《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
201800820179 赵呈亮	校内建筑数据集的组建；单张图片三维重建的实现及讲解；微信小程序的编写	
201800820133 盛靖斐	校内建筑数据集的组建；论文的编写；多张图片三维重建的实现	
201800820236 闫文超	校内建筑数据集的组建；多张图片三维重建的实现及讲解；	

山东大学

2021 年 3 月

基于深度学习方法的单目相机 RGB 彩色图像三维重建方法

摘　要： 本文基于两种深度学习方法实现了单目视觉的物体三维重建。其中，LSTM 方法通过从 RGB 彩色图像到三维体素的映射实现单张图片的三维重建，MVS 方法通过训练卷积神经网络利用多张图片实现多视图的三维重建，两种重建方法结果分别用体素与点云展现。我们对比了算法的特点并运用到学校内的建筑、雕塑等测试效果。

关键词： 单目视觉；三维重建；深度学习；LSTM；卷积神经网络；多视图重建；单张重建

Monocular Camera Based on Deep Learning Method

RGB Color Image 3D Reconstruction Method

Abstract: In this paper, we implement 3D reconstruction of objects based on two deep learning methods for monocular vision. Among them, the LSTM method realizes 3D reconstruction of a single image by mapping from RGB color images to 3D voxels, and the MVS method realizes 3D reconstruction of multiple views by training convolutional neural networks using multiple images, and the results of the two reconstruction methods are presented in voxels and point clouds, respectively. We compare the characteristics of the algorithms and apply them to test the effect on buildings and sculptures in the school.

Key words: monocular vision; 3D reconstruction; deep learning; LSTM; convolutional neural network; multi-view reconstruction; single-sheet reconstruction

1 研究背景

从二维图像中恢复失去的维度一直是经典的多视角立体和在不同角度看形状方法的目标，这些方法已经被广泛研究了几十年。第一代方法从几何学的角度来处理这个问题；他们专注于从数学上理解和规范三维到二维的投影过程，目的是为这个不理想的反问题设计数学或算法解决方案。有效的解决方案通常需要使用精确校准的相机拍摄多幅图像。第二代三维重建方法试图通过将三维重建问题制定为一个识别问题来利用这种先验知识。深度学习技术的出现，以及更重要的是大量训练数据集的不断增加，导致了新一代的方法，能够从一张或多张 RGB 图像中恢复物体的三维几何和结构，而无需复杂的相机校准过程。

2 国内外研究现状

传统方法是基于优化的方法，如 sfm，sfs 等，不基于关键点等信息，仅仅基于单目图像的方法。这类方法的核心是 blend shape 模型，最常见的就是 3DMM 模型，也就是将三维的图像按照各个维度进行分离并降维表示，随后线性叠加表示三维模型，它要解决的就是要分别求取相关的参数，通常从二维图像重建 3 维，然后从 3 维投影回 2 维计算误差。

目前的方法主要是基于深度学习的方法，主要分为四种：深度图(depth)，点云(point cloud)，体素(voxel)，网格(mesh)，主要步骤为深度估计、三维预测等，采用深度学习从 2D 图像到其对应的 3D voxel 模型的映射，提高了重建精度。

3 基于 3D-LSTM 的图像三维重建

3.1 算法原理

该算法利用深度学习把基于 RGB 图像进行三维重建，把 2D 图像到其对应的 3D 体素（Voxel）模型建立映射关系。该网络结构从大量的合成数据中学习从物体图像到其基础三维形状的映射，从任意视角接收一个 RGB 彩色图像，并以三维网格（体素）的形式输出该物体的重建。

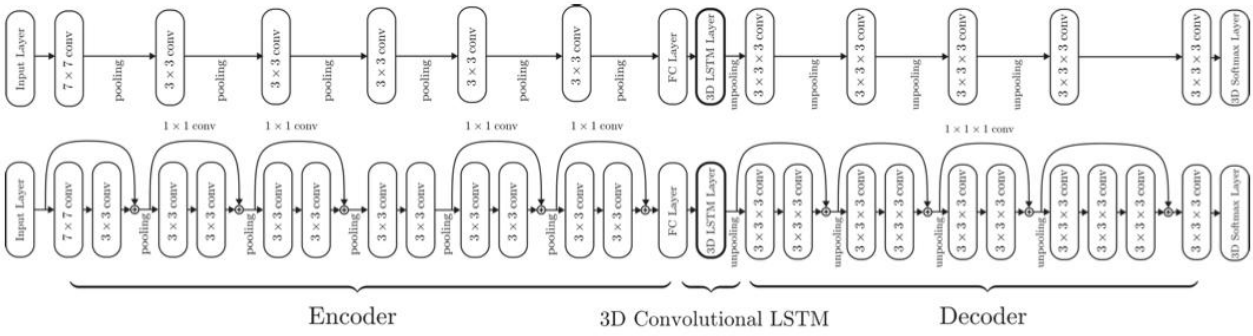


图 1 2D-LSTM 网络结构

该网络由三个组成部分组成：2D 卷积神经网络（2D-CNN），一种名为 3D 卷积 LSTM（3D-LSTM）的新型架构，以及一种 3D 反卷积神经网络（3D-DCNN）。首先利用一个标准的 CNN 结构对原始输入图像进行编码；再利用一个标准的反卷积神经网络对其解码。中间用 LSTM 进行过渡连接，LSTM 单元排列成 3D 网格结构，通过卷积每个单元接收一个从编码器和邻居的隐藏状态中获得的特征向量，并将他们输送到 Decoder 中。这样每个 LSTM 单元重构输出体素的一部分。总之，通过这样的 Encoder-3DLSTM-Decoder 的网络结构就建立了单张 2D RGB 彩色图像到 3D 体素模型的映射。

编码器：2D-CNN

我们使用 CNN 将图像编码为特征,其中有两个不同的 2D-CNN 编码器：一个标准的前馈 CNN 和一个深度残差的变体。第一个网络由标准卷积层、池化层和漏整流线性单元组成，然后是一个全连接层。另一个深度残差变体在标准卷积层和深层残差层之间添加残差连接，有效地改善并加快了深度网络的优化过程。为了与卷积层后的通道数量相匹配，我们使用 1×1 的卷积层来处理残差连接。编码器的输出被扁平化，并传递给一个全连接层。连接层将输出压缩成 1024 维的特征向量。

3D-LSTM

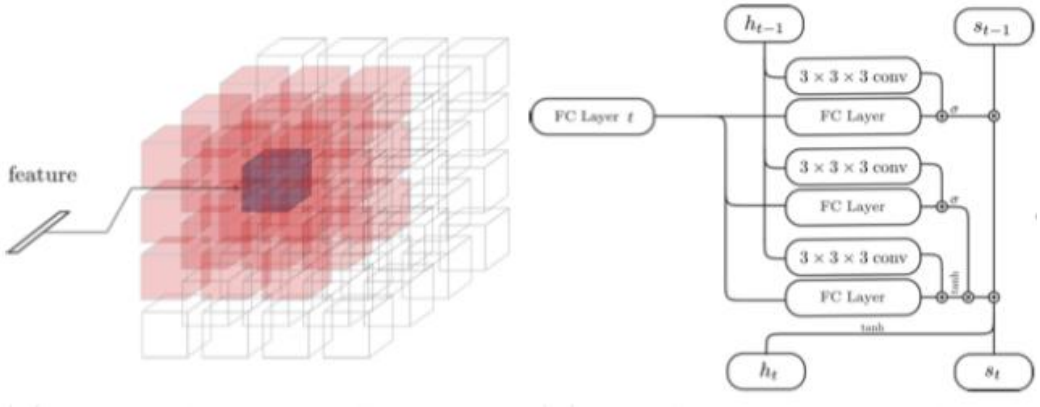


图 2 LSTM 单元结构图

这是该网络的核心部分，该网络由一组具有受限连接的结构化 LSTM 单元组成。3D-LSTM 单元在空间上分布在 3D 网格结构中，每个单元负责重建最终输出的特定部分。其输入为 Encoder 部分末尾全连接层提取出的特征，在每个时间步骤，3D-LSTM 中的每个单元（中心）通过 3×3×3 卷积从 Encoder 接收相同的特征向量，以及从其空间相邻接收隐藏状态作为输入，最终输出决定了三维网格空间的某个网格是否存在

重建物体中。

解码器：3D 反卷积神经网络

在接收到输入图像序列$x_1, x_2, \ldots, x_T$之后，3D-LSTM 将隐藏状态h_T传递给解码器，这增加了隐藏状态分辨率。应用 3D 卷积，该解码器通过应用三维卷积、非线性和三维解聚来提高隐藏状态的分辨率，直到达到目标输出分辨率。

与编码器一样，我们提出了一个简单的解码器网络，有 5 个卷积和一个深度残差版本，有 4 个残差连接，然后是一个最后的卷积。在激活达到目标输出分辨率的最后一层之后
输出分辨率，我们使用体素方向 softmax 将最终激活$V \in R^{N_{vox} \times N_{vox} \times N_{vox} \times 2}$转换为（i，j，k）处的体素单元的占用概率 p(i,j,k) 。

3.2 算法实现过程

3.2.1 损失函数

网络的损失函数被定义为体素的交叉熵之和：

$$L(X,y) = \sum_{i,j,k} y(i,j,k) log(p(i,j,k)) + (1 - y(i,j,k)) log(1 - p(i,j,k))$$

该损失函数计算原图片与三维重建体素模型中每一体素的存在性差别，由于体素的位置较为规整，该损失函数计算简单。

3.2.2 其他参数

表 1 3D-LSTM 网络训练参数

数据集	ShapeNet：https://shapenet.org PASCAL 3D+： https://cvgl.stanford.edu/projects/pascal3d.html	
训练参数	学习率	0.0001
	epoch	60000
	batch	24
硬件	Nivida v100	
环境	Pytorch、anaconda	

3.3 实验结果

3.3.1 效果图

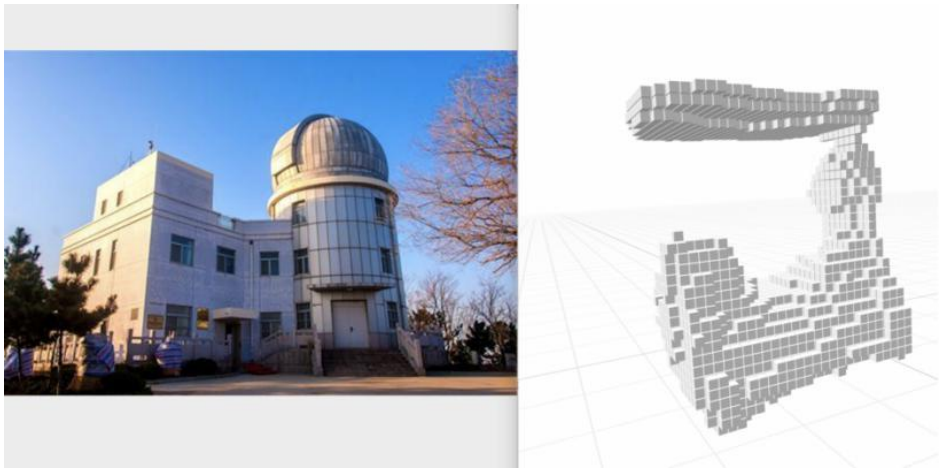


图 3（1） 天文台三维重建效果

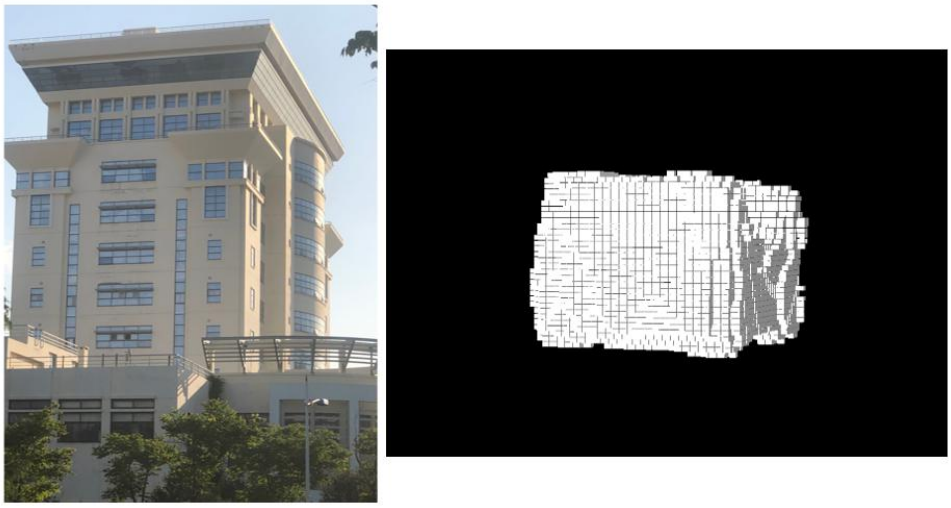


图 3（2） 图书馆三维重建效果

可以看到此方法可以有效恢复物体结构，但对于物体表面细节处理并不理想。

3.3.2 其他算法对比

我们使用体素交叉联合（IoU）比较 PASCAL VOC 的每类重建。其中除了 Kar 等人的方法之外，以相同的配置进行实验。将地面实况对象分割掩模和关键点标签作为训练和测试的附加输入，结果如下：

表 2 LSTM 与其他算法对比结果

	自行车	船	沙发	电视	飞机	平均值
Kar 等人的	0.144	0.188	0.149	0.492	0.298	0.254
LSTM	0.330	0.466	0.251	0.438	0.472	0.391

可见此算法相较其他算法重建的准确率较高。

4 基于 MVSNet 的图像三维重建

4.1 算法原理

MVS 是一种从具有一定重叠度的多视图视角中恢复场景的稠密结构的技术，传统方法利用几何、光学一致性构造匹配代价，进行匹配代价累积，再估计深度值。虽然传统方法有较高的深度估计精度，但由于存在在缺少纹理或者光照条件剧烈变化的场景中的错误匹配，传统方法的深度估计完整度还有很大的提升空间。近年来卷积神经网络已经成功被应用在特征匹配上，提升了立体匹配的精度。在这样的背景下，香港科技大学 Yaoyao 等人，在 2018 年提出了一种基于深度学习的端到端深度估计框架——MVSNet。

基于 MVSNet 的深度图估计步骤如下：深度特征提取，构造匹配代价，代价累计，深度估计，深度图优化，最后利用网络输出的深度图进行稠密重建。

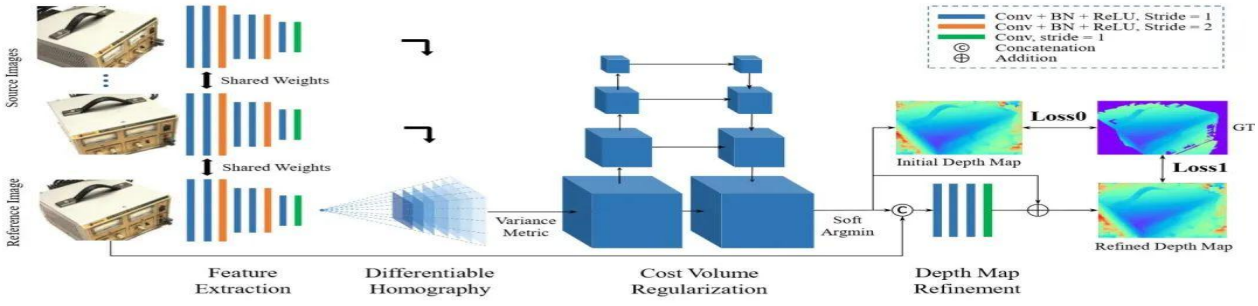


图 4 MVSNet 网络结构

深度特征提取

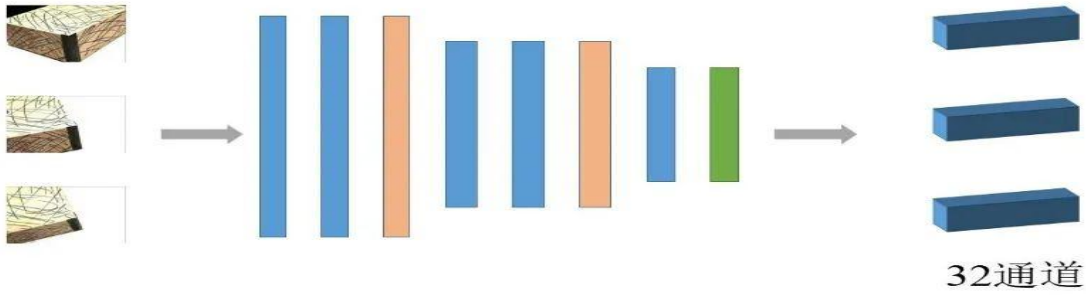
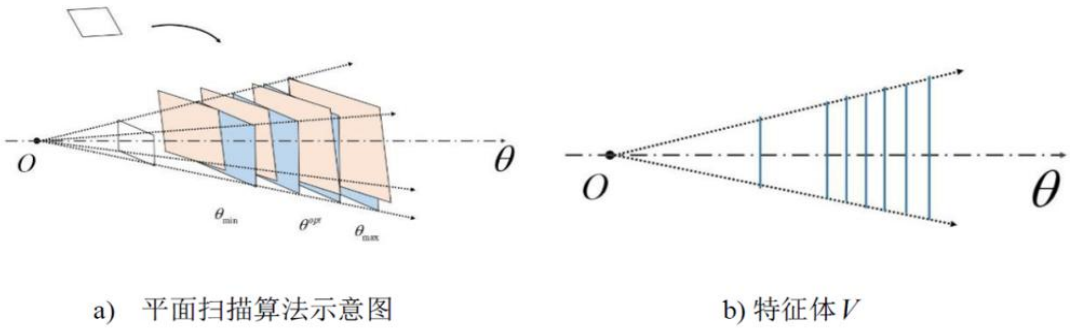


图 5 深度提取单元

深度特征指通过神经网络提取的影像特征，相比传统 SIFT、SURF 的特征有更好的匹配精度和效率。经过视角选择之后，输入已经配对的 N 张影像，即参考影像和候选集。首先利用一个八层的二维卷积神经网络提取立体像对的深度特征 Fi，输出 32 通道的特征图，并且各个图像提取过程的网络是权值共享的。

构造匹配代价

MVSNet 利用平面扫描算法构造参考影像的匹配代价，因为平面扫描算法适用于无纠正影像，且能达到实时计算差图的效果。通过深度特征抽取后，每张影像$I_i \in \{I_{ref} \cup A\}$可以得到一张对应的特征图$F_i$，根据先验的深度范围信息，对于参考影像$I_{ref}$，以其主光轴为扫描方向，将参考影像按照某一深度间隔θ_{scale}，从最小深度处θ_{min}，一直映射到最大深度处θ_{max}。可以得到一个处于不同深度间隔的相机锥体，为了方便计算光学一致性，利用插值的方法，使得每张投影的长宽一样。



a) 平面扫描算法示意图　　b) 特征体V

图 6 构造匹配代价单元

已知参考影像I_{ref}、候选集中影像的相机参数为：$\{K_i,\ R_i,\ t_i\},\ i = I_{ref} \cup A$。对于$I_{ref}$，将候选集中代表$I_j \in A$的特征图$F_j$投影到该相机锥体的不同深度中定义这个投影变换为：$X_{ref} = H_j(\theta)x_j$，$x_j$表示特征图坐标，$H_j(\theta)$表示对于第 j 个特征图，映射到深度θ的参考影像上的单应性矩阵。和式（1）类似：

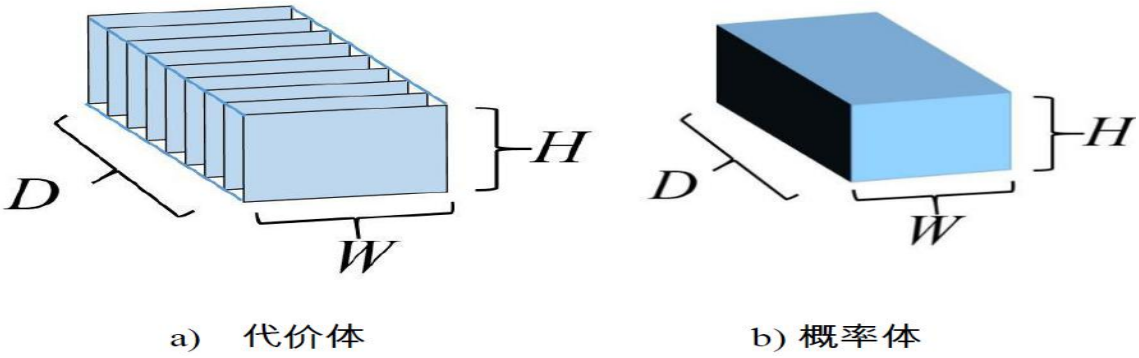
$$H_j(\theta) = K_j R_j (I - \frac{(t_{ref} - t_j)n_{ref}^T}{\theta}) R_{ref}^T K_{ref}^T$$

考虑到对亚像素的深度估计，以保证深度图平滑，该单应性矩阵是完全可以微分的。通过投影变换，N 张影像可形成 N 个特征体$\{V_i\}_{i=1}^N$,这个特征体就是匹配代价的表示。

代价积累

MVSNet 的代价累积通过构造代价体实现的。代价体是一个由长、宽与参考影像长宽一样的代价图在深度方向连接而成的三维结构(图 a)，在深度维度，每一个单位表示一个深度值。其中，某一深度的代价图

上面的像素表示参考影像同样的像素在相同深度处，与候选集影像的匹配代价。



a) 代价体　　b) 概率体

图 7 代价积累单元

深度估计

MVSNet 的深度估计是通过神经网络直接学习的。网络训练方法是，输入代价体 V 和对应深度图真值，利用 SoftMax 回归每一个像素在深度 0 处的概率，得到一个表示参考影像每个影像沿深度方向置信度的概率体 P，以此完成从代价到深度值的学习过程。

当已知概率体时，最简单的方法可以获取参考影像的所有像素在不同深度的概率图，按照赢者通吃原则直接估计深度图。然而，赢者通吃原则无法在亚像素级别估计深度，造成深度突变、不平滑情况。所以需要沿着概率体的深度方向，以深度期望值作为该像素的深度估计值，使得整个深度图中的不同部分内部较为平滑。

$$\theta_l = \sum_{\theta=\theta_{min}}^{\theta_{max}} \theta \times P(\theta)$$

其中，$P(\theta)$表示特征 l 在深度θ置信度。

4.2 算法实现过程

4.2.1 损失函数

大多数深度立体网络使用软参数操作对差异/深度输出进行回归。输出使用软参数操作，它可以被解释为沿深度方向的期望值。如果深度值在深度范围内被均匀采样，期望值的表述是有效的。然而在递归 MVSNet 中，我们应用反深度来采样。为了有效地处理深度范围较宽的重建，我们采用反深度法对深度值进行采样。我们将网络训练成一个具有交叉熵损失的多类分类问题，而不是将该问题作为一个回归任务。

分类问题的交叉熵损失：

$$Loss = \sum_{p}\left(\sum_{i=1}^{D} -p(i,p)\cdot logQ(i,p)\right)$$

其中p是空间图像坐标，$p(i,p)$是概率体积p中的一个体素。概率体积p中的一个体素。Q是地面实况二元占位体积，它是由地面真实深度图的单次编码产生的。$Q(i,p)$是与$p(i,p)$对应的体素。

4.3.2 实验参数

表 4 MVSNet 实验参数

数据集	DTU:http://roboimagedata.compute.dtu.dk	
训练参数	学习率	0.001

	epoch	100000
环境	TensorFlow	

4.3 实验结果

4.3.1 其他算法对比

表 5 MVSNet 与其他算法对比结果

	Mean Acc.	Mean comp.
Tola	0.342	1.19
Gipuma	0.283	0.873
Colmap	0.400	0.664
MVSNet	0.383	0.452

4.3.2 效果图

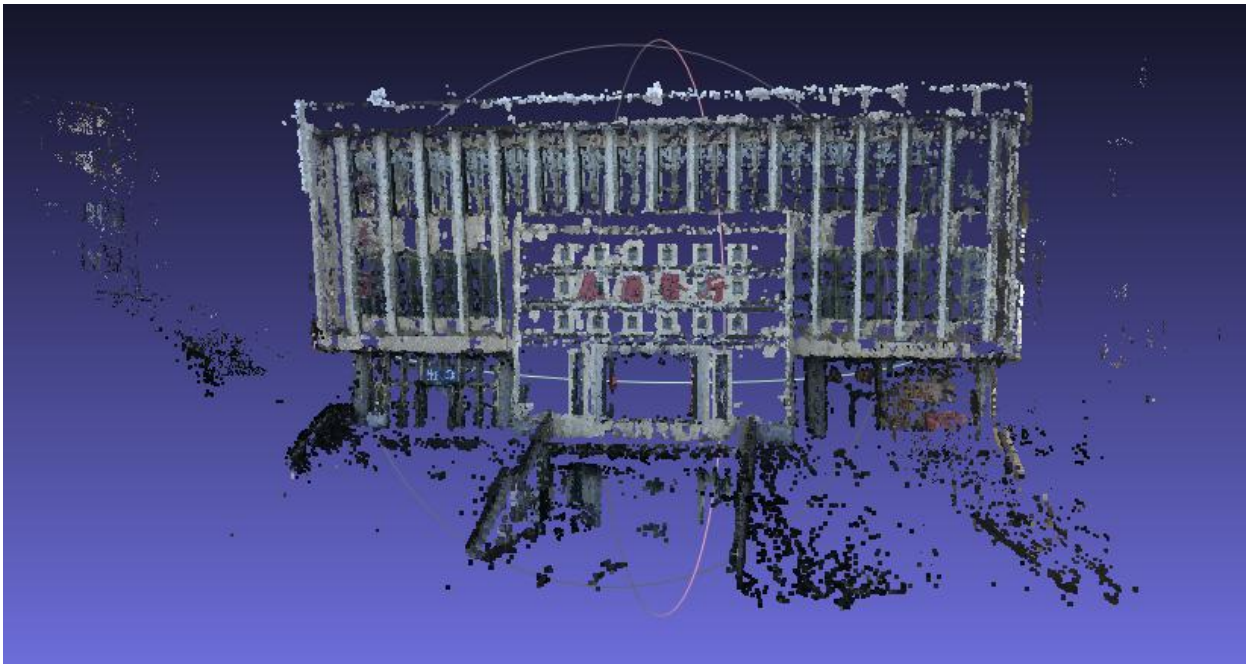


图 8 泰园重建效果图

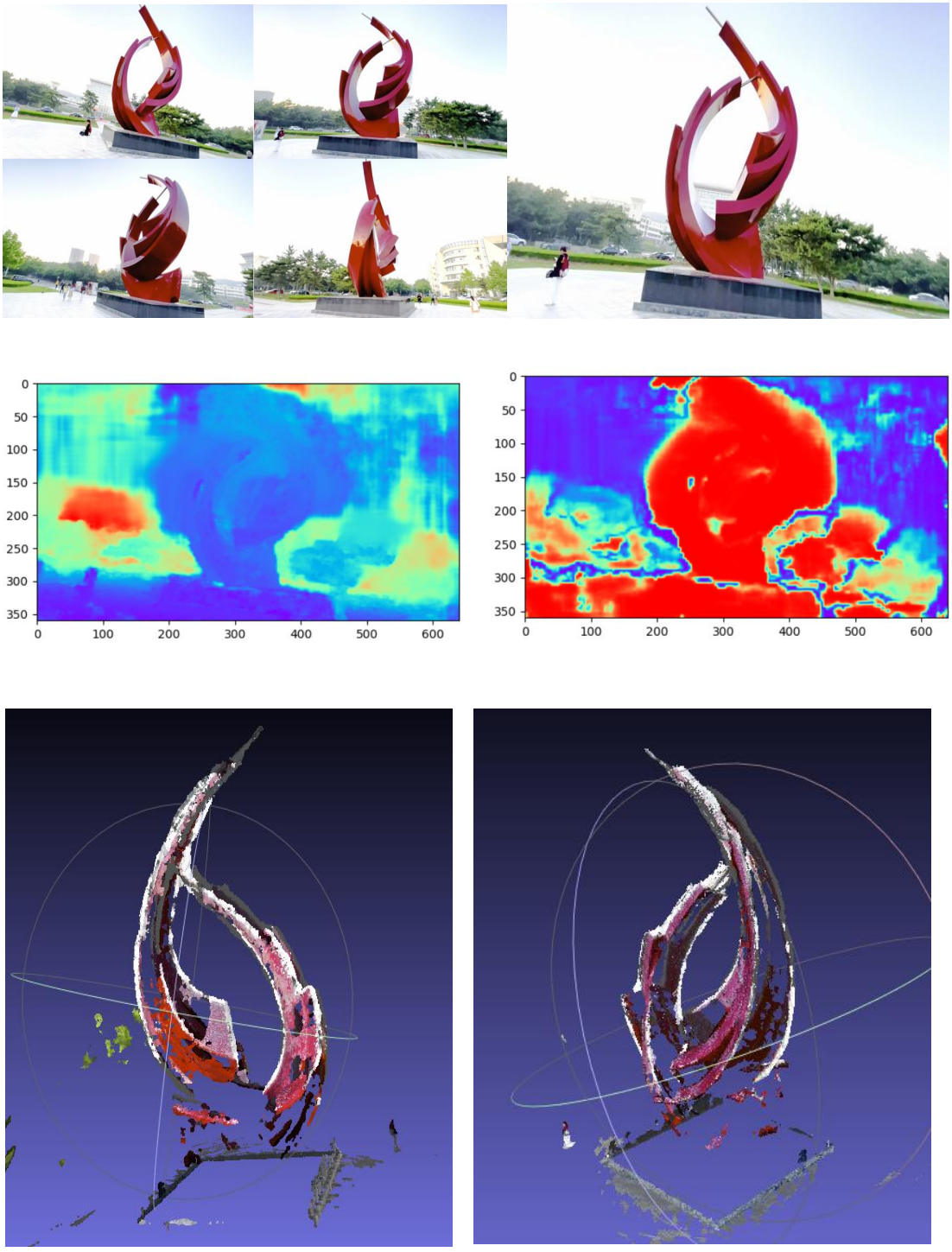


图 9 从左至右、从上至下依次是：输入的多角度图片中的几张、某一角度图片、上图对应的深度图、概率图，重建的点云图

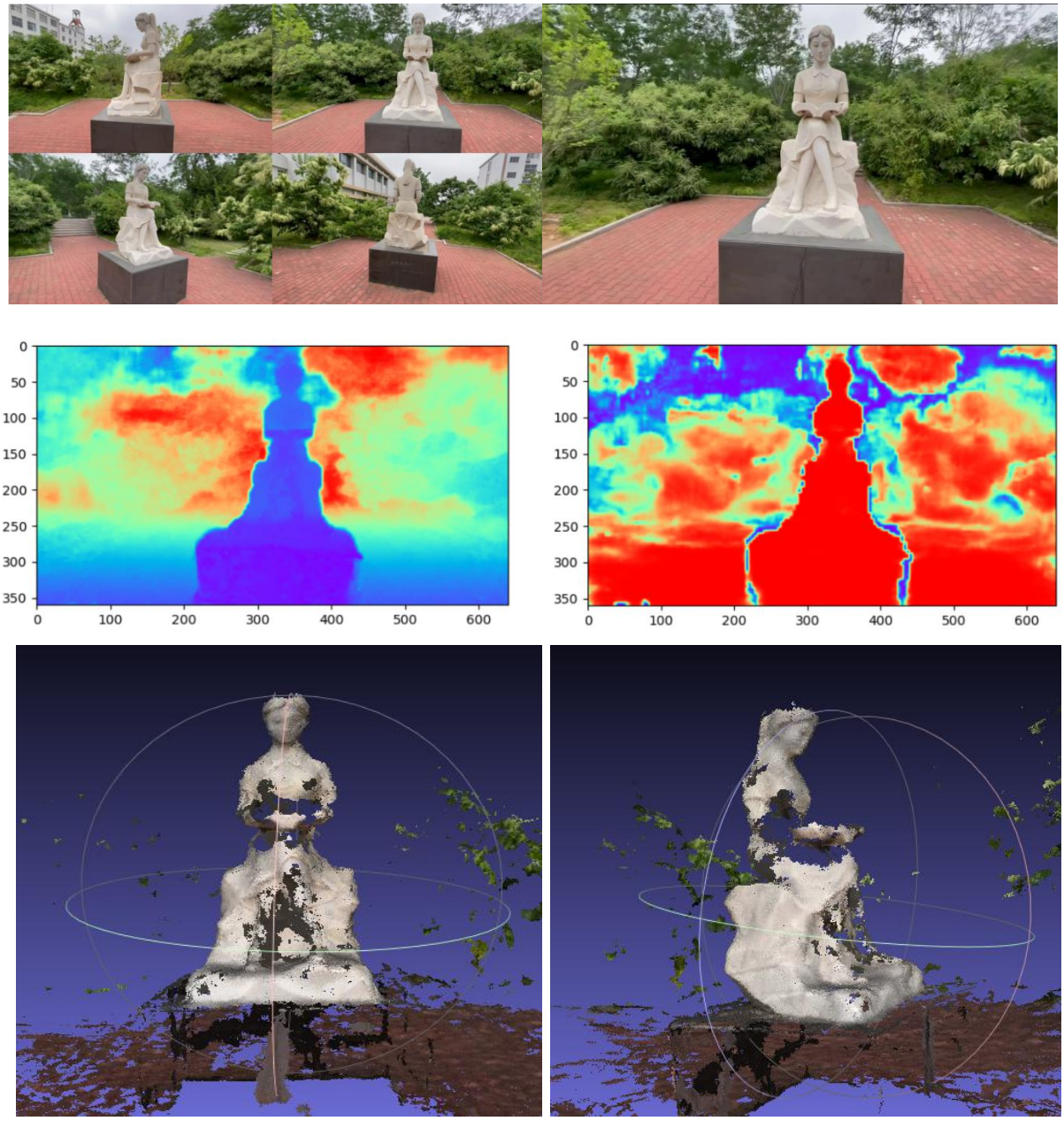


图 10 从左至右、从上至下依次是：输入的多角度图片中的几张、某一角度图片、上图对应的深度图、概率图，重建的点云图

5 算法对比

	LSTM	MVSNet
输入	单张图片，限制较小	多角度、多张图片，对输出效果影响大
速度	快	较慢
输出效果	较好反映物体结构，但缺乏表面细节	物体结构与表面效果较好

适用场景	对物体结构要求高且需要输出速度较快	要求真实反映物体结构与细节

附录

参考文献

[1] 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 International Publishing, 2016.

[2] Yao Y， Luo Z， Li S，et al. Recurrent MVSNet for High-resolution Multi-view Stereo Depth Inference[J]. IEEE, 2019.

[3] N. D. Campbell, G. Vogiatzis, C. Hern´andez, and R. Cipolla.Using multiple hypotheses to improve depth-maps for multiview stereo. European Conference on Computer Vision(ECCV), 2008.

[4] J.-R. Chang and Y.-S. Chen. Pyramid stereo matching network. Computer Vision and Pattern Recognition (CVPR),2018.