基于双流加权 Gabor 卷积网络融合的室内 RGB-D 图像语义分割

王旭初[1,2*]，刘辉煌[2]，牛彦敏[3]

[1]重庆大学光电技术及系统教育部重点实验室，重庆 400040；

[2]重庆大学光电工程学院，重庆 400040；

[3]重庆师范大学计算机与信息科学学院，重庆 401331

摘要 针对室内场景下光照变化、物体相互遮挡以及类别复杂等问题，提出了一种基于双流加权 Gabor 卷积网络融合的彩色-深度(RGB-D)图像语义分割方法。为了获得方向和尺度不变性特征，设计了一种加权 Gabor 方向滤波器用于构建深度卷积网络(DCN)，提取对方向和尺度变化具有适应性的特征信息。为了构建轻量级特征提取网络，采用宽残差-加权 Gabor 卷积网络分别提取彩色和深度双流图像特征，并利用金字塔池化模块对提取的深度特征进行多尺度融合以丰富图像上下文信息。对所提语义分割方法在 NYUDv2 数据集上进行实验，分别设置不同的对比方法。结果表明所提方法具有合理性和有效性，并在分割效果上具有一定的竞争性。

关键词 图像处理；语义分割；加权 Gabor 卷积网络；宽残差模块；多尺度特征融合；室内 RGB-D 图像

中图分类号 TP391.41 **文献标志码** A **doi**：10.3788/AOS202040.1910001

Indoor RGB-D Image Semantic Segmentation Based on Dual-Stream Weighted Gabor Convolutional Network Fusion

Wang Xuchu[1,2*], Liu Huihuang[2], Niu Yanmin[3]

[1] *Key Laboratory of Optoelectronic Technology and Systems of Ministry of Education, Chongqing University, Chongqing* 400040, *China*;

[2] *College of Optoelectronic Engineering, Chongqing University, Chongqing* 400040, *China*;

[3] *College of Computer and Information Science, Chongqing Normal University, Chongqing* 401331, *China*

Abstract To handle the problems of illumination change, mutual occlusion of objects, and complicated semantic categories in indoor scenes, a color-depth (RGB-D) image semantic segmentation method based on the dual-stream weighted Gabor convolutional network fusion is proposed in this work. In order to obtain direction and scale invariant features, a weighted Gabor direction filter is designed to construct a deep convolution network (DCN) to extract feature information that is adaptive to direction and scale changes. In order to build a lightweight feature extraction network, a wide residual weighted Gabor convolutional network module is used to extract color and depth dual-stream image features, and a pyramid pooling module is used to fuse the extracted depth features to enrich the image context information. The proposed semantic segmentation method is tested on NYUDv2 dataset, and different comparison methods are set up. The results show that the proposed method is reasonable and effective, and the segmentation effect is competitive.

Key words image processing; semantic segmentation; weighted Gabor convolution network; wide residual module; multiscale feature fusion; indoor RGB-D image

OCIS codes 100.2960; 330.1400; 150.0155; 200.4260

1 引 言

室内图像语义分割是场景理解和解译的重要步骤，对室内机器人移动定位与环境交互、安防监控领域的事件检测等起到关键作用。由于室内场景具有背景复杂、光照不均、低层视觉特征辨识力弱以及物

收稿日期：2020-04-26；**修回日期**：2020-06-08；**录用日期**：2020-06-19

基金项目：重庆市基础与前沿研究计划(cstc2016jcyjA0317)

* **E-mail**：xcwang@cqu.edu.cn

体之间存在大量遮挡的问题，要准确识别图像中内容及其位置往往面临巨大的挑战。

传统语义分割方法一般利用分类器对人工特征进行像素级分类，并采用条件随机场（CRF）进行精细化处理。但是，分类器的设计一般针对单一类别，分类器用于多类别分割任务时具有较大的训练难度，且计算复杂度高。此外，手动设计的特征具有较大局限性，导致模型泛化能力不强，分割精度不高。目前基于深度学习的语义分割方法已展示出较大的优势，不仅能实现多类别分割，还可以进行端到端训练，其类型可大致分为基于编解码、结合CRF、彩色-深度（RGB-D）图像融合的分割框架三类。

基于编解码的分割方法分为编码和解码两部分，编码部分用于提取特征，而解码部分则用于逐渐恢复丢失的空间信息，包括U-Net[1]、全卷积神经网络（FCN）[2]、SegNet[3]以及DeepLabv3＋[4]等。Noh等[5]将语义分割问题视为实例分割问题，提出一种深度反卷积网络DeconvNet。该方法首先进行逐像素类别标签识别，预测分割掩码，然后将其送入网络中，通过训练得到分割结果的组合，因此该方法能处理不同尺度的物体，增强对图像细节的处理。但是，目标候选框提取需要耗费大量的时间和存储空间，且过程比较复杂，难以实现快速准确的分割。Liu等[6]将全局上下文信息加入全卷积网络中，提出一种ParseNet网络，用网络中任意一层平均特征表示每个位置的特征，用于捕获图像中的全局语义信息，提高模型的分割性能。张哲晗等[7]提出一种改进的对称编码-解码网络SegProNet，利用池化索引与卷积融合语义信息及图像特征，构建端到端的语义分割网络，该方法在CCF卫星数据集上展示了较好的分割性能。Yu等[8]为了解决全卷积网络中上采样操作无法弥补损失信息的问题，采用空洞卷积实现多层次上下文聚合，特征图像分辨率未降低，该方法在FCN基础上提升了性能。吴止锾等[9]提出一个考虑类别不均衡的FCN模型，用于高分辨率遥感图像的语义分割，为了提高小类预测的准确率，采用加权交叉熵损失函数和自适应阈值方法，实现端到端的精确分类。为了尽可能恢复下采样过程中损失的特征信息，研究人员提出了反卷积和空洞卷积等方法，但是利用反卷积无法恢复低层特征，空洞卷积会消耗很大的计算和存储空间，二者均不利于精确快速的语义分割。Lin等[10]提出一种多阶段提炼网络RefineNet以融合不同分辨率的特征图，通过大范围残差连接将下采样过程丢失的信息进行融合。此外，采用残差连接和恒等映射实现端到端训练，可使各层特征具有不同针对性，进而提升分割效果。例如，胡涛等[11]提出的极化合成孔径雷达图像语义分割方法融合了各层深度特征。Wang等[12]设计了密集上采样卷积（DUC）和混合空洞卷积结构（HDC），前者主要用来产生像素级预测，捕获和解码双线性上采样中缺失的信息，后者能扩大网络感受野以聚合全局信息，避免空洞卷积运算所导致的网格问题。虽然基于编解码的分割方法能恢复损失的部分信息，但是一般情况下无法将丢失的信息完全恢复，因此最终分割结果仍然不够理想。

基于CRF的分割方法将CRF集成到深度网络，通过产生与图像视觉特征一致的结构化输出来改善分割结果。Zheng等[13]将CRF引入网络模型后端以形成一种新的分割框架CRFasRNN，该方法可改善卷积核感受野过大和池化造成像素分类粗糙的问题。Lin等[14]提出的CRF语义分割框架结合了图像区域之间“碎片-碎片”上下文和“碎片-背景”上下文，采用分段训练方式，引入多尺度输入和滑动空间金字塔来避免反向传播过程中重复的、代价高昂的CRF推理。上采样操作使得图像局部信息丢失，不易获得更为精细化的结果。传统CRF仅考虑相邻像素之间的类别关联性，而全连接CRF考虑了任意两个像素之间的类别关联性以获得平滑结果。不过，这些方法并未采用更高阶的势能，在精细化语义分割方面还存在一些不足。Arnab等[15]认为高阶势函数能显著提高分割性能，并证明了基于目标检测和超像素的高阶势能可以包含于深层网络的CRF，通过设计高阶微分算子的位势函数并使用可微平均场算法进行推理，以实现较好的细分性能。

上述方法大多采用反卷积网络预测每个像素的类别，并利用CRF进行后端处理，这限制了物体边缘的分割精度。近年来，结合彩色和深度图像的语义分割也借鉴类似思路，采用相等权值并融合颜色和深度信息进行分割，由于没有考虑颜色和深度信息对于不同场景、不同类别表达能力的差异，该方法对于复杂室内场景的分割效果欠佳。Ren等[16]从特征构造层面出发，提出结合树状模型与马尔可夫随机场（MRF）概率图模型的方法，该方法利用超像素区域层次化结构构建树状模型，然后生成像素级特征描述子，最后通过线性支持向量机进行分类；由于传统特征在信息表达方面存在不足，该方法相较于深度学习方法具有一定的局限性。Silberman等[17]从室内场景语义标注结果推断结构类别间的

支撑关系，将标注的目标类别转换为4种结构类别。虽然构建标注类别与结构类别的关系具有一定的现实应用价值，但是无法显著提升语义分割性能。He等[18]提出一种基于超像素的多视点CNN语义分割方法，该方法利用同一场景中的其他视图信息，辅助实现图像高质量语义分割。此外，该方法构造一种新颖的时空池化层用于聚合信息，能够通过时空数据驱动池化层，进行多视图聚合，提升语义分割精度。Cheng等[19]提出一种局部感知反卷积网络，以门的方式融合彩色图像和深度图像的上下文信息，从而提高边界分割的准确率，该网络通过灵活学习边界信息和门式融合来提高精度，但需要采用人工特征进行过分割预处理。Yurdakul和Yemez[20]通过研究深度及时间信息对卷积和递归神经网络架构的视频分割任务的贡献，将深度信息加入到语义分割框架中，以实现高质量语义分割。Hu等[21]针对彩色(RGB)和深度(depth)图像的特征分布在不同场景中具有较大差异的问题，提出一种注意力互补网络ACNet，选择性地从RGB分支和深度分支提取特征。具体地，他们所提出的注意力互补模块(ACM)主要基于通道注意力机制，从RGB和深度分支中提取加权特征，在不同信息流中挖掘更多高质量的特征信息。Lin等[22]提出一种切换上下文的网络，将深度通道用于图像区域中的目标辨识，通过分析区域特性来选择网络分支并进行上下文表示，语义分割结果具有更好的一致性，使用梯度等底层特征产生超像素的方法限制了弱结构区域的分割效果。

总体而言，结合RGB和depth图像并利用深度通道可克服光照影响，可为室内场景语义分割提供更为丰富的边缘和空间信息。但是，目前大多数语义分割方法要么忽略对深度信息的利用，要么所采用的深度信息较为单一，仅用于构造区域级特征，不考虑利用深度通道进行上下文推断。而且，用于训练的滤波器缺乏对室内目标方向、尺度等先验信息的利用，增加了网络训练的复杂度。

对此，本文提出一种基于双流Gabor卷积网络融合的室内RGB-D场景语义分割方法，通过构造可学习的宽残差加权Gabor卷积网络，分别提取RGB和depth图像的双流特征，然后采用金字塔池化分别对双流深度特征进行多尺度融合，利用不同尺度特征来缓解物体差异性问题。将融合后的不同尺度特征进行上采样，将以解码方式级联的特征输入到Softmax进行语义分类。

传统的卷积神经网络无法提取图像中的旋转不变性特征，导致提取到的特征信息表达不足。Gabor滤波器作为一种多尺度多方向的特征提取工具，在频率表达上具有优良特性，很适用于提取图像中的纹理等信息，且提取的特征具有尺度和旋转不变性。基于此，本文提出一种可学习的加权Gabor方向滤波器，替代传统卷积滤波器对网络模型进行训练，提取图像中方向和尺度不变性特征，增加特征的表达能力。此外，所提方法还能减少网络的参数量，加快网络训练速度。语义分割方法一般需要构建较深的网络来提取图像中丰富的语义以及上下文信息，但是网络层数越多意味着模型训练越复杂，容易产生过拟合，因此构建轻量级的语义分割网络尤为必要。本文采用宽残差模块(WRB)构建语义分割网络模型，实现网络轻量化的同时获得较好的深度特征，从而可以提高模型的训练效率。图像中物体的多尺度问题一直是语义分割中的难点，导致目前方法难以实现对小目标的准确分割。对此，本文通过构建金字塔池化模块(PPM)来对RGB和depth图像特征进行多尺度融合，以适应图像中不同尺度的物体，增强网络模型的分割性能。

2　本文方法

2.1　方法概述

本文模型总体结构如图1所示，其核心在于基于加权Gabor方向滤波器进行网络中的卷积操作，并将特征融合与编解码框架相结合，从而提升模型对特征的抽象表达能力。具体而言，本文模型以RGB和depth图像作为输入，分别通过新型宽残差Gabor卷积网络提取图像深度特征，然后采用金字塔池化模块分别对RGB和depth图像特征进行池化，得到不同尺度的双流特征，以缓解物体差异性的问题。进一步地，将不同尺度RGB和depth图像特征进行级联融合，其中 $f_1^{\mathrm{r}}, f_2^{\mathrm{r}}, \cdots, f_n^{\mathrm{r}}$ 表示不同尺度RGB图像的深度特征，$f_1^{\mathrm{d}}, f_2^{\mathrm{d}}, \cdots, f_n^{\mathrm{d}}$ 表示不同尺度depth图像的深度特征，n 表示金字塔池化中的内核个数。最后，将融合后的不同尺度特征进行上采样，以解码形式进行特征级联，得到包含不同尺度的融合特征，并将其送入Softmax进行分类。其中，宽残差加权Gabor卷积网络可从RGB图像中提取纹理和颜色等特征，可从depth图像中提取边缘和空间轮廓信息，且可以采用轻量级网络提取深度特征。

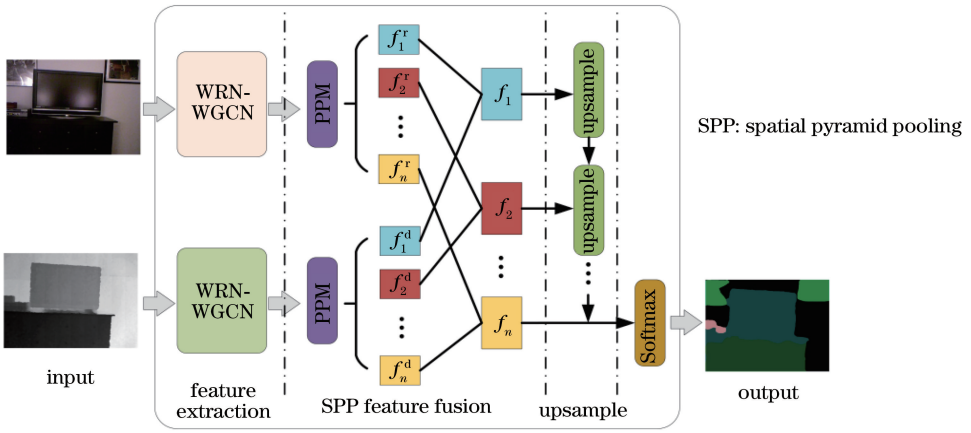

图 1　双流加权 Gabor 卷积网络融合的 RGB-D 图像语义分割

Fig. 1　RGB-D image semantic segmentation by double-stream weighted Gabor convolution network fusion

2.2　加权 Gabor 方向滤波器

深度卷积神经网络提取的特征难以适应方向和尺度变化，自身参数无法根据特征差异性自动调节，而且训练时间较长、空间复杂度较高。传统滤波器在特征提取方面具有一定优势，通过对图像进行针对性处理，提取对空间变换具有不变性质的特征，特征冗余度往往小于深度卷积神经网络。Gabor 滤波器[23]作为一种与简单细胞视觉刺激响应非常相似的滤波器，在提取目标的局部空频信息方面具有良好特性。从可视化结果来看，Gabor 滤波器类似于卷积神经网络的浅层滤波器。

本文在 Luan 等[24]构造的 Gabor 方向滤波器基础上进行改进，提出一种加权 Gabor 方向滤波器，使用 Gabor 滤波器对卷积滤波器进行调制，通过调节卷积滤波器特征提取过程，增强特征对方向和尺度的适应性，同时减少网络参数。本文方法的具体处理过程是：首先生成不同方向和尺度的 Gabor 滤波器，并针对不同方向的滤波器学习一个权重系数，以突出不同方向特征之间的差异性，然后对卷积滤波器进行调制，产生各个方向的加权 Gabor 方向滤波器（WGoF），使输出特征对图像方向和尺度具有适应性。其中，WGoF 作为一个参数可调的滤波器，通过 Gabor 滤波器来调节卷积滤波器，以增强特征图的表达。

加权 Gabor 方向滤波器的计算流程如下：首先生成 U 个方向和 V 个尺度的 Gabor 滤波器，通过学习一个权重向量 $\boldsymbol{W}$ 对每个方向进行加权，然后对尺寸为 $N\times M\times M$ 的可学习滤波器进行调制，$M\times M$ 表示二维滤波器尺寸，N 表示 Gabor 滤波器方向数，调制过程为

$$C_{i,u}^{v}=C_{i,\circ}^{\circ}[\boldsymbol{W}\cdot\boldsymbol{G}(u,v)],\tag{1}$$

式中：u 和 v 分别为方向和尺度索引；$\boldsymbol{G}(u,v)$ 为对应的 Gabor 滤波器；$C_{i,\circ}$ 表示可学习滤波器；$^{\circ}$ 为点积操作；$C_{i,u}^{v}$ 为调制滤波器。WGoF 可以表示为

$$\boldsymbol{C}_{i}^{v}=(C_{i,1}^{v},C_{i,2}^{v},\cdots,C_{i,U}^{v})。\tag{2}$$

由于 Gabor 滤波器具有多个方向，可将 WGoF 视为三维卷积滤波器，其中滤波器尺度在不同层具有不同表现。Gabor 滤波器对可学习滤波器的调制过程如图 2 所示，其中 w_1，w_2，w_3，w_4 为学习到的权重。在获取 Gabor 方向滤波器后，将其与输入特征图像进行卷积，得到输出特征 $\hat{F}$，处理过程如图 3 所示，对应的计算方式为

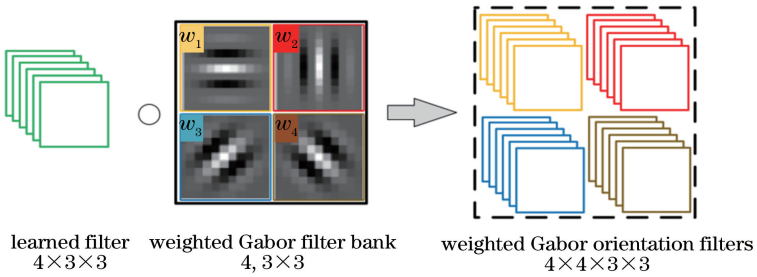

图 2　WGoFs 调制过程

Fig. 2　Modulation process of WGoFs

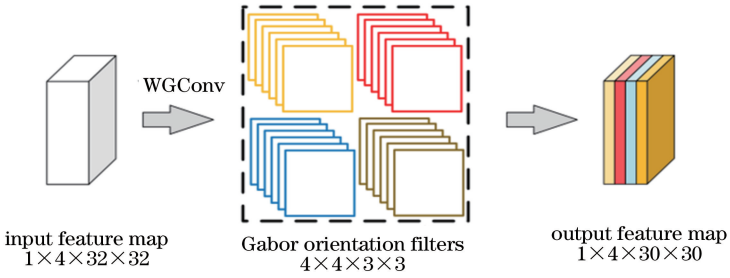

图 3　WGoFs 卷积过程

Fig. 3　Convolution process of WGoFs

$$\hat{F}=\mathrm{GConv}(F,\boldsymbol{C}_i),\tag{3}$$

式中：$\boldsymbol{C}_i$ 表示第 i 个 WGoF；F 表示输入特征；WGConv 表示可学习 Gabor 滤波器卷积。

本文采用反向传播法进行训练，设第 n 次迭代计算得到的可学习滤波器为 $C_{i,\mathrm{o}}^{(n)}$，则第 $n+1$ 次迭代计算得到的可学习滤波器为 $C_{i,\mathrm{o}}^{(n+1)}$，进行如下更新：

$$\delta=\frac{\partial L'}{\partial C_{i,\mathrm{o}}^{(n)}}=\sum_{u=1}^{U}\frac{\partial L'}{\partial C_{i,u}^{(n)}}\circ[\boldsymbol{W}\cdot\boldsymbol{G}(u,v)],\tag{4}$$

$$C_{i,\mathrm{o}}^{(n+1)}=C_{i,\mathrm{o}}^{(n)}-\eta\delta,\tag{5}$$

式中：L' 表示损失函数；η 表示学习率。通过对可学习滤波器进行参数更新，简化了训练过程，模型也更为紧凑和有效，使其对方向和尺度的变化具有更好的鲁棒性。

2.3　宽残差-加权 Gabor 卷积网络模块

深度神经网络模型层数不断加深有助于增强学习能力，提取更为丰富的特征。但是，当网络层数增加到一定数量后，由于受到梯度消失的影响，模型的测试精度难以再有提升，甚至随着训练过程的推进，测试精度逐渐下降，最终导致损失函数无法收敛到最小值。为了缓解该问题，残差模块应运而生，其主要思想是采用快捷连接跳过多层卷积层，通过不断堆叠模块，可使网络继续加深而不受梯度消失的影响。但是，不断加深的网络使得衰减特征重用的问题逐渐凸显，导致残差模块中只有部分参数参与更新。对此，Zagoruyko 和 Komodakis[25] 提出一种 WRB，由该模块堆叠的网络模型具有浅而宽的特性，因此能使用较浅模型表示较深的网络。

室内场景背景较为复杂，存在多种光照的干扰，特征提取具有较大的难度。为了构造出具有高分辨率的特征，往往需要采用较深的网络模型，且需要将各层之间的特征图像不断进行融合，以缓解多尺度物体的差异性过大问题。为了在构建轻量级网络的同时提取到较好的特征，本文采用 WRB 构建特征提取网络，分别提取 RGB 和 depth 图像特征。WRB 与普通残差模块的最大区别在于增加了系数 k 和卷积核的数量，使得在减少网络层数的同时保证参数数量，达到加快模型训练的目的。不同残差模块结构对比如图 4 所示，其中图 4(a)所示为原始残差模块，包含两个 3×3 的卷积层、批归一化层和 ReLu 层。图 4(b)和 4(c)表示两个不同的 WRB，其中 x_l 和 x_{l+1} 分别为第 l 层和第 $l+1$ 层的输入特征；FM 为采用不同卷积核数量及宽度的特征学习。相较于原始宽残差仅增加不同的结构系数，WRB 既增加了不同数量的卷积层，也增加了不同的特征

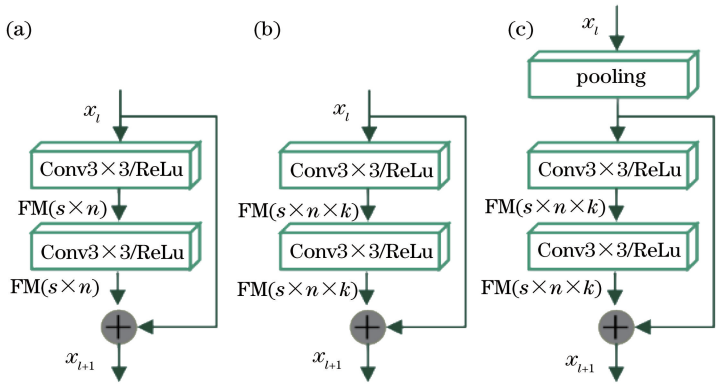

图 4　宽残差模块。(a)原始残差模块；(b)宽残差模块 1；(c)宽残差模块 2

Fig. 4　Wide residual blocks. (a) Original residual block; (b) wide residual block 1; (c) wide residual block 2

图数量，进而构建了宽而浅的网络模型。

本文采用 WRB 构建网络时所采用的卷积滤波器均为 WGoFs，一方面通过采用 WRB 使得模型较浅，另一方面利用 WGoFs 提取图像中方向和尺度不变性特征，以更好地关注 RGB 图像中的纹理、颜色等特征，提取 depth 图像中的边缘轮廓信息，增强模型的信息表达能力。为了构建轻量级模型，利用浅层神经网络实现深层神经网络同等性能，本文将宽残差-加权 Gabor 卷积网络（WRN-WGCN）模块的网络层数设置为 13，该模块由三个宽残差组构成，每个残差组的宽度 k 设置为 4。在每个残差组中，深度系数 L 决定了残差组的结构，第一个残差组 GCConv2 和第二个残差组 GCConv3 采用图 4(b)的结构，第三个残差组 GCConv4 采用图 4(c)的结构。具体结构参数如表 1 所示。WRN-WGCN 模块结构示意图如图 5 所示。首先利用 3×3 的 GCConv 对输入图像进行卷积操作，然后通过两个宽残差组提取特征，最后输出对应的特征图像。

表 1　WRN-WGCN 结构参数设置

Table 1　Structural parameter setting of WRN-WGCN

Group name	Output feature size	Block type
GCConv1	$N\times N$	$[3\times 38]$
GCConv2	$N\times N$	$\begin{bmatrix}3\times3 & 16\times k\\ 3\times3 & 16\times k\end{bmatrix}\times L$
GCConv3	$N\times N$	$\begin{bmatrix}3\times3 & 16\times k\\ 3\times3 & 16\times k\end{bmatrix}\times L$
GCConv4	$(N/2)\times(N/2)$	$\begin{bmatrix}3\times3 & 32\times k\\ 3\times3 & 32\times k\end{bmatrix}\times L$

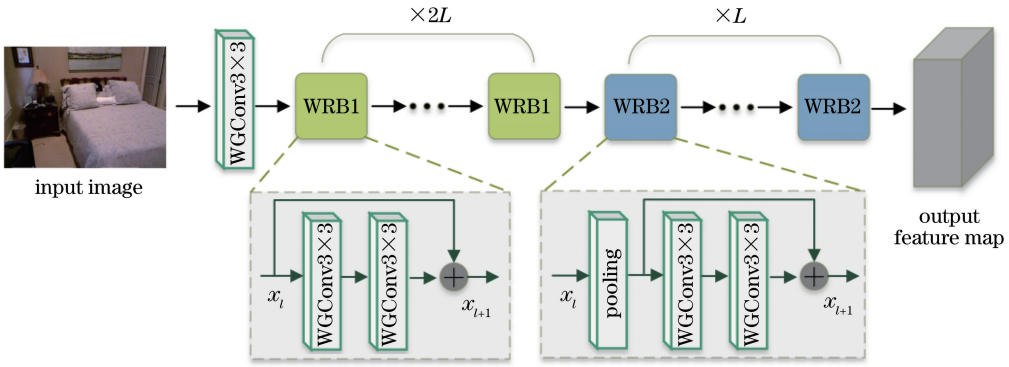

图 5　WRN-WGCN 模块结构

Fig. 5　Architecture of WRN-WGCN module

2.4　金字塔池化特征融合模块

在图像分割任务中，上下文信息作为一种关键特征，在整个分割过程中起到至关重要的作用。金字塔池化[26]作为提取上下文信息方法中最常见的一种，具体结构如图 6 所示，其核心思想是通过修改池化层的层数和尺寸来适应图像，从全局捕获图像

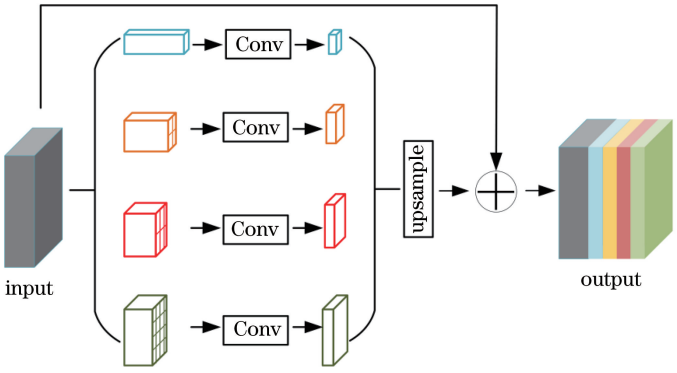

图 6　金字塔池化模块

Fig. 6　Pyramid pooling module

上下文信息。

图 6 中分别采用 4 种不同尺寸的池化内核对输入特征进行 1×1 卷积，以减小特征的通道数量，利用双线性插值进行上采样，以保证输出特征图与输入特征图大小一致，并在通道上进行拼接。本文主要采用金字塔池化对 RGB 和 depth 图像进行多尺度融合。如图 7 所示，首先采用池化内核对 RGB 和 depth 图像进行池化，然后将对应尺度的特征图进行拼接，最后通过上采样将不同尺度的图像进行融合，以有效聚合两类图像不同尺度上的信息，提升网络的特征提取能力。为了较好地表示空间尺度信息，本文采用的 4 种内核尺寸是 1×1、2×2、3×3 和 6×6。

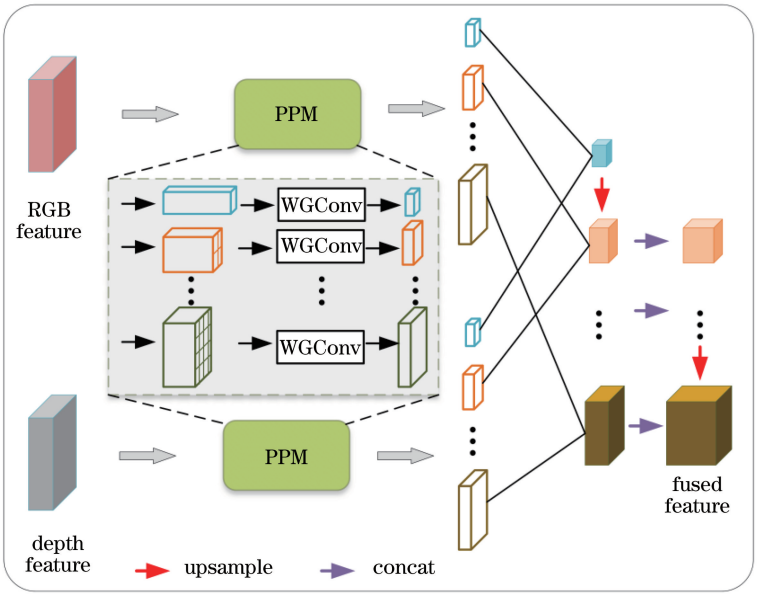

图 7　本文金字塔池化特征融合模块

Fig. 7　Proposed pyramid pooling feature fusion module

3　实验结果及分析

3.1　数据集与实验平台

1) NYUDv2 数据集

本文实验在主流语义分割数据集 NYUDv2[27] 上进行，该数据集包含 1449 对稠密标注的 RGB 和 depth 图像对(图像分辨率为 640 pixel×480 pixel)，而且包含 464 种场景以及 35064 个目标，具有 894 个目标类别。图 8 为数据集中的某一 RGB 图像、depth 图像以及语义标签。实验中采用 40 类语义标签，并按照文献[17]的标准划分策略对数据集进行划分，采用 795 张图像进行训练，654 张图像进行测试。在训练过程中，为了解决数据量不足的问题，对图像进行翻转、平移、裁剪以及色彩抖动等操作。

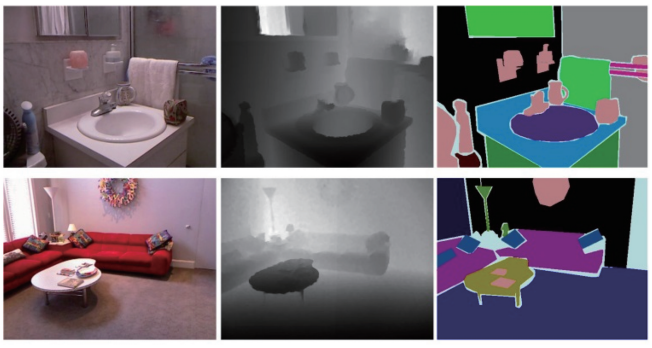

图 8　数据集中的彩色图像、深度图像以及语义标签。(a) RGB 图像；(b)深度图像；(c)语义标签

Fig. 8　RGB and depth images and their corresponding semantic labels in dataset.
(a) RGB images; (b) depth images; (c) semantic labels

2）实验平台

本文所有实验均在 2.4 GHz Intel Xeon CPU E5-2640 v4、48 GB RAM、NVIDIA RTX 2080Ti、Linux 64 位计算机上进行，编程环境采用 Anaconda 5.0.1（Python 3.6）、TensorFlow 1.4、Keras 2.0.8、Pytorch 0.4.0。

3.2　本文方法相关参数设置

1）训练参数设置

在训练过程中，采用随机梯度下降优化器进行参数更新，将训练和验证过程重复一定次数。通过不断尝试不同参数值，将批尺寸设置为 8，初始学习率 l_{init} 设置为 0.001。为了适应不同阶段权值修正幅度，在迭代后期逐渐减小学习率，当迭代次数为 q 时，对应学习率为 $l_{\mathrm{init}}(1-q/l_{\mathrm{maxiter}})^{0.9}$，其中 l_{maxiter} 表示最大迭代次数。采用上述设置进行实验，训练损失曲线和验证损失曲线如图 9 所示，当模型训练约 3200 个 epoch 时曲线接近收敛。

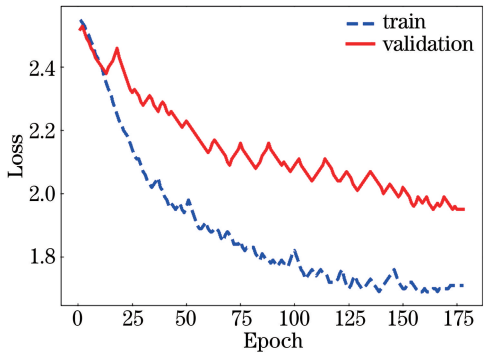

图 9　训练过程中的损失曲线

Fig. 9　Loss curves in training process

2）Gabor 方向滤波器参数设置

加权Gabor方向滤波器的方向数U和尺度数V对分割结果会产生影响，本文实验采用基于不同尺度数和方向数的训练模型得到的最佳测试精度来设置参数。尺度数对测试精度的影响曲线如图 10(a)所示，其中卷积滤波器尺寸分别设为 5×5 和 3×3，方向数设为 1，4，7，不同模型大小如表 2 所示。方向数对测试精度的影响曲线如图 10(b)所示，方向数设为 2，3，4，5，6，7。从图中可以看出，当方向数为 4、尺度数为 4 时模型具有较高的分割精度。

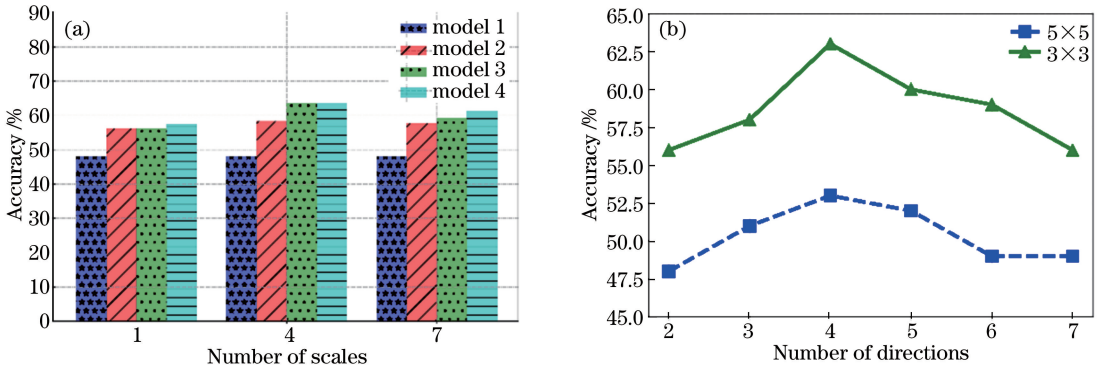

图 10　不同尺度数量和方向数量对测试精度的影响。(a)不同尺度数量下的测试精度；(b)不同方向数量下的测试精度

Fig. 10　Test accuracy versus number of scales and number of directions. (a) Test accuracy under different number of scales; (b) test accuracy under different number of directions

表 2　不同滤波器大小下的模型大小

Table 2　Model sizes with different filter sizes

Model name	Filter size	Model size /MB
Model 1	5×5	163
Model 2	5×5	124
Model 3	3×3	148
Model 4	3×3	117

3.3　评估指标

本文实验采用像素精度（A_{cc}）、均像素精度（m_{Acc}）、均交并比（m_{IoU}）以及频权交并比（F_{WIoU}）4 种指标评价语义分割效果。假设 n_{ij} 表示标签类别为 i 的像素点被预测为类别 j 的个数，n_{C} 为总类别数量，s_i 为标签类别为 i 的像素个数，则总的像素数量 $s=\sum_i s_i$，上述指标分别定义为 $A_{\mathrm{cc}}=\sum_i \frac{n_{ii}}{s}$，$m_{\mathrm{Acc}}=\frac{1}{n_{\mathrm{C}}}\sum_i \frac{n_{ii}}{s_i}$，$m_{\mathrm{IoU}}=\frac{1}{n_{\mathrm{C}}}\sum_i \frac{n_{ii}}{s_i+\sum_j n_{ji}-n_{ii}}$，$F_{\mathrm{WIoU}}=\frac{1}{s}\sum_i \frac{n_{ii}}{s_i+\sum_j n_{ji}-n_{ii}}$。

3.4　实验结果及讨论

本文方法包含多个不同功能模块，如宽残差卷积模块、金字塔池化特征融合模块以及加权 Gabor 方向滤波器。为了验证各个模块的有效性，对所提方法进行拓展，生成 4 种变体方法。为了对比各个模块的有效性，设置变体模型 1 为 Baseline；以 RGB

图像和深度图像为输入，然后通过常规卷积神经网络提取图像特征，并采用特征级联的方式进行融合，网络中卷积操作均采用常规卷积滤波器。为了验证所提宽残差特征提取网络的有效性，设置变体方法2为WRN-CNN：输入为RGB图像和深度图像，分别采用宽残差网络进行特征提取，并将两种特征直接进行级联和融合，其中宽残差网络由常规卷积滤波器构成。为了验证所提加权Gabor方向滤波器的有效性，设置变体方法3为WGCN：以RGB图像和深度图像为输入，采用加权Gabor方向滤波器作为卷积滤波器，并将获得的两种特征直接进行级联和融合。相比于变体模型1而言，该方法将加权Gabor方向滤波器替换为常规卷积滤波器。为了验证本文所提特征融合方法的有效性，设置变体方法4为PP-Fusion：以RGB图像和depth图像为输入，通过常规卷积神经网络提取特征，并采用金字塔池化特征融合模块进行融合。

本文基于NYUDv2数据集构造训练集，训练得到室内场景下的语义分割模型，然后在测试集上进行量化分析。为了验证模型的泛化性能，本文模型及相关方法还在SUN-RGBD数据集上进行测试，并进行结果评估和可视化分析。

3.4.1　NYUDv2数据集

为了验证各模块设计的有效性，本文设计消融实验，根据4种变体方法构建不同网络模型，并对模型进行训练和测试，得到评估指标。此外，将本文方法与现有经典的采用FCN和SegNet语义分割的方法进行对比，得到在NYUDv2数据集上的量化结果如表3所示。

表3　不同分割算法在NYUDv2数据集上的结果对比

Table 3　Comparison of results for different segmentation algorithms on NYUDv2 dataset

Method	Module			A_{cc}/%	m_{Acc}/%	m_{IoU}/%	F_{WIoU}/%
	WRN-CNN	WGCN	PP-Fusion				
Ours	√	√	√	**66.3**	**50.8**	**40.0**	**53.1**
Variant 1				58.3	41.6	30.1	45.8
Variant 2	√			58.6	42.4	31.9	45.3
Variant 3		√		60.8	48.2	35.8	50.4
Variant 4			√	63.2	45.8	36.4	46.6
FCN[2]				65.4	45.1	34.3	48.6
SegNet[3]				56.2	47.6	35.1	50.1

从NYUDv2数据集的量化结果可看出，相比于传统卷积滤波器网络，基于加权Gabor方向滤波器的网络模型更具有优势，在4个评价指标上相对于基本模型分别提升了2.5%、6.6%、5.7%和4.6%，这说明采用不同方向和尺度的Gabor滤波器对可学习卷积滤波器进行调制能有效提取图像中的特征，使分割结果不易受方向和尺度变化的干扰。而将RGB图像特征和depth图像特征进行多尺度融合，也能在一定程度上提升网络模型的分割性能。相较于经典语义分割模型FCN和SegNet，所提方法的精度有明显的提升，在F_{WIoU}指标上分别提升了4.5%和3.0%，这不仅说明了该方法结合双流图像的有效性，也说明通过多尺度特征融合并提取不变性特征能够丰富图像的信息表达。

为了对实验结果进行直观展示，将分割图像以不同颜色标注，得到如图11所示的语义结果。从图中可以看出，经典语义分割方法FCN和SegNet总体上能分割不同类型的目标，这表明编码-解码结构具有一定的特征恢复能力，但在细节上表现较为粗糙，对尺度较小的物体更是如此。而所提金字塔池化特征融合模块在多尺度处理方面具有一定优势，对于部分尺度较小物体的分割准确度相对较高，且对部分物体边缘的分割更加精细化。在方向和尺度不变性方面，本文所提基于加权Gabor方向滤波器的模型在细节上具有较好的体现，这说明提取方向和尺度不变性特征对分割效果具有一定的提升作用。而采用WRB构建的彩色和深度图像的双流网络模型，充分利用图像中的多种不同信息表达，在目标完整性和精细边缘分割方面具有明显优势。

3.4.2　SUN-RGBD数据集

在图像语义分割研究中，跨数据集实验结果是检验不同方法泛化性能的重要手段。为了验证所提方法的跨数据集分割结果，本文在NYUDv2数据集上训练模型，然后在SUN-RGBD数据集上进行测试。SUN-RGBD数据集中共有10335对已配准的彩色和深度图像对[27-28]，包含37类语义类别。由于SUN-RGBD数据集中标签集合是NYUDv2数据集中标签集合的子集，因此仅对37类共同的标签进行量化计算，不同方法的量化结果如表4所示。

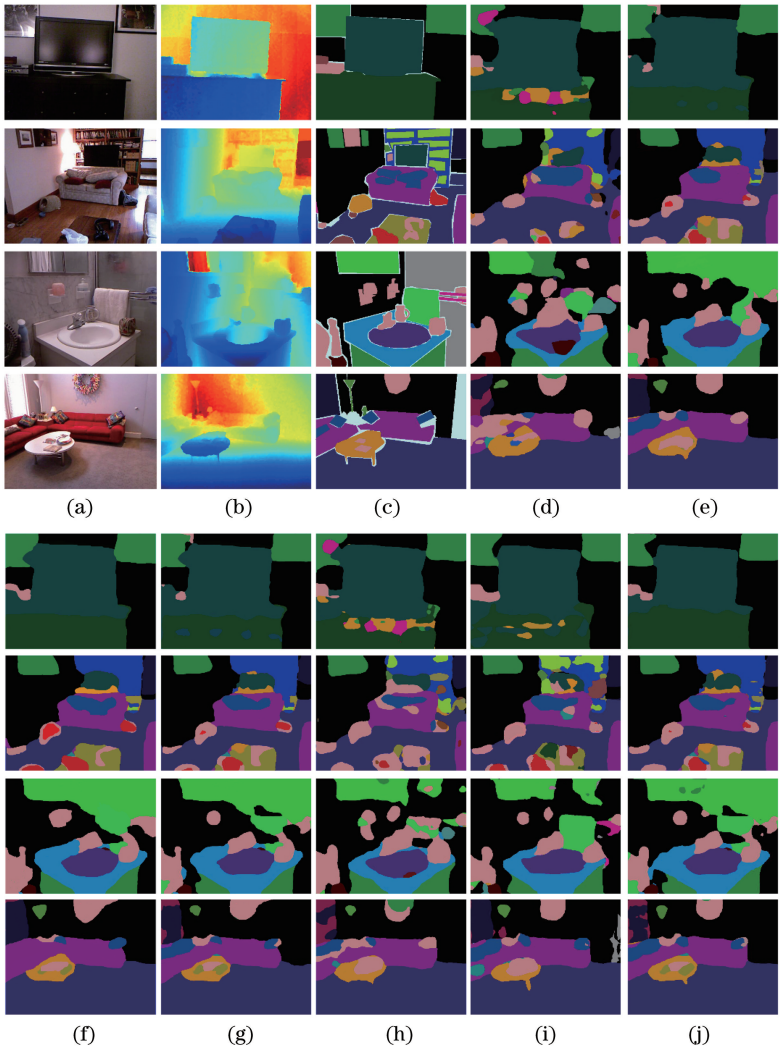

图 11　NYUDv2 数据集上各种方法得到的语义分割结果。(a) RGB;(b) depth;(c) GT;(d) baseline;(e) WRN-CNN;(f) WGCN;(g) PP-Fusion;(h) FCN;(i) SegNet;(j) ours

Fig. 11　Semantic segmentation results obtained by various methods on NYUDv2 dataset. (a) RGB; (b) depth; (c) GT; (d) baseline; (e) WRN-CNN; (f) WGCN; (g) PP-Fusion; (h) FCN; (i) SegNet; (j) ours

表 4　不同分割算法在 SUN-RGBD 数据集上的结果对比

Table 4　Comparison of results for different segmentation algorithms on SUN-RGBD dataset

Method	Module			A_{cc}/%	m_{Acc}/%	m_{IoU}/%	F_{WIoU}/%
	WRN-CNN	WGCN	PP-Fusion				
Ours	√	√	√	**58.2**	**38.5**	**28.2**	**42.0**
Variant 1				45.2	33.7	21.8	37.4
Variant 2	√			44.8	34.5	23.1	38.6
Variant 3		√		54.6	35.1	27.3	37.7
Variant 4			√	56.1	34.6	26.0	36.3
FCN[2]				49.5	36.5	23.7	35.8
SegNet[3]				47.8	34.6	26.2	38.2

虽然 NYUDv2 和 SUN-RGBD 两个数据集均是在室内场景下构建的，但二者仍具有一定的差异性。对比不同方法的评估结果可以发现，本文构建的语义分割方法具有一定的竞争性，相较于经典的编解码语义分割框架 FCN 和 SegNet 具有一定的性能提升。除此之外，通过对不同模块进行消融实验，可发现各个模块，尤其是加权 Gabor 方向滤波器以及金字塔池化特征融合模块相对具

有优势。

图 12 列出了采用不同方法得到的可视化结果，其中不同颜色表示不同语义类别。从图中可以看出，本文方法在结果上具有更好的细节表现，不仅能对尺度差异较大的目标进行分割，还能减小环境光照的影响，具有较强的场景适应性。

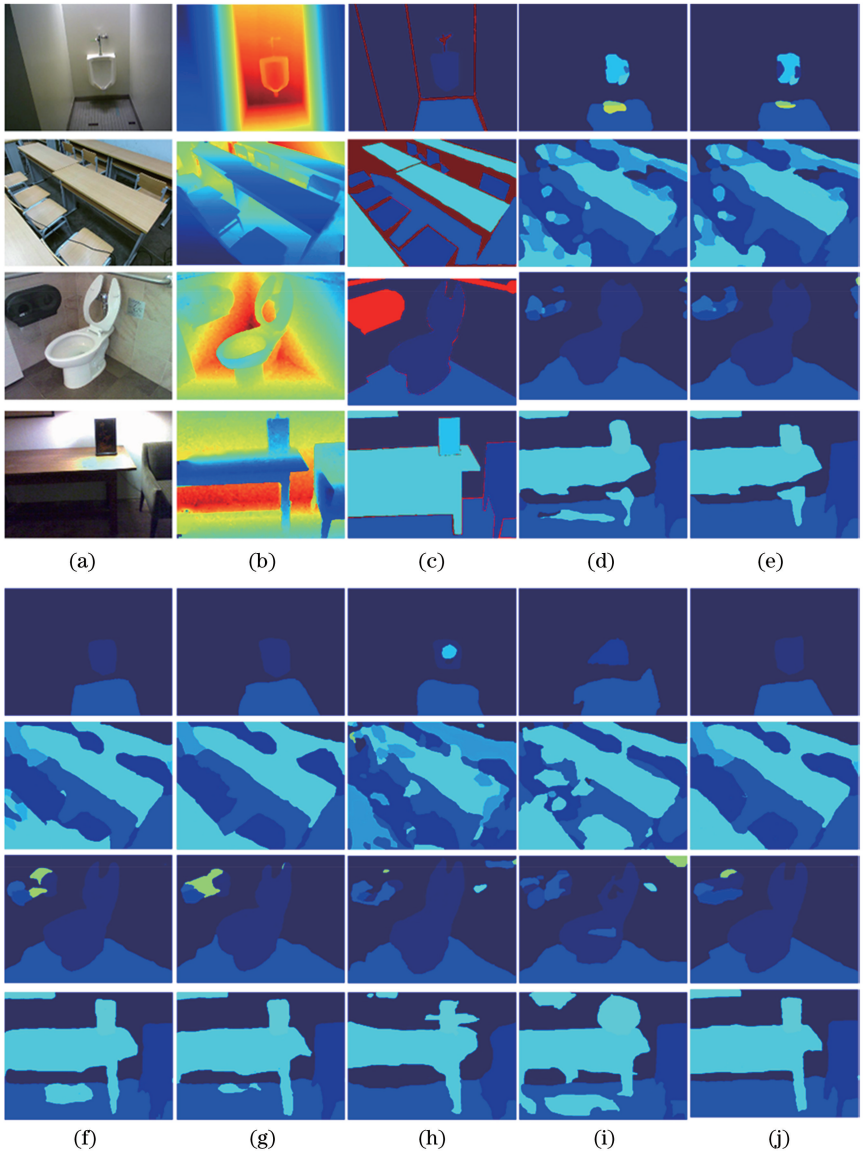

图 12　采用各种方法在 SUN-RGBD 数据集上得到的语义分割结果。(a) RGB；(b) depth；(c) GT；(d) baseline；(e) WRN-CNN；(f) WGCN；(g) PP-Fusion；(h) FCN；(i) SegNet；(j) ours

Fig. 12　Semantic segmentation results obtained by various methods on SUN-RGBD dataset. (a) RGB; (b) depth; (c) GT; (d) baseline; (e) WRN-CNN; (f) WGCN; (g) PP-Fusion; (h) FCN; (i) SegNet; (j) ours

3.4.3　模型复杂度评估

本文所提方法在考虑语义分割精度的同时，还兼顾模型复杂度，由此设计出相对轻量级的网络模型。表 5 对比了不同算法模型的空间复杂度和 m_{IoU} 值。从模型空间复杂度考虑，SegNet 采用池化索引进行非线性上采样，无需对上采样过程进行参数学习，因此其参数量远少于 FCN 方法，但是两者的分割性能相当。本文采用宽残差网络进行特征提取，明显减少了参数量，加之金字塔池化过程的参数量也较少，因此本文模型的复杂度较小。此外，Gabor 方向滤波器也有利于网络的轻量化，使得较简单的网络能学习到复杂的特征表示。从模型时间复杂度考虑，采用传统的卷积滤波器构建的网络的推理时间较长，而采用宽残差网络构建的模型较浅，在推理过程中具有一定优势，结合 Gabor 方向滤波器提取方向和尺度特征，能有效减少模型推理的时间。

表 5　不同算法的推理时间和空间复杂度对比

Table 5　Comparison of reasoning time and space complexity for different algorithms

Method	Module			Model size /MB	Reasoning time /ms
	WRN-CNN	WGCN	PP-Fusion		
Ours	√	√	√	117	42
Variant 1				381	76
Variant 2	√			115	35
Variant 3		√		187	48
Variant 4			√	245	51
FCN[2]				549	43
SegNet[3]				126	58

4　结　论

针对室内场景中背景复杂、光照变化以及目标尺度差异等问题，提出一种基于双流加权 Gabor 卷积网络融合的 RGB-D 图像语义分割方法，模型以 RGB 和 depth 图像作为输入，分别通过宽残差加权 Gabor 卷积网络提取图像特征，然后采用金字塔池化分别对 RGB 和 depth 图像特征进行多尺度处理，之后通过金字塔池化特征融合模块对提取的双流特征进行多尺度深度融合。最后，将融合特征进行上采样，以解码形式级联得到包含不同尺度的融合特征，将其输入 Softmax 进行分类。实验结果表明，所提方法在室内场景语义分割任务中具有较好的表现，能适应不同尺度和方向的变化，且对物体细节部分和边缘轮廓具有较好的分割结果。

参　考　文　献

[1] Ronneberger O, Fischer P, Brox T. U-net: convolutional networks for biomedical image segmentation [M] // Lecture Notes in Computer Science. Cham: Springer International Publishing, 2015: 234-241.

[2] Shelhamer E, Long J, Darrell T. Fully convolutional networks for semantic segmentation [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 39(4): 640-651.

[3] Badrinarayanan V, Kendall A, Cipolla R. SegNet: a deep convolutional encoder-decoder architecture for scene segmentation [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 39(12): 2481-2495.

[4] Chen L C, Zhu Y K, Papandreou G, et al. Encoder-decoder with atrous separable convolution for semantic image segmentation[M]//Computer Vision-ECCV 2018. Cham: Springer International Publishing, 2018: 833-851.

[5] Noh H, Hong S, Han B. Learning deconvolution network for semantic segmentation[C]//2015 IEEE International Conference on Computer Vision (ICCV), December 7-13, 2015, Santiago, Chile. New York: IEEE Press, 2015: 1520-1528.

[6] Liu W, Rabinovich A, Berg A C. ParseNet: looking wider to see better[EB/OL]. (2015-11-19)[2020-04-26]. https://arxiv.org/abs/1506.04579.

[7] Zhang Z H, Fang W, Du L L, et al. Semantic segmentation of remote sensing image based on encoder-decoder convolutional neural network [J]. Acta Optica Sinica, 2020, 40(3): 0310001.
张哲晗, 方薇, 杜丽丽, 等. 基于编码-解码卷积神经网络的遥感图像语义分割[J]. 光学学报, 2020, 40(3): 0310001.

[8] Yu F, Koltun V. Multi-scale context aggregation by dilated convolutions[EB/OL]. (2016-04-30)[2020-04-26]. https://arxiv.org/abs/1511.07122.

[9] Wu Z H, Gao Y M, Li L, et al. Fully convolutional network method of semantic segmentation of class imbalance remote sensing images [J]. Acta Optica Sinica, 2019, 39(4): 0428004.
吴止锾, 高永明, 李磊, 等. 类别非均衡遥感图像语义分割的全卷积网络方法[J]. 光学学报, 2019, 39(4): 0428004.

[10] Lin G S, Milan A, Shen C H, et al. RefineNet: multi-path refinement networks for high-resolution semantic segmentation[C]//2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), July 21-26, 2017, Honolulu, HI, USA. New York: IEEE Press, 2017: 5168-5177.

[11] Hu T, Li W H, Qin X X. Semantic segmentation of polarimetric synthetic aperture radar images based on multi-layer deep feature fusion[J]. Chinese Journal of Lasers, 2019, 46(2): 0210001.
胡涛, 李卫华, 秦先祥. 基于多层深度特征融合的极化合成孔径雷达图像语义分割[J]. 中国激光, 2019, 46(2): 0210001.

[12] 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.

[13] Zheng S, Jayasumana S, Romera-Paredes B, et al. Conditional random fields as recurrent neural networks[C]//2015 IEEE International Conference on Computer Vision (ICCV), December 7-13, 2015, Santiago, Chile. New York: IEEE Press, 2015: 1529-1537.

[14] Lin G S, Shen C H, van den Hengel A, et al. Efficient piecewise training of deep structured models for semantic segmentation [C] // 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 27-30, 2016, Las Vegas, NV, USA. New York: IEEE Press, 2016: 3194-3203.

[15] Arnab A, Jayasumana S, Zheng S, et al. Higher order conditional random fields in deep neural networks[M]//Computer Vision-ECCV 2016. Cham: Springer International Publishing, 2016: 524-540.

[16] Ren X F, Bo L F, Fox D. RGB-(D) scene labeling: features and algorithms[C]//2012 IEEE Conference on Computer Vision and Pattern Recognition, June 16-21, 2012, Providence, RI, USA. New York: IEEE Press, 2012: 2759-2766.

[17] Silberman N, Hoiem D, Kohli P, et al. Indoor segmentation and support inference from RGBD images[M]//Computer Vision-ECCV 2012. Berlin, Heidelberg: Springer Berlin Heidelberg, 2012: 746-760.

[18] He Y, Chiu W C, Keuper M, et al. STD2P: RGBD semantic segmentation using spatio-temporal data-driven pooling [C] // 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), July 21-26, 2017, Honolulu, HI, USA. New York: IEEE Press, 2017: 7158-7167.

[19] Cheng Y H, Cai R, Li Z W, et al. Locality-sensitive deconvolution networks with gated fusion for RGB-D indoor semantic segmentation [C] // 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), July 21-26, 2017, Honolulu, HI, USA. New York: IEEE Press, 2017: 1475-1483.

[20] Yurdakul E E, Yemez Y. Semantic segmentation of RGBD videos with recurrent fully convolutional neural networks [C] // 2017 IEEE International Conference on Computer Vision Workshops (ICCVW), October 22-29, 2017, Venice, Italy. New York: IEEE Press, 2017: 367-374.

[21] Hu X X, Yang K L, Fei L, et al. ACNET: attention based network to exploit complementary features for RGBD semantic segmentation [C] // 2019 IEEE International Conference on Image Processing (ICIP), September 22-25, 2019, Taipei, Taiwan, China. New York: IEEE Press, 2019: 1440-1444.

[22] Lin D, Zhang R M, Ji Y F, et al. SCN: switchable context network for semantic segmentation of RGB-D images [J]. IEEE Transactions on Cybernetics, 2020, 50(3): 1120-1131.

[23] Han J, Ma K K. Rotation-invariant and scale-invariant Gabor features for texture image retrieval [J]. Image and Vision Computing, 2007, 25(9): 1474-1481.

[24] Luan S Z, Chen C, Zhang B C, et al. Gabor convolutional networks [J]. IEEE Transactions on Image Processing, 2018, 27(9): 4357-4366.

[25] Zagoruyko S, Komodakis N. Wide residual networks [EB/OL]. (2017-06-14) [2020-04-26]. https://arxiv.org/abs/1605.07146.

[26] He K M, Zhang X Y, Ren S Q, et al. Spatial pyramid pooling in deep convolutional networks for visual recognition[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2015, 37(9): 1904-1916.

[27] Janoch A, Karayev S, Jia Y Q, et al. A category-level 3-D object dataset: putting the Kinect to work [C] // 2011 IEEE International Conference on Computer Vision Workshops (ICCV Workshops), November 6-13, 2011, Barcelona, Spain. New York: IEEE Press, 2011: 1168-1174.

[28] Xiao J X, Owens A, Torralba A. SUN3D: a database of big spaces reconstructed using SfM and object labels[C]//2013 IEEE International Conference on Computer Vision, December 1-8 2013, Sydney, NSW, Australia. New York: IEEE Press, 2013: 1625-1632.