融合 RGB 图像特征的多尺度深度图像补全方法

郑柏伦, 冼楚华*, 张东九

(华南理工大学计算机科学与工程学院 广州 510064)
(chhxian@scut.edu.cn)

摘要: 针对目前因缺少配对的“缺失-完整”RGB-D数据集而不能直接训练端对端深度图像补全模型的问题, 提出基于随机掩码构造对应的缺失-完整数据, 结合真实数据集与合成数据集交替训练模型策略。基于随机掩码生成不同缺失比例的深度图像, 并且利用合成数据集构造具有可靠真值的深度图像缺失数据, 从而得到具有可靠数据的缺失-完整RGB-D数据集。以此策略为基础, 搭建融合对应RGB图像特征的多尺度深度图像补全网络, 该网络分别从RGB图像特征提取分支和深度图像特征提取分支提取不同尺度的RGB图像特征和深度图像特征, 再经过特征融合分支在不同尺度上对RGB图像特征和深度图像特征进行融合, 进而能够充分地学习RGB图像丰富的语义信息和深度图像的信息补全缺失深度。在NYU-Depth V2数据集的实验表明, 该方法在不同缺失比例深度图像补全任务中, 阈值精度平均值为0.98, 平均相对误差约为0.061, 与现有基于神经网络和优化稀疏方程组的方法相比, 其在阈值精度上平均提升了0.02, 平均相对误差平均下降了0.027。

关键词: 深度图像补全; 融合网络; 深度重建

中图分类号: TP391.41 **DOI:** 10.3724/SP.J.1089.2021.18861

Multi-Scale Fusion Networks with RGB Image Features for Depth Map Completion

Zheng Bolun, Xian Chuhua*, and Zhang Dongjiu

(School of Computer Science and Engineering, South China University of Technology, Guangzhou 510064)

Abstract: Currently, researchers cannot directly train end-to-end model for depth image completion because of lacking paired “incomplete-complete” RGB-D datasets. To address this problem, a random mask-based method which is combined with “real-synthetic” data for joint training strategy is proposed to construct paired incomplete-complete RGB-D data. This method generates depth maps with different missing ratios based on random masks, and uses synthetic scene datasets to construct missing regions of depth map with high fidelity truth values. Based on this strategy, a multi-scale depth map completion network is constructed, which fuses the corresponding RGB image features. The proposed network extracts RGB image features and depth map features with different scales from RGB image branches and depth map branches. Then, in the feature fusion branch, the RGB image features and depth map features are fused at different scales, which makes that the rich semantic features of RGB images and the information of depth maps can be effectively integrated for depth map completion. Experiments on the NYU-Depth V2 dataset show that in depth completion tasks with different missing ratios, the average threshold accuracy of this method is 0.98, and the mean relative error is about 0.061. Compared with the existing methods based on neural networks and optimizing sparse equations, this method improves the threshold accuracy

收稿日期: 2020-12-22; 修回日期: 2021-04-02. 基金项目: 广东省自然科学基金(2019A1515011793); 中央高校基本科研业务费(2020ZYGXZR042). 郑柏伦(1996—), 男, 硕士研究生, 主要研究方向为计算机图形学; 冼楚华(1982—), 男, 博士, 副教授, 硕士生导师, CCF 会员, 论文通讯作者, 主要研究方向为几何建模与处理、计算机图形学、计算机视觉、智能图形处理; 张东九(1995—), 男, 硕士, 主要研究方向为计算机图形学。

by an average of 0.02, and the mean relative error decreases by an average of 0.027.

Key words: depth map completion; fusion networks; depth restoration

如何理解室内场景并进行三维重建一直是计算机图形学重点研究方向,而如何采集完整的深度图像是室内场景三维感知和重建的关键。通常,对深度图像的采集有 2 种方式:(1) 使用 RGB-D 相机主动拍摄对象,从而获得对象的深度信息。但如果被拍摄对象表面十分光滑、存在透明区域和过于细小,或者在高光条件下,所获取的深度信息会不完整,存在缺损。(2) 利用数据学习“RGB 图像-深度图像”的关联关系,从 RGB 图像中重建出相应的深度图像。近年来,很多研究通过构建神经网络来学习这样的关联关系。但是当前的 RGB-D 数据集通常包含残缺的深度,而且这些残缺的深度信息不存在对应的真值数据,因此难以利用当前的神经网络模型完成诸如深度估计和深度图像补全之类的任务。

对于上述问题,本文提出基于随机掩码的“真实-合成”混合数据集联合训练策略,对现存的 RGB-D 数据集,使用随机掩码生成不同缺失比例的深度图像,同时针对真实数据集中固有深度缺失区域,使用合成数据集和随机掩码一同生成具有可靠深度真值的深度缺失区域,最终构造“缺失-完整”的深度数据。同时,基于以此策略生成的数据集,本文提出了融合 RGB 图像信息和深度图像信息的多尺度深度图像补全网络,采用 2 个分支分别提取 RGB 图像特征和深度图像特征,并用一个融合分支对二者进行融合,从而对缺失的深度图像进行有效的高精度补全。

1 相关工作

图像补全是指利用图像本身或图像库的信息填补图像缺损部分,并且要求补全的部分纹理自然且边缘平滑,符合图像整体的语义约束。本节主要介绍 RGB 图像的补全方法和深度图像补全方法的相关工作。

1.1 RGB 图像补全方法

传统的 RGB 图像补全方法有基于传播的方法^[1-3]和基于图像块的方法^[4-7]。基于传播的方法是指将缺损部分的边缘根据原来的角度将邻域信息传播到缺损部分之中。由于该方法是从边缘对内部进行信息传播,因此只对细小范围的缺损部分

有效。而基于图像块的方法修补效果受图像的低维信息影响巨大,因此遇到较为复杂的缺失区域则不能够很好地补全。

近年来,神经网络相关模型被用来解决图像补全问题。Pathak 等^[8]将自动编码解码器和生成对抗网络相结合,以提高图像补全能力。该方法首先将缺失图像编码到高维的特征空间,再使用解码器对生成的特征进行解码并恢复图像,生成对抗网络用来增加图片的细节。与传统方法相比,该方法在性能指标上有了很大的提高,因此衍生出一批对该方法改进的工作^[9-10]。这些方法均用于补全 RGB 图像,深度图像多数以灰度图或 16 位单通道图形式存储,直接应用上述 RGB 图像补全方法很明显得不到有效的结果。

1.2 深度图像补全方法

近年来,学界越发重视应用基于优化的方法解决深度图像补全问题。最初由 Levin 等^[11]提出该类方法,一开始用于解决图像着色问题。图像着色问题是指给定 Y 通道,重建图像的 U, V 通道。应用到深度图像补全问题时,通过把图像从 RGB 空间迁移到 YUV 空间以得到 Y 通道图像,此时如何修复深度缺失区域即相当于如何对图像进行着色。此类方法没有考虑几何约束,因此对于具有丰富纹理信息的部分或者缺失范围较大的部分,不能有效地补全深度图像。

Ku 等^[12]使用传统的图像处理方法对深度数据进行填充,它使用不同形状和大小的核对图像进行膨胀处理,但是它最后会对结果进行模糊处理以减少噪声,最终得到的深度图像精度并不高。Atapour-Abarghouei 等^[13]预先将深度图的空洞部分分类为 12 种情况,利用先验语义信息对空洞进行补全。Atapour-Abarghouei 等^[14]训练一个空洞生成网络来获取深度缺失区域的深度真值,但是通过该空洞生成网络生成的深度缺失区域符合一定的模型分布,不具有随机性,泛化性不足。Sun 等^[15]应用合作收益随机森林算法对深度图像进行补全,输入 RGB 图像和深度图像,不断迭代直至空洞消失。

Zhang 等^[16]首先提出了基于深度卷积网络的深度图像补全方法,利用表面网格重新生成完整的深度图像,以解决缺少训练数据的问题;然后应用一个深度神经网络预测表面法线和遮挡边界,

并根据表面法线和遮挡边界进行稀疏方程优化,求解得出图像的深度。但是,如果预测表面法线的精度不足会很大程度地影响最终的结果,稀疏方程优化也需要大量的运算时间,导致该模型的运行时间较长。

Liu等^[17]利用Zhang等^[16]提出的网络进行单幅图像的深度图像补全,对同一场景多角度拍摄深度图,并运用深度一致性验证和补全后的相邻深度图信息进行补全。该方法适用于多目RGB-D数据集,不适合本文研究的单目深度图像修复问题,其同样存在与Zhang等^[16]方法一样的耗时间问题。Shih等^[18]提出先对图像进行边缘补全,然后将缺失的深度图像和补全后的边缘图像输入到深度图像补全网络中,孟天禹等^[19]提出估计完整的曲面法线图像后,再将叠加曲面法线图像以及深度图像输入到神经网络中,利用曲面法线图指导深度图像修复工作。上述2种方法都忽略了RGB图像,没有充分利用RGB图像丰富的纹理信息。郑博伟等^[20]提出了根据统计特性一致的深度补全方法,但是其所提取的固有身份信息主要应用在人脸数据集中,本文着重于研究室内场景的深度图像补全,因此该方法并不适用于本文研究场景。

2 本文方法

2.1 基于随机掩码的真实-合成混合数据

本文基于随机掩码的真实-合成混合数据集联合训练策略主要包含以下2个方面。

2.1.1 随机掩码

通过二值图像(每个像素的取值只为0或1)来构造掩码样例,随机选取一个掩码样例与深度图像进行逐像素相乘,移除掩码取值是0的像素的数据,从而获取缺失-完整深度图像补全数据集。因为相机采集到的RGB图像一般情况下不会出现缺失区域,所以本文方法并不对RGB图像进行掩码操作。本文通过真实数据集收集了10万张缺失掩码图像。然而,真实场景中缺失区域越大,图片数量越少,这将导致不同缺失比例的样本数量不均衡,进而对模型的性能造成影响。为此,本文通过多种方式增加收集数据的样本数量,使得不同缺失程度的掩码样例数量均衡。增加样本数量的方式包括4种途径:0-1反转、垂直和水平翻转、旋转一定角度(90°以及270°),并设置扩增后的掩码样例分辨率为480×640像素,最后得到总数超过60万包括多种缺失比例的随机掩码数据集。

2.1.2 真实-合成数据集

使用上述随机掩码数据集可以构造具有可靠深度信息的缺失-完整深度图像补全数据集,能够有效地利用现有的大规模RGB-D数据集训练深度图像补全模型。但是,对于现有RGB-D数据集,其中一些高光和透明等区域,仍然缺少有效、真实的深度数据。为了解决这个问题,本文引入合成数据集,构造真实-合成数据集为训练提供监督和指导。本文采用IRS数据集^[21]作为合成数据的来源,其包含丰富的室内场景,如房屋、店铺、办公室和饭店等,且同一场景包含双目RGB图像和视差图像。本文通过视差信息和数据集中的相机参数计算出深度信息,如图1所示,通过几何知识可以根据视差图像推导出物体的深度,其转换关系为 $D = B \times f \div d$ 。其中, D 是转换后的深度值; d 是视差值; B 是双目相机基线长度, $B=0.1$ m; f 是相机焦距, $f=480$ 像素。

图1 双目图像与视差图转换深度图示例

本文同时采用NYU-Depth Raw数据集^[22]和ScanNet数据集^[23]作为训练集。NYU-Depth是最庞大的RGB-D数据集之一,包含多样化的场景,ScanNet^[23]是另外一个包含多个场景的RGB-D数据集,上述数据集都拥有多于50万张配对的RGB-D图像,它们都通过Kinect采集深度信息,因而很多深度图像都存在深度缺失的情况。本文对IRS数据集进行深度转换后生成的数据集,可以加入可信的深度真值,从而提高对这些固有缺失区域的补全能力。因此,本文将这2个真实数据集与转换后的IRS数据集共同作为训练集,构建真实-合成混合数据集进行联合训练,生成配对的缺失-完整深度数据。

本文提出的真实-合成混合数据集联合训练策略如图2所示,首先利用掩码样例对真实场景和合成场景生成缺失深度图,利用真实数据和合成数据对深度图像补全模型进行交替训练。

对于真实场景固有的深度缺失区域,本文用Zhang等^[16]方法进行初步的补全,获取完整但粗糙

图 2 联合训练策略示意图

不精确的深度图, 合成场景的深度图像是经过视差图转换得到, 深度信息是准确的, 因此能够作为补充数据对深度图像补全模型训练进行监督, 增强对固有缺失区域的深度图像补全效果。但是由于合成场景的风格、光照等特征与真实场景有一定的差别, 如果在真实场景中直接运行根据合成场景训练的深度图像补全模型, 效果并不理想; 而使用现在常见的迁移学习算法微调模型的参数以应用到新的数据集, 又会浪费真实数据集中完整深度区域。因此, 本文将真实数据集与合成数据集 1:1 混合, 构造出真实-合成混合数据集进行交替训练模型。这样, 既能充分利用大量的真实数据集中的 RGB-D 图像, 又能利用具有完整且可靠的深度

信息的合成数据集, 增强深度图像补全模型修复能力。

2.2 融合多尺度特征的深度图像补全模型

训练时直接拼接 RGB 图像和深度图像, 效果并不理想。这是因为不同的深度图像缺失的比例不一样, 导致在计算损失时会影响到收敛性。为解决这个问题, 本文提出分别对 RGB 图像和深度图像进行多层次的特征提取, 通过分支网络以不同尺度融合 RGB 图像特征和深度图像特征, 避免深度信息缺失部位和比例的不同造成的影响。如图 3 所示, 本文构建的是一个端到端的基于深度学习的模型, 包含 3 个分支, 分别是 RGB 图像特征提取分支, 深度图像特征提取分支和特征融合分支。

图 3 融合 RGB 图像特征与深度图像特征的多尺度深度图像补全网络

本文采用 DenseNet-169 提取 RGB 图像特征; 使用级联式卷积-池化层多层次提取深度图像特征; 最后特征融合分支合并不同层次相同尺度的 RGB 图像特征和深度图像特征, 并通过连续的“卷积-上采样”网络层将补全的深度图恢复到原始分辨率, 最后输出补全后的深度图像。

2.2.1 多尺度深度图像补全网络

较大尺度特征图具有更大范围的感受野, 更适合缺失范围较大的区域; 而较小尺度的特征图能够更好地感知细节, 更适合处理细小的深度缺失区域。因此, 本文在不同的网络层中提取不同尺度的特征, 然后通过上采样-融合模块拼接不同尺度的特征并进行上采样; 如此不断地变换尺度获取多尺度信息, 能够涵盖全局特征和局部特征, 最终达到恢复缺失深度的目的。整个多尺度深度图像补全网络处理流程如下。

Step1. 通过 ImageNet 数据集预训练的 DenseNet-169 提取 RGB 图像特征, 获得 6 张不同尺寸的 RGB 特征图, 第 1 张特征图的尺寸与输入相同, 其后的特征图尺寸为前一张特征图的 1/2。

Step2. 同时使用深度图像特征提取分支提取输入的不完整深度图像的特征, 获得 6 张不同尺寸深度特征图, 各个特征图的尺寸与 RGB 特征图一致。

Step3. 通过特征融合分支把相同大小的 RGB 图像特征与深度图像特征, 联合上一层级上采样-融合模块经过输出的融合特征进行拼接; 然后通过卷积层融合当前层级的三支特征, 当前层级输出的融合特征经过上采样后输入到下一层级的上采样-融合模块中, 最终获得与输入大小一致的补全深度图像。

本文模型的 3 个分支在数据集上进行联合训练, 使得 RGB 图像特征和深度图像特征相互补充。

2.2.2 RGB 图像特征提取分支

实验证明, 层数越多的神经网络可以获得更高维和更细粒度的特征, 但当层数变深, 神经网络会出现梯度消失问题。为此, Huang 等^[24]提出密集连接网络。与残差网络^[25]不同, 密集连接网络将传递到当前层的特征进行拼接, 而不是求和组合特征。本文采用预训练的 DenseNet-169 模型、1 个卷积层和 1 个池化层构成 1 个 RGB 图像特征提取模块。具体做法是从 DenseNet-169 模型中提取了尺寸分别为输入图像的 1/1, 1/2, 1/4, 1/8, 1/16 和 1/32 的特征图; 将它们的 RGB 图像特征通过特征融合模块与对应尺度的深度图像特征进行特征拼接。网络结构如图 3 “彩色图分支”所示。

2.2.3 深度图像特征提取分支

该分支使用级联式“卷积-池化”网络提取输入

的深度图像的多尺度特征。如图 3 的深度图分支所示, 该分支的主体部分为输入模块和 5 个下采样模块, 下采样模块的特征提取功能通过级联式卷积-池化操作完成。本文以 ReLU 函数作为卷积层的激活函数。在该分支中, 所有卷积核的大小相同, 设置为 3×3 , 步长设置为 1, 第 1 个下采样模块拥有 4 个卷积核, 之后的模块卷积核数目以 2 的倍数增长。本文的深度图像特征提取分支结构图如图 4 所示, 不同颜色的色块表示不同维度的卷积层, 它们可以提取不同尺寸的特征。

图 4 深度图像特征提取分支结构示意图

2.2.4 特征融合分支

RGB 图像特征和深度图像特征差异较大, 直接将拼接特征输入到 U-Net 结构^[26]不能够充分地挖掘数据中的信息。为得到更好的深度重建结果, 本文提出创建特征融合分支, 建立三支深度图像补全模型。特征融合分支包括多个上采样-融合模块。每个模块结构如图 5 所示, 模块有 3 个输入, 分别为 RGB 图像特征、深度图像特征和上层的融合特征。该特征融合模块首先对上层的融合特征进行上采样, 尺寸与 RGB 图像特征和深度图像特征相同, 然后将三者进行拼接, 再通过连续 2 个卷积层后输出, 输出的融合特征则是下层上采样-融合模块的输入。第 1 个上采样-融合模块的卷积层有 256 个卷积核, 后续模块的卷积核数量以 2 的倍数减少, 本文中特征融合分支共有 5 个上采样-融合模块。

图 5 上采样-融合模块

2.3 损失函数

现实场景的深度值是连续的, 因此回归分析的思想被应用于深度图像补全任务中. 平均绝对误差(mean absolute error, MAE)和均方误差(mean squared error, MSE)常被应用为解决回归问题的损失函数. 上述损失函数可以衡量模型补全性能, 但是, 上述损失函数都会忽略深度图像各个像素点之间存在的关联信息. 为了提高本文模型的深度图像补全性能, 挖掘各个像素点深度值的关联关系, 并尽量保证缺失区域补全后的边缘平滑过渡, 本文提出了考虑多方面约束的联合损失函数. 联合损失函数考虑的约束有点约束、梯度约束和结构约束. 联合损失函数表示为

$$L_{\text{total}} = \lambda_1 L_{\text{point}} + \lambda_2 L_{\text{grad}} + \lambda_3 L_{\text{ssim}}.$$

其中, L_{total} 为本文的联合损失函数; L_{point} 为点约束; L_{grad} 为梯度约束; L_{ssim} 为结构约束; λ_1, λ_2 和 λ_3 为权重系数.

(1) 点约束. MAE 和 MSE 都被应用为回归分析中的损失函数. 因为 MSE 是误差平方和的平均数, 平均效应可能导致缺失区域在修复后变得模糊. 所以本文使用 MAE 作为联合损失函数的点约束. 输入一幅完整深度图像, 其点约束定义为

$L_{\text{point}} = \frac{1}{W \times H} \sum_{i=1}^W \sum_{j=1}^H \|I_{i,j}^{\text{pred}} - I_{i,j}^{\text{gt}}\|_1$. 其中, W 表示深度图像的宽; H 表示深度图像的高; $I_{i,j}^{\text{pred}}$ 表示预测深度值; $I_{i,j}^{\text{gt}}$ 表示深度真值; $\|\cdot\|_1$ 表示 L_1 范数.

(2) 梯度约束. 与 RGB 图像不同, 深度图像的边缘特征更突出. 具体表现为深度值连续的区域, 其梯度约等于 0; 深度值间断的区域, 则梯度一般极大. 考虑上述情况, 本文将梯度约束加入联合损失函数之中, 分别计算预测深度的梯度与深度真值的梯度, 梯度约束使用二者的 MAE. 梯度约束定义为

$$L_{\text{grad}} = \frac{1}{W \times H} \sum_{i=1}^W \sum_{j=1}^H \left\| \nabla_x (I_{i,j}^{\text{pred}}) - \nabla_x (I_{i,j}^{\text{gt}}) \right\|_1 + \frac{1}{W \times H} \sum_{i=1}^W \sum_{j=1}^H \left\| \nabla_y (I_{i,j}^{\text{pred}}) - \nabla_y (I_{i,j}^{\text{gt}}) \right\|_1.$$

其中, ∇_x 表示图像垂直方向梯度; ∇_y 表示图像水平方向梯度.

(3) 结构约束. 结构相似度(structural similarity measurement, SSIM)反映了 2 幅图像之间的相似性, 它在计算图像相似性时考虑了结构等多方面

因素, 其值越大越好. 其定义为

$$\text{SSIM}(x, y) = \frac{(2\mu_x \mu_y + c_1)(2\sigma_{xy} + c_2)}{(\mu_x^2 + \mu_y^2 + c_1)(\sigma_x^2 + \sigma_y^2 + c_2)}.$$

其中, x, y 分别表示深度预测值和深度真值; μ_x, μ_y 分别表示深度预测值和深度真值的均值; σ_x^2, σ_y^2 分别为深度预测值和深度真值的方差; σ_{xy} 表示 x, y 的协方差; $c_1 = (k_1 L)^2$, $c_2 = (k_2 L)^2$ 是 2 个常数, 避免除零的情况出现. 在本文中设置 L 为像素范围, $k_1 = 0.01$, $k_2 = 0.03$. 通过滑动窗口来计算最后的全局 SSIM, 给定滑动窗口的大小为 $K \times K$, 设置 $K=11$, 最后取平均值. 结构约束定义为

$$L_{\text{SSIM}} = \frac{1}{W \times H} \sum_{i=1}^W \sum_{j=1}^H \text{SSIM}(I_{i,j}^{\text{pred}}, I_{i,j}^{\text{gt}}, K).$$

其中, W 和 H 分别表示深度图像的宽度和高度.

3 实验及结果分析

本文从 NYU-Depth, ScanNet 这 2 个真实数据集, 以及转化为深度图像后的 IRS 合成数据集中, 分别随机抽取 6 万幅、4 万幅和 6 万幅图像构成训练集. 此外, 本文采用水平和垂直翻转、裁切、重组颜色通道以及随机丢弃部分深度 4 种方式扩充训练集. 实验使用 Adam 优化算法对梯度进行优化, 训练的批样本数量设定为 4, 学习率设定为 0.000 1, 在超过 20 万幅图像的数据集上训练模型, 模型共迭代 50 万次后收敛. 用 4 个衡量指标平均相对误差(mean relative error, MRE), 均方根误差(root mean squared error, RMSE), 平均对数误差(mean log₁₀ error, $E_{\lg}$)和阈值精度(threshold accuracy, δ_i)评价实验结果. 4 种衡量指标分别定义为

$$\text{MRE} = \frac{1}{n} \sum_p \frac{|y_p - \hat{y}_p|}{y};$$

$$\text{RMSE} = \sqrt{\frac{1}{n} \sum_p (y_p - \hat{y}_p)^2};$$

$$E_{\lg} = \frac{1}{n} \sum_p |\lg(y_p) - \lg(\hat{y}_p)|;$$

$$\% \text{ of } y_p \text{ s.t. } \max\left(\frac{y_p}{\hat{y}_p}, \frac{\hat{y}_p}{y_p}\right) = \delta < t^i.$$

其中, t 的取值为 1.25; i 的取值为 1, 2, 3; y_p 是单个像素的深度真值; $\hat{y}_p$ 是对应像素的预测深度值; n 是输入深度图像中具有可靠深度值的像素

数目. 为了避免不同补全方法得到的深度不同对实验结果的影响, 只考虑输入深度图像中具有可靠深度的像素.

3.1 不同骨架网络的性能影响

本文采用DenseNet作为RGB图像特征提取分支的骨架网络. 由于DenseNet存在多个版本的骨架网络, 如DenseNet169和DenseNet121. 这些不同骨架网络的深度和参数都各不相同, 应用在相同的数据集上也有着不同的性能表现. 为了获得最好的特征提取效果, 对RGB图像特征提取分支, 本文首先使用不同骨架网络对深度图像补全模型的性能影响进行对比实验. 分别以DenseNet-121和DenseNet-169作为RGB图像特征提取分支的骨架网络, 使用NYU-Depth V2评测本文模型的性能. 从表1可以看出, DenseNet-169在6个性能衡量指标上都要优于DenseNet-121. 该实验说明, 选用DenseNet-169提取RGB图像特征, 能够获得更好的深度图像补全效果.

表1 DenseNet参数对深度图像补全能力的影响

DenseNet	$\delta_1 \uparrow$	$\delta_2 \uparrow$	$\delta_3 \uparrow$	MRE $\downarrow$	RMSE $\downarrow$	$E_{lg} \downarrow$
121	0.93	0.98	0.99	0.066	0.26	0.027
169	0.94	0.99	0.99	0.057	0.24	0.025

注. 加粗数字表示最好的评价结果.

3.2 生成的缺失-完整数据集有效性验证

目前的RGB-D数据集都存在固有的深度缺失部分, 这些缺失部分又没有可靠的深度真值. 为了能够利用这些大规模的RGB-D数据集, 很多研究都会运用各自的方法补全缺失区域或创建新的完整深度图像. 例如, Zhang等^[16]利用多角度表面重建技术, 从场景中提取出上百万个三角网格, 并在场景中对RGB-D图像进行采样, 从图像视点的摄像机姿态渲染重建网格, 最终获得完整深度数据. 本文采用第2.1节介绍的基于随机掩码的真实-合成混合数据集. 这些模型在自己生成的训练集训

练以后均在公开的数据集上进行实验, 以证明其有效性, 本文也不例外. 本节选用NYU-Depth V2数据集进行补全实验, 证明本文的基于随机掩码的真实-合成混合数据集的有效性.

NYU-Depth V2因客观原因存在固有缺损部分, 本文从中选取了缺损比例低于5%的部分图像作为实验的测试集, 测试集共有400对RGB-深度图像对应组合, 再从掩码数据集抽取缺失比例分别为30%, 50%, 70%以及90%的4个级别的掩码样例, 各个缺失比例均抽取4000幅掩码样例图像, 然后将掩码样例与测试集的图像逐元素相乘, 最终生成16000对RGB-深度图像对, 作为最终的测试集来评估本文方法和对比方法的实验结果. 实验结果如表2所示, 其中对比方法是当前深度图像补全中效果最好的方法. 从表2可以看出, 无论是哪种缺失比例, 本文方法的6个性能指标都比Zhang等^[16]的方法更好. 这充分说明了本文方法能够有效地修复缺失的深度信息, 本文的基于随机掩码的真实-合成混合数据集能有效地训练深度补全模型. 另外要特别指出, 从表2中可以看出, Zhang等^[16]方法与本文方法的差距随着缺失比例的增大而增大, 说明本文方法能够更有效地修复大范围的缺失区域.

为了方便地观察本文方法的效果, 本文从数据集中抽取了4个具体例子进行可视化操作, 可视化效果如图6所示. 其中, 第3列和第4列各幅图像上的数值是该图像预测深度和真值的MSE, 可以客观衡量不同深度图像补全方法的效果. 观察图6可以发现, 本文方法大幅度优于Zhang等^[16]方法. 从图6第1行场景和第3行场景中可以看出, Zhang等^[16]方法对于大规模缺失部分基本不能进行有效的深度图像补全, 而本文方法仍然有比较好的效果; 对于如第4行部分缺失的区域, 本文的补全效果也优于Zhang等^[16]方法.

表2 NYU-Depth V2数据集对照实验结果

方法	缺失比例/%	$\delta_1 \uparrow$	$\delta_2 \uparrow$	$\delta_3 \uparrow$	MRE $\downarrow$	RMSE $\downarrow$	$E_{lg} \downarrow$
Zhang等 ^[16]	30	0.994	0.999	0.999	0.014 7	0.086 3	0.002 81
本文		0.998	0.999	0.999	0.014 4	0.070 1	0.002 49
Zhang等 ^[16]	50	0.992	0.998	0.999	0.027 5	0.114 2	0.004 54
本文		0.997	0.999	0.999	0.024 5	0.087 5	0.004 08
Zhang等 ^[16]	70	0.983	0.995	0.997	0.061 4	0.177 1	0.010 68
本文		0.991	0.997	0.998	0.047 2	0.129 1	0.007 95
Zhang等 ^[16]	90	0.864	0.960	0.988	0.246 9	0.483 9	0.042 76
本文		0.937	0.990	0.998	0.156 9	0.289 9	0.027 41

注. 加粗数字表示最好的评价结果.

图 6 对随机移除深度信息的场景进行补全的可视化实验结果

3.3 真实场景深度图像补全能力对比

为了衡量本文方法在真实场景补全固有缺失区域的效果,本节设计实验对真实场景进行深度图像补全.因为固有缺失区域缺少深度真值,所以本节对补全结果进行可视化处理并且从可视化结果评估模型性能.对真实场景的补全结果如图 7 所示.值得注意的是,如图 7 第 1 行场景以及第 2 行场景的红色方框标记所示,Zhang 等^[16]方法对这些区域全部失效,补全的结果不可信,由此可以说明其方法存在不稳定性,不稳定的结果应用到实际生活中会带来很大的影响.相比之下,本文方法的补全效果更为稳定,可以避免同一区域深度完全错误的情况.观察第 3 行和第 4 行红色方框标记区域可以发现,与 Zhang 等^[16]方法相比,本文方法能够更好地补全大范围的深度缺失区域.观察第 5 行的红色方框标记区域可以发现,本文方法在边缘的深度图像补全结果更加平滑和准确.上述结果说明,无论是补全结果的稳定度、更大比例缺失的补全准确度,还是边缘的平滑度,本文方法都要优于对比方法.实验结果证明,本文方法能够有效地对真实场景进行深度图像补全,且补全效果要优于 Zhang 等^[16]方法.

3.4 稀疏深度图像补全能力对比

对修复稀疏深度图像同样是计算机图形学的研究热点.该问题要解决的是对于稀疏深度图像,如何根据 RGB 图像的纹理和光照等丰富的特征,

修复得到精确完整的深度信息.稀疏的深度图通常出现在通过激光雷达对室外场景获取深度信息的过程中,室内场景基本通过深度相机采集得到稠密深度图像.因此,目前大多数研究通过对室内的深度图像随机采样生成稀疏深度图像.

为了证明本文深度图像补全模型具有良好的泛化性,将本文方法直接应用到室内场景稀疏深度补全任务中.与目前大多数研究相似,本文对完整的深度图像随机取样,模拟稀疏深度图像,应用本文方法对其进行补全.实验结果如表 3 所示,对比方法中, Lee 等^[27]以及 Cheng 等^[28]的方法仅能够补全稀疏深度, Zhang 等^[16]与本文的方法理论上可以对所有深度图像进行补全.表 3 中采样率表示随机抽样的像素与总像素的比例,通常,采样率与方法的性能成正比.在计算效率上,本文方法的计算耗时比 Zhang 等^[16]缩减了数百倍.而 Cheng 等^[28]方法目前在室内稀疏深度修复任务中表现最好,但由于其在深度补全过程中需要对结果进行数次迭代优化,因此在时间效率上不如本文方法.在表 3 展示的实验结果中可以看到,与稠密深度补全任务相比, Zhang 等^[16]方法性能表现明显退化.本文方法同样直接应用到稀疏深度图像补全任务中,没有特定的优化,采样率也远低于 Zhang 等^[16]和 Lee 等^[27],但性能表现依然优于其他方法.上述实验结果也证明了本文方法具有良好的泛化性和鲁棒性.

图7 真实场景深度图像补全结果样例

同样,为了更直观地展示稀疏深度图像补全效果,本文从数据集里抽取了6个例子进行对比,结果如图8所示。其中,图8c和图8d图像上的数值为该样例的MSE,可以客观地衡量不同深度图像补全方法的效果。从图8可以看出,虽然本文模型没有对稀疏深度图像进行特别的训练和优化,但是本文方法不管是完整性还是准确性,在稀疏深度图像补全任务中均有良好的表现。观察图8可

以发现,Cheng等^[28]的深度图像补全结果普遍平滑性不足,且其不能有效地补全物体中细小的区域,如第2行方框标示的椅子腿和桌腿,第3行方框标示的椅子腿部分和第4行方框标示的车把和前后轮连接处;而本文方法能够很好地补全此类区域的深度。实验结果证明,本文方法能够应用到各种缺失程度的深度图像补全任务中,并且具备应用到补全室外场景稀疏深度的可能性。

表3 不同方法稀疏深度图像补全实验结果

方法	$\delta_1 \uparrow$	$\delta_2 \uparrow$	$\delta_3 \uparrow$	MRE $\downarrow$	RMSE $\downarrow$	$E_{lg} \downarrow$	采样率/%	运行时间/s
Zhang等 ^[16]	0.971			0.0409	0.168		1.01	12.3
Lee等 ^[27]	0.983			0.0403	0.142		1.01	
Cheng等 ^[28]	0.982	0.993	0.998	0.0316	0.187	0.0141	0.45	0.262
本文	0.994	0.998	1.000	0.0280	0.121	0.0119	0.45	0.033

注:加粗数字表示最好的评价结果。

图 8 稀疏深度图像补全任务可视化实验结果

4 结 语

针对深度图像缺失区域补全的问题, 本文提出了融合 RGB 图像特征和深度图像特征的多尺度深度图像补全方法. 针对当前大量 RGB-D 数据集由于固有的深度缺失区域而不能直接训练端对端深度补全模型的问题, 提出结合掩码数据集和真实-合成混合数据集生成缺失-完整深度数据集, 并基于此数据集构建了融合多尺度特征的深度图像补全模型. 实验表明, 本文方法的重建效果在常见的性能衡量指标中都优于对比方法, 在选取的方法中表现最好. 这说明了本文提出的深度图像补全模型在室内场景深度图像补全任务的有效性和鲁棒性.

参考文献(References):

- [1] Bertalmio M, Sapiro G, Caselles V, *et al.* Image inpainting[C] //Proceedings of the 27th Annual Conference on Computer Graphics and Interactive Techniques, New York: ACM Press, 2000: 417-424
- [2] Ballester C, Bertalmio M, Caselles V, *et al.* Filling-in by joint interpolation of vector fields and gray levels[J]. IEEE Transactions on Image Processing, 2001, 10(8): 1200-1211
- [3] Levin A, Zomet A, Weiss Y. Learning how to inpaint from global image statistics[C] //Proceedings of the IEEE International Conference on Computer Vision. Los Alamitos: IEEE Computer Society Press, 2003: 305-312
- [4] Efros A A, Freeman W T. Image quilting for texture synthesis and transfer[C] //Proceedings of the 28th Annual Conference on Computer Graphics and Interactive Techniques. New York: ACM Press, 2001: 341-346
- [5] Kwatra V, Essa I, Bobick A, *et al.* Texture optimization for example-based synthesis[J]. ACM Transactions on Graphics, 2005, 24(3): 795-802
- [6] Barnes C, Shechtman E, Goldman D B, *et al.* The generalized patch match correspondence algorithm[C] //Proceedings of European Conference on Computer Vision. Heidelberg: Springer, 2010: 29-43
- [7] Huang J B, Kang S B, Ahuja N, *et al.* Image completion using planar structure guidance[J]. ACM Transactions on Graphics,

- 2014, 33(4): 1-10
- [8] Pathak D, Krahenbuhl P, Donahue J, *et al.* Context encoders: feature learning by inpainting[C] //Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2016: 2536-2544
- [9] van den Oord A, Kalchbrenner N, Espeholt L, *et al.* Conditional image generation with PixelCNN decoders[C] //Proceedings of the 30th Annual Conference on Neural Information Processing Systems. Cambridge: MIT Press, 2016: 4797-4805
- [10] Goodfellow I J, Pouget-Abadie J, Mirza M, *et al.* Generative adversarial nets[C] //Proceedings of the 27th Annual Conference on Neural Information Processing Systems. Cambridge: MIT Press, 2014: 2672-2680
- [11] Levin A, Lischinski D, Weiss Y. Colorization using optimization[J]. ACM Transactions on Graphics, 2004, 23(3): 689-694
- [12] Ku J, Harakeh A, Waslander S L. In defense of classical image processing: fast depth completion on the CPU[C] //Proceedings of the 15th Conference on Computer and Robot Vision. Los Alamitos: IEEE Computer Society Press, 2018: 16-22
- [13] Atapour-Abarghouei A, Breckon T P. DepthComp: real-time depth image completion based on prior semantic scene segmentation[C] //Proceedings of British Machine Vision Conference. Guildford: BMVA Press, 2017: 208.1-208.13
- [14] Atapour-Abarghouei A, Akcay S, de La Garanderie G P, *et al.* Generative adversarial framework for depth filling via wasserstein metric, cosine transform and domain transfer[J]. Pattern Recognition, 2019, 91: 232-244
- [15] Sun J Y, Lin Q, Zhang X G, *et al.* Kinect depth recovery via the cooperative profit random forest algorithm[C] //Proceedings of the 11th International Conference on Human System Interaction. Los Alamitos: IEEE Computer Society Press, 2018: 57-62
- [16] Zhang Y D, Funkhouser T. Deep depth completion of a single RGB-D image[C] //Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2018: 175-185
- [17] Liu H M, Tang X C, Shen S H. Depth-map completion for large indoor scene reconstruction[J]. Pattern Recognition, 2020, 99: 107112
- [18] Shih M L, Su S Y, Kopf J, *et al.* 3D photography using context-aware layered depth inpainting[C] //Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2020: 8028-8038
- [19] Meng Tianyu, Chen Chunyu, Zhao Dandan, *et al.* Depth image inpainting algorithm guided by surface normal[J]. Applied Science and Technology, 2021, 48(2): 58-63(in Chinese)
(孟天禹, 陈春雨, 赵丹丹, 等. 曲面法线指引下的深度图像修复算法[J]. 应用科技, 2021, 48(2): 58-63)
- [20] Zheng Bowei, Li Bin, Li Yanran. Towards statistical consistency for deep image inpainting[J]. Journal of Signal Processing, 2020, 36(9): 1511-1524(in Chinese)
(郑博伟, 李斌, 李炎然. 面向统计特性一致的深度图像修复方法[J]. 信号处理, 2020, 36(9): 1511-1524)
- [21] Wang Q, Zheng S Z, Yan Q S, *et al.* IRS: a large synthetic indoor robotics stereo dataset for disparity and surface normal estimation[OL]. [2020-12-22]. <https://arxiv.org/abs/1912.09678>
- [22] Silberman N, Hoiem D, Kohli P, *et al.* Indoor segmentation and support inference from RGBD images[C] //Proceedings of European Conference on Computer Vision. Heidelberg: Springer, 2012: 746-760
- [23] Dai A, Chang A X, Savva M, *et al.* ScanNet: richly-annotated 3D reconstructions of indoor scenes[C] //Proceedings of the 29th IEEE Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2017: 2432-2443
- [24] Huang G, Liu Z, van der Maaten L, *et al.* Densely connected convolutional networks[C] //Proceedings of the 29th IEEE Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2017: 2261-2269
- [25] He K M, Zhang X Y, Ren S Q, *et al.* Deep residual learning for image recognition[C] //Proceedings of the 29th IEEE Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2016: 770-778
- [26] Ronneberger O, Fischer P, Brox T. U-Net: convolutional networks for biomedical image segmentation[C] //Proceedings of the 18th International Conference on Medical Image Computing and Computer-Assisted Intervention. Heidelberg: Springer, 2015: 234-241
- [27] Lee B, Jeon H G, Im S, *et al.* Depth completion with deep geometry and context guidance[C] //Proceedings of the International Conference on Robotics and Automation. Los Alamitos: IEEE Computer Society Press, 2019: 3281-3287
- [28] Cheng X J, Wang P, Yang R G, *et al.* Depth estimation via affinity learned with convolutional spatial propagation network[C] //Proceedings of European Conference on Computer Vision. Heidelberg: Springer, 2018: 108-125