基于Fast-MVS的三维重建

**摘 要：** 三维重建技术已得到了广泛的应用。方法同样也层出不穷，我们将这些方法依据原理分为两类：基于传统多视图几何的三维重建算法；基于深度学习的三维重建算法。基于深度学习的三维重建算法相比于传统三维重建算法，在抗干扰性与完整性方面有着更好的表现，是随着神经网络技术的兴起，三维重建在自身领域的进步与发展。但由于深度学习框架的复杂性，相关算法耗时较长且效率较低，所以有研究提出了Fast-MVS算法，是在使用最广泛的基于深度学习进行稠密点云构建的MVSNet算法基础上进行的改进。其主要框架通过稀疏高分辨率深度图预测网络、深度图稠密化网络、高斯-牛顿精细化网络层层优化深度图，是一种新的由稀疏到稠密、由粗糙到精细的框架，本项目的实现就采用了该算法。

**关键词：**传统三维重建；特征点匹配；深度学习；Fast-MVS；点云配准

# 1 传统SFM算法

1.1 介绍

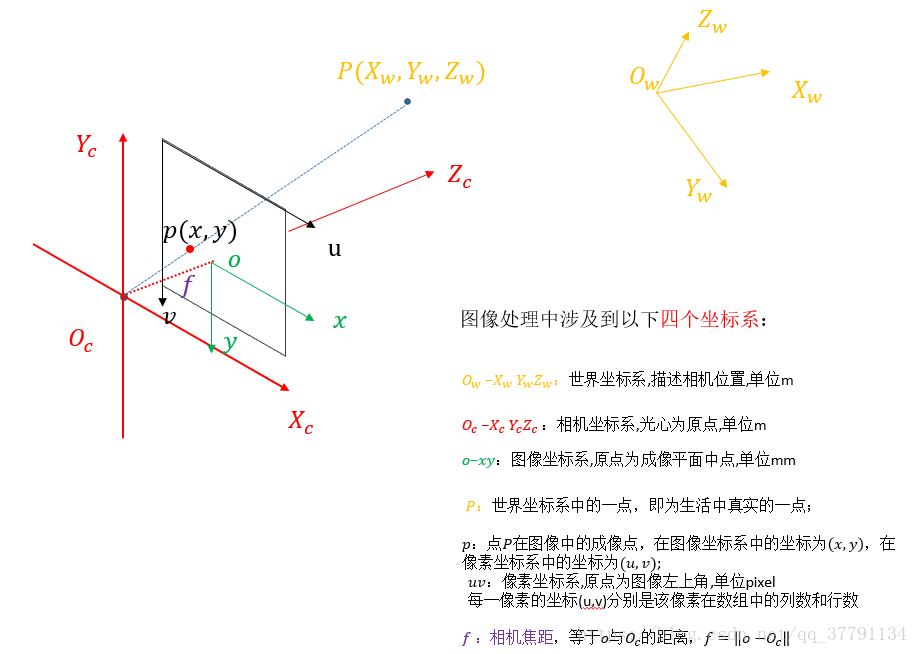
SFM（structure-from-motion）算法是基于多幅二维无序图片进行三维结构估计的离线算法。人的大脑可以从动的物体中取得其三维的信息，是因为大脑在动的2D图像中找到了匹配的地方，即Corresponding area（points）。然后通过匹配点之间的视差得到相对的深度信息，在这一点上，原理和基于Stereo的三维重建相同。

输入一系列二维图片，得到相机估计参数和场景几何信息，即稀疏的3D点云。

1.2 算法原理

相机的成像原理基于小孔成像原理实现，将生活中真实的一点，即世界坐标系的一点，通过层层坐标系的映射，最终形成了我们在计算机中看到的由像素构成的图像。所以想要找到真实物体与计算机图片间的联系，就要新构建出世界坐标系与像素坐标系间的转换关系。

像素坐标系是一个二维直角坐标系，反映了相机CCD/CMOS芯片中像素的排列情况。原点位于图像的左上角，轴、轴分别于像面的两边平行。像素坐标系中坐标轴的单位是像素（整数）。

世界坐标系也称为测量坐标系，是一个三维直角坐标系，以其为基准可以描述相机和待测物体的空间位置。世界坐标系的位置可以根据实际情况自由确定。

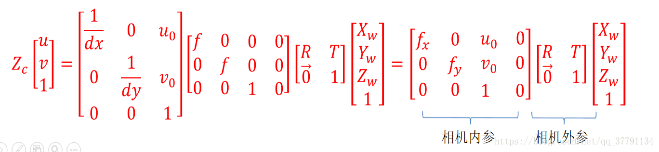
图表 1

相机外参是用于描述从世界坐标系变换到相机坐标系，可以描述相机的位置和观测物体的位置为宏观的坐标系，可以根据实际情况来确定，如果是单目相机，基本可以认为是重叠。如果是涉及多个相机，比如在图片拼接或者三维重建中，涉及到多个世界坐标系，就需要将世界坐标系进行统一，从而将不同的相机拍摄的图像注册到一个坐标系下来。

相机内参是基于相机设置，如焦距、最

高分辨率、传感器尺寸等计算出的内参矩阵。

有了外参矩阵与内参矩阵，就可以通过矩阵变换将世界坐标系转换为像素坐标系。



图表 2

1.3 算法流程

其主要过程是先从图片中提取焦距信息，然后利用SIFT等特征提取算法提取图像特征，用kd-tree模型去计算两张图片特征点之间的欧式距离进行特征点的匹配，从而找到特征点匹配个数达到要求的图像对。

1.3.1 图像的特征点提取

关于图像的特征点检测与提去，常见的算法有Harris, SIFT, SURF, ORB等。本项目使用的是SIFT算法。

SIFT的全称是Scale Invariant Feature Transform，尺度不变特征变换，是由加拿大教授David G.Lowe在1999年提出的特征检测算法，并于2003年对其完善总结。

SIFT特征检测算法的特点有：

1. 该算法检测提取出的SIFT特征是图像的局部特征，对旋转、尺度缩放、亮度变化保持不变性，对视角变化、仿射变换、噪声也保持一定程度的稳定性。

2. 其信息量丰富，适用于在海量特征数据库中进行匹配。

3. 其提取的特征具有多量性，少数物体也可以产生大量SIFT特征。

4. 该算法具有高速性，经优化的SIFT匹配算法甚至可以达到实时性。



图表 3

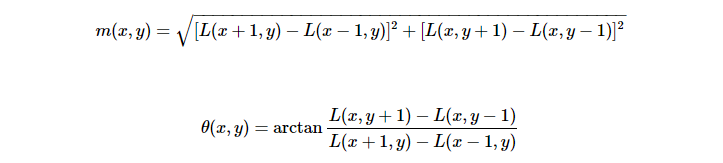
SIFT特征检测的步骤：

1. 在DOG尺度空间中获取极值点。

寻找DOG极值点时，每一个像素点和它所有的相邻点比较，当其大于（或小于）它的图像域和尺度域的所有相邻点时，即为极值点。如下图所示，比较的范围是个的立方体：中间的检测点和它同尺度的8个相邻点，以及和上下相邻尺度对应的 个点——共26个点比较，以确保在尺度空间和二维图像空间都检测到极值点。

2. 特征点方向估计。

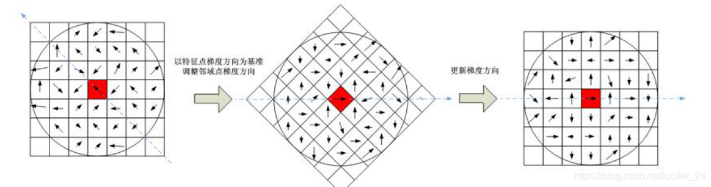
计算以特征点为中心、以为半径的区域图像的幅角和幅值，每个点的梯度的模以及方向。



图表 4

3. 生成特征描述

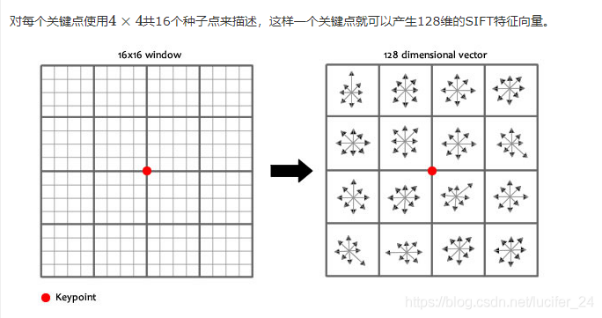
为了保证特征矢量的旋转不变性，要以特征点为中心，在附近邻域内将坐标轴旋转（特征点的主方向）角度，即将坐标轴旋转为特征点的主方向。



图表 5

将旋转后区域划分为（为2或者4，通常取4）个子区域（每个区域间隔为像元），在子区域内计算8个方向的梯度直方图，绘制每个方向梯度方向的累加值，形成一个种子点。

与求主方向不同的是，此时，每个子区域梯度方向直方图将划分为8个方向区间，每个区间为。即每个种子点有8个方向区间的梯度强度信息。由于存在，即个子区域，所以最终共有 个数据，形成128维SIFT特征矢量。



图表 6

1.3.2 特征点匹配

图像的特征点匹配就是在两幅图像中找出有效的匹配点对，它可以分为3个步骤：首先是从图像中检测出特征点，然后根据特征点的邻域信息描述特征点，最后在原图像和待匹配图像中计算特征点之间的对应关系。

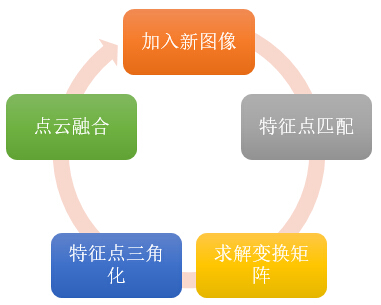
在得到了SIFT特征点的集合后。接下来生成物体目标描述要做的就是对特征点集合进行数据组织，形成一种特殊的表示，其作用是为了加速特征点匹配的过程。所谓的特征点匹配本质上是一个通过距离函数（例如欧式距离）在高维矢量之间进行相似性检索的问题，简单来讲就是范围查询或者K近邻查询的问题。

范围查询就是给定查询点和查询距离阈值，从数据集中找出所有与查询点距离小于查询距离阈值的数据；K近邻查询就是给定查询点和正整数K，从数据集中找到距离查询点最近的K个数据，当K=1时，它就是最近邻查询。

1.3.3 相对变换矩阵

在前面我们已经介绍了外参矩阵和内参矩阵，这里，我们将其分别定义为变化矩阵R与T。通过相关函数获得本征矩阵并将其分解，获取相对变换R和T。

现在已经知道了两个相机之间的变换矩阵R和T，还有每一对匹配点的坐标。三维重建就是通过这些已知信息还原匹配点在空间当中的坐标，然后三角化重建三维模型的过程。



图表 7

# 2 传统MVS算法

2.1 介绍

MVS（Multi-view Stereo），多视角立体视觉，是对立体视觉的推广，其方法即为给定从不同视角拍摄的图像以及其对应的相机几何，通过稠密匹配的方式恢复出物体以及场景的三维结构。是一种从已知相机内外参数的多个视角的彩色影像中，利用立体匹配算法恢复立体结构的三维视觉技术。传统的MVS算法基于handcrafted像素匹配算法具有相当高的重建准确度。

但对于纹理稀疏、非漫反射的物体表面难以得到令人满意的重建效果。 因此，MVS算法的重建完整度还有很大的提升空间。

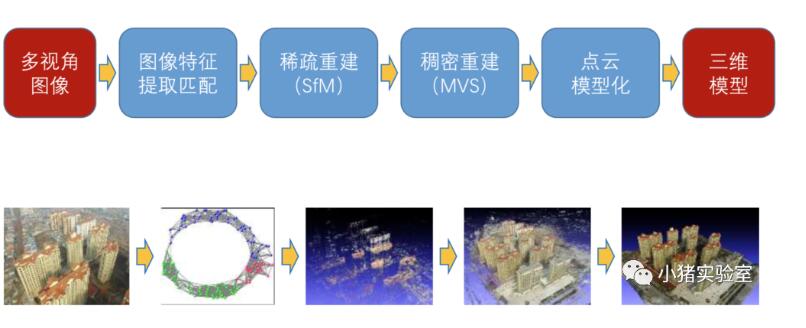
2.2 SFM+MVS的传统三维重建

通过前面的介绍，我们可以了解到，SFM是先从检测图像中提取2D特征表征。这些图像特征表示为图像中的一个小区域，其特征特点依靠高度纹理区域，且拥有粗糙的几何形状。

但是这些场景特征需要在整个场景中唯一。故而通过这些唯一的特征只能生成稀疏的深度图。只有当图像之间找到很多匹配项时，才可以计算出图像之间的3D变换矩阵，从而有效地给出两个相机之间地相对3D位置。

所以基于SFM算法只能实现稀疏点云重建的原因，传统的三维重建算法一般都是通过SFM+MVS实现。

MVS算法可以细化通过SFM技术获得的网格，从而产生密集重构。进而形成稠密点云，完成完整的三维重建。



图表 8

2.3 传统MVS算法的基本流程

对于传统的多视角立体视觉算法来说，PMVS（patch-based MVS）是其中的经典，这里我们就基于此算法对传统MVS算法对三维重建的实现做出介绍。

通过前面的SFM算法，我们已经获得了相机的参数，可以利用这些参数使用基于面片的三维多视角立体视觉算法（PMVS）重建出稠密的点云。

2.3.1 相关概念的定义

首先我们来明确几个概念：

1. 面片（patch）

面片p是一个近似的正切与重建物体表面的一个小矩形，他的一边平行于参考相机的x轴。对于一个面片p，他的几何特征有中心点、单位法向，中心点可定义成 ，单位法向可定义成 ，同时该向量指向相机的光心。

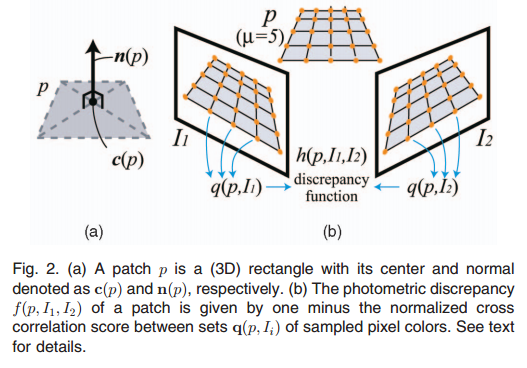
面片p对应一个参考图像 ，在中p是可见的。针对p有扩展矩形，p在 中的投影是 大小的。

2. 灰度一致性函数（Photometric Discrepancy Function）

首先，我们定义图像集合 表示所有在自身图像中可见面片p的图像集合，显然这里有 ，那么，灰度一致性函数则定义为，

其中 。在这里，是指除中去掉的其他元素，是指和的灰度一致性函数。

计算过程是首先把面片p划分为的小格，然后通过双线性差值的方法，对p在的图像上的投影进行差值，得到像素灰度，最后用1减去和的NCC值。



图表 9

3．面片的优化

面片优化的目的就是恢复那些较小的面片，每个面片的重建过程分为以下两步：

1）初始化面片的相关参数中心点 ,法向量 ,可视化图像集 和参考图像。

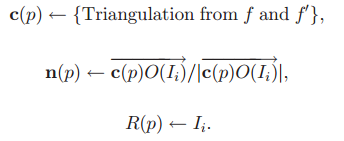
2）优化几何参数和 。

其中几何参数和的优化是通过最小化光度一致性分数而得到的。为了简化计算，将约束在某一条光线上，这样p在其对应的可视图集V∗(p)中某个图像的投影位置就不会变，因此降低了p的自由度和只能求出一个深度。是由欧拉角决定的，可以用共轭梯度法求解这个优化问题。

2.3.1 面片的重构

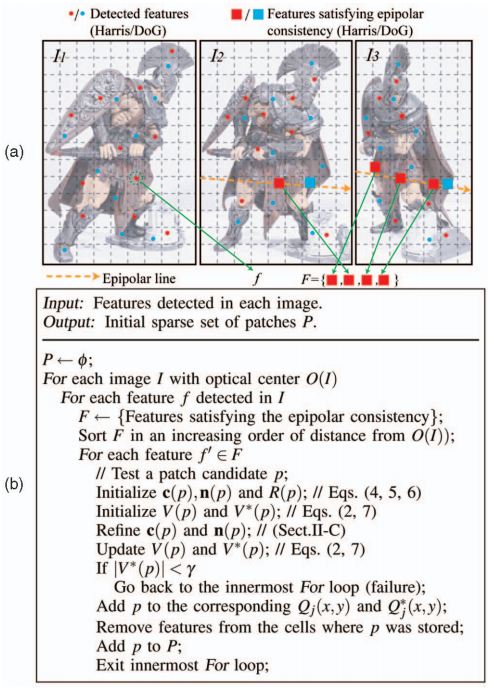
1. 初始化特征匹配

首先用DOG和Harris来提取图像的角点特征，即为每幅图像的特征点。对于图像，以及其对应的光心 ，该图像中的特征点 ，通过允许有两个像素误差的极线约束找到它在其他图像中的同种类型的特征点，构成匹配点对。然后用这些匹配点对使用三角化的方法生成一系列三维空间点，然后将这些点按照距离从小到大顺序进行排列，然后依次尝试生成面片，直到成功。



同样的获得与前面所讲的方法相同。这样面片的信息都已经初始化过了，然后根据前面所讲的对和进行优化，并将优化过后的和带入

中对和进行更新。如果，集认为面片p的可视图像至少为r幅，即接受面片p，重建成功。并将面片p存储到对应的可视图像的图像块中。



图表 10

1. 面片生成（patch expansion）

面片生成的目的就是保证每个图像块至少对应一个面片。通过上面生成的面片，重复的生成新的面片。具体来说就是给定一个面片p，首先获得一个满足一定条件的邻域图像块集合，然后进行面片生成的过程。

当面片和满足如下公式：

的时候，即判定两者为临近关系。

当存在一个面片其所属的图像块满足，同时和属于近邻关系时，此时将从中删除，不对他进行面片生成。同时，即使这个条件没有满足，只要不为空，也不需要再此图像块上执行生成操作。

对于中其余的图像块，将会执行面片生成的操作流程以生成新的面片。首先用的相应变量初始化的、和，对于的初始值为穿过的可视光线与面片所在平面的交点。

使用如下公式：

由得到。

再由上面提到的方法对和进行优化。在优化的过程中，将约束在一条直线上，使得在图像上的位置不会改变，始终对应的是。

优化完成后，给加上一组图像，这些图像块些根据深度测试判断为对其应该是可见的，并根据此更新。

最终如果，则判定是可接受的，即生成成功。

3． 面片过滤（patch filtering）

在面片的重建过程中，可能会生成一些误差较大的面片，因此需要过滤来确保面片的准确性。

第一个过滤器是通过可视一致性进行过滤，另表示与当前可视信息不连续的面片集合，所谓的不连续就是和两个面片不属于近邻关系，但是却存在于同一个中。对于中的面片，如果满足

则将其过滤掉。

直观上来讲，如果p是一个异常值，那么和都会比较小，这样p一般都会被过滤掉。

第二个过滤器同样也是考虑可视一致性，不过会更加严格，对于每个面片p，我们计算他通过深度测试得到的可视图像的总数，如果数目小于r，那么则认为p是异常值，从而过滤掉。

# 3基于深度学习的MVS算法

3.1 介绍

从上文可知，传统的MVS算法通常使用人为设计的指标和规则（如归一化互相关匹配、半全局匹配等）来计算高密度3D点云的对应关系并恢复三维点信息。这些设计方法在理想的实验环境下展现出很高的准确率，但往往具有共同的局限性——易受低纹理、反光和局部反射的干扰，使得实际的匹配步骤困难重重，并最终导致重构的3D立体或深度图不完整。因此，尽管传统方法可以达到令人满意的准确率，但在重构完整性这一方面还有很大的改善空间。

近年来神经网络的兴起，引起了众多学者对立体重构改善的兴趣。站在理论角度，深度学习的方法通常含有全局的语义信息，因此可以学习到高亮度或反射的信息，从而更好地进行2维到3维的匹配；同时，可以基于深度学习对两个视觉图进行立体匹配，代替手工设计标准的方法，便捷高效。

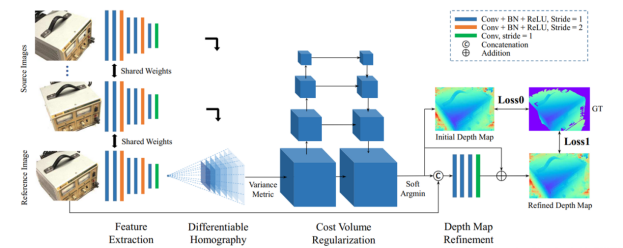
深度学习方法的引入使得MVS展示出令人满意的结果，本文选取最经典的MVSNet进行详细介绍。

3.2 MVSNet

香港科技大学权龙教授团队提出的MVSNet开启了利用深度学习做多视图三维重建的先河。这是一种端到端的学习训练，完成从多张视觉图片推导出深度图的过程。它将复杂的MVS重建，划分为逐视图的深度估计问题，使得大规模重建成为可能。

从MVSNet网络架构图中可看出，在一次运算过程中，只需要输入一张reference image和几张source image，依次经过二维特征提取、单应性变换、代价体正则化、深度图提炼，即可预测出reference image上每个像素的深度信息，并衡量估计的置信度。

MVSNet的训练和评估都是基于大尺寸的DTU数据集，大量的实验证明，无论是在重构的完整性还是质量上，均超出了所有竞争方法。除此之外，在Tanks 和Temples数据集上，还体现了很好的泛化能力，在速度上也完全胜出当前先进的一些算法。



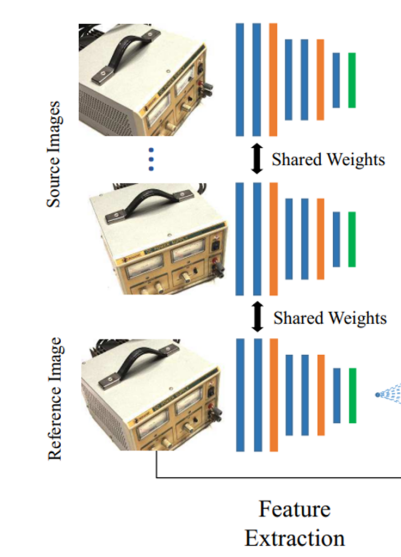
图表 11

3.2.1 二维特征提取

MVSNet的第一步是对每张输入图像进行特征提取。

特征提取网络使用8层的2D卷积神经网络，其第3层、第6层的卷积步长设为2。与一般的匹配网络相同，特征提取网络的参数共享权重。

2D卷积神经网络的输出为N个32通道的特征图。与原图相比，特征图在每个维度上进行了4倍的下采样。虽然特征图的尺寸变小了，但所有图片的像素信息已经被编码到了抽取出来的特征图之中，相比于使用原图进行特征匹配，这种方法明显提高了重构的质量。



图表 12

3.2.2 单应性变换

接下来是基于特征图和相机参数构建代价体（cost volumes）。

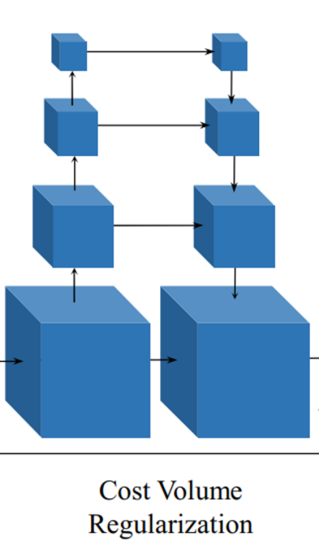
由于我们仅需求reference image对应的深度图，所以要将source image的视角转换为reference image的视角。MVSNet提出了一个基于单应性变换的转换方法，同时将相机的几何关系结合到神经网络中，从而将不同视角转换为同一个视角。

单应性变换的具体过程，即把source image视角上的特征图映射到reference image相机视锥中的256个深度平面上，总存在一个深度与其对应。其中，第i个特征映射到reference image相机视锥中深度d平面的单应矩阵计算公式如下：

N张映射过来的特征图便形成特征体（feature volumes），将多个特征体聚合即为一个代价体（cost volumes）。

3.2.3 代价体正则化

代价体经过一个四级的U-Net结构正则化来生成概率体（probability volume）。其含义就是在每个深度下，每个像素的可能性大小。



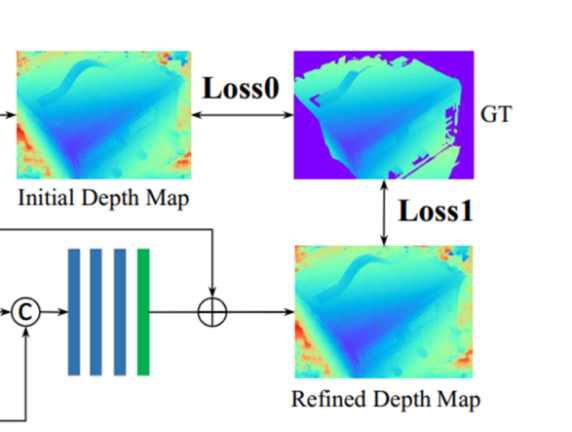
图表 13

受到深度学习以及传统MVS做法的影响，MVSNet采用了多尺度3D卷积层对代价体进行正则化。在该3D卷积层网络中，有四个缩放的尺寸，采用编码-解码的架构，使用较小的内存，去合并周边的信息。为了进一步减少内存的消耗，在第一个3D卷积之后把32通道的代价体减少到8个通道，并且在第2层和第3层间缩小图片的尺寸。最后输出一个单通道的体（volume），并利用softmax操作进行概率归一化。

3.2.4 深度图提炼

选择概率最大的深度作为点的深度，是最简单的一种方法，但这只适用于概率分布只有一个尖锐的波峰的理想情况。在有多个波峰的情况下，该策略就不是特别有效，所以MVSNet采用在深度方向上计算期望，即所有假设深度值的加权和：

尽管从概率体中恢复出深度图是一个直接而合理的过程， 但仍然会存在一些问题。 由于正则化过程中具有较大的感受野，会造成重建深度图的边界过平滑，这也是在语义分割和图像抠图（Image Matting） 中存在的常见问题。受到图像抠图算法的启发， MVSNet在末端添加了一个深度残差学习的网络。



图表 14

该网络把reference image变成和初始深度图的大小相同，接着组合成4个通道的特征图，将其当作网络的输入，经过11层通道数为32的2D卷积结构，输出一个单通道的深度图（Refined Depth Map）。在2D卷积的最后一层没有使用BN，ReLU以及残差网络单元。同时，为了防止一定范围内的偏差，在送入网络的时候，把像素转化到[0,1]，等提炼之后再还原。

另外，注意到损失（loss）的选择，考虑其来源于两方面：GT（ground truth）与初始深度图的计算损失（Loss0），以及与提炼之后的深度图的计算损失（Loss1）。

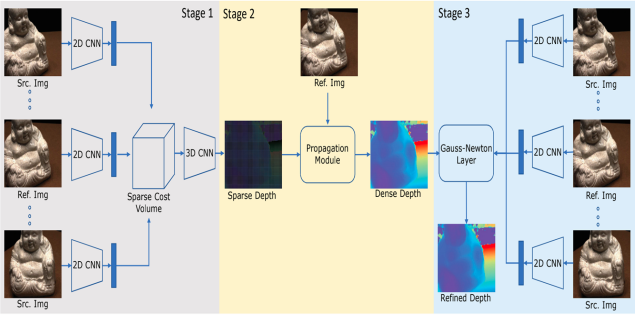
# 4 MVSNet的改进—Fast-MVSNet

4.1 网络概览

基于深度学习的MVS方法在提高重建质量方面取得了巨大成功，但除了重建质量，效率也是现实场景中重建的一个重要因素。因此，我们采取改进的Fast-MVSNet网络进行实际场景的三维重建，这是一种新的由稀疏到稠密、由粗糙到精细的框架，适用于快速和准确的多视图深度估计。

具体而言，在Fast-MVSNet中，首先以较低的成本构造一个稀疏的代价体来学习一个高分辨率的稀疏深度图，接着利用一个小型卷积神经网络对局部区域内像素的深度依赖进行编码，以稠密化该稀疏高分辨率深度图，最后利用可微的高斯-牛顿层来进一步优化深度图，来提高深度估计的准确性。

由此，Fast-MVSNet的网络架构可分为3部分：稀疏高分辨率深度图预测网络、深度图稠密化网络、高斯-牛顿精细化网络。



图表 15

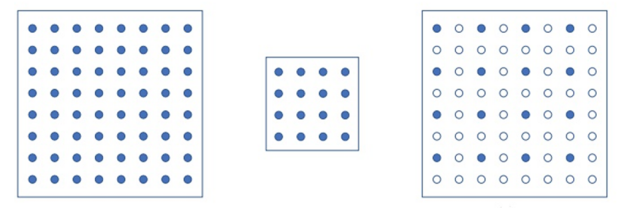
下面将对这三部分进行详细介绍。

4.2 稀疏高分辨率深度图

深度图主要有以下三种构建方法：

1. 高分辨率深度图；
2. 低分辨率深度图；
3. 稀疏的高分辨率深度图；

下图从左至右依次展示了三种深度图的特点。可以看出，高分辨率深度图往往内存成本高，低分辨率深度图容易丢失细节，稀疏的高分辨率深度图则规避了二者的缺陷，在尽可能保持细节的前提下，减小了内存占用和计算成本。另外，稀疏的代价体表示使得3D CNN在空间域的表现类似于dilation为2的空洞卷积，因此在正则化中，有融合更大空间信息的能力。



图表 16

稀疏高分辨率深度图的构建方法与MVSNet获取初始深度图的做法大致相同，使用8层2D卷积网络提取32通道的图像特征，不同之处在于减小了代价体的尺寸和深度采样数：代价体的尺寸缩小至，深度采样平面数为48或96。

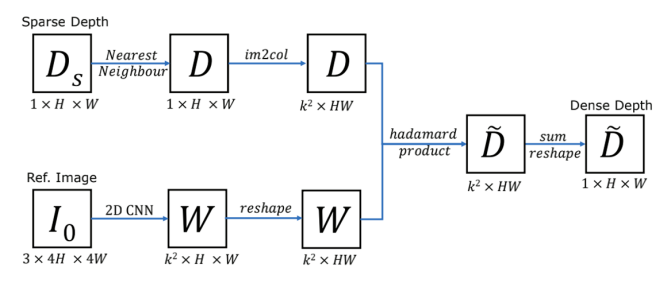
4.3 深度图稠密化

Fast-MVSNet的第二步，对稀疏高分辨率深度图进行扩展，稠密化为密集深度图。最简单的想法是采用最近邻算法，但是它没有考虑原始图像的信息，因此在边界的表现可能不够理想。网络选取了联合双边上采样，它使用原始高分辨率图像的信息作为指导：

其中，是空间滤波器核，是范围滤波器核，为周围的区域，是归一化项。对于不同的场景，这两个滤波器核可能不同，需要手动调整。因此，用权重代替他们，并用简单网络学习权值，所以就有了如下的形式：

是简单网络CNN的输出，以数据驱动的方式学习。虽然没有显式地计算空间信息，但它确实由网络隐式编码。此外，当预测不同位置的不同权重时，可以看作是标准双边上采样器的一般化，它对每个位置应用一个固定的核。

概括来说，最终使用的方法是：从reference image中提取特征图，在特征图的通道中存放卷积核参数，因此最终的通道数。然后用卷积网络提取出的卷积核对稀疏高分辨率深度图进行卷积，得到扩展的稠密深度图。



图表 17

具体步骤细化如下：

1. 使用最临近算法转化为稠密深度图（填补空洞），将每个点邻域的所有点依次写入通道（形成k²的通道，存储其领域点，卷积的一种计算方式）；
2. 同时参考图像提取特征，输出通道数为，得到上每个位置的卷积核，再reshape为的二维，方便计算；
3. Hadamard product：特征图和对应位置相乘；
4. sum相加（实际是卷积运算拆成了两步，先乘积，再相加，将变为1，再reshape为，得到密集深度图；

4.4 高斯-牛顿精细化网络

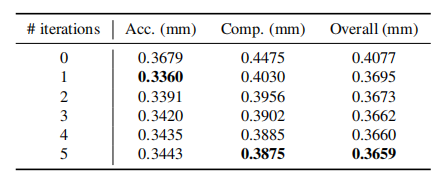
此时，已从reference image和多张source image中提取出了多张特征图，但预测得到的深度图精度是不够的。接着，继续使用扩展深度图预测出的深度，将参考图像上的点投影到每张原图像上处，计算与处特征图取值的差异。若深度估计准确，那么和投影处的特征应该相似，即差值小。

因此，目标即简化为最小化误差函数：

其中(，表示在原图像的投影点，表示reference image在点的深度估计（扩展深度图的取值），和分别为原图像和参考图像的特征图，为特征图在处的特征。

高斯-牛顿法具有可差分、不需要可学习参数、收敛速度快等诸多优点，实验发现，它只需更新一步就可以快速收敛。

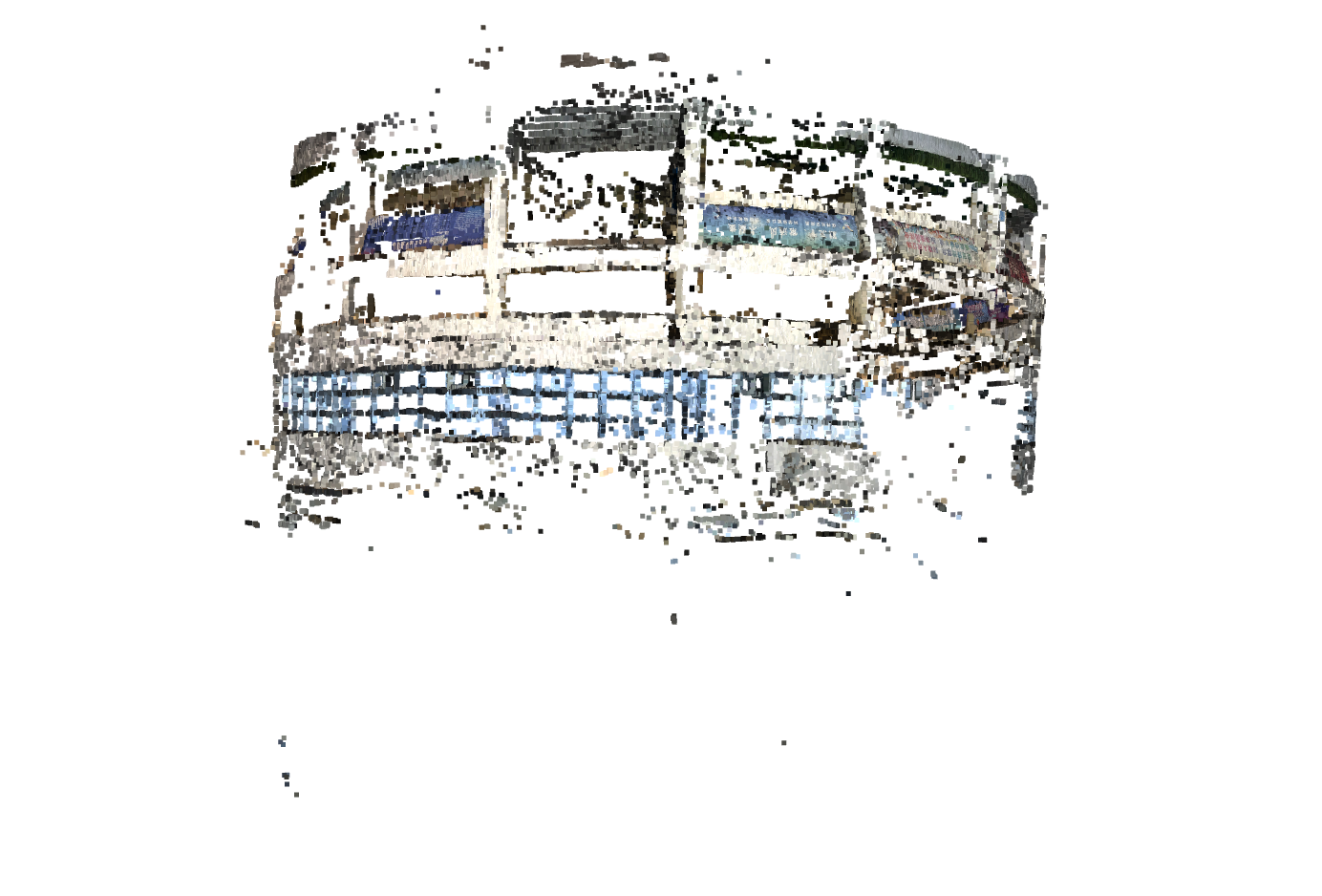
接着，对高斯牛顿精细化进行消融实验，并进行了更多的迭代。结果如下表所示，高斯-牛顿精化能显著提高重建质量。然而，采用高斯-牛顿法迭代次数多，性能改善甚微。因此，在Gauss-Newton精细化中，只使用一次迭代即可。



用高斯-牛顿法最小化的具体步骤如下：

1. 计算残差（观察值与估计值之间的差）；
2. 对每个残差，计算他们对的一阶导数；
3. 最后，可以得到当前深度的增量δ，其中J是雅可比矩阵的叠加，r是残差向量的叠加；
4. 相加得到精细化后的深度图；

# 5 结果展示







参考文献

[1]Yasutaka Furukawa; Carlos Hernández. Multi-View Stereo: A Tutorial[M].Now Foundations and Trends.2015.

[2] Yu Z , Gao S . Fast-MVSNet: Sparse-to-Dense Multi-View Stereo With Learned Propagation and Gauss-Newton Refinement[J]. 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020.

[3] Yao Y , Luo Z , Li S , et al. MVSNet: Depth Inference for Unstructured Multi-view Stereo[J]. 2018.