浙江理工大学学报,第 49 卷,第 3 期,2023 年 5 月
Journal of Zhejiang Sci-Tech University
DOI:10.3969/j.issn.1673-3851(n).2023.03.007

基于通道注意力机制的室内场景深度图补全

任瀚实[a],周志宇[a],孙树森[b]
(浙江理工大学,a.信息科学与工程学院;b.计算机科学与技术学院,杭州 310018)

摘　要:在深度相机获取的室内场景深度图中,部分像素点缺失深度信息;为补全深度信息,设计了端对端的场景深度图补全网络,在此基础上提出了一种基于通道注意力机制的室内场景深度图补全方法。该方法将场景彩色图与缺失部分深度信息的场景深度图作为场景深度图补全网络的输入,首先提取场景彩色图和深度图的联合特征,并根据通道注意力机制将提取到的联合特征进行解码,得到初始预测深度图;然后借助非局部区域上的传播算法逐步优化场景深度的预测信息,得到完整的场景深度图;最后在 Matterport3D 等数据集上进行实验,并将该方法与典型方法进行比较分析。实验结果表明,该方法融合了场景彩色图和深度图特征信息,通过注意力机制提高了深度图补全网络的性能,有效补全了深度相机拍摄室内场景时缺失的深度信息。

关键词:深度图;深度补全;深度学习;注意力机制;室内场景
中图分类号:TP391.41;TP183　　**文献标志码**:A　　**文章编号**:1673-3851 (2023) 05-0344-09
引文格式:任瀚实,周志宇,孙树森. 基于通道注意力机制的室内场景深度图补全[J]. 浙江理工大学学报(自然科学),2023,49(3):344-352.
Reference Format: REN Hanshi, ZHOU Zhiyu, SUN Shusen. Depth map completion for indoor scenes based on the channel attention mechanism[J]. Journal of Zhejiang Sci-Tech University,2023,49(3):344-352.

Depth map completion for indoor scenes based on the channel attention mechanism

REN Hanshi[a], *ZHOU Zhiyu*[a], *SUN Shusen*[b]
(a. School of Information Science and Technology; b. School of Computer Science and Technology, Zhejiang Sci-Tech University, Hangzhou 310018, China)

Abstract: When the depth camera scans the indoor scene, the depth information of some pixels is missing. To solve this problem, we proposed a depth map completion method for indoor scenes based on a channel attention mechanism and an end-to-end scene depth map completion network. With the scene color map and incomplete scene depth map as input, the network first extracted the scene color map features and depth map features, and decoded the extracted combined features based on the channel attention mechanism to obtain the initially predicted depth map. Then the predicted scene depth information was gradually optimized with the help of non-local spatial propagation algorithm to finally obtain the complete scene depth map. Finally, the proposed method was compared with typical methods using datasets such as Matterport3D. Experimental results show that this method, integrating the feature information of the scene color map and the depth map, improves the performance of depth map completion network through the attention mechanism, and effectively complements the lack of depth information when shooting indoor scenes by depth camera.

收稿日期:2022-12-04　　网络出版日期:2023-04-10
作者简介:任瀚实(1997-),女,吉林吉林人,硕士研究生,主要从事计算机图像和机器学习方面的研究。
通信作者:孙树森,E-mail:shusensun@zstu.edu.cn

Key words: depth image; depth completion; deep learning; attention mechanism; indoor scene

0 引　言

目前,场景深度图在自动驾驶[1]、场景重建[2]以及增强现实[3]等领域有着广泛应用。虽然深度传感器技术已经有了很大发展,但 Microsoft Kinect、Intel Real Sense、Google Tango 等商品级 RGB-D 相机,在拍摄场景过亮、物体表面过于光滑、相机与物体之间距离过远或过近等情况下采集到的深度图像会缺失部分像素点的深度数据。深度数据缺失会对自动驾驶、三维重建、目标检测等计算机视觉任务造成不利影响,因此深度图补全研究非常必要。

本文研究商品级相机拍摄的室内场景 RGB-D 图像深度数据补全的问题。深度图补全任务可以归结为一个密集标记问题,因此针对密集标记问题的方法也可以用于深度图补全任务[4]。带有跳跃连接的编码器-解码器结构[4]已广泛用于解决语义分割等密集标记问题。例如,Senushkin 等[4]提出了一种网络,该网络由编码器-解码器和一个轻量级细化网络组成,在 Matterport3D 数据集上取得了最佳效果。此外,解决深度图补全任务的另一种思路是利用传播算法,通过观测到的深度数据补全深度图。例如,Park 等[5]提出了一种网络,采用传播算法,利用非局部区域内相关性较强的像素的深度数据来补全深度图,有效避免了局部区域中无关像素的影响。

单目图像的深度估计是计算机视觉领域中一个重要问题。早期的深度估计方法主要通过人工调整模型、表面和法线[6-7]、单目纹理图或失焦特性[8]等实现;最近的研究主要结合多尺度信息来预测像素级深度[9-10],这可以通过多种方式实现,如融合网络架构中不同层对应的特征映射再使用降维回归。上述方法只适用于估计商品 RGB-D 相机所拍摄的深度图,但不适用于商品级相机拍摄的室内场景 RGB-D 图像深度图补全的问题。由于这些 RGB-D 图像通常缺失部分深度信息,仅在原始深度上训练不能很好地补全缺失的深度图,所以本文采用融合彩色图和深度图的方法来提升深度图补全任务效果。

早期的深度图补全任务一般通过基于压缩感知或滤波器的方法实现。Uhrig 等[11]提出了一种基于卷积神经网络的深度图补全方法,通过彩色图像引导对基于卷积神经网络的深度图补全方法进行了改进。Chen 等[12]提出了一种基于连续卷积的 2D-3D 融合网络结构。Li 等[13]采用多尺度级联沙漏结构进行深度图补全。郑柏伦等[14]使用三分支的多尺度特征融合网络融合彩色图像特征和深度图像特征,进行深度图补全。除了利用彩色图像引导,最近提出的一些方法还使用表面法线和物体边界信息来辅助解决深度图补全任务[15-16]。以上相关算法都属于监督学习,此外处理深度图补全任务也可以使用自监督算法或无监督算法[17-18]。上述大部分工作都是基于室外场景的稀疏深度的 Lidar 数据,并在 KITTI 数据集上验证;另一部分工作针对的是 Kinect 传感器获取的半密集深度数据。最近,Zhang 等[19]引入了用于室内深度图补全任务的大型 RGB-D 数据集 Matterport3D,使用预先训练好的神经网络进行法线估计以及边界检测,再对得到的法向量和边界进行全局优化。虽然该方法在 Matterport3D 数据集上取得良好效果,但其网络的复杂性限制了该方法的实际应用。Huang 等[20]提出的方法在 Matterport3D 数据集上首次超过原始结果。与 Zhang 等[19]的研究相似,他们提出的方法也涉及复杂的图像预处理并通过多阶段处理实现;虽然它不依赖于预先训练的骨干网络,但使用了在外部数据上训练过的深度神经网络。Senushki 等[4]提出的编解码网络训练的模型在 Matterport3D 数据集上表现最佳。该模型仅需要输入室内场景彩色图和不完整的深度图,而不需要法线信息和边界等附加信息,通过编码器进行特征提取后即可利用轻量级的细化网络完成深度图补全任务。

直接进行深度图补全的算法已经具有良好表现;使用彩色图引导稀疏深度图进行传播,并通过稀疏的深度图获得深度信息预测,是另一种完成深度图补全任务的有效途径。空间传播算法从大规模数据中学习特定的亲和力,可应用于包括深度图补全和语义分割在内的计算机视觉任务。最初的空间传播算法通过线性相关的传播模型设计,只针对某像素与其在某个方向上(上、下、左、右)邻近的 3 个像素,没有考虑与该像素相关的所有像素。Cheng 等[21]提出的卷积空间传播网络(Convolutional spatial propagation network,CSPN)克服了这一局限。CSPN 可以预测局部区域内相邻点的相关性,同时更新所有像素的局部上下文信息以提高效率。然而 CSPN 依赖于固定区域内的相邻像素之间的关系,这些固定区域内的相邻像素本身可能来自不同的物体。因此这样的传播方式会因不相关像素导

致深度值混合,也限制了图像中上下文相关信息(即非局部上下文信息)的应用。最近,已有一些研究对非局部上下文信息的应用进行了探索[22-24]。Wang等[25]提出了一种带有非局部块的深度神经网络;非局部块由两两像素间的亲和力计算模块和特征处理模块组成,通过将非局部块嵌入已有的深度网络进行视频分类和图像识别任务。Park等[5]在此基础上提出了一种非局部区域空间传播网络(Non-local spatial propagation network,NLSPN),进一步优化了深度图补全网络。NLSPN计算每个像素的非局部区域内像素间的亲和力,通过深度神经网络得到初始深度图以及置信度图;结合像素间的亲和力关系和置信度,在预测的非局部区域中对初始预测深度图进行迭代传播精化。该网络有效避免了固定区域内不相关像素在传播过程中可能造成的混合深度问题,但在深度边界处的混合深度问题仍未能完全解决。

针对深度相机获取的室内场景深度图,为补全部分像素点缺失的深度信息,本文提出了一种深度图补全方法,设计了一种基于通道注意力(Efficient channel attention,ECA)机制的融合彩色图与深度图特征的深度图补全网络。该补全网络采用编解码器结构,编码器由多个卷积层和通道注意力机制模块组成,对深度图及其对应的彩色图进行多尺度特征提取,在此基础上通过传播算法得到深度补全图像。本文提出的深度图补全网络融合了彩色图与深度图的特征,并且不需要预先提取法线、边界检测等额外信息,降低了网络的结构复杂性,在Matterport3D和NYUv2数据集上取得了较好效果。

1 方法设计

本文提出的深度图补全方法可分为两步:第一步,使用编解码结构获得初始深度预测图,通过编码器学习室内场景彩色图中大量的特征信息,并与室内场景深度图的特征进行融合,以获取丰富图像特征;通过解码器使用多个反卷积操作实现初步补全。第二步,在非局部区域上使用传播算法,细化上一步得到的深度预测图。

1.1 深度图补全网络

本文提出的深度图补全网络采用端对端的卷积神经网络,结合了通道注意力机制和空间传播算法。采用通道注意力机制可以使深度神经网络更加关注关键信息。在提取特征图时融合了彩色图以及深度图信息,能有效地获得信息丰富的特征图,加强彩色图在补全分支上的引导作用。特征图提取采用基于ResNet的深度卷积神经网络,并加入了通道注意力机制模块,以获得不同大小的特征图;在特征图逐步变小的同时增加了特征图的数量,保证了网络结构的有效性。

本文设计的网络结构如图1所示。该络采用编码器—解码器结构,包含两部分:彩色图—深度图特征提取模块和深度图补全模块。彩色图—深度图特征提取模块通过多个卷积层和通道注意力机制模块实现对彩色图和深度图融合特征提取,以获取丰富的特征图关键信息,从而为深度图补全模块作准备。在深度图和彩色图特征提取阶段,先对彩色图和深度图分别通过Conv1进行一次卷积,将二者结果融合,再使用两个通道注意力模块。深度图补全模块采用4个反卷积层,然后分成3个分支,得到初始深度预测图 $\boldsymbol{D}_{\text{init}}$、亲和矩阵 $\boldsymbol{W}$ 以及非局部区域 $N_{m,n}^{\text{NL}}$。通过非局部区域传播算法,最终得到深度预测图 $\boldsymbol{D}_{\text{final\_pred}}$,即:

$$\boldsymbol{D}_{\text{final\_pred}}=f_{\text{NLSPN}}(\boldsymbol{D}_{\text{init}},\boldsymbol{W},N_{m,n}^{\text{NL}}) \tag{1}$$

其中:$f_{\text{NLSPN}}()$表示初始深度预测图 $\boldsymbol{D}_{\text{init}}$ 和最终补全的深度图像 $\boldsymbol{D}_{\text{final\_pred}}$ 之间根据非局部区域 $N_{m,n}^{\text{NL}}$ 和亲和矩阵 $\boldsymbol{W}$ 传播的映射函数。

通过彩色图—深度图特征提取模块对深度图和其对应的彩色图进行特征提取。本文直接将未经处理的深度图和彩色图作为网络的输入。先分别使用一个3×3的卷积层得到一个32通道的彩色图特征图 $\boldsymbol{F}_{\text{c}}$ 和一个16通道的深度图特征图 $\boldsymbol{F}_{\text{d}}$,将二者结合为一个48通道的特征图 $\boldsymbol{F}_2$,再经过多个3×3的卷积层进行多尺度提取,对其中第一步和最后一步得到的特征图 $\boldsymbol{F}_i$,即 $\boldsymbol{F}_2$、$\boldsymbol{F}_4$ 使用通道注意力模块[26],即:

$$\boldsymbol{F}_{\text{c}}=\sigma(\boldsymbol{W}_1*\boldsymbol{I}_{\text{color}}+\boldsymbol{b}_1),$$

$$\boldsymbol{F}_{\text{d}}=\sigma(\boldsymbol{W}_2*\boldsymbol{D}_{\text{depth}}+\boldsymbol{b}_2),$$

$$\boldsymbol{F}_{i,\text{ECA}}=f_{\text{ECA}}(\boldsymbol{F}_i),$$

其中:$\boldsymbol{I}_{\text{color}}$ 表示场景彩色图;$\boldsymbol{D}_{\text{depth}}$ 为场景深度图;σ 表示激活函数Relu;$*$ 表示卷积操作,$\boldsymbol{W}$ 表示卷积的权重;$\boldsymbol{b}_1$、$\boldsymbol{b}_2$ 均为偏置向量;$\boldsymbol{F}_{\text{c}}$ 和 $\boldsymbol{F}_{\text{d}}$ 分别为经过1层卷积后得到的彩色特征图和深度特征图,$f_{\text{ECA}}()$表示通道注意力机制。其中,通道注意力机制模块没有过分增加模型的复杂度,但可以显著提升补全效果。通道注意力机制的作用原理见1.2节。

在进行深度图补全时,使用4个卷积核为3×3,步长为2的反卷积层进行解码。分3个分支得到初始深度预测图 $\boldsymbol{D}_{\text{init}}$、亲和矩阵 $\boldsymbol{W}$ 以及非局部区域 $N_{m,n}^{\text{NL}}$,使用初始深度预测图 $\boldsymbol{D}_{\text{init}}$ 利用非局部区域传

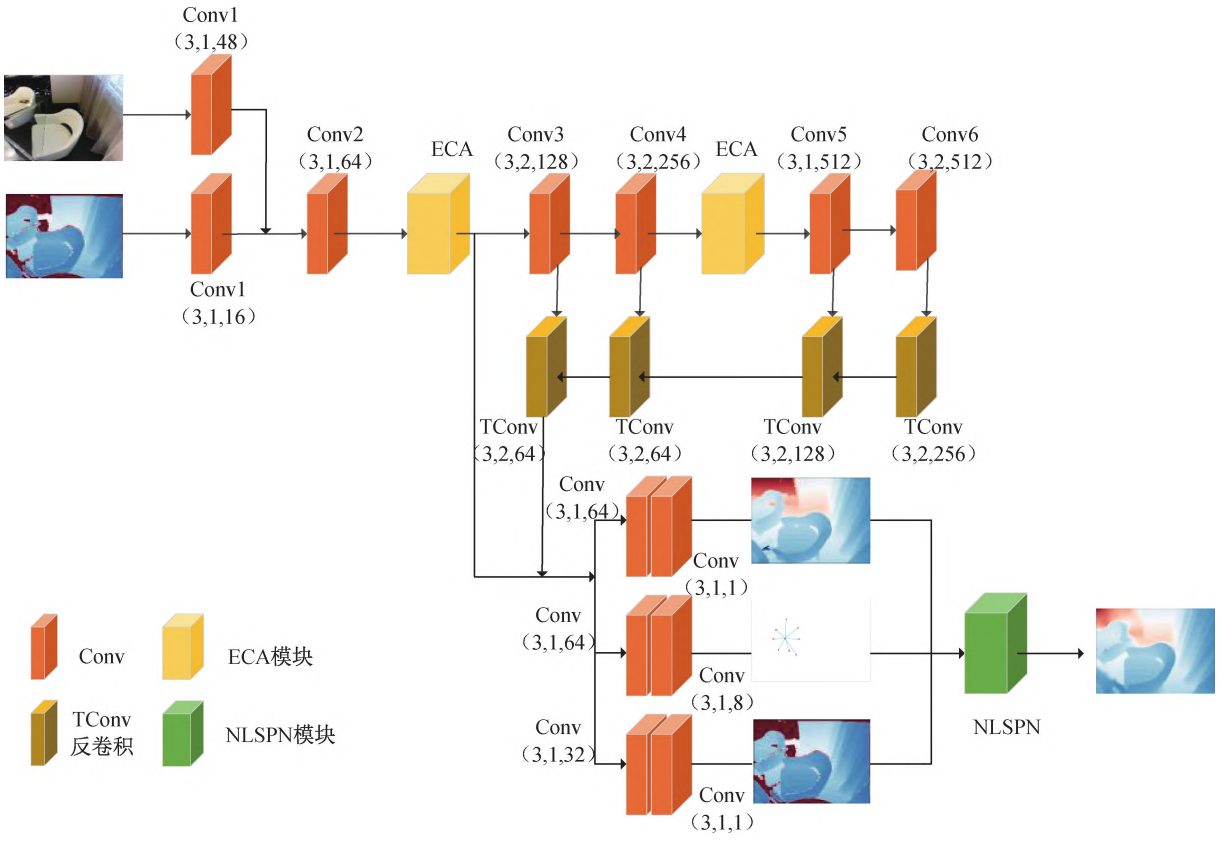


图 1　基于通道注意力机制的深度图补全网络框架

播算法对相关区域 $N_{m,n}^{\mathrm{NL}}$ 中的像素点进行传播赋值,其中相关区域 $N_{m,n}^{\mathrm{NL}}$ 表示为:

$$N_{m,n}^{\mathrm{NL}}=\left\{x_{m+p,n+q}\mid(p,q)\in f_{\varphi}(\boldsymbol{I},\boldsymbol{D},m,n),p,q\in\mathbf{R}\right\}\quad(2)$$

传播算法能使场景内物体前景与背景有效分开,有利于深度图像信息的补全。非局部区域传播算法的实现过程见 1.3 节。

1.2　通道注意力机制模块

为了获得具有更多有效信息的深度特征图,将彩色图—深度图特征提取模块得到的多个特征图输入通道注意力机制模块,对第一步提取特征得到的特征图和最后一个卷积提取特征得到的特征图输入通道注意力机制模块,深度图像补全的效果可以得到显著提升。通道注意力机制模块示意图如图 2 所示。本文在编码器融合彩色图和深度图之后,对其进行多次卷积操作,获得特征图 $\boldsymbol{F}_i$,并对得到的特征图 $\boldsymbol{F}_i$ 先使用全局平均池化层,再利用自适应的卷积核为 k 的 1 维卷积,以捕捉各个通道之间的交互关系,最后得到 $\boldsymbol{F}_{i,\mathrm{ECA}}$。通道注意力机制模块可以在不做降维操作的情况下,以较小的参数量来提升卷积神经网络的性能。

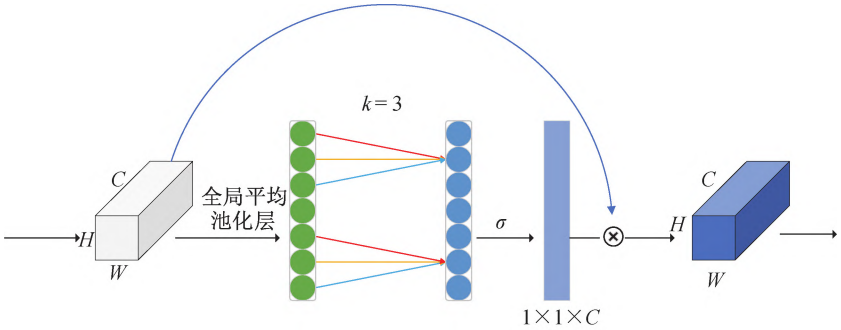


图 2　通道注意力机制模块示意图

通道注意力机制模块的输入为卷积模块提取出的特征图 $\boldsymbol{F}_i$,经过操作得到包含更多有效信息的特征图 $\boldsymbol{F}_{i,\mathrm{ECA}}$。

本文使用矩阵 $\boldsymbol{W}_k$ 来学习通道注意力,

$$\boldsymbol{W}_k=\begin{bmatrix} w^{1,1} & \cdots & w^{1,k} & 0 & 0 & \cdots & \cdots & 0 \\ 0 & w^{2,2} & \cdots & w^{2,k+1} & 0 & \cdots & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & \cdots & 0 & 0 & \cdots & w^{C,C-k+1} & \cdots & w^{C,C} \end{bmatrix},$$

其中:w 是可学习的参数。对于每个通道 y_i 本文只计算它与其相邻的 k 个通道之间的关系,y_i 的权重 ω_i 表示为:

$$\omega_i=\sigma\left(\sum_{j=1}^{k} w^j y_i^j\right),\ y_i^j\in\Omega_i^k,$$

其中:Ω_i^k 表示与 y_i 相邻的一组通道,y_i^j 表示第 j 个与 y_i 相邻的通道。

以上权重计算方式可以通过1维卷积来实现,即:

$$\omega_i=\sigma(\mathrm{Conv1D}_k(y_i)),$$

其中:$\mathrm{Conv1D}_k(\ )$表示卷积核为 k 的1维卷积。使用1维卷积时,通道维度 C 与卷积核 k 之间存在一个线性映射关系,$C=r\times k-b$;为了减小线性函数作为特征关系式所带来的局限性,而且一般通道维度 C 通常为2的幂次方,因此本文将线性函数 $\varphi(k)=r\times k-b$ 扩展为非线性函数,即:

$$C=\varphi(k)=2^{(r\times k-b)}。$$

所以,根据给定的特征图通道数 C,k 可以表示为:

$$k=\psi(C)=\left|\frac{\log_2^C}{r}+\frac{b}{r}\right|_{\mathrm{odd}}。$$

其中:$|x|_{\mathrm{odd}}$ 表示最接近 x 的奇数。本文使用 $r=2$、$b=1$ 时的通道注意力机制模块;当输入的特征为 $\boldsymbol{F}_2$ 时,此时自适应的一维卷积核大小为3,如图2所示。

1.3 非局部区域传播算法

为了进一步细化初步得到的补全深度图,本文对初步获得的深度图使用非局部区域传播算法进行细化处理。传播算法的原理是用具有相似性的相邻观测像素来估计其他深度值缺失的像素,并细化可信度较低区域的深度值。空间传播算法已经被用于各种计算机视觉应用中的关键模块,在深度图补全任务中优势更为明显,比直接回归算法更优越[8,25]。空间传播算法和卷积空间传播算法可以把可信度高的区域信息通过数据间的亲和度依赖有效传播到可信度较低的区域。但是这两种算法会受到局部区域的限制,局部区域像素的深度分布不均,可能导致前景和背景的深度值混合。所以,本文采用非局部区域传播算法[6]改善前背景处深度值混合的问题。

本文定义非局部区域如式(2)所示,其中:$\boldsymbol{I}$ 和 $\boldsymbol{D}$ 分别为彩色图像和深度图像,$f_\varphi()$表示非局部区域预测网络,p、q 为实数(可以为分数)。在此区域中,使用亲和力和置信度的联合学习,增强相关像素的影响,同时抑制不相关像素的影响。$f_\varphi()$网络采用可形变卷积[27]实现。

NLSPN模块的输入为初始预测图 $\boldsymbol{D}_{\mathrm{init}}$、像素间的亲和矩阵和原始深度图的置信度图。置信度图含有原始深度图中每个像素的置信度,每个像素的置信度取值范围为[0,1]。本文将置信度和亲和矩阵相结合,进行归一化得到亲和矩阵 W,即:

$$w_{m,n}^{i,j}=c^{i,j}\cdot\tanh(\hat{w}_{m,n}^{i,j})/\gamma,\ \gamma_{\min}\leqslant\gamma\leqslant\gamma_{\max},$$

其中:$c^{i,j}$ 表示在(i,j)处像素的置信度,取值范围为[0,1];$\hat{w}_{m,n}^{i,j}$ 表示位于(i,j)处像素未经归一化的初始亲和度;γ 表示可学习的归一化参数,$\gamma_{\min}$ 和 $\gamma_{\max}$ 是根据经验设置的最小值和最大值;tanh函数的使用能够减少归一化过程中的偏差。

因此,最后输出结果为:

$$\boldsymbol{D}_{\mathrm{final\_pred}}=f_{\mathrm{NLSPN}}(\boldsymbol{D}_{\mathrm{init}},\boldsymbol{W},N_{m,n}^{\mathrm{NL}})。$$

1.4 损失函数

总损失可以写为:

$$L=\lambda_1 L_1+\lambda_2 L_2+\lambda_{\mathrm{i}} L_{\mathrm{i}}+\lambda_{\mathrm{ssim}} L_{\mathrm{ssim}},$$

$$L_1=\frac{1}{|V|}\sum_{v\in V}\|D(v)-D_{\mathrm{gt}}(v)\|,$$

$$L_2=\frac{1}{|V|}\sum_{v\in V}\|D(v)-D_{\mathrm{gt}}(v)\|^2,$$

$$Li=\frac{1}{|V|}\sum_{v\in V}\|D_{\mathrm{init}}(v)-D_{\mathrm{gt}}(v)\|,$$

其中:$D(v)$表示最终补全得到的深度图中像素 v 处的深度值;$D_{\mathrm{gt}}(v)$表示场景真实深度图中像素 v 处深度值;$D_{\mathrm{init}}(v)$表示初始深度预测图中像素 v 处深度值;V 表示场景真实深度图中有效像素;$|V|$表示有效像素的数量。

本文在总损失中加入了 L_{ssim} 损失,以获得更高质量、结构更完整深度图补全图。L_{ssim} 为结构相似性(Structural similarity,SSIM)损失[21],根据结构信息的退化程度来衡量补全质量。L_{ssim} 越小,补全后深度图的结构与场景真实深度图结构越相似。

2 结果与讨论

本文提出的深度图补全网络使用Pytorch框架搭建训练模型,GPU使用NVIDIA GeForce RTX

3060,内存为 12 GiB。训练数据集使用 Matterport3D 数据集[19] 和 NYUv2 数据集。在 Matterport3D 数据集上只有极少方法能取得较优的深度图补全效果,NYUv2 数据集是深度图像补全任务的常用数据集。

Matterport3D 数据集[19] 由真实的传感器数据以及从官方重构网格获取到的真实深度数据组成。本文的训练集使用其中一个子集,约 2358 张图像,对图像做预处理,将其缩小至 320×240。本文训练模型时使用 50 个 Epoch,使用另一子集约 864 张图像作为测试集,并与典型算法进行比较分析。NYUv2 数据集由 Kinect 传感器捕获的 464 个室内场景的彩色图和深度图像组成。对于训练数据,本文使用了官方训练分割的一个约 50 kib 的图像子集,每张图像缩小到 320×240,然后使用 304×228 中心裁剪。与文献[5, 20]类似,本文从密集深度图像中随机采样 500 个深度像素,并将其与对应的彩色图像联合作为网络输入。训练模型时同样使用 50 个 Epoch,使用官方分割出的 654 张图像测试集进行评估和可视化比较。

2.1　评估指标

本文的室内场景补全任务的验证方式与文献[4-5, 19-20]类似。本文采用均方根误差、平均绝对误差以及精确度对本文的网络进行评估,这些指标在室内场景深度图补全任务的评估中被广泛认可。均方根误差和平均绝对误差直接测量绝对深度精度。均方根误差对异常深度值更为敏感,所以均方根误差通常被视作模型评估的主要指标。另外,精确度 δ_t 表示相对误差小于阈值 t 的预测像素所占的百分比。δ_t 也可以被理解为深度图补全任务的精度,精度越高说明最终预测的深度图与场景真实深度值越接近,补全效果就越好。本文计算了 t 取 1.25、1.25^2、1.25^3 时的 δ_t 值。

均方根误差 E_{RMS} 的计算公式为:

$$E_{\mathrm{RMS}}=\sqrt{\frac{1}{|v|}\sum_{v\in V}\|D(v)-D_{\mathrm{gt}}(v)\|^2};$$

平均绝对误差 E_{MA} 的计算公式为:

$$E_{\mathrm{MA}}=\frac{1}{|v|}\sum_{v\in V}\|D(v)-D_{\mathrm{gt}}(v)\|;$$

δ_t 的计算公式为:

$$\max\left(\frac{D(v)}{D_{\mathrm{gt}}(v)},\frac{D_{\mathrm{gt}}(v)}{D(v)}\right)<t,$$

其中:$t\in\{1.25,1.25^2,1.25^3\}$,$v\in V$。

2.2　实验结果对比分析

为了验证本文提出的深度图像补全网络的性能,选取 Zhang 等[19]、Huang 等[20]、Senushkin 等[4]、Park 等[5]等 4 种较为典型的方法进行实验对比。Zhang 等[19]首次使用深度神经网络用于深度图补全任务,并制作了 Matterport3D 数据集。Huang 等[20]在 Zhang[19]基础上在深度网络中加入了自注意力机制,实现在 Matterport3D 数据集上更优的补全深度图。而 Senushkin 等[4]提出的方法训练出的 DmLrn 模型在 Matterport3D 数据集上取得最佳表现,Park 等[5]提出的方法则在 NYUv2 数据集上取得最佳表现。本文采用均方根误差、平均绝对误差以及精确度作为深度图补全网络性能指标,误差越小、精度越高说明补全效果越好。

由表 1 和表 2 给出的深度图补全误差统计数据可知,DmLrn[4] 和本文方法均能够在不使用法线、物体边缘预测的情况下取得了较优结果。与 Zhang 等[19]、Huang 等[20]、Senushkin 等[4]方法相比,本文方法具有更好的补全结果。在 Matterport3D 数据集上,本文网络均方根误差为 0.309,较该数据集上最优模型 DmLrn[4] 相比降低了 0.652,准确度在阈值为 1.25 时精度提升 11%,在阈值为 1.25^2 时精度提升 6.9%,在阈值为 1.25^3 时精度提升 2.5%;在 NYUv2 数据集上,本文提出的方法与该数据集上的最优模型指标上基本相同。本文方法在编码器的特征提取阶段加入了通道注意力机制模块,在网络训练过程中充分学到了场景的上下文信息,从而提升深度图补全网络的性能。本文方法在 Matterport3D 数据集上取得了更优性能,表明本文方法在前、背景相差较大的大规模室内场景深度图补全方面有较大优势。

表 1　Matterport3D 数据集上的评估指标

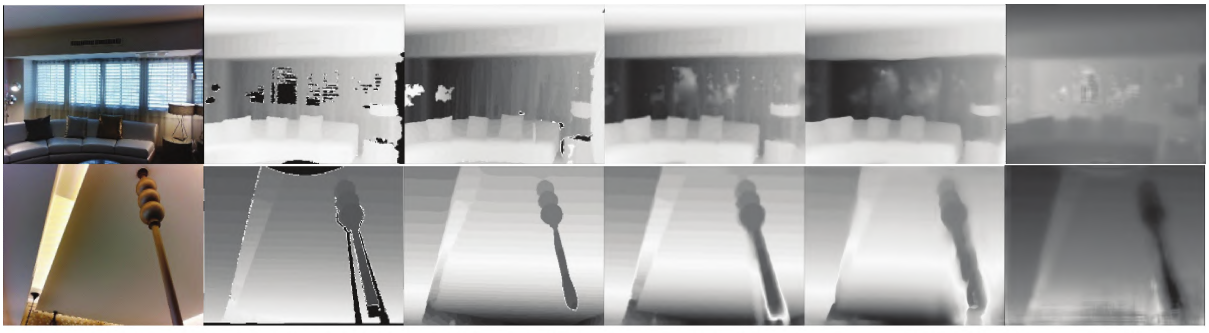
方法	E_{RMS}	E_{MA}	δ_t^1	δ_t^2	δ_t^3
Zhang 等[19]	1.316	0.461	78.1	85.1	88.8
Huang 等[20]	1.092	0.342	85.0	91.1	93.6
Senushkin 等[4]	0.961	0.285	81.3	89.0	94.9
本文方法	0.309	0.162	92.3	95.9	97.4

注:δ_t^1、δ_t^2、δ_t^3 分别表示 t 取 1.25、1.25^2、1.25^3 时深度图补全方法的精确度。下同。

表 2　NYUv2 数据集上的评估指标

方法	E_{RMS}	δ_t^1	δ_t^2	δ_t^3
Senushkin 等[4]	0.272	97.5	99.3	99.8
Park 等[5]	0.092	99.6	99.9	100.0
本文方法	0.116	99.5	99.9	100.0

图3为本文方法与Senushkin等[4]、Huang等[20]方法在Matterport3D数据集上的深度图补全效果。从图3可以看出,本文方法能在保证图像结构信息的基础上达到更优精度,DmLrn[4]在边缘处明显出现深度值混合的现象。本文的方法能够很好地将前景和背景分开,使得补全后的物体边缘更为清晰。图4为本文方法与Park等[5]、Senushkin等[4]方法在NYUv2数据集上的深度图补全效果。从图4可以看出,本文方法物体边缘以及总体补全效果均表现良好,补全后的深度图像具有清晰的边界。综上所述,本文提出的深度图补全网络在不同数据集上均取得较好效果。



(a)彩色图 (b)深度图 (c)场景真实深度图 (d)本文方法补全效果 (e)Senushkin等[4]补全效果 (f)Huang等[20]补全效果

图3 在Matterport3D数据集上不同方法的深度图补全效果

(a)彩色图 (b)深度图 (c)场景真实深度图 (d)本文方法补全效果 (e)Park等[5]补全效果 (f)Senushkin等[4]补全效果

图4 在NYUv2数据集上不同方法的深度图补全效果

3 结 语

本文提出了一种基于注意力机制的场景深度图补全的方法。在场景彩色图像引导下,编码器融合场景彩色图和深度图特征信息,在特征提取过程中使用注意力机制使网络更加关注补全任务所需要的区域,在深度图像补全任务中取得更高的准确度。实验表明,该方法可以较好地补全室内场景深度图中的缺失信息。

为了进一步改善深度图像补全的质量,后续研究可考虑将场景彩色图中的几何结构信息融入场景深度图补全网络,以提升补全后深度图像的边缘准确度。

参考文献:

[1] Fu C, Mertz C, Dolan J M. LIDAR and monocular camera fusion: On-road depth completion for autonomous driving [C]//2019 IEEE Intelligent Transportation Systems Conference (ITSC). Auckland, New Zealand. IEEE, 2019: 273-278.

[2] 梅峰,刘京,李淳秡,等. 基于RGB-D深度相机的室内场景重建[J]. 中国图象图形学报, 2015, 20(10): 1366-1373.

[3] Ping J M, Thomas B H, Baumeister J, et al. Effects of shading model and opacity on depth perception in optical see-through augmented reality[J]. Journal of the Society for Information Display, 2020, 28(11): 892-904.

[4] 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). New York: ACM, 2021: 2181-2188.

[5] Park J, Joo K, Hu Z, et al. Non-local spatial propagation network for depth completion[C]//16th European Conference on Computer Vision-ECCV 2020. Glasgow, UK. Cham: Springer, 2020: 120-136.

[6] 沙浩,刘越,王涌天, 等. 基于二维图像和三维几何约束神经网络的单目室内深度估计方法[J]. 光学学报, 2022,42(19):47-57.

[7] 江俊君,李震宇,刘贤明. 基于深度学习的单目深度估计方法综述[J]. 计算机学报,2022,45(6):1276-1307.

[8] 周萌,黄章进. 基于失焦模糊特性的焦点堆栈深度估计方法[J/OL]. 计算机应用. (2023-02-17)[2023-03-03]. http://kns.cnki.net/kcms/detail/51.1307.TP.20230217.1018.004.html.

[9] Xu D, Wang W, Tang H, et al. Structured attention guided convolutional neural fields for monocular depth estimation [C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. June 18-23, 2018, Salt Lake City, UT, USA. IEEE, 2018: 3917-3925.

[10] 白宇,梁晓玉,安胜彪. 深度学习的 2D-3D 融合深度补全综述[J/OL]. 计算机工程与应用. (2023-03-01)[2023-03-03]. http://kns.cnki.net/kcms/detail/11.2127.TP.20230228.1040.006.html.

[11] Uhrig J, Schneider N, Schneider L, et al. Sparsity invariant CNNs[C]//2017 International Conference on 3D Vision (3DV). October 10-12, 2017, Qingdao, China. IEEE, 2018: 11-20.

[12] Chen Y, Yang B, Liang M, et al. Learning joint 2D-3D representations for depth completion[C]// 2019 IEEE/CVF International Conference on Computer Vision (ICCV). Seoul, Korea (South). IEEE, 2019: 10022-10031.

[13] Li A, Yuan Z J, Ling Y G, et al. A multi-scale guided cascade hourglass network for depth completion[C]// 2020 IEEE Winter Conference on Applications of Computer Vision (WACV). March 1-5, 2020, Snowmass, CO, USA. IEEE, 2020: 32-40.

[14] 郑柏伦,冼楚华,张东九. 融合 RGB 图像特征的多尺度深度图像补全方法[J]. 计算机辅助设计与图形学学报, 2021, 33(9): 1407-1417.

[15] 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]//2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). June 15-20, 2019, Long Beach, CA, USA. IEEE, 2020: 3308-3317.

[16] Xu Y, Zhu X G, Shi J P, et al. Depth completion from sparse LiDAR data with depth-normal constraints [C]//2019 IEEE/CVF International Conference on Computer Vision (ICCV). Seoul, Korea (South). IEEE, 2020: 2811-2820.

[17] 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.

[18] Choi J, Jung D, Lee Y H, et al. SelfDeco: self-supervised monocular depth completion in challenging indoor environments [C]//2021 IEEE International Conference on Robotics and Automation (ICRA). Xi'an, China. IEEE, 2021: 467-474.

[19] Zhang Y, Funkhouser T. Deep depth completion of a single RGB-D image[C]// 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, UT, USA. IEEE, 2018: 175-185.

[20] Huang Y K, Wu T H, Liu Y C, et al. Indoor depth completion with boundary consistency and self-attention [C]//2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW). Seoul, Korea (South). IEEE, 2020: 1070-1078.

[21] Cheng X J, Wang P, Yang R G. Depth estimation via affinity learned with convolutional spatial propagation network[C]//15th European Conference on Computer Vision-ECCV 2020. Munich, Germany. Cham: Springer, 2018: 108-125.

[22] 杨宇翔,曹旗,高明煜, 等. 基于多阶段多尺度彩色图像引导的道路场景深度图像补全[J]. 电子与信息学报, 2022,44(11):3951-3959.

[23] 卢宏涛,罗沐昆. 基于深度学习的计算机视觉研究新进展[J]. 数据采集与处理,2022,37(2): 247-278.

[24] Shim G, Park J, Kweon I S. Robust reference-based super-resolution with similarity-aware deformable convolution [C]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). June 13-19, 2020, Seattle, WA, USA. IEEE, 2020: 8422-8431.

[25] 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. IEEE, 2018: 7794-7803.

[26] Wang Q L, Wu B G, Zhu P F, et al. ECA-net: efficient channel attention for deep convolutional neural networks [C]//2020 IEEE/CVF Conference on

Computer Vision and Pattern Recognition (CVPR). June 13-19, 2020, Seattle, WA, USA. IEEE, 2020: 11531-11539.

[27] Zhu X Z, Hu H, Lin S, et al. Deformable ConvNets V2: more deformable, better results[C]//2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). June 15-20, 2019, Long Beach, CA, USA. IEEE, 2020: 9300-9308.

(责任编辑:康 锋)