文章编号：1005-5630(2023)05-0062-10 **DOI**: 10.3969/j.issn.1005-5630.2023.005.008

一种用于深度补全的双分支引导网络

秦晓飞，胡文凯，班东贤，郭宏宇，于 景
（上海理工大学 光电信息与计算机工程学院，上海 200093）

摘要：深度信息在机器人、自动驾驶等领域中有着重要作用，通过深度传感器获取的深度图较为稀疏，研究人员为了补全缺失的深度信息提出了大量方法。但现有方法大多是针对不透明对象，基于卷积神经网络的强大表征能力，设计了一个双分支引导的编解码结构网络模型，通过针对透明物体的以掩码图为引导的编码分支，提升网络对透明物体特征信息的提取能力，并且使用谱残差块连接编解码部分，提高了网络训练稳定性及获取物体结构信息的能力，除此之外，还加入了注意力机制以提升网络空间和语义信息的特征建模能力。该网络在两个数据集上都达到了领先的效果。

关键词：深度补全；多数据引导；卷积神经网络；谱残差块；注意力机制
中图分类号：TP 391.4 **文献标志码**：A

A dual−branch guided network for depth completion

QIN Xiaofei，HU Wenkai，BAN Dongxian，GUO Hongyu，YU Jing
（School of Optical-Electrical and Computer Engineering, University of Shanghai for Science and Technology, Shanghai 200093, China）

Abstract: Depth information plays an important role in the fields of robotics and autonomous driving. The depth map obtained by the depth sensor is relatively sparse. Researchers have proposed a large number of methods to complement the missing depth values. However, most of the existing methods aim at opaque objects. Based on the powerful representation ability of convolution neural network, this paper designed a dual-branch-guided encoder-decoder structure network. Through mask-guided branch for transparent objects, it improves the ability of the network to extract feature information of transparent objects. And spectral residual blocks improves the stability of network in training process and the ability to obtain object structure information. In addition, attention mechanism is added to improve the feature modeling ability of network space and semantic information. The network achieves state-of-the-art results on all two datasets.

Keywords: depth completion; multiple data guidance; convolution neural network; spectral residual block; attention mechanism

收稿日期：2022−12−17
基金项目：国家自然科学基金重点项目(92048205)；国家留学基金(202008310014)
第一作者：秦晓飞(1982—)，男，高级工程师，研究方向为人工智能算法。E-mail：xiaofei.qin@foxmail.com

引　言

深度信息在计算机视觉领域有着广泛的应用，例如场景理解、自动驾驶、增强现实、移动机器人等[1-5]。这些应用依赖于对物体准确的深度预测，例如机器人抓取要求视觉传感器获取到物体准确的深度信息，从而计算出物体相对于机械夹爪的位置，进而实施抓取。深度图是通过深度传感器来获得的，如激光雷达等传感器，但是由于物体表面反光，或者光发生折射、透射，会使得深度信息缺失，特别是对于透明物体。现代工业中有很多透明的材料，所以机器人抓取存在需要处理透明物体的场景，但是目前的方法大多比较依赖深度传感器获取的深度信息，因此很少能直接应用于有透明物体的场景。透明物体的物理特性会导致光路因反射和折射而失真，从而产生有噪声的深度图，因此，许多基于深度信息的算法无法处理日常生活中随处可见的透明物体，如塑料瓶、玻璃容器等。

深度图是一种表达三维场景信息的表现形式，在三维图形中，深度图在视觉上体现为灰度图。在不考虑硬件、环境等外在因素的影响下，深度图中的每个像素值代表了传感器到场景中各点距离的等比例放缩，所以它可以直接反映物体朝向传感器面的几何形状。对于 RGB-D 相机而言，一般情况下，RGB 图像和深度图像是被校准的，所以彩色通道和深度通道的像素点是一一对应的。

深度补全是一种将稀疏的深度图中的深度值空洞补全的技术。早期对于深度图的补全，有基于传统的图像滤波器，例如，Chen 等[6] 提出使用自适应双边滤波器来补全 Kinect 相机拍摄的稀疏的深度图，消除不匹配的边界区域。Liu 等[7] 提出的三边滤波器，可以在保留深度图像边缘的同时抑制其他模态数据引导信息中的伪影。Alhwarin 等[8] 利用不同立体相机获取的视图差与 RGB-D 相机获取的深度图相融合，来填充深度图中由对象的透明或反射光干扰造成的深度缺失区域。Chiu 等[9] 提出了一种通过加权通道的早期融合与晚期融合的方案，对稀疏的深度图进行补全。Chen 等[10] 利用 RGB-D 相机获取到的图像的上下文语义信息为约束，补全稀疏的深度图。

随着深度学习技术的发展，并且得益于卷积神经网络（convolution neural network，CNN）的表征能力，近些年来提出了大多基于 CNN 的深度补全算法。就输入数据模态而言，深度补全网络分为两大类：单模态数据的深度补全网络，即仅有稀疏的深度图作为网络的输入；多模态数据的深度补全网络，即除深度图外，还有其他模态的数据作为引导。对于多模态数据的算法，例如使用相同场景下彩色相机获取的高质量彩色图像和深度相机获取的稀疏深度图作为网络的输入。Zhang 等[11] 使用 VGG-16 为 backbone 的编解码网络，通过建立物体表面法向量和深度信息之间的联系，从而使用彩色图像的表面法向量来补全稀疏的深度图。Qiu 等[12] 也将类似的表面法线作为引导信息扩展到室外环境，从 LiDAR传感器获取的稀疏深度图中补全缺失的深度值。上述两种方法，都是将物体表面的法向量与稀疏深度图信息进行融合，从而利用了物体表面法线作为另一种引导信息来补全稀疏的深度图。Ma 等[13] 提出了一种自监督的网络，通过彩色图像及深度图的视频帧之间的一致性，来建立从稀疏深度图到密集深度图之间的映射关系。Eldesokey等[14] 提出了一种新的标准在 CNN 层之间传播置信度，并与 RGB 信息相结合补全稀疏深度图。Cheng 等[15] 使用卷积神经网络学习像素之间的亲和力，协助补全缺失的深度值。Huang 等[16] 使用一种自注意力机制和边界一致性的端到端网络进行深度图补全。

但是，大多数的深度补全的方法都是针对室内的家具以及室外的街景，忽略了日常生活常见及现代工业中的透明物体。对于标准的 3D 传感器，如何扫描透明物体是个难题，传统的双目、结构光或 ToF RGB-D 镜头对透明物体难以产生准确的深度估计，在大多数情况下，透明物体会显示为一堆无效的噪点或失真的近似平面。原因是传统的 3D 传感器算法是假设物体的表面符合完全漫反射，即所有方向上的光都是均匀的，但是对于透明物体来说，该假设是不成立的。Sajjan 等[17] 为了将深度补全方法适用于透明物体，提出了 ClearGrasp，该方法预测物体表面法线，透明物体的掩膜和遮挡边界，并使用这些输

出优化和完善透明表面的稀疏深度图。Zhu 等[18]提出了一种两阶段方法，其中包含学习局部隐式深度函数（LIDF）的网络和自校正完善模型，用来针对透明物体的深度补全。

由于仅使用彩色图引导的深度补全方法容易受到图像中物体的阴影和表面的反射影响，受 PENet[19] 启发，本文采用一种双分支输入的编解码结构网络，其中一个分支旨在提取以彩色图为主导的特征信息，另一个分支用于提升网络对透明物体特征信息的提取能力，加入了透明物体的掩膜图作为另一种引导信息，将多尺度两种模态特征信息进行融合，从而补全稀疏的深度图。受 DepthGrasp[20] 启发，本文使用谱残差块堆叠形成的模块来连接编码和解码模块，并且在网络中加入一种注意力机制，从而提高网络对空间和语义信息的特征提取能力。

本文的贡献主要体现在两个方面：首先是设计了一种双分支输入引导的编解码结构的深度补全网络，其中包括利用透明对象的掩膜图为引导的输入提升网络对透明物体特征信息提取能力和不同模态数据特征显著性的方法；除此以外，本文还提出了将注意力机制用于提升网络对数据空间信息和语义信息的建模能力。

1　网络结构和原理

本文设计了一个 Encoder-Decoder 结构的网络，该网络使用不同模态数据引导从而补全稀疏的深度图。网络结构如图 1 所示，整体结构包括编码器部分、解码器部分，以及使用谱残差块（spectral residual block, SRB）进行两部分的连接，同时还加入了注意力机制以提高特征表达能力。对于编码器部分，包括两个分支分别提取以不同数据模态为引导的特征信息，其中一个分支以彩色图作为主导，用于提取主要依赖于 RGB 信息的深度特征图，另一个分支主要以掩码图作为主导，用于提取针对透明物体的深度特征图，并且可以提供更可靠的物体边界。对于谱归一化残差块部分，包含提高网络训练稳定性的谱归一化操作，以及用于获取物体结构信息和区分物体几何形状的残差块。输入解码器部分的特征是经过注意力模块后的融合特征。

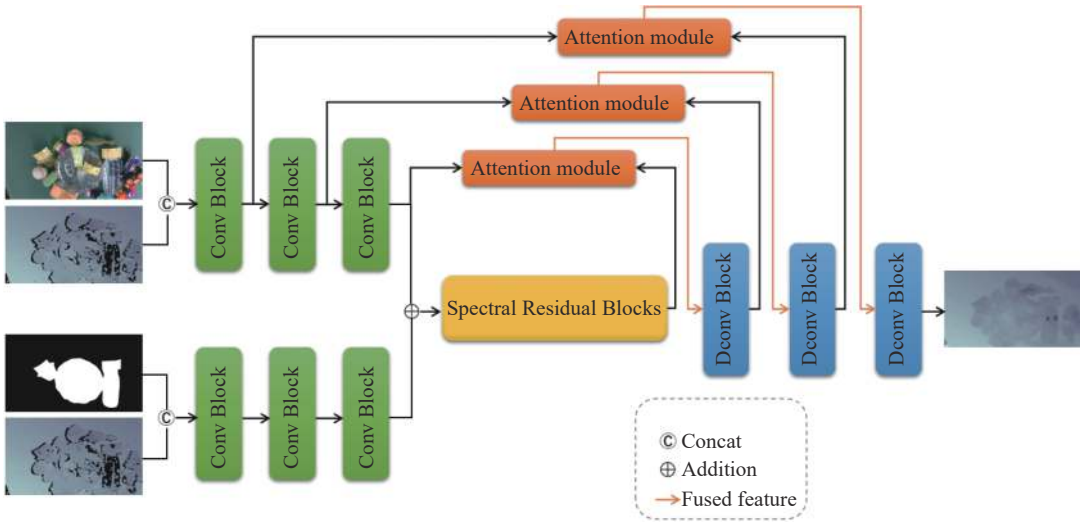

图 1　网络结构

Fig. 1　Model architecture

1.1　编码器模块

对于编码器部分，两分支输入的目的是从各自的分支中彻底利用彩色图和掩码图为主的信息，并且使得两种模态的特征信息能够有效的融合。

以彩色图为主导的分支主要目的在于从 RGB 图像中提取物体结构及几何形状的特征信息，从而有助于预测密集深度图。为了更加有效且准确地进行稀疏深度图的补全，本文将对齐的稀疏深度图与彩色图合并输入到彩色图为主导的

分支中，以帮助对齐的稀疏深度图进行深度预测。在两分支中，解码器具有 3 个 2D 卷积块，每个卷积块中包含有卷积层、批归一化层（BN）和一个 ReLU 激活层。并且输入图像或特征图每经过 1 个卷积块，分辨率大小降为原来的 1/4。针对彩色图引导分支，由于输入的是 RGB-D 数据，故编码器模块的第 1 个卷积层输入 channel 数为 4 通道，64 个卷积核，卷积核大小（kernel size）为 3，步长（stride）为 2，padding 为 1。针对掩膜图引导分支，由于输入是二值掩膜图和灰度图，所以编码器模块的第一个卷积层输入 channel 数为 2 通道，其余参数都一致。在进入连接模块前，两个分支输出的通道数一致，经过逐元素相加后将两分支的特征信息进行融合。

虽然颜色图和稀疏的深度图都用作输入，但是该分支提取了深度预测的颜色优势特征信息，从而可以便于利用颜色图像中的对象结构信息来学习物体边界周围的深度信息。以掩码图为主导的分支目的有两个：首先为了使得网络能够对于透明物体的深度补全效果更好，本文加入了针对透明物体的掩码图，以此让解码器可以更加关注透明物体的特征信息；除此之外，加入掩码图主导的分支可以帮助更好的学习场景中的语义线索，有助于预测具有可靠性边界物体的深度信息，从而减少稀疏深度图中出现的伪影。

1.2　编解码连接模块

本文使用谱残差块堆叠形成的模块连接编码器与解码器部分，谱残差块的作用是有效地捕捉物体结构信息及区分几何形状，模块中的谱归一化操作可以提高网络训练的稳定性。

在网络的训练过程中，由于两个分支提取的不同模态数据的特征信息，数据分布的密度在高维空间中不够准确，所以网络学习目标分布的多模数据结构的能力比较弱，从而导致训练的不稳定性。因此，本文引入谱归一化的方法以稳定网络的训练。此外，SRB 模块中的残差块有助于特征图的传递，从而利于网络获取物体的结构信息以及区分物体的几何形状；也有助于阻止网络梯度消失的情况，从而利于网络的训练。如图 2 所示，本文设计的 SRB 模块，包含了卷积块，谱归一化操作[21]以及 *LeakyReLU* 激活函数。其中使用谱归一化操作代替批归一化操作的原因是，批归一化操作需要通过两次输入数据来计算最小批次的统计值，然后再将输出标准化，因此对于大型网络来说，可能会消耗超过 1/4 的总训练时长。但对于谱归一化而言，不需要额外训练参数，并且在实际操作中不受内存带宽的限制。输入 SRB 模块的特征是来自两个分支融合后的特征图，给定特征图 M，SRB 模块的输出特征图定义为

$$S_{\text{out}} = M \oplus SN(C(LeakyReLU(C(M)))) \tag{1}$$

式中：C 表示 2D 卷积块；SN 表示谱归一化操作；*LeakyReLU* 是一种激活函数，与 ReLU 激活函数的区别在于 ReLU 输入小于 0 的部分值都为 0，而 *LeakyReLU* 输入小于 0 的部分，值为负，且有微小的梯度；⊕表示逐元素相加操作。

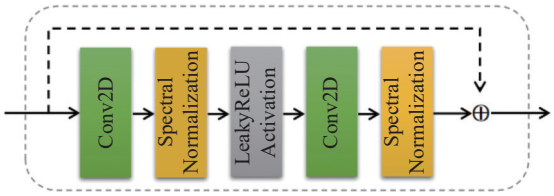

图 2　谱残差块

Fig. 2　Spectral Residual Block

1.3　注意力模块和解码器模块

注意力机制一经提出后，在各种应用场景得到了广泛的应用，近年来也出现了各种不同结构的注意力机制变体。从本质上看，注意力机制的原理是将特征图上的特征值看作是所有特征值的加权和，可以用公式简单表示为

$$Attention(Query, Source) = \sum_i^L Similarity(Query, Key_i) \cdot Value_i \tag{2}$$

式中 *Similarity* 是一个计算相似度权重的函数，是通过网络学习得到的。本文设计的网络中使用到的注意力模块，是受 CBAM[22]的启发，并且针对当前的任务做了相应的改进。首先如果将 CBAM 直接迁移到当前任务，可以得到如图 3 所示的实现方法。

对于原始的注意力模块，在给定输入维度为 $C\times H\times W$ 的特征图 F 时，该注意力模块依次得到维度为 $C\times 1\times 1$ 的通道注意力图 M_c 和维度为

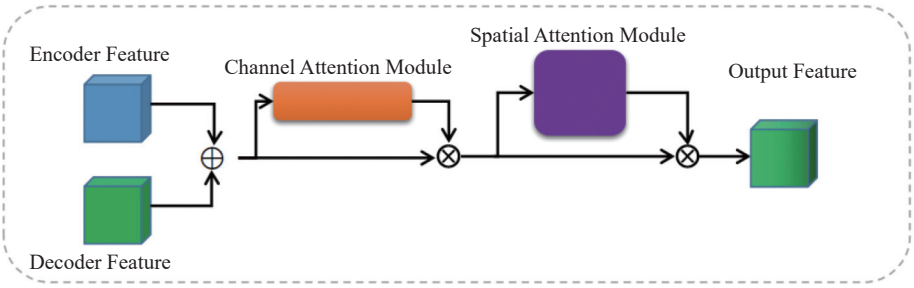

图 3　注意力模块使用方式

Fig. 3　Method of using attention module

$1\times H\times W$ 的空间注意力图 M_s。整体的注意力机制的运行流程可以简单表示为

$$F_{im} = M_c(F)\otimes F \tag{3}$$

$$F_{fo} = M_s(F_{im})\otimes F_{im} \tag{4}$$

式中：F_{im} 是经过通道注意力模块后的中间特征图；F_{fo} 是最终特征。

如图 3 所示，将编码器与解码器中相同分辨率的特征图按通道进行拼接后的的特征图送入通道注意力模块，将得到的通道注意力图与输入特征图进行逐元素相乘后得到中间特征图，再将中间特征图送入空间注意力模块中，最终再将得到的空间注意力图与中间特征图进行逐元素相乘后得到最终的特征图。如图 1 所示，再将经过 CBAM 后的特征图输入到解码器部分的下一个反卷积块中，每个反卷积块中包括有 2D 反卷积层、批归一化层（BN）和 ReLU 激活层。对于反卷积层，输入通道数与注意力模块的输出通道数对齐，反卷积核数为 64，kernel size 为 3，stride 为 2，padding 为 1，output padding 为 1。最终输出会经过一个卷积层，输入通道数与最后一层反卷积层的输出通道数对齐，输出通道数为 1，kernel size 为 2。

由于在网络的浅层，即编码器部分，特征图数据中包含的空间信息更加丰富，而在网络的深层，即解码器部分，特征图数据中蕴含的语义信息更加丰富。因此如图 4 所示，本文将空间注意力模块仅用于处理编码器部分的输出特征图，从而提升网络对图像数据中的物体结构细节信息的建模能力，而将通道注意力模块仅用于处理解码器部分的输出特征图，从而提升网络对特征中语义信息的建模能力。经过实验对比，图 4 所示的处理编解码输出特征的方法，能更好地补全稀疏深度图。

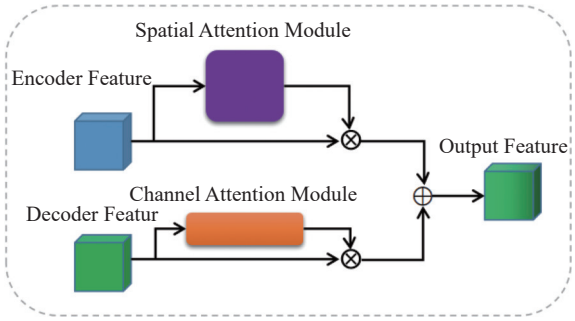

图 4　注意力模块使用方式变体

Fig. 4　The variant method of using attention module

1.4　损失函数

在网络训练时，本文采用均方误差（mean squared error, MSE）计算损失值，损失函数定义为

$$L = \frac{1}{m}\sum_{p\in Q_v}\|G_p - P_p\|^2 \tag{5}$$

式中：G 和 P 分别表示基准深度图和预测的深度图；Q_v 表示在基准深度图中有效的深度值像素集合；m 表示有效的深度值像素的数量。

2　实 验

2.1　数据集及评估标准

ClearGrasp[17] 是一个包含虚拟合成和真实透明物体的数据集。有 9 类包含透明物体的合成图像，有 10 类包含真实世界中的透明物体图像，其中有 7 类存在透明物体类型的重叠。除了 RGB-D 图像外，数据集还提供了透明物体的表

面法线图、分割掩码图、遮挡边界图。本文使用其中 5 个重叠类的图像作为训练集，使用其中 5 个真实透明物体类的图像作为测试集。

TransCG[23] 是一个大规模的用于透明物体深度补全的真实物体数据集。数据集总共包含 51 类透明对象的 57715 张 RGB-D 图像，以及从现实世界设置下的 130 个场景的不同角度拍摄的许多不透明物体。并且数据集还提供了透明对象的 3D 网格模型。

对于本文深度补全任务，采用如文献 [16]、[17] 中的一些常用评估指标，包括 RMSE、REL、MAE 及 Threshold δ。各评估指标描述如下。

RMSE 为算法预测的深度图与基准深度图之间的根均方误差。公式如下

$$\sqrt{\frac{1}{|obj|}\sum_{p\in obj}\|G(p)-P(p)\|^2} \tag{6}$$

式中：G 表示深度图的基准值；P 表示经过算法补全后的深度图；p 表示物体的像素；obj 表示图中物体所在区域的全部像素。

REL 为相对误差。绝对误差与深度图的基准值相比所得。公式为

$$\frac{1}{|obj|}\sum_{p\in obj}\left|\frac{G(p)-P(p)}{G(p)}\right| \tag{7}$$

MAE 为平均绝对误差，公式为

$$\frac{1}{|obj|}\sum_{p\in obj}|G(p)-P(p)| \tag{8}$$

Threshold δ 为带有阈值的精度。δ_t 表示误差范围在 t 以内的像素百分比，公式为

$$\max\left(\frac{P(p)}{G(p)},\frac{G(p)}{P(p)}\right)<t \tag{9}$$

其中，根据先前的方法[17-18]，t 可被设为 1.05，1.10，1.25。

2.2　实验细节

本文设计的算法是基于 PyTorch[24] 实现的，并且是在两块 NVIDIA A30 卡上进行训练，使用一块 A30 卡进行测试的。在实验过程中并未使用任何预训练模型权重，本文采用 Adam 优化器[25] 并设置初始学习率为 0.001，网络训练了 40 个 epoch，并且分别在 5，15，25，35 个 epoch 时将学习率衰减为原来的 1/10，权重衰减系数设置为 0.000 1。在训练过程中每种数据集的 batch size 设置为 64。

2.3　消融分析

本文提出的方法有两个分支的输入，其中以掩码图为主导的分支，目的是让网络关注透明物体的特征信息，从而对透明物体有更好的深度补全效果。因此，为了检验以掩码图为主导的分支对于透明物体的深度补全效果，首先使用 ClearGrasp 数据集来验证加入该分支的有效性。

因为在以彩色图为主导的输入分支上加上了注意力模块，但是以掩码图为主导的输入分支并未加入注意力模块，所以为了保证对比实验不受注意力模块的影响，在对比实验时，将注意力模块去除，并且 SRB 模块堆叠数为 5。Image-Guided 表示网络中只有一个以彩色图引导的输入分支，Mask-Guided表示网络中只有一个以掩码图引导的输入分支，Joint-Guided 表示网络中包含双分支输入。

从表 1 可以看出，如果只使用 Mask-Guided 的输入分支，效果相较于只使用 Image-Guided 的输入分支差，而对于 Joint-Guided 时，深度补全的效果从得到了较大提升。所以即便是对于透明对象而言，对象本身也是具有一定的色彩和结构信息，如果只使用 Mask-Guided 的输入分支，网络对于透明物体的深度补全效果较差。但 Joint-Guided 的网络对于透明对象的深度补全效果比只有 Image-Guided 要好，因此，加入 Mask-Guided 的分支，可以有效地提高网络对透明对象的深度补全效果。

表 1　针对输入分支的评估参数对比表

Tab. 1　The comparison of metrics parameters for the input branch

算法	RMSE ↓	REL ↓	MAE ↓	δ ($t=1.05$) ↑	δ ($t=1.10$) ↑	δ ($t=1.25$) ↑
Image-Guided	0.049	0.065	0.043	52.21	68.51	92.87
Mask-Guided	0.078	0.097	0.056	44.76	60.20	89.63
Joint-Guided	0.044	0.060	0.038	57.22	76.45	95.67

对于注意力机制和谱残差块的分析，将使用 TransCG 数据集来进行实验对比，因为该数

据集中既包含透明对象也包含不透明对象，所以为了验证本文设计算法的通用性，采用该数据集来进行检验。其中对于谱残差块堆叠个数对网络性能的影响如表 2 所示，此时网络中并未加入注意力模块。当 SRB 堆叠得越深，网络的性能越来越好。但是当 N 大于 5 时，测试集的性能不再提升，反而有所下降。这应该是网络太复杂，过度拟合训练集所导致的。

表 2　SRB 堆叠数量变化的评估参数对比表
Tab. 2　The comparison of metrics parameters between different number of SRB

N	RMSE↓	REL↓	MAE↓	$\delta(t=1.05)$↑	$\delta(t=1.10)$↑	$\delta(t=1.25)$↑
3	0.063	0.095	0.058	49.21	67.12	93.11
4	0.050	0.073	0.042	56.35	75.31	95.27
5	**0.042**	**0.057**	**0.035**	**60.48**	**81.20**	**96.83**
6	0.041	0.057	0.035	60.71	81.32	96.80

对于注意力模块的分析，如表 3 所示，此时的网络除了注意力模块之外，编码部分采用双分支输入，而 SRB 的堆叠数量为 5，以下 Joint-Guided 简写为 JG。从表 3 中可以看出，将常规的 CBAM 的注意力机制设计方法迁移到本文设计的网络中时，相较于没有注意力模块的网络有一定的提升。而本文针对网络特性改进的注意力机制使用方法，具有更大的提升。

表 3　有无注意力机制的评估参数对比表
Tab. 3　The comparison of metrics parameters with or without attention mechanism

算法	RMSE↓	REL↓	MAE↓	δ (t=1.05)↑	δ (t=1.10)↑	δ (t=1.25)↑
JG	0.042	0.057	0.035	60.48	81.20	96.83
JG+Att	0.036	0.046	0.025	73.67	88.92	97.23
JG+Att-improved	0.032	0.045	0.024	74.35	89.71	97.90

2.4　对比先前的方法

本部分主要为了检验本文设计的算法在两个公共数据集上的效果，并与之前的方法进行对比。如表 4 所示，对于 ClearGrasp 数据集来说，本文的方法得益于谱残差块以及以掩码图为引导的输入分支，从而提升获取透明对象的结构信息的能力，并提高网络训练的稳定性，使得网络对透明对象的深度补全效果相较于之前的方法有所提升。

表 4　ClearGrasp 数据集的评估参数对比表
Tab. 4　The comparison of metrics parameters on ClearGrasp dataset

算法	RMSE↓	REL↓	MAE↓	δ (t=1.05)↑	δ (t=1.10)↑	δ (t=1.25)↑
JBF[26]	0.389	0.530	0.358	27.61	37.28	51.32
MRF[27]	0.347	0.497	0.311	38.35	52.63	65.19
AD[28]	0.315	0.489	0.297	41.26	61.29	71.24
DenseDepth[30]	0.270	0.428	0.259	18.67	34.34	58.29
DM[29]	0.049	0.075	0.038	59.67	75.85	95.96
DeepCompletion[11]	0.054	0.081	0.045	44.53	69.71	95.77
ClearGrasp[17]	0.038	0.048	0.027	72.94	87.88	97.17
Ours	**0.032**	**0.045**	**0.024**	**74.35**	**89.71**	**97.90**

本文单独对 TransCG 数据集做了测试，如表 5 所示，由于现有对透明对象的深度补全算法较少，但相较于其他方法，本文方法的性能评估参数大多优于其他方法。为了检验网络的泛化性，本文将设计的网络使用 ClearGrasp 数据集和 TransCG 数据集进行交叉训练和测试，比如使用前者训练，后者测试。结果如表 6 所示，可以看出本文设计的算法有良好的泛化性。

表 5　TransCG 数据集的评估参数对比表
Tab. 5　The comparison of metrics parameters on TransCG dataset

算法	RMSE↓	REL↓	MAE↓	δ (t=1.05)↑	δ (t=1.10)↑	δ (t=1.25)↑
ClearGrasp[17]	0.054	0.083	0.037	50.48	68.68	95.28
LIDF-Refine[18]	0.019	0.034	0.015	78.22	94.26	**99.80**
DFNet[23]	0.018	0.027	0.012	83.76	95.67	99.71
Ours	0.018	**0.025**	0.012	**85.69**	**96.49**	99.78

2.5　自采深度图补全

本部分主要是对使用 RealSense D435i 深度相机采集的稀疏深度图进行补全后的结果进行分析，如图 5 所示。彩色图中下方的装置是机械夹爪，数据是由安装于机械臂末端的深度相机进行

表 6　跨域数据集的评估参数对比表

Tab. 6　The comparison of metrics parameters on cross-domain datasets

训练/测试	算法	RMSE↓	REL↓	MAE↓	$\delta(t=1.05)$↑	$\delta(t=1.10)$↑	$\delta(t=1.25)$↑
ClearGrasp/ TransCG	ClearGrasp[17]	0.061	0.108	0.049	33.59	54.73	92.48
	LIDF-Refine[18]	0.146	0.262	0.115	13.70	26.39	57.95
	DFNet[23]	0.048	0.088	0.039	38.65	62.42	95.28
	Ours	**0.036**	**0.070**	**0.034**	**45.47**	**75.21**	**96.02**
TransCG/ ClearGrasp	ClearGrasp[17]	0.085	0.095	0.052	47.26	70.76	92.54
	LIDF-Refine[18]	0.152	0.225	0.139	9.86	20.63	46.02
	DFNet[23]	0.041	0.054	0.031	62.74	83.31	97.33
	Ours	**0.032**	**0.047**	**0.029**	**69.23**	**88.14**	**97.80**

采集的。进行深度补全的训练模型并未用图中的数据进行 fine tuning，图 5 中第二行深度信息是直接使用上述相机进行拍摄获取的稀疏深度图，图 5 最后一行是经过深度补全后的密集深度图，由于机械夹爪装置的存在，使得获取的稀疏深度图下方的深度信息严重缺失，这种情况其实是因为 RealSense D435i 相机的深度信息获取的有效距离是大于一定阈值的，所以距离相机过近时，深度信息缺失严重。图中的物体由于物体表面反光或者物体边缘隆起部分非常纤细，导致深度信息的缺失或深度信息的错误。经过算法模型的深度补全后，可以得到如图 5(a)、(b)、(d)、(e)列不错的补全效果，但是对于这些图补全后的结果，补全图的底部可能是由于深度缺失，导致训练的模型并未“见”过此深度信息，所以对这一部分的补全效果差，甚至还有一些误补全的部分，其次由于模型也并未“见”过这些对象以及背景，所以对深度图进行错误的补全操作，如(c)列对老虎钳的深度补全效果不如直接使用深度相机获取的深度图。因此，可以看出算法模型的泛化性还有待提高。

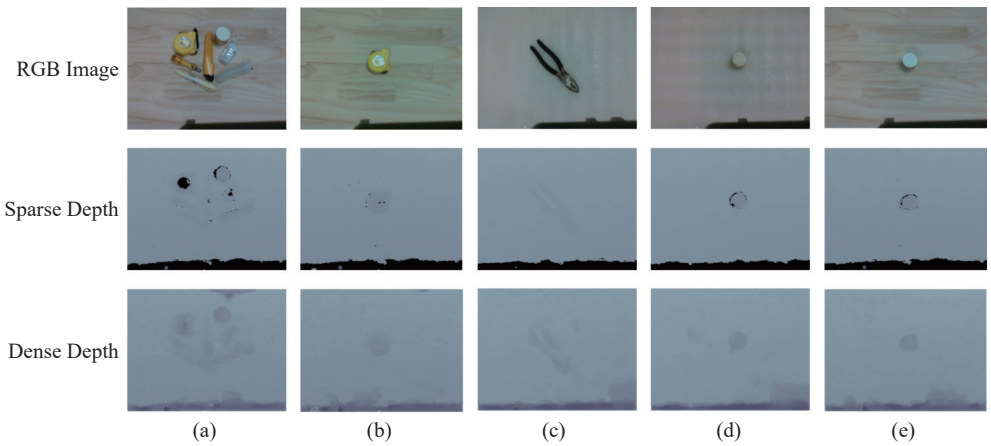

图 5　深度图补全结果

Fig. 5　The results of depth completion.

3　结　论

深度信息在机器人抓取，三维环境重建，自动驾驶等领域有着越来越重要的作用。而由于深度传感器的缺陷，获取到的原始深度图往往是比较稀疏的，导致使用原始深度图无法满足现实任务的需求。许多研究人员基于卷积神经网络的强大表征能力，提出了多种深度补全算法，然而先前大部分方法针对的是不透明对象。本文设计了

一种双分支输入的 Encoder-Decoder 结构的网络，通过以掩码图为引导的输入分支，使得网络能够更加理解在场景中的透明对象的几何特征，使用 SRB 堆叠形成的模块连接编码与解码部分，使网络有效的捕捉物体信息及区分几何形状，并提高网络训练的稳定性。并且将改进后的注意力机制使用方法添加到网络中，进一步提升网络对图像中物体细节信息和特征中语义信息的建模能力。通过实验验证了算法的有效性。

深度补全网络的作用是构建一种原始稀疏深度图与高质量稠密深度图之间的映射关系，然而要将深度补全算法应用到实际落地项目中，对算法的实时性能和精度的要求较高，所以如何设计一个高效的深度补全网络仍是一项挑战。由于机器人抓取这类实时性要求比较高的任务，需要追求算法的实时性能，未来需要在对精度影响不大的情况下，尽量将深度补全网络的推理时间提升，从而更加适用于抓取类任务。

参考文献：

[1] JARITZ M, DE CHARETTE R, WIRBEL E, et al. Sparse and dense data with CNNs: depth completion and semantic segmentation[C]//International Conference on 3D Vision (3DV). Verona: IEEE, 2018: 52 − 60.

[2] SONG Z B, LU J F, YAO Y Z, et al. Self-supervised depth completion from direct visual-LiDAR odometry in autonomous driving[J]. IEEE Transactions on Intelligent Transportation Systems, 2022, 23(8): 11654 – 11665.

[3] DU R F, TURNER E, DZITSIUK M, et al. DepthLab: real-time 3D interaction with depth maps for mobile augmented reality[C]//Proceedings of the 33rd Annual ACM Symposium on User Interface Software and Technology. ACM, 2020: 829 − 843.

[4] MA F C, CARLONE L, AYAZ U, et al. Sparse depth sensing for resource-constrained robots[J]. The International Journal of Robotics Research, 2019, 38(8): 935 – 980.

[5] TEIXEIRA L, OSWALD M R, POLLEFEYS M, et al. Aerial single-view depth completion with image-guided uncertainty estimation[J]. IEEE Robotics and Automation Letters, 2020, 5(2): 1055 – 1062.

[6] CHEN L, LIN H, LI S T. Depth image enhancement for Kinect using region growing and bilateral filter[C]//Proceedings of the 21st International Conference on Pattern Recognition. Tsukuba: IEEE, 2012: 3070 − 3073.

[7] LIU S J, LAI P L, TIAN D, et al. Joint trilateral filtering for depth map compression[C]//Proceedings of SPIE 7744, Visual Communications and Image Processing 2010. Huangshan: SPIE, 2010: 77440F.

[8] ALHWARIN F, FERREIN A, SCHOLL I. IR stereo Kinect: improving depth images by combining structured light with IR stereo[C]//13th Pacific Rim International Conference on Artificial Intelligence. Gold Coast: Springer, 2014: 409 − 421.

[9] CHIU W W C, BLANKE U, FRITZ M. Improving the Kinect by cross-modal stereo[C]//British Machine Vision Conference. Dundee: BMVC, 2011: 1 − 10.

[10] CHEN K, LAI Y K, WU Y X, et al. Automatic semantic modeling of indoor scenes from low-quality RGB-D data using contextual information[J]. ACM Transactions on Graphics, 2014, 33(6): 208.

[11] ZHANG Y D, FUNKHOUSER T. Deep depth completion of a single RGB-D image[C]//Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018: 175 − 185.

[12] QIU J X, CUI Z P, ZHANG Y D, et al. DeepLiDAR: Deep surface normal guided depth prediction for outdoor scene from sparse LiDAR data and single color image[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2019: 3308 − 3317.

[13] MA F C, CAVALHEIRO G V, KARAMAN S. Self-supervised sparse-to-dense: self-supervised depth completion from LiDAR and monocular camera[C]// 2019 International Conference on Robotics and Automation (ICRA). Montreal: IEEE, 2019: 3288 − 3295.

[14] ELDESOKEY A, FELSBERG M, KHAN F S. Confidence propagation through CNNs for guided sparse depth regression[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2020, 42(10): 2423 – 2436.

[15] CHENG X J, WANG P, YANG R G. Depth estimation via affinity learned with convolutional spatial propagation network[C]//Proceedings of the 15th European Conference on Computer Vision (ECCV).

Munich: Springer, 2018: 108 − 125.

[16] HUANG Y K, WU T H, LIU Y C, et al. Indoor depth completion with boundary consistency and self-attention[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision Workshop. Seoul: IEEE, 2019: 1070 − 1078.

[17] SAJJAN S, MOORE M, PAN M, et al. Clear grasp: 3D shape estimation of transparent objects for manipulation[C]//2020 IEEE International Conference on Robotics and Automation (ICRA). Paris: IEEE, 2020: 3634 − 3642.

[18] ZHU L Y, MOUSAVIAN A, XIANG Y, et al. RGB-D local implicit function for depth completion of transparent objects[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville: IEEE, 2021: 4647 − 4656.

[19] HU M, WANG S L, LI B, et al. PENet: towards precise and efficient image guided depth completion[C]//2021 IEEE International Conference on Robotics and Automation (ICRA). Xi'an: IEEE, 2021: 13656 − 13662.

[20] TANG Y J, CHEN J H, YANG Z G, et al. DepthGrasp: depth completion of transparent objects using self-attentive adversarial network with spectral residual for grasping[C]//2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). Prague: IEEE, 2021: 5710 − 5716.

[21] MIYATO T, KATAOKA T, KOYAMA M, et al. Spectral normalization for generative adversarial networks[C]//6th International Conference on Learning Representations. Vancouver: ICLR, 2018.

[22] WOO S, PARK J, LEE J Y, et al. CBAM: convolutional block attention module[C]//Proceedings of the 15th European Conference on Computer Vision (ECCV). Munich: Springer, 2018: 3 − 19.

[23] FANG H J, FANG H S, XU S, et al. TransCG: a large-scale real-world dataset for transparent object depth completion and a grasping baseline[J]. IEEE Robotics and Automation Letters, 2022, 7(3): 7383 – 7390.

[24] PASZKE A, GROSS S, MASSA F, et al. PyTorch: an imperative style, high-performance deep learning library[C]//Proceedings of the 33rd International Conference on Neural Information Processing Systems. Vancouver: NeurIPS, 2019: 721.

[25] KINGMA D P, BA J. Adam: a method for stochastic optimization[C]//3rd International Conference on Learning Representations. San Diego: ICLR, 2014.

[26] SILBERMAN N, HOIEM D, KOHLI P, et al. Indoor segmentation and support inference from RGBD images[C]//12th European Conference on Computer Vision. Florence: Springer, 2012: 746 − 760.

[27] HARRISON A, NEWMAN P. Image and sparse laser fusion for dense scene reconstruction[C]//7th International Conference on Field and Service Robotics. Cambridge: Springer, 2010: 219 − 228.

[28] LIU J Y, GONG X J. Guided depth enhancement via anisotropic diffusion[C]//14th Pacific-Rim Conference on Advances in Multimedia Information Processing. Nanjing: Springer, 2013: 408 − 417.

[29] SENUSHKIN D, ROMANOV M, BELIKOV I, et al. Decoder modulation for indoor depth completion[C]//2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). Prague: IEEE, 2021: 2181 − 2188.

[30] ALHASHIM I, WONKA P. High quality monocular depth estimation via transfer learning[J]. arXiv:, 1812, 11941: 2018.

（编辑：张　磊）