基于多尺度特征提取和全连接条件随机场的图像语义分割方法

董永峰[1,2]，杨雨沂[1]，王利琴[1,2*]

[1]河北工业大学人工智能与数据科学学院，天津 300401；

[2]河北省大数据计算重点实验室，天津 300401

摘要 针对图像语义分割中图像的上下文信息利用不充分、边缘分割不清等问题，提出一种基于多尺度特征提取与全连接条件随机场的网络模型。分别以多尺度形式将 RGB 图像和深度图像输入网络，利用卷积神经网络提取图像特征；将深度信息作为补充信息添加到 RGB 特征图，得到语义粗分割结果；采用全连接条件随机场优化语义粗分割结果，最终得到语义精细分割结果。实验结果表明，所提方法提高了图像语义分割的精度，优化了图像语义分割的边缘，具有实际应用价值。

关键词 图像处理；图像语义分割；卷积神经网络；多尺度特征；深度学习；全连接条件随机场

中图分类号 TP391 **文献标识码** A **doi**：10.3788/LOP56.131007

Image Semantic Segmentation Based on Multi-Scale Feature Extraction and Fully Connected Conditional Random Fields

Dong Yongfeng[1,2], Yang Yuxin[1], Wang Liqin[1,2*]

[1] *School of Artificial Intelligence, Hebei University of Technology, Tianjin* 300401, *China*;

[2] *Hebei Provincial Key Laboratory of Big Data Computing, Tianjin* 300401, *China*

Abstract Aiming at the problems of insufficient usage of context information and unclear image edge segmentation in image semantic segmentation, a network model based on multi-scale feature extraction and fully connected conditional random fields is proposed. RGB and depth images are input into the network in a multi-scale form, and their features are extracted by a Convolutional neural network. Depth information is added to supplement the RGB feature map and obtain a rough semantic segmentation, which is optimized by the fully connected conditional random fields. Finally, fine semantic segmentation results are obtained. This proposed method improves the precision of semantic segmentation and optimizes the image edge segmentation, which has a practical application.

Key words image processing; image semantic segmentation; Convolutional neural network; multi-scale feature; deep learning; fully connected conditional random field

OCIS codes 100.6890; 110.2960; 150.1135

1 引 言

图像语义分割是像素级的密集分类问题，目标是对图像中的每个像素进行语义信息标注，从而从整体上理解图像。其在目标检测、场景理解和三维重建等计算机视觉任务中已广泛应用[1-2]，具有重要的理论研究意义和实际应用价值。图像语义分割的难点主要来源于物体层次、类别层次和背景层次[3-4]三个方面。

在深度学习出现之前，传统的图像语义分割方法主要包括像素级别阈值法、基于像素聚类的分割方法以及基于图论划分的分割方法[5]。基于卷积神

收稿日期：2018-11-13；**修回日期**：2018-12-23；**录用日期**：2019-01-30

基金项目：天津市基础研究计划(17JCTPJC55400)、天津市基础研究计划 (17JCTPJC55600)、河北省自然科学基金(F2017202145)

* **E-mail**：wangliqin@scse.hebut.edu.cn

经网络(CNN)的方法不需要人工参与,能够直接从大量样本图像数据集中自动学习与语义相关的有用特征,并获得精确的结果,较传统语义分割方法具有明显的优势。Long 等[6]提出的全卷积网络(FCN)是语义分割模型的开山之作,DeepLab[7]、SegNet[8]等网络模型结构与 FCN 相同;与基于 FCN 的方法相比,Noh 等[9]提出的反卷积网络克服了物体尺度带来的高识别错误问题;为了充分利用物体—图像和物体—物体之间存在的上下文关系,提高语义分割的精度,文献[10-11]使用扩张卷积聚合多尺度的上下文信息;Zhao 等[12]提出的 PSPNet 通过空间池化模块和空间场景解析网络,利用不同区域的上下文信息聚合全局上下文信息;Lin 等[13]将多尺度图像作为输入,产生了不同尺度的特征图,提高了语义分割的精度;熊志勇等[14]将缩放得到的三个不同尺度图像作为网络输入,通过多尺度融合算法生成输出图;蒋应锋等[15]提出多尺度交替迭代训练,对每个像素进行类别标定,并应用超像素描绘分割图的轮廓;刘丹等[16]以超像素为中心,提取不同尺度的图像块作为网络的输入,将多尺度 CNN 模型结构用于图像语义分割;Zheng 等[17]提出的 CRFasRNN 模型和 Lin 等[18]提出的结合 CNN 和条件随机场(CRF)的模型利用 CNN 的特征提取和 CRF 概率图形建模的优势,提高了不同图像区域之间的语义相关性。

为了解决普通 CNN 方法无法处理的边缘分割不清等问题,通常采用条件 CRF[19]、全连接条件随机场(FullCRF)[20]、马尔科夫随机场(MRF)或高斯条件随机场(G-CRF)[21]优化网络输出,从而得到更加精确的分割结果。

为充分利用物体的上下文信息,并清晰描绘图像的边缘轮廓,本文在 FuseNet[22]的基础上,设计了一种多尺度的 CNN 模型——多尺度特征提取和全连接条件随机场相结合的图像语义分割方法(MSF-CRF)。该方法采用语义粗分割网络(MSF-Net)得出语义粗分割结果,再利用 FullCRF 优化,得到语义精细分割结果。MSF-CRF 模型分别以多尺度形式将 RGB 图像和深度图像输入到对应的尺度分支,并将 RGB 图像的深度信息以元素求和的方式添加到 RGB 特征图,而后采用多尺度特征融合的方式将三个尺度分支的信息进行融合,进一步改善语义分割的性能,最后采用 FullCRF 优化图像的边界。对所提方法进行了详细研究,并在 NYUv2 数据集上进行了实验验证。

2 MSF-CRF 语义分割模型

2.1 多尺度输入

为了利用图像的多尺度信息,提取不同尺度下的不变特征,更好地感知图像中过大或者过小的物体,避免损失过多的图像信息,提高语义分类的精度。采用双线性插值的方式将 RGB 图像和深度图像分别转化为三种不同的尺度,然后输入到网络中。以 320 pixel×240 pixel 的图像为基准,分别以基准图像的 0.6、0.8 和 1 倍进行处理,得到128 pixel×96 pixel、192 pixel × 144 pixel 和 320 pixel×240 pixel三种不同的尺度。三个尺度的图像共用同一个网络模型,采用共享权重的方式训练网络。另外,采用最邻近插值方式将真实标签图处理成大小为 320 pixel×240 pixel 的图像。

2.2 语义粗分割网络—MSF-Net

MSF-Net 为编码-解码式网络,其将输入图像同时输入到三个尺度分支中,生成不同尺度的特征图,每个尺度的图像经过网络后,会得到三个得分图,进行多尺度特征融合后,即可得到语义粗分割结果。

图 1 所示为多尺度特征融合前的一个尺度分支的结构图,表 1 所示为该分支编码器部分参数设置。通过输入不同尺度的图像,产生不同尺度的特征图,以保证预测图的效果。每个尺度分支的编码部分由 RGB 图像和深度图像的特征提取分支组成,解码部分与编码部分相对应。

RGB 图像特征提取分支的第一部分由 CBR 块、融合层和池化层组成,该部分执行两次,其中 CBR 块由卷积层(Conv)、批归一化层(BN)和激活函数(ReLU)组成;第二部分由 CBR 块、融合层、池化层和 Dropout 层组成,加入 Dropout 层可以防止网络出现过拟合现象;第三部分同样由 CBR 块、融合层、池化层和 Dropout 层组成,该部分执行一次。

深度图像特征提取分支的第一部分由 CBR 块和池化层组成,该部分执行两次;第二部分由 CBR 块、池化层和 Dropout 层组成,该部分执行两次;第三部分为 CBR 块。

2.2.1 MSF-Net 网络结构定义

将类别标签定义为集合 $\zeta=\{1,2,\cdots,K\}$,其中 K 为标签序号。对于 S 个具有相同长度和宽度的 RGB-D 图像和真实标签图,$X_1,X_2,\cdots,X_S$ 表示输入到每个尺度分支的具有 4 个通道的 RGB-D 图像,$G_1,G_2,\cdots,G_S$ 表示与之相对应的真实标签图,i 为

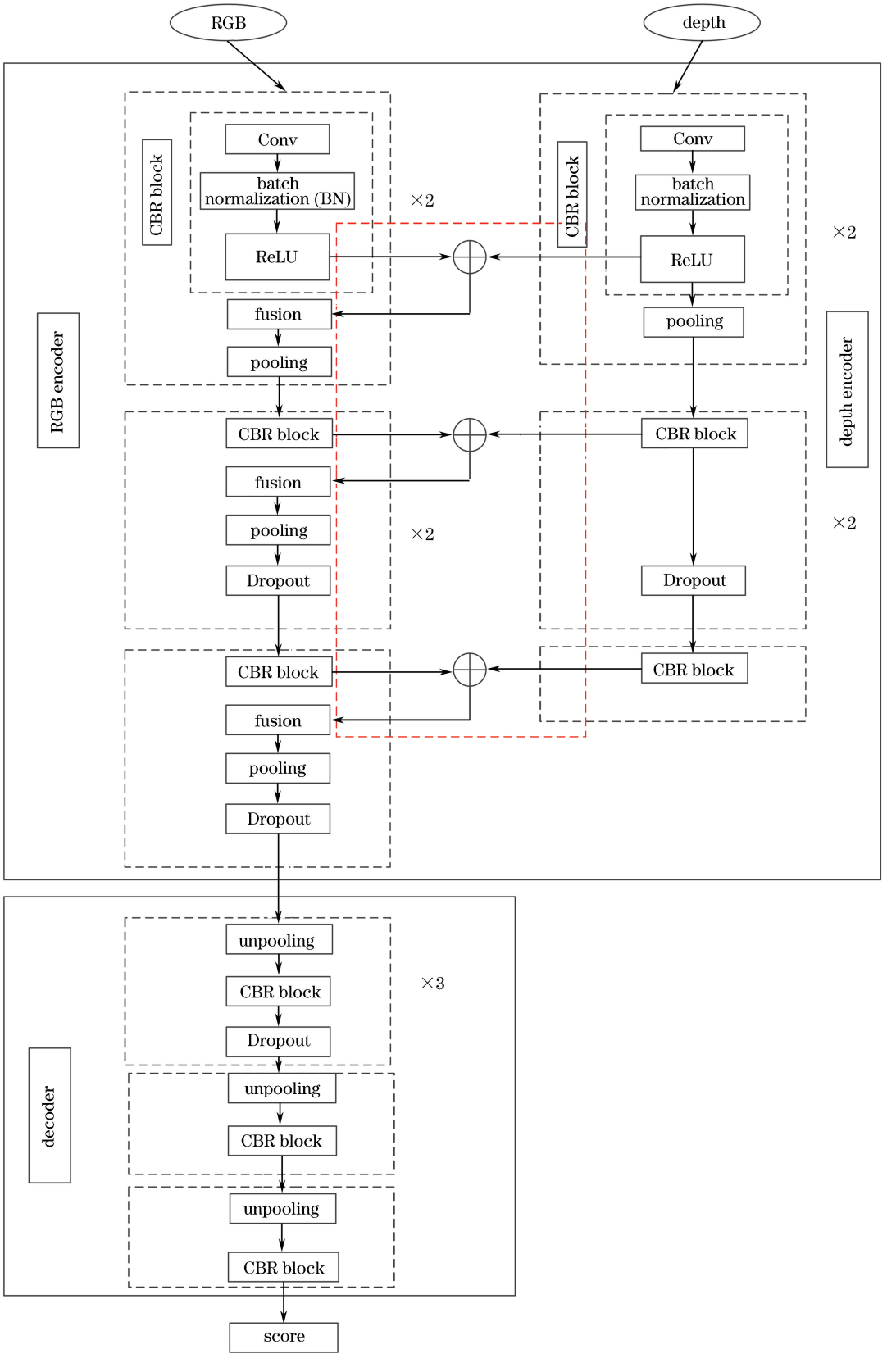

图 1 特征融合前的单分支网络结构

Fig. 1 Single-branch network structure before feature fusion

RGB-D 图像和真实标签图的索引。网络参数为 $W=[w^{(1)},w^{(2)},\cdots,w^{(L)}]$，其中 L 为网络层数，

$w^{(L)}$为第 L 层的网络参数，网络的各层定义为 $t^{(L)}(x,w^{(L)})$，x 为像素，则整个网络的定义为

$$f(x,W)=t^{(L)}\{t^{(L-1)}\{\cdots t^{(2)}[t^{(1)}(x,w^{(1)}),w^{(2)}]\cdots,w^{(L-1)}\},w^{(L)}\},\tag{1}$$

式中：$f(x,W)$的第 q 个分量 $f_q(x,W)$表示像素 x 属于类别 q 的得分。

利用 Softmax 函数将得分映射为输入图像中所含像素类别的概率分布，该过程称为 Softmax 的回归过程，定义为

$$p(q\mid x,W)=\frac{\exp[f_q(x,W)]}{\sum_{k=1}^{K}\exp[f_k(x,W)]},\tag{2}$$

式中：p 为每个像素点的预测概率。

为了得到最优参数 W^*，采用交叉熵损失函数计算预测值和真实值之间的差，则有

$$W^*=\arg\min_W\frac{1}{2}\|W\|^2-\frac{\lambda}{SHW}\sum_{i=1}^{S}\sum_{j=1}^{HW}\log[p(g_{ij}\mid x_{ij},W)],\tag{3}$$

式中：x_{ij}和 g_{ij}分别为第 i 张训练图像的第 j 个像素及其对应的真实标签；超参数 $\lambda>0$ 表示 W 的 L2 范数。

2.2.2 RGB 图像与深度图像融合

深度图像包含图像的场景结构信息，将其与 RGB 图像进行融合，可以使网络学到更多的特征，从而更好地预测。RGB 图像与深度图像融合通常有两种方法：一种方法是将 RGB 图像和深度图像堆叠成 4 个通道，然后输入到网络中进行训练；另一种方法是将 RGB 图像和深度图像分为两个同步的网络分支，分别提取特征，在每一步池化操作之前，通过元素求和的方式将深度信息融合进 RGB 特征图，如图 1 中 RGB encoder 分支与 depth encoder 分支中间的虚框处所示。

2.2.3 多尺度得分图融合

多尺度得分图融合过程如图 2 所示。通过滤波计算出每个尺度分支中各个标签的得分后，采用双线性插值方法将得分图统一成同样大小，具体过程为

$$F(X,K)=\sum_{v=1}^{V}[w_v\times f_L^v(x,W)],\tag{4}$$

表 1 特征融合前单个分支编码器部分的参数设置表

Table 1 Parameter setting table of single branch encoder before feature fusion

RGB encoder			Depth encoder		
Conv block1:	Conv block2:	Conv block3:	Conv block1:	Conv block2:	Conv block3:
3×3 Conv 64	3×3 Conv 128	3×3 Conv 256	3×3 Conv 64	3×3 Conv 128	3×3 Conv 256
3×3 Conv 64	3×3 Conv 128	3×3 Conv 256	3×3 Conv 64	3×3 Conv 128	3×3 Conv 256
2×2 maxpooling	2×2 maxpooling	2×2 maxpooling	2×2 maxpooling	2×2 maxpooling	2×2 maxpooling
Conv block4:	Conv block5:		Conv block4:	Conv block5:	
3×3 Conv 512	3×3 Conv 512		3×3 Conv 512	3×3 Conv 512	
3×3 Conv 512	3×3 Conv 512		3×3 Conv 512	3×3 Conv 512	
3×3 Conv 512	3×3 Conv 512		3×3 Conv 512	3×3 Conv 512	
2×2 maxpooling	2×2 maxpooling		2×2 maxpooling		

式中：$F(X,K)$为多尺度特征融合后的结果；X 表示输入图像；V 表示网络的尺度总分支数，采用双线性插值的方式处理后，可以得到 V 个尺度的输入图像；w_v 为每个尺度分支的输出权重，本文将其设置为 1/3，表示每个尺度的重要性相同；$f_L^v(x,W)$为每个尺度分支的最终得分，$v=\{1,2,\cdots,V\}$。

2.3 基于 FullCRF 的语义精细分割-MSF-CRF

为了清晰地描绘目标区域的轮廓，利用 FullCRF 对 MSF-Net 产生的语义粗分割结果进行优化，过程如图 3 所示。

全连接条件随机场的能量函数 $E(a)$由两部分组成，即

$$E(a)=\sum_y\varphi_u(a_y)+\sum_{y<z}\varphi_p(a_y,a_z),y,z\in\{1,2,\cdots,N\},\tag{5}$$

式中：a 为每个像素 y 的标签；N 为图像中的像素总数；$\sum_y\varphi_u(a_y)$为一元势能函数，主要计算输入图像的第 y 个像素点属于类别 a_y 的概率，可以直接从卷积神经网络中获得，其可以预测像素的标签，而不考虑标签分配的平滑度和一致性；$\sum_{y<z}\varphi_p(a_y,a_z)$为能量函数的二元势能函数，主要计算像素间的相互影响，为相似的像素分配相似的标签。将成对势能函数化为高斯加权后，可得

$$\varphi_p(a_y,a_z)=\mu(a_y,a_z)\sum_{m=1}^{K}\boldsymbol{\omega}^{(m)}k^{(m)}(\boldsymbol{f}_y,\boldsymbol{f}_z),\tag{6}$$

式中：$\boldsymbol{f}_y$ 和 $\boldsymbol{f}_z$ 为特征空间中像素 y 和 z 的特征向量，来源于图像特征中的空间位置和 RGB 值；$k^{(m)}$

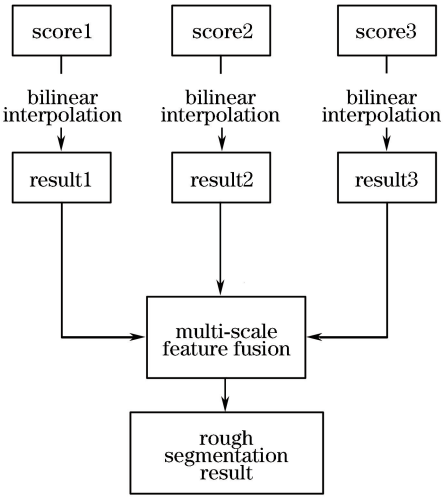

图 2 多尺度特征融合

Fig. 2 Multi-scale feature fusion

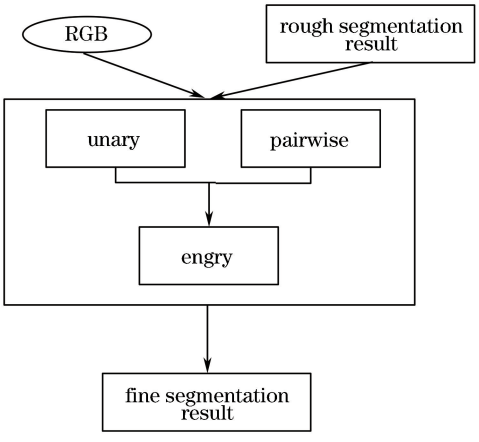

图 3 FullCRF 优化语义粗分割结果

Fig. 3 FullCRF optimization semantic rough segmentation result

为高斯核；$k^{(m)}(\boldsymbol{f}_y,\boldsymbol{f}_z)$为高斯核函数；$\omega^{(m)}$为可学习的参数，表示线性组合权重；$\mu(a_y,a_z)$为标签兼容性函数，仅取决于标签 a_y 和 a_z，而非输入图像。

为了实现多类图像的分割和标记，FullCRF 使用两个对比明显的高斯核函数。第一个核函数使用了像素的位置信息和颜色信息，用 p_y 和 p_z 来表示像素 y 和 z 的位置，用 X_y 和 X_z 表示像素 y 和 z 的原始颜色值，用可学习参数 θ_α 和 θ_β 来决定空间接近度和颜色相似度；第二个核函数只使用了像素的位置信息，用来去除孤立的区域。

$$k(\boldsymbol{f}_y,\boldsymbol{f}_z)=\omega^{(1)}\exp\left(-\frac{|p_y-p_z|^2}{2\theta_\alpha^2}-\frac{|X_y-X_z|^2}{2\theta_\beta^2}\right)+\omega^{(2)}\exp\left(-\frac{|p_y-p_z|^2}{2\theta_\gamma^2}\right),\tag{7}$$

式中：$\omega^{(1)}$、$\omega^{(2)}$及 θ_α、θ_β 和 θ_γ 均为模型可以学习的参数。

3 实验结果及分析

采用 NYUv2[23]室内场景数据集验证本文所提方法的有效性和可靠性。该数据集由 1449 个成对 RGB-D 图像组成，包括 3 个城市 26 个场景类别下 464 个不同的场景。同时，该数据集中包括 35064 个不同的对象，跨越 3894 个不同的类。根据 Gupta 等[24]的分类标准，将 NYUv2 数据集分为 795 张训练集和 654 张测试集。同时，参考该文献中的映射标准，将原始标签映射为 40 类(0 代表无效)。

为了评价图像语义分割结果的精度，以像素精度(PA)、平均类别精度(MA)和平均 IoU(MIoU)三个指标作为评价标准，分别利用 f_{PA}、f_{MA}、f_{MIoU} 表示 PA、MA 和 MIoU，则有

$$f_{\mathrm{PA}}=\frac{1}{N}\sum_q f_{\mathrm{TP}_q},q\in\{1,2,\cdots,K\},\tag{8}$$

$$f_{\mathrm{MA}}=\frac{1}{K}\sum_q\frac{f_{\mathrm{TP}_q}}{f_{\mathrm{TP}_q}+f_{\mathrm{FP}_q}},\tag{9}$$

$$f_{\mathrm{MIoU}}=\frac{1}{K}\sum_q\frac{f_{\mathrm{TP}_q}}{f_{\mathrm{TP}_q}+f_{\mathrm{FP}_q}+f_{\mathrm{FN}_q}},\tag{10}$$

式中：f_{TP_q}为分类正确的像素个数；f_{FP_q}为被预测为 q 类，但却不属于 q 类的像素个数；f_{FN_q}为被预测为非 q 类，但却属于 q 类的像素个数。

基于 PyTorch 深度学习框架进行实验，使用 NVIDIA GeForce 1080ti GPU 进行计算，cuDnn7.0 库加速。参数设置：学习率为 0.01，批训练样本个数为 4，动量为 0.9，权重衰减为 0.0005，Dropout 层的概率为 0.5。

3.1 MSF-CRF 与其他方法对比

为验证所提方法的可行性，将 MSF-CRF 分别与多个图像语义分割方法进行比较，如 FCN[6]、FuseNet[22]等。表 2 所示为不同网络模型在 NYUv2 数据集上的语义分割结果。

实验结果表明，与仅使用深度图像或者 RGB 图像作为 FCN 输入的结果相比，MSF-CRF 的效果明显更优，证明了深度图像可以为 RGB 图像添加辅助信息，提高了分类的准确性；使用 RGB 图像和深度图像作为网络输入时，所提方法较文献[25]方法在 PA 方面有很大程度的提升，结果高出 6.6%；较文献[26]方法在 PA 和 MA 方面均有较大程度的提升，结果分别高出 3.1%和 12.7%；较 FCN 在 PA 和 MA 均有较大程度的提升，结果分别高出5.4%和

表 2 不同网络在 NYUv2 数据集上的结果

Table 2 Results of different networks on NYUv2 dataset

Method	Input data type	PA / %	MA / %	MIoU / %
Method in Ref. [6]	RGB	60.0	42.2	29.2
Method in Ref. [6]	Depth	57.1	35.2	24.2
Method in Ref. [25]	RGB-depth	60.3	—	28.6
Method in Ref. [26]	RGB-depth	63.8	31.5	—
Method in Ref. [6]	RGB-depth	61.5	42.4	30.5
Method in Ref. [22]	RGB-depth	65.6	42.2	27.8
MSF-CRF	RGB-depth	66.9	44.2	30.2

1.8%；与 FuseNet 相比，在 PA、MA 和 MIoU 三个评价标准上都有一定程度的提升，结果分别高出 1.3%、2% 和 2.4%。实验表明本方法具有可行性。

3.2 分割性能对比

为验证所提方法的有效性，采用所提方法与 FuseNet 分别对数据集中的 40 个类别进行实验所得到的类别精度和 IoU 分数如表 3 所示。

将 RGB 图像和深度图像作为 FuseNet 和 MSF-CRF 的输入进行实验，所得结果如表 3～4 所示，可以看到：在类别精度方面，MSF-CRF 较 FuseNet 在 24 个类别中都有提升，如墙、地板、柜橱、沙发、桌子、书架、电视机等，其中在沙发、书架、电视机等具有明显边界的物体上提升得最多，分别为 12.1%、12.7%和 11.7%；在 IoU 方面，MSF-CRF 较 FuseNet 在 26 个类别中都有 1%～12%的提升，如床、沙发、桌子、镜子、床头柜等，其中在镜子上提高得最多，为 18.2%。同时，橱柜、沙发、桌子等物体在类别精度方面和 IoU 方面均有所提升。

表 3 40 个类别的类别精度对比表

Table 3 Comparison of classification accuracy of 40 categories

Dataset	Wall	Floor	Cabinet	Bed	Chair	Sofa	Table	Door
FuseNet	89.2	95.7	67.9	75.7	74.6	71.0	49.3	34.8
MSF-CRF	91.8	96.5	71.0	73.7	73.5	83.1	49.5	27.1
Dataset	Window	Bookshelf	Picture	Counter	Blinds	Desk	Shelf	Curtain
FuseNet	52.9	48.0	68.1	56.4	67.2	15.1	12.6	56.5
MSF-CRF	53.8	60.7	66.6	63.5	45.6	26.0	17.3	58.5
Dataset	Dresser	Pillow	Mirror	Floormat	Clothes	Ceiling	Books	Fridge
FuseNet	28.4	44.3	30.7	38.8	22.9	75.5	21.2	11.9
MSF-CRF	45.3	49.3	54.9	19.0	15.9	69.2	10.7	21.0
Dataset	TV	Paper	Towel	Shower	Box	White board	Person	Nightstand
FuseNet	39.1	5.7	23.0	34.9	7	32.5	23.2	35.1
MSF-CRF	50.8	4.3	29.6	30.6	3.3	24.3	49.4	54.0
Dataset	Toilet	Sink	Lamp	Bathtub	Bag	Other struct	Other furniture	Other prop
FuseNet	75.0	32.4	40.1	51.9	1.6	19.8	10.8	45.7
MSF-CRF	78.7	32.9	40.2	50.1	1.0	9.3	18.7	46.8

表 4 40 个类别的 IoU 对比表

Table 4 Comparison of IoU of 40 categories

Dataset	Wall	Floor	Cabinet	Bed	Chair	Sofa	Table	Door
FuseNet	59.5	70.8	44.7	59.3	41.2	47.5	31.8	19.6
MSF-CRF	57.2	70.4	45.0	63.7	43.8	50.2	35.4	15.4
Dataset	Window	Bookshelf	Picture	Counter	Blinds	Desk	Shelf	Curtain
FuseNet	27.5	30.0	44.1	34.4	42.5	11.3	5.8	34.8
MSF-CRF	32.7	30.8	48.0	38.5	36.3	17.0	6.1	43.1
Dataset	Dresser	Pillow	Mirror	Floormat	Clothes	Ceiling	Books	Fridge
FuseNet	23.7	29.6	24.3	29.5	8.5	42.3	14.8	8.9
MSF-CRF	32.1	34.3	42.5	17.0	9.4	39.8	9.5	14.0
Dataset	TV	Paper	Towel	Shower	Box	White board	Person	Nightstand
FuseNet	31.5	3.8	18.5	20.3	4	22.4	14.8	26.6
MSF-CRF	39.1	3.7	21.8	26.1	2.4	20.7	32.9	40.1
Dataset	Toilet	Sink	Lamp	Bathtub	Bag	Other struct	Other furniture	Other prop
FuseNet	49.1	24.3	28.8	41.1	1.1	11.1	7.9	21.9
MSF-CRF	50.1	21.2	31.2	39.8	0.9	7.3	13.4	25.0

图 4 所示为 MSF-CRF 与 FuseNet 的语义分割结果图，从上到下依次是 RGB 图像、深度图、真实标签图、FuseNet 模型预测图和 MSF-CRF 模型预测图。从图 4 可以看到，MSF-CRF 能够提高分类的准确度，并优化图像的分割边缘，使分割边界更清晰平滑。第 1 列中，MSF-CRF 对床头柜和床的分割结果明显更优，更加接近真实标签；MSF-CRF 对于第 2 列的图画、第 3 列的柜橱、第 4 列的枕头以及第 5 列的床头柜和灯的分割结果明显更优，所得结果边界描绘清晰，分类更加准确，语义分割结果优于 FuseNet 结果。

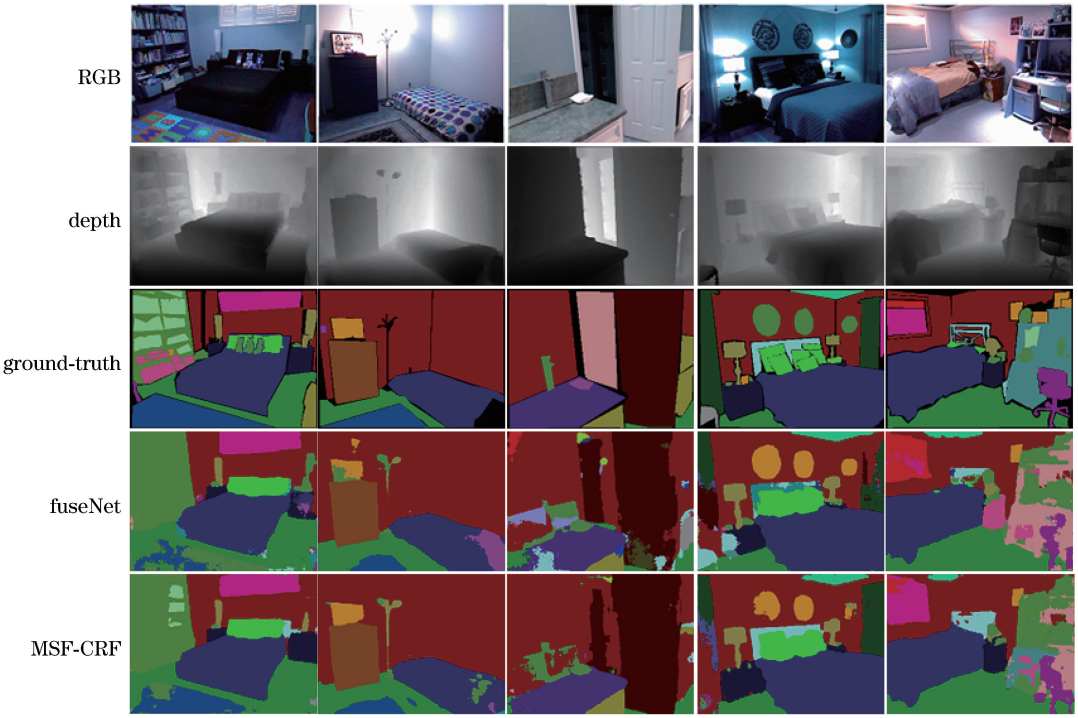

图 4 分割结果对比图

Fig. 4 Comparison of segmentation results

4 结 论

结合多尺度特征提取与全连接条件随机场，提出了一种利用不同尺度图像作为网络输入的 CNN 模型，融合图像的彩色信息与深度信息训练网络，利用 FullCRF 描绘图像的边界。在 NYUv2 数据集上的实验结果表明，多尺度的图像特征对图像语义分割结果具有重要影响，多尺度 CNN 模型对图像语义分割具有更好的表现力。该模型不仅更大程度地结合了物体的上下文信息，还优化了图像语义分割的边缘，使得网络在语义分割类别和边界方面都得到了更好的效果，提高了语义分割的准确性。

参 考 文 献

[1] Guo Y M, Liu Y, Georgiou T, *et al*. A review of semantic segmentation using deep neural networks [J]. International Journal of Multimedia Information Retrieval, 2018, 7(2): 87-93.

[2] Guo C C, Yu F Q, Chen Y. Image semantic segmentation based on Convolutional neural network feature and improved superpixel matching[J]. Laser & Optoelectronics Progress, 2018, 55(8): 081005.
郭呈呈, 于凤芹, 陈莹. 基于卷积神经网络特征和改进超像素匹配的图像语义分割[J]. 激光与光电子学进展, 2018, 55(8): 081005.

[3] Wei Y C, Zhao Y. A review on image semantic segmentation based on DCNN[J]. Journal of Beijing Jiaotong University, 2016, 40(4): 82-91.
魏云超, 赵耀. 基于 DCNN 的图像语义分割综述[J]. 北京交通大学学报, 2016, 40(4): 82-91.

[4] Zhang X M, Zhu X B, Cai Q, *et al*. Survey of the deep learning models for image semantic segmentation[J]. Chinese High Technology Letters, 2017, 27(9): 808-815.
张新明, 祝晓斌, 蔡强, 等. 图像语义分割深度学习模型综述[J]. 高技术通讯, 2017, 27(9): 808-815.

[5] Rother C, Kolmogorov V, Blake A. GrabCut: interactive foreground extraction using iterated graph

cuts[J]. ACM Transactions on Graphics, 2004, 23(3): 309-314.

[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, 2015: 3431-3440.

[7] Chen L C, Papandreou G, Kokkinos I, *et al*. Semantic image segmentation with deep Convolutional nets and fully connected CRFs[EB/OL]. (2016-06-07)[2018-10-25]. https://arxiv.org/abs/1412.7062.

[8] 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.

[9] 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, 2015: 1520-1528.

[10] Yu F, Koltun V. Multi-scale context aggregation by dilated Convolutions[EB/OL]. (2016-04-30)[2018-10-25]. https://arxiv.org/abs/1511.07122.

[11] 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.

[12] 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), July 21-26, 2017, Honolulu, HI, USA. New York: IEEE, 2017: 6230-6239.

[13] 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, 2016: 3194-3203.

[14] Xiong Z Y, Zhang G F, Wang J Q. Multi-scale feature extract for image sematic segmentation[J]. Journal of South-Central University for Nationalities (Natural Science Edition), 2017, 36(3): 118-124.
熊志勇, 张国丰, 王江晴. 基于多尺度特征提取的图像语义分割[J]. 中南民族大学学报(自然科学版), 2017, 36(3): 118-124.

[15] Jiang Y F, Zhang H, Xue Y B, *et al*. A new multi-scale image semantic understanding method based on deep learning[J]. Journal of Optoelectronics • Laser, 2016, 27(2): 224-230.
蒋应锋, 张桦, 薛彦兵, 等. 一种新的多尺度深度学习图像语义理解方法研究[J]. 光电子 • 激光, 2016, 27(2): 224-230.

[16] Liu D, Liu X J, Wang M Z. Semantic segmentation with multi-scale Convolutional neural network[J]. Remote Sensing Information, 2017, 32(1): 57-64.
刘丹, 刘学军, 王美珍. 一种多尺度 CNN 的图像语义分割算法[J]. 遥感信息, 2017, 32(1): 57-64.

[17] 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, 2015: 1529-1537.

[18] Lin G S, Shen C H, van den Hengel A, *et al*. Exploring context with deep structured models for semantic segmentation[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2018, 40(6): 1352-1366.

[19] Lafferty J D, McCallum A, Pereira F C N. Conditional random fields: probabilistic models for segmenting and labeling sequence data[C]// Proceedings of the Eighteenth International Conference on Machine Learning, June 28-July 01, 2001. San Francisco: Morgan Kaufmann Publishers Inc., 2001: 282-289.

[20] Krähenbühl P, Koltun V. Efficient inference in fully connected CRFs with Gaussian edge potentials[C]// Proceedings of the 24th International Conference on Neural Information Processing Systems, December 12-15, 2011, Granada, Spain. USA: Curran Associates Inc., 2011: 109-117.

[21] Vemulapalli R, Tuzel O, Liu M Y, *et al*. Gaussian conditional random field network for semantic segmentation[C]// 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 27-30, 2016, Las Vegas, NV, USA, 2016: 3224-3233.

[22] Hazirbas C, Ma L N, Domokos C, *et al*. FuseNet: incorporating depth into semantic segmentation via fusion-based CNN architecture[M]// Lai S H, Lepetit V, Nishino K, *et al*. Computer vision-ACCV 2016. Lecture notes in computer science. Cham:

Springer, 2017, 10111: 213-228.

[23] Silberman N, Hoiem D, Kohli P, *et al*. Indoor segmentation and support inference from RGBD images[M] // Fitzgibbon A, Lazebnik S, Perona P, *et al*. Computer vision-ECCV 2012. Lecture notes in computer science. Berlin, Heidelberg: Springer, 2012, 7576: 746-760.

[24] Gupta S, Arbeláez P, Malik J. Perceptual organization and recognition of indoor scenes from RGB-D images [C] // 2013 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 23-28, 2013, Portland, OR, USA. New York: IEEE, 2013: 564-571.

[25] Gupta S, Girshick R, Arbeláez P, *et al*. Learning rich features from RGB-D images for object detection and segmentation[M] // Fleet D, Pajdla T, Schiele B, *et al*. Computer vision-ECCV 2014. Lecture notes in computer science. Cham: Springer, 2014, 8695: 345-360.

[26] Deng Z, Todorovic S, Latecki L J. Semantic segmentation of RGBD images with mutex constraints[C] // 2015 IEEE International Conference on Computer Vision (ICCV), December 7-13, 2015, Santiago, Chile. New York: IEEE, 2015: 1733-1741.