《计算机视觉》大作业

基于单目视觉的三维重建算法

实验小组成员 （学号+班级+姓名）	分工及主要完成任务	成绩
都一凡	M^3VSNet 自监督代码 小论文 登台展示 视频录制	
刘润迪	Monodepth2 自监督代码 小论文 视频录制 小程序前端	
许函嘉	colmap 传统方法稀疏重建 小论文 视频录制 小程序后端	

(上述工作为协调分工，贡献相同，不区分先后)

山东大学

2021 年 6 月

基于无监督学习的单目视觉三维重建

都一凡[1]，刘润迪[1]，许函嘉[1]

（1. 山东大学（威海）数学与统计学院，山东威海 264209）

摘 要： 随着三维重建技术的发展，多视角立体（MVS）的方法逐渐成熟，而在 MVSNet 之后，许多基于有监督学习的深度神经网络也得到了广泛应用。但是在三维重建领域存在一个显著的问题：用于训练的真实数据的标签如：点云、深度图等的获取需要耗费大量的成本。因此不借助数据标签的无监督学习方法开始逐渐得到人们的关注。本文梳理了无监督学习在三维重建中的一系列方法，包括单目视频、单目立体、多视角立体等，这些均属于单目视觉在三维重建中的应用。其中我们详细介绍了 M^{3}VSNet，它在 MVSNet 的基础上使用无监督的方法进行深度图估计。为了保证重建点云的完整性和鲁棒性，M^{3}VSNet 提出了一种结合了像素级损失和特征级损失的多度量损失函数，能同时兼顾细粒度的像素值和粗粒度的特征。除此之外，在深度图的精调方面，M^{3}VSNet 也引入了法线深度一致性纠正来对 MVSNet 进行改进。在标准数据集 DTU 上的进行实验的结果表明 M^{3}VSNet 在无监督学习领域达到了 SOTA 的效果，并且在准确性上超越了基于有监督学习的经典 MVSNet。在我们采集的数据集上，也得到了同样的结果。

关键词： 单目视觉；三维重建；深度学习；无监督学习；多视角立体

中图分类号：分类号 1；分类号 2 **文献标识码：**A **文章编号：**0372-2112（xxxx）xx-xxxx-xx *由编辑填写*

网络出版地址：http://www.ejournal.org.cn

DOI：10.12263/DZXB.xxxxxxxx *由编辑填写* 分类号

Monocular 3D Reconstruction Based On unsupervised learning

Yifan Du[1], Rundi Liu[1], Hanjia Xu[1]

(*1. Department of Mathematics and Statistics, Shandong University Weihai Campus, Weihai, Shandong Province , China;*

Abstract: With the development of 3D reconstruction technology, multi-view stereo (MVS) methods have gradually matured. After the proposal of MVSNet, many deep neural networks based on supervised learning have been widely used. However, there is a significant problem in the field of 3D reconstruction: the acquisition of labels such as: point clouds, depth maps, etc. used for training is costly. Therefore, unsupervised learning methods that do not rely on data labels are gaining attention. In this paper, we sort out a series of methods of unsupervised learning in 3D reconstruction, including monocular video, monocular stereo, multi-view stereo, etc., which are all applications of monocular methods in 3D reconstruction. Among them, we introduce M3VSNet in detail, which uses an unsupervised approach for depth map estimation based on MVSNet. To ensure the completeness and robustness of the reconstructed point cloud, M3VSNet proposes a multi-metric loss function that combines pixel-level loss and feature-level loss, which can take into account both fine-grained pixel values and coarse-grained features. In addition, M3VSNet also introduces a normal depth consistency correction to improve MVSNet in terms of fine-tuning the depth map. The results of experiments on the standard dataset DTU show that M3VSNet achieves SOTA results in the unsupervised learning domain and outperforms the classical MVSNet based on supervised learning in terms of accuracy. The same results are obtained on our collected dataset.

Key words: monocular; 3D reconstruction; deep learning; unsupervised learning; multi-view stereo

---

收稿日期：年-月-日；修回日期：年-月-日；责任编辑：某某 *投稿时不填写此项*

基金项目：基金中文完整名称(No.项目号，No.项目号)；基金中文完整名称(No.项目号)；…；基金中文完整名称(No.项目号)

1 引言

三维重建是指在二维图像信息中获得图像的信息，并通过对这些信息进行处理得到图像的三维信息的技术，它是计算机视觉的重要组成部分。近几年来，随着计算机视觉领域各个方向的快速发展，三维重建技术逐渐成为计算机领域的一个研究热点。三维信息也被广泛应用于不同的领域。比如，机器人领域的自动导航、计算机视觉中的识别物体、建筑学中通过高精度的三维重建来实现预知的建成效果、考古学中通过三维重建实现对文物的保护以及修复。因此，来自众多领域的需求推动着三维重建技术向易实现、高精度的方向发展。

目前，计算机视觉的估计方法主要有两类，分别为主动向目标发射光束的主动视觉法和通过摄像系统摄取二维图像的被动视觉法[1]。在被动视觉的估计方法中，若选用多目视觉方法，即利用多个机位拍摄目标物体的多幅图像，将增加测距系统的成本，也会增大其系统安装和图像采集等技术的要求。然而基于被动视觉法的单目视觉研究仅利用一台手持式数码摄像机，就可实现三维重建，具有结构简单、使用方便以及成本低廉等优点，且非常容易推广应用至不同领域，具有优秀的拓展性。与双目和多目视觉系统比较，单目视觉的方法大大的降低了系统工作量，节省了计算机处理时间。

2 背景与方法

2.1 监督学习

自 2014 年深度学习掀起第三次人工智能浪潮以来，基于有监督学习的深度学习算法受到了人们极大的关注，并在计算机视觉领域得到广泛应用。有监督学习只适用于训练集和测试集都已经经过人工标注，使得模型在这些标签的监督下进行损失函数的优化。

但是在三维重建领域，使用有监督学习进行三维重建存在如下缺点：

（1）对于数据真实标签的采集需要耗费高昂的成本，例如点云的采集需要三维激光扫描仪，深度图的采集需要能接受反射光的传感器等，这些都需要耗费大量的人力、物力。

（2）由于有监督学习的模型是在某个具体的训练集上训练的，因此它能够较好地拟合某个特定的数据分布。但是如果此时将模型用于其他场景，很可能出现重建失败的情况。

（3）有监督学习的目标是将输入映射成一个特定的输出，这其中很难理解中间过程是否学习到了高质量的向量表示。因此当最终重建结果缺乏完整性和鲁棒性时，较难找出有缺陷的地方做出改进。

2.2 无监督学习

随着深度学习技术的发展和应用的推广，越来越多的场景要求使用无标签的数据。与此同时，无监督的数据十分廉价易得，如文本、图片、语音等，可谓“取之不尽，用之不竭”。因此，将无监督学习的范式应用到三维重建中的方法应运而生。

无监督学习的基本思路是：根据源图片的不同视角，预测参考图片的视角。这二者之间通过一个深度图作为中间变量来连接。预测的参考图片和参考图片本身越接近，说明作为中间变量的深度图越能较好地代表源图片的信息。当模型训练好之后，将图片映射到深度图的编码器便可以用来生成深度图。

在三维重建领域，无监督学习通常在网络结构上会采取某一种有监督学习的框架结构，它们之间最大的不同在于训练方式上的不同。无监督学习并非真正意义上的“没有”监督，只是不再利用数据标签，而是利用数据本身作为监督信号，也可以认为是我们通常所说的“自监督”，二者之间在定义上并无显著差别。

2.3 单目视觉三维重建

利用单目图像进行三维重建已经成为计算机视觉领域的研究热点。单目视觉三维重建是指：使用单个摄像头捕获的一幅或多幅图像进行三维重建。它利用图像的二维特征（例如纹理、轮廓和阴影）获得物体在空间中的三维视觉信息，只需要一台相机就足够。从技术讲，该方法避免了立体对应，需要解决的主要问题是如何确定单个摄像头在两个不同时间点的空间转换关系。

该类方法主要分为基于传统算法（SfM）的三维重建和基于深度学习算法的三维重建，除此之外还有将传统算法与深度学习算法融合的三维重建方法。

传统的三维重建算法按照传感器是否主动向物体照射光源可以分为主动式和被动式，重建步骤包括特征点提取、配准、相机标定、数据融合等；

深度学习算法进行三维重建的步骤主要包括使用深度学习模型获取单个图像的深度，再根据相机参数从深度图重建点云。与传统方法相比，深度学习算法能够更轻松有效的地恢复场景的三维信息，三维模型也具有更好的真实性。

因此，基于单目视觉的三维重建任务就是采用单目摄像机根据摄像机的内参数来估算三维信息，或者采用图像匹配和摄像机自标定技术从序列图像中估算三维信息。

3 研究现状

三维重建技术作为计算机视觉的重要研究方向，经过国内外研究工作者多年的努力，已经取得了许多成果。

Pollefeys[2]等人的三维重建系统使用自带的单目摄像机绕着目标拍摄一系列的连续图像，通过处理这些图像来重建出三维模型，获取的图片越多重建出效果就越真实。但是该算法比较复杂，耗时较长。Snavely[3]等人提出了一种基于无序互联网图像序列的视觉重建方法，但是其并不能获得一个致密的三维重建结果。A Geiger[4]等人发明了一种基于双目立体视觉的实时三维重建方法，从而获得致密的三维重建结果，但是立体视觉系统结构比较复杂。

目前，国内在三维重建技术方面也进行了大量研究，在理论和算法方面都取得了一定进展。

CVsuite 是由中科院自动化研究所机器人视觉组研究并实现的一款软件，该软件使用方便，具有特征点提取、匹配、摄像机的自标定以及模型的三维显示功能，可以处理不同来源的二维图像。但其缺点也比较明显，它在实际操作过程中匹配效率不高、速度较慢。

上海交通大学计算机系马利庄教授于 2009 年在三维重建方面提出了一种基于构建 Visual Hull 求取物体表面形状及其反射属性的算法[5]。通过计算得到的物体表面以及反射参数来渲染三维重建的真实感。

微软研究院(Microsoft Research)在 2013 年推出了 Kinect Fusion 项目[6]，Kinect Fusion 利用 Kinect 传感器(深度相机)可以直接得到目标场景的深度图像数据，将传感器绕着目标物体或场景移动就可以获得不同角度下的多组深度图 像数据并实时地构建场景三维模型。

可以看出，国外工作者很早就对三维重建展开了研究，而我国相关研究起步较晚。由于我国今年在现代化生产的快速发展，三维重建具有的广泛应用价值。因此进行相关的三维重建研究尤为重要。

4 相关工作

4.1 自监督单目视频

使用自监督学习进行三维重建的一种方式来自于立体配对。即在训练过程中可以使用同步立体对，通过预测这对之间的像素差异，可以训练一个深度网络，从而在测试时进行单目深度估计。针对新视图合成问题，提出了一种计算离散深度的模型。通过预测连续的视差值扩展了这种方法[7]，通过包含左右深度一致性项产生了优于目前有监督方法的结果。基于立体视觉的方法已经被扩展到使用半监督的数据[8]，生成式对抗网络[9]，附加一致性[10]，时间信息[11]，以及用于实时使用[12]。

4.2 自监督单目立体几何

一种约束较少的自监督形式是使用单目视频，其中连续的时间帧提供训练信号。在这里，除了预测深度，网络还必须估计帧间的摄像机姿态，这在物体随着相机运动的情况下是一个巨大的挑战。估计出的摄像机姿态只需要在训练时使用，以帮助约束深度估计网络计算出的深度值。

在第一个将自监督学习用到单目视觉的方法中，[13]训练了一个深度估计网络和一个单独的姿态网络。为了处理非刚性场景运动，一个额外的运动解释掩模允许模型忽略违反刚性场景假设的特定区域。然而，他们后来的在线迭代模型禁用了这个掩模，获得了更好的性能。受[14]的启发，[15]提出了一个使用多个运动掩模的更复杂的运动模型。然而，这个模型没有在现有数据集上进行充分的评估，因此很难确定它的实用性如何。[16]还将运动分解为刚性和非刚性两个部分，利用深度和光流来解释物体运动。这改善了对流的估计，但他们报告说，在联合训练流和深度时，效果并没有得到改善。在光流估计的情况下，[17]表明它有助于明确地对遮挡进行建模。

4.3 多视角立体几何MVS

4.3.1 传统 MVS 方法

多视角立体领域已经提出了许多传统的方法，

如基于体素的方法[18]、特征点扩散方法[19]和深度图融合方法[20]。首先，基于体素的方法消耗大量计算资源，其精度主要依赖于体素的分辨率[21]。其次，在特征点扩散方法中，空白区域存在严重的纹理缺失问题。第三，最常用的方法是深度图的融合，它首先得到深度图，然后融合所有的深度图再得到最后的点云。此外，前人还提出了许多改进 MVS 的方法。Silvano[22]在 3D 空间中提出了一种斑点匹配算法，这种算法可以大规模并行运算，极大提高了计算效率。Johannes[23]同步估计深度和法线图，并使用光度先验和几何先验来改进基于图像的深度和法线融合。然而，在处理不理想环境如无纹理或纹理重复区域和非朗伯表面时，算法的鲁棒性有待提高。

4.3.2 有监督学习 MVSNet

自从 Yao Yao 在 2018 年提出 MVSNet 以来，许多基于 MVSNet 的有监督学习网络也纷纷诞生。为了降低 GPU 内存的消耗，Yao Yao 又提出了 R-MVSNet，将门控循环单元引入到 MVSNet 中。Gu 使用了级联的概念来缩小代价体素。Yi 引入了两种新的自适应的金字塔多尺度图像视图聚合来增强无纹理区域的点云。Luo 利用具有各向同性和各向异性三维卷积的平面扫描体得到了更好的结果[24]。在这类任务中，代价体素和 3D 正则化的计算过程对内存的消耗是巨大的。更重要的是，获取真实的深度图标签来训练一个有监督的网络是需要耗费大量人力物力的。

5 模型与算法

在这一节中，将详细介绍一种基于无监督学习的多视角立体几何的深度学习方法——M³VSNet。首先在 5.1 节描述网络的架构，这个网络用来生成初始深度图；然后在 5.2 节说明如何通过法线与局部表面切线的正交性得出深度一致的法线，对初始的深度图进行精调。最后，通过考虑匹配不同角度来提高点云重构的鲁棒性和完整性；5.3 节中详细介绍了 M³VSNet 所提出的多度量损失，这也是无监督学习和有监督学习最大的不同之处。

5.1 模型架构

M³VSNet 的基本架构由三部分组成——金字塔特征聚合、基于方差的体素生成和三维 U-Net 正则化，如图 1 所示。金字塔特征聚合从低层次到高层次依次提取特征，从而能够包含更多的上下文信息。然后使用与 MVSNet[25]相同的基于方差的代价体素生成和 3D U-Net 正则化来生成初始深度图。M3VSNet 的创新之处在于它的体系结构由两部分组成，即计算深度的一致性和多度量损失。在生成初始深度图之后，我们结合法线深度一致性来考虑法线和局部表面切线之间的正交性。更重要的是，该模型构建了多度量损失，包括像素层面的损失和特征层面的损失。

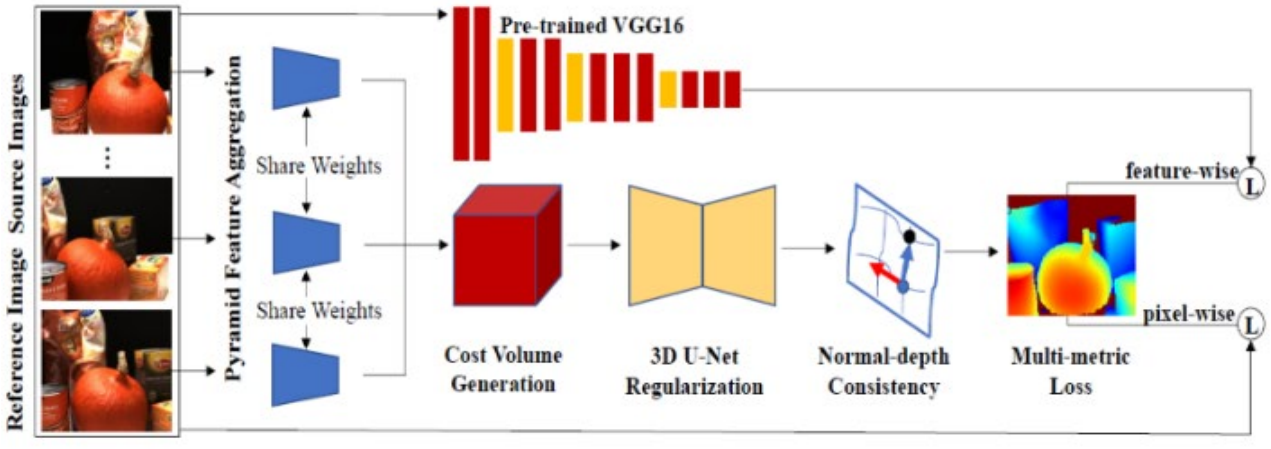

图 1 M³VSNet 架构

5.1.1 金字塔型特征提取

在先前的基于监督学习的网络中，如 MVSNet[25]，只采用了 1/4 特征(1/4 表示原始参考图像大小的 1/4)。1/4 特征缺少多尺度的上下文信息来匹配源图像和参考图像之间对应的像素。因此，为了对此进行改进，M³VSNet 使用金字塔形网络进行特征聚合，将不同尺度的特征与不同感受野[26]的上下文信息聚合在一起。图 2 描述了这个模块的详细结构。对于每一张输入图像，构建一个特征提

取网络，提取聚合后的 1/4 特征。在自底向上的过程中，将第 3、6、9 层的步幅设置为 2，得到 11 层 2 维 CNN 中的 4 个尺度特征。每个卷积层后面是 BatchNorm 和 ReLU 的结构。在从上到下的过程中，每一层的特征都是由较高层的上采样和同一层中通道较少的特征串联而成。为了减小最后 1/4 特征的维数，每个级联采用 1x 1 卷积。最后，我们得到了 32 个通道的最后一个特征，它是上下文信息从底层到高层的融合。

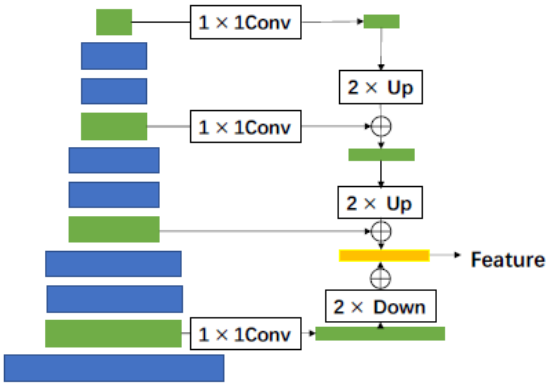

图 2 金字塔型结构

5.1.2 代价体素和 3D UNet 正则化

单应性变换是多视角匹配过程中的重要一步，它的作用是将物品从一个视角变换到另外一个视角。在最好的传统方法中，计算代价是使用参考图片与多个源图片通过两两配对进行计算的，一般会采用单应性变换差异的均值作为图像匹配的依据，此处 M³VSNet 也采用了和 MVSNet 相同的处理方式，用方差代替均值，能够更好地刻画视觉上的差异。此处在中的单应性变换也用到了 MVSNet 中不同的深度假设 D。

3D UNet 是用来对代价体素进行正则化的网络，在集成特征的时候十分有效。最终使用 soft argmin 操作得到初始的深度估计。

5.2 一致性法线纠正

由于初始深度仍然包含图片之间一些不正确的匹配关系，因此，为了进一步提高深度估计图的质量，M³VSNet 基于法线和局部表面切线[27]之间的正交性，加入了法线-深度一致性的纠正。这种纠正将对 3D 空间中不合理的深度估计值进行精调，使其变得更加合理。法线-深度一致性纠正可分为两个步骤：首先，通过深度值和各个像素深度值的正交性计算出某一点的法线。然后根据投影关系，通过法向深度和初始深度推断出纠正后的精细深度。该模块与 3D U-Net 配合，将二维和三维空间的深度细化，提高了深度的准确性和一致性。

通常对于一个像素点，我们选择与其相邻接的 8 个邻居来推断中心点的法线。由于邻居之间的正交性，此处应采用叉乘运算。对于每个中心点p_i，它的所有相邻点构成的集合记为p_{ix}和p_{iy}。如果p_i的深度Z_i和相机的内参数K已知，则法线N_i可以通过下式计算:

$$P_i = K^{-1} Z_i p_i \tag{1}$$

$$\tilde{N_i} = \vec{P_i P_{ix}} \times \vec{P_i P_{iy}} \tag{2}$$

公式（1）即为通过源图像的深度值和像素反解出参考图像的像素值，它与已知源图像和参考图像，求深度值互为反问题。公式（2）即为通过叉乘计算法线。为增加最终法线估计N_i的可信度，对 8 个邻域计算的法线取平均值来代表最终真正的法线:

$$N_i = \frac{1}{8}\sum_{i=1}^{8}\left(\tilde{N_i}\right) \tag{3}$$

当有法线深度图和初始深度图时，将二者结合，就可以得到最终的精细化深度图。对于每个像素$p_i(x_i, y_i)$，都需要对邻居$P_{neiahbor}$的深度经过一次精调，它们对应的 3D 坐标下的点是P_i和$P_{neiahbor}$。P_i的法线为$\vec{N}(n_x, n_y, n_z)$，深度是Z_i，$P_{neiahbor}$的深度是$Z_{neighbor}$。而且由于局部表面的正交性和表面一致性，显然可以得出$\vec{N} \perp \overrightarrow{P_i P_{neighbor}}$。综上所述，通过中心点的深度和法线可以推断出邻域的深度$Z_{neighbor}$:

$$\left(K^{-1} Z_i p_i - K^{-1} Z_{\text{neighbor}}\, p_{\text{neighbor}}\right) \begin{bmatrix} n_x \\ n_y \\ n_z \end{bmatrix} = 0 \tag{4}$$

在对深度进行精调时，M³VSNet 还考虑了中心点的 8 个邻居。考虑到一些边缘或不规则表面法线的不连续性，因此对参考图像I_i设置一个权重w_i，使得最终的深度更符合几何上的一致性。权重定义如下:

$$w_i = e^{-\alpha_1 |\nabla I_i|} \tag{5}$$

权重w_i取决于P_i和$P_{neiahbor}$之间的梯度，梯度越大表示细化深度的可靠性越低。针对一个中心点的 8 个邻居，最终细化深度$\tilde{Z}_{neighbor}$是 8 个不同方

向的加权和的组合。在此之后，还需要对深度在三维空间中进行正则化，从而提高估计深度图的精度和连续性：

$$\tilde{Z}_{\text{neighbor}} = \sum_{i=1}^{8} w_i' Z_{\text{neighbor}} \quad (6)$$

$$w_i' == \frac{w_i}{\sum_{i=1}^{8} w_i} \quad (7)$$

5.3 损失函数

M³VSNet 提出了一种新的多度量损失函数，不同于先前深度学习模型的损失函数，这个损失函数在考虑像素级别损失的同时，还考虑了特征级别的损失，这对于重建是非常关键的。像素级别的损失可以保证匹配对有更多的纹理细节，而特征级别的损失可以充分利用照片中的语义信息。

多度量损失函数的关键思想是多张视图间光度的一致性。给定参考图像I_{ref}和源图像I_{src}，对应拍摄这两张图像的摄像机的内参数分别表示为K_{ref}和K_{src}（通常情况下这二者相同）。同时，从I_{ref}到I_{src}的外参数变换记作T，即相机之间的姿态变换。那么对于I_{ref}中的像素$pi(xi, yi)$，它在 I_{src}中对应的像素$pi'(xi', yi')$可以通过下式计算:

$$p_i' = KT\left(K^{-1}\tilde{Z}_i p_i\right) \quad (8)$$

这也是图像和深度之间的反问题，与已知两点像素和内外参数矩阵求深度互为反解。而从参考图像到源图像的重叠区域I'_{src}可以利用双线性插值方法进行采样得到：

$$I'_{src} = I_{\text{src}}(p_i') \quad (9)$$

对于遮挡区域，I'_{src}中的像素值设为 0。显然，将p_i投影到I_{src}的外部区域可以得到掩模 M，即不在所有图像中都出现的区域。基于先前的约束，可以将多度量损失函数L表示为像素级损失L_{pixel}和特征级损失L_{feature}的加权和。

$$L = \sum\left(\gamma_1 L_{\text{pixel}} + \gamma_2 L_{\text{feature}}\right) \quad (10)$$

5.3.1 像素级别的损失

对于像素级损失，我们只考虑参考图像I_{ref}和其他源图像之间的光度一致性。这个损失函数主要有三个部分。首先，光度层面的损失用于比较I_{ref}和I'_{src}在像素值上的差异。为了减轻光照变化的影响，损失函数L_{photo}也考虑了每一个像素处的梯度：

$$L_{\text{photo}} = \frac{1}{m}\sum\left(\left(I_{ref} - I'_{src}\right) + \left(\nabla I_{ref} - \nabla I'_{src}\right)\right) \cdot M \quad (11)$$

式中，m 为掩模 M 中有效点个数之和。

其次，设置结构相似度损失(SSIM)函数 L_{SSIM}来衡量I_{ref}和I'_{src}之间的相似度。当I_{ref}与I'_{src}相同时，S 函数将为 1。这里的 S 函数可以用任意一种相似性度量函数来代替：

$$L_{SSIM} = \frac{1}{m}\sum\frac{1 - S\left(I_{ref}, I'_{src}\right)}{2} \cdot M \quad (12)$$

第三，考虑到模型输出的深度值可能不符合人眼的视觉感知，因此有必要对其进行平滑处理。最终的精细化深度图的平滑损失函数迫使一阶邻域和二阶邻域的梯度减小，从而使得生成的深度图更加平滑：

$$L_{\text{smooth}} = \frac{1}{n}\sum\left(e^{-\alpha_2|\nabla I_{\text{ref}}|}\left|\nabla\tilde{Z}_i\right| + e^{-\alpha_3|\nabla^2 I_{\text{ref}}|}\left|\nabla^2\tilde{Z}_i\right|\right) \quad (13)$$

式中，n 为参考图像I_{ref}中的点个数之和。

最后，将上述三个损失函数结合，得到总的像素级别的损失L_{pixel}如下:

$$L_{\text{pixel}} = \lambda_1 L_{\text{photo}} + \lambda_2 L_{SSIM} + \lambda_3 L_{\text{smooth}} \quad (14)$$

5.3.2 特征级别的损失

仅使用像素级损失函数的神经网络在无纹理和纹理有大量重复区域的情况下会出现图片之间的匹配错误，因为它只关注了图像在像素层面上这种细粒度的差异。因此除了像素级别的损失之外，M³VSNet 的主要改进之一是在损失函数中加入了特征层面的损失。就像在图像风格迁移中一样，像素级别的损失与感知层面的损失能够更好地提升风格迁移的效果[27]。特征层面的损失将使得模型能够学习到更多的语义层面的信息，从而能够匹配不同的图像中的语义特征。

由于估计的深度与 5.1 节中提到的金字塔特征网络之间具有很强的相关性，所以高层次的特征提取是用经过预训练的 VGG16，而不是金字塔特征网络来做的。通过预训练的 VGG16 网络，如图 3 所示，参考图像I_{ref}可以提取更多的高层面的语义信息来构建特征级别的损失函数。在这里，M³VSNet 提取了 8、15 和 22 层的输出，它们的尺寸分别是原始输入图像的一半、四分之一和八分之一。事实上，

之所以不用第三层的输出是因为它的尺寸与原始输入图像大小相同，这实际上是重复使用了像素级别的损失函数，并没有带来效果上的提升。

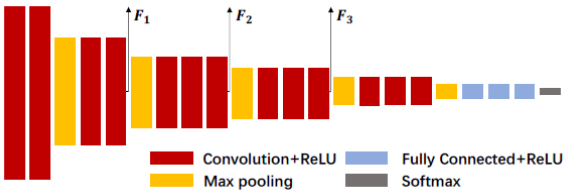

图 3 用预先训练的 VGG16 提取特征

对于 VGG16 中的每个特征，M³VSNet 基于跨多视图的概念来构造损失函数。类似于 3.3.1 节，可以通过求解反问题得到F_{src}中对应的像素p'。从F_{ref}变换到F_{src}对应的特征可以通过下式计算:

$$F'_{src} = F_{src}(p'_i) \qquad (15)$$

受人类视觉系统通过特征而不是单个像素来感知场景的启发，特征层面可以有更大的感受野。因此，可以在一定程度上缓解不理想区域对重建造成的负面影响。由于 VGG16 提取了丰富的语义信息，因此估计的最终深度可以反映出除了像素纹理值这种细粒度特征以外的特征相似性。VGG16 某一层的损失函数L_F可以表示成下式:

$$L_F = \frac{1}{m}\sum\left(F_{ref} - F'_{src}\right)\cdot M \qquad (16)$$

最后的特征层面的损失函数是不同尺度特征的加权和，例如：L_{F_i}表示预训练的 VGG16 的第i层的输出：

$$L_{\text{feature}} = \beta_1 L_{F_8} + \beta_2 L_{F_{15}} + \beta_3 L_{F_{22}} \qquad (17)$$

6 实验

6.1 在DTU数据集上的表现

DTU 数据集是一个多视图立体数据集，包含 124 个不同的场景，每个场景被机械臂携带的相机扫描了 49 次。我们使用与 MVSNet[25]和 MVS2[29]相同的训练集-测试集的分割方式。即，场景 1, 4, 9, 10, 11, 12, 13, 15, 23, 24, 29, 32, 33, 34, 48, 49, 62, 75, 77, 110，114，118 用于测试，其余用于训练。

6.1.1 实验细节

我们使用 pytorch 深度学习框架来搭建网络。在训练阶段，我们只使用 DTU 的训练集，没有任何真实的深度图作为标签。输入图像的分辨率是原始图像的经过裁剪得到的 640 × 512。由于金字塔特征聚合对原始图像的降维，因此最终深度的分辨率为 160 × 128。此外，生成的深度图的范围控制在 425mm 到 935mm，深度图的采样数 D 设置为 192。该模型的 batch_size 设置为 4 。通过使用 adam 优化器进行 10 个 epoch 的训练，初始学习率设置为 1e-3，每两个 epoch 衰减为原来的 0.5。为了平衡损失函数中各种不同的权值，我们设$\gamma_1 = 1$，$\gamma_2 = 1, \alpha_1 = 0.1, \alpha_2 = 0.5, \alpha_3 = 0.5, \lambda_1 = 0.8, \lambda_2 = 0.2, \lambda_3 = 0.067$。除此之外，$\beta_1 = 0.2, \beta_2 = 0.8, \beta_3 = 0.4$,。在每次迭代中，使用一个参考图像和两个源图像。在测试阶段，输入图像的分辨率为 1600 x 1200。

6.1.2 实验结果

M³VSNet 使用了一系列的官方指标[30]用于评估 M3VSNet 在 DTU 数据集上的性能。其中有三个度量标准分别是准确性、完整性和整体性。整体性即为准确性和完整性的平均值。为了证明 M3VSNet 的有效性，我们将 M3VSNet 与 Furu[31]、Tola[32]和 Colmap[23]这三种传统的经典方法进行了比较，并与 SurfaceNet[21]和 MVSNet 这两种经典的基于监督学习的方法进行了比较。以及其他两种基于无监督学习的方法，如 Unsup MVS[34]和 MVS2[29]。

如表 1 所示，M³VSNet 优于现有的两种基于无监督学习的方法[29]。M³VSNet 在所有指标上超过 Unsup_MVS[17，34]，在准确性和整体性上超过了 MVS2(除了完整性没有超过)。因此，M³VSNet 为多视点立体重建建立了最先进的无监督学习方法。此外，在相同的设置深度假设 D = 192 的情况下，M³VSNet 在点云重建的总体性能上超过了基于监督学习的 MVSNet[25]。与传统的 MVS 方法相比[23，32，35]，M³VSNet 在点云重建的完整性上有了明显的提高，整体质量优于 Furu 和 Tola。

表 1 不同算法在 DTU 数据集上的对比

算法	平均距离（mm）		
	准确度	完整度	总计
Furu	0.612	0.939	0.775
Tola	**0.343**	1.190	0.766
Colmap	0.400	**0.664**	**0.532**
SurfaceNet	0.450	1.043	0.746

MVSNet（D=192）	**0.444**	**0.741**	**0.592**
Unsup_MVS	0.881	1.073	0.977
MVS2	0.760	**0.515**	0.637
M3VSNet（D=192）	**0.636**	0.531	**0.583**

6.2 在自采数据集上的表现

我们使用手机对山东大学威海校区的四个标志性建筑（分别是图书馆、校训石、女孩雕塑和“航”雕塑）进行图像拍摄。其中，拍摄者手持手机对校训石、女孩雕塑、“航”雕塑进行围绕建筑物一周的连续拍摄，并从中选取 30 张不同角度的照片；由于图书馆的地理位置较为特殊，因此采取的拍摄手法是：拍摄者手持手机在图书馆前广场移动并进行连续拍照，从而获得图书馆约 120 度范围的正面拍摄图。

6.2.1 实验结果

同样地，使用 pytorch 深度学习框架搭建网络。首先将自采图片裁剪为 1600 × 1200 的大小并统一进行编号，然后将处理过的自采图片集放入训练好的模型输出结果，最后可以得到每张图片的深度图和融合后的点云文件，效果如下所示。

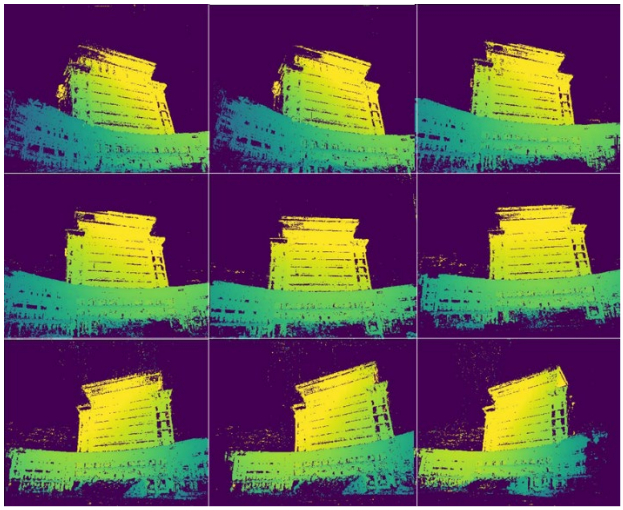

图 4 山东大学威海校区图书馆深度图

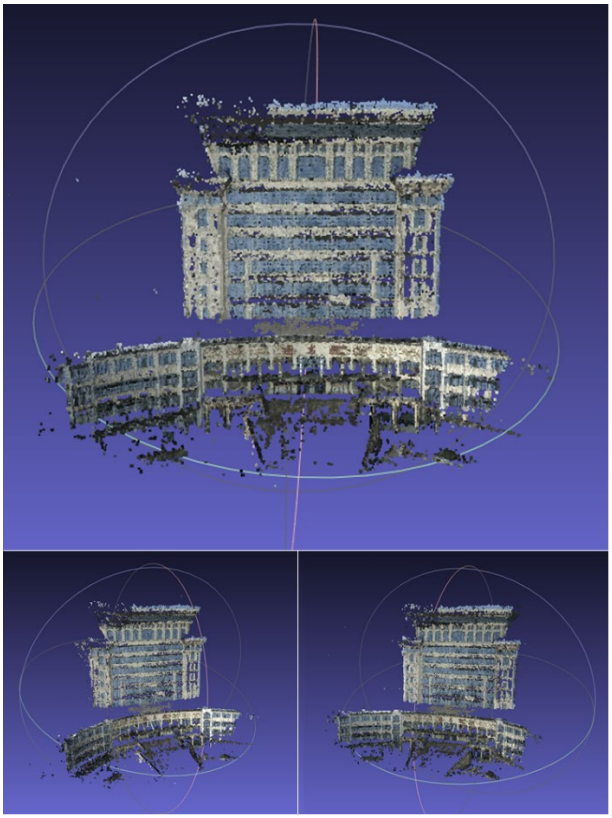

图 5 山东大学威海校区图书馆点云图

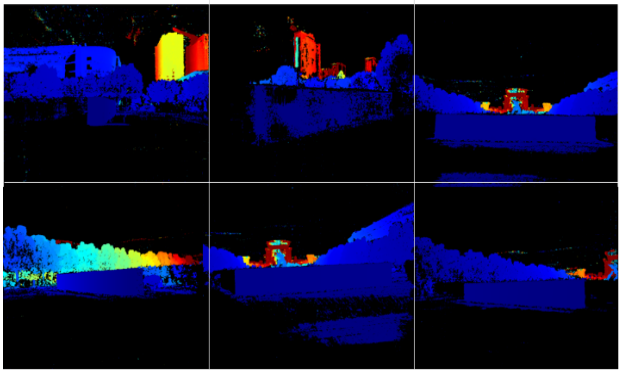

图 6 山东大学威海校区校训石深度图

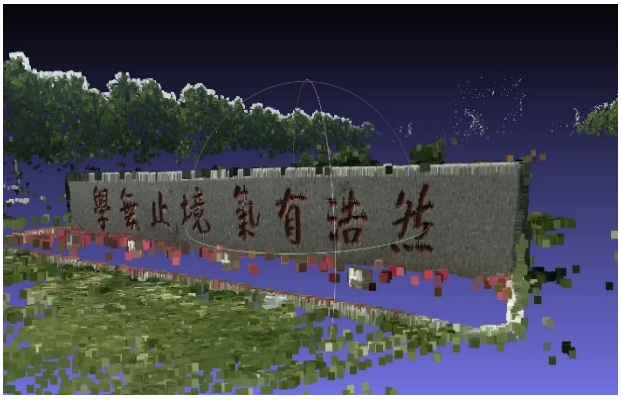

图 7 山东大学威海校区校训石点云图

6.2.2 复杂条件下的实验效果

我们在强太阳光下对不规则、具有弱纹理的

“航”雕塑进行拍摄，将图像输入模型，结果显示我们仍然可以得到良好的效果。

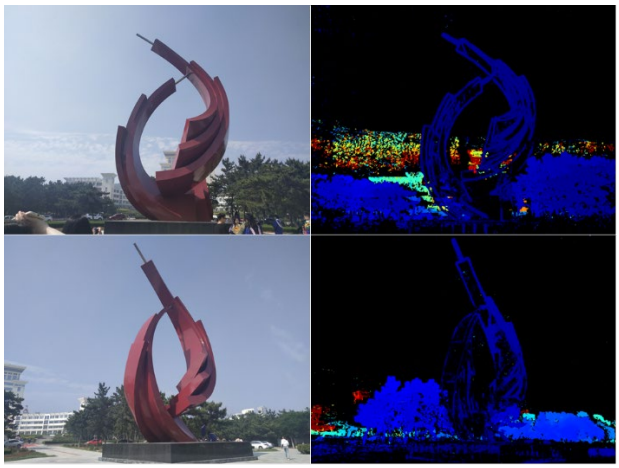

图 8 “航”雕塑拍摄图片与深度图

6.2.3 结果对比

在算法的对比方面，我们选取了 MVSNet，M³VSNet，以及传统的 colmap 方法进行对比，具体的评价指标已经在表 1 中列出，由于我们自采的数据集没有激光扫描仪等设备采集真实的深度图，因此无法从量化的指标上对其进行评估，只能从视觉效果上进行简单评估。深度估计结果如下所示：

我们选择了背景较为杂乱的一个雕塑进行深度估计，来检测模型的鲁棒性和泛化能力。可以看出，传统方法（图 a）在自采图片上的仍然是最好的，MVSNet（图 b）和 M³VSNet（图 c）在视觉效果上难以区分，从细节上来看 M³VSNet 略胜一筹。

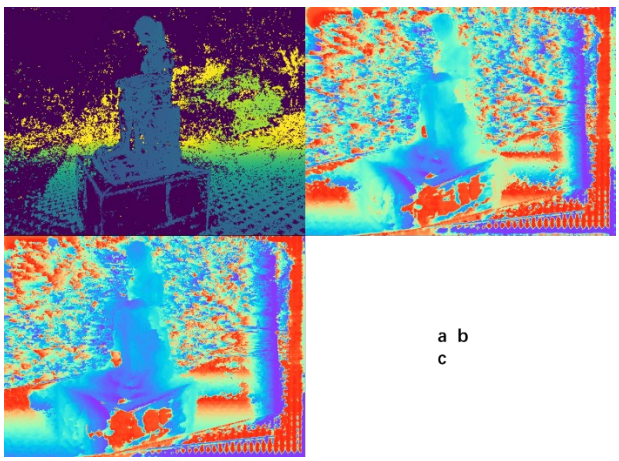

图 9 杂乱背景下图片与深度图

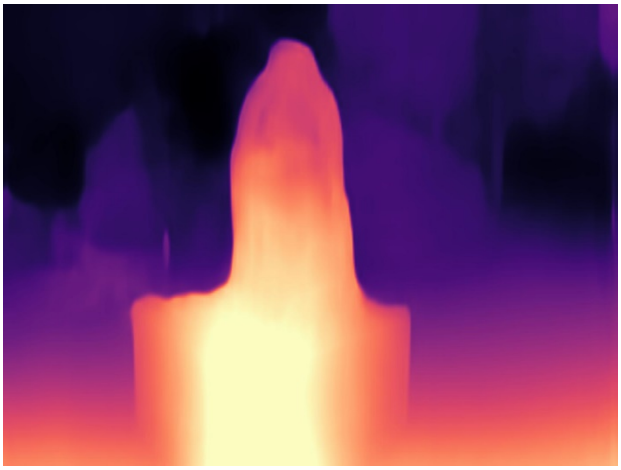

图 10 monodepth2 拍摄图片与深度图

单独将单张图片的模型 monodepth2 提取出来观察，可以发现由于该模型只使用一张图片对建筑进行重建，因此无法捕捉到细节特征，只能对建筑的轮廓进行重建， 同时也是速度最快，占用内存最小，最适合部署在边缘设备上的模型。

如表 2 所示，我们对不同的算法的运行时间进行了简单对比。可以看出，M³VSNet 和 MVSNet 在时间上并无太大差别，但是传统方法 Colmap 耗费的时间是这二者的几十倍，精准的深度估计以耗费大量的时间为代价。

表 2 不同算法在 DTU 数据集上的对比

算法	时间（张/s）
M³VSNet	2
MVSNet	3
Colmap	120
Monodepth2	2

7 结论

通过对模型的一系列讨论及实验对比，最后关于无监督学习的方法我们可以得出如下结论：

1) 无监督的网络架构适用于一些背景杂乱、弱纹理、强光等不理想的场景，尤其是在数据缺乏真实标签时，能够更好地学习数据本身的向量表示。
2) 使用了多尺度的损失函数，无监督学习可以同时兼顾 pixel-wise 的特征和 feature-wise 的特征，更符合人类的视觉感知，重建效果更好。

3) 由于模型在学习时不借助于标签，而只需要数据本身的像素值和特征，因此模型具有更强的泛化能力。

参考文献

[1] 江静，张雪松. 基于计算机视觉的深度估计方法[J]. 光电技术应用. 2011，26(1)：51-55.
Jiang J, Zhang Xuesong. Depth estimation method based on computer vision [J]. Applications of optical technology. 2011, 26(1): 51-55.

[2] Pollefeys, M., Van Gool, L. and Vergauwen, M.(2004)Visual Modeling with a Hand-Held Camera. International Journal of Computer Vision,59,207-232.

[3] Snavely, N., Seitz, S.M. and Szeliski, R.(2008)Modeling the World from Internet Photo Collections. International Journal of Computer Vision,80,189-210.

[4] Geiger,A.,Ziegler, J. and Stiller, C.(2011) StereoScan: Dense 3d Reconstruction in Real-Time. Intelligent Vehicles Symposium,Baden-Baden,5-9 June 201,963-968.

[5] Zheng Z Y, Ma L Z, Li Z, Chen Z H. Reconstruction of shape and reflectance properties based on the visual hull. In: Proc of the Int Conf on Computer Graphics and Virtual Reality. Las Vegas,2009,29-32

[6] Newcombe, R.A., Izadi, S. and Hilliges, O.(2013)Kinect Fusion: Real-Time Dense Surface Mapping and Tracking Mixed and Augmented Reality,233,430-436.

[7] Michael Goesele, Brian Curless, and Steven M Seitz. 2006. Multi-view stereo revisited. In 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06), Vol. 2. IEEE, 2402–2409.

[8] Yevhen Kuznietsov, Jorg St¨ uckler, and Bastian Leibe. Semi- ¨supervised deep learning for monocular depth map prediction. In CVPR, 2017.

[9] Yue Luo, Jimmy Ren, Mude Lin, Jiahao Pang, Wenxiu Sun,Hongsheng Li, and Liang Lin. Single view stereo matching. In CVPR, 2018.

[10] Filippo Aleotti, Fabio Tosi, Matteo Poggi, and Stefano Mattoccia. Generative adversarial networks for unsupervised monocular depth prediction. In ECCV Workshops, 2018.

[11] Matteo Poggi, Fabio Tosi, and Stefano Mattoccia. Learning monocular depth estimation with unsupervised trinocular assumptions. In 3DV, 2018.

[12] Ruihao Li, Sen Wang, Zhiqiang Long, and Dongbing Gu. UnDeepVO: Monocular visual odometry through unsupervised deep learning. arXiv, 2017.

[13] Tinghui Zhou, Matthew Brown, Noah Snavely, and David Lowe. Unsupervised learning of depth and ego-motion from video. In CVPR, 2017

[14] Arunkumar Byravan and Dieter Fox. Se3-nets: Learning rigid body motion using deep neural networks. In ICRA, 2017.

[15] Sudheendra Vijayanarasimhan, Susanna Ricco, Cordelia Schmid, Rahul Sukthankar, and Katerina Fragkiadaki. SfMNet: Learning of structure and motion from video. arXiv, 2017.

[16] Zhichao Yin and Jianping Shi. GeoNet: Unsupervised learning of dense depth, optical flow and camera pose. In CVPR, 2018.

[17] Joel Janai, Fatma Guney, Anurag Ranjan, Michael Black, ¨and Andreas Geiger. Unsupervised learning of multi-frame optical flow with occlusions. In ECCV, 2018.

[18] Sudipta N Sinha, Philippos Mordohai, and Marc Pollefeys. 2007. Multi-view stereo via graph cuts on the dual of an adaptive tetrahedral mesh. In 2007 IEEE 11th International Conference on Computer Vision. IEEE, 1–8.

[19] Yasutaka Furukawa and Jean Ponce. 2007. Accurate, Dense, and Robust MultiView Stereopsis. 2007 IEEE Conference on Computer Vision and Pattern Recognition(2007), 1–8.

[20] Connelly Barnes, Eli Shechtman, Adam Finkelstein, and Dan B Goldman. 2009. PatchMatch: A randomized correspondence algorithm for structural image editing. In ACM Transactions on Graphics (ToG). ACM, 24

[21] Mengqi Ji, Juergen Gall, Haitian Zheng, Yebin Liu, and Lu Fang. 2017. Surfacenet: An end-to-end 3d neural network for multiview stereopsis. In Proceedings of the IEEE International Conference on Computer Vision. 2307–2315.

[22] Silvano Galliani, Katrin Lasinger, and Konrad Schindler. 2015. Massively parallel multiview stereopsis by surface normal diffusion. In Proceedings of the IEEE International Conference on Computer Vision. 873–881.

[23] Johannes L Schönberger, Enliang Zheng, Jan-Michael Frahm, and Marc Pollefeys. 2016. Pixelwise view selection for unstructured multi-view stereo. In European Conference on Computer Vision. Springer, 501–518.

[24] Keyang Luo, Tao Guan, Lili Ju, Haipeng Huang, and Yawei Luo. 2019. P-mvsnet:Learning patch-wise matching confidence aggregation for multi-view stereo. In Proceedings of the IEEE International Conference on Computer Vision. 10452–10461.

[25] Yao Yao, Zixin Luo, Shiwei Li, Tian Fang, and Long Quan. 2018. Mvsnet: Depth inference for unstructured multi-view stereo. In Proceedings of the European Conference on Computer Vision (ECCV). 767–783.

[26] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. 2017. Feature pyramid networks for object detection. In Proceedings of the IEEE conference on computer vision and pattern recognition. 2117–2125.

[27] Zhenheng Yang, Peng Wang, Wei Xu, Liang Zhao, and Ramakant Nevatia. 2018. Unsupervised learning of geometry from videos with edge-aware depth-normal consistency. In Thirty-Second AAAI Conference on Artificial Intelligence

[28] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. 2016. Perceptual losses for realtime style transfer and super-resolution. In European conference on computer vision. Springer, 694–711.

[29] Yuchao Dai, Zhidong Zhu, Zhibo Rao, and Bo Li. 2019. MVS2: Deep Unsupervised Multi-View Stereo with Multi-View Symmetry. In 2019 International Conference on 3D Vision (3DV). IEEE, 1–8.

[30] Rasmus Jensen, Anders Dahl, George Vogiatzis, Engin Tola, and Henrik Aanæs. 2014. Large scale multi-view stereopsis evaluation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 406–413.

[31] Yasutaka Furukawa and Jean Ponce. 2007. Accurate, Dense, and Robust MultiView Stereopsis. 2007 IEEE Conference on Computer Vision and Pattern Recognition (2007), 1–8.

[32] Engin Tola, Christoph Strecha, and Pascal Fua. 2011. Efficient large-scale multiview stereo for ultra high-resolution image sets. Machine Vision and Applications 23 (2011), 903–920.

[33] Johannes L Schönberger, Enliang Zheng, Jan-Michael Frahm, and Marc Pollefeys. 2016. Pixelwise view selection for unstructured multi-view stereo. In European Conference on Computer Vision. Springer, 501–518.

[34] Tejas Khot, Shubham Agrawal, Shubham Tulsiani, Christoph Mertz, Simon Lucey, and Martial Hebert. 2019. Learning Unsupervised Multi-View Stereopsis via Robust Photometric Consistency. arXiv preprint arXiv:1905.02706 (2019).

[35] Jan-Michael Frahm, Pierre Fite-Georgel, David Gallup, Tim Johnson, Rahul Raguram, Changchang Wu, Yi-Hung Jen, Enrique Dunn, Brian Clipp, Svetlana Lazebnik, et al. 2010. Building rome on a cloudless day. In European Conference on Computer Vision. Springer, 368–381.