

基于半监督生成对抗网络的三维重建云工作室

余翀

(英伟达半导体科技(上海)有限公司, 上海 201210)

摘要: 由于固有的问题复杂性和计算复杂度, 三维重建是计算机视觉研究和应用领域非常重要且富有挑战性的课题。目前已有的三维重建算法往往会导致重建的三维模型上存在着明显的空洞、扭曲失真或者模糊不清的部分, 而基于机器学习的三维重建算法往往又只能重建简单的分离物体, 并表示成三维体元形式。所以这些算法框架对于实际应用来说都远远不够。从 2014 年起, 生成对抗网络被广泛应用于半监督学习, 以及产生非真实数据集的应用中。所以本文的重点是采用生产对抗网络原理, 来获得高质量的三维重建效果。提出了一种新颖的半监督三维重建算法架构, 命名为 SS-GAN-3D。该算法通过训练生成对抗网络模型, 使其达到收敛状态, 以此来迭代式地提高原始三维重建模型的质量。SS-GAN-3D 只需要将事先观测的二维图像作为弱监督样本, 对于三维结构外形的先验知识或者参考观测都没有任何依赖。最终通过定性和定量实验, 以及对实验结果的分析, 该算法框架在 Tanks & Temples 和 ETH3D 标准三维重建测试集上, 比目前最先进的三维重建方法有明显优势。基于 SS-GAN-3D 算法, 又提出了三维重建云工作室解决方案。

关键词: 三维重建; 生成对抗网络; 半监督学习; 云工作室

中图分类号: TP18, TP39

文献标识码: A

doi: 10.11959/j.issn.2096-6652.201909

Three-dimensional reconstruction cloud studio based on semi-supervised generative adversarial networks

YU Chong

NVIDIA Semiconductor Technology Co., Ltd., Shanghai 201210, China

Abstract: Because of the intrinsic complexity in computation, three-dimensional (3D) reconstruction is an essential and challenging topic in computer vision research and applications. The existing methods for 3D reconstruction often produce holes, distortions and obscure parts in the reconstructed 3D models. While the 3D reconstruction algorithms based on machine learning can only reconstruct voxelized 3D models for simple isolated objects, they are not adequate for real usage. From 2014, the generative adversarial network (GAN) is widely used in generating unreal dataset and semi-supervised learning. So the focus of this paper is to achieve high quality 3D reconstruction performance by adopting GAN principle. A novel semi-supervised 3D reconstruction framework, namely SS-GAN-3D was proposed, which can iteratively improve any raw 3D reconstruction models by training the GAN models to converge. This new model only takes 2D observation images as the weak supervision, and doesn't rely on prior knowledge of shape models or any referenced observations. Finally, through qualitative and quantitative experiments and analysis, this new method shows compelling advantages over the current state-of-the-art methods on Tanks & Temples and ETH3D reconstruction benchmark datasets. Based on SS-GAN-3D, the 3D reconstruction studio solution was proposed.

Key words: three-dimensional reconstruction, generative adversarial network, semi-supervised learning, cloud studio

1 引言

在计算机视觉和计算机图形学领域, 三维重建是一项复原真实物体形状、结构和外观的技术。由于其具有丰富直观的表现力, 三维重建技术被广泛应用于工程建设^[1]、地理测绘^[2]、考古学^[3]、游戏设计^[4]、虚拟现实^[5]等领域中。

本文提出了一种基于半监督生成对抗网络的三维重建算法(SS-GAN-3D), 该方法将传统三维重建技术的优势与最新的生成对抗网络的机器学习原理相结合。通过对三维生成模型和三维判别模型的对抗训练过程进行同步细调, 本文提出的框架可以以一种半监督学习的方式稳定地完善所重建的三维物体的重建质量。在此算法的基础上, 还搭建了三维重建云工作室, 向广大用户提供便捷可访问的三维重建云服务系统。本文的主要工作包括如下几个方面。

- SS-GAN-3D 是一个半监督学习的算法框架, 它只需要将事先观测的二维图像作为监督样本。对于三维物体结构外形的先验知识、计算机辅助设计模型库或者参考观测都没有任何依赖。
- 目前绝大多数基于机器学习的三维重建算法只能重建简单的分离物体, 例如桌子、椅子、汽车等, 并且重建出来的模型只能以离散化的体元形式表示。而 SS-GAN-3D 可以重建复杂的三维物体以及场景, 同时重建出来的模型能以高精度的结构加外观方式呈现出来。
- 通过建立三维重建模型评价标准与生成对抗网络之间的联系, SS-GAN-3D 优化了整个网络的训练过程。这项工作使得生成对抗网络在复杂场景的重建中得以实际应用。

2 相关研究工作

三维重建的目标可以是一些分离物体^[6-7], 也可以是大尺度的场景^[8-9]。对于不同的重建目标, 研究人员会尝试用不同的方式去展现重建出来的 3D 模型。常见的展现形式包括立体体元^[6]、点云^[8]以及网状骨架和表面纹理^[10]结合。近年来, 研究人员在三维重建技术的新方法研究中, 取得了长足的进步。目前最先进的三维重建方法主要可以分成以下 4 类。

(1) 基于立体多视角和运动结构恢复的三维重建方法

该类算法首先会根据两张图片进行特征匹配,

然后用得到的双视角重建结果去初始化三维模型, 添加新的匹配图像并反复迭代, 进行三角特征匹配, 并用光束平差法进行运动结构恢复。该类算法的时间复杂度是 $O(n^4)$, 其中, n 代表观测相机的个数。该类算法中最有代表性的是 VisualSFM^[10-11], 该算法进一步提高了运算性能, 优化了包括光束平差法在内的大量耗时步骤, 时间复杂度降低到 $O(n)$ 。该算法同时也提供了简单易用的图形用户界面。

然而, 这类算法也有着明显的局限性, 它们都基于一个重要的假设, 那就是特征信息在多个视角上是可以完美匹配的。如果各个视角之间的空间距离较大, 由于局部外观变化或者相互遮挡, 特征匹配将变得异常困难。另外一个局限之处在于, 如果待重建物体表面缺少纹理信息, 或者表面存在镜面反射, 特征匹配过程就很可能完全失效。

(2) 基于深度摄像头的三维重建方法

该类算法中最著名的是 KinectFusion^[12], 通过探测到的深度信息, 该方法能够连续地追踪和求解深度摄像头在 6 个自由度上的姿态信息。该方法的追踪精度要明显优于基于运动结构恢复的三维重建方法(因为这种方式只能通过彩色图片帧与帧之间的特征匹配来进行摄像头姿态的追踪)。通过将深度和姿态信息迭代式地融合到全局的稠密立体模型中, 最终输出构建好的三维模型。在 Whelan 的工作^[8]中, 他在 KinectFusion 的基础上进一步提升了追踪精度、顽健性以及重建质量。改进算法采用了密集式图像帧对应到模型的摄像头追踪、滑窗式点元融合以及非刚体表面变形等技术, 得到了更高质量的三维重建模型。

这类算法的局限性主要在于, 自遮挡、光线反射以及深度传感器融合误差等因素的存在, 会导致重建出的三维模型上存在着明显的空洞、扭曲失真或者模糊不清的部分。

(3) 基于形状先验知识的三维重建方法

该类算法的代表作是三维循环重建神经网络(3D-R2N2)^[6]算法, 它利用深度卷积神经网络, 从大量的训练数据集中去学习观测的二维图像与相应目标物体的三维形状之间的映射关系。结合长短时记忆网络的优势, 3D-R2N2 算法可以根据从任意视角观测到的一张或多张图片, 输出目标物体对应的重建结果。重建出的模型采用三维体元形式进行表示。

3D-R2N2 算法的主要优势在于其框架能够同时适用于单视角和多视角的三维重建。通过控制神经网络输入门和遗忘门的状态,可以选择性地更新对模型的隐式表达。甚至在不同视角得到的观测信息有部分冲突的情况下,该算法仍然能够输出重建的三维模型。然而,该算法的局限性也很明显。3D-R2N2 算法的最终效果完全取决于三维计算机辅助设计模型和相应的二维观测共同构成的训练集的质量。另外,该算法只能重建一些分离物体,并且模型的输出形式为离散化的三维体元,很难展现物体表面的真实材质和纹理信息。

(4) 基于生成对抗网络的三维重建方法

该类算法中最具代表性的是 3D-GAN^[7]算法。3D-GAN 算法引入了生成对抗损失,并将其作为区分一个物体是真实的还是重建的评判标准。因为三维物体是高度结构化的,采用生成对抗损失,比传统的体元级别独立启发式评判标准的效果更好,能够更加准确地捕捉目标物体的三维结构的细微差异。

3D-GAN 算法最明显的优势来自生成对抗网络本身的特性。在此架构下,3D-GAN 算法可以学到低维度概率空间和实际物体三维空间之间的映射关系。所以整个三维重建的过程不依赖于三维计算机辅助设计模型库和相应的二维参考观测。但是,算法的局限性同样也来自生成对抗网络本身的特性。就目前来说,虽然许多研究工作已经改善了生成对抗网络的训练过程,但是即使

是对于二维图片来说,相比于普通的深度学习网络来说,它还是比较难以收敛的。鉴于三维空间的复杂度,3D-GAN 算法也只能重建简单的分离物体,并表示成三维体元形式,在模型大小、颜色、纹理样式和质量方面都有局限性。

3 基于半监督生成对抗网络的三维重建算法

3.1 算法原理

试想一下这样一个例子,一个观察者希望将一个真实的场景和一个人工重建的场景模型区分出来。首先他会在真实的三维场景中进行观察,然后他在重建的三维场景模型中也进行观察,而且每个观察的位置和视角和他在真实场景中的时候是一样的。如果他在重建的三维场景模型中观察到的一系列二维图片,与他在真实的三维场景中观察所得完全一致,那么这个观察者自己其实是很难区分到底哪个是真实的三维场景,哪个是重建的三维场景模型。

为了构建三维重建算法,可以将每一组在真实场景中观测到的二维图片,与在重建的场景模型中观测到的二维图片之间的差异进行累加。如果在每个观察位置和视角上,这样的差异都足够小,就可以认为这个重建的三维模型质量很高。而且从量化的角度来看,累计的差异越小,重建的三维模型质量越高。可以把这个作为最终评判三维重建模型的标准。对于这个概念更加直观的表述如图1所示。

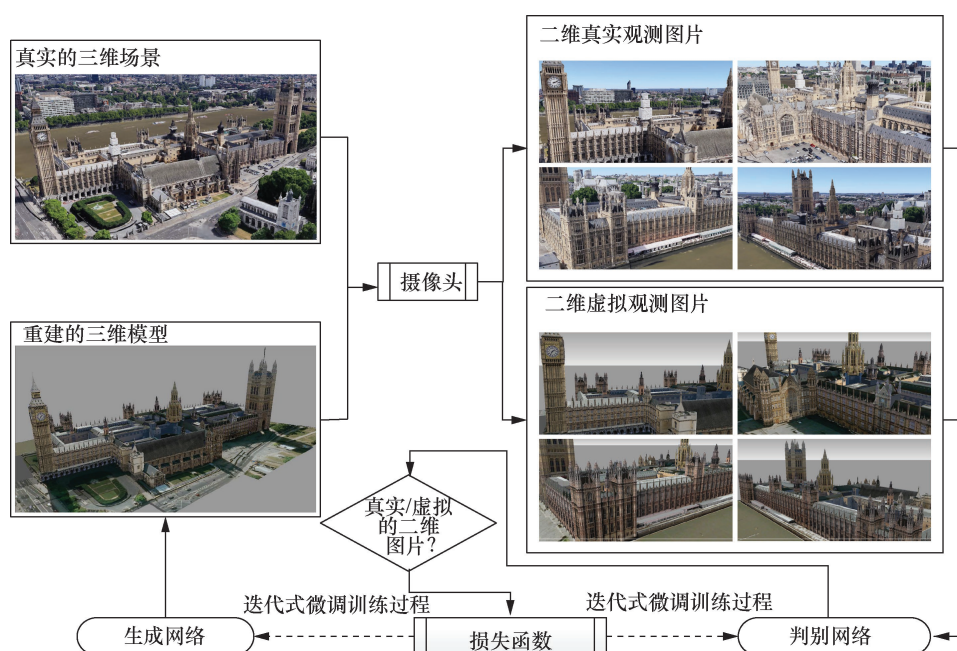


图1 基于半监督生成对抗网络的三维重建算法的原理与算法流程

生成对抗网络是由一个生成网络和一个判别网络共同组成的。生成网络的作用是产生和真实样本非常相似的新样本，并试图用这些新样本使判别器的判断结果产生混淆；而判别网络的目标是清楚地分辨出哪些是真实样本，哪些是生成网络合成的样本。同时判别网络还会估计出，某个特定样本是生成网络合成出来的样本的可能性有多大。当整个生成对抗网络模型达到纳什均衡，也就是说，生成网络可以产生特性和分布与真实样本完全一致的新样本，同时，判别网络对于每一对真实和生成样本组，输出的判别概率达到 0.5 时，整个生成对抗网络模型就完成训练，达到收敛状态。

将三维重建的目标与生成对抗网络模型相结合，本文开发出了一种新的三维重建架构：基于半监督生成对抗网络的三维重建网络（SS-GAN-3D）。SS-GAN-3D 是由一个三维模型生成网络和一个三维模型判别网络共同组成的。这里可以把判别网络想象成上面例子中提到的那个观察者。这样一来，生成网络的目标是重建和真实三维场景非常相似的三维模型，并且试图用这个三维场景模型去迷惑判别网络。而判别网络的目标是清楚地分辨真实三维场景和重建的三维模型之间的区别。当 SS-GAN-3D 达到纳什均衡，也就是说，三维模型生成网络可以重建特性和分布与真实三维场景完全一致的三维模型，同时，三维模型判别网络对于每一对真实场景和三维重建模型组，输出的判别概率达到 0.5。这样一来，也符合上文给出的三维重建模型质量的衡量标准。总之，本文新提出的架构将传统的三维重建求解问题等价转化成训练 SS-GAN-3D，并达到收敛的机器学习问题。

3.2 算法流程

对 SS-GAN-3D 进行训练时，会首先生成一个非常粗糙的三维模型，作为对三维模型生成网络的初始化。这里采用“.ply”格式表示粗糙三维模型。顶点、边缘和色彩信息分别以三元组形式存储。这个初始三维模型的构建过程是：先用摄像头对目标场景进行拍摄，生成的视频流作为观测真值；然后利用该视频流，生成二维观测图像，估计出摄像头运行轨迹，并且通过空间立体匹配的方法^[13]来构建原始的粗糙三维模型。空间立体匹配方法通过比较相邻观测图像帧之间的差异，对图像上各点在空间中的深度信息进行估计。同时从视频流中截取出的二维观测图片，也被用来构成真值

图片数据集。

初始化完成之后，就可以对 SS-GAN-3D 中的生成网络和判别网络进行迭代式微调训练。整个训练流程如图 1 所示。

由于 SS-GAN-3D 需要从重建的三维模型中得到二维观测图片，将重建出的三维模型导入专业开源三维引擎软件 Blender 和 OpenDR^[14]中。OpenDR 是一个可微分的渲染器，它能逼真地近似给出从三维模型到二维图像的真实渲染效果，同时能够提供反向传播算法所需要的从二维图像到三维模型的梯度变化情况。渲染器的可微分性是非常必要的，因为生成对抗网络的结构需要整个网络都是完全可微的，这样才能将判别网络的梯度变化传递回来，更新生成网络，形成完整的循环迭代结构。

在 Blender 中，可以设定一个虚拟摄像头，虚拟摄像头的光学参数和在真实三维场景中用于采集视频流的真实摄像头完全一致。在处理真实视频流的时候，摄像头的运行轨迹已经计算出来了。所以在 Blender 中，让虚拟摄像头沿着这条运行轨迹移动，并且利用 OpenDR 渲染器，在与真实场景中相同的位置和视角进行观测，并渲染生成二维图片。这样一来，可以分别从重建的三维模型和真实的三维场景中得到相同数目的二维虚拟和真实观测图片。

有了二维虚拟和真实观测图片集合之后，用判别网络来区分它们到底是来自对真实三维场景还是对重建三维模型的观测结果。同时根据损失函数来计算整个网络的损失值。通过网络损失值，SS-GAN-3D 可以继续微调训练过程，产生新的三维生成网络和三维判别网络。新训练得到的三维生成网络将重建一个新的三维模型，供虚拟摄像头进行观测。新观测得到的虚拟观测图片与原先的真实观测图片，会被一起送进新的三维判别网络中进行分类。SS-GAN-3D 会这样迭代地训练，并不断生成新的三维生成和判别网络，直到总体的损失值收敛到一个理想的阈值。

3.3 损失函数的定义

SS-GAN-3D 的整体损失函数包含两部分：重建损失 L_{Recons} 和交叉熵损失 $L_{\text{SS-GAN-3D}}$ 。所以损失函数可以写成

$$L_{\text{Overall}} = L_{\text{Recons}} + \lambda L_{\text{SS-GAN-3D}} \quad (1)$$

其中, λ 是调节重建损失和交叉熵损失权重的参数值。

在 SS-GAN-3D 算法框架中, 模型重建的质量是通过判别网络进行评价的。所以, 重建损失是通过计算送进判别网络的二维真实和虚拟观测图片集之间的差异得到的。本文选取了 3 种衡量图片质量的定量指标^[15], 用于计算差异。峰值信噪比 (PSNR) 从灰度值保真度的角度来定量衡量图片差异。结构相似度 (SSIM)^[16] 从结构级保真度的角度来定量衡量图片差异, 同时该指标参考和模拟了人眼系统对于结构性图案的判断标准。归一化相关性 (NC) 则是表示相同维度图片的矩阵相似性。这 3 种评价定量指标的表达式如下所示。

$$PSNR(\mathbf{x}, \mathbf{y}) = 10 \lg \left(\frac{(MAX_I)^2}{MSE(\mathbf{x}, \mathbf{y})} \right) \quad (2)$$

其中, MAX_I 代表图片 $\mathbf{x}$ 和 $\mathbf{y}$ 中, 各个像素点能取到的最大值。 $MSE(\mathbf{x}, \mathbf{y})$ 代表图片 $\mathbf{x}$ 和 $\mathbf{y}$ 的均方误差。

$$SSIM(\mathbf{x}, \mathbf{y}) = \frac{(2\mu_x\mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)} \quad (3)$$

其中, $\mu_x = \frac{1}{N} \sum_{i=1}^N x_i$ 和 $\mu_y = \frac{1}{N} \sum_{i=1}^N y_i$ 表示图片 $\mathbf{x}$ 和 $\mathbf{y}$ 的平均灰度值。 σ_x 和 σ_y 表示图片 $\mathbf{x}$ 和 $\mathbf{y}$ 的方差。 σ_{xy} 表示图片 $\mathbf{x}$ 和 $\mathbf{y}$ 的协方差。参数 C_1 和 C_2 是两个常量, 当 $\mu_x^2 + \mu_y^2$ 或者 $\sigma_x^2 + \sigma_y^2$ 非常接近于 0 的时候, C_1 和 C_2 可以防止最终 SSIM 产生发散的结果。

$$NC(\mathbf{x}, \mathbf{y}) = \frac{\mathbf{x} \cdot \mathbf{y}}{\|\mathbf{x}\| \|\mathbf{y}\|} \quad (4)$$

其中, $\mathbf{x} \cdot \mathbf{y}$ 表示矩阵 $\mathbf{x}$ 和 $\mathbf{y}$ 的内积, 运算符 $\|\cdot\|$ 表示向量的欧几里得范数。

显然, 两张图片的结构相似度指标取值为 0~1, 归一化相关性指标的取值为 -1~1。如果 SSIM 指标或 NC 指标非常接近于 1, 说明图片 $\mathbf{x}$ 和 $\mathbf{y}$ 之间差距很小。对于峰值信噪比指标, 常见图片的取值为 20~70 dB, 需要用广义 Sigmoid 函数将它的值归一化。

$$E_Sig(PSNR(\mathbf{x}, \mathbf{y})) = \frac{1}{1 + e^{-0.1(PSNR(\mathbf{x}, \mathbf{y}) - 45)}} \quad (5)$$

所以最终重建损失可以写成如下形式。

$$L_{Recons} = \sum_{j=1}^N \left\{ \alpha \cdot \left[1 - E_Sig(PSNR_{G_j F_j}) \right] + \beta \cdot \left(1 - SSIM_{G_j F_j} \right) + \gamma \cdot \left(1 - NC_{G_j F_j} \right) \right\} \quad (6)$$

其中, α 、 β 、 γ 是调节总体损失值中 PSNR、SSIM 和 NC 指标之间所占比例的参数。下标 $G_j F_j$ 表示一对真实和虚拟观测二维图片。上标 N 表示图片集中这种图片对的总数。3.4 节将会结合网络结构, 详细讨论 SS-GAN-3D 的交叉熵损失。

3.4 SS-GAN-3D 的网络结构

对于 SS-GAN-3D 来说, 判别网络需要很强的分类性能来处理三维空间投影产生的复杂二维图片。所以本文采用了 ResNet-101 网络^[17] 作为判别网络的主要结构。生成网络的结构和判别网络类似, 但是因为生成网络需要重建三维模型, 所以把其中的二维平面卷积层都替换成三维体积全卷积层。典型的 ResNet 网络都采用批量归一化, 使得整个训练过程比较稳定。但是批量归一化操作的引入, 使得判别网络判断的是一批输入和一批输出之间的映射关系。在 SS-GAN-3D 中, 希望在训练过程中保证单一输入和单一输出之间的映射关系。所以对于生成和判别网络, 都把批量归一化替换成层级归一化, 以此来避免引入输入样本之间的相关性。为了提高训练效果, 还将 ReLU 层替换成参数化 ReLU 层。为了提高收敛性能, 实际采用 Adam 求解器来替代随机梯度下降 (SGD) 求解器。在实际应用中, Adam 求解器可以使 SS-GAN-3D 在一个比较大的学习率下进行训练。详细的网络分层结构如图 2 所示。

根据研究者的实验, 就目前来说, 只有加上梯度惩罚限制的 Wasserstein GAN (WGAN)^[18] 结构才能成功地训练类似 ResNet 结构的复杂生成和判别网络。所以本文借鉴了 WGAN 的改进训练算法, 将其应用于 SS-GAN-3D 的训练过程中。训练生成网络 G 和判别网络 D 的目标函数如下。

$$\min_G \max_D \mathbb{E}_{x \sim \mathbb{P}_r} [D(x)] - \mathbb{E}_{\tilde{x} \sim \mathbb{P}_g} [D(\tilde{x})] \quad (7)$$

其中, $\mathbb{P}_r$ 代表的是真实图片的分布情况, $\mathbb{P}_g$ 代表的是生成图片的分布情况。 $\tilde{x}$ 是生成网络 G 的隐式输出结果。在最初版本的 WGAN 的训练过程中, 对权重值的裁剪很容易导致优化失败, 包括网络性能下降, 梯度爆炸或者梯度消失。在改进的版本中, 梯度惩罚作为一种更松的约束条件, 替代了简单的

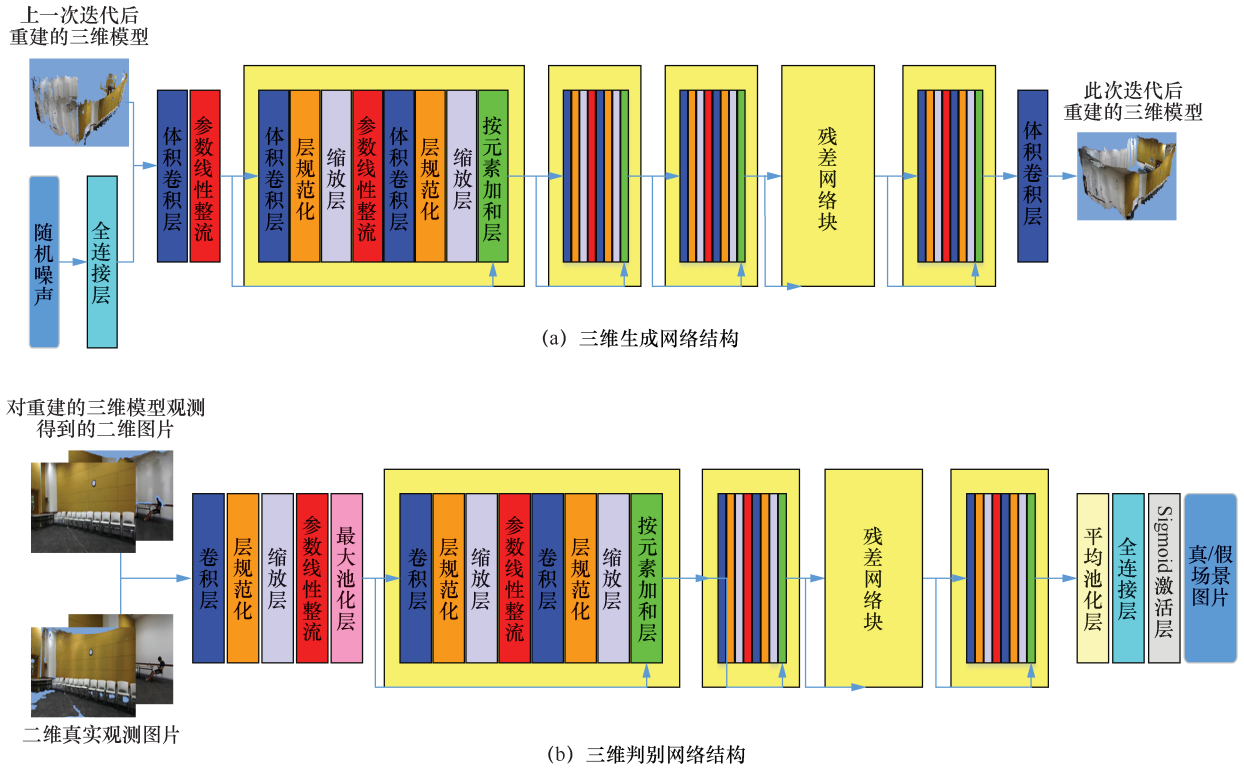


图 2 SS-GAN-3D 的三维生成网络和三维判别网络结构

权值裁剪。所以最终 SS-GAN-3D 的交叉熵损失为

$$L_{\text{SS-GAN-3D}} = \mathbb{E}_{x \sim \mathbb{P}_r} [D(x)] - \mathbb{E}_{\tilde{x} \sim \mathbb{P}_g} [D(\tilde{x})] - \theta \mathbb{E}_{\tilde{x} \sim \mathbb{P}_g} \left[\left(\|\nabla_{\tilde{x}} D(\tilde{x})\|_2 - 1 \right)^2 \right] \quad (8)$$

其中, θ 是调节梯度惩罚在交叉熵损失中所占百分比的参数。 $\mathbb{P}_x$ 表示的是, 从 $\mathbb{P}_r$ 和 $\mathbb{P}_g$ 分布中得到的样本点对之间的直线上, 均匀采样形成的数据集。交叉熵损失的值可以定量反映 SS-GAN-3D 的训练过程。这个值越小, 说明真实和虚拟二维观测图片之间的 Wasserstein 距离越小。

4 实验结果

4.1 定性实验

在定性实验中, 采用 ZED 立体摄像机作为图片和视频数据的采集工具。真实的数据集是通过立体摄像机扫描一个大礼堂得到的。立体摄像机的左视摄像头和右视摄像头会同步采集两段视频流。从采集到的视频流中, 可以提取一组场景图片作为观测真值。同时, 基于立体摄像机对深度信息的估测, 能够计算出相机的运行轨迹。有了这两项信息, 可以用空间立体匹配的方法构建网状骨架和表面纹

理组成的初始粗糙三维模型。

空间立体匹配法将目标场景的几何结构信息, 表示成三维三角网状骨架形式。三角网状骨架又是由顶点表面和法向量组成的。同时, 为了恢复三维模型的表面材质, 需要将采集过程中得到的二维图片映射到骨架表面上。所以在空间立体匹配的过程中, 会从观测图片集中记录下一组图片子集, 再将子集中的图片处理聚合一张表面纹理映射图。最终, 这张表面纹理图会通过自动生成的 UV 坐标^[19], 映射到三维网状骨架的每个表面上。

有了初始三维重建模型并对损失函数的参数进行初始化之后, 就可以对 SS-GAN-3D 进行迭代式微调训练。各参数的取值为: $\lambda = 0.7$, $\alpha = 0.25$, $\beta = 0.6$, $\gamma = 0.15$, $\theta = 10$ 。三维生成和判别网络的学习率为 0.063。选取了有代表性的三维模型重建结果, 如图 3 所示。需要被重建的三维场景是英特尔公司的大礼堂。礼堂长约 23 m, 宽 11 m, 高 5 m。图 3 (a) 展示的是在初始化阶段, 通过空间立体匹配法产生的初始粗糙三维场景模型。图 3 (b) 至图 3 (f) 展示的是在 SS-GAN-3D 迭代式微调训练过程中, 重建出的三维场景模型。可以看出, 重建的三维场景模型质量由非常粗糙逐步变得精细。

明显的空洞、扭曲失真和模糊不清的部分被 SS-GAN-3D 大大削减。图 3 (f) 也是当损失函数取很小值的时候, 最终重建的三维场景模型。在重建的三维模型中观测的二维图片与真实场景观测图片之间的对比结果如图 4 所示。这里在每个重建的三维场景模型中, 各取 4 张不同视角的二维观测图片作为例子来直观展示重建模型的质量情况。图 4 (a1) 至图 4 (a4) 这 4 张二维观测图片对应图 3 (a) 展示的三维重建场景模型。图 4 (b1) 至图 4 (f4) 的二维观测图片与图 3 展示的三维重建模型的对应关系也是类似的。图 4 (g1) 至图 4 (g4) 是在真实场景中观测的真值图片。位于同一列的图片是在相同位置和视角进行观测得到的。图 3 和图 4 所示结果可以定性证明, SS-GAN-3D 重建出的三维模型质量很高。

4.2 定量实验

本文将 SS-GAN-3D 方法与目前最先进的三维重建方法在各种不同场景的基准测试集上进行定量实验。下面简单介绍本文中所用到的两个基准测试集。

Tanks and Temples 基准测试集^[20]是由英特尔实验室建立的, 用于评估基于图像和基于视频的三维重建算法效果。该测试集包括室外场景和室内环境。真值数据用工业激光扫描仪 (FARO Focus 3D X330 HDR) 进行采集。共有 14 个不同种类的大尺度场景数据集, 其中 8 个是中等难度的, 6 个是困难度的场景。Tanks and Temples 基准测试集也提供了三维模型真值, 包括表面和几何结构信息, 因此该测试集可以精确地定量评估三维重建算法的重

建精确度。在 Tanks and Temples 基准测试集上的重建结果如图 5 所示。

ETH3D 基准测试集^[21]是由苏黎世联邦理工学院建立的, 用于评估双视角和多视角的立体三维重建算法效果。该测试集包括 7 个室外场景和 7 个室内场景。不同视角的高精度图像真值, 以及同步的低精度立体视频流, 用高精度的激光扫描仪进行采集。测试集包括各种丰富的场景细节, 比如森林中的各种树木枝叶以及游乐场上的绳子等。有些场景还有大块的同质和反射性表面, 比如输送管道以及办公室的玻璃幕墙。这些都给被测试的三维重建算法带来了巨大挑战。

本文选择了以下一些较先进的三维重建方法来和 SS-GAN-3D 进行对比: VisualSFM^[11]、PMVS^[21]、MVE^[22]、Gipuma^[23]、COLMAP^[24]、OpenMVG^[25]以及 SMVS^[26]。除此之外, 也评估了一些有兼容接口算法的组合效果。

关于对比性评估, 首先需要将重建的三维模型与真值场景对齐。因为各个重建算法都会去估计采集摄像头的位置, 所以可以通过与真实摄像头位置一一对应^[20], 完成整个对齐工作。接下来用与真值点云同样密度的体元栅格, 对重建的三维模型进行采样。如果有多个点落在同一个体元栅格中, 就采用这些点的几何平均值作为采样的结果。

采用以下 3 个指标来评估重建模型的质量。

- 精确度指标是对重建准确度的量化, 该指标的值代表重建模型上的点与真值场景中的点的靠近程度。 R 代表从重建模型上采样得到的点集, 用 G 代表真值点集。对于点集 R 上的一个点 r , 它

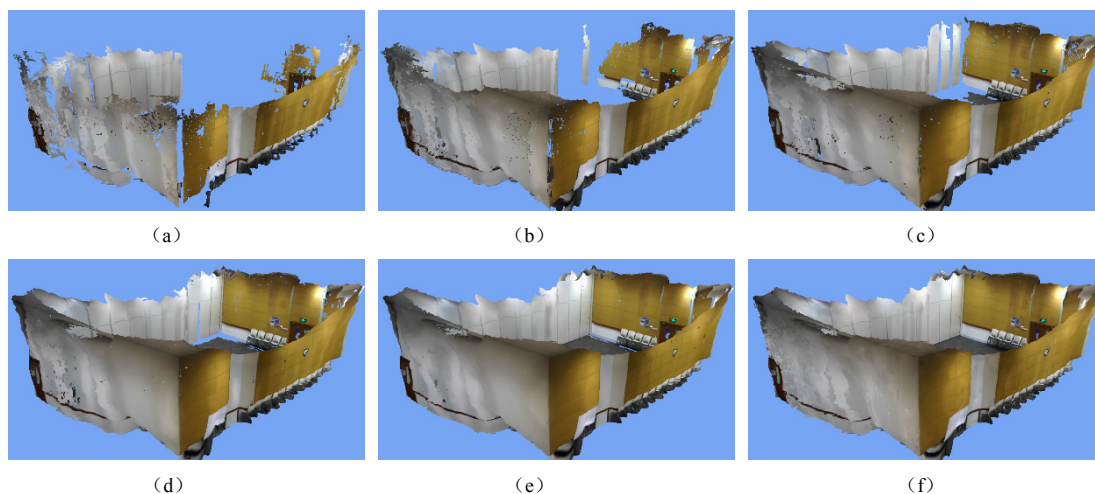


图3 SS-GAN-3D 的重建结果

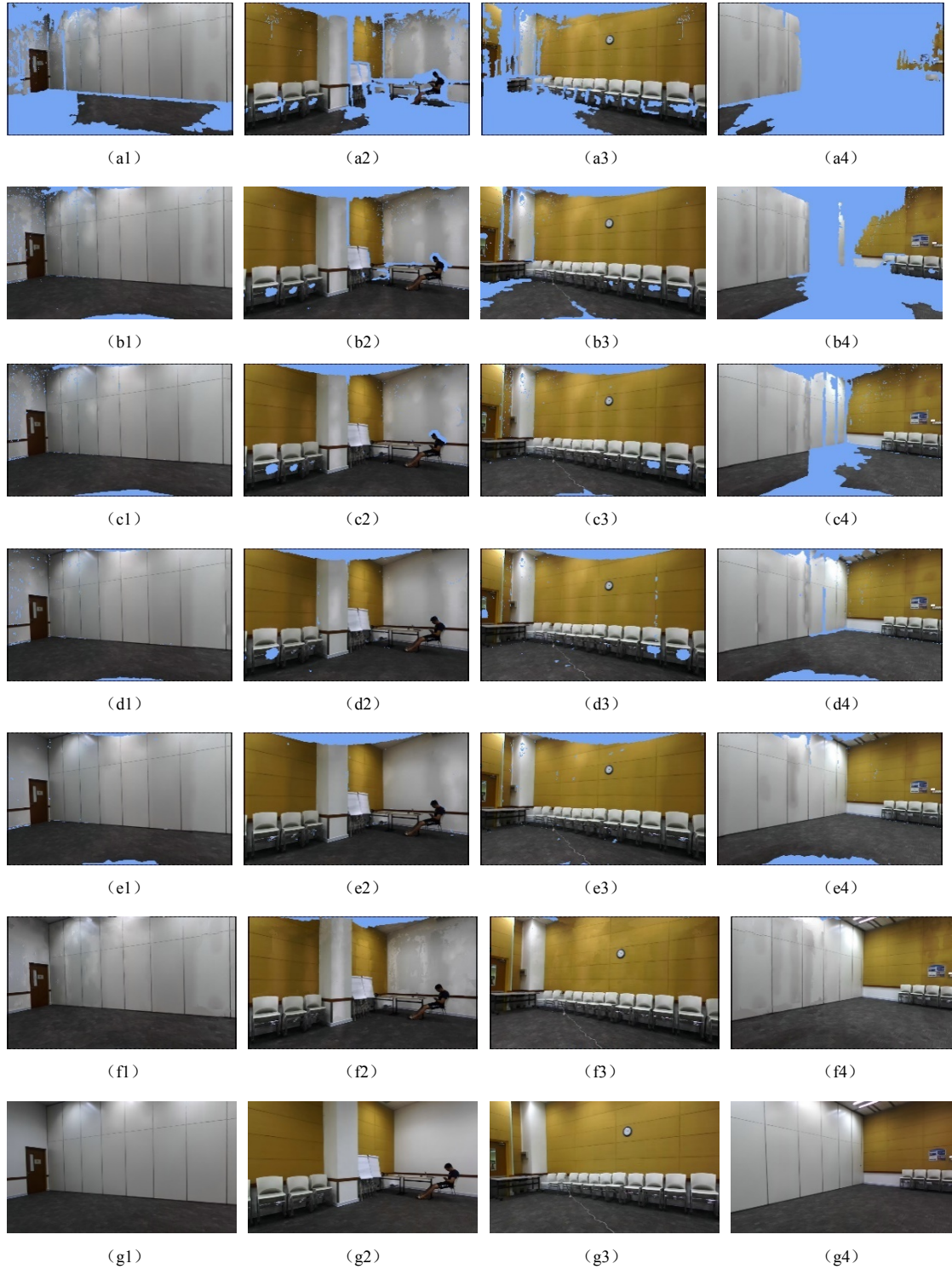


图 4 在重建出的三维模型中观测的二维图片与真实场景中观测的图片

到真值点集的距离定义为

$$d_{r \rightarrow G} = \min_{g \in G} \|r - g\| \quad (9)$$

这样一来，对于任意给定的距离阈值 e ，重建模型的精确度指标定义为

$$P(e) = \frac{\sum_{r \in R} [d_{r \rightarrow G} < e]}{|R|} \quad (10)$$

• 召回率指标是对重建完整性的量化，该指标的值代表有多少真值点集中的点被重建模型所



图 5 在 Tanks and Temples 基准测试集上的重建结果

包含。对于点集 G 上的一个点 g ，它到重建模型采样点集的距离定义为

$$d_{g \rightarrow R} = \min_{r \in R} \|g - r\| \quad (11)$$

这样一来，对于任意给定的距离阈值 e ，重建模型的召回率指标定义为

$$R(e) = \frac{\sum_{g \in G} [d_{g \rightarrow R} < e]}{|G|} \quad (12)$$

• 假设有一个非常稀疏的点集，其中每个点的位置和真值点集中对应点的位置精确重合，那么单一的精确度指标数值可以变得很大。假设整个空间中布满了点，那么单一的召回率指标数值也可以变得很大。为了避免这些极端情形，将精确度和召回率指标结合在一起，构成一个综合性的指标： F 分数。

$$F(e) = \frac{2P(e)R(e)}{P(e) + R(e)} \quad (13)$$

这样一来，前面提到的两种特殊情况都会使得 F 分数趋近于 0。只有当重建出的三维模型既准确又完整的时候， F 分数才能取得一个比较大的值。

各重建算法在 Tanks and Temples 和 ETH3D 基准测试集精确度、召回率以及 F 分数指标见表 1 至表 6。根据综合性的 F 分数指标，SS-GAN-3D 的效果优于其他的先进三维重建方法。

在 Tanks and Temples 基准测试集上，对于精确度指标，和 SS-GAN-3D 比较接近的是 COLMAP 算法以及 VisualSFM 与 PMVS 的组合算法。对于召回率指标，比较接近的是 OpenMVG 与 MVE 的组合算法。但是对于综合性的 F 分数指标，SS-GAN-3D 比排名第二的算法还有 1.1~1.5 倍的提升。

在 ETH3D 基准测试集上，对于精确度指标，和 SS-GAN-3D 比较接近的是 COLMAP 算法以及 PMVS 算法。对于召回率指标，比较接近的是 COLMAP 算法。但是，对于综合性的 F 分数指标，SS-GAN-3D 比排名第二的算法还有 1.02~1.2 倍的提升。

5 三维重建云工作室

基于 SS-GAN-3D，结合实际商业应用需求，本文提出了三维重建云工作室。之所以提出这样的解决方案，是因为本文分析了目前工业级和消费级三

表 1 各重建算法在 Tanks and Temples 基准测试集上的精确度

算法/场景	Family	Francis	Horse	Lighthouse	M60	Panther	Playground	Train	Auditorium	Ballroom	Courtroom	Museum	Palace	Temple
COLMAP	56.02	34.35	40.34	41.07	53.51	39.94	38.17	41.93	31.57	24.25	38.79	45.12	27.85	34.30
MVE	37.65	18.74	11.15	27.86	3.68	25.55	12.01	20.73	6.93	9.65	21.39	25.99	12.55	14.74
MVE+SMVS	30.36	17.80	15.72	29.53	34.54	29.59	11.42	22.05	8.29	10.62	21.24	18.57	11.45	12.76
OpenMVG+MVE	38.88	22.44	18.27	31.98	31.17	31.48	23.32	26.11	14.21	19.73	25.94	28.33	10.79	17.94
OpenMVG+PMVS	61.26	49.72	37.79	47.92	47.10	52.88	41.18	37.20	26.79	29.10	42.70	47.82	23.78	28.58
OpenMVG+SMVS	31.87	21.36	16.69	31.63	34.71	33.83	32.61	26.32	16.45	14.72	22.92	20.05	12.81	15.07
SS-GAN-3D	66.63	48.99	42.15	50.07	53.35	52.89	46.30	41.21	38.01	29.08	43.04	48.23	30.59	33.45
VisualSFM+PMVS	59.13	38.67	35.25	48.92	53.20	53.74	46.02	33.69	37.57	29.75	41.31	40.36	31.16	18.69

表 2 各重建算法在 Tanks and Temples 基准测试集上的召回率

算法/场景	Family	Francis	Horse	Lighthouse	M60	Panther	Playground	Train	Auditorium	Ballroom	Courtroom	Museum	Palace	Temple
COLMAP	45.82	16.46	18.79	49.34	59.69	57.01	66.61	42.15	10.73	26.29	31.40	38.44	13.36	23.56
MVE	68.52	32.75	14.74	68.59	8.14	75.40	3.83	49.32	2.92	18.26	40.21	52.05	14.79	19.51
MVE+SMVS	30.47	15.62	7.82	41.06	45.20	52.71	1.34	20.86	0.51	4.96	14.13	21.03	5.84	5.80
OpenMVG+MVE	69.70	37.91	24.01	73.21	71.15	77.41	84.71	57.69	15.22	39.72	43.42	55.74	2.20	31.41
OpenMVG+PMVS	30.85	10.77	7.73	28.73	30.04	24.19	25.88	22.58	2.48	7.63	13.93	20.99	3.94	8.18
OpenMVG+SMVS	31.99	18.66	13.65	43.16	45.51	54.02	39.91	24.02	4.41	9.54	17.46	24.11	6.82	10.35
SS-GAN-3D	69.31	38.11	25.12	72.89	69.97	77.60	83.55	55.72	15.47	37.66	43.59	54.83	14.74	32.28
VisualSFM+PMVS	28.02	7.77	6.73	27.83	34.36	25.07	28.86	8.25	2.49	6.63	10.20	13.30	4.15	1.13

表 3 各重建算法在 Tanks and Temples 基准测试集上的 F 分数

算法/场景	Family	Francis	Horse	Lighthouse	M60	Panther	Playground	Train	Auditorium	Ballroom	Courtroom	Museum	Palace	Temple
COLMAP	50.41	22.26	25.64	44.83	56.43	46.97	48.53	42.04	16.02	25.23	34.71	41.51	18.06	27.93
MVE	48.60	23.84	12.70	39.63	5.07	38.17	5.81	29.19	4.11	12.63	27.93	34.67	13.58	16.79
MVE+SMVS	30.41	16.64	10.44	34.35	39.16	37.90	2.40	21.44	0.96	6.76	16.97	19.72	7.73	7.98
OpenMVG+MVE	49.92	28.19	20.75	44.51	43.35	44.76	36.57	35.95	14.70	26.36	32.48	37.57	3.65	22.84
OpenMVG+PMVS	41.04	17.70	12.83	35.92	36.68	33.19	31.78	28.10	4.54	12.09	21.01	29.17	6.76	12.72
OpenMVG+SMVS	31.93	19.92	15.02	36.51	39.38	41.60	35.89	25.12	6.96	11.58	19.82	21.89	8.90	12.27
SS-GAN-3D	67.94	42.87	31.48	59.36	60.54	62.91	59.58	47.38	21.99	32.82	43.31	51.32	19.89	32.85
VisualSFM+PMVS	38.02	12.94	11.30	35.48	41.75	34.19	35.47	13.25	4.67	10.84	16.36	20.01	7.32	2.13

表 4 各重建算法在 ETH3D 基准测试集上的精确度

算法/场景	Botanical garden	Boulders	Bridge	Door	Exhibition hall	Lecture room	Living room	Lounge	Observatory	Old computer	Statue	Terrace
COLMAP	93.68	86.15	93.37	95.10	90.28	92.11	95.03	89.75	94.93	79.95	98.27	95.04
Gipuma	84.20	73.57	89.55	90.44	84.24	84.38	94.89	76.35	80.40	76.14	96.80	82.36
MVE	62.66	51.91	31.61	39.28	40.65	50.76	74.98	56.24	44.67	33.94	64.11	71.04
PMVS	96.71	92.97	90.31	89.57	88.51	90.11	95.12	90.98	88.92	77.55	97.07	83.13
SS-GAN-3D	94.84	91.27	93.70	95.42	90.63	93.69	95.33	89.57	93.49	81.76	97.64	94.35

表 5 各重建算法在 ETH3D 基准测试集上的召回率

算法/场景	Botanical garden	Boulders	Bridge	Door	Exhibition hall	Lecture room	Living room	Lounge	Observatory	Old computer	Statue	Terrace
COLMAP	81.44	53.00	83.75	75.53	48.34	48.81	81.40	24.14	90.30	32.95	60.52	75.65
Gipuma	70.99	39.64	19.85	72.56	20.10	16.16	21.47	5.08	71.16	13.07	43.70	25.08
MVE	34.56	26.92	26.57	38.87	11.73	11.37	27.44	1.37	50.42	11.41	14.60	36.93
PMVS	25.87	27.58	50.92	49.78	21.37	14.17	40.81	2.01	66.62	9.61	38.88	34.49
SS-GAN-3D	83.99	52.87	84.34	83.63	56.59	57.60	86.62	34.76	94.52	42.87	74.76	81.92

表 6 各重建算法在 ETH3D 基准测试集上的 F 分数

算法/场景	Botanical garden	Boulders	Bridge	Door	Exhibition hall	Lecture room	Living room	Lounge	Observatory	Old computer	Statue	Terrace
COLMAP	87.13	65.63	88.30	84.19	62.97	63.81	87.69	38.05	92.56	46.67	74.91	84.24
Gipuma	77.03	51.52	32.50	80.52	32.46	27.13	35.02	9.53	75.50	22.31	60.22	38.45
MVE	44.55	35.45	28.87	39.07	18.21	18.58	40.18	2.67	47.37	17.08	23.78	48.60
PMVS	40.82	42.54	65.12	63.99	34.43	24.49	57.12	3.93	76.17	17.10	55.52	48.75
SS-GAN-3D	89.09	66.95	88.77	89.14	69.67	71.34	90.77	50.08	94.00	56.25	84.68	87.70

维重建的方法,发现过程既繁琐又复杂。目前的解决方案至少需要经过 5 个主要步骤,才能重建所需的三维模型。第一步,需要选择特定的传感器或摄像头,进行初始数据采集;第二步,需要选择特定的三维建模引擎,进行三维模型的展示与计算工作;第三步,需要选择合适的硬件设备,进行运算的加速。在实际应用中,用户还需要考虑各种硬件设备的成本和性价比;第四步,需要将上面 3 个部分进行集成,并且解决集成过程中的兼容性问题。第五步,还需要设计和实现一个用户交互界面,使终端用户可以方便地进行他们特定的三维重建任务。这样繁琐的流程(如图 6 所示)还需要用户有比较强的三维重建技术背景,非常不利于大规模推广和应用。

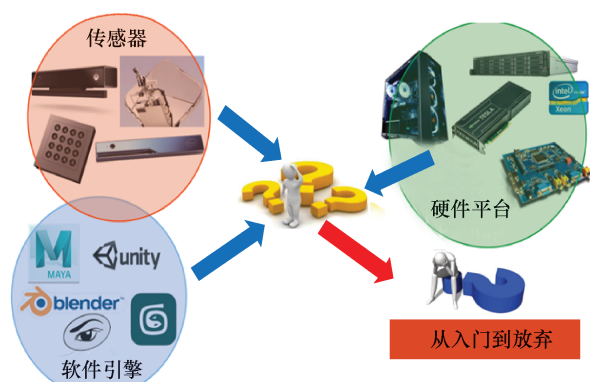


图 6 目前三维重建解决方案的复杂性示意

基于对目前三维重建方案的分析,本文对提出

的三维重建云工作室进行了改进,总体只需要 3 个步骤。首先,用户可以用各种常见的视频拍摄设备对需要重建的物体进行环绕拍摄;然后将拍摄好的视频通过用户界面上传到云端运算服务器,同时,在配置页面上勾选重建的模型所需要的输出格式等;最后,等待云端运算结束之后,自动下载得到所需格式的三维模型。重建的三维模型也可以方便地导入常用的三维建模软件(如 Maya、Unity3D)中,进行更多的创意制作。解决方案的流程如图 7 所示。



图 7 三维重建云工作室解决方案的流程

本文的云工作室解决方案的优势在于以下几点。

- 终端用户只需要提供他们的重建需求和配置,我们会帮他们选择合适的三维建模引擎和硬件平台,并且一并解决各部分集成中遇到的冲突和兼容性问题。

- 基于 SS-GAN-3D 算法框架的优良性能,对于图像和视频数据采集的设备的精度和性能要求大大降低,用户只需要用常见的拍摄设备,如手机、运动相机、DV 或是数字相机,就可以采集需要重建的物体和场景信息,因而降低了实际应

用的壁垒。

• 解决方案基于云计算,重建效率大大提高,而重建所需的成本会显著下降,有利于在实际商业应用中进行推广。

三维重建云工作室的行业应用介绍如下。

(1) 长城保护

众所周知,长城是世界中古七大奇迹之一,也是首批列入《世界遗产名录》的文化遗产。由于部分长城位于险峰断崖之上,且周边草木茂密,修缮人员到达施工现场步履维艰,导致大段大段长城年久失修,破损情况严重。北京箭扣长城就是一个典型的案例,它充分展现了古老长城的惊、险、奇、特,但该段长城自然风化严重,修缮和保护迫在眉睫。

如果利用传统手段勘察和修缮,短期内几乎不可能实现。英特尔公司与中国文物保护基金会合作,借助技术手段实施长城保护项目。

在本项目中,英特尔猎鹰 8+无人机对城墙进行近距离航拍和测绘,获取高分辨率图像。得到的数据会借助 SS-GAN-3D 三维重建云工作室的相关技术,进行清晰准确的三维建模。此后再通过一些人工智能算法,对多形态数据进行分析、处理,并结合重建出来的三维模型,为实际长城的修缮和维护提供指导与预测数据。相关的图片如图 8 所示。

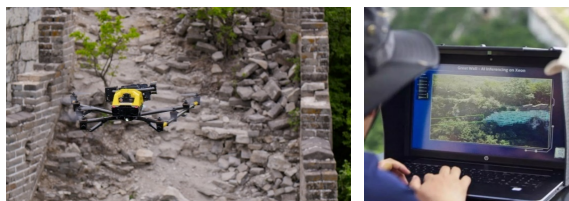


图8 三维重建云工作室在长城保护项目中的应用

(2) 其他行业应用

作者所在团队的云工作室解决方案在 2018 年分别参加了阿里巴巴的云栖大会以及腾讯的云+未来大会。在大会上,我们与来自各个行业的用户进行了广泛交流,目前已经与虚拟故宫博物院、虚拟旅游、建筑家装、智能制造、机器人、游戏设计以及服装定制等行业的客户进行了需求沟通与合作。一些重建模型案例如图 9 所示。

6 结束语

本文提出了对于复杂场景进行高精度三维重

建的算法架构。SS-GAN-3D 将传统的三维重建问题转化成了对生成对抗网络的训练与收敛性问题。由于算法基于半监督学习原理,SS-GAN-3D 对于三维形状的先验知识没有任何依赖。所以该算法框架对于实际商业中的工业和消费级三维重建应用都相当适用。

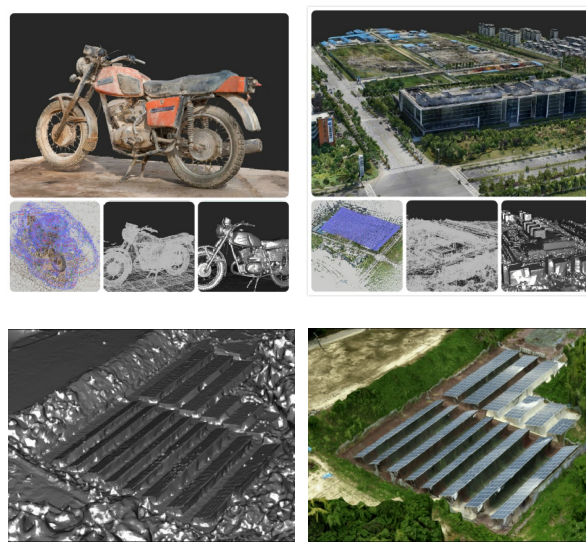


图9 三维重建云工作室在其他行业项目中的应用

SS-GAN-3D 从人观察环境的角度,提供了衡量与评价三维重建模型的量化指标。所以该算法也可以用来指导目前在三维建模领域需要人工设计和参与的一系列工作,例如电脑游戏中的角色建模,电影中的特技特效设计,无人驾驶领域中仿真模拟器的设计等。

参考文献:

- [1] DAI A, NIEBNER M, ZOLLHÖFER M, et al. BundleFusion: real-time globally consistent 3D reconstruction using on-the-fly surface reintegration[J]. ACM Transactions on Graphics (TOG), 2017, 36(4): 76a.
- [2] NEX F, REMONDINO F. UAV for 3D mapping applications: a review[J]. Applied Geomatics, 2014, 6(1): 1-15.
- [3] JOHNSON-ROBERSON M, BRYSON M, FRIEDMAN A, et al. High-resolution underwater robotic vision-based mapping and three-dimensional reconstruction for archaeology[J]. Journal of Field Robotics, 2017, 34(4): 625-643.
- [4] GOTARDO P F U, SIMON T, SHEIKH Y, et al. Photogeometric scene flow for high-detail dynamic 3D reconstruction[C]//Proceedings of the IEEE International Conference on Computer Vision, 2015: 846-854.
- [5] SRA M, GARRIDO-JURADO S, SCHMANDT C, et al. Procedurally generated virtual reality from 3D reconstructed physical space[C]//Proceedings of the 22nd ACM Conference on Virtual Reality Software and Technology, ACM, 2016: 191-200.

- [6] CHOY C B, XU D, GWAK J Y, et al. 3D-R2N2: a unified approach for single and multi-view 3D object reconstruction[C]//European conference on computer vision, Springer, Cham, 2016: 628-644.
- [7] WU J, ZHANG C, XUE T, et al. Learning a probabilistic latent space of object shapes via 3D generative-adversarial modeling[C]//Advances in Neural Information Processing Systems, 2016: 82-90.
- [8] WHELAN T, SALAS-MORENO R F, GLOCKER B, et al. ElasticFusion: real-time dense SLAM and light source estimation[J]. The International Journal of Robotics Research, 2016, 35(14): 1697-1716.
- [9] ZHOU Q Y, MILLER S, KOLTUN V. Elastic fragments for dense scene reconstruction[C]//Proceedings of the IEEE International Conference on Computer Vision, 2013: 473-480.
- [10] WU C. Towards linear-time incremental structure from motion[C]//2013 International Conference on 3D Vision, IEEE, 2013: 127-134.
- [11] WU C. Visual SFM: a visual structure from motion system[J]. 2011.
- [12] NEWCOMBE R A, IZADI S, HILLIGES O, et al. KinectFusion: real-time dense surface mapping and tracking[C]//Mixed and augmented reality (ISMAR), 2011 10th IEEE international symposium on. IEEE, 2011: 127-136.
- [13] PILLAI S, RAMALINGAM S, LEONARD J J. High-performance and tunable stereo reconstruction[C]//Robotics and Automation (ICRA), 2016 IEEE International Conference on. IEEE, 2016: 3188-3195.
- [14] LOPER M M, BLACK M J. OpenDR: an approximate differentiable renderer[C]//European Conference on Computer Vision, Springer, Cham, 2014: 154-169.
- [15] YU C. Steganography of digital watermark based on artificial neural networks in image communication and intellectual property protection[J]. Neural Processing Letters, 2016, 44(2): 307-316.
- [16] WANG Z, BOVIK A C, SHEIKH H R, et al. Image quality assessment: from error visibility to structural similarity[J]. IEEE transactions on image processing, 2004, 13(4): 600-612.
- [17] HE K, ZHANG X, REN S, et al. Deep residual learning for image recognition[C]//Proceedings of the IEEE conference on computer vision and pattern recognition, 2016: 770-778.
- [18] GULRAJANI I, AHMED F, ARJOVSKY M, et al. Improved training of Wasserstein GANs[C]//Advances in Neural Information Processing Systems, 2017: 5767-5777.
- [19] BOGO F, ROMERO J, LOPERM, et al. FAUST: dataset and evaluation for 3D mesh registration[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2014: 3794-3801.
- [20] 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): 78.
- [21] FURUKAWA Y, PONCE J. Accurate, dense, and robust multiview stereopsis[J]. IEEE transactions on pattern analysis and machine intelligence, 2010, 32(8): 1362-1376.
- [22] FUHRMANN S, LANGGUTH F, GOESELE M. MVE-a multi-view reconstruction environment[C]//GCH. 2014: 11-18.
- [23] GALLIANI S, LASINGER K, SCHINDLER K. Massively parallel multiview stereopsis by surface normal diffusion[C]//Proceedings of the IEEE International Conference on Computer Vision. 2015: 873-881.
- [24] SCHONBERGER J L, FRAHM J M. Structure-from-motion revisited[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016: 4104-4113.
- [25] MOULON P, MONASSE P, PERROT R, et al. OpenMVG: open multiple view geometry[C]//International Workshop on Reproducible Research in Pattern Recognition. Springer, Cham, 2016: 60-74.
- [26] LANGGUTH F, SUNKAVALLI K, HADAP S, et al. Shading-aware multi-view stereo[C]//European Conference on Computer Vision. Springer, Cham, 2016: 469-485.

[作者简介]



余翀 (1989—)，男，复旦大学硕士研究生，英伟达半导体科技有限公司高级架构师，英特尔亚太研发有限公司高级工程师，主要研究领域为人工智能与机器学习、机器视觉与图像处理、无人机与导航系统、智能控制理论与机器人系统、物联网技术与可穿戴设备等。