计算机科学
COMPUTER SCIENCE

基于注意力机制与密集邻域预测的轻量化图像语义分割

王国刚, 董志豪

引用本文

王国刚, 董志豪. 基于注意力机制与密集邻域预测的轻量化图像语义分割[J]. 计算机科学, 2024, 51(6A): 230300204-8.

WANG Guogang, DONG Zhihao. Lightweight Image Semantic Segmentation Based on Attention Mechanism and Densely AdjacentPrediction [J]. Computer Science, 2024, 51(6A): 230300204-8.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.230300204

基于注意力机制与密集邻域预测的轻量化图像语义分割

王国刚　董志豪
山西大学物理电子工程学院　太原 030006

摘　要　DeepLabv3+计算复杂度高,空洞空间金字塔池化模块难以突出重要通道特征,解码器生成的高语义化特征图缺乏足够的细节信息。针对上述问题,提出一种基于注意力机制与密集邻域预测的轻量化图像语义分割模型。该模型把 MobileNet V2 作为主干网络,减少了模型参数量;利用通道空洞空间金字塔池化提取多尺度信息,并对特征图的各通道加权,强化重要通道特征的学习;采用密集邻域预测融合高级特征与低级特征,细化分割结果。在 PASCAL VOC 2012 增强数据集上进行实验,结果表明,所提方法的平均交并比和平均像素精确度均高于其他 7 种主流对比算法。与 DeepLabv3+相比,参数量与计算量分别减少 184.82×10^6 和 90.83GFLOPs,该算法在提升分割精度的同时减少了计算开销。

关键词:深度学习;语义分割;DeepLabv3+;注意力机制
中图分类号　TP391

Lightweight Image Semantic Segmentation Based on Attention Mechanism and Densely Adjacent Prediction

WANG Guogang and DONG Zhihao
College of Physics and Electronic Engineering, Shanxi University, Taiyuan 030006, China

Abstract　A novel algorithm named as lightweight image semantic segmentation based on attention mechanism and densely adjacent prediction is proposed to avoid the disadvantages of the difficulty in highlighting important channel features for atrous spatial pyramid pooling module, higher computational complexity and lacking of sufficient detailed information for the high level semantic feature map generated by the decoder in DeepLabv3+ algorithm. The lightweight MobileNetV2 is regarded as the backbone network to reduce model parameters. After the multi-scale information is extracted by the channel atrous spatial pyramid pooling, each channel of the feature map is weighted to reinforce the learning of important channel features. Moreover, the segmentation results are refined since densely adjacent prediction is utilized to combine high-level and low-level features. Experiments are performed on the PASCAL VOC 2012 augmented dataset, and the experimental results show that both mean Intersection over union and mean pixel accuracy of the proposed method are higher than the state-of-the-art algorithms. Compared with DeepLabv3+, the parameters and calculation amount are decreased by 184.82×10^6 and 90.83GFLOPs respectively. The proposed algorithm not only improves the segmentation accuracy, but also reduces the computation cost compared to the baseline algorithm.

Keywords　Deep learning, Semantic segmentation, DeepLabv3+, Attention mechanism

1　引言

语义分割是计算机视觉的一项重要任务,其研究目的是对图像的每个像素点进行分类并分配与之对应的类别标签。它在自动驾驶[1-3]、医学影像[4-6]、3D 重建[7-9]等领域具有广阔的发展前景。

传统图像分割算法包括基于阈值、基于边缘、基于区域以及结合特定理论工具的方法,其分割精度和效率难以满足实际应用的需求。近年来,随着深度学习[10]的兴起,出现了以全卷积神经网络[11](Fully Convolutional Networks, FCN)为代表的语义分割算法。FCN 首次使用卷积层替代全连接层,构建端到端的语义分割网络。

当前流行的语义分割模型大多由 FCN 演变而来,主要分为编码-解码结构和空间金字塔池化模型两种。编码-解码结构模型的主要代表有 SegNet[12]和 U-Net[13],其编码端通过卷积和下采样提取图像特征,解码端经上采样操作恢复空间信息。该类方法一定程度上解决了语义分割中的空间细节丢失问题,但难以获取多尺度上下文信息。相比之下,以 PSPNet[14],DeepLabv2[15],DeepLabv3[16]为代表的金字塔池化模型采用不同膨胀率的空洞卷积[17],获得了丰富的上下文信息,增强了对不同尺度目标的预测能力。该类方法常以池化或带步长的卷积获取上下文信息,因此空间细节信息丢失问题比较严重。

上述方法仅聚合空间上下文信息,难以体现像素之间的语义关联关系。近年来,众多研究工作将注意力机制与卷积神经网络结合,提升了语义分割任务的性能。DANet[18]在语义分割模型中引入 PAM(Position Attention Module)和 CAM

基金项目:国家自然科学基金(11804209);山西省自然科学基金(201901D111031,201901D211173)
This work was supported by the National Natural Science Foundation of China(11804209) and Natural Science Foundation of Shanxi Province, China(201901D111031,201901D211173).
通信作者:王国刚(kingguogang@sxu.edu.cn)

(Channel Attention Module)，利用 PAM 获取特征图任意两个位置特征的相互依赖关系，并通过 CAM 为特征图的各个通道赋予权重值，强化重要通道特征的学习。NLNet[19] 和 DNLNet[20] 采用空间非局部自注意力机制，通过建立像素之间的非局部长距离依赖关系，获得了较好的分割效果，但模型空间复杂度较高。CCNet[21] 仅在每个像素的水平方向和垂直方向上建立与其他像素之间的语义依赖，一定程度上减少了模型参数量。

在 DeepLabv3 的基础上，DeepLabv3＋[22] 添加了一个恢复空间信息的解码器，融合了编码-解码结构和空间金字塔池化两种模型的优势，提升了语义分割效果。但是，Xception[23] 作为 DeepLabv3＋提取特征的主干网络，参数量多，计算开销大；空洞空间金字塔池化(Atrous Spatial Pyramid Pooling，ASPP)模块难以突出重要通道特征；解码器生成的高语义化特征图缺乏足够的图像细节信息。

针对上述问题，提出一种基于注意力机制与密集邻域预测的轻量化图像语义分割算法(Lightweight Image Semantic Segmentation Based on Attention Mechanism and Densely Adjacent Prediction，LAD)。LAD 的编码器将轻量级 MobileNetV2[24] 作为主干网络，降低了模型参数量；利用通道空洞空间金字塔池化提取多尺度信息，并对特征图的各通道加权，强化重要通道特征的学习。解码器利用密集邻域预测[25] 融合高级特征与低级特征，丰富了图像的细节信息，细化了分割结果。实验结果表明，LAD 算法分割精度优于其他 7 种主流对比算法。与基准算法相比，LAD 在提升分割精度的同时减少了计算开销。

2 相关模型

作为一种典型的语义分割网络，DeepLabv3＋在 Deeplabv3 的基础上增加了一个简单而有效的解码器来细化分割结果，其网络架构如图 1 所示。DeepLabv3＋编码器由骨干网络 Xception 和 ASPP 模块组成。Xception 网络提取图像特征；ASPP 模块通过不同膨胀率的并行空洞卷积生成多尺度特征图，并在通道维度上整合、降维后得到高级语义特征图。解码器对编码器输出的特征张量进行 4 倍上采样，再与主干网络 Xception 中对应层级的特征图拼接以丰富图像的语义信息和细节信息，最后通过 3×3 卷积细化特征，并 4 倍上采样得到语义分割图。

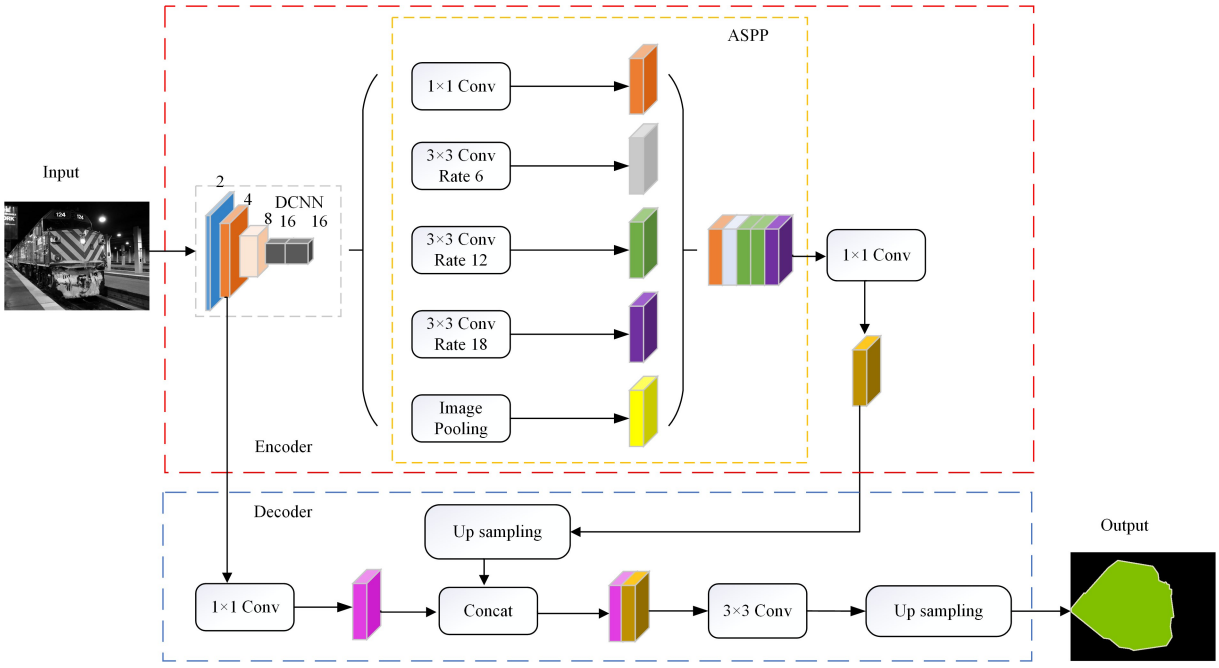

图 1 DeepLabv3＋网络架构

Fig. 1 DeepLabv3＋ network structure

3 基于注意力机制与密集邻域预测的轻量化图像语义分割

3.1 基于 MobileNetV2 网络特征提取

MobileNetV2 是 2018 年由谷歌团队提出的轻量级网络。相比于 MobileNetV1，MobileNetV2 引入具有线性瓶颈层的逆残差结构，如图 2 所示。与 ResNet 残差结构相反，逆残差结构经 1×1 卷积通道升维、3×3 深度可分离卷积提取特征后，使用 1×1 卷积压缩通道，以保证在高维空间提取丰富特征。由于 ReLu 激活函数易增加高维特征的非线性表达，可能会破坏低维特征，MobilenetV2 引入线性瓶颈层(Linear Bottleneck)，在降维的卷积层后利用线性卷积替换原始卷积和 ReLu 函数的组合，以保留更多的特征信息。

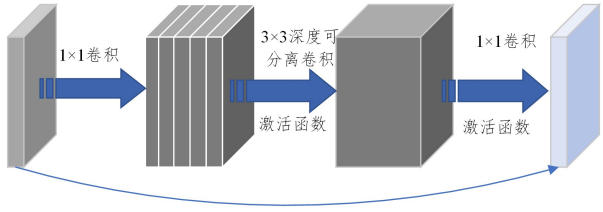

图 2 逆残差结构

Fig. 2 Inverted residual structure

MobileNetV2 特征提取网络，经 5 次下采样得到分辨率为初始图像 1/32 的特征图。由于该特征图尺寸小，含有的特征信息少，LAD 主干网络将 MobileNetV2 第 5 次下采样的卷积步长调整为 1，并保持通道数不变，以得到分辨率为初始图像 1/16 的语义特征图。LAD 主干网络如表 1 所列，其中，

t 和 n 分别表示 bottleneck 扩展因子和重复次数，c 和 s 分别为输出通道数和步长。

表 1　LAD 主干网络结构

Table 1　LAD backbone network structure

Operator	t	c	n	s
Conv2d	—	32	1	2
bottleneck	6	16	1	1
bottleneck	6	24	2	2
bottleneck	6	32	3	2
bottleneck	6	64	4	2
bottleneck	6	96	3	1
bottleneck	6	160	3	1
bottleneck	6	320	1	1

DeepLabv3＋使用 Xception 作为主干网络。Xception 网络层数多，卷积层最多通道数可达 2048 个，且该网络结构复杂，导致系统开销大。相比之下，MobileNetV2 网络卷积层最多通道数仅为 320，模型更加轻量。因此，LAD 使用 MobileNetV2 网络提取图像特征。MobileNetV2 利用深度可分离卷积降低了模型复杂度。深度可分离卷积可分解为深度卷积和逐点卷积，计算量如式(1)－式(4)所示。

$$C_1 = D_K \times D_K \times M \times D_H \times D_H \tag{1}$$

$$C_2 = M \times N \times D_H \times D_H \tag{2}$$

$$C_3 = C_1 + C_2 \tag{3}$$

$$C_4 = D_K \times D_K \times M \times N \times D_H \times D_H \tag{4}$$

其中，D_K 为卷积核大小，M 和 N 分别为输入和输出的通道数，D_H 代表特征图的高度或宽度。C_1，C_2，C_3，C_4 分别表示深度卷积、逐点卷积、深度可分离卷积和标准卷积的计算量。深度可分离卷积与普通卷积计算量之比为：

$$\frac{C_3}{C_4} = \frac{D_K \times D_K \times M \times D_H \times D_H + M \times N \times D_H \times D_H}{D_K \times D_K \times M \times N \times D_H \times D_H} = \frac{1}{N} + \frac{1}{D_K^2} \tag{5}$$

由式(5)可知，LAD 主干网络采用的深度可分离卷积比普通卷积减少约(D_K^2－1)倍的计算量。

3.2　通道空洞空间金字塔池化

ASPP 模块生成的语义特征图各通道权重相同，没有考虑不同通道重要程度的差异，难以凸显富含有用特征的通道信息。为此，LAD 算法构建通道空洞空间金字塔池化模块(Channel Atrous Spatial Pyramid Pooling，CASPP)，如图 3 所示。CASPP 采用全局平均池化、1×1 卷积、膨胀率分别为 6，12，18 的空洞卷积生成多尺度特征图，再利用通道注意力模块(Channel Attention，CA)对拼接后的特征图进行通道加权，强化权值高的通道特征，并抑制权值低的通道特征，以增强不同通道特征间的区分度。

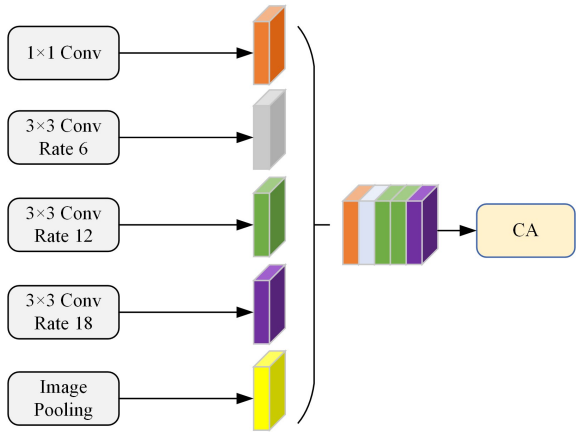

图 3　通道空洞空间金字塔池化模块

Fig. 3　Channel atrous spatial pyramid pooling module

通道注意力机制可为特征图每个通道分配权重，有效捕获通道间的相互依赖性。CASPP 的通道注意力模块 CA 如图 4 所示。

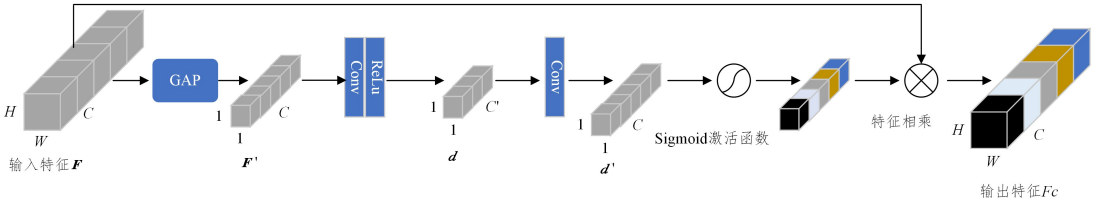

图 4　通道注意力模块

Fig. 4　Channel attention module

给定输入特征图 $\boldsymbol{F} \in R^{H \times W \times C}$ 后，对 $\boldsymbol{F}$ 使用全局平均池化得到每个通道的全局特征，如式(6)所示。

$$\boldsymbol{F}_c' = \frac{1}{WH} \sum_{i=1}^{H} \sum_{j=1}^{W} F_c(i,j) \tag{6}$$

其中，H 和 W 表示特征图的高度以及宽度，$\boldsymbol{F}_c'$ 和 $F_c(i,j)$ 分别表示第 c 个通道的全局特征和(i,j)处的特征值。

$\boldsymbol{F}$ 经全局平均池化得到 $\boldsymbol{F}' \in R^{1 \times 1 \times C}$ 后，采用 1×1 卷积融合通道信息并降维生成特征张量 $\boldsymbol{d} \in R^{1 \times 1 \times C'}$ ($C/C'=8$)，再升维 $\boldsymbol{d}$ 以得到与 $\boldsymbol{F}'$ 相同通道数的 $\boldsymbol{d}' \in R^{1 \times 1 \times C}$。最后利用 Sigmoid 函数生成各通道权重值，并由式(7)得到输出特征 $\boldsymbol{F}_c$。

$$\boldsymbol{F}_c = \boldsymbol{F}_c^{\text{in}} \times \sigma_c(\boldsymbol{d}') \tag{7}$$

其中，σ 代表 Sigmoid 激活函数，$\boldsymbol{F}_c^{\text{in}}$ 和 $\sigma_c(\boldsymbol{d}')$ 分别表示第 c 个通道的输入和权重值。

相比于 ASPP，CASPP 利用 CA 对拼接后的特征图进行通道加权，强化重要通道特征的学习，增强不同类别特征之间的区分度，获得更加丰富的语义信息。

3.3　密集邻域预测

DeepLabv3＋解码器输出的高级特征图具有丰富的语义信息，但缺乏足够的空间信息，难以捕获丰富的图像细节，导致分割精度较低。为此，LAD 利用密集邻域预测模块(Densely Adjacent Prediction，DAP)编码更多的空间信息，在高级特征中嵌入低级特征，使每个像素点的预测结果包含其 $k \times k$ 邻域内像素点的信息，从而提升分割精度。

DAP 模块由卷积、标准化、子像素卷积[26](Pixel Shuffle)、平均池化 4 个模块组成，如图 5 所示。经卷积和标准化操作后，LAD 解码器输出特征图的通道数由 21 个扩展为 $21 \times k^2$ 个(k 取 3)；再利用子像素卷积使特征图的高度与宽度变为原来的 k 倍；最后采用 $k \times k$ 池化核进行平均池化操作以得到语义分割图，如式(8)所示。

$$r_{i,j}=\frac{1}{k\times k}\sum_{0\leqslant l,m<k}x_{i+l-\lfloor k/2\rfloor,j+m-\lfloor k/2\rfloor}^{(l\times k+m)} \tag{8}$$

其中，$r_{i,j}$ 表示位于 (i,j) 像素点处的预测结果，$x_{i,j}^{(c)}$ 表示位于 (i,j) 处第 C 组通道的特征值。

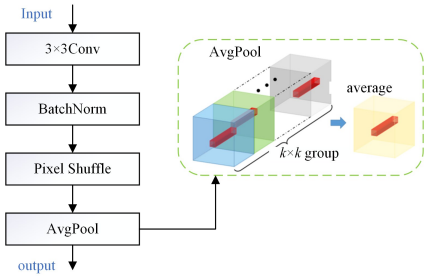

图 5 密集邻域预测模块

Fig. 5 Densely adjacent prediction module

经 DAP 处理后，输出特征图包含了丰富的空间信息和语义信息。每个像素点的预测结果都参考了周围 $k\times k$ 邻域内像素点的空间信息，各像素点不再是独立预测，这有利于得到更加精细准确的分割结果。

3.4 LAD 网络架构

LAD 网络由编码器和解码器组成，如图 6 所示。编码器包括主干网络 MobileNetV2 和通道空洞空间金字塔池化模块(CASPP)。MobileNetV2 用于提取图像特征并降低网络复杂度；CASPP 可对主干网络得到的特征图进行多尺度采样和通道加权，以得到高级语义特征图。解码器将编码器得到的高级特征图与骨干网络对应层级的低级特征图融合，再经 3×3 卷积、上采样和 DAP 操作得到更为精细鲁棒的分割结果。

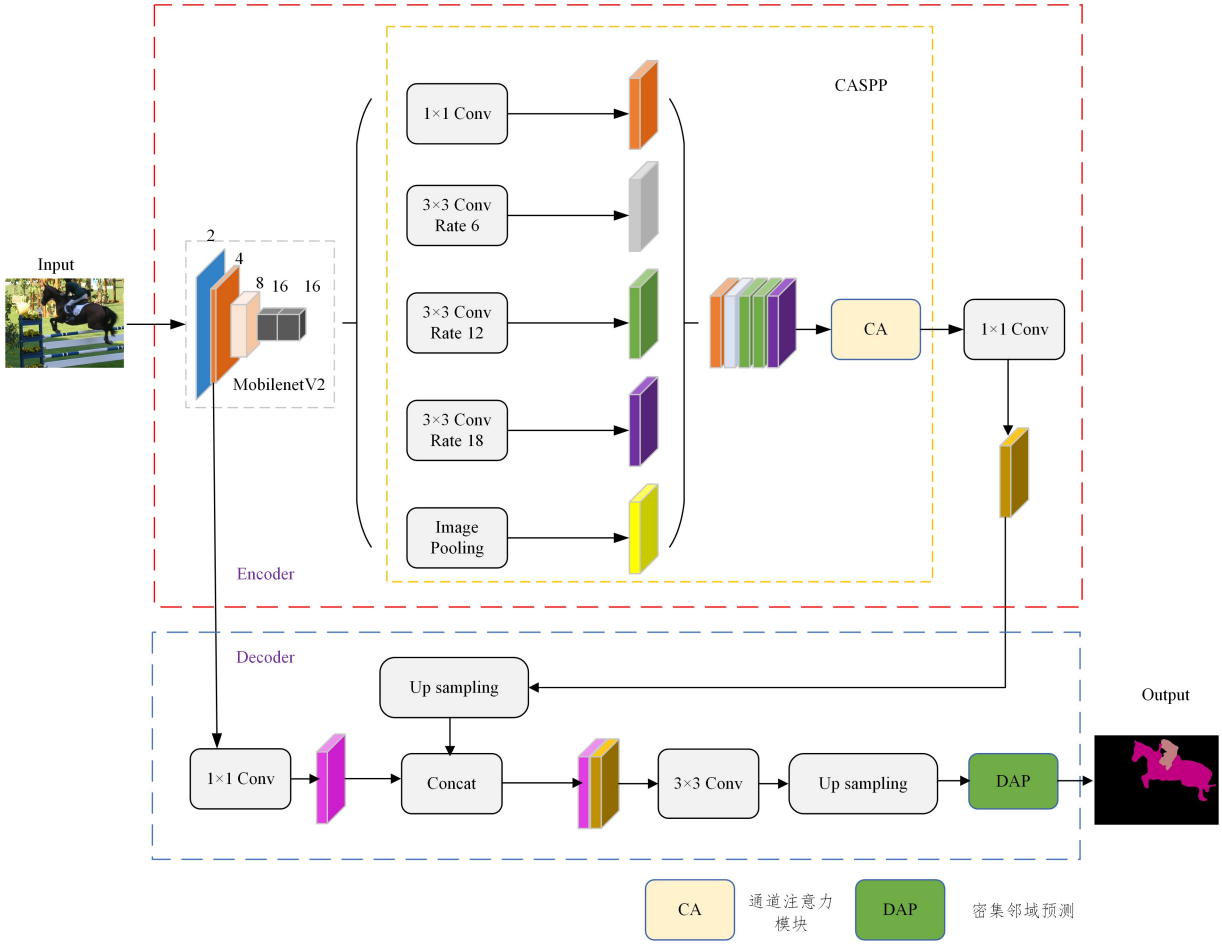

图 6 LAD 网络架构

Fig. 6 LAD network structure

4 实验

4.1 数据集

实验采用 PASCAL VOC 2012 增强数据集评估算法性能。该数据集由 PASCAL VOC 2012 标准数据集和 Semantic Boundaries Dataset(SBD)数据集组成。合并后的数据集包含 21 个语义分类、10582 张训练图像、1449 张验证图像和 1456 张测试图像。

4.2 评价指标

实验采用平均像素精确度(Mean Pixel Accuracy，MPA)、平均交并比(Mean Intersection over Union，MIoU)、浮点计算量 FLOPs(Floating point Operations)和参数量(Parameters)以衡量算法性能。

MPA 表示物体类别像素精度(Pixel Accuracy，PA)的平均值，计算公式如式(9)所示。

$$MPA=\frac{1}{k+1}\sum_{i=0}^{k}\frac{p_{ii}}{\sum_{j=0}^{k}p_{ij}} \tag{9}$$

交并比(Intersection over Union，IoU)是预测结果图和真实标签图交集的测度与并集测度的比值，用来计算分割结果图与真实标记图的相似度。MIoU 为各类别交并比的代数

平均，计算公式如式(10)所示。

$$MIoU=\frac{1}{k+1}\sum_{i=0}^{k}\frac{p_{ii}}{\sum_{j=0}^{k}p_{ij}+\sum_{j=0}^{k}p_{ji}-p_{ii}} \tag{10}$$

其中，$k+1$ 为类别总数，p_{ij} 表示真实类别和预测类别分别为 i，j 的像素数量。

4.3 实验设置

实验环境配置如下：操作系统为 Ubuntu、CPU 为 Intel(R) Xeon(R) Platinum 8255C @ 2.50 GHz、显卡为 11 GB 的 NVIDIA GeForce RTX 2080 Ti。实验采用 Pytorch1.8.1 深度学习框架，通过 Adam 优化器来调整网络参数，使用 Poly 策略动态更新学习率大小。参数设置如表 2 所列。

表 2 参数设置

Table 2 Parameter settings

参数	值
Base Learning rate	0.0002
Power	0.92
Batch size	8
Weight decay	0.0005
Input size	512×512
Epoch	50

在 PASCAL VOC 2012 增强数据集上，LAD 模型损失函数的变化曲线如图 7 所示。由图可见，训练集和验证集的损失值均随迭代次数增加而减小，40 轮迭代后趋于收敛。

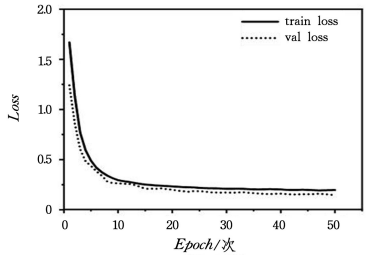

图 7 LAD 模型损失曲线

Fig. 7 Loss curves of LAD module

4.4 实验结果和分析

为验证 LAD 算法性能，在 PASCAL VOC 2012 增强数据集上将其与其他语义分割算法进行对比，包括 FCN，SegNet，PSPNet，GCN[27]，DeepLabv3+，DANet 和 CCNet。8 种算法中，LAD 算法 5 类目标的 *IoU* 值为最优，9 类目标的 *IoU* 值为次优；且 *MIoU* 值达到了 77.51%，与 FCN，SegNet，GCN，PSPNet，DeepLabv3+，DANet，CCNet 相比，LAD 算法 *MIoU* 值分别提升 17.44%，11.36%，2.53%，8.92%，0.54%，1.27%，0.25%。

表 3 LAD 各类别 *IoU* 值与其他算法的对比

Table 3 *IoU* comparison between each class of LAD and other algorithms

(%)

类别	FCN	SegNet	GCN	PSPNet	DeepLabv3+	DANet	CCNet	LAD
background	86.26	90.28	93.59	92.33	93.84	92.01	94.27	**94.33**
aeroplane	69.54	76.81	86.71	82.30	**88.00**	83.59	87.21	86.86
bicycle	29.56	38.53	40.69	35.75	40.82	**48.54**	42.80	45.60
bird	70.3	76.76	85.64	76.28	87.44	83.80	**88.69**	88.47
boat	46.09	41.57	64.43	46.59	67.57	58.20	**71.28**	64.62
bottle	40.04	56.63	62.11	67.87	76.28	**82.24**	72.52	76.69
bus	85.55	90.20	92.66	90.46	**93.96**	92.37	93.49	93.52
car	77.34	83.95	85.39	81.84	**88.202**	85.08	86.02	86.94
cat	79.65	85.58	89.03	86.52	**91.62**	88.2	90.65	91.07
chair	32.25	30.24	42.97	32.14	39.81	48.36	46.17	**48.42**
cow	66.44	76.32	81.39	76.25	**90.68**	74.59	85.54	84.95
dining table	52.71	51.50	64.55	50.26	50.22	**72.36**	65.98	66.86
dog	72.37	80.40	83.62	77.41	87.24	82.83	86.08	**87.45**
horse	56.05	75.06	77.95	77.02	**89.05**	78.06	83.28	79.74
motorbike	52.82	76.14	81.67	75.16	84.29	**85.79**	82.27	83.19
person	73.74	76.31	83.14	79.01	**84.85**	80.77	83.61	83.21
potted plant	33.74	43.31	54.93	54.00	**64.35**	60.39	57.56	52.57
sheep	74.74	57.09	80.50	74.83	**86.65**	82.13	79.65	83.50
sofa	48.14	47.75	63.06	42.96	48.58	64.95	64.60	**65.88**
train	66.84	73.75	88.99	80.51	88.49	87.93	87.96	**89.46**
monitor	50.20	60.92	71.58	60.97	**74.56**	68.98	72.96	74.48
MIoU	60.07	66.15	74.98	68.59	76.97	76.24	77.26	**77.51**

注：加粗、下划线数字分别表示每行最优与次优结果。

表 4 给出了各类别的 *PA* 值。由表 4 可知，8 种算法中，LAD 算法 6 类目标的 *PA* 值为最优，6 类目标的 *PA* 值为次优；且 LAD 算法 *MPA* 值为 89.17%，与 FCN，SegNet，GCN，PSPNet，DeepLabv3+，DANet，CCNet 相比，分别提升了 20.93%，10.54%，1.46%，9.43%，2.72%，1.85%，0.72%。

表 5 给出了 8 种算法在客观性评价指标上的结果。由表 5 可知，相比于 DeepLabv3+基准算法，LAD 算法的 *MIoU* 和 *MPA* 值分别提升 0.54%和 2.71%，且参数量、计算量分别减少 184.82×10^6 和 90.83 GFLOPs。在分割精度方面，LAD 算法 *MIoU* 与 *MPA* 值均高于其他 8 种对比算法。由表 5 还可看出，尽管 PSPNet 参数量与计算量最少，但是 LAD 的分割精度明显高于 PSPNet。具体说来，LAD 的 *MIoU* 和 *MPA* 值比 PSPNet 分别提升了 8.32%和 9.43%。此外，CCNet 与 DANet 分割精度略低于 LAD，但参数量与计算量明显高于 LAD。综上可知，LAD 算法在保持高精度的同时，降低了计算复杂度，语义分割性能最优。

在 PASCAL VOC 2012 增强数据集中选取 5 幅图像，并与其他算法进行定性对比，实验结果如图 8 所示。

表 4 LAD 各类别 *PA* 值与其他算法的对比

Table 4 *PA* comparison between each class of LAD and other algorithms

类别	FCN	SegNet	GCN	PSPNet	DeepLabv3+	DANet	CCNet	LAD
background	95.70	95.32	96.18	96.45	**97.13**	95.02	94.06	96.12
aeroplane	81.35	93.15	94.01	93.05	**97.02**	96.01	95.94	94.72
bicycle	70.69	80.67	89.45	81.53	**92.51**	82.45	80.36	88.40
bird	77.69	86.58	**94.69**	87.49	93.87	92.87	93.61	93.50
boat	64.88	63.59	86.02	64.50	83.81	85.48	87.49	**89.48**
bottle	59.65	81.21	**92.52**	80.79	90.13	90.57	91.36	92.18
bus	76.63	93.42	98.43	93.07	97.24	96.78	98.47	**98.51**
car	75.10	89.56	92.50	90.31	92.22	93.12	**93.56**	91.99
cat	85.02	91.47	93.02	92.26	**97.20**	94.08	94.72	94.87
chair	32.91	41.22	65.45	44.03	50.47	65.21	**68.74**	66.07
cow	48.44	84.23	85.78	87.43	95.56	94.28	**96.22**	90.74
dining table	53.75	53.06	71.56	53.48	54.56	68.24	**79.19**	75.50
dog	70.88	90.32	93.16	90.61	**94.00**	91.60	92.85	92.61
horse	72.59	84.13	**94.95**	85.52	93.82	93.89	94.17	94.23
motorbike	78.33	84.52	90.72	87.71	**94.86**	94.57	94.22	91.10
person	87.10	86.53	89.55	88.83	**91.54**	87.85	86.70	90.29
potted plant	49.25	68.12	72.40	71.06	78.09	77.12	**78.44**	73.23
sheep	66.51	86.32	93.71	79.96	90.04	93.56	92.73	**95.77**
sofa	37.35	54.23	73.15	51.76	58.69	65.23	63.61	**79.56**
train	76.82	89.18	93.70	90.28	92.30	94.21	95.36	**95.41**
monitor	72.55	54.32	81.03	64.53	80.37	81.65	85.78	**88.30**
MIoU	68.24	78.63	87.71	79.74	86.45	87.32	88.45	**89.17**

注：加粗、下划线数字分别表示每行最优与次优结果。

表 5 8 种算法客观评价指标的对比

Table 5 Comparison of objective evaluation indices of eight algorithms

Models	Backbone Networks	*MIoU*/%	*MPA*/%	参数量	GFLOPs
FCN	VGG-16	60.07	68.24	77.20×10^6	237.63
SegNet	VGG-16	66.15	78.63	112.36×10^6	327.09
PSPNet	MobileNetV2	68.59	79.74	$\mathbf{9.20\times10^6}$	**6.11**
GCN	ResNet152	74.98	87.71	231.64×10^6	142.53
DeepLabv3+	Xception	76.97	86.45	208.72×10^6	167.00
CCNet	ResNet50	77.26	88.45	182.23×10^6	440.24
DANet	ResNet50	76.24	87.32	178.37×10^6	1025.00
LAD	MobileNetV2	**77.51**	**89.17**	23.90×10^6	76.17

注：加粗、下划线数字分别表示每列最优与次优结果。

待分割原图 8(a)由人物、桌子和背景组成。由图可知，SegNet 未能识别出桌子类别，FCN 和 PSPNet 对桌子类别存在错误分割现象，GCN 和 DeepLabv3+将人的手部错误分割成桌子的一部分。只有 LAD 算法可以正确分割桌子与人物，没有漏分割、错误分割问题，分割结果最优。

待分割原图 8(b)由鸟类和背景组成。由图可知，FCN 分割结果不连续，鸟的翅膀与躯干分离。SegNet 对鸟的翅膀边缘分割不完整，PSPNet 在鸟的轮廓处分割比较粗糙。GCN，DeepLabv3+，LAD 均能分割出鸟和背景，但在一些细节方面，如鸟的尾部，分割结果更加精细。

待分割原图 8(c)由沙发和背景组成。对比发现，FCN，SegNet，PSPNet，GCN，DeepLabv3+，LAD 均能正确分割出物体类别，但在沙发左侧边缘以及轮廓处，LAD 的分割结果更接近于标签图。

待分割原图 8(d)由 3 个瓶子和背景组成。由图可知，DeepLabv3+没有分割出中间的瓶子，FCN，SegNet，GCN 在瓶子边缘处存在错误分割现象，PSPNet 对瓶子边缘分割比较粗糙。相比之下，LAD 没有漏分割、错误分割现象，且分割结果更准确。

待分割原图 8(e)由自行车和背景组成。由图可以看出，FCN，SegNet，GCN，PSPNet，DeepLabv3+存在分割不均衡问题，对车轮部分分割较为模糊。相比之下，LAD 对车轮与车座的分割更加细化，分割结果优于其他算法。

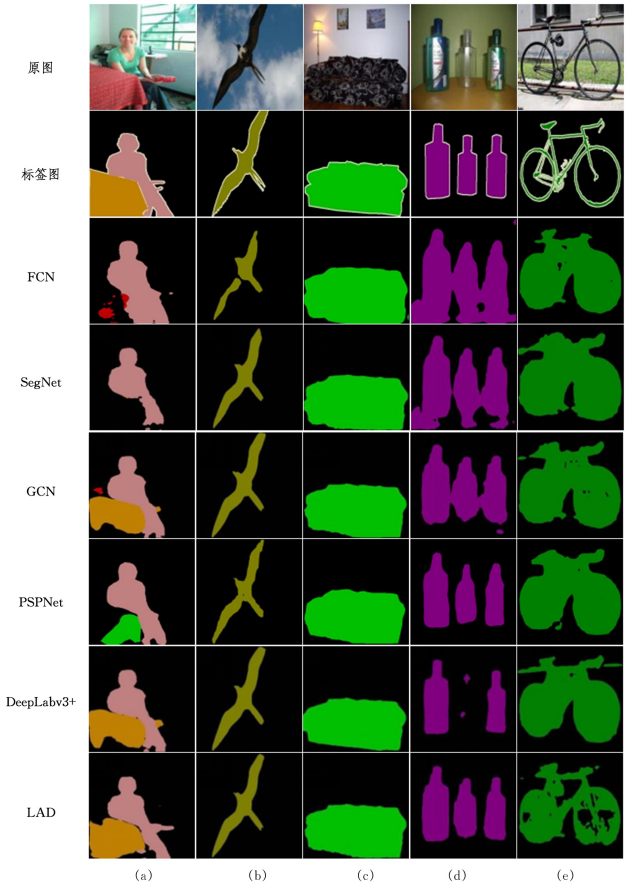

图 8 不同算法在 PASCAL VOC 2012 增强数据集上的结果

Fig. 8 Results of different algorithms on PASCAL VOC 2012 augmented dataset

4.5 消融实验

为验证 CA 模块中不同池化方式以及通道压缩比例(C/C')对语义分割性能的影响，设置了 4 组对比实验，实验结果如表 6 所列。由表中第 1，2 行可以发现，使用全局平均池化所得 *MIoU* 值比全局最大池化高 0.48%，这是由于网络深层的高级语义信息有助于像素的分类，而使用最大池化会损失部分语义信息，因此采用平均池化有利于保持语义分割的类别一致性，分割效果更佳。此外，对比表中第 2－4 行可以发现，通道压缩比例C/C'值为 16 和 32 时，模型的 *MIoU* 值分别下降 0.14%和 0.47%，参数量分别减少 0.79×10^6 和 1.18×10^6，这表明通道压缩比例较大时，通道压缩导致模型参数量减少，但也会导致部分通道信息丢失，造成分割精度下降。因此，选取全局平均池化，且通道压缩比例C/C'为 8 时，分割性能最优。

表 6 CA 模块消融实验

Table 6 Ablation experiment results of CA module

Group	*MIoU*/%	参数量
全局最大池化，$C/C'=8$	77.03	23.90×10^6
全局平均池化，$C/C'=8$	77.51	23.90×10^6
全局平均池化，$C/C'=16$	77.37	23.11×10^6
全局平均池化，$C/C'=32$	77.04	22.72×10^6

为比较 DAP 模块中不同邻域大小(k)对分割性能的影响，进行 4 组对比实验，实验结果如表 7 所列。由表可知，当 k 取 3 时 *MIoU* 值为 77.51，分割性能最优。此外，从表中还可发现，模型参数量随 k 值变大而逐渐增加，当 k 分别取 4 和 5 时，与 k 取 3 时相比，*MIoU* 值分别降低 5.42%和 6.57%，这表明 k 值过大会导致某像素点的预测结果参考周围过多冗余像素信息，不利于分割效果的提升。

表 7 DAP 模块消融实验

Table7 Ablation experiment results of DAP module

Group	*MIoU*/%	参数量
$k=2$	76.29	23.82×10^6
$k=3$	77.51	23.90×10^6
$k=4$	72.09	24.00×10^6
$k=5$	70.94	24.14×10^6

为验证 LAD 算法中 MobileNetV2 轻量化网络、CASPP、DAP 方案的有效性，利用控制变量法在 PASCAL VOC 2012 增强数据集上进行逐层消融实验，实验结果如表 8 所列。表中第 1 行是主干网络为 Xception 的 DeepLab v3＋算法，第 4 行为本文提出的 LAD 算法。

表 8 消融实验结果

Table8 Ablation experiment results

Group	MobileNetV2	CASPP	DAP	*MIoU*/%	*MPA*/%	参数量
1				76.97	86.45	208.72×10^6
2	√			74.59	84.01	**22.19×10^6**
3	√	√		76.65	87.64	23.76×10^6
4	√	√	√	**77.51**	**89.16**	**23.90×10^6**

注：加粗数字表示最优结果。

由表中第 1，2 行可以看出，引入 MobileNetV2 后，*MIoU* 和 *MPA* 分别下降 2.38%和 2.44%，参数量减少 186.53×10^6。这表明 MobileNetV2 方案以微小精度为代价，换取了模型参数量近 90%的减少。由表中第 2，3 行可以看出，引入 CASPP 之后，*MIoU* 和 *MPA* 分别提升 2.06%和 3.36%，参数量增加 1.57×10^6。这是由于 CASPP 能捕获多尺度信息，强化重要特征通道的学习。由表中第 3，4 行可以看出，引入 DAP 之后，算法的 *MIoU* 和 *MPA* 值分别提升 0.86%和 1.52%，参数量增加 0.14×10^6。这是因为 DAP 能在高级特征中融入低级特征，获得更多图像细节信息。

综上可知，相比于基准算法 DeepLabv3＋，本文提出的 LAD 算法利用 MobileNetV2 网络提取特征，虽损失了部分分割精度，但大幅减少了模型参数量。在此基础上，LAD 引入的 CASPP 与 DAP 弥补了分割精度的损失，且没有增加过多参数量。因此，本文所提 LAD 算法在提升分割精度的同时减少了计算开销。

结束语 针对 DeepLabv3＋参数量多、ASPP 模块难以突出重要通道特征、解码器生成的特征图缺乏足够的细节信息的问题，提出了 LAD 网络模型。该模型的编码器将 MobileNetV2 作为主干网络，减少了模型参数量；利用通道空洞空间金字塔池化提取多尺度信息，并对特征图的通道加权，强化重要通道特征的学习；解码器利用密集邻域预测融合高级特征与低级特征，丰富了图像的细节信息，细化了分割结果。实验结果表明，LAD 的 *MIoU* 与 *MPA* 值均高于其他 7 种对比算法；与基准模型相比，LAD 在提升分割精度的同时减少了计算开销。然而，LAD 在分割包含目标类别较多、存在遮挡的图像时分割效果较差，且分割效率难以满足实时性要求，后续将从以上角度出发，进一步研究分割精度与效率的高性能网络。

参 考 文 献

[1] CAI Y F，DAI L，WANG H，et al. Multi-Target Pan-Class Intrinsic Relevance Driven Model for Improving Semantic Segmentation in Autonomous Driving[J]. IEEE Transactions on Image Processing，2021，30：9069-9084.

[2] ZHOU W，BERRIO J S，WORRAL S，et al. Automated Evaluation of Semantic Segmentation Robustness for Autonomous Driving[J]. IEEE Transactions on Intelligent Transportation Systems，2020，21(5)：1951-1963.

[3] YI D W，FANG H，HUA Y N，et al. Improving Synthetic to Realistic Semantic Segmentation With Parallel Generative Ensembles for Autonomous Urban Driving[J]. IEEE Transactions on Cognitive and Developmental Systems，2022，14(4)：1496-1506.

[4] YIN P S，YUAN R，CHENG Y M，et al. Deep Guidance Network for Biomedical Image Segmentation[J]. IEEE Access，2020，8：116106-116116.

[5] ZHANG M，LI X，XU M J，et al. Automated Semantic Segmentation of Red Blood Cells for Sickle Cell Disease[J]. IEEE Journal of Biomedical and Health Informatics，2020，24(11)：3095-3102.

[6] GAO Z J，HE Y，LI Y. A Novel Lightweight Swin-Unet Network for Semantic Segmentation of COVID-19 Lesion in CT Images[J]. IEEE Access，2023，11：950-962.

[7] IBRAHIM M，AKHTAR N，WISE M，et al. Annotation Tool and Urban Dataset for 3D Point Cloud Semantic Segmentation[J]. IEEE Access，2021，9：35984-35996.

[8] SHI W J，XU J W，ZHU D C，et al. RGB-D Semantic Segmentation and Label-Oriented Voxelgrid Fusion for Accurate 3D Semantic Mapping[J]. IEEE Transactions on Circuits and Systems for Video Technology，2022，32(1)：183-197.

[9] YANG J F，ZHOU B C，QIU H D，et al. MLFNet-Point Cloud

Semantic Segmentation Convolution Network Based on Multi-Scale Feature Fusion[J]. IEEE Access, 2021, 9: 44950-44962.

[10] MINAEE S, BOYKOV Y, PORIKLI F, et al. Image Segmentation Using Deep Learning: A Survey[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2022, 44(7): 3523-3542.

[11] LONG J, SHELHAMER E, DARRELL T, et al. Fully convolutional networks for semantic segmentation[C]// 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Boston, MA, USA, 2015: 3431-3440.

[12] 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.

[13] RONNEBERGER O, FISCHER P, BROX T. U-Net: Convolutional Networks for Biomedical Image Segmentation[C]// International Conference on Medical image computing and computer-assisted intervention. Cham: Springer, 2015: 234-241.

[14] ZHAO H S, SHI J P, QI X J, et al. Pyramid Scene Parsing Network[C]// 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Honolulu, HI, USA, 2017: 6230-6239.

[15] 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.

[16] CHEN L C, PAPANDREOU G, SCHROFF F, et al. Rethinking atrous convolution for semantic image segmentation[EB/OL]. (2017-12-05) [2023-02-24]. https://arxiv.org/abs/1706.05587.

[17] 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) [2023-02-24]. https://arxiv.org/abs/1412.7062.

[18] FU J, LIU J, TIAN H J, et al. Dual Attention Network for Scene Segmentation[C]// 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Long Beach, CA, USA, 2019: 3141-3149.

[19] WANG X L, GIRSHICK R, GUPTA A, et al. Non-local neural networks[C]// 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, UT, USA, 2018: 7794-803.

[20] YIN M L, YAO Z H, CAO Y, et al. Disentangled non-local neural networks[C]// Proceedings of the European Conference on Computer Vision. 2020: 191-207.

[21] HUANG Z L, WANG X G, WEI C C, et al. CCNet: Criss-Cross Attention for Semantic Segmentation[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2023, 45(6): 6896-6908.

[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). 2018: 801-818.

[23] CHOLLET F. Xception: Deep Learning with Depthwise Separable Convolutions[C]// 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Honolulu, HI, USA, 2017: 1800-1807.

[24] SANDLER M, HOWARD A, ZHU M, et al. MobileNetV2: Inverted Residuals and Linear Bottlenecks[C]// 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, UT, USA, 2018: 4510-4520.

[25] ZHANG Z L, ZHANG X Y, PENG C, et al. ExFuse: Enhancing Feature Fusion for Semantic Segmentation[C]// European Conference on Computer Vision. Charm: Springer, 2018: 273-288.

[26] SHI W, CABALLERO J, HUSZAR F, et al. Real-Time Single Image and Video Super-Resolution Using an Efficient Sub-Pixel Convolutional Neural Network[C]// 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Las Vegas, NV, USA, 2016: 1874-1883.

[27] PENG C, ZHANG X Y, YU G, et al. Large Kernel Matters—Improve Semantic Segmentation by Global ConvolutionalNetwork[C]// 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Honolulu, HI, USA, 2017: 1743-1751.

WANG Guogang, born in 1977, Ph. D, associate professor, is a member of CCF (No. K7194M). His main research interests include the image processing and computer vision, and artificial intelligence.