融合深度图像的卷积神经网络语义分割方法

王孙平[1,2] 陈世峰[1]

[1](中国科学院深圳先进技术研究院 深圳 518055)
[2](中国科学院大学 北京 100049)

摘 要 该文提出了一种基于深度学习框架的图像语义分割方法，通过使用由相对深度点对标注训练的网络模型，实现了基于彩色图像的深度图像预测，并将其与原彩色图像共同输入到包含带孔卷积的全卷积神经网络中。考虑到彩色图像与深度图像作为物体不同的属性表征，在特征图上用合并连接操作而非传统的相加操作对其进行融合，为后续卷积层提供特征图输入时保持了两种表征的差异。在两个数据集上的实验结果表明，该法可以有效提升语义分割的性能。

关键词 语义分割；深度学习；深度图像
中图分类号 TG 156 **文献标志码** A

Depth-Aware Convolutional Neural Networks for Semantic Segmentation

WANG Sunping[1,2] CHEN Shifeng[1]

[1](*Shenzhen Institutes of Advanced Technology*, *Chinese Academy of Sciences*, *Shenzhen* 518055, *China*)
[2](*University of Chinese Academy of Sciences*, *Beijing* 100049, *China*)

Abstract In this paper, a deep learning-based image semantic segmentation method was studied. A neural network trained by point pair annotations of relative depth was used to predict depth images from common color images. By feeding the color and depth images into a fully convolutional networks with atrous convolution, accurate segmentation of the images could be obtained. As different representations of object properties, concatenate operation on the feature maps instead of traditional adding operation was used to fuse them. The differences between these two representations could be preserved when they were feed into the next convolutional layers. Experimental results on two different datasets show that, performance of semantic segmentation can be improved by the proposed method.

Keywords semantic segmentation; deep learning; depth image

1 引 言

图像的语义分割是计算机视觉中的一个基础问题，作为图像理解的重要一环，在自动驾驶系统、地理信息系统、医疗影像分析及机械臂物体抓取等实际应用中都有关键作用。其中，地理信

收稿日期：2018-03-19 **修回日期**：2018-04-14
基金项目：国家自然科学基金-深圳机器人基础研究中心项目(U1713203)
作者简介：王孙平，硕士研究生，研究方向为计算机视觉与机器学习；陈世峰(通讯作者)，博士生导师，研究方向为计算机视觉、图像处理，E-mail：shifeng.chen@siat.ac.cn。

息系统中的卫星遥感图像可使用语义分割的方法自动识别道路、河流、建筑物、植物等。在无人驾驶系统中，车载摄像头和激光雷达采集的图像，经语义分割可以发现道路前方的行人、车辆等，以辅助驾驶和避让。在医疗影像分析领域，语义分割主要用于肿瘤图像分割和龋齿诊断等。

图像的语义分割任务是指为一幅输入图像的每个像素分配一个语义类别，从而完成像素级别的分类。传统的语义分割主要使用手工设计的特征和支持向量机、概率图模型等方法。随着深度卷积神经网络在计算机视觉任务中刷新多项记录，包括图像分类[1-3]、物体检测[4-6]等，深度学习的方法也在语义分割任务中被广泛使用[7-9]。

卷积神经网络本身具有一定的对局部图像变换的不变性，可以很好地解决图像分类问题。但在语义分割任务中，分类的同时还需要得到精确的位置，这与局部图像变换的不变性相矛盾。在典型的图像分类模型中，多层网络组成了一个从局部到全局的金字塔结构。其中，顶层的特征图分辨率最低，虽然它包含全局的语义信息，但却无法完成精确的定位。全卷积神经网络[7]利用端到端、像素到像素的方法进行训练，对于顶层特征图定位不够精细的问题，采用跳跃结构综合了浅层精细的表观信息和深层粗糙的语义信息。

Chen 等[8]使用了另一种方案，直接在网络结构中减少了下采样的操作以得到更高的分辨率，并且利用了带孔的卷积，在不增加网络参数数量的前提下增大卷积核的感受野，从而获取更多关于图像像素的上下文信息。在信号处理领域，类似的方法最初用于非抽样小波变换的高效计算[10]。此外，还使用全连接的条件随机场方法[11]对卷积神经网络的输出结果进行后处理，达到了更精细的分割结果。

Zhao 等[12]在带孔卷积的网络模型基础上，提出了金字塔池化模块。该研究使用全局平均池化(Global Average Pooling，GAP)操作结果作为一个全局的上下文信息表征，与之前的特征图连接，使组合后的特征图同时包含全局的上下文信息和局部信息，是目前在 Pascal VOC 2012 数据集[13]上分割结果最好的方法之一。

图像中物理属性(如深度、表面法向量、反射率)的估计属于中层视觉任务，并可对高层视觉任务有所帮助。目前已经有许多数据驱动的深度估计方法[14-17]被提出，但这些方法受限于由深度传感器采集的图像数据集。尽管近年来消费级深度图像采集设备，如微软 Kinect、华硕 Xtion Pro 和英特尔 RealSense 等得到了大量使用，但仍主要局限于室内场景。对于镜面反射、透明或较暗物体等情况，常常会得到失败的结果。因此，在非受限的场景中难以用深度传感器得到可靠的深度图像。而对于语义分割任务而言，明确、清晰的边缘比精确的深度测量值本身更重要。有经验证据表明，相对于场景中某点的测量值，人类更擅长于估计两点之间的次序关系[18]。对于图像中两点的深度而言，“相等”“更深”“更浅”三种关系具有对单调变换的不变性，而且由人类对其标注，不存在场景受限的问题。Chen 等[19]构建了一个人类标注的“相对深度”点对数据集，并提出了一种以此标注端到端的训练卷积神经网络，从彩色图像预测深度图像的方法，显著改善了非受限场景下的单图深度感知。本文提出将彩色图像预测出的深度图像融入语义分割的卷积神经网络，利用深度图像的特性改善分割性能。

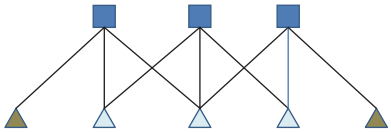
本文的主要创新点为：(1)使用从彩色图像预测的深度图像作为语义分割网络的输入；(2)用多分支输入、特征图合并连接融合深度图像特征的方法改善语义分割性能。实验结果表明，融合深度图像的特征可以显著提升语义分割性能。

2 融合深度图像的语义分割

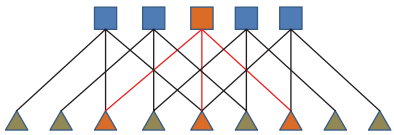
2.1 语义分割的卷积神经网络

典型的用于分类任务的卷积神经网络主要包含卷积层、激活函数、池化层和全连接层。一张输入图像经过网络由全连接层输出一个一维向量，再使用 Softmax 函数归一化后作为物体分类的得分。语义分割任务的卷积神经网络利用分类网络预训练得到的权重参数，采用全卷积的网络结构，直接对输入的三通道彩色图像和像素级的标注掩膜进行端到端的训练。由于取消了全连接层，可以适应任意尺寸的输入图像，并输出与之相同尺寸的分割结果。

卷积神经网络某一层输出的特征图中像素的位置对应于其在原图像中的位置称为“感受野”。由于网络结构中存在池化层或卷积层的下采样操作，最后卷积层输出的特征图分辨率往往很低。如果减少下采样操作来增加最后一个卷积层的特征图分辨率，那么会使卷积核的感受野变小，并带来更大的计算代价。而带孔的卷积操作在不改变网络权重参数数量的前提下，可以增大卷积核的感受野。图 1(a)为卷积核尺寸为 3 的普通卷积操作。图 1(b)是比率参数 r 为 2 的带孔卷积操作，在与图 1(a)相同的参数数量情况下，处理并输出了更高分辨率的特征图。



(a) 普通卷积(r=1)



(b) 带孔卷积(r=2)

图 1 卷积操作

Fig. 1 Convolution operation

对于一个一维的信号输入 $x[i]$ 和一个长度为 K 的卷积核 $w[k]$，比率参数为 r 的条件下，带孔卷积的输出 $y[i]$ 定义如下：

$$y[i]=\sum_{k=1}^{K}x[i+r\cdot k]w[k] \tag{1}$$

其中，比率参数 r 表示对输入信号的采样步长，普通卷积可视为比率参数 r=1 的特例。

本文使用的语义分割网络在使用带孔卷积的基础上，进行全局平均池化操作。其意义首先在于将特征图的所有信息合并到多个通道的单个点，形成一种全局的上下文先验信息；然后，再将其缩放回原特征图大小，与原特征图连接形成双倍通道数量的特征图，经过若干卷积层输出分割结果。由于特征图综合了这样的全局上下文信息，分割结果可得到明显改善[12]。

图 2 是本文使用语义分割模型的网络结构。其中，“彩色图像网络”以 VGG-16[2]作为基础模型，将 conv5 替换成 3 个比率参数为 2 的带孔卷积层，conv6 为一个比率参数为 12 的带孔卷积层，最后输出通道数量为 256 的特征图。“深度图像网络”分支仅包含 3 个卷积核尺寸为 3 的普通卷积层，通道数分别为 64、128、256。两个分支分别进行全局平均池化、缩放到原尺寸及合并连接操作，得到 512 个通道的特征图。网络中其他部分的作用在下面几个小节中介绍。

2.2 从彩色图像预测深度图像

目前使用稀疏的“相对深度”标注进行学习并预测出稠密的深度图像主要有两种方法，分别由 Zoran 等[20]和 Chen 等[19]提出。其中，Zoran 等[20]首先训练一个在图像的超像素中心之间预测深度次序的分类器，然后用能量最小化的方法恢复整体的深度，使这些次序关系达到一致，最后在超像素中进行插值来得到像素级别的深度图像。Chen 等[19]直接使用全卷积神经网络实现了彩色图像到深度图像的端到端训练，并提出了一种使用相对深度标注来训练网络的方法。对于相对深度标注需要设计一个合适的损失函数，基于

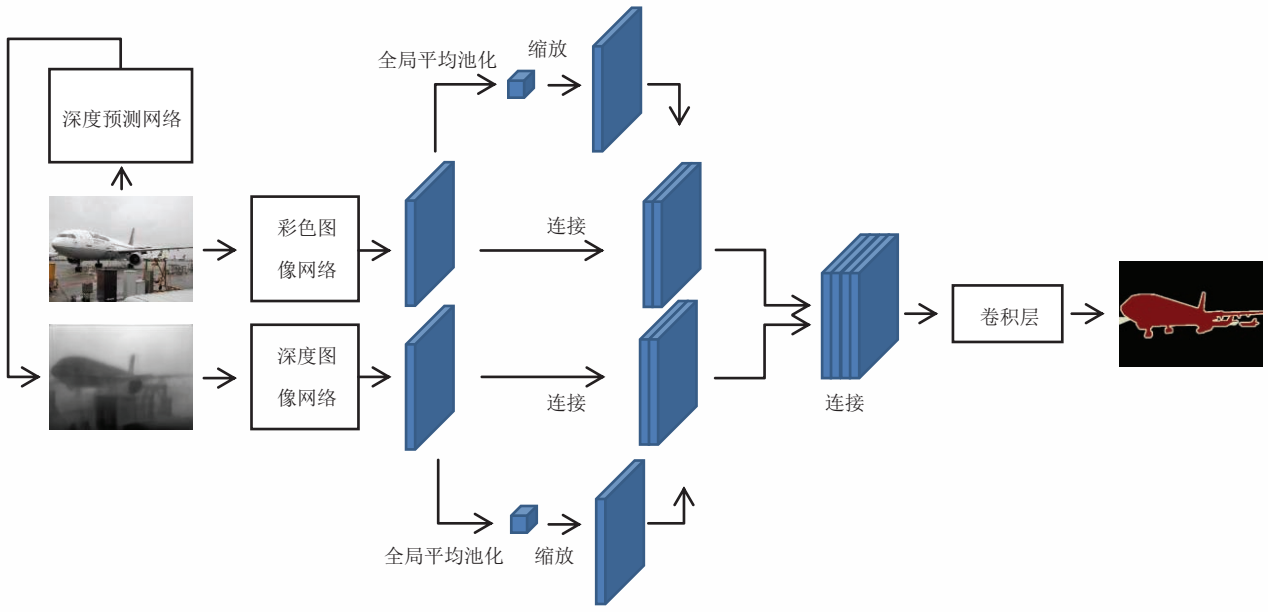


图 2　网络结构

Fig. 2　Network architecture

这样的原则：真实深度次序为“相等”时，预测的深度值差距越小越好；否则差距越大越好。

假设训练集的图像为 I，对其的 K 次查询 $R=\{(i_k, j_k, r_k)\}$, $k=1,\cdots,K$。其中，i_k、j_k 分别是第 k 次查询中两个点的位置。$r_k\in\{+1,-1,0\}$ 是对两点深度次序关系的标注，预测的深度图像为 z，则 i_k、j_k 对应的深度值为 z_{i_k}、z_{j_k}。定义如下损失函数：

$$L(I,R,z)=\sum_{k=1}^{K}\psi_k(I,i_k,j_k,r,z) \tag{2}$$

其中，$\psi_k(I,i_k,j_k,r,z)$ 是第 k 次查询的损失。

$$\psi_k(I,i_k,j_k,r,z)\begin{cases}\log\left[1+\exp\left(-z_{i_k}+z_{j_k}\right)\right], r_k=+1\\ \log\left[1+\exp\left(z_{i_k}-z_{j_k}\right)\right], r_k=-1\\ \left(z_{i_k}-z_{j_k}\right)^2, r_k=0\end{cases} \tag{3}$$

对于人类标注的相对深度点对，只需直接使用这个损失函数。对于深度传感器获取的深度图像，随机采样若干个点对即可转换为相同的形式。本文使用 Chen 等[19]的“相对深度”网络模型从彩色图像预测深度图像。该模型使用一种“沙漏”形的网络结构[21]，首先用深度传感器采集深度图像数据集进行预训练，然后在相对深度点对数据集上精调，预测的深度图像如图 3(b)所示。

相对深度的标注点对选择在很大程度上会影响网络训练的结果。如果随机在二维平面内选取两个点，会造成严重的偏置问题[19]：假设一个算法简单地认为底部的点比上方的点深度更近，有 85.8% 的概率会与人类标注的结果相同。一个更好的采样方法是从同一水平线上随机选取两个点，但这同样会造成简单认为中心的点深度更近的算法与人类标注结果有 71.4% 的概率相同。因此，一个合适的采样策略是从一条水平线上随机选取两个与其水平线中心对称的点，这样左边的点比右边的点深度更近的概率为 50.03%。

2.3　彩色与深度图像特征的融合

获得了估计的深度图像后，如何将深度图像与彩色图像的特征融合也是一个重要问题。一种简单的方法是将彩色图像的 3 个通道与深度图像的 1 个通道堆叠，形成 4 个通道的输入。然而，深度图像对物体的几何意义与彩色图像代表的光学意义并不相同，Long 等[7]实验也表明这种方式

并不能对性能有明显的改善。Gupta 等[22]提出了一种由深度信息导出的称为 HHA 的表征，由水平视差、距地面高度和局部表面法线与重力方向夹角组成，取得了更好的结果。但这种表征过于复杂，且未包含比深度图像本身更多的信息[23]。

本文提出的融合方法是：首先，分别用两个网络分支处理彩色图像和深度图像，得到 a 和 b 个通道的特征图；然后，用类似 PSPNet[12] 中金字塔池化模块的合并连接操作将两个分支的特征图合并成 $a+b$ 个通道的特征图；最后，经过若干卷积层输出分割结果。与特征图融合常用的相加操作相比，用合并连接操作可以使两个分支网络输出的特征更加独立，而非只为后续卷积层提供相同表征形式的特征图。如图 2 所示，将彩色图像和深度图像分支输出的两个通道数为 512 的特征图合并连接，得到 1 024 个通道的特征图。

初步实验发现，使用与最后卷积层输出的相同尺寸的较低分辨率深度图像和少量卷积层，可以取得比使用较高分辨率的深度图像和更多卷积层与池化层更好的结果。一方面，由于深度图像的预测网络输出的分辨率本身较低，高分辨率的深度图像仅仅是通过缩放得到；另一方面，不使用池化层更有利于网络输入和输出像素之间的位置对应。

3 实验

3.1 数据集

本文在 Pascal VOC 2012 数据集和 SUN RGB-D 数据集[24]上进行实验。其中，Pascal VOC 2012 数据集的图像包含 20 种类别的物体和一个背景类别，语义分割数据集被分成 3 个部分：训练集(1 464 张图像)、验证集(1 449 张图像)和测试集(1 456 张图像)。其中，验证集和测试集不包含训练集的图像。我们遵循惯例使用增加的包含 10 582 张训练图像的标注数据[25]，在 1 449 张图像上进行验证。SUN RGB-D 数据集是一个适用于场景理解的数据集，包含 4 种不同传感器获取的彩色图像与深度图像，包括 NYU Depth v2[26]，Berkeley B3DO[27]和 SUN3D[28]等数据集，共有 10 335 张 RGB-D 图像和其像素级的语义分割标注，其中包含 5 285 张训练图像和 5 050 张测试图像。

3.2 数据集处理

本文对两个数据集采取了适合自然图像的常用数据增强方法：随机缩放、镜像和裁剪填充。其中，(1)随机缩放：将图像随机缩放为原来的 0.5～1.5 倍；(2)镜像：以 50% 的概率对图像进行水平翻转；(3)裁剪填充：以 500×500 的固定尺寸裁剪或填充图像(若尺寸不足则填充灰色)。

网络的输入包括彩色图像和深度图像。由于 Pascal VOC 2012 数据集不含深度传感器采集的深度图像，本文使用从彩色图像预测得到的深度图像作为输入。对于 SUN RGB-D 数据集，本文对深度传感器采集的深度图像、彩色图像预测得到的深度图像均作为输入进行了实验。

3.3 实验过程及参数

本文使用如图 2 所示的网络结构，首先使用深度预测的网络从彩色图像预测出深度图像，然后将彩色图像和深度图像分别输入两个卷积神经网络分支。其中，彩色图像的分支是以 VGG-16 模型为基础的包含带孔卷积的网络，权重由 ImageNet[29]上预训练的 VGG-16[2]的权重进行初始化，其他卷积层均为 Xavier 随机初始化[30]。两个网络分支经过合并连接后，再通过两个卷积层输出分割结果。

网络训练的批尺寸(Batch Size)参数为 10，输入的彩色图像大小为 500×500，深度图像和用于对比的灰度图像大小为 63×63。初始学习率为 0.000 1(最后一个层为 0.001)，按照多项式函

数衰减，训练迭代 20 000 次后停止。动量参数为 0.9，权重衰减参数为 0.000 5。实验均在 NVIDIA GeForce TITAN X GPU 上进行。分割性能以各个类别的像素交并比 IoU（Intersection-over-Union）得分平均数作为评价指标。

本文在两个数据集上设计了 5 个实验，将输入图像分为：

（1）VOC 数据集，彩色图像和预测的深度图像；

（2）VOC 数据集，彩色图像和灰度图像；

（3）SUN 数据集，彩色图像和预测的深度图像；

（4）SUN 数据集，彩色图像和深度传感器采集的深度图像；

（5）SUN 数据集，彩色图像和灰度图像。

其中，灰度图像由彩色图像转换而成，用于替代深度图像输入网络作为对照。

4　实验结果

4.1　Pascal VOC 数据集实验对比

为了对比有无深度图像信息的效果，我们比较了实验（1）、（2）中不同类别的分割性能，结果如表 1 所示。由表 1 可以看出，对于大多数类别，融合预测的深度图像特征都能对分割性能有效提升，只有颜色特征明显、图像中尺寸较小的盆栽植物（plant）类别下降了 0.1%。原因是深度预测模型的输出分辨率较低，对于图像中尺寸小的物体深度预测结果较差。其中，结构特征明显且图像中尺寸较大的物体提升明显，如飞机（aero）、船（boat）和沙发（sofa）等，与深度图像本身物理意义的作用相符，证实了该方法的有效性。

Pascal VOC 数据集上的分割结果如图 3 所示。由图 3 可以观察到，即使对于室外的场景，深度图像仍能捕获到清晰的物体轮廓。在包含深度图像输入的情况下，由于深度图像较为清晰的边缘，物体边界处的分割也达到了更好的效果。

表 1　Pascal VOC 数据集分割 IoU 结果

Table 1　IoU results on Pascal VOC dataset

类别	IoU（%）	
	彩色＋深度	彩色＋灰度
bkg	**91.1**	90.8
aero	**78.6**	76.6
bike	**33.6**	31.4
bird	**79.6**	78.5
boat	**57.3**	55.3
bottle	**66.1**	64.2
bus	**87.1**	86.0
car	**77.7**	77.4
cat	**83.1**	82.6
chair	**32.0**	30.4
cow	**69.1**	67.4
table	**52.3**	51.5
dog	**75.7**	**75.7**
horse	**71.2**	71.0
mbike	**69.2**	68.7
person	**76.7**	76.1
plant	44.5	**44.6**
sheep	**70.2**	69.7
sofa	**42.7**	40.3
train	**79.7**	79.2
tv	**60.7**	59.0
mIoU	**66.6**	65.5

注：mIoU 表示均交并比；粗体代表较好的结果

4.2　SUN RGB-D 数据集实验对比

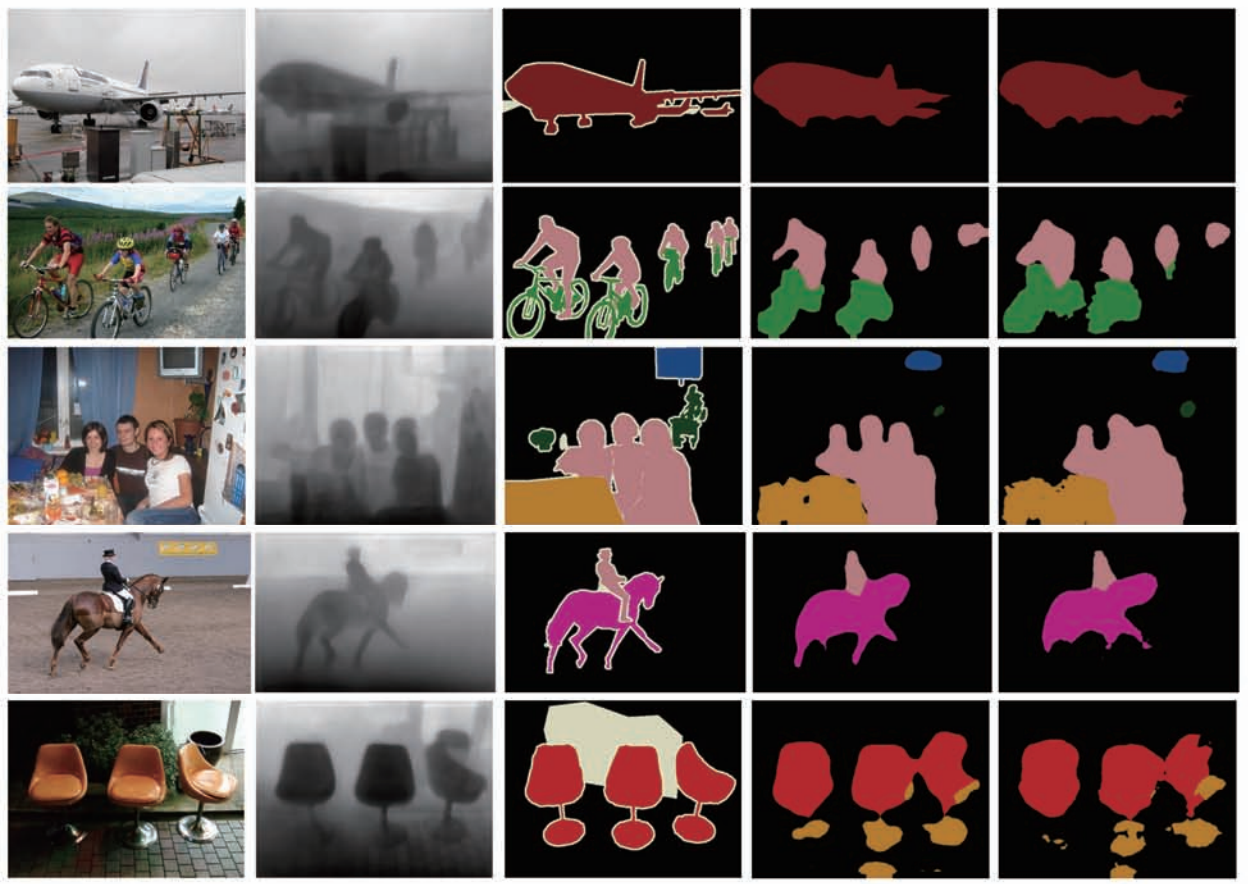
表 2 比较了在 SUN RGB-D 数据集上预测的深度图像、使用传感器采集的深度图像和无深度

表 2　SUN RGB-D 数据集的 mIoU 结果

Table 2　The mIoU results of SUN RGB-D dataset

方法	mIoU（%）
彩色＋灰度	38.1
彩色＋传感器深度	38.4
彩色＋预测深度	**38.5**

注：粗体代表较好的结果



(a) 原彩色图像 (b) 预测的深度图像 (c) 图像标注 (d) 实验(1)的结果 (e) 实验(2)的结果

图 3 Pascal VOC 数据集上的分割结果

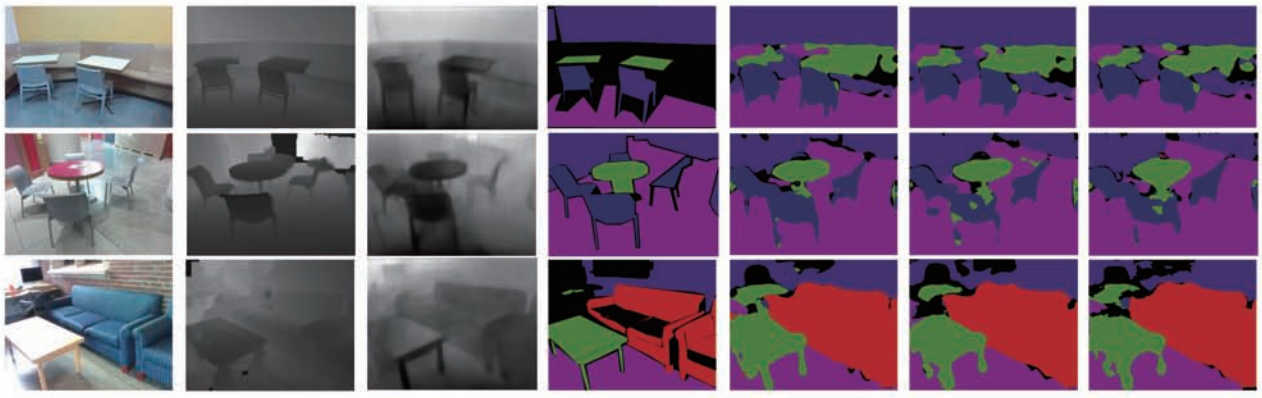
Fig. 3 Segmentation results on Pascal VOC dataset

信息 3 种情况下，即实验(3)、(4)、(5)的分割结果。由图 3 可以看出，使用深度图像的分割结果较好，而且使用预测的深度图像结果稍好于使用传感器深度图像的结果。这说明对于语义分割任务，预测的深度图像能够起到替代传感器采集的深度图像的作用。

SUN RGB-D 数据集上的分割结果如图 4 所示。由图 4 可以看到，第一行深度图像能清晰地分辨出椅脚，表明使用深度图像的实验对椅脚部分的分割效果较好。第二、三行的传感器深度图像存在一些像素值缺失的区域和噪声，而预测的深度图像虽然深度测量值不够精确，但保持了比较完整的物体形态。这是预测的深度图像能够取得稍好的分割结果的一个原因。

5 讨 论

图像中物体的语义和深度具有密切的联系，获取并利用深度图像可以对语义分割任务起到很大的辅助作用。但非受限环境下深度图像的获取是一个挑战。深度传感器获取的深度图像数据集局限于室内环境和固定场景(如公路等)，而且目前在语义分割任务中对深度信息的利用方法仍存在很多缺陷[22,23]。本文使用卷积神经网络从彩色图像中预测出深度图像，以带孔卷积的语义分割网络为基础设计了一个多分支网络，用特征图合并连接的方式融合彩色图像和深度图像的特征进行语义分割。带孔的卷积在不增加网络参数数量的前提下增大了卷积核的感受野，使其包含更



(a) 原彩色图像　(b) 传感器采集的深度图像　(c) 预测的深度图像　(d) 图像标注　(e) 实验(3)结果　(f) 实验(4)结果　(g) 实验(5)结果

图 4　SUN RGB-D 数据集上的分割结果

Fig. 4　Segmentation results on SUN RGB-D dataset

多的图像上下文信息，从而改善分割性能[8]。在其他条件相同的情况下，本文提出的含有深度图像信息与合并连接操作的网络和不含深度图像信息(以灰度图像作为替代)的网络相比，在 Pascal VOC 数据集上的均交并比(mIoU)提升了 1.1%。在 SUN RGB-D 数据集上的分割结果表明，使用预测的深度图像训练的网络与使用传感器获取的深度图像的网络性能接近，且都好于不含深度图像的网络。这说明预测的深度图像可以代替传感器采集的深度图像改善语义分割的结果。但当前方案所使用的相对深度点对数据集标注数量较少，网络模型也有很大的改进空间[19]。在卷积神经网络中利用深度图像仍然是一个非常值得研究的问题。

6　结　论

本文提出一种多分支网络和特征图连接的方法融合深度图像特征，使用彩色图像预测的深度图像解决非受限场景下深度图像获取困难的问题。利用金字塔池化模块中使用的合并连接操作连接彩色图像和深度图像的特征图，使两种类型的特征互为补充且保持独立的表征。在两个数据集上的分割结果表明，该方法能够利用深度图像细化物体的边缘，提升语义分割的性能。目前，仍然没有很好的方法在卷积神经网络中充分利用深度图像，下一步将尝试对语义分割模型的损失函数和网络结构进行改进。

参 考 文 献

[1] Krizhevsky A, Sutskever I, Hinton GE. Imagenet classification with deep convolutional neural networks [C] // Proceedings of the 25th International Conference on Neural Information Processing Systems, 2012: 1097-1105.

[2] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition [J]. Computer Science, 2014, arXiv:1409.1556.

[3] He KM, Zhang XY, Ren SQ, et al. Deep residual learning for image recognition [C] // IEEE Conference on Computer Vision and Pattern Recognition, 2016: 770-778.

[4] Girshick R, Donahue J, Darrell T, et al. Rich feature hierarchies for accurate object detection and semantic segmentation [J]. Computer Science, 2013: 580-587.

[5] Girshick R. Fast R-CNN [J]. Computer Science, 2015, arXiv:1504.08083.

[6] Ren S, He K, Girshick R, et al. Faster R-CNN: towards real-time object detection with region proposal networks [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2015, 39(6): 1137-1149.

[7] Long J, Shelhamer E, Darrell T. Fully convolutional

networks for semantic segmentation [C] // IEEE Conference on Computer Vision and Pattern Recognition, 2015: 3431-3440.

[8] Chen LC, 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, 2016, 40(4): 834-848.

[9] Zheng S, Jayasumana S, Romera-Paredes B, et al. Conditional random fields as recurrent neural networks [J]. Computer Science, 2015, doi: 10.1109/ICCV.2015.179.

[10] Holschneider M, Kronland-Martinet R, Morlet J, et al. A real-time algorithm for signal analysis with the help of the wavelet transform [M] // Wavelets. Springer Berlin Heidelberg, 1990: 286-297.

[11] Krähenbühl P, Koltun V. Efficient inference in fully connected CRFs with gaussian edge potentials [J]. Computer Science, 2012: 109-117.

[12] Zhao HS, Shi JP, Qi XJ, et al. Pyramid scene parsing network [C] // IEEE Conference on Computer Vision and Pattern Recognition, 2017: 6230-6239.

[13] Everingham M, Gool LV, Williams CKI, et al. The pascal visual object classes (VOC) challenge [J]. International Journal of Computer Vision, 2010, 88(2): 303-338.

[14] Karsch K, Liu C, Kang SB. Depth transfer: depth extraction from videos using nonparametric sampling [M] // Dense Image Correspondences for Computer Vision. Springer International Publishing, 2016: 775-788.

[15] Saxena A, Sun M, Ng AY. Make3D: learning 3D scene structure from a single still image [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2009, 31(5): 824-840.

[16] Eigen D, Fergus R. Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture [C] // IEEE International Conference on Computer Vision, 2015: 2650-2658.

[17] Li B, Shen CH, Dai YC, et al. Depth and surface normal estimation from monocular images using regression on deep features and hierarchical CRFs [C] // IEEE Conference on Computer Vision and Pattern Recognition, 2015: 1119-1127.

[18] Todd JT, Norman JF. The visual perception of 3-D shape from multiple cues: are observers capable of perceiving metric structure? [J]. Perception & Psychophysics, 2003, 65(1): 31-47.

[19] Chen WF, Fu Z, Yang DW, et al. Single-image depth perception in the wild [C] // Advances in Neural Information Processing Systems, 2016: 730-738.

[20] Zoran D, Isola P, Krishnan D, et al. Learning ordinal relationships for mid-level vision [C] // IEEE International Conference on Computer Vision (ICCV), 2015: 388-396.

[21] Newell A, Yang K, Deng J. Stacked hourglass networks for human pose estimation [M] // Stacked Hourglass Network for Human Pose Estimation. Springer International Publishing, 2016: 483-499.

[22] Gupta S, Girshick R, Arbeláez P, et al. Learning rich features from RGB-D images for object detection and segmentation [C] // European Conference on Computer Vision, 2014: 345-360.

[23] Hazirbas C, Ma L, Domokos C, et al. Fusenet: incorporating depth into semantic segmentation via fusion-based cnn architecture [C] // Asian Conference on Computer Vision, 2016: 213-228.

[24] Song SR, Lichtenberg SP, Xiao JX. SUN RGB-D: a RGB-D scene understanding benchmark suite [C] // IEEE Conference on Computer Vision and Pattern Recognition, 2015.

[25] Hariharan B, Arbeláez P, Bourdev L, et al. Semantic contours from inverse detectors [C] // IEEE International Conference on Computer Vision (ICCV), 2011: 991-998.

[26] Silberman N, Hoiem D, Kohli P, et al. Indoor segmentation and support inference from rgbd images [C] // European Conference on Computer Vision, 2012: 746-760.

[27] Janoch A, Karayev S, Jia Y, et al. A category-level 3D object dataset: putting the kinect to work [C] // IEEE International Conference on Computer Vision, 2011: 1168-1174.

[28] Xiao JX, Owens A, Torralba A. SUN3D: a database of big spaces reconstructed using SfM and object labels [C] // IEEE International Conference on Computer Vision, 2013: 1625-1632.

[29] Russakovsky O, Deng J, Su H, et al. Imagenet large scale visual recognition challenge [J]. International Journal of Computer Vision, 2015, 115(3): 211-252.

[30] Glorot X, Bengio Y. Understanding the difficulty of training deep feedforward neural networks [J] Journal of Machine Learning Research, 2010, 9: 249-256.