文章编号:1671-6833(2024)04-0011-08

基于轻量化深度卷积循环网络的 MVS 方法

佘 维[1,2,3], 孔祥基[1,3], 郭淑明[2,4], 田 钊[1,3], 李英豪[1,2,3]

(1. 郑州大学 网络空间安全学院,河南 郑州 450002;2. 嵩山实验室,河南 郑州 450046;3. 郑州市区块链与数据智能重点实验室,河南 郑州 450002;4. 国家数字交换系统工程技术研究中心,河南 郑州 450002)

摘 要: 针对基于深度学习的 MVS 方法存在网络参数量大、显存占用较高的问题,提出一种基于轻量化深度卷积循环网络的 MVS 方法。首先,采用轻量化多尺度特征提取网络提取图像的高层语义特征图,构建稀疏代价体减小计算体积;其次,使用卷积循环网络对代价体进行正则化,一次平面扫描完成正则化过程,减少显存占用;最后,通过深度图扩展模块扩展稀疏深度图为稠密深度图,并结合优化算法保证重建精度。在 DTU 数据集上与最近的方法进行对比,包括传统 MVS 方法 Camp、Furu、Tola、Gipuma,基于深度学习的 MVS 方法 SurfaceNet、PU-Net、MVSNet、R-MVSNet、Point-MVSNet、Fast-MVSNet、GBI-Net、TransMVSNet。实验结果表明:所提方法在精度上与其他方法保持较小差距的前提下,能够将预测时显存开销降低至 3.1 GB。

关键词: 轻量化; 深度卷积循环网络; MVS 方法; 正则化; DTU 数据集

中图分类号: TP39;TP751.1 **文献标志码**: A **doi**: 10.13705/j.issn.1671-6833.2024.04.003

立体视觉是计算机视觉领域的基本问题之一,其目标是从一个或多个图像中恢复拍摄场景的三维结构,恢复图像中像素点对应三维场景的相对深度,在三维重建、增强现实、自动驾驶、机器人技术等领域应用广泛。

单视图立体视觉从单个图像中推断 3D 形状的关键几何和结构信息。Yan 等[1]从学习代理的角度研究了单视图三维对象重建,提出一种具有由投影变换定义的新投影损失的编码器-解码器网络。Sun 等[2]提出一种基于深度学习单视图三维重建和姿态估计的模型。从单视图恢复三维结构是一个不适定问题,由于图像与三维模型间存在的表示模式差异,通常存在物体自遮挡、低光照等情况。

多视图立体视觉(MVS)[3]从一组校准过的重叠图像中恢复场景的密集三维结构,能够有效减少物体遮挡、光照等因素带来的影响。根据输出表示形式,可分为直接点云重建[4],体积重建[5]和深度图重建。根据最近的 MVS 基准,目前最好的 MVS 算法都是基于深度图的方法。虽然深度图可以被视为点云表示的一种特殊情况(如像素级点云),但它将重构简化为逐视图深度的估计问题。此外,可以很容易地将深度图融合到点云或体积重建。

虽然传统方法在 MVS 上取得了优异的性能,但它们也有一些共同的局限性。例如,场景的低纹理、镜面和反射区域使得密集匹配难以处理,从而导致重建结果不完整。最近的研究表明,使用 Deep CNN 可以进一步提高 MVS 的性能。Huang 等[6]和 Ji 等[7]使用多视图图像构建代价体,并使用 CNN 学习该代价体的正则化过程。Yao 等[8]提出了一种 MVS 深度学习网络 MVSNet,用于从多视图图像中推断深度图。该网络架构是一种端到端的 MVS 架构,基于卷积神经网络特征构建成本量,并使用 3D 卷积神经网络学习成本量正则化。该网络不仅性能显著优于以往的方法,而且速度效率大大提高。Yao 等[9]和杜弘志等[10]提出了基于门控循环单元(GRU)的方法,沿深度方向顺序正则化代价体。Chen 等[11]

收稿日期:2023-11-10;**修订日期**:2023-12-25

基金项目:嵩山实验室预研项目(YYYY022022003);国家自然科学基金资助项目(62206252);河南省科技攻关项目(212102310039)

作者简介:佘维(1977—),男,湖南常德人,郑州大学教授,博士,博士生导师,主要从事复杂系统建模与仿真、机器学习、区块链、数据智能的研究,E-mail:wshe@zzu.edu.cn。

通信作者:李英豪(1987—),男,河南郑州人,郑州大学副教授,博士,主要从事数字图像处理、模式识别、机器学习的研究,E-mail:yinghaoli@zzu.edu.cn。

引用本文:佘维,孔祥基,郭淑明,等. 基于轻量化深度卷积循环网络的 MVS 方法[J]. 郑州大学学报(工学版),2024,45(4):11-18. (SHE W, KONG X J, GUO S M, et al. MVS method based on lightweight deep convolutional recurrent network[J]. Journal of Zhengzhou University (Engineering Science),2024,45(4):11-18.)

提出了 Point-MVSNet 网络从粗到细的方式预测深度图,并处理点云以估计每个点的 3D 流,有效提高了 MVS 过程的计算效率。Yu 等[12]提出了 Fast-MVSNet 网络,构造高分辨率稀疏深度图,并使用高斯牛顿层对结果进行优化,其计算效率和内存占用均优于以上模型。此外,最近的优化算法[13-14]也表现出了巨大的潜力。

然而,这些基于深度学习的方法大多使用多尺度 3D 卷积网络来预测深度图,因此,随着三维空间体积的增长,内存需求立方级增长,基于深度学习的方法只能在高计算资源设备上进行运算。本文针对基于深度学习的 MVS 网络参数量大、显存占用较高的问题,提出了一种由稀疏到稠密、由粗糙到精细的轻量化卷积循环网络 LSD-MVSNet。采用轻量化多尺度特征提取网络提取图像的高层语义信息,构建稀疏代价体减小计算体积;正则化过程使用卷积循环网络逐深度平面处理以减少显存占用,使得本文方法参数量大大减少,运行时的显存开销显著降低。对于特征图的通道聚合过程,充分利用了二维卷积特征,保证了重建质量。本文工作主要包含:设计了一种轻量化的多尺度特征提取网络;提出了一种充分利用所有特征的稀疏代价体构造方式;设计了一种双向卷积门控循环单元正则化网络。

1 相关知识

1.1 多视图立体视觉

多视图立体视觉重建的方法中,基于点云的方法直接在三维点上操作,通常依赖传播策略来逐步强化重建。由于点云的传播是按顺序进行的,这些方法难以完全并行化,处理时间长。基于体积的方法将三维空间划分为规则网格,然后估计每个体素是否在表面附近。这种表示的缺点是空间离散化错误和高内存消耗。相比之下,深度图是最灵活的表示法,它将复杂的 MVS 问题解耦为相对较小的逐视图深度图估计问题,每次只关注一个参考图像和几个源图像。

1.2 基于深度图的多视图立体视觉

深度图表示为每个观测视图的 2.5D 形式用来显示 3D 几何图形,利用深度融合技术可以将深度图融合为 3D 点云恢复三维模型。传统的方法在理想的朗伯表面场景下表现出良好的效果,但它们存在一些共同的局限性,例如场景的低纹理、镜面和反射区域使密集匹配难以处理,从而导致不完整的重建。

近年来卷积神经网络研究的成功,引发了人们对改进多视角立体视觉的兴趣如 DeepMVS。基于学习的方法引入诸如镜面先验和反射先验等全局语义信息,实现更健壮的匹配。具体来说,基于平面扫描的方法经过特征提取、代价体构建、代价体正则化、深度图回归得到预测深度图,深度融合后得到 3D 点云。

1.3 卷积门控循环单元

门控循环单元是循环神经网络的一种,是长短期记忆的简化版本,能够更好地处理长时序列中每个阶段之间的信息流动。

多视图立体视觉问题中,物体在深度方向是连续的,对于物体上的任意一点,深度方向上距离较近的相邻点对其有较大贡献,深度方向上距离较远的相邻点对其有较小贡献,因此门控循环单元可以应用到本文的正则化过程中。本文模型中的卷积门控循环单元结构如图 1 所示。

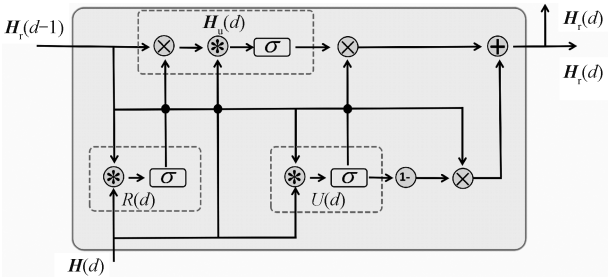

图 1 卷积门控循环单元结构

Figure 1 Convolutional gated recurrent unit architecure

图 1 中,$R(d)$ 为重置门;$U(d)$ 为更新门;σ 为逐元素运算的 sigmoid 函数;$\boldsymbol{H}_r(d-1)$ 为深度为 $d-1$ 时的隐藏状态;$\boldsymbol{H}_u(d)$ 为深度为 d 时的候选隐藏状态;$\boldsymbol{H}(d)$ 为深度为 d 时的输入状态;$\boldsymbol{H}_r(d)$ 为深度为 d 时隐藏状态;$\otimes$ 为按元素相乘;$*$ 为卷积运算符;$\oplus$ 为按元素相加运算符。图 1 省略了偏置值的体现。

卷积门控循环单元主要包括重置门 $R(d)$ 与更新门 $U(d)$:

$$R(d)=\sigma_g(\boldsymbol{W}_r*[\boldsymbol{H}(d),\boldsymbol{H}_r(d-1)]+\boldsymbol{b}_r);\quad(1)$$

$$U(d)=\sigma_g(\boldsymbol{W}_u*[\boldsymbol{H}(d),\boldsymbol{H}_r(d-1)]+\boldsymbol{b}_u)。\quad(2)$$

式中:$\boldsymbol{W}$ 为学习的权重参数;$\boldsymbol{b}$ 为学习的偏置值;[,] 为连接 2 个向量运算。

输入经过重置门后与前一平面 $d-1$ 的隐藏状态计算出当前平面 d 的候选隐藏状态 $\boldsymbol{H}_u(d)$:

$$\boldsymbol{H}_u(d)=\sigma_c(\boldsymbol{W}_c*[\boldsymbol{H}(d),R(d)\otimes\boldsymbol{H}_r(d-1)]+\boldsymbol{b}_c)。\quad(3)$$

由候选隐藏状态 $\boldsymbol{H}_u(d)$ 与更新门输出可构造当前平面 d 的隐藏状态 $\boldsymbol{H}_r(d)$:

$$\boldsymbol{H}_r(d)=(1-U(d))\otimes\boldsymbol{H}_r(d-1)+U(d)\otimes\boldsymbol{H}_u(d)。\quad(4)$$

卷积门控循环单元相较于经典门控循环单元的区别在于引入了卷积层,二维卷积计算不仅在空间上优化深度图,而且沿着深度方向聚合空间的全局信息。

2　LSD-MVSNet 模型

2.1　模型架构

本文提出的模型整体架构如图 2 所示。首先,参考图像及 $N-1$ 个源图像经过 1 个 6 层二维卷积网络输出特征图;其次,特征图经过可微单应性变换构造稀疏代价体;再次,代价体正则化后经过 softmax 输出稀疏深度图;最后,参考图像经过深度图扩展网络扩展稀疏深度图为稠密深度图。另外,本文模型沿用了 Fast-MVSNet[12] 的优化层,参考图像经过优化层与稠密深度图融合后输出优化后的深度图。

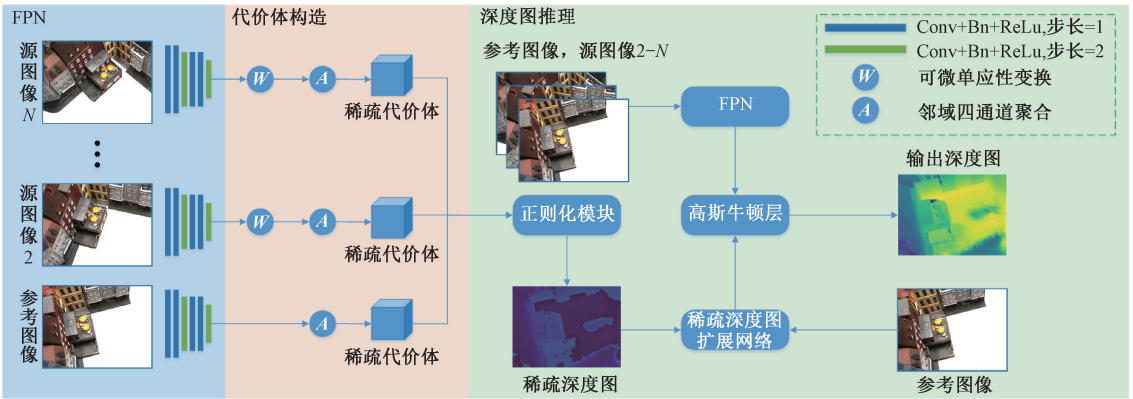

图 2　LSD-MVSNet 模型架构图

Figure 2　LSD-MVSNet model architecture

2.2　特征提取模块

特征提取主要通过多尺度二维卷积网络完成。给定一个参考图像 $\boldsymbol{I}_1$ 和一组相邻的源图像 $\{\boldsymbol{I}_i\}_{i=2}^{N}$,对输入图像进行密集匹配。多尺度设计借鉴了 MVSNet[8] 的 3 个尺度。本文方法设计了一个轻量化的多尺度二维卷积网络,第 3 层和第 6 层的步幅设置为 2 层,将特征金字塔划分为 3 个尺度。在前 2 个尺度内,应用了 2 个卷积层来提取更高级别的图像表示。每个卷积层之后都有 1 个批归一化层和 1 个用于校正的 ReLu 函数,权重参数在特征金字塔之间共享。

特征提取网络的输出为 N 个 8 通道的特征图 $\{\boldsymbol{F}_i\}_{i=1}^{N}$,每个维度为输入图像的 1/4。虽然特征提取后图像尺度缩小,但每个像素的原始相邻信息已经被编码到 8 通道特征图中,避免了在密集匹配过程中丢失全局特征。

2.3　代价体构建模块

基于 MVSNet 的网络大多沿用了平面扫描算法[15]。平面扫描是一种多视图立体算法,解决了二维到三维的立体问题。参考图像 $\boldsymbol{I}_1$ 与源图像 $\{\boldsymbol{I}_i\}_{i=2}^{N}$ 经过特征提取网络输出 N 个 8 通道特征图 $\{\boldsymbol{F}_i\}_{i=1}^{N}$。给定深度假设 $\{D_d\}_{d=1}^{D}$,匹配特征图 $\{\boldsymbol{F}_i\}_{i=2}^{N}$ 中像素 $\boldsymbol{p}$,在参考特征图 $\boldsymbol{F}_1$ 中的相对应像素 $\boldsymbol{p}'$ 在第 d 个深度假设平面经过可微单应性变换可以表示为

$$\boldsymbol{p}'_{i,d}=\boldsymbol{K}_i\cdot(\boldsymbol{R}_{1,i}\cdot(\boldsymbol{K}_1^{-1}\cdot\boldsymbol{p}\cdot D_d)+\boldsymbol{t}_{1,i})。\quad(5)$$

式中:$i\in[1,N]$;$\boldsymbol{K}_i$,$\boldsymbol{R}_{1,i}$,$\boldsymbol{t}_{1,i}$ 分别为特征图对应相机的内参矩阵、参考特征图到匹配特征图的旋转矩阵和位移向量。

将所有特征图变换到参考特征图的平面后组成特征体 $\{\boldsymbol{V}_i\}_{i=1}^{N}$,特征体大小为

$$V=\frac{W}{4}\times\frac{H}{4}\times D\times F。\quad(6)$$

式中:W、H、D、F 分别为输入图像的宽度、高度、深度采样数和特征图的通道数。

将多个特征体聚合为一个代价体 $\boldsymbol{C}$:

$$\boldsymbol{C}=M(\boldsymbol{V}_1,\boldsymbol{V}_2,\cdots,\boldsymbol{V}_N)=\frac{\sum_{i=1}^{N}(\boldsymbol{V}_i-\overline{\boldsymbol{V}}_i)^2}{N}。\quad(7)$$

式中:N 为代价体个数;$\boldsymbol{V}_i$ 为以第 i 个特征图作为参考特征图时的特征体;$\overline{\boldsymbol{V}}_i$ 为以第 i 个特征图作为参考特征图时的所有特征体均值。

特征图的维度直接影响了代价体的维度,最近的研究通过减小特征图维度来降低代价体的维度,从而加速后续的正则化过程。几种代价体构造方案如图 3 所示。

图 3(a)为 MVSNet 和 R-MVSNet 所使用的特征图。Point-MVSNet 为了降低正则化成本,缩小了特

征图的维度,如图3(b),其特征体大小 $V=\frac{W}{8}\times\frac{H}{8}\times D\times F$。Fast-MVSNet构造了一个稀疏代价体如图3(c)所示,经过正则化阶段后,使用1个轻量的扩展网络将获得的稀疏深度图扩展为原始分辨率深度图,降低了正则化成本,提高了处理速度,并保证了重建质量。

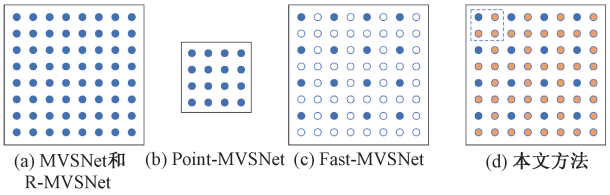

图3 不同代价体构造方案

Figure 3 Different cost volume construct schemes

本文设计了与Fast-MVSNet类似的稀疏代价体构造方式,与前者直接丢弃75%特征点不同的是,本文方法充分利用了所有特征,如图3(d)所示。本文特征提取阶段输出8通道特征图,对于特征图的蓝色像素点,将其相邻的3个橙色特征像素点扩充至蓝色像素点的通道维度,构造了与Fast-MVSNet相似的32通道稀疏代价体。同时本文放弃了在深度方向上均匀采样的方法,在逆深度方向上进行均匀采样。在处理大尺度重建时,可以确保特征聚集在一个相对集中的区域,因此本文网络具有更好的处理大尺度数据的能力。

2.4 正则化模块

对代价体进行全局正则化的一种方法是沿深度方向进行顺序处理。MVSNet采用三维卷积正则化提取整个空间的成本信息,这种方式精度较高,但是运行时显存开销达到模型分辨率的立方级别。大多数改进MVSNet网络均延续了三维卷积正则化过程。本文使用卷积门控循环单元来进行正则化。如图4所示,红色体素为当前感兴趣区域,蓝色体素为代价体正则化过程接受野,图4(b)中绿色平面为当前处理的深度平面,卷积门控循环单元正则化时考虑前 d 个平面。三维卷积正则化显存占用为 $H\times W\times D$,而卷积门控循环单元正则化显存占用为 $H\times W\times 1$。

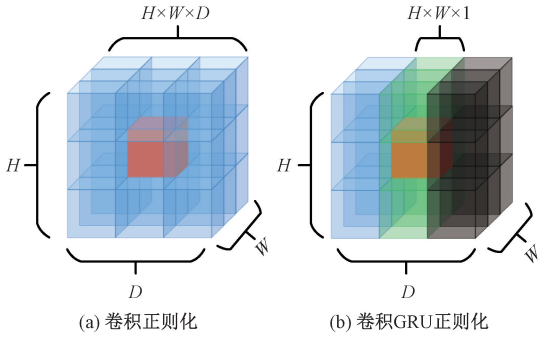

图4 不同正则化方案

Figure 4 Different regularization schemes

本文方法在深度方向上按顺序对代价体进行正则化。如图5所示,代价体 $\boldsymbol{C}$ 可看作 D 个代价图在深度方向上的拼接。由于卷积门控循环单元沿深度方向单向处理代价图,对于代价体 $\boldsymbol{C}$ 的代价图 $\boldsymbol{C}(i)$,$i\in[1,D]$,正则化过程中仅能感受前 $i-1$ 个代价图。为了更好地感知全局信息,本文提出一种双向卷积门控循环单元正则化方法,该方法使用一个卷积门控循环单元网络,沿深度前、后2个方向对代价体进行正则化。具体来说,本文模型使用前向、后向2个正则化过程,分别从代价图 $\boldsymbol{C}(1)$ 到代价图 $\boldsymbol{C}(D)$、从代价图 $\boldsymbol{C}(D)$ 到代价图 $\boldsymbol{C}(1)$ 对代价体进行正则化,最后2个正则化后的代价体按元素求均值,得到最终正则化后的代价体,理论上可以将运行时内存开销降低至模型分辨率的平方级别。基本的门控循环单元模型由1个单层组成,为了进一步提高正则化能力,本文使用3层堆叠卷积门控循环单元,形成更深层次的网络。需要说明的是,相较于三维卷积正则化网络,本文方法在深度方向上迭代处理特征图,在减小显存占用的同时降低了处理速度。正则化过程将正则化代价图定义为 $\{\boldsymbol{C}'(i)\}_{i=1}^{D}$,对于第 d 个平面的顺序处理理想状态下,$\boldsymbol{C}(d)$ 应当只依赖于当前平面的代价图 $\boldsymbol{C}(d)$

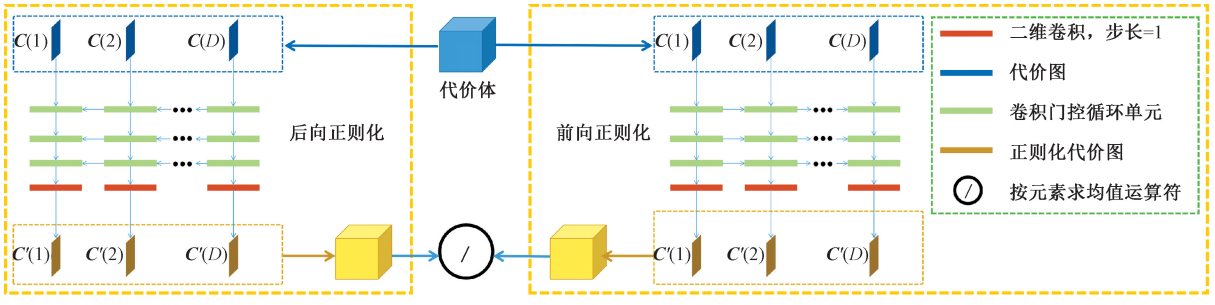

图5 卷积门控循环单元正则化

Figure 5 Convolutional GRU regularization

以及之前所有平面的代价图 $\{\boldsymbol{C}'(i)\}_{i=1}^{d-1}$。在本文的模型中,使用卷积门控循环单元在深度方向上聚合了这些上下文信息,对应了自然语言处理领域中的时间方向。

为了进一步增强正则化能力,本文采用 3 层叠加的卷积门控循环单元结构。首先应用二维卷积层将 32 通道的代价图 $\boldsymbol{C}(t)$ 缩小到 16 通道,作为第一个门控循环单元层的输入。每一层门控循环单元的输出作为下一层门控循环单元的输入,3 层的输出通道数分别设置为 16、4、1。正则化的代价图 $\{\boldsymbol{C}'(i)\}_{i=1}^{D}$ 最后经过一个 softmax 层,生成用于计算训练损失的概率体积。

2.5　深度图扩展模块

稀疏深度图 $\boldsymbol{D}$ 扩展得到稠密深度图 $\widehat{\boldsymbol{D}}$。本文选择与 Fast-MVSNet 相同的联合双侧上采样器,使用原始高分辨率图像的信息作为指导。稀疏深度图 $\boldsymbol{D}$ 首先利用最近邻扩展到稠密深度图。然后,一个轻量级卷积网络以参考图像作为输入,输出每个位置的 $k \times k$ 权重 $\boldsymbol{w}$。最后,计算出扩展的深度图 $\widehat{\boldsymbol{D}}$:

$$\widehat{\boldsymbol{D}}(\boldsymbol{p}) = \frac{1}{Z_p} \sum_{q \in N(\boldsymbol{p})} \boldsymbol{D}(\boldsymbol{q}) \cdot \boldsymbol{w}_{p,q}。 \tag{8}$$

式中:$N(\boldsymbol{p})$ 为像素点 $\boldsymbol{p}$ 附近的局部 $k \times k$ 邻域;Z_p 为一个归一化项;$\boldsymbol{w}_{p,q}$ 为轻量级卷积网络学习权重参数。

轻量级卷积网络简单地使用 2.1 节网络结构来提取图像特征,并附加 1 个 2 层 3×3 卷积网络来预测具有 $k \times k$ 通道的特征图。

2.6　损失函数

本文使用估计深度图与地面真实深度图之间平均绝对差作为模型的训练损失,初始深度图 $\widetilde{\boldsymbol{D}}$ 和优化深度图 $\widetilde{\boldsymbol{D}}'$ 都考虑在训练损失中:

$$L = \sum_{\boldsymbol{p} \in P_{\text{valid}}} \|\widetilde{\boldsymbol{D}}(\boldsymbol{p}) - \widehat{\boldsymbol{D}}(\boldsymbol{p})\| + \|\widetilde{\boldsymbol{D}}'(\boldsymbol{p}) - \widehat{\boldsymbol{D}}(\boldsymbol{p})\|。 \tag{9}$$

式中:P_{valid} 为有效真实深度的集合。

3　实验

3.1　数据集

本文实验使用公开的 MVS 数据集:DTU 数据集。DTU 数据集是被广泛应用在多视图三维重建中的经典大规模数据集,共包含 80 个场景,且场景多样性较大。每个场景都在 49 或 64 个精确的相机位置上拍摄,并有 7 种不同的照明条件。每张图片的分辨率为 1 600×1 200 像素,该数据集提供了由精确的结构光扫描仪以及高分辨率 RGB 图像获得的参考模型。

3.2　对比方法

本文方法与不同的 MVS 方法进行对比,包括传统 MVS 方法 Camp[16]、Furu[17]、Tola[18]、Gipuma[19],基于深度学习的 MVS 方法 PUNet[20]、MVSNet[8]、RMVSNet[9]、PointMVSNet[11]、SurfaceNet[7]、Fast-MVSNet[12]、GBI-Net[21]、TransMVSNet[22]。

3.3　评价标准

本文实验结果评价标准采用 MVS 领域中的 3 个评价指标:精确度误差、完整度误差和总体误差。计算 MVS 点云与参考点云中最近点距离,采样距离阈值为 0.2 mm、异常值拒绝阈值为 20 mm、射线扩展阈值为 10 mm。指标值单位均为 mm,指标值越小效果越好。

精确度为 MVS 点云到参考点云的距离差,反映了 MVS 点云的重建质量:

$$\text{精确度误差} = \frac{\sum_{\boldsymbol{p} \in P} \|d(\boldsymbol{p}, \boldsymbol{p}')\|}{n}。 \tag{10}$$

式中:P 为重建点云集合;$d(,)$ 为求点云距离函数;$\boldsymbol{p}$ 为 P 中 0.2 mm 采样点;$\boldsymbol{p}'$ 为 $\boldsymbol{p}$ 对应的真实点云集合 P' 中 0.2 mm 采样点;n 为重建点云个数。

完整度为参考点云到 MVS 点云的距离差,反映了 MVS 表面的重建质量:

$$\text{完整度误差} = \frac{\sum_{\boldsymbol{p}' \in P'} \|\mathrm{d}(\boldsymbol{p}', \boldsymbol{p})\|}{n'}。 \tag{11}$$

式中:n' 为真实点云个数。

总体误差为精确度误差和完整度误差的总体评价:

$$\text{总体误差} = \frac{\text{精确度误差} + \text{完整度误差}}{2}。 \tag{12}$$

此外,对于基于深度学习的 MVS 方法,本文实验计算了效率指标,包括深度图分辨率、显存开销和帧率。

3.4　实验方案

本文使用基于深度学习的 MVS 领域中使用的 DTU 训练集作为实验的训练集,遵循相同的过程来生成渲染深度图进行训练。使用 Pytorch 框架实现本文的模型,输入图像的分辨率为 640×512 像素,视图数 N 设置为 3。使用了与 MVSNet 相同的视图选择策略,为参考图像选择源图像进行训练。在稀疏深度图预测中,沿用了 Fast-MVSNet 的深度采样数 $D = 48$,本文模型采用逆深度采样,在 425 ~ 921 mm 进行逆深度均匀采样,以便更好地处理高分辨率输入图像。使用初始学习率为 0.001 的

RMSProp 优化器,每 2 个周期降低学习率 0.9。在 1 台 Nvidia Tesla T4 GPU 上进行训练,批量大小为 4。首先对稀疏深度图预测模块和传播模块进行了 4 轮的预训练,然后对整个模型进行 12 轮端到端训练。

在 DTU 测试集预测时,使用分辨率为 1 280×960 像素的 $N=5$ 个图像作为输入,设置深度平面的采样数 $D=96$。首先为每个参考图像预测一个深度图,然后使用后处理 fusibile 将预测的深度图融合到点云中,对输出点云进行定量测试得到实验结果。在 1 台 Nvidia Tesla T4 GPU 上进行预测。

4 结果分析

4.1 精度分析

本文方法在 DTU 测试集上与其他 MVS 方法精度结果对比如表 1 所示。从表 1 可以看出,本文所提轻量化方法的平均精确度误差与最好的 Gipuma 对比仅增加 0.080 mm,平均完整度误差与 GBI-Net 对比仅增加 0.219 mm,平均总体误差与最好的 GBI-Net 对比仅增加 0.133 mm。图 6 为 DTU 测试集中场景 scan23 和 scan34 的重建点云结果。

表 1 不同方法在 DTU 测试集上的精度结果对比

Table 1 Accuracy result comparison of different methods on DTU test dataset

方法	精确度误差/mm	完整度误差/mm	总体误差/mm
Camp[16]	0.835	0.554	0.695
Tola[18]	0.342	1.190	0.766
Gipuma[19]	**0.283**	0.873	0.578
Furu[17]	0.613	0.941	0.777
PU-Net[20]	1.220	0.667	0.943
MVSNet[8]	0.456	0.646	0.551
R-MVSNet[9]	0.385	0.452	0.417
Point-MVSNet[11]	0.361	0.421	0.391
SurfaceNet[7]	0.450	1.040	0.745
Fast-MVSNet[12]	0.359	0.441	0.400
GBI-Net[21]	0.315	**0.262**	**0.289**
TransMVSNet[22]	0.321	0.289	0.305
本文方法	0.363	0.481	0.422

(a) scan23的真实点云 (b) scan23的MVSNet预测点云 (c) scan23的Fast-MVSNet预测点云 (d) scan23的本文方法预测点云

(e) scan34的真实点云 (f) scan34的MVSNet预测点云 (g) scan34的Fast-MVSNet预测点云 (h) scan34的本文方法预测点云

图 6 DTU 测试集 scan23 和 scan34 点云预测结果

Figure 6 Point cloud evaluation results on DTU test dateset scan23 and scan34

4.2 效率分析

在 DTU 测试集上与部分基于深度学习的 MV 方法进行效率对比,结果如表 2 所示。

表 2 不同方法在 DTU 测试集上的效率结果对比

Table 2 Efficiency result comparison of different methods on DTU test dataset

方法	深度图分辨率/像素	显存开销/GB	帧率/(帧·s^{-1})
MVSNet[8]	288×216	10.8	1.05
R-MVSNet[9]	400×300	6.7	9.10
Point-MVSNet[11]	640×480	8.7	3.35
Fast-MVSNet[12]	640×480	5.5	**0.60**
GBI-Net[21]	**1 600×1 152**	4.3	—
TransMVSNet[22]	1 182×864	3.7	—
本文方法	640×480	**3.1**	3.60

从表 2 可以看出,本文所提轻量化方法在显存开销上优于其他 MVS 方法。由于预测时显存开销与 CUDA Context 有关,不同显卡的 CUDA Context 占用不同,且不同版本的 Pytorch 框架也会产生较小影响,因此在不同显卡和不同 Pytorch 版本上测试时显存开销可能有微小差别。

4.3 消融实验

为了验证本文的轻量化特征提取 FPN 模块、稀疏深度图模块和正则化模块的有效性,本文设计了 3 个变种方法,如表 3 所示。

方法(a)为 FPN 模块+稀疏深度图模块+3DCNN 正则化模块架构,使用基于深度学习的 MVS 常用 FPN 模块提取图像特征,使用 Fast-MVSNet 中稀疏深度图模块构造代价体,使用 3DCNN 模块对代价体进行正则化,组成骨干网络架构。方法(b)

为轻量化 FPN 模块+稀疏深度图模块+3DCNN 正则化模块架构,在方法(a)的基础上,轻量化 FPN 模块替换了原 FPN 模块。可以看出,所提轻量化 FPN 模块大幅度降低了显存占用,精度上几乎没有损失。方法(c)为轻量化 FPN 模块+本文方法稀疏深度图模块+3DCNN 正则化模块架构,在方法(b)的基础上,本文所提稀疏深度图模块替换了原稀疏深度图模块。可以看出,本文方法稀疏深度图模块的 MVS 精确度误差、完整度误差和总体误差均有所下降。方法(d)为轻量化 FPN 模块+本文稀疏深度图模块+本文正则化模块架构,即本文所提方法。在方法(c)的基础上,本文所提正则化模块替换了原 3DCNN 正则化模块,显存开销大幅降低至 3.1 GB,精度略微有所下降。

表 3　消融实验

Table 3　Ablation study

方法	架构	精确度误差/mm	完整度误差/mm	总体误差/mm	深度图分辨率/像素	显存开销/GB	帧率/(帧·s^{-1})
(a)	FPN+稀疏深度图+3DCNN 正则化	0.359	0.441	0.400	640×480	5.5	**0.60**
(b)	轻量化 FPN+稀疏深度图+3DCNN 正则化	0.361	0.441	0.401	640×480	3.8	1.23
(c)	轻量化 FPN+本文方法稀疏深度图+3DCNN 正则化	**0.357**	**0.429**	**0.393**	640×480	6.7	1.25
(d)	轻量化 FPN+本文稀疏深度图+本文正则化	0.363	0.481	0.422	640×480	**3.1**	3.60

5　结论

本文针对基于深度学习的 MVS 网络参数量大、显存占用较高的问题,提出了 LSD-MVSNet 网络。实验结果表明:本文所提轻量化多尺度特征提取网络能够充分提取图像的高层语义信息,降低显存开销;邻域特征通道聚合充分利用图像特征,保证了一定的精度;正则化过程中卷积循环网络能够进一步降低显存开销,使得本文方法参数量大大减少,运行时显存开销显著降低。

参考文献:

[1] YAN X C, YANG J M, YUMER E, et al. Perspective transformer nets: learning single-view 3D object reconstruction without 3D supervision[C]//Proceedings of the 30th International Conference on Neural Information Processing Systems. New York: ACM, 2016: 1704-1712.

[2] SUN X Y, WU J J, ZHANG X M, et al. Pix3D: dataset and methods for single-image 3D shape modeling[C]// 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2018: 2974-2983.

[3] FURUKAWA Y, HERNÁNDEZ C. Multi-view stereo: a tutorial[J]. Foundations and Trends in Computer Graphics and Vision, 2015, 9(1/2): 1-148.

[4] 纪勇, 刘丹丹, 罗勇, 等. 基于霍夫投票的变电站设备三维点云识别算法[J]. 郑州大学学报(工学版), 2019, 40(3): 1-6, 12.
JI Y, LIU D D, LUO Y, et al. Recognition of three-dimensional substation equipment based on Hough transform [J]. Journal of Zhengzhou University (Engineering Science), 2019, 40(3): 1-6, 12.

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

[6] HUANG P H, MATZEN K, KOPF J, et al. DeepMVS: learning multi-view stereopsis [C] // 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2018: 2821-2830.

[7] JI M Q, GALL J, ZHENG H T, et al. SurfaceNet: an end-to-end 3D neural network for multiview stereopsis[C]//2017 IEEE International Conference on Computer Vision (ICCV). Piscataway: IEEE, 2017: 2326-2334.

[8] YAO Y, LUO Z X, LI S W, et al. MVSNet: depth inference for unstructured multi-view stereo[C]//European Conference on Computer Vision. Cham: Springer, 2018: 785-801.

[9] YAO Y, LUO Z X, LI S W, et al. Recurrent MVSNet for high-resolution multi-view stereo depth inference[C]// 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2019: 5520-5529.

[10] 杜弘志, 张腾, 孙岩标, 等. 基于门控循环单元的立体匹配方法研究[J]. 激光与光电子学进展, 2021, 58(14): 387-394.
DU H Z, ZHANG T, SUN Y B, et al. Stereo matching method based on gated recurrent unit networks[J]. Laser & Optoelectronics Progress, 2021, 58(14): 387-394.

[11] CHEN R, HAN S F, XU J, et al. Point-based multi-view stereo network[C]//2019 IEEE/CVF International Conference on Computer Vision (ICCV). Piscataway: IEEE, 2019: 1538-1547.

[12] YU Z H, GAO S H. Fast-MVSNet: sparse-to-dense multi-view stereo with learned propagation and Gauss-Newton refinement[C]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Pis-

cataway: IEEE, 2020: 1946-1955.
[13] MA L B, LI N, YU G, et al. Pareto-wise ranking classifier for multi-objective evolutionary neural architecture search[J]. IEEE Transactions on Evolutionary Computation, 2023: 1-12.
[14] LI N, MA L B, YU G, et al. Survey on evolutionary deep learning: principles, algorithms, applications, and open issues[J]. ACM Computing Surveys, 2024, 56(2): 1-34.
[15] COLLINS R T. A space-sweep approach to true multi-image matching[C]//Proceedings CVPR IEEE Computer Society Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2002: 358-363.
[16] CAMPBELL N D, VOGIATZIS G, HERNÁNDEZ C, et al. Using multiple hypotheses to improve depth-maps for multi-view stereo[C]// 10th European Conference on Computer Vision. New York: ACM, 2008: 766-779.
[17] FURUKAWA Y, PONCE J. Accurate, dense, and robust multiview stereopsis[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2010, 32(8): 1362-1376.
[18] 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.
[19] GALLIANI S, LASINGER K, SCHINDLER K. Massively parallel multiview stereopsis by surface normal diffusion[C]//2015 IEEE International Conference on Computer Vision (ICCV). Piscataway: IEEE, 2015: 873-881.
[20] YU L Q, LI X Z, FU C W, et al. PU-net: point cloud upsampling network[C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2018: 2790-2799.
[21] MI Z X, DI C, XU D. Generalized binary search network for highly-efficient multi-view stereo[C]//2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2022: 12981-12990.
[22] DING Y K, YUAN W T, ZHU Q T, et al. TransMVSNet: global context-aware multi-view stereo network with transformers[C]//2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2022: 8575-8584.

MVS Method Based on Lightweight Deep Convolutional Recurrent Network

SHE Wei[1,2,3], KONG Xiangji[1,3], GUO Shuming[2,4], TIAN Zhao[1,3], LI Yinghao[1,2,3]

(1. School of Cyber Science and Engineering, Zhengzhou University, Zhengzhou 450002, China; 2. Songshan Laboratory, Zhengzhou 450046, China; 3. Zhengzhou Key Laboratory of Blockchain and Data Intelligence, Zhengzhou 450002, China; 4. China National Digital Switching System Engineering & Technological R&D Center, Zhengzhou 450002, China)

Abstract: Based on deep learning MVS methods, neural networks suffered from a large number of parameters and high GPU memory consumption. To address this issue, a lightweight deep convolutional recurrent network recurrent network-based MVS method was proposed. Firstly, the original images passed through a lightweight multi-scale feature extraction network to obtain high-level semantic feature maps. Then, a sparse cost volume to reduce the computational workload was constructed. Next, GPU memory consumption was reduced by using a simple plane sweeping technique that utilized by a convolutional recurrent network for cost volume regularization. Finally, sparse depth maps were extended to dense depth maps using an extension module. With a refinement algorithm, the proposed approach achieved a certain level of accuracy. The proposed approach was compared to state-of-the-art methods on the DTU dataset including traditonal MVS methods Camp, Furu, Tola, and Gipuma, and also including deep learning-based MVS methods SurfaceNet, PU-Net, MVSNet, R-MVSNet, Point-MVSNet, Fast-MVSNet, GBI-Net, and TransMVSNet. The results demonstrated that the proposed approach reduced GPU consumption to approximately 3.1 GB during the prediction stage, and the differences in precision compared to other methods were relatively small.

Keywords: lightweight; deep convolutional recurrent network; MVS method; regularization; DTU dataset