基于 Unity3D 三维多视角虚拟数据集构建

郑义桀,罗健欣,陈卫卫,潘志松,张艳艳,孙海迅
(陆军工程大学 指挥控制工程学院,江苏 南京 210007)

摘 要:基于深度学习的多视角三维重建(Multi View Stereo,MVS)是计算机视觉领域的研究热点。但构建高质量的多视角三维重建数据集需要消耗大量时间、人力和财力成本,因此当前可直接应用于多视角三维重建的数据集相对较少。为了降低数据集制作成本、提高制作效率,文章提出了一种有效的虚拟世界仿真现实世界的方法。通过 Unity3D 虚拟引擎,融合域适应和域随机方法,搭建三维虚拟场景,自动高效生成三维多视角虚拟数据(相机图像、相机参数和场景深度图),在此基础上构建了多视角三维重建虚拟数据集 Visual DTU。实验结果表明,使用虚拟数据集可大幅降低数据集制作的经济和时间成本,且基本能取得与采用真实数据集训练相同的三维重建效果;通过增加虚拟数据集训练样本或混合虚拟数据集和真实数据集进行模型训练,可进一步提升模型性能。

关键词:计算机视觉;三维重建;Unity3D;虚拟数据集;DTU
中图分类号:TP391.9 **文献标识码**:A **文章编号**:1673-629X(2023)05-0173-07
doi:10.3969/j.issn.1673-629X.2023.05.026

3D Multi-view Virtual Dataset Construction Based on Unity3D

ZHENG Yi-jie,LUO Jian-xin,CHEN Wei-wei,PAN Zhi-song,ZHANG Yan-yan,SUN Hai-xun
(School of Command and Control Engineering,Army Engineering University,Nanjing 210007,China)

Abstract:Multi View Stereo (MVS) based on deep learning is a hot research topic in computer vision field. However,it takes a lot of time,manpower and financial cost to construct high-quality multi-view 3D reconstruction data sets,so there are relatively few data sets that can be directly applied to multi-view 3D reconstruction at present. In order to reduce data set production costs and improve production efficiency,we put forward an effective method of virtual world simulating real world. Through Unity3D virtual engine,domain adaptation and domain randomization methods are integrated to build 3D virtual scenes and automatically and efficiently generate 3D multi-view virtual data (camera image,camera parameters and scene depth map). On this basis,Visual DTU is constructed for multi-view 3D reconstruction virtual data set. The experimental results show that using virtual data sets can greatly reduce the economic and time cost of data set making,and basically achieve the same effect of 3D reconstruction as using real data sets. The model performance can be further improved by adding training samples of virtual data sets or mixing virtual data sets and real data sets for model training.

Key words:computer vision;three-dimensional reconstruction;Unity3D;virtual data set;DTU

0 引 言

数据集在深度学习模型的训练和测试中起着重要的作用,但现实世界的数据集制作需要耗费大量时间和精力,深度学习多视角三维重建(Multi View Stereo,MVS)的数据集相比于二维图像需要的信息更多,通常需要增加相机姿态和场景深度,制作也更为复杂。

当前已有一些三维重建有关的数据集,主要来源于现实世界,但一些不包含相机标定或深度信息,可作为深度学习 MVS 训练的数据集并不多,目前常用的主要有 DTU[1]、ETH3D[2] 和 Tanks and Temples[3] 等。这些数据集为获取高精度 3 维场景图像和 3 维点云,制作均有较高的硬件要求。例如,3 种数据集均采用专业高清单反相机拍摄场景图像;ETH3D 和 Tanks and Temples 使用亚毫米级激光扫描仪获取场景深度;DTU 在一个 6 轴工业机器臂上安装了一个相机和一个结构光扫描仪,同时获取图像、深度图和相机姿态。数据集制作也消耗大量人力和时间成本。例如,在 Tanks and Temples 制作中,每个场景必须进行多次扫描才能密集覆盖表面。简单的物体从 4 个位置扫描,如小雕像。中型结构物体从 8 到 10 个位置进行扫描,

收稿日期:2022-07-14 **修回日期**:2022-11-16
基金项目:国家自然科学基金(62076251)
作者简介:郑义桀(1991-),男,硕士,助理工程师,研究方向为计算机视觉、SLAM、三维重建;通信作者:罗健欣(1984-),男,博士,讲师,研究方向为计算机视觉、计算机图形学、网络与多媒体通信。

如火车;最大的室外场景从 14 个和 17 个位置进行扫描,如宫殿和寺庙,尤其宫殿数据集的数据采集持续了两天。且后期还需要进行的点云对齐和优化。

近年来计算机图形学的发展使得虚拟图像设计更加真实。各种渲染引擎的出现,例如 UE4,Unity3D 和 CryEngine,使得逼真的虚拟图像的获取更加方便可行。不需要过高配置的硬件和大量的人力、时间投入,使得虚拟数据集的研究愈加火热,也产生了各种虚拟数据集,例如:用于自动驾驶 Virtual KITTI[4]、用于对象检测 ParallelEye[5]、用于语义分割 SYNTHIA[6]等,大多取得了不错的效果。

大量研究证明了虚拟数据集的可行性,且制作了许多虚拟数据集,但当前还没有专门应用于深度 MVS 的虚拟数据集。受上述研究启发,该文基于 Unity3D 提出了一种自动生成三维多视角虚拟数据的方法,并制作了虚拟数据集 Visual DTU,通过实验证明了该数据集的有效性。主要贡献有:

(1)提出了一种使用虚拟世界仿真现实世界的方法。融合域适应和域随机方法,通过 Unity3D 引擎搭建虚拟三维场景,设置虚拟相机,自动控制虚拟相机位置和方向的变换以获取三维多视角虚拟数据,主要包括相机图像、相机参数及场景深度图。

(2)提供了基于 Unity3D 引擎制作的可用于深度 MVS 模型训练的虚拟数据集 Visual DTU 及详细制作过程。该数据集包含了 128 组图像,每组图像包含了 49 个相机视角和 7 种光照强度变化,及每个相机视角的内外参数和每个视角的深度图。

(3)通过实验验证了 Visual DTU 的有效性,证明了使用虚拟引擎生成的高质量相机图像、场景深度图和相机参数可用于深度学习多视角三维重建模型的训练,同时可大幅降低了数据集制作时间和成本。将该数据集应用于当前主流的几种深度 MVS 模型,如:CVP - MVSNet[7]、M3VSNet[8]、PatchmatchNet[9]、JADCS-MS[10]等。证明其训练效果基本与真实数据集相当,并在增加训练样本或混合真实数据集和虚拟数据集的情况下可进一步提高模型性能。

1 MVS 数据集研究发展

1.1 真实数据集

Seitz 等[11]提出的多视角数据集是最早的 MVS 评估数据集,它仅包含两个具有低分辨率图像和校准相机的室内物体。Strecha 等[12]获取建筑立面的真实模型,并为 MVS 评估提供高分辨率图像和真实点云。为了评估不同光照条件下的算法性能,DTU 数据集用固定的相机轨迹获取了 128 个室内小物体的图像和点云。点云被进一步三角化成网格模型,并被渲染成不同的视点,以生成相机深度图。当前基于深度学习的 MVS 网络通常使用 DTU 数据集进行训练。Tanks and Temples 使用高清摄像机和激光扫描仪获取室内外不同大小的物体和场景,然而,它们的训练集只包含 7 个具有真实点云的场景。ETH3D 包含一个低分辨率集和一个高分辨率集,涵盖了从自然场景到人造的室内和室外环境不同的视角和场景类型,也是第一个涵盖了手持移动设备的重要使用情况,但是类似 Tanks and Temples,ETH3D 只为网络训练提供了少量的真实数据。Yao 等[13]为提高 MVS 模型泛化性,制作了 BlendedMVS 数据集,其中包含了超过 17 000 张高分辨率图像,涵盖了各种场景,包括城市、建筑、雕塑和小型物体。

1.2 虚拟数据集

相比于真实数据集的高人力和财力需求,虚拟数据集制作更加简单高效,因此近年来使用虚拟数据集训练和测试深度神经网络的做法越来越流行。

当前已有一些虚拟数据集,主要应用于目标检测[14-17]、自动驾驶[4,18]和语义分割[6,19-20]等方面,都取得了不错的效果。由于虚拟数据集与真实数据集之间存在域差,使得虚拟数据集训练的模型在真实数据集的测试不易取得很好的性能。为了提高模型训练效果,目前主要分为两种方法:

一是域适应,即让虚拟数据更接近真实数据。Nogues 等[21]利用 CycleGAN 网络对源域图像进行处理,使其与目标域图像外观上更加接近,这样使用风格转换之后的虚拟数据训练出的网络可以在真实数据集上取得更好的性能。Chen 等[22]利用梯度反转层对 Faster R-CNN 中的图像级特征和实例特征进行对抗训练处理,使得两个特征在源域和目标域的差距尽可能小。Saito 等[23]则通过对抗训练方法对全局特征进行弱对齐,对局部特征进行了强对齐,进而实现 Faster R-CNN 的域适应。

二是域随机。域随机[24]用一种相反的思路,不刻意追求虚拟数据的真实性,而是通过以非照片真实感的方式随机扰动环境,故意放弃照片真实感,迫使网络学习关注图像的基本特征。Tremblay 等[25]提出了一个利用合成图像训练深度神经网络进行目标检测的系统。为了处理真实世界数据中的可变性,系统依赖于域随机技术,其中模拟器的参数(如照明、姿势、对象纹理等)以非真实方式随机化。结果显示域随机不仅优于更真实的照片级数据集,而且改善了仅使用真实数据获得的结果。在车辆识别和状态估计实验中[26],将强度响应、模糊、噪声三种因素随机化引入到虚拟图像中,通过特殊的场景车辆布置方法和丰富的背景图片来保证虚拟图片的丰富变化性,以 2D 包围盒检测

作为实例验证虚拟数据集的性能,结果显示该虚拟数据集相比其他虚拟数据集有着明显优势。

2　虚拟场景仿真及数据集生成

在近几年大部分著名的深度学习 MVS 研究中,均采用了 DTU 数据集进行模型训练,因此该文参照 DTU 数据集设置虚拟世界场景并制作虚拟数据集 Visual DTU。DTU 数据集包含了 128 组场景,每组场景包含了 49 个相机视角及 7 个不同光照的 343 张图像,以及对应的相机参数和深度图。

如图 1 所示,首先通过 Unity3D 设置虚拟场景,而后控制相机视角,同步生成多视角相机图像、相机深度图和相机参数。再对生成的相机图像进行中心剪裁、深度图格式转换、相机参数的坐标系转换,最后组合成虚拟数据集 Visual DTU。

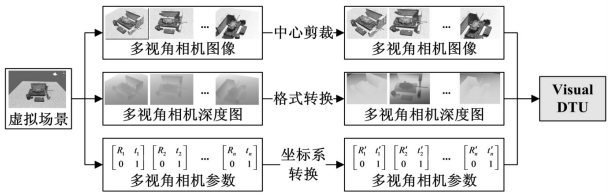


图 1　Visual DTU 生成流程

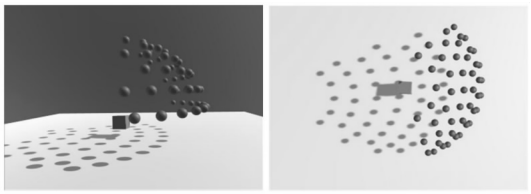
2.1　虚拟仿真场景设置

Visual DTU 参照 DTU 数据集格式制作,DTU 数据集中共使用了 49 个相机位置,对应 49 个相机参数文件,每个相机参数文件包含 1 个 4 × 4 外参数矩阵和 1 个 3 × 3 内参数矩阵。由公式(1)可计算出各个相机在世界坐标系中的位置 o_{cam}^{w} 。

$$o_{\mathrm{cam}}^{w} = -\boldsymbol{R}^{\mathrm{T}}\boldsymbol{t} \tag{1}$$

其中, $\boldsymbol{R}$ 为旋转矩阵, $\boldsymbol{t}$ 为平移向量。

计算发现 49 个相机基本部署在同一球面上,因此可拟合出球心坐标,作为物体放置的位置。图 2 展示了相机和物体位置关系。



(a)侧视图　　(b)顶视图

图 2　虚拟场景中相机(球体)和物体(立方体)位置

Visual DTU 共 128 组场景,使用已构建好的人物、动物、车辆、建筑、家具摆件等各类 3D 模型。在设置不同的场景时,分别采用了域适应和域随机的方法。

为实现域适应,部分场景设置采用高分辨率高仿真性物体模型,以尽可能仿真真实世界,图 3 展示了 4 组真实物体和虚拟物体的对比,每组图像中,左侧为真实物体,右侧为虚拟物体。同时每个相机视角均采集 7 种不同光照强度和方向的图像,以增强网络模型对不同光照条件下的适应性。图 4 展示了不同光照下的场景设置。

图 3　真实物体与虚拟物体对比

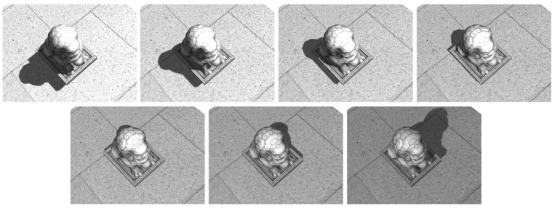


图 4　不同光照强度和方向的场景设置

为实现域随机效果,部分场景采用低分辨率低仿真性的简易物体模型,或采用非真实卡通造型,或在场景中随机加入一些三维物体,或者随机改变物体颜色,以强化网络模型学习图像的基本特征。同时,采用域随机的方式可以无限增加数据集,提高数据集生成速度。图 5 展示了部分场景中的域随机效果。

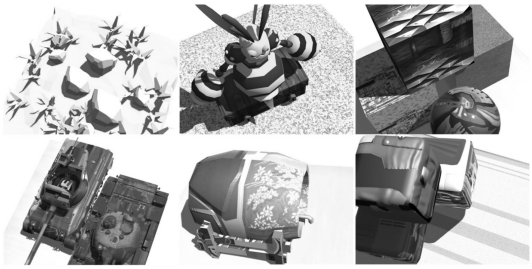


图 5　域随机场景

2.2　三维多视角虚拟数据生成

该文通过控制虚拟世界中相机位置和方向,获得三维多视角虚拟数据,主要包括不同相机视角的图像、相机参数和深度图。图像和深度图的生成方法采用文献[27]提供的函数“ImageSynthesis”,该函数通过相机挂载可直接生成任意尺寸的相机图像(该文设置 1 600× 1 200 尺寸)和以单通道灰度图表示的深度图(与相机图像尺寸相同)。

相机外参矩阵可由 C#库函数“matrix. SetTRS”获得,但有几点区别:(1)该函数采用相机坐标系转换为世界坐标系的外参矩阵,而深度 MVS 模型采用世界坐标系转换为相机坐标系的外参矩阵;(2)Unity3D 使用左手坐标系,而深度 MVS 模型使用右手坐标系;(3)Unity3D 中使用米(m)作为长度单位,而深度 MVS 模型使用毫米(mm)作为长度单位。因此外参矩阵需要转换。

原外参矩阵为：

$$\begin{bmatrix} \boldsymbol{R} & \boldsymbol{t} \\ 0 & 1 \end{bmatrix} = \begin{bmatrix} r_{00} & r_{01} & r_{02} & t_0 \\ r_{10} & r_{11} & r_{12} & t_1 \\ r_{20} & r_{21} & r_{22} & t_2 \\ 0 & 0 & 0 & 1 \end{bmatrix} \tag{2}$$

转换后的外参矩阵为：

$$\begin{bmatrix} \boldsymbol{R}' & \boldsymbol{t}' \\ 0 & 1 \end{bmatrix} = \begin{bmatrix} r_{00} & -r_{10} & r_{20} & t_0' \\ -r_{01} & r_{11} & -r_{21} & t_1' \\ r_{02} & -r_{12} & r_{22} & t_2' \\ 0 & 0 & 0 & 1 \end{bmatrix} \tag{3}$$

其中：

$$t_0' = -(t_0 \times r_{00} + t_1 \times r_{10} + t_2 \times r_{20}) \tag{4}$$

$$t_1' = t_0 \times r_{01} + t_1 \times r_{11} + t_2 \times r_{21} \tag{5}$$

$$t_2' = -(t_0 \times r_{02} + t_1 \times r_{12} + t_2 \times r_{22}) \tag{6}$$

相机内参数需单独设定，可直接在 Unity3D 中设置。内参矩阵一般表示为：

$$\boldsymbol{K} = \begin{bmatrix} f/dx & 0 & u_0 \\ 0 & f/dy & v_0 \\ 0 & 0 & 1 \end{bmatrix} \tag{7}$$

其中，f 代表相机焦距，单位毫米；$1/dx$ 代表 x 轴方向 1 毫米内的像素个数；$1/dy$ 代表 y 轴方向 1 毫米内的像素个数；(u_0, v_0) 代表图像原点位置。由于虚拟相机图像和相机传感器的宽和高可以直接设置，因此相机内参矩阵可由公式(8)计算得出。

$$\boldsymbol{K} = \begin{bmatrix} fW_{\text{pic}}/W_{\text{sen}} & 0 & W_{\text{pic}}/2 \\ 0 & fH_{\text{pic}}/H_{\text{sen}} & H_{\text{pic}}/2 \\ 0 & 0 & 1 \end{bmatrix} \tag{8}$$

其中，W_{pic} 和 H_{pic} 代表相机图像的宽和高，单位像素；W_{sen} 和 H_{sen} 代表相机传感器的宽和高，单位毫米。

2.3 多视角三维重建虚拟数据集构建

为能适用当前大多数深度 MVS 模型的训练需要，按照文献[28]中图像的处理方法，将原始 1 600×1 200尺寸的相机图像降采样为 800×600 尺寸的图像，然后进行中心裁剪得到640×512 尺寸的图像。与图像大小变化相适应，相机外参数保持不变，内参数减小 1 倍。

深度图根据不同的深度 MVS 模型训练需求进行相应降采样和剪裁，但初始深度图是单通道灰度图，每个像素点对应 1 个 0 ~ 255 的灰度值，因此需要将其转化为真实深度值。通过公式(9)可进行灰度值与深度值的转化。

$$\text{depth} = \frac{V_{\text{grey}}}{255(\text{far} - \text{near})} + \text{near} \tag{9}$$

其中，V_{grey} 为像素灰度值，far 和 near 分别为相机的远、近剪裁平面的距离。为方便计算，该文设置为 far = 10，near = 0.01。

3 实 验

将 Visual DTU 与当前流行的几种真实 MVS 数据集进行对比，以验证 Visual DTU 的有效性。主要采用 2 个评价指标：一是数据集制作的时间和经济成本；二是采用不同数据集训练的 MVS 网络模型的三维重建效果。三维重建效果采用 DTU 数据集提供方法评估。主要包括准确性(Acc.)、完整性(Comp.)和总体性(Overall)。准确性(Acc.)以重建的三维点云与真实物体点云的距离来衡量，表示重建的点的精度；完整性(Comp.)表示物体表面被重建的完整程度；总体性(Overall)是准确性和完整性的平均值，是一个综合的误差指标。3 种指标的值越小代表误差越小。

然后，通过改变训练样本数量、混合真实数据集和虚拟数据集的方式分析 Visual DTU 的使用规律，以达到最好的训练效果。

3.1 数据集制作成本对比

该文主要对比了 Visual DTU 与当前流行的几种真实数据集(ETH3D、Tanks and Temples、DTU)在数据采集过程中的经济和时间成本。结果如表 1 所示，其中价格是依据各数据集进行数据采集所需的主要设备计算得来，具体是参照该型号设备的当前价格，对于原始论文中没有说明设备型号的情况，则按照同类型设备的均价计算。

表 1 数据集制作成本对比

数据集	时间	场景类型	主要设备	价格(万元)	平均获取 1 个场景数据的时间
ETH3D[2]	2017	室内外大场景	三维激光扫描仪、单反相机、全局快门相机	>40	约 1 天
Tanks and Temples[3]	2017	室内外大、小场景	三维激光扫描仪、单反相机	>48	小场景 2–4 小时 大场景 1–2 天
DTU[1]	2016	室内小场景	6 轴工业机械臂、工业相机、结构光扫描仪	>4.5	约 1 小时
Visual DTU	2022	小场景	笔记本电脑	0.5	约 6 分钟

从经济成本看,真实数据集所需设备成本都比较高。例如,ETH3D 和 Tanks and Temples 所采用的如 Focus3D X330 激光扫描仪需花费几十万,各种高清相机也需数万元。而且,该文并未将一些辅助设备列入。例如,DTU 为了产生光照变化的场景而设置了 16 个 LED 灯,Tanks and Temples 为了获取更稳定的图像而使用相机云台。而相比之下,制作 Visual DTU 仅需一台普通的笔记本电脑(该文采用的是联想 Air15),可大幅度降低经济成本。

从时间成本看,真实数据集时间消耗普遍较大,尤其 Tanks and Temples 和 ETH3D 需采集室外大范围场景数据,而室外容易受到自然环境或人为运动的影响而出现采集失败的情况。DTU 主要是室内小场景,通过机器臂完成视角转换,速度相对较快。而 Visual DTU 完全由电脑自动计算完成,不会受到外界的影响,可快速完成数据采集工作;时间消耗与场景类型无关,且时间消耗将随着主机性能的提高而降低。

3.2　Visual DTU 和 DTU 训练效果对比

使用当前比较流行的几种有监督(CVP-MVSNet[7]、Patchmatchnet[9])和无监督(M3VSNet[8]、JDACS-MS[10])深度 MVS 模型进行对比实验。有监督深度 MVS 模型训练需要输入相机图像、相机参数和场景深度图,无监督深度 MVS 模型训练仅需要输入相机图像和相机参数,不需要输入场景深度图,因此在两类模型上比较更具有一般性。实验均使用 Pytorch 实现,在 4 个 NVIDIA GTX 1080Ti GPU 上进行训练。和原模型训练方法相同,采用 79 组图像作为训练集,除因主机内存限制而降低个别模型训练的批大小外,其余各训练参数设置均与原模型相同。为方便对比,训练后的模型均采用 DTU 数据集中的 22 组图像进行测试。模型在深度估计后均重建了场景的三维点云。

表 2　Visual DTU 与 DTU 数据集训练效果对比

模型与数据集	Acc./mm	Comp./mm	Overall/mm
CVP-MVSNet(DTU)[7]	0.296	0.406	0.351
CVP-MVSNet(Visual DTU)[7]	0.379	0.331	0.354
Patchmatchnet(DTU)[9]	0.427	0.277	0.352
Patchmatchnet(Visual DTU)[9]	0.453	0.295	0.374
M3VSNet(DTU)[8]	0.636	0.531	0.583
M3VSNet(Visual DTU)[8]	0.650	0.532	0.591
JADCS-MS(DTU)[10]	0.398	0.318	0.358
JADCS-MS(Visual DTU)[10]	0.424	0.318	0.379

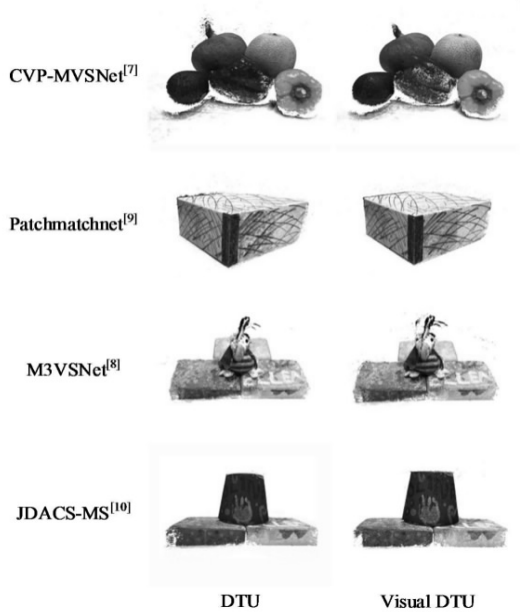


图 6　不同数据集重建效果对比

表 2 展示了测试结果,括号内表示使用的数据集。从表中数据可以看出,采用虚拟数据集 Visual DTU 训练的模型的重建效果基本与采用真实数据集训练的模型相当,有监督模型总体重建效果平均差距仅 0.012 5 mm,无监督模型总体重建效果平均差距仅 0.014 5 mm。证明由 Unity3D 生成的相机图像、场景深度图和相机参数与实际比较相符,可以作为深度 MVS 模型的训练输入,而该文制作的虚拟数据集 Visual DTU 一定程度上可代替真实数据集 DTU 用于模型的训练。图 6 展示了各深度 MVS 模型采用不同数据集训练后重建的点云效果,直观上可看出两种数据集训练的模型重建效果基本相当,只在一些细微的地方有所差距。

3.3　Visual DTU 训练样本数量变化对比

上文证实,当采用相同数量的 Visual DTU 样本进行深度 MVS 网络训练,基本可达到采用真实数据集 DTU 训练的效果,但还是有一些差距。因此,本节分析了增加训练样本的情况下三维重建效果的变化。将 Visual DTU 训练的场景数量从 80 逐步增加至 130,总体性误差随训练样本数量变化如图 7 所示。由图 7 可以看出,训练样本数量逐渐增加时,总体性误差承下降趋势,最终总体误差能小于原先采用 79 个 DTU 场景训练的情况。由此可以看出,虽然虚拟数据集 Visual

DTU 与真实数据集 DTU 还有一些差距,但可以通过增加训练样本数量来弥补。但从总体误差变化趋势也能看出,增加训练样本并不能无限降低误差,误差的决定因素还是深度 MVS 网络的结构设计,仅通过增加训练样本很难突破深度 MVS 网络本身的不足。

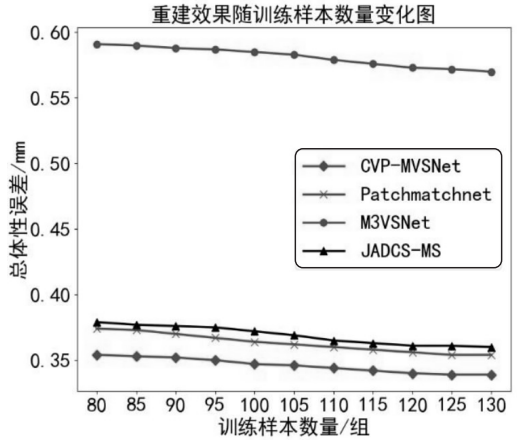


图 7 重建效果随样本数量变化

3.4 混合数据集训练

本节研究了深度 MVS 模型在 Visual DTU 与 DTU 两种数据集上共同训练的情况。分别采用 3 种混合训练方式:(a)两数据集随机混合成一个更大的整体,而后在这个更大的数据集上进行训练;(b)先用 DTU 数据集训练,后用 Visual DTU 数据集训练;(c)先用 Visual DTU 数据集训练,后用 DTU 数据集训练。选用 CVP-MVSNet 进行对比实验。训练时,除第一种方式中训练样本扩大一倍意外,其余各参数设置均与 3.2 节相同。测试同样采用 DTU 测试集的 22 组图像,表 3 展示了仅用一种数据集训练和混合数据集训练的重建效果对比。

表 3 混合数据集训练效果对比

训练方式	Acc./mm	Comp./mm	Overall/mm
DTU	0.296	0.406	0.351
Visual DTU	0.379	0.331	0.354
(a)	0.336	0.354	0.345
(b)	0.374	0.412	0.393
(c)	0.373	0.299	0.336

从表 3 可以看出,3 种混合训练方式的重建结果有很大区别。采用整体混合数据集训练的重建效果虽然精确性误差比 DTU 数据集要低,完整性误差比 Visual DTU 低,但总体性误差要优于仅采用一种数据集训练。先用 Visual DTU 数据集训练、后用 DTU 数据集训练的方式的重建效果最好,各个指标均比仅使用 Visual DTU 训练的模型更好,整体性误差比仅使用 DTU 数据集降低了 4.27%,比仅使用 Visual DTU 降低了 5.08%。但是,先用 DTU 数据集训练、后用 Visual DTU 数据集训练的方法的重建结果反而比仅使用一种数据集训练更差。分析主要原因,是由于测试采用的是真实数据集 DTU,因此模型后期采用 Visual DTU 训练反而会使预测值与真实值增大偏差。由此可得出结论,先采用虚拟数据集预训练,后用真实数据集训练可以达到最好的模型训练效果。

4 结束语

基于 Unity3D 提出了一种虚拟世界仿真现实世界的方法,通过 Unity3D 引擎搭建虚拟场景,设置虚拟相机,自动同步生成三维多视角虚拟数据,并制作了虚拟数据集 Visual DTU。通过大量实验证明,虚拟引擎生成的高质量相机图像、场景深度图和相机参数可用于深度 MVS 模型训练,且该方法可大幅度降低了数据集制作的成本和时间。证明了虚拟数据集 Visual DTU 基本可以代替真实数据集,并可通过增加样本数量可弥补与真实数据集直接的差距。同时,采用虚拟数据集预训练、后用真实数据集训练的方式可进一步提高训练效果。

在未来的工作中,将进一步研究数据集与深度 MVS 网络性能的内在关系,以达到可以通过设计特定的数据集来提高特定网络的重建效果。并进一步通过 Unity3D 设置更复杂的大场景或运动场景,提高深度 MVS 网络的鲁棒性和泛化性。

参考文献:

[1] AANÆS H, JENSEN R R, VOGIATZIS G, et al. Large-scale data for multiple-view stereopsis[J]. International Journal of Computer Vision, 2016, 120(2): 153-168.

[2] SCHOPS T, SCHONBERGER J L, GALLIANI S, et al. A multi-view stereo benchmark with high-resolution images and multi-camera videos[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. Honolulu: IEEE, 2017: 3260-3269.

[3] KNAPITSCH A, PARK J, ZHOU Q Y, et al. Tanks and temples: benchmarking large-scale scene reconstruction[J]. ACM Transactions on Graphics (ToG), 2017, 36(4): 1-13.

[4] GAIDON A, WANG Q, CABON Y, et al. Virtual worlds as proxy for multi-object tracking analysis[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. Las Vegas: IEEE, 2016: 4340-4349.

[5] TIAN Y, LI X, WANG K, et al. Training and testing object detectors with virtual images[J]. IEEE/CAA Journal of Automatica Sinica, 2018, 5(2): 539-546.

[6] ROS G, SELLART L, MATERZYNSKA J, et al. The synthia dataset: a large collection of synthetic images for semantic segmentation of urban scenes[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. Las

Vegas:IEEE,2016:3234-3243.

[7] YANG J,MAO W,ALVAREZ J M,et al. Cost volume pyramid based depth inference for multi-view stereo[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. Seattle: IEEE/CVF, 2020: 4877-4886.

[8] HUANG B,YI H,HUANG C,et al. M3VSNet:unsupervised multi-metric multi-view stereo network[C]//2021 IEEE international conference on image processing (ICIP). Anchorage:IEEE,2021:3163-3167.

[9] WANG F,GALLIANI S,VOGEL C,et al. PatchmatchNet: learned multi-view patchmatch stereo[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. Nashville:IEEE/CVF,2021:14194-14203.

[10] XU H,ZHOU Z,QIAO Y,et al. Self-supervised multi-view stereo via effective co-segmentation and data-augmentation [C]//Proceedings of the AAAI conference on artificial intelligence. New York:AAAI,2021:2-6.

[11] 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). New York: IEEE,2006:519-528.

[12] STRECHA C,VON H W,VAN G L,et al. On benchmarking camera calibration and multi-view stereo for high resolution imagery[C]//2008 IEEE conference on computer vision and pattern recognition. Anchorage:IEEE,2008:1-8.

[13] YAO Y,LUO Z,LI S,et al. Blendedmvs:a large-scale dataset for generalized multi-view stereo networks[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. Seattle: IEEE/CVF, 2020: 1790-1799.

[14] XU J,VÁZQUEZ D,LÓPEZ A M,et al. Learning a part-based pedestrian detector in a virtual world[J]. IEEE Transactions on Intelligent Transportation Systems,2014,15(5): 2121-2131.

[15] PENG X,SUN B,ALI K,et al. Learning deep object detectors from 3d models[C]//Proceedings of the IEEE international conference on computer vision. Santiago:IEEE,2015: 1278-1286.

[16] ŽIDEK K,LAZORÍK P,PITEL J,et al. An automated training of deep learning networks by 3D virtual models for object recognition[J]. Symmetry,2019,11(4):496.

[17] 杨　壮. 面向 Bin Picking 的虚拟数据集构建及智能识别方法的研究[D]. 上海:华东理工大学,2019.

[18] ALHAIJA H A,MUSTIKOVELA S K,MESCHEDER L,et al. Augmented reality meets computer vision: efficient data generation for urban driving scenes[J]. International Journal of Computer Vision,2018,126(9):961-972.

[19] RICHTER S R, HAYDER Z, KOLTUN V. Playing for benchmarks[C]//Proceedings of the IEEE international conference on computer vision. Venice: IEEE, 2017: 2213-2222.

[20] HOLLÓSI J,KRECHT R,MARKO N,et al. Improving the efficiency of neural networks with virtual training data[J]. Hungarian Journal of Industry and Chemistry,2020,48(1): 3-10.

[21] NOGUES F C,HUIE A,DASGUPTA S. Object detection using domain randomization and generative adversarial refinement of synthetic images[C]//IEEE conference on computer vision and pattern recognition. Salt Lake City: IEEE, 2018:4321-4328.

[22] CHEN Y H,LI W,SAKARIDIS C,et al. Domain adaptive faster R-CNN for object detection in the wild[C]//IEEE conference on computer vision and pattern recognition. Salt Lake City:IEEE,2018:3339-3348.

[23] SAITO K,USHIKU Y,HARADA T,et al. Strong-weak distribution alignment for adaptive object detection[C]//IEEE conference on computer vision and pattern recognition. Long Beach:IEEE,2019:6956-6965.

[24] TOBIN J. Domain randomization for transferring deep neural networks from simulation to the real world[C]//International conference on intelligent robots and systems (IROS). Vancouver:IEEE/RSJ,2017:23-30.

[25] TREMBLAY J,PRAKASH A,ACUNA D,et al. Training deep networks with synthetic data:bridging the reality gap by domain randomization[C]//Proceedings of the IEEE conference on computer vision and pattern recognition workshops. Salt Lake City:IEEE,2018:969-977.

[26] 王　也. 基于深度学习与虚拟数据的车辆识别与状态估计研究[D]. 长春:吉林大学,2019.

[27] PATRICK R. Generating synthetic data for image segmentation with unity and PyTorch/fastai[EB/OL]. [2019-02-20]. https://blog. stratospark. com/category/programming. html.

[28] YAO Y,LUO Z,LI S,et al. Mvsnet:depth inference for unstructured multi-view stereo[C]//Proceedings of the European conference on computer vision (ECCV). Munich: Springer International Publishing,2018:767-783.