语义增强的多视立体视觉方法

韩　燮[1,2,3],王若蓝[1,2,3],赵　融[1,2,3]
(1. 中北大学 大数据学院,山西 太原 030051;
2. 山西省视觉信息处理及智能机器人工程研究中心,山西 太原 030051;
3. 机器视觉与虚拟现实山西省重点实验室,山西 太原 030051)

摘　要:针对在基于深度学习技术的特征提取网络中,深层次的卷积神经网络提取的特征缺乏低级语义信息的问题,该文提出了语义增强的多视立体视觉方法。首先,提出了一种 ConvLSTM(Convolutional Long Short-Term Memory)语义聚合网络,通过使用 ConvLSTM 网络结构,对多个卷积层提取的特征图进行预测,得到融合每层语义信息的特征图,有助于在空间上层层抽取图像的高级特征时,利用长短期记忆神经网络结构的记忆功能来增强高层特征图中的低级语义信息,提高了弱纹理区域的重建效果,提高了 3D 重建的鲁棒性和完整性;其次,提出了一种可见性网络,在灰度图的基础上,通过突出特征图上可见区域的特征,加深了可见区域在特征图中的影响,有助于提高三维重建效果;最后,提取图像的纹理信息,并进入 ConvLSTM 语义聚合网络提取深层次特征,提高了弱纹理区域的重建效果。与主流的多视立体视觉重建方法相比,重建效果较好。

关键词:三维重建;深度学习;多视立体视觉;特征提取;语义聚合网络
中图分类号:TP391　　**文献标识码**:A　　**文章编号**:1673-629X(2024)03-0041-08
doi:10.3969/j.issn.1673-629X.2024.03.007

Semantic-enhanced Multi-view Stereo Vision Approach

HAN Xie[1,2,3],WANG Ruo-lan[1,2,3],ZHAO Rong[1,2,3]
(1. School of Data Science and Technology,North University of China,Taiyuan 030051,China;
2. Shanxi Province's Vision Information Processing and Intelligent Robot Engineering Research Center,Taiyuan 030051,China;
3. Shanxi Key Laboratory of Machine Vision and Virtual Reality,Taiyuan 030051,China)

Abstract:We propose a semantic-enhanced multi-view stereo vision method that aims to address the issue of deep convolutional neural networks lacking low-level semantic information in their feature extraction. Firstly, we propose a ConvLSTM (Convolutional Long Short-Term Memory) semantic aggregation network that uses the ConvLSTM network structure to predict the feature map extracted by multiple convolutional layers. This approach results in a feature map that integrates the semantic information of each layer, allowing us to extract high-level features layer by layer in space. The long short-term memory neural network structure's memory function enhances the low-level semantic information in the high-level feature map, leading to improved reconstruction in weak texture regions and greater robustness and integrity in 3D reconstruction. Secondly, we propose a visibility network that highlights the visible area's characteristics on the feature map and deepens the visible area's influence, resulting in better three-dimensional reconstruction. Finally, the texture information of the image is extracted and entered into the ConvLSTM semantic aggregation network to extract the deep-level features, which improves the reconstruction effect of the weak texture area. Compared with mainstream multi-view stereo vision reconstruction methods, the reconstruction effect is better.

Key words:3D reconstruction;deep learning;multi-view stereo vision;feature extraction;semantic aggregation network

0　引　言

计算机视觉领域的三维重建技术[1-2]已经应用于生活中的各个领域,如自动驾驶、文物修复、虚拟现实、智能家居等。多视立体视觉(Multi-Vision Stereo,

收稿日期:2023-04-06　　**修回日期**:2023-08-09
基金项目:国家自然科学基金(62272426,62106238)
作者简介:韩　燮(1964-),女,教授,博士,通信作者,研究方向为计算机视觉、仿真与可视化、智能信息处理;王若蓝(1999-),女,硕士研究生,CCF 会员(D7980G),研究方向为人工智能与计算机视觉。

MVS)[3]方法作为三维重建技术中的一种,使用人工设计的特征计算立体匹配的代价,聚合代价并进行视差计算和优化,得到像素的深度值,实现场景的三维重建,是计算机视觉领域中的经典方法。

传统的 MVS 方法主要分为四类,分别是基于点云扩散的 MVS 方法[4-5]、曲面演化的 MVS 方法[6]、基于体素的 MVS 方法[7-8]和基于深度图的 MVS 方法[9-10]。

基于点云的 MVS 方法是将初始的点云生成面片,并对面片迭代地传播以进行三维重建,难以并行化操作;曲面演化的 MVS 方法需要对待恢复场景的表面进行猜测,并根据多视图之间的光度一致性进行迭代重建表面,易出现离散型错误;基于体素的 MVS 方法将场景空间表示为 3D 体素,通过标记场景表面的体素以实现重建,占用内存较大,不适用于大场景空间的重建任务;基于深度图的方法将三维重建分解为多个根据立体图像对进行单深度图估计的小任务,快速灵活并且占用内存较小,适用于海量图像的大场景重建任务。目前很多优秀的 MVS 方法都是基于深度图的 MVS 方法。

随着深度学习技术在计算机视觉领域中的快速发展,将深度学习技术与多视立体视觉方法结合以提高重建效果也成为一种趋势。深度学习技术中的卷积神经网络[11](Convolutional Neural Networks,CNN)能够从图像中抽取高层语义特征,通过学习并利用场景全局语义信息,获得更稳健的匹配和实现更完整的重建。卷积神经网络在许多基于深度学习技术的 MVS 方法中展现出了很大的优势。PVAMVSNet(Pyramid Multi-view Stereo Net with Self-adaptive View Aggregation)方法[12]提出了一种有效且高效的多视图立体网络,使用基于 2D 全连接卷积神经网络的 UNet 网络进行特征提取,并在自适应视图聚合网络的基础上,配合多尺度图像的输入进行准确和完整的三维重建。TransMVSNet 方法[13]对特征提取步骤进行研究,提出了一个基于 2D CNN 的特征匹配转换器,利用注意力聚合图像内和图像之间的远程上下文信息,改善了三维重建的效果。CasMVSNet[14](Cascade Cost Volume for High-resolution Multi-view Stereo)在图像的特征提取阶段,使用特征金字塔网络进行图像特征的提取,在提取图像的高级特征的同时,融合了多种不同尺度下的信息,有利于深度值的准确预测。目前,虽然基于深度学习的 MVS 方法利用卷积神经网络大大提高了重建效果,取得了巨大的成功,但仍有一定的问题。在卷积神经网络中,不同层次产生的特征图的语义特征具有巨大差异。浅层的特征图语义信息较少,但细节特征较多,目标位置准确;深层的特征图语义信息丰富,但目标位置模糊。上述基于深度学习的 MVS 方法使用卷积神经网络提取的包含高级语义特征但缺乏低级更精细表示的特征图构建成本体,忽视了深层特征图的语义信息缺陷,导致在弱纹理区域的重建效果较差。

针对上述问题,该文提出使用基于卷积层的长短期记忆神经网络结构代替普通的 CNN 提取图像特征,这有助于在空间上层层抽取图像的高级特征时,利用长短期记忆神经网络结构的记忆功能来增强高层特征图中的低级语义信息,弥补了单纯 CNN 提取的特征图中低级特征缺失,提高了精细特征的表达,提升了弱纹理区域的重建效果。

与此同时,许多基于深度学习的主流 MVS 方法也致力于研究如何让匹配代价或者深度图预测更加精确[15-18]。其中,2019 年 Luo 等人[16]设计了匹配置信度模块,先学习聚合提取的图像特征的像素级对应信息,再通过匹配置信度模块对不同采样平面匹配置信度进行聚合,该模块可以有效地抑制数据的噪声部分。Fast-MVSNet 方法[17]在局部区域中利用 CNN 编码像素之间的深度相关性,产生密集的高分辨深度,并利用高斯牛顿层(Gaussian Newton)对深度进行持续优化。Luo 等人又提出了 AttMVS(Attention-Aware Deep Neural Network)方法[18],设计了一种新的注意增强匹配置信度模块,在一般由图像特征得到的像素级匹配置信度的基础上加入了局部区域的上下文信息,逐层次聚合并正则化为概率体,增强了匹配置信度。在上述研究的基础上,该文提出可见性网络与纹理特征提取,通过提取纹理特征增加纹理信息,并将源图像与参考图像转换为灰度图,消除颜色对光照变化的敏感,然后使用 CNN 计算相似度之后归一化,通过可见度值区别可见区域与不可见区域,并作用在特征提取网络中,增强可见区域的特征,提高 3D 重建的鲁棒性和完整性。

综上所述,该文在 VA-MVSNet 方法的基础上进行改进,主要贡献体现在以下三个方面:

(1)提出了一种使用 ConvLSTM 网络的新型特征提取网络。在对图像进行特征提取的同时,通过使用每层的特征图来预测最终的特征图,在深层的特征图语义中丰富缺失的浅层的语义信息,进行稳健的全局语义信息聚合。

(2)提出了一种可见性网络,首先进行灰度图像的提取,排除光照变化的影响,其次突出可见区域的特征,并指导特征图的生成,加深了可见区域在特征图中的影响,有助于提高三维重建效果。

(3)对图像进行纹理特征的提取,增加纹理特征的影响,提升目标场景的三维结构的重建效果。

1　网络架构

文中方法的整体网络架构如图 1 所示。假设图像大小为 ($H \times W \times 3$),深度假设的范围为 D,方法的输入由 1 个参考图像和 $N-1$ 个源图像构成。该方法在 VA-MVSNet 方法的基础上主要通过 ConvLSTM 语义聚合网络、可见性网络和纹理特征提取模块进行语义特征的提取和增强,之后通过可微单应性变换扭曲到参考图像的相机视锥内建立 3D 特征体 ($H \times W \times D \times C$),之后使用 VA-MVSNet 方法的自适应视图聚合方法得到代价体,最后使用 3D 正则化网络进行深度图的预测。

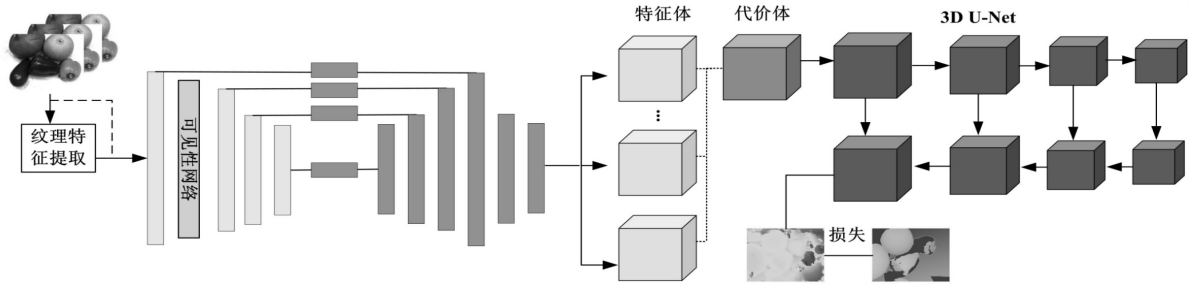

图 1　网络架构

该方法主要通过在特征提取阶段利用语义增强的思想进行改进,即特征提取时的 ConvLSTM 语义聚合网络(第 1.1 节)、可见性网络(第 1.2 节)、纹理特征提取模块(第 1.3 节),以增强特征图中的语义信息,提升三维结构的重建效果。

1.1　ConvLSTM 语义聚合网络

在基于深度学习的多视图立体重建方法中,特征提取网络通常由深层卷积神经网络组成,以提取更抽象的高层语义特征。虽然这些特征更高级、更抽象,但缺少了低层语义信息所具备的优势,过分强调高层特征会影响到细节语义信息的提取和准确位置的获取,不利于提升整体三维重建效果。ConvLSTM 是一种特殊的循环神经网络,可以记忆很久之前的信息,捕捉和学习长期依赖关系,并提取空间特征,非常适合于处理和预测具有长时间间隔和延迟的卷积神经网络之间的图像[19]。

因此,该文提出了一个基于 ConvLSTM 的语义聚合网络。该网络利用 ConvLSTM 的记忆功能和空间相关性提取功能,将高层语义和低层语义聚合,生成特征图。这样生成的特征图,除了具备更高层次的语义信息之外,还具备更低层次的语义信息,避免了普通多层卷积神经网络的缺点,从而提高了整体场景空间的重建效果。ConvLSTM 的内部运行过程如图 2 所示。

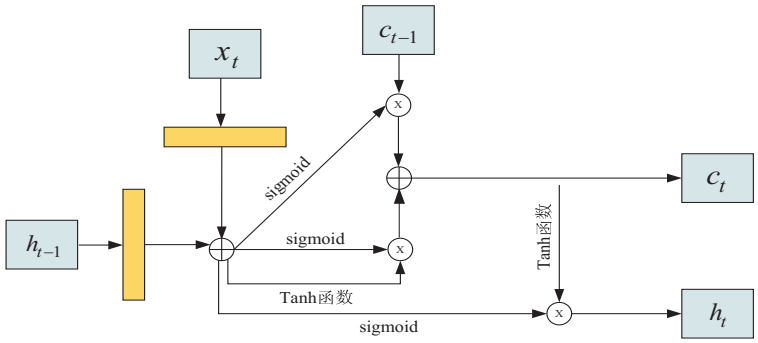

图 2　ConvLSTM 内部运行图

该方法利用 ConvLSTM 的优势通过叠加多个 ConvLSTM 层设计了基于 ConvLSTM 语义在每个时间步上产生的隐藏状态,可以在这些隐藏状态上应用一个卷积层来提取中级特征,然后将这些特征输入到后续的层中进行处理,进而增加底层语义信息的权重,如图 3 所示。

该方法在 VA-MVSNet 方法的 2D U-Net 网络的基础上,在低层次提取特征的过程中,加入了四级 ConvLSTM 网络。在 2D U-Net 网络中图像特征大小为 ($H \times W \times 4$) 时,加入三层的 ConvLSTM 网络,均采

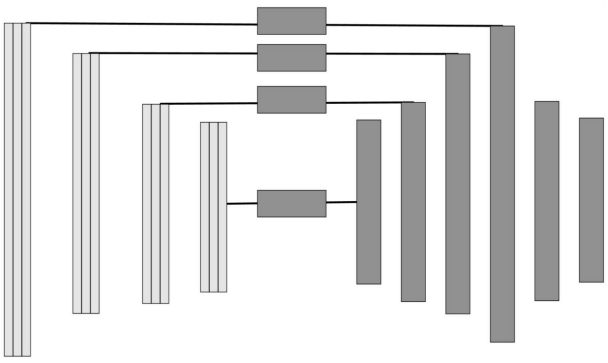

图 3　ConvLSTM 语义聚合网络图

用(3,1,1)的卷积核,并设置输出的通道数均为 8;在 2D U-Net 网络中图像特征大小为($\frac{1}{2}H \times \frac{1}{2}W \times 8$)时,加入三层的 ConvLSTM 网络,均采用(3,1,1)的卷积核,并设置输出的通道数均为 16;在 2D U-Net 网络中图像特征大小为($\frac{1}{4}H \times \frac{1}{4}W \times 16$)时,使用三层的 ConvLSTM 网络,均采用(3,1,1)的卷积核,并设置输出的通道数均为 32;在 2D U-Net 网络中特征图大小为($\frac{1}{8}H \times \frac{1}{8}W \times 32$)时,使用三层的 ConvLSTM 网络,均采用(3,1,1)的卷积核,并设置输出的通道数均为 64。基于 ConvLSTM 的语义聚合网络在聚合高层语义特征与低层语义特征的同时,也聚合不同尺度产生的语义特征的信息,以丰富高层特征图中不同的语义信息,构建出更准确的匹配成本体,提升三维重建的整体效果。

1.2 可见性网络结构

图像内容随着不同视角的变化而有所不同。不同视角可能会产生遮挡,这可能会大大减少图像中的有用信息,进而影响良好的三维重建效果。此外,颜色对光照变化的敏感性也增加了三维重建的难度。为了克服这些问题,该文提出一种使用灰度图像以及卷积神经网络来表达可见性的方法,这有助于增强可见区域的判别特征,并抑制不可见区域的影响,从而提高三维重建的效果,如图 4 所示。

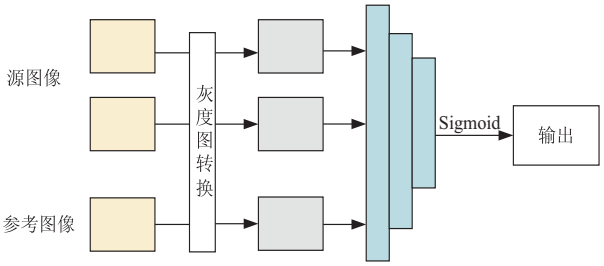

图 4 可见性网络结构

卷积层的输出特征图的通道数对模型的性能和效果有着重要的影响。首先,较大的输出通道数可以增加卷积层的自由度,从而使得模型能够学习到更多、更复杂的特征,这样可以提高模型的判别性和泛化能力,使得模型更加适用于复杂的任务。其次,卷积核的输出通道数也可以控制特征图的维度和质量。具体来说,较大的输出通道数可以使得特征图的维度更高,包含更多的特征信息,从而提高模型的表达能力,同时,增加输出通道数还可以降低特征图中的噪声和冗余信息,使得特征图的质量更高,进而提高模型的准确率和鲁棒性,与此同时,使用权值共享的方式,即对输入数据中的不同位置使用相同的卷积核,从而可以使模型具有平移不变性,能够识别输入数据中的不同位置的相同特征。

图 5 展示了一个 3 通道的数据经过一个大小为 3 的卷积核作用的过程。

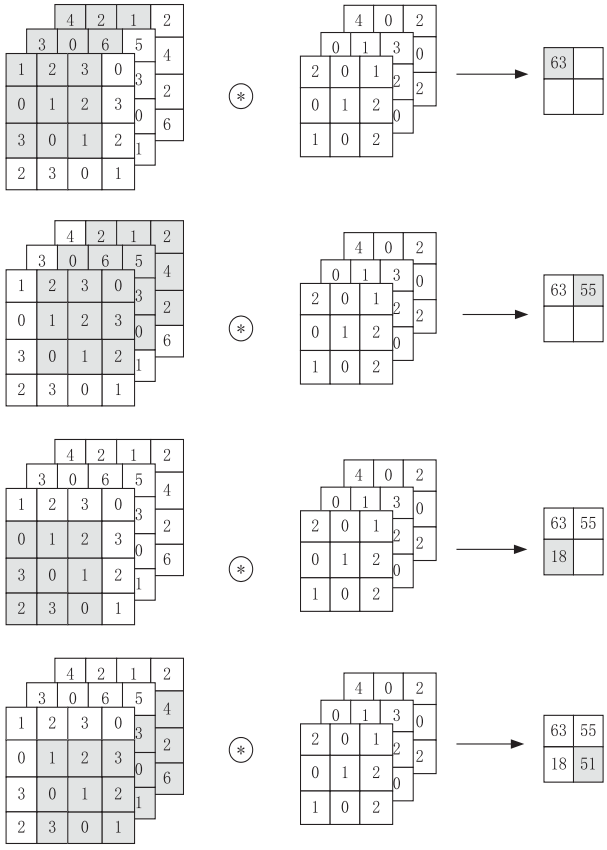

图 5 卷积计算过程

因此,该方法使用 PIL 库内的函数对 RGB 图像进行灰度的转换之后,使用了 5 层卷积网络进行提取,通道数逐渐叠加,卷积核使用(3,1,1)的卷积核,然后经过激活函数形成可见性度量值并与 ConvLSTM 语义聚合网络提取的低级特征相乘,继续进行网络中高级特征的提取。

1.3 纹理特征提取

在特征抽取过程中,该文采用经典的局部二值化模型(Local Binary Pattern,LBP)抽取图像中的纹理特征,以改善纹理区域的重构结果,体现图像中每个像素与局部区域内像素之间的关系。LBP 算法可以使用不同的邻域半径和邻域像素数量来生成不同尺度的纹理特征。一般来说,LBP 算法会计算每个像素的 LBP 值,并将其与周围像素的 LBP 值进行比较,从而得到该像素的纹理特征向量。在实际应用中,可以使用 LBP 算法提取图像的纹理特征,并将其作为分类器的输入,从而实现图像分类和识别等任务。

该方法使用圆形范围半径为 1 的 LBP 算子对输入图像进行纹理特征的提取,并设置领域像素点的数量为 8,LBP 算子如图 6 所示。

该方法将彩色的三维图像($H \times W \times 3$)转成灰度

图像,并对灰度图像进行 LBP 算子运算,得到对应的 LBP 特征图,部分 LBP 特征图如图 7 所示,然后将 LBP 特征图与输入图像在通道方向上进行拼接,作为 ConvLSTM 语义聚合网络的输入($H \times W \times 4$)。

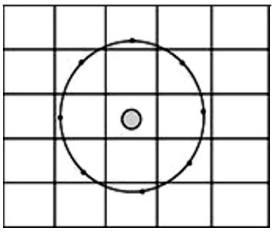

图 6　LBP(8,1)算子

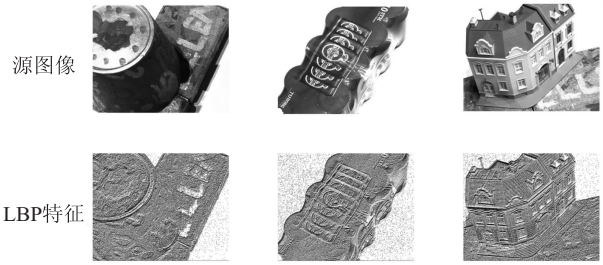

图 7　LBP 特征图示例

2　实验结果与分析

该方法在 DTU[20] 数据集上训练,按照一般的做法,在训练时,将所输入图像的尺寸设定为 $W \times H = 640 \times 512$,并且设置输入图像的数量为 5。深度假设的范围是 425 mm 至 935 mm,从这个范围内采样假定的深度值,深度平面的数量是 192,当前的网络是在 PyTorch 上实现的,并且利用 Adam 进行 16 个 epoch 的端到端的训练,初始的学习速率是 0.001,每个 epoch 衰减是 0.9。批量长度设为 4。对于测试,该方法在原始输入图像分辨率为 1 600×1 184 的 DTU 数据集上评估,使用 7 个图像视图,$D = 192$ 进行深度平面扫描,并使用 MVSNet 方法中使用的训练损失,如式 1 所示。对于 Tanks and Temples 数据集[21],输入图像的分辨率设置为 1 920×1 056。

$$\text{Loss} = \sum_{p \in \Omega} \| D(p) - D_{\text{GT}}(p) \|_1 \tag{1}$$

其中,$D(p)$ 表示预测深度图,$D_{\text{GT}}(p)$ 表示真值深度图。

2.1　实验环境

该方法的研究与实验均在 Linux 环境下进行,操作系统是 Ubuntu 20.04,使用 2 块 NVIDIA RTX TITAN 的 GPU 卡进行实验,具体信息如表 1 所示。

表 1　实验环境信息

类别	类型或版本
操作系统	Ubuntu 20.04
CPU	Intel Core i9-9900k
GPU	NVIDIA RTX TITAN
GPU0 内存	24 212 M
GPU1 内存	24 220 M

2.2　数据集

DTU 数据集是一个大规模的 MVS 数据集,包括 49 或 64 个视角,并且提供了真实点云和深度图,可以将通过实验得到的预测值与其进行比较。每个视角由一幅图像和对应的相机参数组成,图像的原始尺寸为 1 600×1 200 像素,包含 8 位 RGB 颜色。124 个场景中包含 80 个测试场景。其中,59 个场景包含 49 个摄像机位置,21 个场景包含 64 个摄像机位置。

为了提升 MVS 方法在大规模场景重建中的泛化能力,Arno Knapitsch 等人提供了 Tanks and Temples 数据集。该数据集基准更大,更多样化,包括完整的大型室外结构以及复杂的室内环境。这个数据集包括两组不同的场景:中等和高级。中等组包括各种建筑,如雕像、大型车辆和房屋,镜头路径从外向内。该方法选择中等组场景进行泛化实验。

2.3　评价指标

点云的准确度、完整度、F-Score 等是评价三维重构算法性能的重要指标。在不同的算法或数据集下,点云准确度和完整度的度量并不完全一致,F-Score 的计算方法也不完全一样。

(1)准确度(Accuracy)。

假设 G 为真实点云集,R 为预测点云集,对于一个预测点云 $r \in R$,当满足式 2 的条件时,对预测点云中的任何点,将被视为与重构点云中的点相匹配的很好:

$$\| G - \text{argmin} \| g - R \|_2 \|_2 \leqslant \lambda \tag{2}$$

其中,λ 是与场景有关的被指派给该数据组的参数。将较大的场景设定为较大的数值。在 DTU 数据集,其准确度并非以百分数来度量,而以绝对平均数来度量。平均绝对值愈小,准确度愈小时,则表示该方法的精度愈高,如式 3 所示。

$$\text{Acc} = \frac{1}{|R|} \sum_{r \in R} \min_{g \in G} \| r - g \|_2 \tag{3}$$

(2)完整度(Completeness)。

完整度是指真实点云在重建后的点云中可以匹配的像素点百分比的度量。同准确度一样,对于一个真实点云 $r \in R$,就会认为对于重建后的点云,该点具有良好的匹配性,如式 4 所示。

$$\| R - \text{argmin} \| r - G \|_2 \|_2 \leqslant \lambda \tag{4}$$

其中,λ 是由数据集分配的与场景相关的参数。对于比较大的场景设置为较大的值。距离的定义与倒角距离定义相同,在 DTU 数据集中,完整度是指真实点云

与预测点云之间的相对距离,平均绝对距离越小,认为其准确度越优秀,如式 5 所示。

$$\text{Comp} = \frac{1}{|G|}\sum_{g\in G}\min_{r\in R}\|g-r\|_2 \tag{5}$$

(3)F-score。

F-score 是衡量多视立体视觉方法性能的重要指标,是对准确度与完整度的调和平均值。在 DTU 数据集中,F-score 设置为完整度与准确度两者的平均数,称作平均值;在 Tanks and Temples 数据集中,F-score 设置为真实点云 G 与预测点云 R 之间的差异百分比与预测点云 R 与实际点云 G 之间的差异百分比的调和平均值,值越高,表明该方法的性能越优秀。

对于一个预测点云 $r \in R$,它到真实点云的距离定义如式 6 所示,则设预测点云 R 与实际点云 G 之间的差异小于某一特定阈值 d 的比例为 $P(d)$,如式 7 所示。

$$e_r = \min_{g\in G}\|r-g\| \tag{6}$$

$$P(d) = \frac{100}{|R|}\sum_{r\in R}[e_r < d] \tag{7}$$

同样,对于一个真实点云 $g \in G$,它到预测点云的距离定义如式 8 所示,则真实点云 G 与预测点云之间差异程度的指标可以用式 9 所示。

$$e_g = \min_{r\in R}\|g-r\| \tag{8}$$

$$R(d) = \frac{100}{|G|}\sum_{g\in G}[e_g < d] \tag{9}$$

则 F-score 的计算公式如式 10 所示。

$$f(d) = \frac{2P(d)R(d)}{P(d)+R(d)} \tag{10}$$

2.4 结果分析

该方法在 DTU 数据集的定量结果如表 2 所示,定性结果如图 8、图 9 所示。

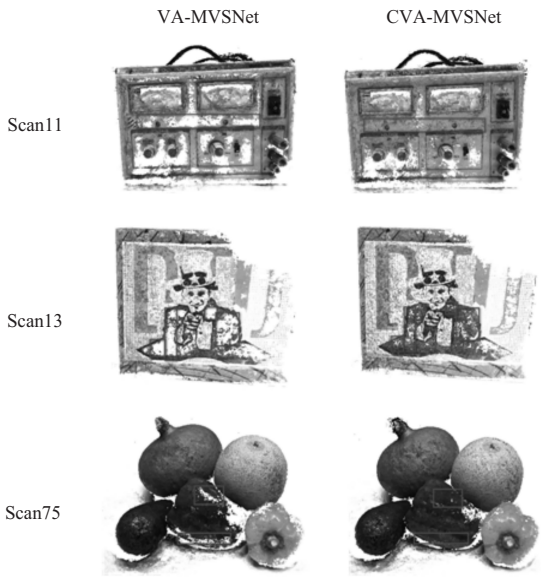

图 8 点云结果对比

图 9 局限性展示

由表 2 可知,该方法在完整度上面表现不错,比 VA-MVSNet 方法提高了 0.038,同时实现了具有竞争力的整体性能,但通过与一些经典的方法以及先进方法的比较,还有所差距,与基础网络性能以及所用方法有关,还需进一步改善与提高。

表 2 DTU 数据集指标结果对比

方法	准确度	完整度	平均值
Gipuma[22]	0.283	0.873	0.578
MVSNet[23]	0.396	0.527	0.462
Point-MVSNet[24]	0.361	0.421	0.391
CasMVSNet[14]	0.325	0.385	0.355
]PVA-MVSNet[12]	0.379	0.336	0.357
VA-MVSNet[12]	0.378	0.356	0.369
RC-MVSNet[25]	0.373	0.354	0.363
文献[26]	0.355	0.301	0.328
文献[27]	0.372	0.302	0.337
文中方法	0.376	0.318	0.347

除了在重建的三维点云模型上的定量比较,图 8、图 9 展示了文中方法与其他先进方法的定性比较。

得益于集成了高层语义特征信息和低层语义特征信息以及多尺度特征信息的 ConvLSTM 语义聚合网络,该方法能够为弱纹理区域估计更完整和更精细的空间表面。由图 8 可看出,该方法在精细结构处的重建效果更好,验证了该方法的有效性,但是由图 9 也可以看出,该方法也具有一定的局限性,重建的精准度不足,重建场景模糊。

与此同时,该文使用在 DTU 数据集上训练好的参数模型在 Tanks and Temples 数据集的中级数据集上进行泛化能力的验证。由表 3 的数据对比可以看出,该方法在一些场景如 Family, Francis, Light-house, Panther 上的重建结果较好,具有一定的泛化能力,在其他场景泛化能力不足,可以分析得到其他方法如 PVA-MVSNet 方法结合多尺度信息进行三维重建是有可取之处的。

表 3　Tanks and Temples 数据集指标结果对比

方法	Mean	Family	Francis	Horse	Light-house	M60	Panther	Play-ground	Train
MVSNet[23]	43.48	55.99	28.55	25.07	50.79	53.96	50.86	47.90	34.69
R-MVSNet[28]	48.40	69.96	46.65	32.59	42.95	51.88	48.80	52.00	42.38
M3VSNet[29]	37.67	47.74	24.38	18.74	44.42	43.45	44.95	47.39	30.31
Fast-MVSNet[17]	47.39	65.18	39.59	34.98	47.81	49.16	46.20	53.27	42.91
SelfsupCVP-MVSNet[30]	46.71	64.95	38.79	24.98	49.73	52.57	51.53	50.66	40.45
PVA-MVSNet	54.46	69.36	46.80	46.01	55.74	57.23	54.75	56.70	49.06
文中方法	53.99	70.17	53.05	39.69	58.34	55.65	56.07	50.69	48.27

除此之外,在本节中,进行消融实验来定量分析该方法中关键组件的优势和有效性。对于以下所有研究,均在 DTU 数据集上进行实验和评估,并使用准确性和完整性衡量重建质量。

本节比较了 ConvLSTM 语义聚合网络、可见性网络以及纹理特征提取的有效性。定量结果如表 4 所示。由表 4 可知,ConvLSTM 语义聚合网络、可见性网络都可以显著提高重建标准的结果,并且互补,以取得最佳效果。

表 4　不同组件的消融实验在 DTU 数据集的结果

ConvLSTM 语义聚合网络	可见度网络	准确度	完整度	平均值
√	×	0.379	0.327	0.353
×	√	0.375	0.341	0.358
×	×	0.378	0.359	0.369
√	√	0.373	0.325	0.349

3　结束语

该文提出了一种具有语义增强功能的多视立体视觉方法。该方法通过使用 ConvLSTM 语义聚合网络增强高层特征图中的低层语义信息,增强了细节语义特征以及多尺度语义信息,利于弱纹理区域的重建,同时通过纹理特征提取模块提取图像的纹理特征,加深了纹理语义信息在高层特征图的影响,提升了三维重建的整体效果,有效地提高了低纹理表面的性能,这两个模块是轻量级、有效和互补的。除此之外,提出了一种可见性网络,一定程度上排除了光照变化的影响,突出了可见区域的特征,加深了可见区域在特征图中的影响,有助于提高三维重建效果。在 DTU 数据集和 Tanks and Temples 数据集上的实验,证实了该方法的合理性和有效性。

后续的工作将考虑不同层次的语义特征与匹配成本体之间的关系,进一步研究不同语义特征对于匹配代价的影响,以更准确地进行代价匹配。

参考文献:

[1]　马银平,彭　如. 基于变换矩阵的三维重建算法研究[J]. 计算机技术与发展,2011,21(8):78-81.

[2]　于　勇,张　晖,林茂松. 基于双目立体视觉三维重建系统的研究与设计[J]. 计算机技术与发展,2009,19(6):127-130.

[3]　SEITZ S M, CURLESS B, DIEBEL J, et al. A comparison and evaluation of multi-view stereo reconstruction algorithms [C]//2006 IEEE computer society conference on computer vision and pattern recognition (CVPR´06). Ames: IEEE, 2006:519-528.

[4]　FURUKAWA Y, PONCE J. Accurate, dense, and robust multiview stereopsis[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2009, 32(8):1362-1376.

[5]　LHUILLIER M, QUAN L. A quasi-dense approach to surface reconstruction from uncalibrated images [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2005, 27(3):418-433.

[6]　ESTEBAN C H, SCHMITT F J. Silhouette and stereo fusion for 3d object modeling[J]. Computer Vision and Image Understanding, 2004, 96(3):367-392.

[7]　SEITZ S M, DYER C R. Photorealistic scene reconstruction by voxel coloring[J]. International Journal of Computer Vision, 1999, 35(2):151-173.

[8]　KUTULAKOS K N, SEITZ S M. A theory of shape by space carving[J]. International Journal of Computer Vision, 2000, 38(3):199-218.

[9]　TOLA E, STRECHA C, FUA P. Efficient large-scale multi-view stereo for ultra high-resolution image sets[J]. Machine Vision and Applications, 2012, 23(5):903-920.

[10] GOESELE M, CURLESS B, SEITZ S M. Multi-view stereo revisited[C]//IEEE computer society conference on computer vision and pattern recognition (CVPR'06). Jeju: IEEE, 2006: 2402-2409.

[11] WU J. Introduction to convolutional neural networks[R]. Nanjing: Nanjing University, 2017.

[12] YI H, WEI Z, DING M, et al. Pyramid multi-view stereo net with self-adaptive view aggregation[C]//European conference on computer vision. Glasgow: Springer, 2020: 766-782.

[13] DING Y, YUAN W, ZHU Q, et al. Transmvsnet: global context-aware multi-view stereo network with transformers [C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR). New Orleans: IEEE, 2022: 8585-8594.

[14] GU X D, FAN Z W, ZHU S Y, et al. Cascade cost volume for high-resolution multi-view stereo and stereo matching [C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR). Seattle: IEEE, 2020: 2495-2504.

[15] KHOT T, AGRAWAL S, TULSIANI S, et al. Learning unsupervised multi-view stereopsis via robust photometric consistency[J]. arXiv: 1905.02706, 2019.

[16] LUO K, GUAN T, JU L, et al. P-mvsnet: l-earning patch-wise matching confidence aggregation for multi-view stereo [C]//Proceedings of the IEEE/CVF international conference on computer vision (ICCV). Seoul: IEEE, 2019: 10452-10461.

[17] YU Z, GAO S. Fast-mvsnet: sparse-to-dense multi-view stereo with learned propagation and gauss-newton refinement [C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. Seattle: IEEE, 2020: 1949-1958.

[18] LUO K, GUAN T, JU L, et al. Attention-aware multi-view stereo[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. Seattle: IEEE, 2020: 1590-1599.

[19] 豆浩冉. 基于 ConvLSTM 深度时空网络的短临降水预报研究[D]. 南京: 南京邮电大学, 2022.

[20] AANS H, JENSEN R R. Large-scale data for multiple-view stereopsis[J]. International Journal of Computer Vision, 2016, 120(2): 153-168.

[21] KNAPITSCH A, PARK J, ZHOU Q Y, et al. Tanks and temples: benchmarking large-scale scene reconstruction[J]. ACM Transactions on Graphics, 2017, 36(4): 78.

[22] GALLIANI S, LASINGER K, SCHINDLER K. Gipuma: massively parallel multi-view stereo reconstruction[J]. Publikationen der Deutschen Gesellschaft für Photogrammetrie, Fernerkundung und Geoinformation e. V, 2016, 25(361-369): 2.

[23] YAO Y, LUO Z X, LI S W, et al. MVSNet: depth inference for unstructured multi-view stereo[C]//European conference on computer vision. Munich: Springer, 2018: 767-783.

[24] CHEN R, HAN S F, XU J, et al. Point-based multi-view stereo network[C]//IEEE/CVF international conference on computer vision (ICCV). Seoul: IEEE, 2019: 1538-1547.

[25] CHANG D, BOŽIČ A, ZHANG T, et al. RC-MVSNet: unsupervised multi-view stereo with neural rendering[C]//European conference on computer vision: Tel Aviv: Springer, 2022: 665-680.

[26] 王若蓝, 赵 融, 韩 燮. 融合梯度和高斯过程回归的多视图重建方法[J]. 微电子学与计算机, 2023, 40(3): 37-45.

[27] 常益凡. 基于深度图估计的多视图三维重建[D]. 成都: 电子科技大学, 2022.

[28] YAO Y, LUO Z, LI S, et al. Recurrent mvsnet for high-resolution multi-view stereo depth inference[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. Long Beach: IEEE, 2019: 5525-5534.

[29] PASZKE A, GROSS S, CHINTALA S, et al. Autom-atic differentiation in pytorch[C]//Proceedings of the 27th international conference on neural information processing systems (NIPS). Long Beach: Curran Associates, 2017: 1-4.

[30] XU H B, ZHOU Z P, QIAO Y U, et al. Self-supervised multi-view stereo via effective co-segmentation and data-augmentation[C]//AAAI conference on artificial intelligence. Palo Alto: AAAI, 2021: 3030-3038.