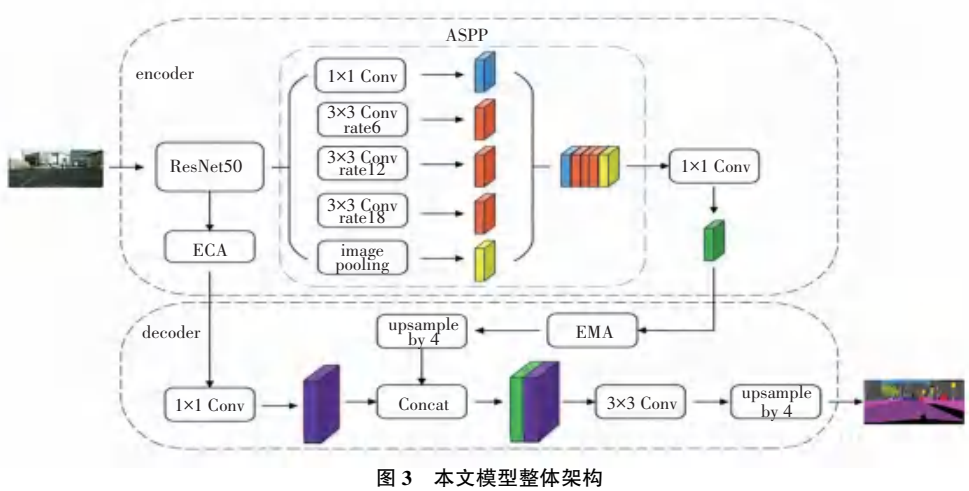


图 3　本文模型整体架构

Fig. 3　Overall architecture of the model in this paper

2　实验

2.1　实验数据集和参数配置

本实验使用的数据集是 CityScapes 数据集。CityScapes 数据集[23]、即城市景观数据集，是一个包含大量城市街景图像的大型数据集，涵盖了 50 个不同城市的街道场景。CityScapes 拥有 5 000 张在城市环境中驾驶场景的图像，其中训练集图片 2 975 张，验证集图像 500 张，测试集图像 1 525 张。该数据集包含有 19 个类别的密集像素标注，其中 8 个精度达到了实例级分割。

CityScapes 数据集中图像数据丰富，且有 fine 和 coarse 两套评测标准，前者提供 5 000 张精细标注的图像，后者提供 20 000 张粗糙标注的图像和 5 000 张精细标注的图像。为方便实验，将 CityScapes 数据集按 9∶1 的比例划分为训练集和验证集[24]。

实验使用的操作系统为 Windows 10，使用的深度学习框架为 PyTorch，使用的服务器为 Intel(R) Xeon(R) Gold 6330 CPU、显存为 24 GB、GPU 为 NVIDIA GeForce RTX 3090 Laptop GPU；编程软件为 PyCharm2023、Pytorch2.0.0、Python3.8(ubuntu20.04)、CUDA11.8 等。

2.2　实验评价指标

为了定量评估语义分割网络的性能，实验设置了 2 个评估指标：平均像素精度(MPA)和平均交并比(mIoU)。其中，MPA 表示每个类别中被正确识别的像素总数占该类别的像素总数的比例的均值，MPA 越大，预测正确的像素越多；mIoU 表示真实分割结果与预测分割结果的交集与并集的商，即预测结果和真实结果的重合度。指标计算公式如下：

$$\mathrm{MPA}=\frac{1}{N}\sum_{k=1}^{N}\frac{\mathrm{TP}_k}{\sum_{k=1}^{N}\mathrm{FP}_k} \tag{10}$$

$$\mathrm{mIoU}=\frac{1}{N}\sum_{k=1}^{N}\frac{\mathrm{TP}_k}{\mathrm{TP}_k+\mathrm{FP}_k+\mathrm{FN}_k} \tag{11}$$

其中，N 表示类别总数；k 表示第 k 类；TP_k、TN_k、FP_k、FN_k 分别表示真正例、真负例、假正例和假负例。

2.3　对比实验

为验证本文模型优势，将本模型与几种经典的语义分割网络进行对比实验，实验使用的对比模型包括 HRNet、PSPNet、EfficientNet、EfficientPS[25] 和

DeepLabv3+。模型在 CityScapes 数据集上的对比实验结果见表 1。

表 1 各模型在 CityScapes 测试集上的对比结果

Table 1 Comparison results of various models in the CityScapes test set

Model	Backbone	MAP/%	mIoU/%
PSPNet	ResNet50	78.8	70.6
EfficientNet	EfficientNet	79.2	70.8
HRNet	HRNet	80.3	71.7
EfficientPS[25]	EfficientNet	80.6	72.1
DeepLabv3+	Xception	78.9	71.2
本文模型	ResNet50	82.2	74.5

实验结果表明,本文模型的 MPA 指标和 mIoU 指标相对于对比模型有较为显著的提升,与基准模型 DeepLabv3+相比,本文模型的 MPA 指数提升了 3.3%,mIoU 指数提升了 3.3%,提升显著。与 PSPNet 模型相比,本文的模型将 MPA 值和 mIoU 值分别提高了 3.4%和 3.9%;与 EfficientNet 模型相比,本文模型的 MPA 高出 3.0%,mIoU 则提高了 3.7%;与 HRNet 模型进行比较,本文模型的 MPA 值和 mIoU 值分别高出 1.9%和 2.8%;与 EfficientPS 模型进行比较,本文模型的 MPA 值和 mIoU 值分别高出 1.6%和 2.4%。

2.4 消融实验

为了验证本文模型所添加的模块的有效性,在 CityScapes 数据集上做了消融实验,实验结果见表 2。

从实验结果可以发现,DeepLabv3+基准模型在仅添加 EMA 注意力模块时,MPA 指数提升至 80.4%,mIoU 指数提升至 73.3%,这是因为经 ASPP 模块处理后的高水平特征,而后又经 EMA 模块处理,在保留原有维度的同时,通过跨维度交互作用有效保留每个通道的信息,提升了对输入图像的全局信息,特别是长距离依赖信息的提取和利用。模型在仅添加 ECA 模块时,MPA 指数提升至 80.0%,mIoU 指数提升至 72.9%,因为 ECA 模块通过不降维的局部跨信道交互策略和自动学习快速实现空间和通道信息的有效融合,加强了模型对上下文信息的利用,同时抑制了无关特征对模型的干扰。

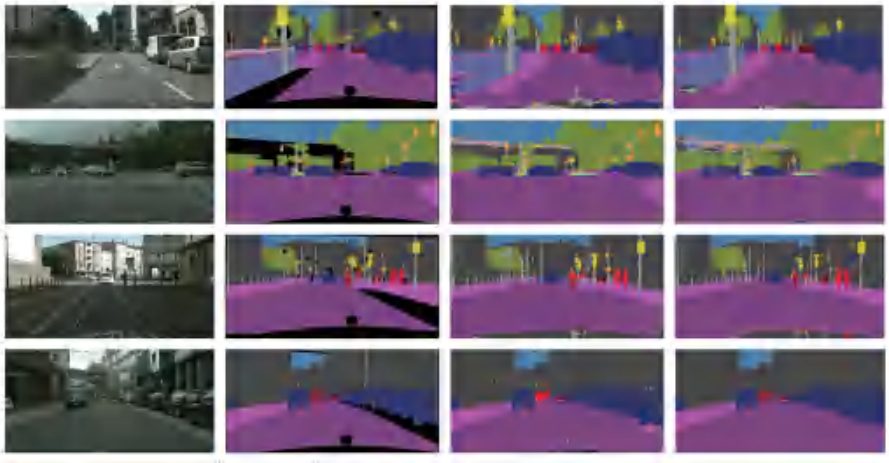
当基准模型同时添加了 EMA 注意力模块和 ECA 注意力模块后,模型的性能得到显著提升,MPA 指数提升至 82.2%,mIoU 指数提升至 74.5%,这是因为添加 EMA 模块后编码器提取的高水平特征更有效保留了全局信息,而 ECA 模块又实现了空间和通道信息的高效融合,使低水平特征图包含了更丰富的低级特征,并剔除了更多无关特征,从而进一步提升了模型的精度。

表 2 消融实验结果

Table 2 Results of ablation experiment %

model	MAP	mIoU
Baseline	78.9	71.2
Baseline+EMA	80.4	73.3
Baseline+ECA	80.0	72.9
Baseline+EMA+ECA	82.2	74.5

使用测试集对模型分割效果进行测试,部分结果如图 4 所示。从左到右依次为原图像、标签、基准模型 DeepLabv3+的分割结果,以及本文所提方法的分割结果。从图 4 中有红色圆圈圈出的部分可以看出本文算法在分割精度方面优势显著。



(a) 原图像 (b) 标签 (c) 基准模型 DeepLabv3+的分割结果 (d) 本文所提方法的分割结果

图 4 分割结果的部分可视化示例

Fig. 4 Partial visual examples of the segmentation results

3　结束语

注意力机制通过为输入数据按规定的方式添加权重的方式，提高了深度神经网络对输入图像特定区域信息的关注。为了解决现有语义分割模型在执行城市街景语义分割任务时面临的难以提取全局信息、捕捉长距离依赖信息和分割边缘不准确的问题，本文提出了一种融合注意力机制的 DeepLabv3+语义分割算法。为了提高模型对输入图像全局信息和长距离依赖信息的捕捉利用，在模型的 ASPP 空间空洞金字塔池化模块后添加了一个 EMA 模块关注全局信息，在解码器获取的低水平特征图后添加了一个 ECA 模块。ECA 模块在关注通道信息的同时嵌入了位置信息，加强了对目标上下文信息和边缘区域的识别，将基准模型的损失函数替换为 Focal Loss 函数以减轻数据集中类别不平衡的影响，提高了模型分割精度。

实验结果表明，本文提出的模型有效提高了分割精度。同时，消融实验证明了本文所添加模块的有效性。虽然本文模型通过添加注意力机制提升了模型对全局信息的把握和边缘信息的利用，但因受限于硬件性能，对原始图片的裁剪幅度稍大，影响了模型的精度。未来将考虑使用更好的硬件设备或更有效的训练方法进一步提升模型精度。

参考文献

[1] 王柳，梁铭炬. 融合深度信息的室内场景分割算法[J]. 计算机系统应用，2024，33(3)：111-117.

[2] 张新钰，高洪波，赵建辉，等. 基于深度学习的自动驾驶技术综述[J]. 清华大学学报(自然科学版)，2018，58(4)：438-444.

[3] 田娟秀，刘国才，谷珊珊，等. 医学图像分析深度学习方法研究与挑战[J]. 自动化学报，2018，44(3)：401-424.

[4] 陶显，侯伟，徐德. 基于深度学习的表面缺陷检测方法综述[J]. 自动化学报，2021，47(5)：1017-1034.

[5] 姜长泓，梁奇强，李秀华，等. 城市街景图像语义分割方法研究[J]. 长春工业大学学报，2022，43(Z1)：447-453.

[6] 刘洪民，刘炜炜. 智慧城市建设理论与实践研究综述[J]. 浙江科技学院学报，2020，32(2)：89-95.

[7] Chen L C，Murphy K，Kokkinos I，et al. DeepLab：semantic image segmentation with deep convolutionalnets，atrous convolution，and fully Connected CRFs[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence，2017，40(4)：834-848.

[8] Zhao H S，Shi J P，Qi X J，et al. Pyramid scene parsing network[C]// Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway，NJ：IEEE，2017：2881-2890.

[9] Wang J，Sun K，Cheng T，et al. Deep high - resolution representation learning for visual recognition [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence，2020，43(10)：3349-3364.

[10] Tan M，Le Q. Efficientnet：rethinking model scaling for convolutional neural networks [J]. arXiv preprint arXiv，1905.11946，2019.

[11] Tan M X，Pang R M，Le Q V. EfficientDet ：scalable and efficient object detection [C1// Proceddings of 2020 IEEE / CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway，NJ：lEEE ，2020：10778-10787.

[12] He K，Zhang X，Ren S，et al. Deep residual learning for image recognition[J]. arXiv preprint arXiv，1512.03385，2016.

[13] 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. Piscataway，NJ：IEEE，2017：2980-2988.

[14] Simon M，Rodner E，Denzler J. ImageNet pre-trained models with batch normalization[J]. arXiv preprint arXiv，1612.01452，2016.

[15] Wang X，Girshick R，Gupta A，et al. Non-local neural networks[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway，NJ：IEEE，2018：7794-7803.

[16] Ouyang D，He S，Zhang J，et al. Efficient multi-scale attention module with cross-spatial learning[C]//Proceedings of the IEEE International Conference on Acoustics，Speech and Signal Processing (ICASSP). Piscataway，NJ：IEEE，2023：1-5.

[17] Hu J，Shen L，Sun G. Squeeze-and-excitation networks[J]. arXiv preprint arXiv，1709.01507，2017.

[18] Woo S，Park J，Lee J Y，et al. CBAM：Convolutional Block Attention Module[J]. arXiv preprint arXiv，1807.06521，2018.

[19] Wang Q，Chen P，Wang X，et al. ECA-Net：Efficient channel attention for deep convolutional neural networks [J]. arXiv preprint arXiv，1910.03151，2019.

[20] 赵永强，饶元，董世鹏，等. 深度学习目标检测方法综述[J]. 中国图象图形学报，2020，25(4)：629-654.

[21] 陈超，齐峰. 卷积神经网络的发展及其在计算机视觉领域中的应用综述[J]. 计算机科学，2019，46(3)：63-73.

[22] Chen L C，Zhu Y，Papandreou G，et al. Encoder-decoder with atrous separable convolution for semantic image segmentation[C]// Proceedings of the European Conference on Computer Vision (ECCV). Cham：Springer，2018：801-818.

[23] Highmore B. Cityscapes：Cultural readings in the material and symbolic city [J]. Journal of Tourism and Cultural Change，2007，5(2)：128-139.

[24] 周华平，邓彬. 融合多层次特征的 deeplabV3+轻量级图像分割算法[J]. 计算机工程与应用，2024，60 (16)：269-275.

[25] Mohan R，Valada A. EfficientPS：Efficient panoptic segmentation [J]. International Journal of Computer Vision，2021，129(5)：1551-1579.