https://doi.org/10.1051/jnwpu/20244210129

基于多视图传播的无监督三维重建方法

罗锦锋[1], 袁冬莉[1], 张蓝[2], 屈耀红[1], 宿世鸿[1]

(1.西北工业大学 自动化学院, 陕西 西安 710072; 2.西安电子科技大学 网络与信息安全学院, 陕西 西安 710071)

摘 要:提出一种端到端的深度学习框架,从多视图中计算深度图从而重建出三维模型。针对目前大多数研究方法通过 3D 卷积实现 3D 成本体积正则化并回归得到初始深度图带来巨大的 GPU 内存消耗,以及由于设备受限导致有监督的方法中深度图真值难以获取的问题,提出一种多视图传播的无监督三维重建方法。该方法借鉴 Patchmatch 算法思想,在深度范围内将深度划分 n 层,通过多视图传播得到深度假设,并利用多个视图之间的光度一致性、结构相似性和深度平滑度构建多指标损失函数,作为网络中学习深度预测的监督信号。实验表明,文中提出的方法在 DTU、Tanks & Temples 和自制数据集上的性能和泛化性非常有竞争力,比采用 3D 成本体积正则化的方法快 1.7 倍以上,内存使用量减少 75%。

关 键 词:多视图传播;无监督;三维重建;Patchmatch 算法;多指标损失函数
中图分类号:TP181 **文献标志码**:A **文章编号**:1000-2758(2024)01-0129-09

三维重建技术作为计算机视觉领域的热门技术在医学治疗、航天航海、自动驾驶、VR 虚拟现实等领域应用相较成熟,一直受到大家重点关注。出于对重建成本和设备易得性考虑,利用视觉传感器获得多张已知相机参数的不同视角图像重建出真实三维场景成为研究人员常用手段之一,即多视图立体匹配(multi-view stereo,MVS)。尽管,MVS 已经被研究了几十年,但它仍面临一些挑战,如遮挡、照明变化以及无纹理区域等问题使得 MVS 的实际应用更加具有挑战性。

随着深度学习的高速发展,越来越多的研究人员将端到端的神经网络引入到 MVS 方法中,将三维重建看作分类和回归问题,从而利用神经网络在训练过程中得到的先验知识,解决经典 MVS 模型无法解决的问题。2018 年,Yao 等[1] 首次将深度学习引入到 MVS 方法中,提出 MVSNet,其基于 Plane-sweeping 算法[2] 和可微单应性变换构建 3D 代价体,使用 3D CNN 对其进行正则化,然后回归出深度图,最后通过融合估计的深度图来获得 3D 几何体,但 3D CNN 的使用需要消耗巨大的 GPU 内存。为此,R-MVSNet 采用 GRU 时序网络以顺序方式正则化代价体[3],以增加大量运行时间为代价,减少了 GPU 内存需求。Yang 等[4] 提出了 CVP-MVSNet,采用从粗到细的网络构建代价体金字塔,而不是像 MVSNet 方法一样以固定分辨率构建代价体,与上述所提方法相比,有着更加紧凑、轻量级的网络结构。这些通过构建代价体的方法虽然在重建质量上取得了不错的效果,但在运行时间和消耗内存方面仍不令人满意。

几种传统的 MVS 方法[5-7] 舍弃了构建代价体的思想,基于 Patchmatch 算法思想,消耗较少的内存就能完成深度图估计。Patchmatch 算法最早于 2009 年由 Barnes 等[8] 提出,用于寻找图像块之间的近似最近邻匹配,寻找最近邻的速度是一次质的飞跃。2011 年,文献[9]提出了 Patchmatch Stereo 算法,其核心思想是基于迭代传播为所有的像素点分别找一个属于该像素的视差平面,利用这个视差平面计算该像素点的视差值。2015 年,Galliani 等[5] 提出了 Gipuma 算法,该算法基于 Patchmatch Stereo 算法提出了一种新的、大规模并行的高质量多视点匹配方法,它使用红黑棋盘模式并行化进行过程中的消息

收稿日期:2023-03-17
基金项目:国家自然科学基金(61473229)与航空科学基金(20181353013)资助
作者简介:罗锦锋(1999—),硕士研究生
通信作者:袁冬莉(1966—),副教授 e-mail:yuandongli@nwpu.edu.cn

传递,更适合多核 GPU,并从双目匹配扩展到多视图匹配。2016 年,Schönberger 等[6] 基于 Patchmatch 算法提出了 Colmap 算法,它联合逐像素视图选择、深度图和表面法线进行深度估计。2019 年,Xu 等[7]在 Gipuma 算法基础上,提出了高效准确的 ACMH算法,效率是 Colmap 算法的 8~10 倍;针对低纹理区域的深度估计,他进一步提出了多尺度几何一致性引导框架(ACMM),并将 ACMH 与 ACMM 相结合,获得较粗尺度下低纹理区域可靠的深度估计值,并确保其可以传播到更细尺度。

文献[10]结合深度学习和 Patchmatch 方法思想,提出了 PatchmatchNet 算法,它继承 Patchmatch 算法效率的同时,也利用深度学习提高了性能。相比于 3D CNN 正则化 3D 代价体和 GRU 时序网络正则化 2D 代价图的方法,运行速度快 2.5 倍以上,内存使用量减少 66.67%。然而,这些方法虽然取得了不错的精度,但依赖于 3D 地面真实数据的监督。

针对 3D 地面真实数据获取困难的问题,不少研究人员开始研究无监督的 MVS 方法,通过建立自我监督损失以估计该视图上重建的图像和真实图像之间的差异。Khot 等[11] 提出了基于鲁棒光度一致性的无监督 MVS 方法(Unsup_MVS),为克服视图之间的遮挡和照明变化,提出一种稳健的损失公式,即强制执行一阶一致性和针对每个点选择性地强制执行与某些视图一致性的方法。Huang 等[12] 提出了一种新的无监督多度量 MVS 网络(M^3VSNet),提出一种新的多度量损失函数,该函数结合了像素和特征损失函数,以从匹配对应的不同视角学习固有约束。但是,这些无监督 MVS 方法中大部分采用了经典的方法,即采用 3D CNN 对 3D 代价体进行正则化并回归出深度图的方法,耗费了大量的内存和时间。

针对上述问题,借鉴 PatchmatchNet 方法,本文提出一种基于多视图传播的无监督三维重建方法,命名为 Uns-PMVSNet,改进了 PatchmatchNet 网络内部结构,不用依赖于 3D 地面真实数据并使用较少的内存和较短的时间就能重建出较高质量的三维模型,使网络的普适性得到大幅度提升。

1 Patchmatch 算法理论

文献[9]提出 Patchmatch 方法核心思想是随机搜索和迭代更新,主要包括 4 个步骤:①初始化:每个像素根据深度范围随机初始化相应的深度值;②传播:把初始的所有深度平面内的正确的深度平面传播至同一深度平面内的其他像素,迭代多次;③匹配代价计算与聚合:计算每个像素不同深度估计的匹配代价,选取代价最小的深度值作为深度估计值;④随机分配:在每次传播后,对当前像素的平面在一定范围内随机生成一个调整量,寻找代价最小的那个平面,加快找到真实的平面。

基于 Patchmatch 算法思想,文献[10]提出了基于学习的 Patchmatch 方法,即 PatchmatchNet。该方法主要流程为:①深度图初始化:借鉴 Plane-sweeping 算法[2] 思想,预定义深度范围为[$d_{\min}$, $d_{\max}$],第一次迭代时,在逆深度范围内对每个像素的深度假设采样 48 个间隔得到 48 个深度层,这些深度层组成一个采样深度图;②自适应传播:采用 2D 卷积学习每个像素 p 的 K 个邻域的额外偏移量,然后根据这个偏移量得到自适应邻域的像素坐标,将邻域对应的深度值传给中心像素作为中心像素的深度假设,然后加上初始化得到的深度图为中心上下采样几层(局部扰动)后的深度假设作为新的深度假设;③自适应匹配代价计算:根据第一步得到深度假设后,计算邻域帧的特征映射到参考帧对应的深度层下的特征相似性,然后根据计算得到的特征相似性采用 1×1×1 卷积和 sigmoid 函数计算每个像素邻域帧的权重,根据该权重值和特征相似性计算最终的匹配代价;④自适应匹配代价聚合:自适应采样一些邻域内像素对应的深度假设,计算邻域像素相对于中心像素的深度图相似性,根据特征相似性和深度图相似性计算代价聚合的权重值,最后采用 1×1×1 卷积进行匹配代价聚合;⑤深度图回归:采用 softmax 函数将匹配代价聚合值回归得到各个像素对应的深度值;⑥深度图优化:将估计的深度图分辨率从 $W/2\times H/2$ 提高到 $W\times H$,并基于 MSG-Net[13],设计一个深度残差网络,使用 RGB 图像优化估计的深度图。

考虑 PatchmatchNet 方法在深度图随机初始化带来的较大计算量以及该方法模型需要依赖深度图真值监督的局限性,本文对此做出一些改进。

2 多视图传播无监督网络设计

本文提出的多视图传播无监督三维重建方法继承了 PatchmatchNet 方法利用从粗到细的框架估计

深度图的思想,整体网络结构框架分为多尺度特征提取、深度图初始化、自适应传播、自适应匹配代价计算、自适应匹配代价聚合、深度图回归和多指标损失函数,相比于 PatchmatchNet 方法主要做两点改进。

从第 1 节可知 PatchmatchNet 方法[10]在深度图初始化阶段采用的是随机初始化方法,该方法需要比较大的计算量;考虑到采用 Colmap[6]、SLAM[14]等方法获取图像的相机参数(内参和外参),在这过程中也获得了三维模型的稀疏点。基于这些稀疏点,本文提出通过三角剖分[15]获得深度图的方法取代 PatchmatchNet 网络中深度图随机初始化方法,简化了网络结构,加快了网络的运行时间。

另外,PatchmatchNet 方法[10]属于有监督的三维重建方法,在网络训练过程中,需要采用深度图真值进行监督训练,这些深度图真值的质量取决于激光雷达、深度相机等设备的质量,且存在设备受限获取不了深度图的问题。因此,本文提出多指标的损失函数,从大规模的无监督数据中挖掘自身的监督信息,使得该网络不需要依赖 3D 地面真实数据监督就能获得较好的精度,且网络的普适性得到大幅度提升。

综上所述,整体网络结构如图 1 所示,由多尺度特征提取器、基于学习的 Patchmatch 网络(网络结构详见图 2)和空间细化模块组成。

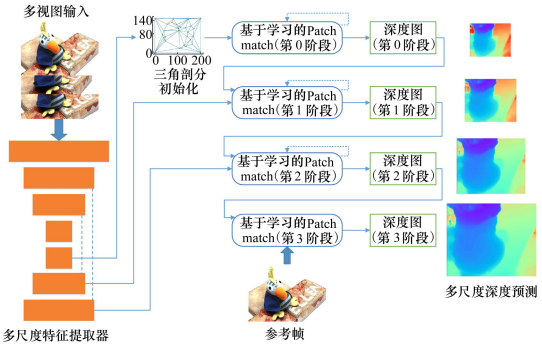


图 1　整体网络结构

2.1　多尺度特征提取

在 N 张尺寸为 $W \times H$ 图片中,定义 I_0 为参考图像,$\{I_i\}_{i=1}^{N-1}$ 表示邻域图像。采用 FPN[16]特征提取网络从输入的图像中,以多个分辨率分层提取特征,用于基于学习的 Patchmatch 网络框架中,以从粗到细的方式不断推进深度图估计。

2.2　基于学习的 Patchmatch 网络

基于学习的 Patchmatch 网络主要分为:初始化深度图、深度假设自适应传播、自适应匹配代价计算、聚合和深度图计算。相比于传统的 Patchmatch 使用倾斜平面的方法(每像素对应的深度假设包括深度值和对应的法向量),为了减小计算量,采用 MVSNet 方法[1]中使用的前向平行平面(假设像素在相机坐标系下平行于图像平面,对应的深度假设只包含深度值),具体结构如图 2 所示。

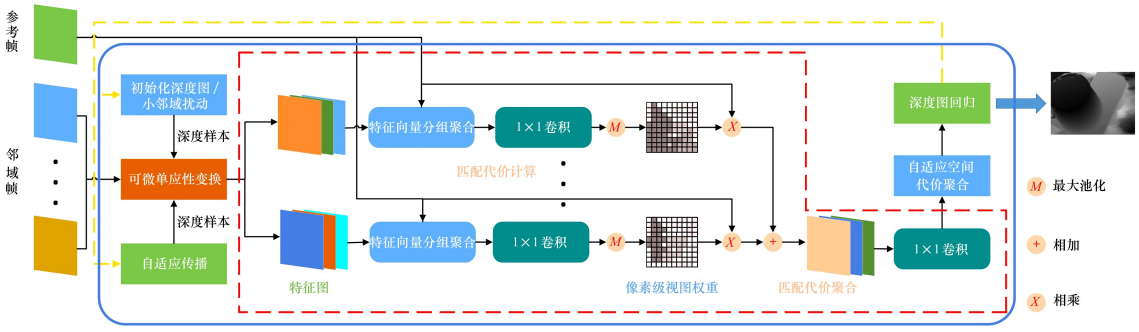


图 2　基于学习的 Patchmatch 网络结构

2.2.1　深度图初始化

PatchmatchNet 方法[10]的第一次迭代中,预定义一个深度范围 $[d_{\min}, d_{\max}]$,并在逆深度范围中随机初始化深度图。本文提出的方法在这方面进行了改进,考虑实际过程中,通过 Colmap[6]、SLAM[14]等方法获取图像的相机参数(内参和外参)的同时,也可以获得模型的稀疏点,因此本文通过这些稀疏点初始化深度图,具体方法是:①根据深度范围定义深度图的 4 个角点,界定感兴趣区域;②基于定义的深度图区域,将稀疏点通过几何变换投影到对应帧上,

获得稀疏深度图;③采用三角剖分[15]算法对投影后的的稀疏点进行处理,即根据 3 个稀疏点构成的三角形平面,在该平面内插值获得其他点的深度信息进而获得完整的深度图,具体流程如图 3 所示。

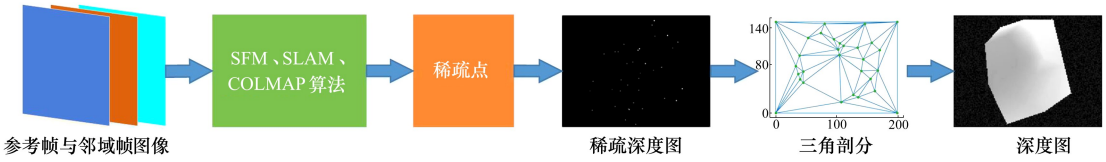


图 3 深度图初始化流程图

三角剖分算法主要分为 4 步:

1) 根据(1)式的转换关系,将稀疏点的像素 $\boldsymbol{P}_u$ 的齐次坐标转换到相机坐标下,求得三角形 3 个顶点的相机坐标$[X \quad Y \quad Z]'$。

$$\begin{bmatrix} X \\ Y \\ Z \end{bmatrix} = \begin{bmatrix} f_x & 0 & c_x \\ 0 & f_y & c_y \\ 0 & 0 & 1 \end{bmatrix}^{-1} \begin{bmatrix} u \\ v \\ 1 \end{bmatrix} \cdot Z = \boldsymbol{K}^{-1} \cdot \boldsymbol{P}_u \cdot Z \quad (1)$$

式中:$\boldsymbol{K}$ 为相机内参;f_x,f_y 为焦距;c_x,c_y 为相机光轴在图像坐标系的偏移量;像素 $\boldsymbol{P}_u$ 的齐次坐标$[u \quad v \quad 1]'$;Z 为深度值。

2) 通过三角形 3 个顶点的相机坐标求得三角形组成的平面的法向量 $\boldsymbol{n}$,如(2) 式所示。

$$\boldsymbol{n} = \boldsymbol{a} \times \boldsymbol{b} = (\boldsymbol{p}_1 - \boldsymbol{p}_0) \times (\boldsymbol{p}_2 - \boldsymbol{p}_0) \quad (2)$$

式中:$\boldsymbol{p}_0$,$\boldsymbol{p}_1$,$\boldsymbol{p}_2$ 是平面内 3 个点的坐标。

3) 根据三角形平面的法向量 $\boldsymbol{n}$ 和一个顶点的坐标 $\boldsymbol{p}_0$ 求得该三角形平面方程 $\boldsymbol{L}$,如(3) 式所示。

$$\boldsymbol{L} = \frac{\boldsymbol{n}_c}{\boldsymbol{n}_c \cdot \boldsymbol{p}_0} = \frac{\boldsymbol{n}}{\boldsymbol{n} \cdot \boldsymbol{p}_0} \quad (3)$$

式中:$\boldsymbol{n} = (n_1, n_2, n_3)$;$\boldsymbol{n}_c = \dfrac{\boldsymbol{n}}{|\boldsymbol{n}|} = \dfrac{\boldsymbol{n}}{\sqrt{n_1^2 + n_2^2 + n_3^2}}$。

4) 将每个三角形进行栅格化得到面内投影的每一个像素坐标 $\boldsymbol{P}_u$,再根据三角形平面方程 $\boldsymbol{L}$ 和像素坐标与相机坐标的转换关系 $\boldsymbol{K}^{-1}$ 在每个三角形平面内插值得到对应的深度值,如(4) 式所示。

$$Z = \frac{1}{\boldsymbol{L} \cdot \boldsymbol{K}^{-1} \cdot \boldsymbol{P}_u} \quad (4)$$

本文提出通过稀疏点进行三角剖分[15]获得深度图的方法取代 PatchmatchNet[10] 网络中深度图随机初始化方法,从而缩短网络的运行时间。

2.2.2 自适应传播

大部分传统的方法从静态的邻域中传播深度假设,但是深度图的空间一致性往往只对一个表面上的像素成立,所以希望在采样邻域的过程中只采样和当前像素在同一表面上的邻域。即采用 2D 卷积学习每个像素 $\boldsymbol{p}$ 的 K_p 个邻域的额外偏移量 $\{\Delta \boldsymbol{o}_i(\boldsymbol{p})\}_{i=1}^{K_p}$,并通过双线性插值获得深度假设,实现自适应传播,如(5) 式所示,加快算法的收敛速度,提升精度。

$$\boldsymbol{D}_p(\boldsymbol{p}) = \{\boldsymbol{D}(\boldsymbol{p} + \boldsymbol{o}_i + \Delta \boldsymbol{o}_i(\boldsymbol{p}))\}_{i=1}^{K_p} \quad (5)$$

式中:$\{\boldsymbol{o}_i\}_{i=1}^{K_p}$ 为固定的偏移量;$\boldsymbol{D}_p(\boldsymbol{p})$ 为每个像素 $\boldsymbol{p}$ 深度假设集合;$\boldsymbol{D}$ 为上个阶段输出的深度图。

2.2.3 自适应匹配代价计算

自适应匹配代价计算即计算图像特征相似度,通过自适应传播获得所有深度假设后,在参考帧的深度假设下建立前向平行平面,采用可微单应性变换将邻域帧的图像特征映射到这些平面中,可微单应性变换如(6)式所示。

$$\boldsymbol{p}_{i,j} = \boldsymbol{p}_i(d_j) = \boldsymbol{K}_i \cdot (\boldsymbol{R}_{0,i} \cdot (\boldsymbol{K}_0^{-1} \cdot \boldsymbol{p} \cdot d_j) + \boldsymbol{t}_{0,i}) \quad (6)$$

式中:$\boldsymbol{p}_{i,j}$ 为参考帧在深度 d_j 下的像素 $\boldsymbol{p}$ 映射到邻域帧 i 下的像素位置;$\{\boldsymbol{K}_i\}_{i=0}^{K}$ 为参考帧 0 和邻域帧 i 的相机内参;$\{[\boldsymbol{R}_{0,i} | \boldsymbol{t}_{0,i}]\}_{i=1}^{K}$ 为参考帧 0 和邻域帧 i 的相机外参;$d_j = d_j(\boldsymbol{p})$ 为像素 $\boldsymbol{p}$ 对应深度假设。

根据邻域帧映射到参考帧上的图像特征和参考帧的图像特征计算相似度,进而获得每个像素的匹配代价。具体是将图像特征通道分为 G 组以降低内存,然后计算第 g 组的特征相似度,如(7)式所示。

$$\boldsymbol{S}_i(\boldsymbol{p}, j)^g = \frac{G}{C} \langle \boldsymbol{F}_0(\boldsymbol{p})^g, \boldsymbol{F}_i(\boldsymbol{p}_{i,j})^g \rangle \quad (7)$$

式中:$\langle \cdot, \cdot \rangle$ 表示内积;C 为特征通道数;$\boldsymbol{F}_0(\boldsymbol{p}) \in \mathbf{R}^C$ 为像素 $\boldsymbol{p}$ 的特征;$\boldsymbol{F}_i(\boldsymbol{p}_{i,j}) \in \mathbf{R}^C$ 为像素 $\boldsymbol{p}$ 映射到邻域帧 i 下对应的特征;$\boldsymbol{S}_i(\boldsymbol{p}, j) \in \mathbf{R}^G$ 为群相似性向量。

2.2.4 自适应空间代价聚合和深度回归

根据每组的特征相似度,采用 1×1×1 卷积和 sigmoid 函数计算每个像素的邻域帧的权重 $\boldsymbol{\omega}_i(\boldsymbol{p})$,然后根据这些像素级视图权重为像素 $\boldsymbol{p}$ 和第 j 个深

度假设计算每组特征相似性 $\bar{\boldsymbol{S}}(\boldsymbol{p},j)$,如(8)式所示。

$$\bar{\boldsymbol{S}}(\boldsymbol{p},j)=\frac{\sum_{i=1}^{N-1}\boldsymbol{\omega}_i(\boldsymbol{p})\cdot\boldsymbol{S}_i(\boldsymbol{p},j)}{\sum_{i=1}^{N-1}\boldsymbol{\omega}_i(\boldsymbol{p})} \tag{8}$$

式中,$N-1$ 为邻域帧数量。

考虑多尺度特征提取器在空间域中聚合了来自大感受野的相邻信息,为防止跨表面边界聚合问题,采用 2D 卷积学习每个像素 $\boldsymbol{p}$ 的额外 2D 偏移量 $\{\Delta\boldsymbol{p}_k\}_{k=1}^{K_e}$,并使用特征权重 ω_k 和深度权重 d_k 分配第 k 个邻域的贡献权重,然后采用 $1\times1\times1$ 卷积聚合 G 组匹配代价 $\widetilde{\boldsymbol{C}}(\boldsymbol{p},j)$,定义为

$$\widetilde{\boldsymbol{C}}(\boldsymbol{p},j)=\frac{1}{\sum_{k=1}^{K_e}\omega_k d_k}\sum_{k=1}^{K_e}\omega_k d_k\boldsymbol{C}(\boldsymbol{p}+\boldsymbol{p}_k+\Delta\boldsymbol{p}_{k,j}) \tag{9}$$

式中:$\boldsymbol{C}$ 为邻域像素的匹配代价,通过采样位置 $(\boldsymbol{p}+\boldsymbol{p}_k+\Delta\boldsymbol{p}_{k,j})_{k=1}^{K_e}$ 处的特征相似度双线性插值到邻域位置获得;K_e 为邻域像素的个数。

在深度图回归中,使用 softmax 函数将代价 $\widetilde{\boldsymbol{C}}(\boldsymbol{p},j)$ 转化成置信度 $\boldsymbol{P}$,然后根据每层深度假设值与对应的置信度值求期望获得像素 $\boldsymbol{p}$ 处的深度值 $\boldsymbol{D}(\boldsymbol{p})$,如(10)式所示。

$$\boldsymbol{D}(\boldsymbol{p})=\sum_{j=0}^{m-1}d_j\cdot\boldsymbol{P}(\boldsymbol{p},j) \tag{10}$$

式中,m 为深度层数。

2.3　深度图优化

深度图优化首先将估计的深度图的分辨率从 $W/2\times H/2$ 提高到 $W\times H$,然后使用 RGB 图像优化估计的深度图。基于 MSG-Net[13],设计一个深度残差网络,网络输出一个残差值加到 Patchmatch 的上采样估计中,以获得细化的深度图。为了避免对某个深度比例产生偏差,先将深度图归一化到[0,1],并在细化后将深度图反归一化回原始深度范围。

2.4　多指标损失函数

PatchmatchNet[10]方法采用数据集的深度图真值监督网络预测的深度值,本文在这方面做出改进,提出多指标损失函数,实现无需借助深度真值进行训练也能重建得到较好的精度。多指标损失函数包括光度一致性损失、结构相似性损失和深度平滑损失。

光度一致性损失的关键思想是在同一视图上计算邻域帧与参考帧的颜色差异,即根据(6)式将 $N-1$ 个邻域帧 I_i 映射到参考帧 I_1 上以计算颜色差异,另考虑到一些像素可能被映射到图像的外部区域,定义映射过程中的二进制有效掩码为 M_i,光度一致性损失计算如(11)式所示。

$$L_{\text{photo}}=\sum_{i=2}^{N}\frac{\|(I_i'-I_1)\odot M_i\|_2+\|(\nabla I_i'-\nabla I_1)\odot M_i\|_2}{\|M_i\|_1} \tag{11}$$

式中:I_i' 为邻域帧 I_i 映射到参考帧 I_1 的图像;∇ 表示梯度算子;$\odot$ 表示点积;$\|\ \|_1$ 表示 1- 范数;$\|\ \|_2$ 表示 2- 范数。

结构相似性损失的关键思想是在同一视图上计算邻域帧与参考帧之间的相似性,同理,根据(6)式将邻域帧映射到与参考帧同一视图上,根据(12)式计算 2 个图像之间的相似度。

$$S_{\text{SIM}}(x,y)=\frac{(2\mu_x\mu_y+c_1)(2\sigma_{xy}+c_2)}{(\mu_x^2+\mu_y^2+c_1)(\sigma_x^2+\sigma_y^2+c_2)} \tag{12}$$

式中:μ_x 为参考帧的均值;μ_y 为邻域帧的均值;σ_x 为参考帧的标准差;σ_y 为邻域帧的标准差;σ_{xy} 为参考帧与邻域帧之间的协方差;c_1,c_2 为常数,是避免分母为 0 的平滑因子。

定义结构相似性损失为

$$L_{S_{\text{SIM}}}=\sum_{i=2}^{N}[1-S_{\text{SIM}}(I_1,I_i')]M_i \tag{13}$$

深度平滑损失的关键思想是深度图变化较大的区域对应的 RGB 值变化也相对较大,定义深度平滑损失为

$$L_{\text{smooth}}=\sum_{i,j}\|\partial_x D^{ij}\|\mathrm{e}^{-\|\partial_x X^{ij}\|}+\|\partial_y D^{ij}\|\mathrm{e}^{-\|\partial_y X^{ij}\|} \tag{14}$$

式中:$\partial_x D^{ij}$ 表示深度图 x 方向的梯度;$\partial_x X^{ij}$ 表示 RGB 图 x 方向的梯度;$\partial_y D^{ij}$ 表示深度图 y 方向的梯度;$\partial_y X^{ij}$ 表示 RGB 图 y 方向的梯度。

综上所述,定义无监督三维重建网络的损失为

$$L_{\text{loss}}=\lambda_1 L_{\text{photo}}+\lambda_2 L_{S_{\text{SIM}}}+\lambda_3 L_{\text{smooth}} \tag{15}$$

式中,$\lambda_1,\lambda_2,\lambda_3$ 为权重参数。

3　实验与分析

本文所提出的 Uns-PMVSNet 方法在 DTU 数据集[17]、Tanks & Temples 数据集[18]和自制数据集下进行评估,验证本方法的性能和泛化性。

3.1 实验数据

DTU 数据集[17]是具有 128 个不同场景的室内多视图立体数据集,利用一个搭载可调节亮度灯的工业机械臂对每个物体拍摄 49 个视角,每个视角共有 7 个不同亮度,且所有场景有相同的相机轨迹;Tanks & Temples 数据集[18]是室外大型场景高分辨率图像,由现实环境中的视频序列提供,其中测试数据分为中级组和高级组。由于上述 2 个数据集都是通过专业的设备,且是固定距离拍摄的,因此为更好地验证本文方法的泛化能力,确保结果的可靠性,采用自制数据集进行验证。自制数据集是西北工业大学的标志性雕像,数据集包含 300 张水平 360°视角的图片,采用 1 200 万像素的手机在阳光下任意距离拍摄获得,且未对图像做滤波、曝光补偿等任何处理。

3.2 实验细节

本文使用了 Pytorch 框架实现该模型,并在 DTU 数据集[17]下进行训练和测试,选择场景:1,4,9,10,11,12,13,15,23,24,29,32,33,34,48,49,62,75,77,110,114,118 作为测试集,其他作为训练集,DTU 训练集是分辨率为 640×512 的图像,测试集是 1 600×1 200 的图像;输入图像数量设置为 5,将阶段 3,2,1 的 Patchmatch 迭代次数分别设置为 2,2,1,传播的邻域个数设置为 16,8,0,学习率设置为 0.001;结构相似性损失中图像均值、标准差和协方差采用平均池化实现,平滑因子 c_1 为 0.000 1,c_2 为 0.000 9;多指标损失函数权重参数(λ_1,λ_2,λ_3)分别设置为 12,6,0.05;网络在 NVIDIA RTX A6000 服务器上进行训练,在 NVIDIA GTX 1650 GPU 上进行测试。

为进一步验证本文方法的泛化能力,在更复杂的室外数据集 Tanks & Temples[18]和自己制作的数据集下测试了本文方法。另外,2.2.1 节提出采用稀疏点进行深度图初始化的方法中,这些稀疏点以及相机内参、外参和深度范围是由 Colmap[6]方法计算得到。

3.3 实验结果

为了验证本文提出的 Uns-PMVSNet 方法的性能,在 DTU 数据集[17]上进行测试并与其他 MVS 方法进行重建效果对比,如图 4 所示。从图 4 可以看出,本文提出的 Uns-PMVSNet 方法重建效果比 3D 激光扫描得到的地面真实点更好,比有监督的 PatchmatchNet 方法重建效果略差一点,相比于无监督的 JDACS[19]方法重建效果更好。

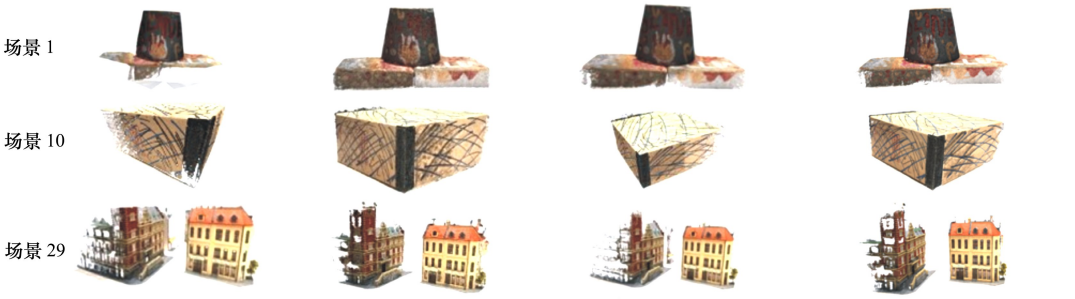


图 4 DTU 数据集部分场景重建结构对比图

本文方法在 DTU 测试集[17]上的部分重建结果如图 5 所示,另外,为验证本方法的泛化能力,在 Tanks & Temples 数据集[18]和自己制作的数据集上进行实验,在 Tanks & Temples 数据集上的部分重建重建结果分别如图 6 所示,在自制数据集上重建结果如图 7 所示。在图 6 中,可以观察到重建结果呈现出较好的细节保留和几何形状的精确重建,图中的物体轮廓清晰可见,纹理细节也得到了良好的恢复。图 7 表明,本文方法在自制数据集下仍能展现出令人满意的重建效果,具备较强的泛化性能。

图 5 DTU 数据集上的部分重建结果

图 6 Tanks & Temples 数据集上的部分重建结果

图 7　自制数据集上的重建结果

3.4　实验评估

将本文方法在整体质量、显存消耗和运行时间上与几种先进的重建方法进行对比,证明本文方法的先进性能。

3.4.1　质量评估

使用距离度量[17](越低越好)评估本文提出的 Uns-PMVSNet 在 DTU 数据集[17]上的重建质量,规定输入图像的分辨率为 1 600×1 200,评估结果如表 1 所示。

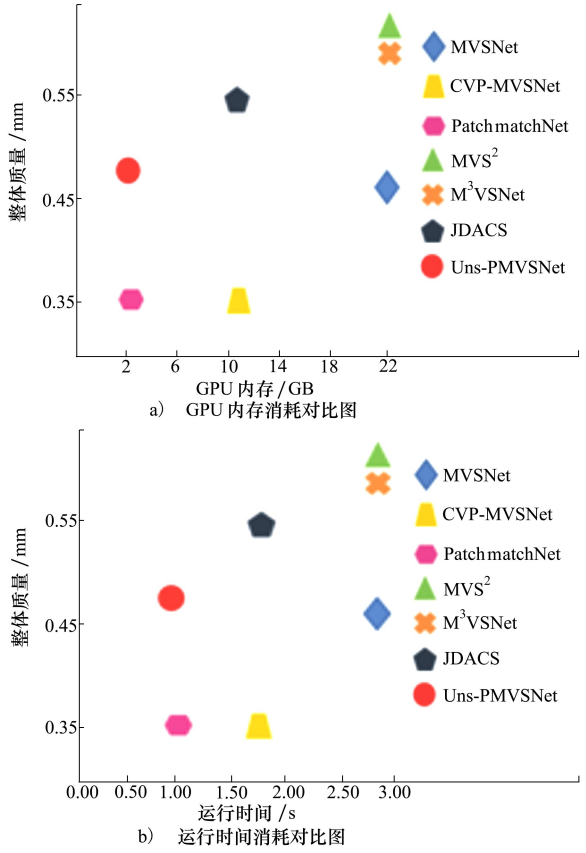
表 1　DTU 数据集上的评估结果

分类	方法	平均距离/mm		
		精度	完整度	整体质量
传统几何方法	Furu[20]	0.613	0.941	0.777
	Camp[21]	0.835	0.554	0.695
	Gipuma[5]	0.283	0.873	0.578
	Colmap[6]	0.400	0.664	0.532
有监督方法	Surfacenet[22]	0.450	1.040	0.745
	MVSNet[1]	0.444	0.741	0.592
	R-MVSNet[3]	0.383	0.452	0.417
	Fast-MVSNet[23]	0.336	0.403	0.370
	PatchmatchNet[10]	0.427	0.277	0.352
	CVP-MVSNet[4]	0.296	0.406	0.351
无监督方法	Unsup_MVS[11]	0.881	1.073	0.977
	MVS^2[24]	0.760	0.515	0.637
	M^{3}VSNet[12]	0.636	0.531	0.583
	JDACS[19]	0.571	0.515	0.543
	Uns-PMVSNet	0.436	0.515	0.476

距离度量包括精度度量、完整度度量和整体质量度量。精度度量是测量重建点与地面真值的平均距离,衡量重建点的质量;完整度度量是测量地面真实点到重建的平均距离,衡量重建模型表面的完整性;整体质量度量是综合精度和完整度,计算精度和完整度的平均值。从表 1 可以看出,本文提出的 Uns-PMVSNet 方法重建的整体质量仅略低于部分有监督方法,在传统几何方法和无监督方法中具有较强的竞争力。

3.4.2　显存和时间消耗对比

本文提出的 Uns-PMVSNet 方法在运行所耗显存和时间上较以往的大部分方法有很大的竞争力,规定输入图像的分辨率为 1 600×1 200,运行网络所耗 GPU 内存为 2.6 GB,运行时间为 1.0 s 左右,如图 8 所示。在图像分辨率同为 1 600×1 200 情况下,在有监督的方法中,如 MVSNet[1]、CVP-MSVNet[4]、PatchmatchNet[10]等,本文提出的方法在消耗显存方面比 MVSNet[1]少 88.6%左右,比 CVP-MVSNet[4]少 75%左右,比 PatchMatchNet[10]少 0.1 GB 显存;在运行时间方面比 MVSNet[1]少 63%左右,比 CVP-MVS-Net[4]少 41.2%左右;重建 DTU 数据集中每一个场景三维模型时,PatchmatchNet[10]耗时为 80 s 左右,而本文方法只需要 74 s 左右,比 PatchmatchNet[10]具有更快的运行速度。在同为无监督的方法中,如 MVS[24]、M^{3}VSNet[12]、JDACS[19]等,本文提出方法在显存消耗方面比 MVS[24]和 M^{3}VSNet[12]少 88.9%左右,比 JDACS[19]少 75%左右;在运行时间方面比 MVS[24]和 M^{3}VSNet[12]少64.3%左右,比 JDACS[19]少 44.4%左右。综上所述,本文提出的 Uns-PMVSNet 实现了优越的性能,在显存和运行时间消耗上具备较强的竞争力。



a)　GPU 内存消耗对比图

b)　运行时间消耗对比图

图 8　不同网络的 GPU 内存和时间消耗对比图

4 结 论

本文提出了 Uns-PMVSNet,这是一种基于多视图传播的无监督三维重建方法,该网络改进了 PatchMatchNet 深度图初始化这部分结构,采用三角剖分方法进行深度图初始化,降低了网络的复杂度,提高了网络运行时间。同时,基于该改进网络,提出无监督的三维重建方法。在 DTU、Tanks & Temples 和自己制作的数据集上的实验表明,本文提出的方法相较于采用 3D 成本体积正则化的方法,在运行速度上少快 1.7 倍,在内存使用量方面少 75%;相较于当前已知的无监督的 MVS 方法,本文方法在 GPU 内存消耗和耗时方面展现出显著的优势;此外,在重建精度、完整度和整体质量方面超过了许多已有的无监督 MVS 方法,具备较强的竞争力。

参考文献:

[1] YAO Y, LUO Z X, LI S W, et al. MVSNet: depth inference for unstructured multi-view stereo[C]//15th European Conference on Computer Vision, 2018: 785-801

[2] GALLUP D, FRAHM J M, MORDOHAI P, et al. Real-time plane-sweeping stereo with multiple sweeping directions[C]//2007 IEEE Conference on Computer Vision and Pattern Recognition, 2007: 1-8

[3] YAO Y, LUO Z X, LI S W, et al. Recurrent MVSNet for high-resolution multi-view stereo depth inference[C]//32nd IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019: 5520-5529

[4] YANG J, MAO W, ALVAREZ J, et al. Cost volume pyramid based depth inference for multi-view stereo[J]. IEEE Trans on Pattern Analysis and Machine Intelligence, 2022, 44(9): 4748-4760

[5] GALLIANI S, LASINGER K, SCHINDLER K. Massively parallel multiview stereopsis by surface normal diffusion[C]//2015 IEEE International Conference on Computer Vision, 2015: 873-881

[6] SCHÖNBERGER J L, FRAHM J M. Structure-from-motion revisited[C]//2016 IEEE Conference on Computer Vision and Pattern Recognition, 2016: 4104-4113

[7] XU Q, TAO W. Multi-scale geometric consistency guided multi-view stereo[C]//2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019: 5478-5487

[8] BARNES C, SHECHTMAN E, FINKELSTEIN A, et al. PatchMatch: a randomized correspondence algorithm for structural image editing[J]. ACM Transactions on Graphics, 2009, 28(3): 24

[9] BLEYER M, RHEMANN C, ROTHER C. PatchMatch stereo-stereo matching with slanted support windows[C]//British Machine Vision Conference, 2011

[10] WANG F, GALLIANI S, VOGEL C, et al. PatchmatchNet: learned multi-view patchmatch stereo[C]//IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021: 14189-14198

[11] KHOT T, AGRAWAL S, TULSIANI S, et al. Learning unsupervised multi-view stereopsis via robust photometric consistency[J/OL]. (2019-05-07)[2023-03-17]. https://arxiv.org/abs/1905.02706

[12] HUANG B, YI H, HUANG C, et al. M3VSNET: unsupervised multi-metric multi-view stereo network[C]//IEEE International Conference on Image Processing, 2021: 3163-3167

[13] HUI T W, LOY C C, TANG X O. Depth map super-resolution by deep multi-scale guidance[C]//14th European Conference on Computer Vision, 2016: 353-369

[14] MUR-ARTAL R, TARDÓJ D. ORB-SLAM2: an open-source slam system for monocular, stereo, and RGB-D cameras[J]. IEEE Trans on Robotics, 2017, 33(5): 1255-1262

[15] SHEWCHUK J R. Delaunay refinement algorithms for triangular mesh generation[J]. Computational Geometry & Applications, 2014, 47(1/2/3): 741-778

[16] LIN T Y, DOLLÁR P, GIRSHICK R, et al. Feature pyramid networks for object detection[C]//2017 IEEE Conference on Computer Vision and Pattern Recognition, 2017: 936-944

[17] AANAES H, JENSEN R R, VOGIATIZIS G, et al. Large-scale data for multiple-view stereopisis[J]. International Journal of Computer Vision, 2016, 120(2): 153-168

[18] 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): 1-13

[19] XU H, ZHOU Z, QIAO Y, et al. Self-supervised multi-view stereo via effective co-segmentation and data-augmentation[C]//35th AAAI Conference on Artificial Intelligence, 2021: 3030-3038

[20] FURUKAWA Y, PONCE J. Accurate, dense, and robust multiview stereopsis[J]. IEEE Trans on Pattern Analysis and Machine Intelligence, 2010, 32(8): 1362-1376
[21] CAMPBELL N, VOGIATZIS G, HERNÁNDEZ C, et al. Using multiple hypotheses to improvedepth-maps for multi-view stereo [C]//10th European Conference on Computer Vision, 2008: 766-779
[22] JI M, GALL J, ZHENG H, et al. SurfaceNet: an end-to-end 3D neural network for multiview stereopsis[C]//2017 IEEE International Conference on Computer Vision, 2017: 2326-2334
[23] YU Z, GAO S. Fast-MVSNet: sparse-to-dense multi-view stereo with learned propagation and gauss-newton refinement[C]// IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020: 1946-1955
[24] DAI Y, ZHU Z, RAO Z, et al. MVS2: deep unsupervised multi-view stereo with multi-view symmetry[C]//2019 International Conference on 3D Vision, 2019: 1-8

Unsupervised 3D reconstruction method based on multi-view propagation

LUO Jingfeng[1], YUAN Dongli[1], ZHANG Lan[2], QU Yaohong[1], SU Shihong[1]
(1.School of Automation, Northwestern Polytechnical University, Xi'an 710072, China;
2. School of Cyber Engineering, Xidian University, Xi'an 710071, China)

Abstract: In this paper, an end-to-end deep learning framework for reconstructing 3D models by computing depth maps from multiple views is proposed. An unsupervised 3D reconstruction method based on multi-view propagation is introduced, which addresses the issues of large GPU memory consumption caused by most current research methods using 3D convolution for 3D cost volume regularization and regression to obtain the initial depth map, as well as the difficulty in obtaining true depth values in supervised methods due to device limitations. The method is inspired by the Patchmatch algorithm, and the depth is divided into n layers within the depth range to obtain depth hypotheses through multi-view propagation. What's more, a multi-metric loss function is constructed based on luminosity consistency, structural similarity, and depth smoothness between multiple views to serve as a supervisory signal for learning depth predictions in the network. The experimental results show our proposed method has a very competitive performance and generalization on the DTU, Tanks & Temples and our self-made dataset; Specifically, it is at least 1.7 times faster and requires more than 75% less memory than the method that utilizes 3D cost volume regularization.

Keywords: multi-view propagation; unsupervised; 3D reconstruction; Patchmatch algorithm; multi-metric loss function

引用格式:罗锦锋, 袁冬莉, 张蓝, 等. 基于多视图传播的无监督三维重建方法[J]. 西北工业大学学报,2024,42(1): 129-137
LUO Jingfeng, YUAN Dongli, ZHANG Lan, et al. Unsupervised 3D reconstruction method based on multi-view propagation [J]. Journal of Northwestern Polytechnical University, 2024, 42(1): 129-137 (in Chinese)