

清华 大学

综合 论文 训 练

题目：基于机器视觉的三维室内场景
建模方法研究

系 别：计算机科学与技术系

专 业：计算机科学与技术

姓 名：黄家晖

指导教师：胡事民教授

2018 年 6 月 10 日

关于学位论文使用授权的说明

本人完全了解清华大学有关保留、使用学位论文的规定，即：学校有权保留学位论文的复印件，允许该论文被查阅和借阅；学校可以公布该论文的全部或部分内容，可以采用影印、缩印或其他复制手段保存该论文。

(涉密的学位论文在解密后应遵守此规定)

签 名: 黄永晖 导师签名: 胡秉民 日 期: 2018.6.19

中文摘要

三维室内场景建模是图形学领域研究中的热点问题，同时也与虚拟现实、室内导航等实际商业应用有着密切的联系。然而，现有的三维重建方法大多需要复杂昂贵的数据采集设备，难以普及到普通用户。本文提出了一种基于机器视觉进行建模的方法，仅需要单张 RGB 图像作为输入，利用机器视觉的方法自动估计房间几何形状、识别家具位置并预测其形状和姿态，最终输出室内三维模型。

关键词：图形学；三维重建；室内场景；机器视觉

ABSTRACT

3D indoor scene modeling is a subject undergoing intense study in Graphics community. Nevertheless, most 3D reconstruction algorithms available nowadays rely heavily on expensive devices and sensors, rendering them barely applicable to everyday use. In the thesis, a novel modeling approach based on Computer Vision is proposed. The algorithm takes in RGB image as the single input. The room geometry, location and pose of the furnitures are automatically estimated and optimized.

Keywords: Graphics; 3D Reconstruction; Indoor Scene; Computer Vision

目 录

第 1 章 引言	1
1.1 研究背景	1
1.2 图形学与三维场景建模	2
1.3 机器视觉与图形学的交叉	4
1.4 课题意义与价值.....	4
第 2 章 文献综述	6
2.1 三维场景重建	6
2.1.1 基于深度图像的场景重建.....	6
2.1.2 基于彩色图像的场景重建.....	7
2.2 房间布局估计	8
2.3 物体检测与姿态计算	9
2.4 模型检索与生成.....	10
第 3 章 房间几何估计	12
3.1 问题分析	12
3.2 关键点预测网络.....	12
3.2.1 分割中继监督	14
3.2.2 损失函数定义	15
3.3 相机参数估计与场景重构	15
3.3.1 综合语义信息与低阶特征.....	16
3.3.2 相机内参矩阵估计	17
第 4 章 家具检索与摆放优化	20
4.1 家具识别、检索与位姿估计.....	20
4.1.1 图像和形状共存隐空间	21
4.1.2 多任务姿态预测.....	22
4.1.3 三维坐标恢复	23
4.2 基于概率的摆放优化	25
4.3 物件与场景贴图.....	27

第 5 章 实验验证	29
5.1 几何布局估计	29
5.1.1 实现细节	29
5.1.2 对比分析	29
5.1.3 重建效果展示	32
5.2 家具检索与优化	32
5.2.1 隐式空间构造	34
5.2.2 多任务预测	36
5.2.3 摆放优化	37
5.3 示例展示	39
第 6 章 结论	47
插图索引	49
表格索引	51
参考文献	52
致 谢	57
声 明	58
附录 A 外文资料的调研阅读报告或书面翻译	59
A.1 简介	60
A.2 相关工作	61
A.2.1 三维重建的颜色优化	62
A.3 方法概览	63
A.4 基于基本几何体的抽象	64
A.4.1 基于单帧的平面检测	64
A.4.2 基本体分类	65
A.4.3 结构优化	65
A.5 纹理优化	66
A.5.1 基于颜色的几何优化	66
A.5.2 颜色迁移优化	67
A.5.3 纹理映射对齐优化	68
A.5.4 纹理锐化	69

A.6 场景补全	71
A.6.1 几何补全	71
A.6.2 纹理补全	72
A.7 网格模型生成	73
A.8 结果	75
A.8.1 方法局限	77
A.9 总结	78
在学期间参加课题的研究成果	81

第1章 引言

1.1 研究背景

随着计算机学科的发展以及计算能力的不断增强，计算机算法被应用于各式各样的生活场景中，为方便人们的生活提供帮助。在这些应用中，自动驾驶、室内导航、虚拟现实（VR）和增强现实（AR）等是目前研究的热点问题，受到了国内外许多科研工作者们的关注。

为了实现这些应用，计算机必须采集和获取周边环境的三维表示，并对三维数据进行分析和处理，最终才能反馈给用户准确的结果。例如，在增强现实应用中，最终任务是将虚拟的三维模型和真实的摄像机录像进行结合，并对三维模型加以正确的光照、阴影，算法需要准确地计算模型的旋转角度和摆放平面，并实时地根据真实世界的视频算出物体的移动情况以及相机角度的变化情况；在电视游戏或是虚拟现实应用中，游戏的设计者往往需要花费很长的时间精心设计游戏的场景内容，有时建模师需要根据给定的照片和说明进行建模，在仅有图片参考的应用场景中，建模算法不仅需要找出与图片最为接近的内容组合，还需要求解不同物体的偏转角度和大致形状。

这些高级的应用都对现有的算法提出了很大的挑战，但同时也带来了新的研究课题和商业机遇。微软（Microsoft）公司在 2010 年推出了 Kinect 输入设备，可以实现动作的捕捉，同时也可以实现深度图像的获取。Kinect v2 设备采用“Time of Flight”技术通过红外光获取每一个视频像素的深度值（这里的深度代表像素点对应物理世界中的三维点到相机的测量距离），并且由于其低廉的价格让最终的民用变为可能，同时催生了许多诸如 KinectFusion^[1], BundleFusion^[2] 等根据 RGB-D 图像重建三维场景研究。著名科技公司 Matterport，目前提供专业的三维重建解决方案，通过该公司的独有技术和专有设备 Matterport PRO 3D 进军销售、游戏、房地产、旅游等各个行业，为用户提供高质量的重建服务。

除了上述专有设备之外，还有许多研究试图通过随处可得的彩色图像直接进行三维重建。这些算法无需借助复杂设备，通过提取图像中的稀疏特征，并跨图像地进行匹配和计算，最终能够得到各帧的相机姿态和物体的几何模型。对于小型的物体来说，一般通过 SfM 技术围绕物体拍摄多张照片进行重建；而对于大型的场景，可以借助 SLAM 技术同时进行相机定位和场景制图，得到稀疏的点

云，目前较好的研究包括：ShapeFit^[3]，鲁棒相机位置估计算法^[4]，LSD-SLAM^[5]，ORB-SLAM2^[6]等。

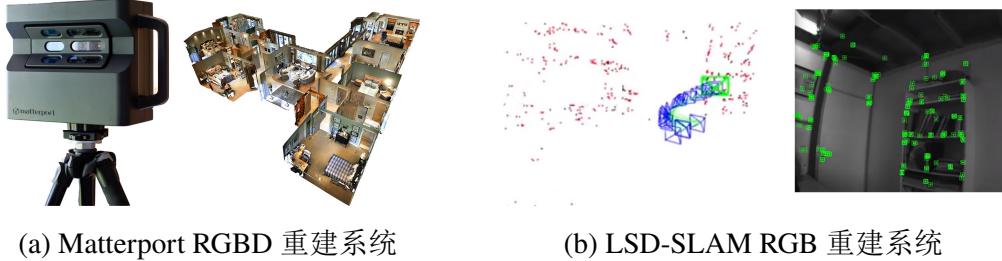


图 1.1 现有的三维重建算法及系统效果展示

然而，上述两种三维重建的大方向都有其存在的问题：对于能够采集深度数据的专业采集设备而言，虽然能够获得较为精确的几何数据，但他们往往拥有高昂的成本：Matterport 最新的 PRO 3D 售价高达 3000 美元，而微软 Kinect 虽然表面上价格低廉，但近日已经由于长期的销售入不敷出而停产；此外，现有的算法很难修复采集设备引入的噪声或漏洞，并需要特殊处理回环、定位不准确等问题。对于业界常用的多张图像恢复模型的传统算法，可以通过普通的设备获取大量的数据，但这些传统算法往往需要在图像中寻找可以匹配的稀疏特征，很难完全利用单帧的完整信息，这也就意味着这些算法需要大量的输入以及计算量。

随着机器学习和视觉近几年来的蓬勃发展，深度学习算法框架已经在单张图像识别^[7]、分类任务^[8]上表现得很优秀了。这也同时给以我们启发：能否将视觉算法应用到重建任务中，从而摆脱对昂贵采集设备的依赖，同时也取得较优的重建效果呢？这引出了本研究的主题：希望借用现有的机器视觉与深度学习算法，来最大限度地从单张图片中挖掘已有信息，从而对图片中的三维场景、物体进行恢复，实现最终的三维重建任务。

1.2 图形学与三维场景建模

图形学是研究计算机如何表示、处理、渲染甚至生成三维数据的学科，与人类认知世界的过程相通，其研究成果在游戏、电影、建筑、制图等艺术或工程领域都有广泛应用。其中三维数据的采集和建立是图形学领域的热门研究课题：与二维数据不同，三维数据的表示多种多样，针对这些不同的表示也有不同的

算法。

在单个三维物体层面，其表示方法包括点云（Point Cloud）、体素（Voxel Grid）和网格（Mesh）等等。而上升到三维场景的层面，表示则更为多样：场景是物体的组合，既可以使用表示一般三维模型的方式表示场景，也可以使用一些简单的几何体（例如圆柱体、长方体、平面）对场景中的物体进行形状层面的抽象、或是为场景中的每一个单独的语义区域（例如桌子、椅子）分配对应大小和方位的 CAD 模型，这些表示方式的选择取决于具体的应用场景，也同时影响着相应算法的复杂度和重建效果。

本文希望研究的话题是根据单张图片进行三维场景建模，这就为建模过程的本身提出了更高的要求。现有的渲染技术通过模拟相机的成像原理来进行三维场景到二维的转换，世界坐标首先被转换成以相机坐标系为参考坐标系的本地坐标，再经过仅和相机参数有关的投影矩阵将三维齐次坐标降维到二维齐次坐标：这种降维的转换损失了深度的相关信息，为从二维到三维的恢复带来了很大的不确定性和模糊性。如图1.2所示，展示的是著名的“大小恒常性错觉”实验，在不含背景信息的情况下，我们很难仅根据物体本身的二维形状来判断其深度和大小。庆幸的是，有了场景的上下文信息以及一些先验知识，原本问题的解空间被进行了大幅度的缩减。在获取到具体的场景后，人类可以根据自己的经验来大致判断洞穴的深度，以及三位探险家身材的大小关系。

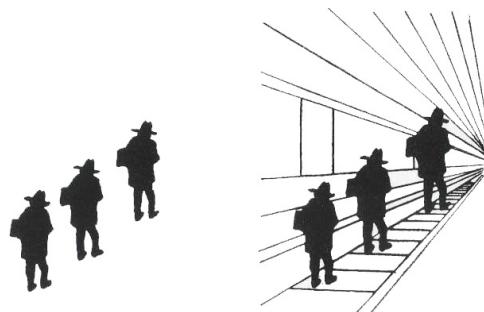


图 1.2 经典的大小恒常性错觉实验，在加入了上下文信息之后，人们更容易判断三维物体的相对位置和大小

利用上下文信息是目前根据二维信息恢复三维模型的重要突破点，而目前这方面的研究并未成熟，学界仍然在逐步推进的过程之中。

1.3 机器视觉与图形学的交叉

机器视觉领域的研究主要集中在图像处理和识别的领域，希望通过分析提取图像的特征获得高层的语义信息，实现计算机对图片真正的“理解”。而随着近年来人工智能领域的不断发展，视觉专家们将这些学习算法应用到图像处理领域取得了很大的成功^[7-13]，在图像分类、物体识别、人脸识别、风格化等方向均取得了不小的突破。一些研究者通过神经网络算法成功从含有复杂背景的二维图像中恢复出了大致的三维表示。这也启发了许多图形学研究者考虑是否能将机器学习算法应用于几何处理、三维重建等领域。这便促成了视觉、图形学以及人工智能三个领域的交叉。

具体来说，一些工作^[14]尝试使用三维渲染的方式，辅助以大型的模型数据库 ShapeNet^[15]，输出大量合成图片进行机器学习算法的训练，免去了收集、标注数据集的人力劳动；一些课题组通过提取二维脸部特征对三维人脸进行恢复^[16]，实现了有趣的人脸更换、虚拟会议应用；还有另外一些研究将概率生成模型应用于建模的过程^[17]，大大加速了美术师模型构建的效率。这些技术的飞速发展速度给了我们很大的信心，二者的交叉算法在应用上仍有很大的潜力可以挖掘。三维室内场景的恢复在两个领域都有一些尝试性的工作，如果将二者结合，利用视觉技术来辅助三维重建、从而实现智能的重建，将是一个值得研究的议题。

1.4 课题意义与价值

本课题主要研究如何根据单张室内场景的彩色照片，恢复出照片所拍到的部分的三维模型，其中房间的几何形状采用正方体包围盒进行拟合，家具则从大型的模型数据库中进行检索。

从研究意义上来说，从单张图片恢复出场景三维模型的工作研究很少，而本工作将三维重建与视觉相结合，利用深度学习和传统方法的结合来解决三维场景问题，是一次创新型的尝试。

同时该课题也有广泛的实用价值：正如1.1节所说，相比之下，带有能获取深度信息传感器的设备还远远不及普通彩色相机普及，对于普通用户来说，希望能利用手机或数码相机这种最简单的设备来进行三维场景的恢复。且对于室内场景来说，无论从房地产销售、虚拟现实还是游戏制作上都有广泛的需求。

本文的主要贡献是提出了一种根据单张室内场景彩色照片进行房间三维 CAD 模型重建的方法，这种方法最终的输出为以平面表示的房间内能观察到

的部分的几何形状，以及其中所有落地家具的 CAD 模型以及旋转角度。根据恢复出的场景，我们可以进行简单的编辑，并添加纹理，最终实现新视角生成（Novel View Synthesis）、场景漫游、图片编辑等应用。

本论文各章的内容安排如下：第 1 章为研究背景简介以及课题难点与意义，第 2 章将对相关的文献进行综合论述，第 3 章主要介绍房间几何形状估计算法，第 4 章则介绍家具检索和摆放优化算法，第 5 章说明实验的具体实现以及实验结果，第 6 章将总结全文，对方法的优劣进行分析并提出未来的研究方向。

第 2 章 文献综述

本章将首先从整体上对三维场景重建技术进行简单的综述。由于本文提出的方法由房间几何布局估计和三维物体检测与识别两部分组成，还将对这部分的相关文献和方法进行整理。

2.1 三维场景重建

SLAM（Simultaneous Localization and Mapping）技术致力于通过有限的数据采集装置以及对自身运动的有限感知，定位自身位置的同时还能对整个场景建立地图。基于视频的场景重建就是该技术的一个应用。针对不同的输入，人们发明了不同的 SLAM 算法来进行场景重建，直到目前依然是比较热门的研究领域。

2.1.1 基于深度图像的场景重建

基于深度图的场景重建所利用的设备为深度摄像机，该相机获取到的场景信息含有 4 个频道，包括 RGB 信息以及对应每个像素上的深度，该深度度量了从相机到物体表面的距离。因此该设备每时每刻采集的都是三维信息。目前的研究主要集中在如何将视频中多帧的三维信息进行有效融合以消除噪声，并在融合的过程中试图预测场景语义信息。研究深度图像的先驱 Curless 和 Levoy 提出了基于三维像素块的融合算法^[18]，该算法采用增量式更新手段，能对扫描噪声进行处理；而后随着 Kinect 步入市场，该方向研究逐渐步入高潮：KinectFusion 技术^[1] 采用 GPU 流水线，用移动平均的方法减小噪声，取得了实时重建和交互的效果；Voxel Hashing 技术^[19] 则提出了一种高效的数据结构用来存取和读入重建数据、使得大规模重建变得可能；BundleFusion 技术^[2] 利用全局信息来准确估计相机位姿，在大规模重建下也能取得实时且准确的效果；此外，该方向还有基于点云的实时重建算法^[20] 以及 Elastic Fusion^[21] 等有影响力的工作。

在学界，也有许多研究利用单帧深度图像进行语义和几何的联合估计：Chen 等人提出了一种语义建模方法^[22]（Semantic Modeling），通过已有模型库中的模型来适配深度图像中的特征、并通过学习数据库中的上下文信息来对恢复物体的语义信息进行约束，最终对场景进行高效和鲁棒的重建。相似地，Guo 等人^[23]

通过寻找图片的语义分割和数据库中分割信息的匹配，辅助以空间约束和房间整体的几何约束，使得重建效果更自然、更有意义。随着大型场景数据库的逐步完善，来自普林斯顿大学的 Song 等人提出 Im2Pano3D^[24] 等一系列工作，利用现有的场景数据库，通过机器学习的方法补全未观测部分的场景，从而拼合成完整的全景图片。

2.1.2 基于彩色图像的场景重建

失去了深度信息的彩色图像的三维重建问题变得更加复杂有挑战性。基于彩色图像的方法通常获取一段视频，并从视频的每帧中寻找特征匹配来恢复三维结构。ORB-SLAM 系统^[25] 综合了前人的研究成果，提出了一种实时的、室内室外皆适用的恢复算法，其改进版的 ORB-SLAM2 系统^[6] 也被广泛应用于机器人应用中，提供了更加准确的轨迹估计。由 Engel 等人提出的 LSD-SLAM 系统^[5] 在 CPU 上也能进行实时重建，该方法利用了稠密的图像特征，使得建立大规模重建变得可能。

相比于上述基于视频的算法，一个更有挑战性的议题是通过单张彩色图片进行三维恢复，这往往需要对于图片信息更深层的挖掘和提取，以及大量先验知识的加入。Zheng 等人提出了 Interactive Images 方案^[26]，该方案需要人工给定图片场景物体的包围盒，系统根据物体关系自动求取最合适的三维参数，对于一些规则物体（例如沙发、桌子）等，辅助以用户进一步的操作，还能恢复更为精确的模型。为了减少人力成本，得到更加自动化的解决方案，Liu 等人提出了利用模型库自动对家具进行匹配的算法^[27]，在匹配中他们使用星图的方式来整合从图像中提取除了局部特征子在模型库中进行搜索，该方法能够对常见的家具组合进行很好的近似拟合。和本文工作比较相似的为 IM2CAD 工作^[28]，该算法首先对物品进行检测，并使用穷举法匹配相似模型，最后采用“渲染-匹配”的方式进行室内物品的安放和位置优化，展示了一些效果较好的例子；但是该工作提出的解决方案并不完整，由于并未利用到大型数据集，其整体恢复效果并不理想。本文提出的算法是一种创新性的改进，采用鲁棒的学习算法，较为完整地对该问题进行了处理。

另外值得一提的是，目前和三维场景有关的数据集也处于蓬勃发展的阶段，这些数据集的产生从很大程度上帮助了和三维场景重建有关的研究。其中，比较成熟的数据集包括 Matterport3D^[29]，ScanNet^[30]，SUNCG^[31] 等等。

2.2 房间布局估计

由于本工作主要的应用场景为室内，因此选择性地利用一些室内独有的特征不仅能够使最终恢复出的室内几何形状更加合理，还能对重建算法进行约束，更容易求得较优的解。在这些室内独有特征假设中，曼哈顿世界^[32] 假设就是一个应用广泛的假设：由于人造场景具有比较规则的结构特征，因此可以假设所有的墙面、地面和天花板和世界坐标系中的 xy 平面， yz 平面和 zx 平面都仅包含平行和正交的关系。通过这个几何特征，在进行三维重建的时候，可以进行消失点估计，并通过求解优化方程来同时得到几何坐标和相机内参，具体重建方法请参考3.3节。

传统的几何估计方法通过提取图像边缘或区块的特征进行墙面边界提取以及消失点估计。Lee 等人首先提取图像边缘并将其划分成若干线段，并求取场景的三个正交方向消失点，生成方向图^[33]（Orientation Map）；在墙角处，使用几何推断的方法根据墙角周围的线段预测角点类型、判断墙面的先后顺序，以此评价产生的结果候选的合理性。为了解决上述方法只能预测较空旷场景的问题，Hedau 研究组将原来的几何估计问题简化为三维包围盒适配问题，通过迭代式地轮换屋内家具像素位置估计和包围盒重适配这两步，逐步选出最合适的包围盒参数以及像素级别的几何分割，并对空余空间进行估计，最终得到场景的几何上下文分割^[34]（Geometric Context）。

为了进一步利用房屋几何和屋内物体的关系，Lee 等人提出了一种基于物体所占体素的推理方法^[35]，他们同时计算了房屋几何的候选集以及房间内物体所占包围盒的候选集，并联合查找满足实际情况（例如物体的包围盒不能两两相交、物体包围盒必须在房屋包围盒内部）的候选对进行评价确定最终的房间布局。相近地，为了最大限度利用屋内物体信息，Zhao 等人通过对家具的功能、形状和外观进行联合建模，从而从图像中提取提取层次关系，填补未被检测出的物体及其功能，相比以往模型显著提升了准确率^[36]。此外，还有其他研究者^[37-38]着重研究了如何挖掘曼哈顿世界假设、通过概率模型来减少标注难度的议题。

2015 年，来自 UIUC 的 Mallya 等人^[39] 首先尝试使用全卷积神经网络（FCN）模型来进行房间几何预测，是使用深度学习框架处理该问题的首次尝试。与随机森林（Random Forest）相比，深度学习框架的效果明显更优，证明了神经网络在处理和分析此类问题的优越性。后来 Izadinia 学者也对此进行了说明：由于卷积神经网络能够挖掘整张图像的上下文信息，所以在训练时无需输入对物

体的标注，神经网络即可自动通过海量数据学习物体和房屋几何的关系，并通过物体位置来隐式地增强分割效果^[28]。虽然 Mallya 等人使用 FCN 来预测原图的信息边缘（Informative Edge，即墙面交汇处），在最后输入结果的时候依然需要根据消失点产生一系列的候选布局，使得整个系统训练并非端对端，也需要花费比较长的时间进行推断。为此 Dasgupta 设计了 DeLay 系统^[40]，同样通过 FCN 模型直接预测一个 $W \times H \times L$ 的分类图，该分类图的每一个频道 I_i 代表一种墙面的分布，在预测后只需要跨频道选择置信度最高的墙面分类进行输出即可，这种方法相比与信息边缘法准确度和速度上都有了较大的提升。其余的相关工作还包括近年 Magic Leap 公司的 Lee 等人提出的基于关键点的端对端预测框架 RoomNet^[41] 以及 Zou 等人提出的 LayoutNet 全景图预测模型^[42]。

本文提出的房间几何预测算法主要受到了 RoomNet^[41] 的启发，预测关键点信息实际上是对网络的输出增加了几何约束，相比于 [39] 和 [40] 的方法无需进行后续的优化，在预测速度上期望能有一些提升。

2.3 物体检测与姿态计算

物体的检测问题（Detection）是视觉领域的经典问题，现有文献大多通过机器学习的方法来提取图像的特征，并根据特征进行分类，选取特征最明显的区域作为检测输出。

神经网络的应用大大提升了原有检测算法的准确性，其代表性工作为 Girshick 在 2013 年提出的 R-CNN 算法^[43]，通过使用卷积神经网络对候选框进行特征识别，相比原来的方法提高了近 30% 的准确率。该工作的改进版 Faster R-CNN^[44] 则进一步尝试使用网络作为包围盒的提取器，大大提升了识别速度和准确性。最近由 He 等人发明的 Mask R-CNN^[7] 给以预测网络多任务学习使其能预测每个检测到的物体的轮廓，同时通过改进池化方式使得预测更加精确，其预测效果达到了目前最好的水平。

上述 R-CNN 系列工作常被称作二步检测算法（2-Shot Detection），这是由于其池化步骤建立于候选框的基础上，需要对每一个候选框单独进行识别操作，其优点是精度高，缺点是检测速度慢。为了解决这个问题，学界提出了一些单步检测算法（Single-Shot Detection）：其中，YOLO^[45] 和 YOLOv2^[46] 实现了端对端的单步学习，能够实现大型图像的实时跟踪；SSD^[47] 通过金字塔式连接，复用浅层特征，准确率相比以往工作有所提升。

物体检测并非本文的主要贡献，因此在检测室内家具的时候，我们使用学术界已有的检测器^[48]进行检测。

物体姿态识别的主要目的是从彩色图像中恢复出三维模型在相机空间内的旋转和平移角度，即总共 6 个自由度（Degree of Freedom）的参数回归问题。对于已知物体大小以及 n 个在三维空间中的坐标，加上这些三维点在二维平面上的投影，求取相机位姿的问题被定义为 PnP（Perspective-n-Point）问题，在图形学领域已经有较为经典的解决方案^[49]，许多工作利用该方法进行反向物体位姿求解。例如，Pavlakos 等人^[50]通过图片特征首先定位物体的关键点（例如飞机机翼的尖端、汽车的后轮等），通过求解 PnP 问题获得三维位姿，但该方法对于关键点共面的病态情况无法求解；为了改进这一方法，Grabner 等人^[51]通过神经网络直接预测物体包围盒的 8 个顶点，取消共面约束，大大提升了物体姿态估计的效果。

除此之外，姿态识别还包括渲染匹配法^[52]：Mottaghi 等人通过渲染各个方向的 CAD 模型，并与原图的 HOG（Histogram of Oriented Gradients）特征进行匹配，枚举所有的情况并选择最接近的位姿进行输出。同样使用渲染法但更加巧妙的是 Su 等人的工作^[14]，与枚举不同，他们通过渲染的手段为 CNN 提供数据进行训练，用神经网络的记忆模块来提取姿态估计所需的特征，提高了准确率和识别效率。

本文的算法主要采用先检测，后处理的方法，由于室内场景的限制约束，我们假定所有的家具都是落地家具，仅需要估算其在相机空间内的旋转角度即可。这种简化能够在取得较好效果的前提下大大提高估算的准确性。

2.4 模型检索与生成

本节主要论述与文章算法相关的场景：即根据二维彩色图像进行模型检索或生成。在1.2一节中我们提到，对于单个三维模型而言，其表示方法的不同也预示着算法的多种多言。通过模型部分拼接或是大型数据库检索的方法一般被归类到三维检索，而直接生成网格、体素或点云模型则被归类为三维生成。

现阶段，已经有许多为机器学习算法所准备的大型模型库以及数据集：以 ShapeNet^[15] 为例，目前已经收录了来自 55 个不同种类的近 5.2 万个模型，这使得基于机器学习的检索和生成算法成为可能。检索方面，一般通过提取形状或是图片的特征并选择匹配特征最多的一组解作为输出。这其中比较基础的工作

是 Chen 等人的贡献^[53]，他们通过提取形状的球谐（Spherical Harmonics）特征、MPEG7 等特征来测试形状检索的效果，并提出了实用的光场（LightField）描述符为后来的研究奠定基础。近年来，Li 等人^[54]利用了卷积神经网络的净化特性学习从图像到形状隐空间上的映射，首次能做到以较高的相似度快速检索模型库；与该文章采取了相同思路的还有 Shape2Vec 工作^[55]，该算法将草图、彩色图片、深度图片、形状和问题都映射到统一的 Word2Vec 空间中，与传统的分类器不同，由于 Word2Vec 空间包含了丰富的语义信息，这使得映射器能充分掌握到不同种类输入的相同之处，获得更好的检索效果。但是与 [54] 相比，该方法由于抽象层次更高，很难发现同一个类别的模型之间的细微差别。

Wu 等人提出的 3D-GAN^[56] 模型使用对抗生成网络来学习了物体体素表示的概率隐式空间，能够对三维体素物体取得较好的编码效果，通过另一个专门的映射器来进行图片编码学习之后，能够从二维图片生成体素重建；次年 Wu 等人又提出了 MarrNet 模型^[57]，通过分步学习的方式，借用深度图不含无用纹理的特征性质，巧妙地解决了训练数据的迁移问题，并在三维生成领域取得了很高的评价指标。无符号距离场（UDF, Unsigned Distance Field）是一种比体素更加注重细节的表示方式，在固定网格大小的空间内体素仅含 0-1 二值，而 UDF 则在每格内以实数值表示距离表面的最小距离。Han 等人利用该性质得到了相比体素细节更加丰富的重建效果^[58]。点云相对于体素表示具有更高的精度，可以处理旋转、缩放等问题，Fan 等人^[59]提出了可以根据图片生成点云模型的网络，为神经网络生成无序集合做出了第一次尝试。而最近由 Groueix 等人提出的 AtlasNet 模型^[60]则希望更全面地解决模型生成问题，他们预测每一个模型小块的纹理映射，基于三维表面的小块来拼合模型，与体素和点云法相比重建效果更加真实可信。

三维生成领域正处在蓬勃发展的阶段，但是由于缺乏有效的模型表示方法，其重建效果和真实的 CAD 模型还有较大的差距，还远远不能用于真实应用场景中。考虑到现有模型库的庞大，大部分的家具都能够使用数据库中的模型进行表示，对于一些特殊模型的处理也相对鲁棒，因此，本文将主要使用基于检索的三维模型适配方法来表示三维模型。具体来说，我们将首先使用已有的物体检测框架进行目标发现，接着对于每一个发现的家具实体，我们使用检索的方法找到最相近的模型，并提出一种多任务学习框架来同时估计旋转角度，再利用几何方法进行摆放，最后使用大规模室内数据集先验对摆放的位置进行优化。

第3章 房间几何估计

本章将介绍房间几何估计步骤，该步骤的输入是室内图片，输出是各个室内关键点在相机空间内的三维坐标，关于关键点的定义请参考3.1节。

3.1 问题分析

在2.2一节中已经提到，目前的研究成果证明了使用深度学习算法来提取图片高层语义信息的有效性，能够有效提高房屋几何预测的准确性和鲁棒性。因此本文将继续延续该思路并提出新的几何估计框架。

在前人的思路中，直接使用 FCN 来预测图像边缘或是墙壁分类，但是由于神经网络本身的记忆拟合特性，预测结果很难精确用直线段将不同的区块进行分割，这就要求后期必须要重新进行处理之后预测结果才有意义。本文的方法受启发于 [41] 和 [61]，采用直接端对端使用分割网络预测关键点热力图的方法进行几何估计。假设所有输入的照片内容都由正常相机拍摄，则这些照片拍摄出的室内布局总共可以分为如图3.1所示的几类情况，图中灰色和粉色的标签为关键点，被定义为墙角或墙面交线与图片边缘的交点。

这种基于关键点的输出定义方式通过硬性约束使得每个区域之间使用直线段进行分割，避免了复杂的后处理过程。然而，不同的照片拥有不同的关键点个数，为此我们当然可以使用 RNN^[62] 模型来进行可变关键点个数的预测，但是考虑到问题的特殊性，关键点个数在同一个房间种类 (Type) 下是不变的，因此我们可以直接通过网络的不同分支对不同房屋种类进行预测，各司其职，达到相同的效果。

需要解释的是，对于其他情况房屋布局种类（例如包含两面墙即以下的情况），由于缺乏必要的几何约束，为病态问题，无法进行三维重建，这里不进行讨论。

3.2 关键点预测网络

为了训练使得网络提取关键点信息，本文参考 [41] 的方法定义热力图作为标准输出来训练网络。假设输入图像大小为 $W \times H$ ，房间布局类型为 k ，其第 i 个

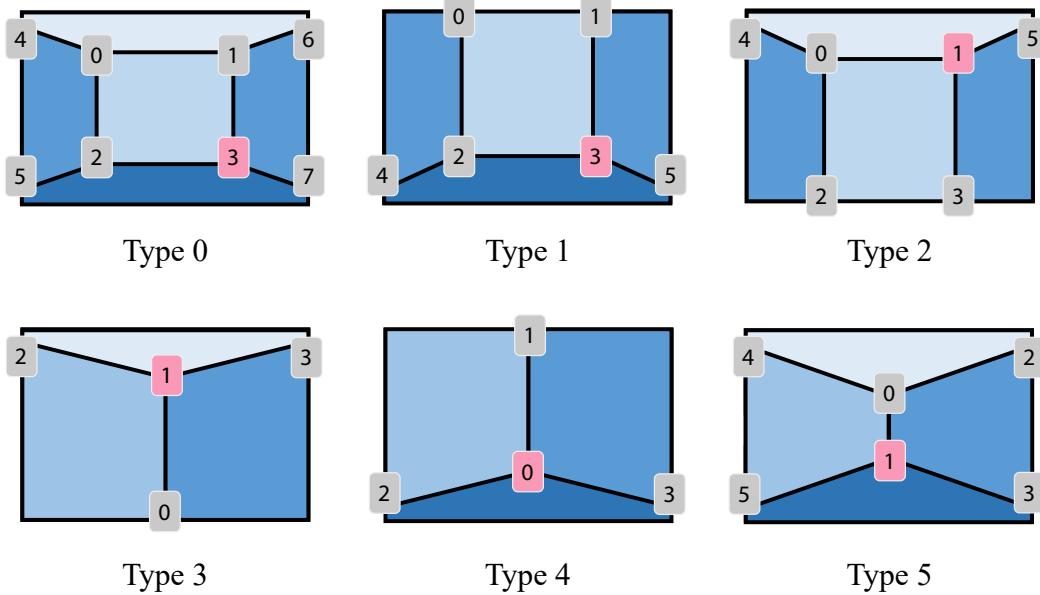


图 3.1 本文处理的共 6 种房间几何类型。不同颜色的区块代表方向不同的墙面，带有数字的灰色或粉色标签代表了关键点的编号，粉色的关键点为定义的坐标基准原点，作为3.3.2节的 $P_{I,\text{down}}$ 来使用。

关键点在原图中的位置为 (m, n) ，对于该特定的关键点，其热力图 \mathcal{M}_i 大小为固定的 $T \times T$ ，热力图上的值是中心为 $\mu = [T \frac{m}{W}, T \frac{n}{H}]^T$ ，协方差矩阵为 $\Sigma = \sigma \mathbf{I}$ 的二维高斯分布。而对于类型 k 的布局的所有关键点的热力图几何的拼接 $\{\mathcal{M}_1, \mathcal{M}_2, \dots\}$ 即为该布局的热力图标准输出。

本文采用图像分割（Image Segmentation）网络来预测热力图。目前在深度学习领域进行图像分割的工作多种多样，例如 FCN^[13]，SegNet^[12]，PSPNet^[11]，RefineNet^[10] 等等，在这些工作中我们选择 PSPNet^[11] 作为算法骨干网络：该网络首先提取图像特征，然后将图像进行不同尺度的金字塔缩放，小尺度特征图负责处理全局信息，而大尺度特征图则负责处理细节信息，最终再将这些信息融合输出，这符合布局估计算法的需要；我们将小尺度的特征图用于房屋类型 k 的分类上，将大尺度特征图用于关键点识别上。

我们的房屋布局预测网络整体架构如图3.2所示。除了使用热力图 $\{\mathcal{M}_i\}$ 来约束输出优化网络，本文还参考了^[9] 中继监督的思路加入了分割监督来帮助网络收敛，请参见3.2.1节的讨论。关于网络训练的损失函数，请参见3.2.2节的描述。

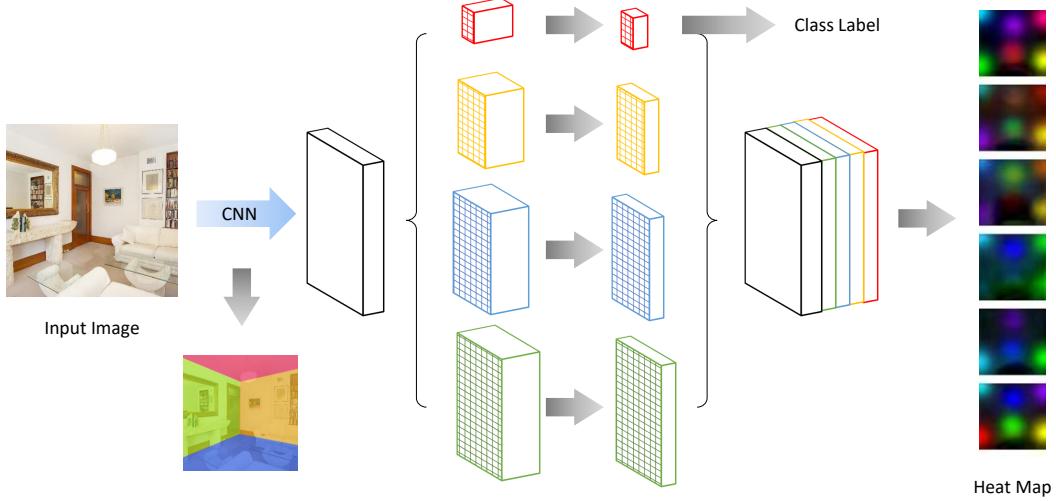


图 3.2 房屋布局预测网络整体架构。单张图片首先经过卷积神经网络的特征提取，而后被分为不同尺度的池化金字塔^[11]，我们在特征提取网络上进行分割监督，为关键点预测提供指导，并根据综合语义信息确定房间布局类型，以此选择最优的热力图进行输出。

3.2.1 分割中继监督

我们引入了分割监督模块对网络的训练加以帮助。当图片输入特征提取深层网络的时候，容易出现梯度消失现象，具体来说算法为特征提取网络中间加入输出层，并通过场景分割结果进行监督。之所以选择在特征提取层进行监督，一方面考虑到后期的网络输出大小不足（我们采用 PSPNet 的 {1, 2, 3, 6} 配置）监督意义不大，另一方面也考虑到前期的监督能够使得网络尽早学习图像语义特征，为后期运算提供帮助。为了消除歧义性（指某些墙面既可以被定义为中间墙，也可以被定义为右侧墙），我们使用统一的分割设计重新定义分割标签。最终该监督项被加入损失函数中一同训练网络，分割监督定义为：

$$\mathcal{L}_{\text{seg}} = \frac{1}{W \times H} \sum_t^{W \times H} \sum_k^6 \mathbb{1}_{k,c} \log(\Phi(t)) \quad (3-1)$$

其中 $\Phi(t)$ 为像素 t 的分类结果， $\mathbb{1}_{k,c}$ 是指示函数，当 k 与真实分割标签 c 相同时为 1，否则为 0。

3.2.2 损失函数定义

假设第 k 种房屋布局 ($k = 1, 2, \dots, 6$) 所含有的关键点数为 N_k , 那么网络最终输出的特征图大小是 $T \times T \times \sum_k N_k$, 其中代表第 k 个关键点热力图的频道起始为 $S_k = \sum_{k'=0}^{k-1} N_{k'}$, 终止为 $E_k = \sum_{k'=0}^k N_{k'} - 1$ 。训练时我们以训练数据标签为准, 把不属于该房屋布局的相关频道权重置 0, 属于房屋布局的频道权重置 1, 关键点热力图损失函数形式化地定义为:

$$\mathcal{L}_{kp} = \sum_{q=S_k}^{E_k} \|\phi(q) - \mathcal{M}_{k,q-S_k}\|^2 \quad (3-2)$$

其中 $\phi(q)$ 为第 q 个频道的热力图预测值。最终网络训练的损失函数为分割监督、类别监督(房屋布局分类)和关键点监督的加权和:

$$\mathcal{L} = \mathcal{L}_{kp} + \alpha \mathcal{L}_{seg} + \beta \mathcal{L}_{cls} \quad (3-3)$$

根据该损失函数直接端对端训练, 测试时通过取输出热力图上的最大值, 即可获得关键点位置。

3.3 相机参数估计与场景重构

虽然使用关键点预测成功约束了墙壁边缘必须为直线, 但作为一个合理的预测还应该满足其他约束, 这些约束包括:

1. 正交约束: 由墙角连出的三条线在世界坐标中是正交的, 这等价于三个正交方向的垂心与图像中心重合^[63] (对于非完整存在三个消失点的情况特殊处理即可);
2. 相机配准合理性约束^[64]: 通过上述约束计算出的相机配准值需要符合其值域。即内参矩阵焦距 $f_x > 0, f_y > 0$ 。

由于神经网络预测的不准确性, 我们很难保证上述两条附加约束同时满足, 对于一些失败的例子而言, 我们希望可以使用原图的低阶^①特征进行预测修正(如图3.3所示)。对于第一个约束, 本文采用 [34] 中的消失点法, 结合网络预测进行融合修正, 请参见3.3.1节; 对于第二个约束, 我们在计算相机内参矩阵时, 假设 $f = f_x = f_y$ 对问题进行简化, 并仅使用两个方向正交方程来求解焦距, 请参见3.3.2节。

① 这里的低阶严谨的含义应该是“低层次”(Low-level), 指的是能简单从图像中提取出的特征, 例如边缘或导数。



图 3.3 利用低阶特征对预测结果进行视角修正的例子，左侧图片中黄色、紫色和蓝色的线代表了利用低阶特征检测出的属于不同方向的小线段，右侧图片的绿色包围盒代表了较差的网络输出，而蓝色包围盒代表了修正后的布局预测

3.3.1 综合语义信息与低阶特征

因为其较大的感受野和较深的架构，使用神经网络获取的是图像综合语义信息；消失点位置约束则直接由图像本身定义而来，属于低阶特征，本节将首先简介相机投影模型以及消失点法，接着说明本文使用的融合算法。

在美术中，消失点法是画家常用的绘制技巧，用于绘制具有几何真实感的画面，例如两条平行的铁轨或是道路会在无限远处交汇于二维画布上的一点。在相机成像投影模型中，常用的即为针孔相机模型，如图3.4所示。成像投影时，需要首先找到三维空间中需要成像的点 P_W 和相机焦点 O (Focal Point) 的连线 OP_W ，该连线和成像平面 $z = f$ (Image Plane) 的交点 P_C 即为该点显示在图像上的位置。这种投影方式被称作透视投影 (Perspective Projection)。

现在，假设有一组平行的射线 $\{l_1, l_2, \dots, l_n\}$ ，他们的方向均为 \vec{d} ，起点分别

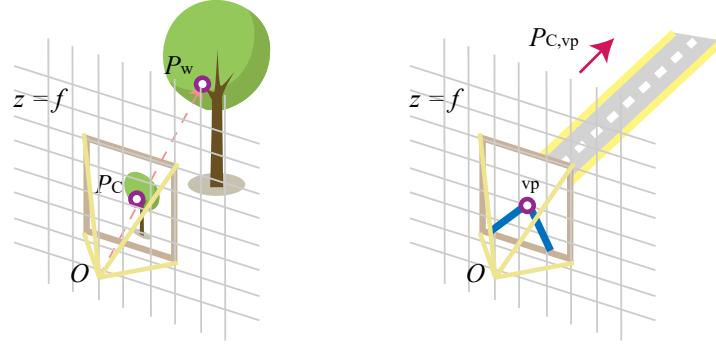


图 3.4 投影相机模型，当所有平行的线投影到 $z = f$ 平面时相交于一点

为 $\{s_1, s_2, \dots, s_n\}$ ，则每一条直线在世界空间中都可以表示为：

$$l_i(t) = s_i + t \vec{d} \quad t \geq 0 \quad (3-4)$$

利用投影方程，可以求出 $t \rightarrow +\infty$ 的时候，直线上的点显示在成像平面上的位置 $P_{C,l}$ ：

$$P_{C,l}x = f \lim_{t \rightarrow +\infty} \frac{s_{i,x} + td_x}{s_{i,z} + td_z} = f \frac{d_x}{d_z}, \quad P_{C,l}y = f \lim_{t \rightarrow +\infty} \frac{s_{i,y} + td_y}{s_{i,z} + td_z} = f \frac{d_y}{d_z} \quad (3-5)$$

可以看出，最终在成像平面上的点和 i 无关，这也就意味着所有平行的直线在投影变换下都将相交于画布上的一点，且我们可以根据该点的坐标直接计算出直线的方向 \vec{d} 。

接着，我们仿照 Hedau 的算法^[34] 对原图像提取小线段，这些线段构成了图像的低阶信息。在原来的算法中，Hedau 等人需要进行复杂度为 $O(N^5)$ 的贪婪算法进行消失点估计，而因为我们有了神经网络输出的大致屋内结构，可以用该信息辅助进行包围盒方向估计，在这里本文采用一种类似 RANSAC 的方法来估计消失点，详见算法1。

在该算法中，置信度 ϵ 为对神经网络预测结果的置信度阈值，其值越小，输出结果就会约倾向于使用网络高阶特征而非低阶小线段特征。对比如图3.5所示。

3.3.2 相机内参矩阵估计

这里统一给出本文使用的相机内参矩阵的定义：将外参矩阵置为单位矩阵，意味着相机位于原点，朝向 z 轴方向，此时内参矩阵 K 将成像平面 ($z = F$) 上

输入: 提取的小线段集合 \mathcal{S} , 方向 $D \in \{x, y\}$, 阈值 ϵ, δ

输出: 方向 D 上的消失点位置 (x_D, y_D)

```

1 根据网络预测的关键点位置计算消失点  $vp_D$ ;
2 过滤  $\mathcal{S}$  中所有距离  $vp_D$  大于阈值  $\epsilon$  的点;
3  $n_{best} \leftarrow 0$ ;
4 while 迭代次数  $< k$  do
5   随机从  $\mathcal{S}'$  中选择一对线段, 计算其交点  $sp$ ;
6   统计所有  $\mathcal{S}'$  中距离  $sp$  小于阈值  $\delta$  点的个数  $n_{cur}$ ;
7   if  $n_{cur} > n_{best}$  then
8      $n_{best} \leftarrow n_{cur}$ ;
9     记录当前的  $sp$ ;
10  end
11 end
12  $(x_D, y_D) \leftarrow$  最后一次被记录的  $sp$ ;
```

算法 1: 高阶预测与低阶特征消失点融合算法

的坐标 P_C 转换为真实图像上的像素坐标 P_I :

$$K = \begin{bmatrix} f & 0 & \frac{1}{2}W \\ 0 & f & \frac{1}{2}H \\ 0 & 0 & 1 \end{bmatrix} \quad (3-6)$$

假设图像平面上的已知消失点为 $P_{I,1}$ 和 $P_{I,2}$, 其对应在成像平面上的坐标为 $P_{C,1}$ 和 $P_{C,2}$, 设函数 $\vec{\mathcal{N}}(\cdot)$ 为向量归一化函数, 则有:

$$\vec{\mathcal{N}}(K^{-1}P_{I,1}) \cdot \vec{\mathcal{N}}(K^{-1}P_{I,2}) = 0 \quad (3-7)$$

因为输入照片均由正常相机拍出, 单像素长宽比为 $1 : 1$, 整体图像宽度 W 和高度 H 均为已知, 故上述方程仅有一个未知数 f , 直接求解即可。最终得到的 $P_{C,i} = K^{-1}P_{I,i}, i \in \{1, 2\}$ 即为两个正交的方向, 第三个垂直方向直接通过叉乘 $P_{C,1} \times P_{C,2}$ 即得。

我们给以神经网络预测的上下墙角点以较高的置信度, 直接将墙角相关的预测点作为最终预测输出, 并以此来决定包围盒的位移, 具体上下角点的三维



图 3.5 修正算法示例：左上图为高阶语义检测结果，粉色为关键点，红色点为左侧墙壁消失点；右上图首先对图像提取段线段，以蓝色表示；下方两幅图的红色线段为过滤出的符合大致方向的线段，而黄色则为算法1确定的线段集合。橙色消失点 $\epsilon = 0.05$ ；绿色消失点 $\epsilon = 0.1$ ，我们统一取 $\delta = 0.1$

坐标 ($P_{W,\text{up}}$ 和 $P_{W,\text{down}}$) 可以由下式计算：

$$\begin{cases} P_{W,\text{down}} = K^{-1}P_{I,\text{down}} \\ P_{W,\text{up}} = P_{W,\text{down}} + hP_{C,3} \\ \begin{bmatrix} P_{C,3} & K^{-1}P_{I,\text{down}} \end{bmatrix} \begin{bmatrix} h \\ -\lambda \end{bmatrix} = -P_{W,\text{down}} \end{cases} \quad (3-8)$$

该式假设相机离地面高度为 1，求解时首先求取第三个式子的最小二乘解，得到 h 带入前两个式子即可得到最终解。

经过了上述推算，就能够得到房屋三个正交面在相机空间内的方向以及上下墙角点的三维坐标，可以完整地输出房屋三维模型了。

第 4 章 家具检索与摆放优化

房间几何布局识别完成之后，算法接下来对室内的家具进行识别和检索、预测其旋转姿态，并使用数据挖掘的方法建立概率模型，对识别出的物体位姿进行优化。

4.1 家具识别、检索与位姿估计

对于家具的识别，本文采用目前效果较好的 Mask R-CNN^[7] 模型，该模型通过更精细的池化操作以及更多的物体信息输入取得了较好的训练和测试效果。其最终输出为每一个家具实例的包围盒 BB 以及置信度 p_{obj} ，其他的输出暂时不使用。目前，在 COCO 数据集上端对端训练的 Mask R-CNN 能够识别包括桌子、椅子、书柜等近 40 个室内物品种类。

对于所有置信度 p_{obj} 高于阈值 η 的检测，算法使用其对应的二维包围盒 BB 截取的区域来进行相似 CAD 模型的检索。我们参考 [55] 和 [54] 的方法来计算二维图片和三维形状的联合隐空间 \mathcal{J} 。

我们希望能尽可能找到和图片中相似的模型，因此采用语义信息来构造 \mathcal{J} 并不能满足需求。此外，由于现阶段能够直接提取传统图片特征的方法都对图片的拍摄角度有要求，很难从单张图片中直接提取出和视角无关的特征。因此算法将从三维模型的角度出发：设模型为 \mathcal{S}_i ，使用传统的确定性映射 \mathcal{F} 将其映射到 \mathcal{J} 中；令集合 $\{\mathcal{I}_{i,j}\}$ 包含了形状 \mathcal{S}_i 在真实世界中的若干张照片（索引为 j ），最终问题转换为学习一个映射 \mathcal{G} ，使得：

$$\forall i, \forall j, \quad \mathcal{G}(\mathcal{I}_{i,j}) = \mathcal{F}(\mathcal{S}_i) \subset \mathcal{J} \quad (4-1)$$

除了找寻相似模型之外，还希望对模型在相机空间中的转角进行预测，在 2.3一小节中已经指出，仅进行一个自由度（1-DoF）的预测不仅能够简化算法结构、提高精度，还能与已有的房间几何估计结果相结合，获得更合理的预测。在这里记图片 $\mathcal{I}_{i,j}$ 的转角为 $\theta_{i,j}$ ，除此之外还需要寻找另外一个映射 \mathcal{H} ，与 \mathcal{G} 形成互补功能：

$$\mathcal{H}(\mathcal{I}_{i,j}) = \theta_{i,j} \in \mathbb{R} \quad (4-2)$$

下面主要说明如何构建上述定义的 \mathcal{F} 、 \mathcal{G} （参见4.1.1节）以及 \mathcal{H} （参见4.1.2小节），并简述如何根据上述信息恢复模型在房间中的三维位置（参见4.1.3小节）。

4.1.1 图像和形状共存隐空间

对于形状隐空间映射来说，其质量决定了相似模型检索的最终效果，我们希望在 \mathcal{J} 中距离相近的数据点，其对应的形状也能尽可能接近；而对于有显著差别的模型而言，希望在 \mathcal{J} 中他们的距离也能足够大。实际上，到目前为止关于“形状相似性”的定义学界并没有统一的标准，对于形状的感知也因人而异，这就为具体的空间映射定义带来了很大的不确定性，本文中我们采用下述两种方式构造隐空间 \mathcal{J} ：

1. 基于多视角图片的特征：这种构造方法基于从不同的角度观察 \mathcal{S} 所形成的二维图片集合的特征相似度，我们直接采用 [54] 中的方法，首先进行光场（Light Field）图片的渲染，接着提取 HOG 特征，并计算相似矩阵，取其中的每一行作为对应形状的向量表示。
2. 基于三维 UDF 的编码特征：这种构造方法直接取 \mathcal{S} 的无符号距离场将其表示在固定的网格大小中 ($32 \times 32 \times 32$)，我们采用和 [56] 相近的网络结构改编为自编码解码（AutoEncoder）网络，进行隐式变量的学习。实践中训练网络直到重建损失最小，提取中间层特征作为形状向量。

之所以选择 UDF 作为模型编码器的输入，主要基于以下考虑：

- 基于网格的表示可以轻松利用现有的编码解码器 CNN 框架进行特征提取；
- 和二值体素相比，距离场存储了更多关于模型的细节信息，这能使得编码效果更好更精细；
- 和 SDF 有符号距离场相比，由于模型库中的模型较差的几何性质，许多模型并非封闭，这会使得基于扫描线的符号随机场提取方法失效；

从实现上来分析，基于三维 UDF 的编码特征在训练编码器时，使用了大量的三维模型，这就使得编码器对于其他相关的模型能够有较为鲁棒的表现，和基于多视角的图片特征相比能更简单地实现新加入形状到隐空间的映射；但由于网络训练的时空复杂度随着网格的边长立方级增长，也就导致很难使用更精确的网格表示（例如 $128 \times 128 \times 128$ ）来表示 \mathcal{S} ，这也可能会导致最终隐空间的编码效果降低。两种编码方式的示例如图4.1所示。本文将在5.2.1节中给出两种空间编码的可视化对比。

上述的两种方法定义了形状空间映射 \mathcal{F} ，对于图片而言，使用卷积神经网

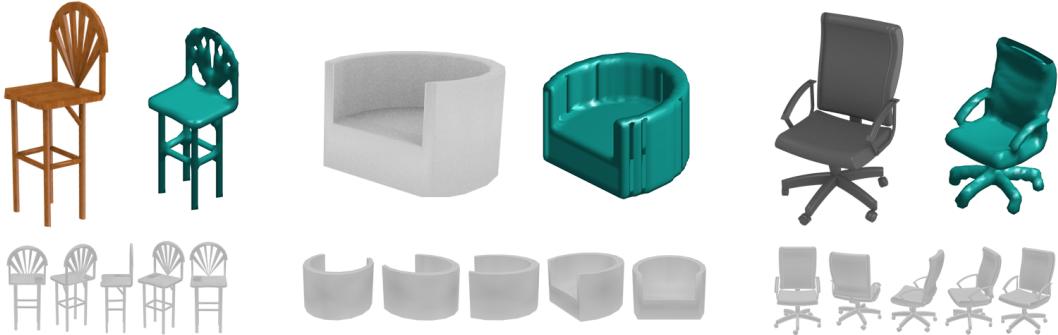


图 4.1 三维形状编码示例，每组左上角为原始网格表示的模型，右上角为 UDF 场恢复出的等值面，而下方一系列则为光场图片的渲染结果。使用 UDF 相比于体素而言能更精确表示三维模型，但相比于光场图片精度稍差一些

络 AlexNet 作为函数 G 的拟合工具。该网络原本用于分类问题，能够较好处理 ImageNet 上的数据，学习出的特征能够适用于各种其他任务。因此，我们希望借助 AlexNet 强大的图片分析功能，学习出和图片中模型视角无关的形状信息。本文将网络最后的分类层去除，更换成隐空间向量值的回归层，最终实现由图片本身到 \mathcal{J} 的端对端训练。

一般而言，为了训练网络必须要有较大的数据量，这里本文仿照 [14] 的方法使用渲染引擎直接渲染 ShapeNet 中的形状，并叠加自然背景来模仿真实照片；为了解决合成图片到真实图片的迁移问题，我们固定了 AlexNet 的浅层特征提取模块，使得已经经过大量数据训练的 AlexNet 能提取出不含有合成信息的特征来。另外，为了确定渲染时的相机角度，我们提取了 ObjectNet3D 数据集^[65] 上已经标注的相机角度信息，并计算每个角度的概率作为渲染时的角度采样参考。

4.1.2 多任务姿态预测

本文采用多任务学习的方法，为上述 AlexNet 模型增加分支，该分支主要输出对物体旋转角度的预测。学习物体形状和物体姿态这两个任务实际上是互补的，对于形状预测而言，希望相同物体的不同姿态都能使得网络输出相同的结果；对于姿态预测而言，则恰恰相反。这对网络来说容易造成较大的歧义性导致学习效果不佳，因此我们将新增的分支添加到卷积层之后，使得网络在使用卷积层提取了必要的特征之后，通过不同的分支将这些特征分解为形状和姿态两部分。新增加的分支和原有网络的深层全连接层结构相同，我们将需要预测的 θ 值平均分为 16 个分箱，其中第 i 个分箱代表了物体在相机空间内的旋转角度为 $\theta \in [\frac{\pi}{8}(i - 0.5), \frac{\pi}{8}(i + 0.5))$ ，这种分箱一方面可以处理角度的连续问题，即我们无

法使用回归问题来很好的定义 0 和 2π 的固有连续性，另一方面使用分类问题取得的求解精度也比回归问题要高^[66]。

最终，我们采用一个统一的架构结合多任务学习方法同时对映射 \mathcal{G} 和 \mathcal{H} 进行了拟合，网络训练的损失函数为：

$$\mathcal{L} = \|\mathcal{G}(\mathcal{I}) - \mathcal{F}(\mathcal{S})\|^2 + \omega \sum_i^{16} \mathbb{1}_{i,i_{\text{real}}} \log \mathcal{H}_i \quad (4-3)$$

其中 ω 是两个损失的平衡系数， $\mathbb{1}_{i,i_{\text{real}}}$ 为指示角度分箱 i 是否为真实分箱 i_{real} 的 0-1 函数。

我们的统一架构如图4.2所示。

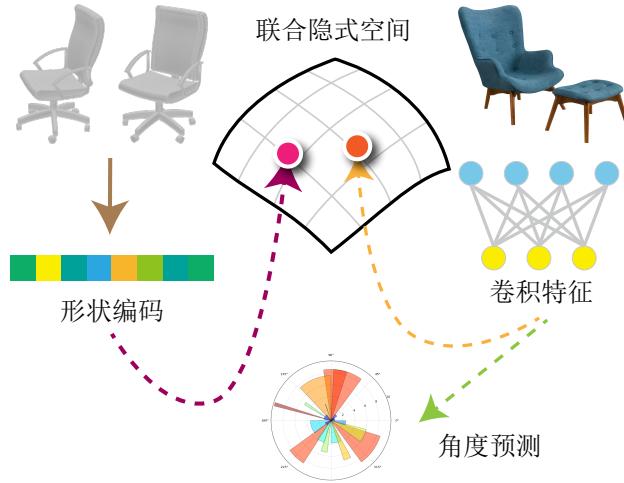


图 4.2 多任务学习架构。我们首先使用形状编码构成联合隐式空间，接着构造神经网络 AlexNet，并学习从卷积特征到隐空间相应形状的映射以及到旋转角度的预测

4.1.3 三维坐标恢复

在3.3中给出的恢复相机和场景三维坐标的方法皆在相机空间内定义，相机位于原点，但整个场景与相机坐标轴并不对齐。这种不对齐会极大复杂化之后算法用到的符号，因此我们将坐标系从相机坐标转换为世界坐标，并假设场景与世界坐标的坐标轴对齐，相机相对于世界零点的变换（逆外参）为 $[\mathbf{R}|\mathbf{T}]$ ，其中 \mathbf{R} 是旋转矩阵，可以由相机坐标系中场景的三个主方向单位向量拼合得到， \mathbf{T} 是位移向量可以由相机坐标系中的场景原点求得：

$$\mathbf{R} = \left[\vec{\mathcal{N}}(P_{C,1}) \quad \vec{\mathcal{N}}(P_{C,2}) \quad \vec{\mathcal{N}}(P_{C,3}) \right]^T \quad \mathbf{T} = -P_{W,\text{down}} \quad (4-4)$$

回忆上述符号定义来自3.3.2小节。

接下来，需要将物体以其三维形式适配到二维包围盒 BBox 中，由于已经做出所有物体皆为落地物体的假设，所以二维包围盒与包围盒内的物体在下框线的相交点 $P_{I,ds}$ 所对应的世界空间内的点 $P_{W,ds}$ 一定在地面上。假设地面墙角在世界空间的坐标为 $P_{W,ground}$ （该点一般为原点），那么我们可以使用直线和平面相交公式轻松计算 $P_{W,ds}$ ：

$$P_{W,ds} = \mathbf{T} + \frac{(\mathbf{P}_{W,ground} - \mathbf{T}) \cdot \vec{\mathcal{N}}(P_{C,3})}{(\mathbf{R}\mathbf{K}^{-1}P_{I,ds}) \cdot \vec{\mathcal{N}}(P_{C,3})} (\mathbf{R}\mathbf{K}^{-1}P_{I,ds}) \quad (4-5)$$

然而，考虑到预测的不准确性以及和模型适配的差异性，我们并没有使用 Mask R-CNN 的物体轮廓。我们将使用上两节预测得到的形状 \mathcal{S} 和角度 θ 信息来确定点 $P_{I,ds}$ ，并最终给出形状的旋转 \mathbf{R}_S 、平移角度 \mathbf{T}_S 和缩放 κ_S ，使之最终适应 BBox 约束。

本文采用一种迭代的方法来求解上述三个参量，详细过程请见算法2。首先将物体在相机空间内旋转预测的角度，然后找出此时物体投影到画面上的渲染图最上方的点和最下方的点，即以当前渲染出的物体的包围盒与物体最下方的交点来对应实际图像中包围盒和物体最下方的交点，根据比例就计算出一个初始的落地点位置。接下来对模型进行缩放，缩放比例以物体的最上方落在图片包围盒的上框上为基准，在算法描述中同样通过最小二乘法完成，该最小二乘的意义为：模型最上方的点在缩放后应该与包围盒对应比例处的点重合，对于这个过定问题采用最小二乘可以解出模型缩放和模型上顶点深度两个值。在完成了第一次适配后，由于初始在确定渲染模型与其包围盒的上下交点时并没有保证其观看角度和图片一致，因此渲染图上下交点占上下框线的比例并不能准确反映图片中的实际情况，这是由透视投影的特殊性决定的。因此，我们重新进行渲染并重新求取交点进行计算，循环迭代固定的次数，保证每一步计算得到的接地点和缩放都比之前更优。

这种方法相比于以往方法，利用到了场景外围的几何信息，能恢复出更合理、更精确的解，但缺点是依赖于模型库本身的数据质量：如果数据库中缺少相关的模型，则适配效果可能只能差强人意。值得一提的是，在我们的算法中，由于系统返回的 \mathcal{S} 实际为一个列表，用户可以手动进行模型选择与适配。

我们给出了一个适配转椅的例子作为该算法迭代时的演示，如图4.3所示。

输入: 网络预测的角度值 θ_C , 形状二维包围盒 BBox, 参数为底部和顶部角点的坐标 (x_b, y_b, x_t, y_t)

输出: 形状 S 最终的旋转 \mathbf{R}_S 、平移角度 \mathbf{T}_S 和缩放 κ_S

- 1 令: $\delta\theta = \arctan(\mathbf{R}\vec{e}_3.x, \mathbf{R}\vec{e}_3.z)$, \mathbf{R}_S 取绕 y 轴旋转 $\theta_C - \delta\theta$ 的四元数;
 - 2 初始化: $\mathbf{T}_{S,0} = \mathbf{R}[0, 0, -\epsilon]^T + \mathbf{T}$, $\kappa_{S,0} = 1$;
 - 3 **for** 进行固定次数的迭代, 本次索引为 k **do**
 - 4 找到 $\mathbf{T}_{S,k-1}, \kappa_{S,k-1}$ 状态下 S 在画面中投影的包围盒以及最上方和最下方的顶点 v_t 和 v_b ;
 - 5 根据上述包围盒的比例计算 BBox 上方和下方的二维点 v_{BBox_t} 和 v_{BBox_d} ;
 - 6 利用公式4-5, 将墙角点更换为 v_{BBox_d} 并计算其世界坐标 P_B ;
 - 7 利用如下算式更新 \mathbf{T}_S 和 κ_S , 第一个算式采用最小二乘法求解即可:
- $$\begin{cases} \begin{bmatrix} v_t - v_b & \mathbf{R}K^{-1}v_{BBox_t} \end{bmatrix} \begin{bmatrix} \kappa_{S,k} \\ -\lambda \end{bmatrix} = \mathbf{T} - P_B \\ \mathbf{T}_{S,k} = P_B - \kappa_S v_b \end{cases}$$
- 8 **end**
 - 9 取最后一次的迭代结果作为输出;

算法 2: 求解三维家具参数的迭代式算法

4.2 基于概率的摆放优化

根据上述步骤计算得到三维模型位置精确性取决于前述步骤的准确性。房屋的几何形状、相机的参数或者是物体检测网络在识别上的误差都有可能造成算法的失败。因此, 本文的算法希望能对不准确的物体位置 (X_i, Y_i) 、旋转角度 θ_i 和缩放 κ_i (即4.1.3节的 κ_S , 这里为了方便后续表示更换了下标) 进行修复, 这种修复仅受到 BBox 的弱约束。一般而言, 可以通过“渲染-匹配”的方法进行部件位置优化, 但考虑到我们的可变参量较多, 且我们希望生成一个合理的结果, 这种合理性约束能够使得恢复出的模型更具有视觉美感。

受启发于 [67] 的工作, 我们通过挖掘已有场景数据库中的家具关系来对摆放进行优化: 如果识别出的椅子和桌子夹角为 85° , 则将其纠正为数据库中占大多数的 90° , 很有可能是正确且更加美观的结果。为了简化问题, 减少参数量,

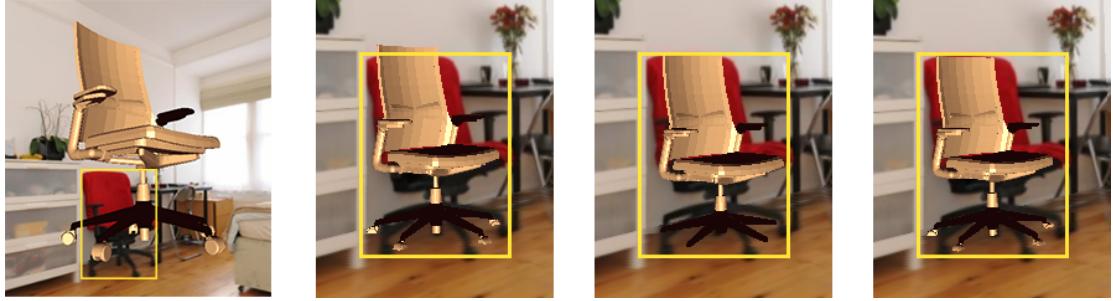


图 4.3 使用算法2进行三维模型位置缩放优化的实例：上述四幅图片分别代表了算法 $k = 1, 2, 3, 4$ 的四步迭代步骤；由于每一次物体都比之前更加接近理想位置，因此其视角差也越来越稳定，估计的最上顶点和最下顶点也会越来越精确，最终更好地适配到 BBox 内部

我们把优化限制在二维空间内。考虑到所有的物体都是接地的，这种简化实际上为优化步骤提供了硬约束。

我们采用的场景数据库为 SUNCG^[31]，该数据库由普林斯顿大学建立，包含了近 4.6 万个场景、近 41 万个房间数据、不同的物体总数达到 2644 个，非常适合我们的分析任务。针对每个房屋，我们主要进行下面两个关系统计：

1. 物体类别间关系 $O_{i,j}(\Delta X, \Delta Y, \Delta \theta)$: 以第 i 种物体的坐标轴为基准，第 j 种物体在该坐标系下的 xy 坐标为 $(\Delta X, \Delta Y)$ ，自转角度为 $\Delta \theta$ ；
2. 物体墙面关系 $\mathcal{W}_i(d, \phi)$: 以第 i 种物体的坐标轴为基准，离它最近的墙面距它的距离为 d ，它相对于这个墙面的转角为 ϕ ；

其中，所有有关距离的度量 ΔX 、 ΔY 以及 d 都应该统一归一化，这里我们除以第 i 种物体的长度。对于这些关系而言，我们首先使用离散的方式对其进行计数，再利用高斯混合模型（Gaussian Mixture Model）对离散值进行拟合以获取连续可导的连续概率密度函数 p_O 和 p_W ：

$$p_{O,i,j}(\Delta X, \Delta Y, \Delta \theta) = \sum_k \rho_k \mathbb{N}(\Delta X, \Delta Y, \Delta \theta | \mu_{O,k}, \Sigma_{O,k}) \rightarrow O_{i,j}(\Delta X, \Delta Y, \Delta \theta) \quad (4-6)$$

$$p_{W,i}(d, \phi) = \sum_k \tau_k \mathbb{N}(d, \phi | \mu_{W,k}, \Sigma_{W,k}) \rightarrow \mathcal{W}_i(d, \phi) \quad (4-7)$$

其中 $\mathbb{N}(\cdot | \mu, \Sigma)$ 代表了均值为 μ 、协方差矩阵为 Σ 的多元高斯分布， ρ 和 τ 为对应高斯分量的权重。

我们使用 `sklearn` 中提供的 EM 算法来进行 GMM 的拟合。假设总共的物品种类数为 C ，则我们提取的 GMM 总共的个数即为 $C(C + 1)$ ，这些混合模型

代表了我们从大规模数据集上挖掘出的分布信息。

此外，在优化时我们还需要限制物体不能过于偏离原来的位置，为了和GMM的概率密度函数定义保持一致，我们同样使用高斯概率模型来对位置偏离约束进行建模。我们希望物体的分布也是一个高斯函数，其“方向”和相机射线方向保持一致。

利用这些信息，即可对输入图片所产生的模型参数集合 $\{X_m\}, \{Y_m\}, \{\theta_m\}$ 进行优化，优化目标以似然函数的形式定义：

$$E(\{X_m\}, \{Y_m\}, \{\theta_m\}) = \prod_m p_{W, c_m}(d_m, \phi_m) p_{\mathcal{L}, m}(X_m, Y_m, \theta_m) \\ \left(\prod_n p_{O, c_m, c_n}^{p_0}(\delta(X_n - X_m), \delta(Y_n - Y_m), \theta_n - \theta_m) \right) \quad (4-8)$$

其中 m 和 n 为房间内所有家具的索引， c_m 和 c_n 代表了家具 m 和家具 n 的种类； p_0 为物体类别关系的初始值，算法鼓励在预测中原本就具有很高概率符合先验约束的物体关系继续跟进，而对于其他的关系来说我们仅仅进行较弱强度的监督（参见本节最开始所举的椅子和桌子夹角的例子），因此采用概率初始值对整体的优化内容加以权值。此外，我们还可以在公式中为所有的概率 p 加入指数项代表其权值，这里由于公式过于复杂没有体现。

公式4-8是一个比较复杂的似然函数，我们采用梯度下降的方法来寻找函数的最大值。最大化似然函数之后继续根据算法2计算缩放 $\{\kappa_m\}$ 即可。

4.3 物件与场景贴图

本部分将简述如何对已有图片进行贴图与智能化编辑。纹理贴图和映射 (Texture Mapping) 原理是为模型的每一个顶点映射一个图片上的位置，这样以来对于模型上的每一个三角形面片，我们可以通过仿射变换求得该三角面片所在平面上的每一个点与纹理贴图二维点的对应矩阵 \mathcal{A} ，该矩阵维数为 2×3 ，支持以齐次坐标形式表示的仿射平移操作，操作过后会保留原有图像的平行关系。这种变换对于拥有许多三角面片的网格模型来说是可行的一种近似。因此在为三维模型映射贴图的时候，对于模型上的每一个顶点 v_i ，我们直接选取其在照片上的投影点 $K\mathbf{R}^{-1}(v_i - \mathbf{T})$ 作为其纹理坐标即可。另外，考虑到模型的边缘可能和图片中的三维物体并非完全匹配，我们将内缩后的模型蒙版应用于图片上，并使用泊松融合技术对纹理进行扩展，使得纹理更加美观。一些贴图效果如图4.4所示。



图 4.4 贴图算法为三维物体计算的贴图效果

但是对于墙面来说，将其简单拆分成两个三角形是不可取的，这是因为我们采用的是投影变换假设，如果依然采用仿射变换，原本应该平行的纹理图案由于其在图片中（图4.5(a)）并不平行，导致恢复出的纹理本身呈扭曲状，如图4.5(b)所示。因此在变换时我们需要使用更高阶的矩阵 \mathcal{P} ，该矩阵维数为 3×3 ，通过引入齐次坐标的方式将仿射变换拓展成投影变换。

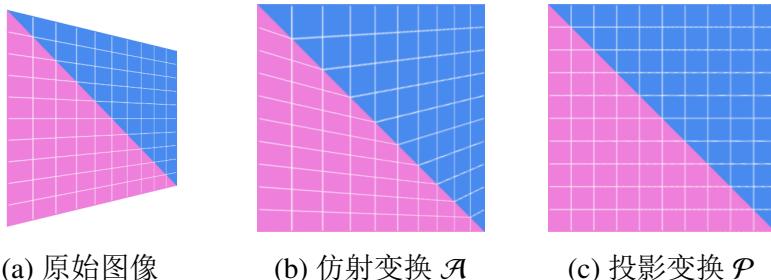


图 4.5 从原始图像映射到纹理贴图时不同变换方式的对比

当用户对恢复出的场景进行室内漫游的时候，需要进行各个不同角度的场景欣赏，需要删除错误投射到地面上的室内家具。算法将找到三维家具投影到画面上的部分，并将这部分作为蒙版进行图像填充，填充使用的方法为 PatchMatch^[68]，该方法能够利用图片蒙版之外的小块信息为蒙版内的缺失内容进行补全，从而使得整个图像连贯且真实。

在完成了家具识别、摆放和纹理叠加之后，我们就能够对于图片中的场景进行完整建模了。在下一章将展示我们的实验结果以及案例演示。

第 5 章 实验验证

5.1 几何布局估计

5.1.1 实现细节

几何估计算法的训练和测试主要采用现有的深度学习框架 Pytorch^[69] 实现，该框架具有较好的灵活性，容易进行错误调试，且理解源码简单直接。本文输入图片统一采用 320×320 的 RGB 通道输入，训练和测试的时候统一减去 ILSVRC 2012 数据集^[8] 上的均值文件，并对图像进行数据增强，增强方式包括左右翻转、比例不超过原图 10% 的放大和缩小以及不超过 5° 的随机旋转。训练在 Nvidia GTX 1080 的 GPU 上进行，优化器为 SGD 优化器，其初始学习率选择为 10^{-6} ，动量值为 0.9，由于后期网络出现了一些过拟合现象，我们选用 80 轮的迭代模型作为最终模型。对于我们提出的修正算法（参见章3.3.1），我们仅将其用于类型 0,1,2,5 中，其余输出遵循网络预测结果。

本文采用的训练和测试用数据集来自 LSUN2015^[61]，其中包含了 4000 张训练图片以及 394 张验证图片。验证时除了使用上述 394 张图片之外，为了取得和 DeLay^[40] 方法的对比数据，我们还加入了来自 Hedau^[34] 数据集中的 314 张验证图片。

5.1.2 对比分析

为了证明我们网络结构的有效性，我们将我们的方法和前人的方法^[34-36,39-41]进行对比，主要的参考指标如下：

- 图片分割像素准确性：直接计算由关键点产生的图片分割结果和真实图片分割结果的像素准确性。由于该匹配具有模糊性，标准采用匈牙利算法（Hungarian Algorithm）计算两张分割图片分割区域的最佳匹配，输出准确性百分比。
- 关键点位置回归误差：该误差为匹配的关键点之间的距离和。由于每张图片大小不同，该距离会除以图片对角线长度作为归一化手段。

除了这些工作之外，我们还设计模型简化测试（Ablation Study），提出分割网络模型，该模型不含有低清的分割监督，仅保留原有 CNN 部分，最终依然输出跨类别的关键点热力图。

表5.1展示了在 Hedau 数据集上的测试效果对比，该数据集虽然对物体进行了标注，但并未给定精确的角点位置，因此我们只提供平均像素差异的结果。表5.2给出了在 LSUN2015 测试集上的效果对比。为了取得公平的对比效果，统一使用 [61] 中的测试代码进行跑分。

表 5.1 房间布局估计网络在 Hedau 数据集^[34] 上的测试结果

方法	平均像素差异
Hedau 法 ^[34]	21.20
Lee 法 ^[35]	16.20
Zhao 法 ^[36]	14.50
Informative Edge ^[39]	12.83
DeLay ^[40]	9.73
RoomNet ^[41]	8.36
本文方法	7.97

表 5.2 房间布局估计网络在 LSUN2015 数据集^[61] 上的测试结果

方法	平均像素差异	角落位置差异
Hedau 法 ^[34]	24.23	15.48
Informative Edge ^[39]	16.71	11.02
DeLay ^[40]	10.63	8.20
RoomNet ^[41]	9.86	6.30
本文方法	9.30	6.07

需要注意的是，在上述标准数据集中，网络允许的预测结果包括总共 11 中房屋布局结果，而由于我们的特殊任务，仅仅允许恢复 6 类布局结果，且取得了比 RoomNet^[41] 法略有提升的结果，这证明了我们的网络以及假设针对多变的房间布局来讲具有一定的鲁棒性。

如图5.1、图5.2可以看出，本文提出的中继监督方法（参见章3.2.1）是有效的。另外由差异矩阵可以看出，即使网络输出了错误的分类结果，但是整体对于房间布局的估计仍然不会太糟。这是因为框架使用了相同的特征提取器，能够提取出房间布局的大致特征，这就使得所有种类的热力图均能学到相关的高阶语义知识，保证热力图的输出始终有意义。

在我们的数据集中仅包含不足 5k 张训练样本，这远远小于 ImageNet^[8] 的数据量，但是我们的方法却能学习到较为鲁棒的高阶语义特征，这是因为所提供的热力图监督和图像分割监督信息量都足够大，使得每一张图片的训练都能切实有效。图5.3展示了分割网络模型以及本文方法在 Hedau 数据集上的迁移效果。

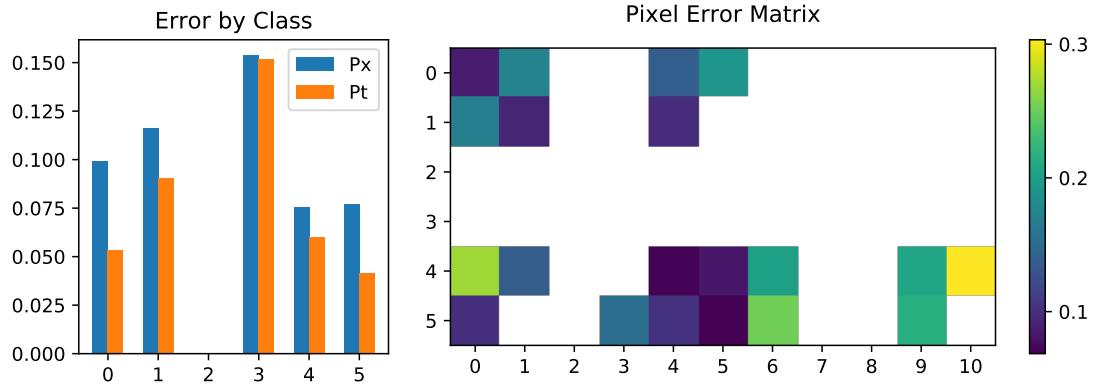


图 5.1 使用本文方法在 LSUN2015 数据集^[61] 上的结果，左图为 6 类房间布局预测的平均像素差异 (Px) 与角落位置差异 (Pt)，右图为实际布局类型被预测为另一类房间类型时的平均像素差异矩阵

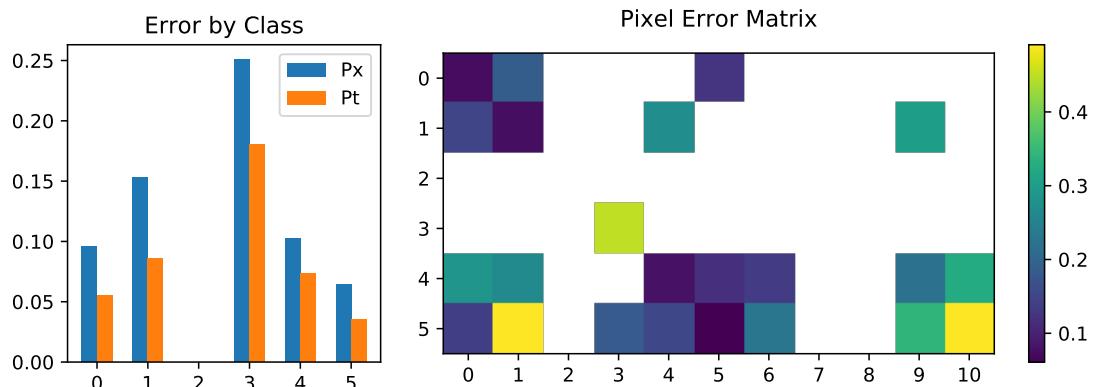


图 5.2 使用分割网络模型在 LSUN2015 数据集^[61] 上的结果，两个图的意义与图5.1相同

此外，图5.4的示例展示了输入了网络可以接受的其他类别布局时，本文的方法依然能输出近似正确的结果。

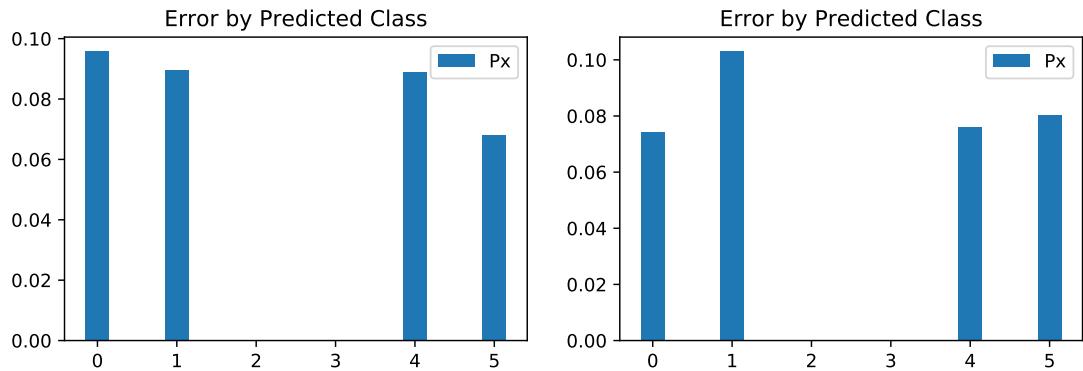


图 5.3 使用分割网络模型（左）和本文方法（右）在 Hedau 数据集^[34] 上的平均像素差异（Px）对比，横轴为预测的房间分类

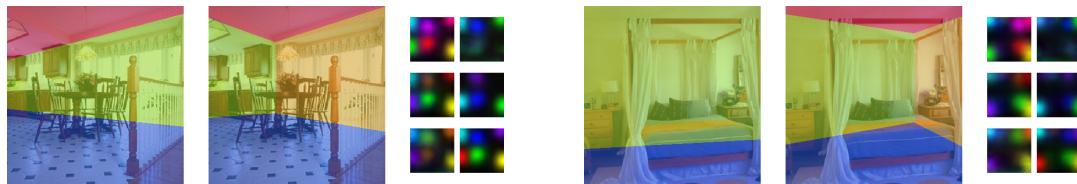


图 5.4 在输入图像布局种类不在网络可预测范围内时，网络的输出结果考虑到了图像本身的歧义性，依然能输出相对合理的结果

5.1.3 重建效果展示

图5.5展示了随机从验证集中选取的样本以及本文所提出方法的结果作为演示。在恢复出三维模型之后，算法对三维模型进行了旋转平移以合成房屋的新视角。由于目前为止还没有对家具进行检测和恢复，因此在旋转视角之后家具会有一些拉伸的不自然效果，稍后会进行解决。

本文使用的三维查看器使用 OpenGL 写成，使用 libqglviewer 库来进行三维内容的查看显示窗口。

5.2 家具检索与优化

本节对应于第4章提出的算法具体实现以及结果展示。第5.2.1节将主要展示隐式空间的构造结果；第5.2.2节展示对比了使用隐空间和物体角度进行联合训练的结果；第5.2.3则主要展示从 SUNCG^[31] 数据集中挖掘出的分布规律以及优化算法的运行结果。

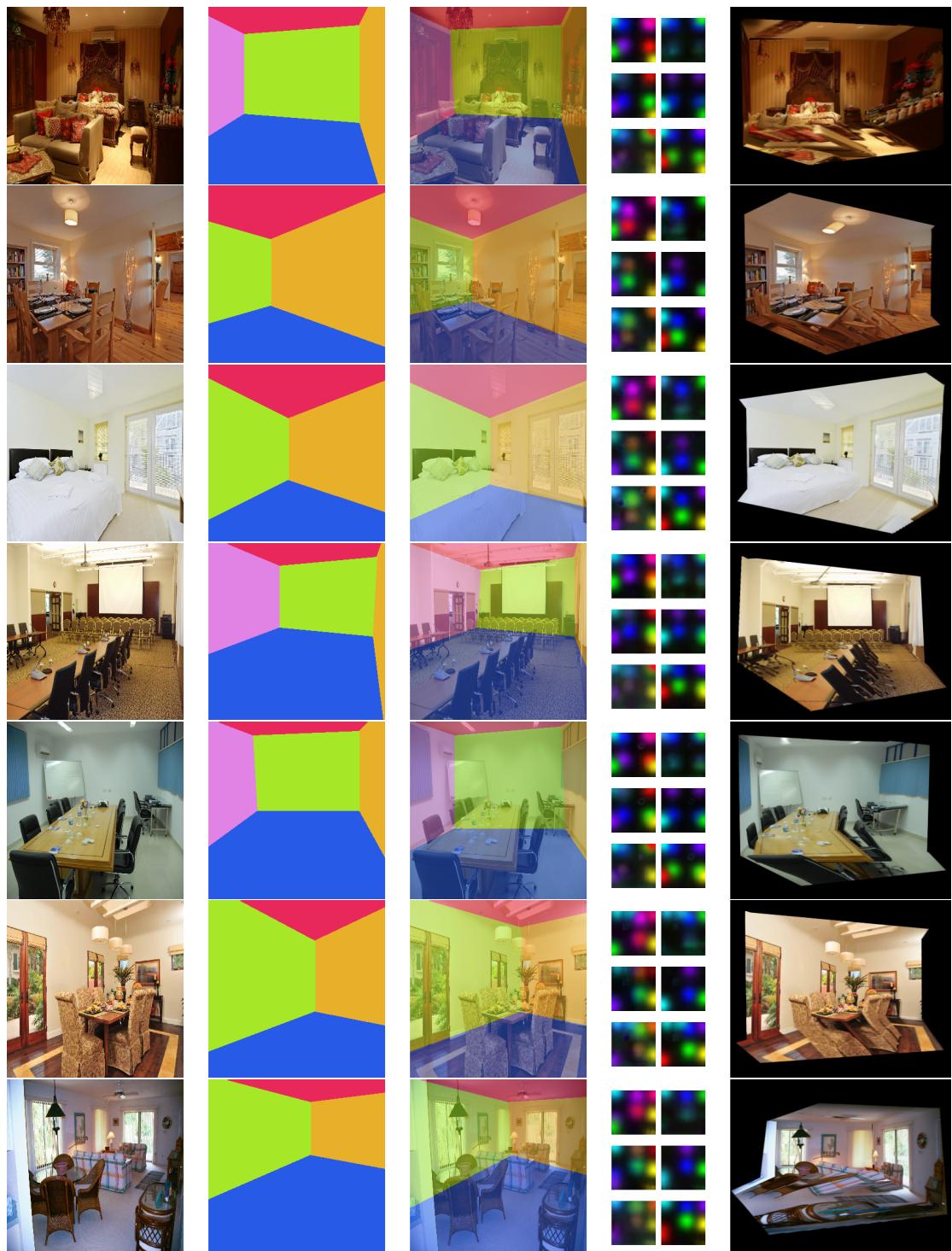


图 5.5 房屋几何恢复结果，列数从左到右分别为：输入图像、基准结果、预测结果、各类别的预测热力图、三维重建效果

5.2.1 隐式空间构造

在第4.1.1节中介绍了两种构造隐空间 \mathcal{J} 的方法。对于第一种方法，我们主要采用 [54] 提供的开源代码进行计算，我们选取的空间向量长度为 128，每个形状渲染 20 张 LFD 图片，我们针对“椅子”类型的空间进行了可视化，如图5.7（红色区域）所示。

对于第二种方法，首先需要提取模型的 UDF。我们使用一种基于宽度优先搜索的方法求取模型的 UDF：对于模型的每个三角面片，我们先计算这些面片周围网格的距离值，接下来从这些已知的网格点出发，计算其相邻格点的距离场值，选取距离最小的距离格点作为确定格点，并以此更新其他非确定格点的值，直到所有网格都含有确定的值为止。每个模型对应的 UDF 场大小为 $32 \times 32 \times 32$ ，在进行可视化的时候，我们使用 MATLAB 中的 `isosurface` 函数进行等值面提取渲染（通常该值可以取为 0.01）。

在为 UDF 编码的时候，我们参考了 3D-GAN^[56] 的编码器和解码器（AutoEncoder）网络，具体来说网络中不含有池化层，避免三维数据过度的池化，仅使用卷积操作进行特征图缩小。中间特征层的输出大小即为隐空间大小。我们直接使用同一类型的模型进行训练，依然采用 Pytorch^[69] 框架，训练 15 个周期直到验证集损失收敛。图5.6展示了使用训练的编码器进行三维重建的效果。解码器对于一个特定的输入向量能保留大致的形状信息，但对于形状细节（例如椅子背部有无孔洞）恢复的并不是特别理想。

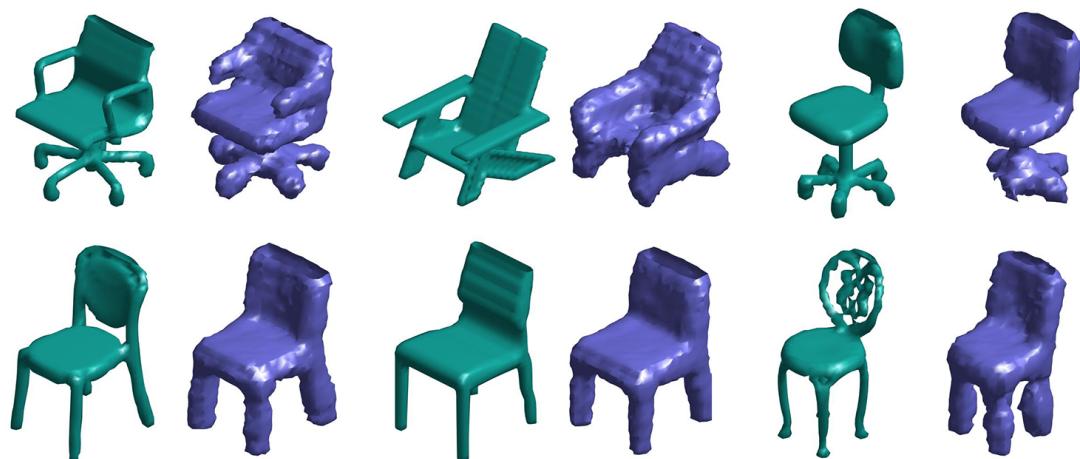


图 5.6 使用自编码解码器进行 UDF 学习与重建的效果。对于输入形状差异较大的图形重建效果可以反映形状差异（第一行）；但对于较相似的形状而言重建差异并不大（第二行）

为了取得和第一种方法的对比，我们同样针对“椅子”一类进行了空间可视化，如图5.7（蓝色区域）所示。我们使用了 `sklearn` 中的 PCA 的方法对于原向量进行降维以显示在画布上。图5.9使用两种编码方式分别训练了模型来测试预测效果。对比两种编码方式，我们可以发现，相比较而言 UDF 对于形状整体把握较好，对于较为厚实的沙发而言，使用 UDF 更容易发现其体积特征；而多视角图片对细节把握较好，能够在排序之后的形状列表中始终给出较为一致稳定的形状预测。

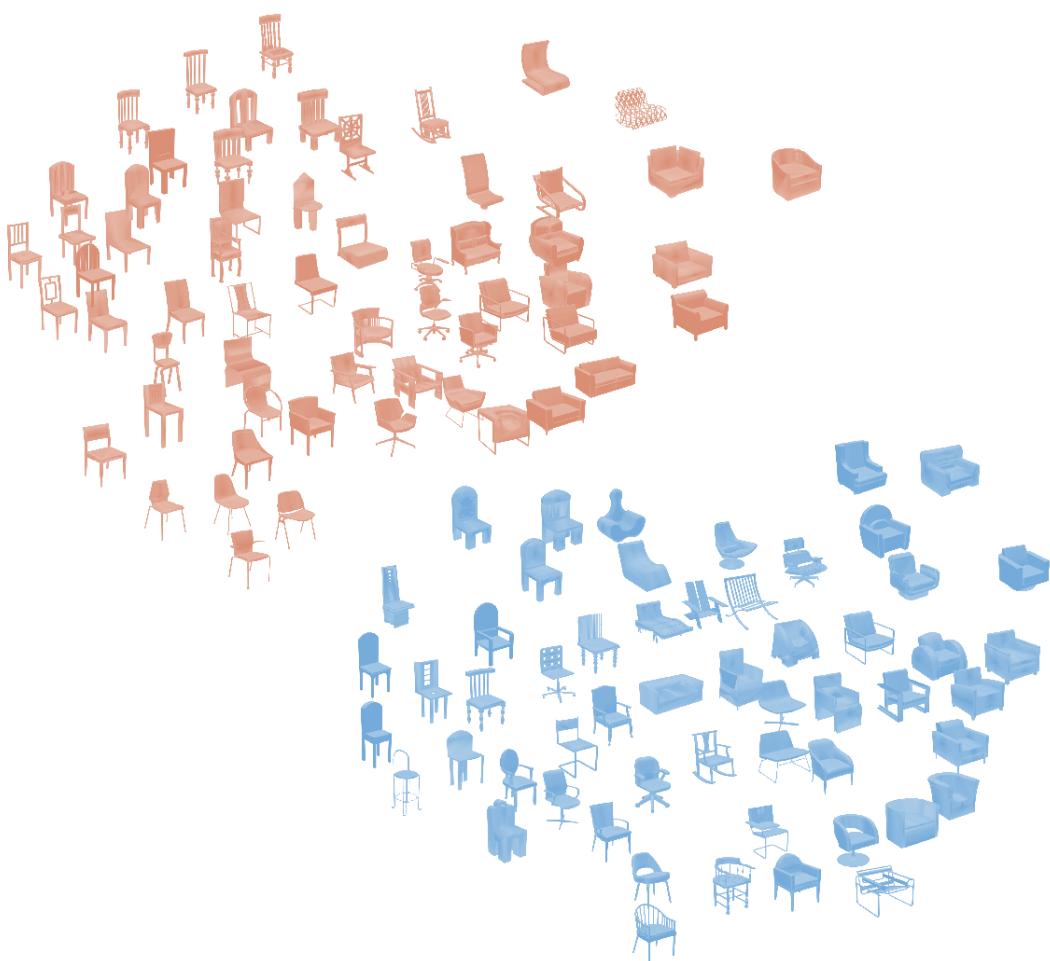


图 5.7 使用多视角图片特征（第一种方法，红色区域）与使用三维 UDF 编码特征（第二种方法，蓝色区域）对比：使用多视角图片特征更注重细节部分，后背带条纹的椅子被单独划分出来；而 UDF 自编码特征对这种细节的分辨能力则稍弱一些

5.2.2 多任务预测

为了生成适用于在第4.1.2所提出框架的数据，我们使用 `blender` 渲染器对所有形状集合进行了渲染。在渲染的时候，需要尽量保证数据的分布和真实世界一致，才能获得在测试集（即我们的输入图片）中的最佳表现。为此本文提取了 ObjectNet3D 数据集^[65] 中相关类别物体的相机拍摄角度，并且仅提取转角 azimuth 值，分箱之后的角度分布如图5.8所示。

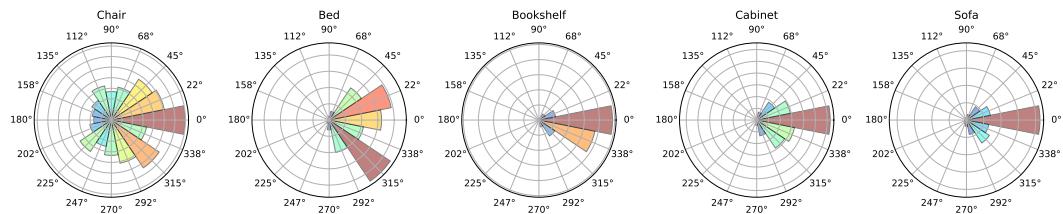


图 5.8 从大规模数据集^[65] 提取的转角分布，是用于生成训练数据的先验分布

完成数据集生成之后，我们使用 `caffe` 框架进行训练，这是因为使用的合成数据量比较大，利用 LMDB 数据库系统进行存储，而 `caffe` 对该数据库系统有着原生支持，能够获得较快的训练速度。本文使用 SGD 优化方式来优化网络，基础学习率为 2×10^{-4} ，动量为 0.9，总共迭代次数为 4 万次。本文总共针对 ShapeNet^[15] 中的 7 个种类进行分别训练，这些种类包括：椅子、沙发、桌子、书柜、床、柜橱和落地灯。目前本框架仅仅能对这些种类的模型进行预测，如果需要增加预测的种类，则需要专门针对新的种类训练网络。本文所使用的训练数据以及训练结果如表5.3所示，对于某些种类来说我们需要消除歧义性。例如对方桌而言，旋转角度为 45° 和 225° 是等价的，因此我们在预测转角的时候将预测空间限制在 180° 以内；而对于落地灯来说，大部分形状都是完全对称图形，所以在实际训练的时候分类效果比较差，这属于正常现象，不影响结果，在最终输出的时候随机选择一个角度进行输出即可。

完成了多任务学习框架的训练之后，就能根据单张物体图片来检索三维模型并预测转角了。图5.9展示了同一张输入图片，使用两个不同的形状编码方式训练的联合模型进行检索时的效果。图5.10则全面展示了不同类别的图片作为输入时网络进行联合形状与转角预测的结果。我们在图上直接以旋转后的三维模型来表示转角以获得更直观的对比。在实际测试中能够发现，对于大部分图片来说，网络输出的结果符合预期，能够在数据库中找到和图片中非常相近的模型，且相似模型的排序也一般出现在列表的前面。

表 5.3 三维模型检索训练及测试数据

种类	WordNet 号	训练图片数	角度准确率	多视角编码误差	UDF 编码误差 ^①
椅子	03001627	677.8k	68.25%	1.30	1.35
沙发	04256520	317.3k	69.83%	1.89	1.25
桌子	04379243	641.8k	61.57%	1.84	1.47
书柜	02871439	46.6k	63.01%	1.09	1.02
床	02818832	25.4k	52.14%	1.20	1.19
柜橱	02933112	157.2k	67.64%	1.42	1.87
落地灯	03636649	231.8k	10.12%	1.94	2.04

① 为了对比本栏的值统一除以 100

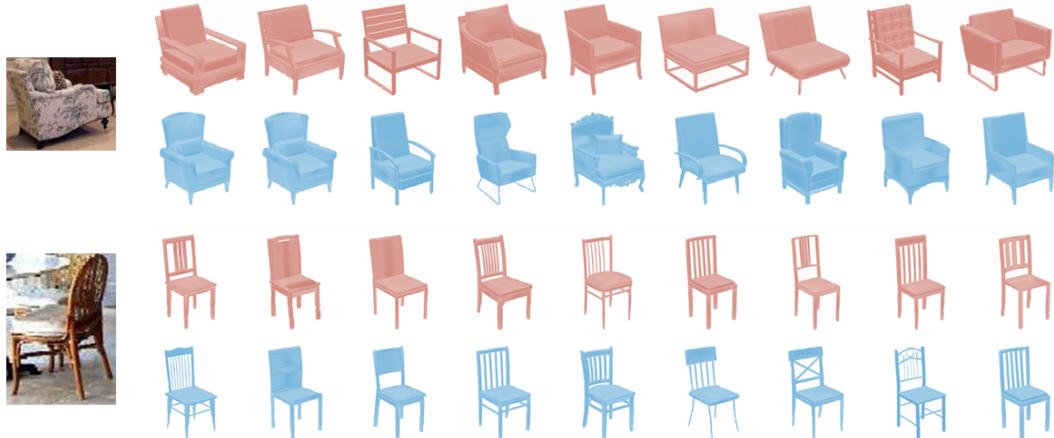


图 5.9 相同输入图片使用不同的形状编码进行检索时的效果：红色模型为多视角图片特征检索结果；蓝色模型为 UDF 编码特征检索，二者均能对相似模型进行检出，且相比较而言 UDF 对于形状整体把握较好，而多视角图片对细节把握较好

5.2.3 摆放优化

首先，展示从 SUNCG^[31] 数据库中提取出的位置先验，我们对 $O_{i,j}$ 和 \mathcal{W}_i 分别进行了可视化。其中，以桌子为中心的椅子分布如图5.11(a) 所示，以床为中心的柜橱分布如图5.11(b) 所示。在图中使用不同的颜色标明了不同散点的转角，同时在二维平面上预拟合一个 GMM 模型并将其概率密度函数叠加使得观察更方便。可以明显看出，对于桌子而言椅子更倾向于分布在它的四个正方向，且摆放的方向均为朝向桌子，这与事实规律相符；对于床而言，除了床头柜的位置会分布在床的两侧之外，还有一部分柜橱分布在床脚处以及其他位置。

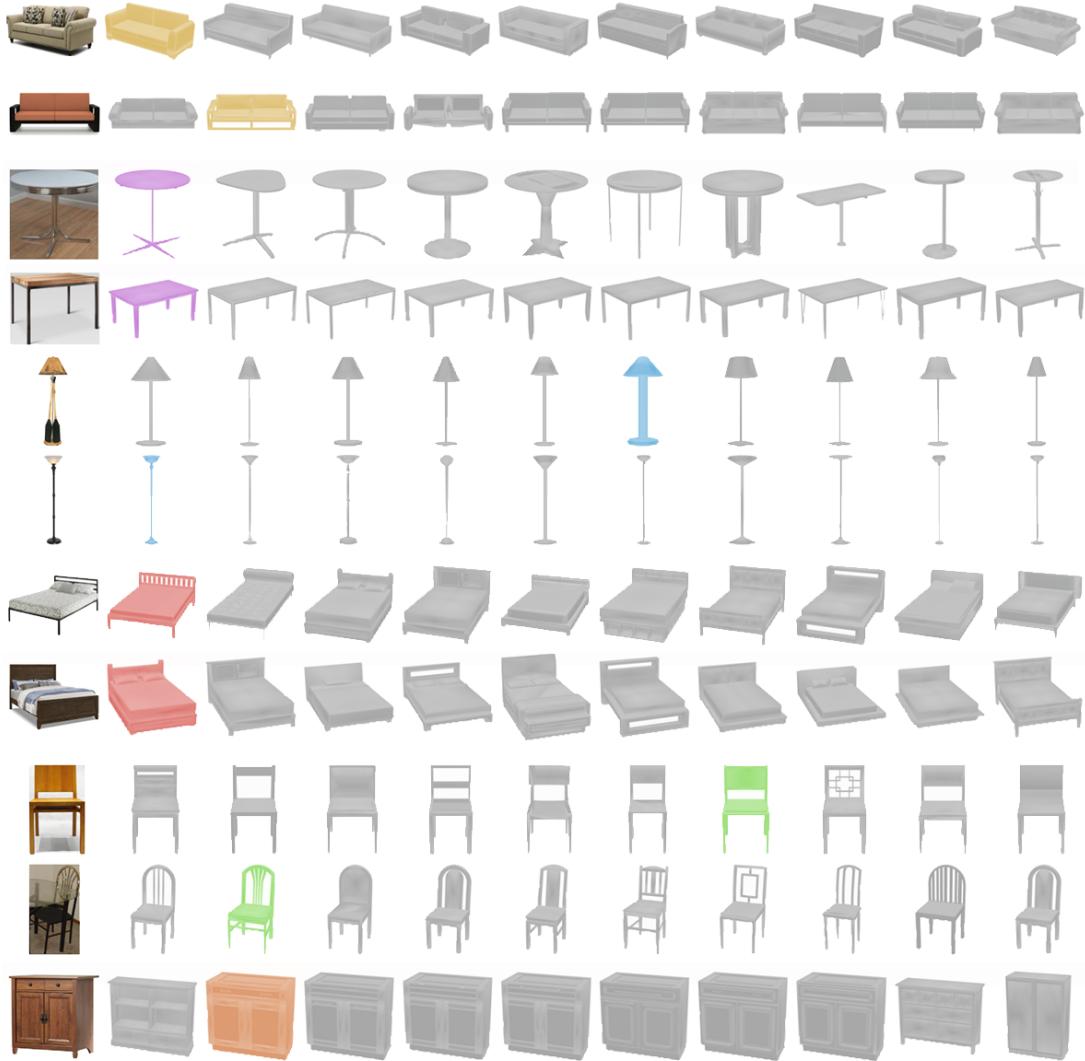
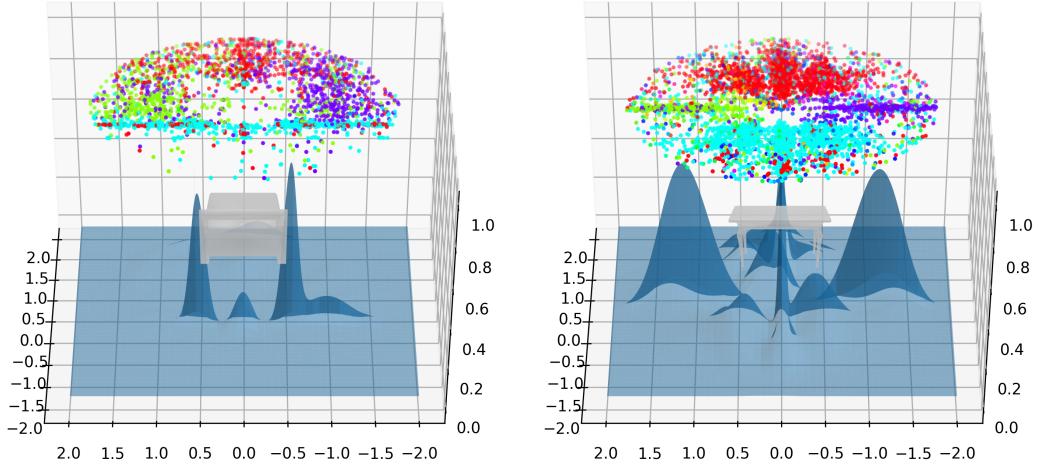


图 5.10 不同类别的图像检索效果，所有的模型均由网络预测的旋转角度进行旋转，每一个检索被高亮的模型是用户手动进行的模型标注，可以发现用户认为正确的模型基本排列在检索结果的前列

图5.12(a)展示了椅子相对其最近墙面的分布，图5.12(b)展示了沙发相对其最近墙面的分布。从图中可以看出这些家具一般是以 $0, \frac{\pi}{2}, \pi, \dots$ 的直角相对墙面的角度分布，且大多数沙发为靠墙放置，这体现了实际房间家具摆放的规则性。通过这些约束，我们可以对原有的预测进行修正，使其更符合设计上的原则和标准。

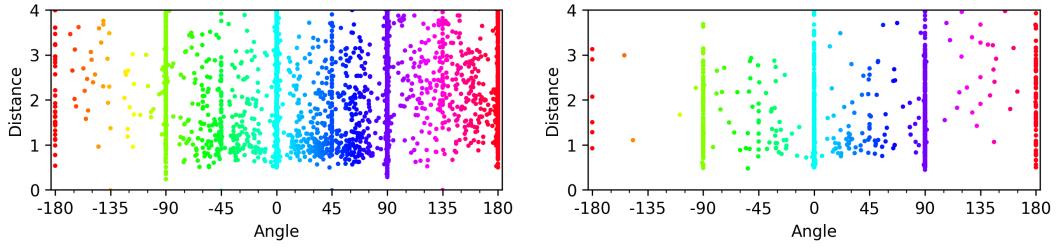
接着，采用4.2中的方法对已有预测进行优化。我们使用 Python 来进行优化框架的实现，并使用 autograd 库来进行较为复杂的求导，在实际优化中我们发现使用 GMM 进行概率拟合时，初始点附近梯度非常小通常难以进行梯度上



(a) 以床为中心的柜橱分配规律

(b) 以桌子为中心的椅子分布规律

图 5.11 $O_{i,j}$ 可视化：每个散点代表了一个数据，点的颜色为 HSV 颜色编码的角度数据，蓝色曲面为空间上拟合的 GMM 概率密度分布函数



(a) 椅子相对其最近墙面的分布

(b) 沙发相对其最近墙面的分布

图 5.12 \mathcal{W}_i 可视化：横轴为相对墙面的旋转角度，纵轴为归一化距离，散点的颜色色环与横轴相同

升，因此设计首先通过各个高斯分量的均值距离进行加权优化，得到合理的初始解之后再使用概率最大似然法进行精细修复。本文截取了优化中的若干个迭代步骤，输出效果如图5.13所示。

5.3 示例展示

在本节中将通过一些例子来展示本文中所提出的方法，这些例子的输入均为单张彩色图像，首先经过房屋布局识别系统进行房间几何识别，由此确定地板和墙面位置、方向以及相机参数，接着对于彩色图像中的每个物体分别进行

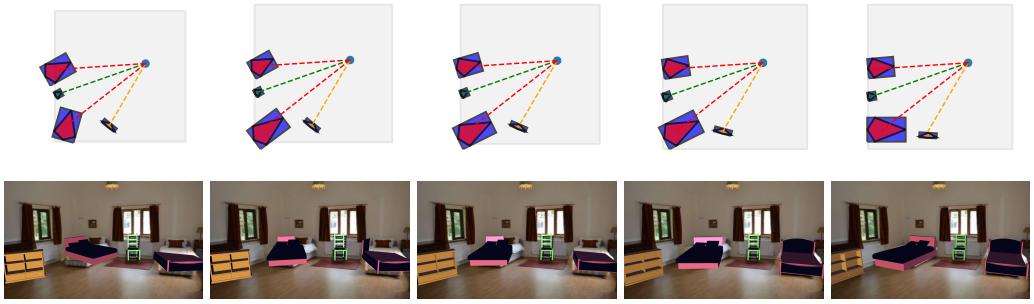


图 5.13 最大化高斯均值加权距离时的迭代可视化，我们对二维位置以及一个自由度的旋转总共三个参量同时进行优化。上面五幅图分别展示了迭代第 0 步、第 5 步、第 10 步、第 20 步以及收敛的结果

检索和摆放，并最终进行了优化步骤。

在所有展示的例子中，我们对于检测不准确的房屋布局加以限制使得每面墙都是方形，且对于 Mask R-CNN 网络漏检的物体进行了补充。为了保证示例的质量，在优化步骤之前允许用户手工对检测的模型进行形状调整，调整范围是隐空间中离图片向量表示最近的 20 个模型。在最终的优化步骤完成之后，本文也提供了一套用户界面使得用户能够对和图片中不相符的物体进行微调，并处理一些算法不能判断的特殊情况。这些特殊情况将在每个例子的描述中阐述，不属于本文解决问题的范畴。

在每个演示中都展示本文提出方法的三种应用：

1. **室内漫游和新视角生成：**对于一张特定角度的输入图像，用户有时希望能从另外一个全新的角度观看图像中展示的场景，甚至希望能够进行场景漫游，即漫步于图片中显示的三维场景中，对场景有更全面的观察。我们的方法使得这种应用成为可能。每个例子的第一行大图及其附属的右侧两个小图展示了这种应用，白色区域为未观测部分；
2. **三维场景理解：**对于机器人应用来说，人们希望其能够通过视觉的手段对自己周围的场景进行理解、并判断各个物体的位置和功能，从而能够在不同的场景进行多样的动作来帮助人类。每个例子中的第二行大图以及其右侧两个小图展示了本方法对于三维场景的理解结果，不同种类的物体以不同颜色标出；
3. **智能图像编辑：**当对场景进行分解之后，可以通过章4.3中介绍的贴图方法为模型进行大致的贴图，这样用户就可以自由编辑贴图后的家具位置甚至是房间布局，并操控相机角度生成新的图片。每个例子最后一行三个小图

给出了图像编辑的一些结果。

整体三维模型查看器使用了基于 C++ 的 Qt 5.10 以及 libqglviewer 写成，处理模型时采用了 OpenMesh 框架，对于图片的变形和贴图我们使用 OpenCV 库中的相应函数实现。我们采用文件读写的方式和 Python 神经网络库和优化库进行对接，绘制模型时采用了多采样的反走样方法使得其视觉效果较好。

示例一 如图5.14所示，展示的是一个中亚风格住宅的三维恢复结果，主要对于床以及两个柜体进行了恢复。本文的摆放优化算法主要将床和后墙面进行了对齐，并由此使得两个柜体也和床变得平行；最终用户手动微调了柜体的位置使得恢复结果与图像对齐。由于床脚柜体主要呈长方体，在视角旋转的时候能够明显看出三维效果。但是对于床头柜来说，检索算法没能在模型库中找到合适大小的匹配模型，导致预测结果得到的床头柜比预期要更长。在图像编辑的时候，用户移除了床头柜，并改变了床脚小柜的摆放位置。

示例二 如图5.15所示，展示了欧美家庭的客厅恢复效果，该场景中主要包括了两个沙发以及一个茶几，由于无法观察到摆放台灯的小桌的接地点，我们没有对其进行恢复。优化算法将两个沙发进行了墙面的对齐，但是用户并不希望棕色沙发和墙面平行，因此手工对棕色沙发的旋转角度进行了修复。另外，对于深绿色的沙发，在模型库中检测到排名靠前的模型沙发靠背高度均比较高，这就导致了最终进行纹理贴图以及多视角观察的时候靠背有一部分实际上取到了墙面的颜色。在图像编辑的时候，用户改变了图像观看视角，并移动、旋转了深绿色沙发、棕色沙发以及茶几，使得他们更加紧凑。从最终的输出图像来看，具有较高的真实性。

示例三 如图5.16所示，展示了儿童房间的恢复效果，场景包含一张床、一个书橱以及一个小柜。对于小柜来说，虽然被正常检测到，但是由于其在画面中的特殊位置我们无法自动对其进行三维摆放，因此对于小柜的位置确定主要由用户完成。此外，对于书橱这种细节较多（层数比较多）的三维物体，算法很难将其每一层和图片都精确对齐，因此在使用 PatchMatch 算法^[68] 去除前景并进行纹理贴图的时候很难精确校准，因此恢复的效果只能是差强人意。在图像编辑的时候，用户对床的角度进行了较大的旋转，拉近了书橱和床的距离，并移除了有视觉干扰的小柜。

示例四 如图5.17所示，为实验室讨论区域的恢复效果，该场景难度比较大，包含了6把椅子和一个桌子。本文使用的检测框架没有检测到最内侧的椅子，因此该椅子由用户进行标出。优化算法倾向于将椅子和桌子对齐，以呈现整齐的排列，而这显然与最靠近画面的椅子方向不一致。因此在优化的时候不对该椅子的位置进行优化，而是直接采用网络预测结果作为输出。由于椅子之间的高度重合和遮挡，为模型检索和贴图带来了很大的挑战。但庆幸的是由于本文提出框架的鲁棒性，对于椅子形状的预测仅出现了微小的偏差，在贴图之后也很难看出模型本身的形式。进行图像编辑的时候，用户旋转了桌子的角度，并移动了各个椅子的位置以获得不同的摆放效果，为了画面的整洁，在图像编辑的前两个示例中用户移除了部分干扰视觉判断的椅子。

上述四个示例展示了本文所提出方法的整体流水线运行效果，能够看出本文的方法对于某些真实世界的彩色图片拥有较好的识别和恢复效果，并能够进行很多高层语义方向的应用。



图 5.14 中亚风格住宅的恢复，包括一个床和两个柜子。通过多个视角的展示能明显看出所恢复的三维效果，在图像编辑时用户移除了床头柜，并改变了床脚小柜的摆放位置



图 5.15 欧美家庭客厅的恢复，包括两个沙发和一个茶几。由于在模型库中缺少这种靠背比较低矮的沙发模型，最终深绿色沙发的贴图有一部分涵盖了后侧的墙壁部分。但整体来说如果视角选取得当，这种不准确性并不影响图像编辑以及场景理解的结果



图 5.16 儿童房间的恢复，包含一张床、一个书橱以及一个小柜，对于小柜的位置确定主要由用户完成。在图像编辑的时候，用户对床的角度进行了较大的旋转，拉近了书橱和床的距离，并移除了有视觉干扰的小柜

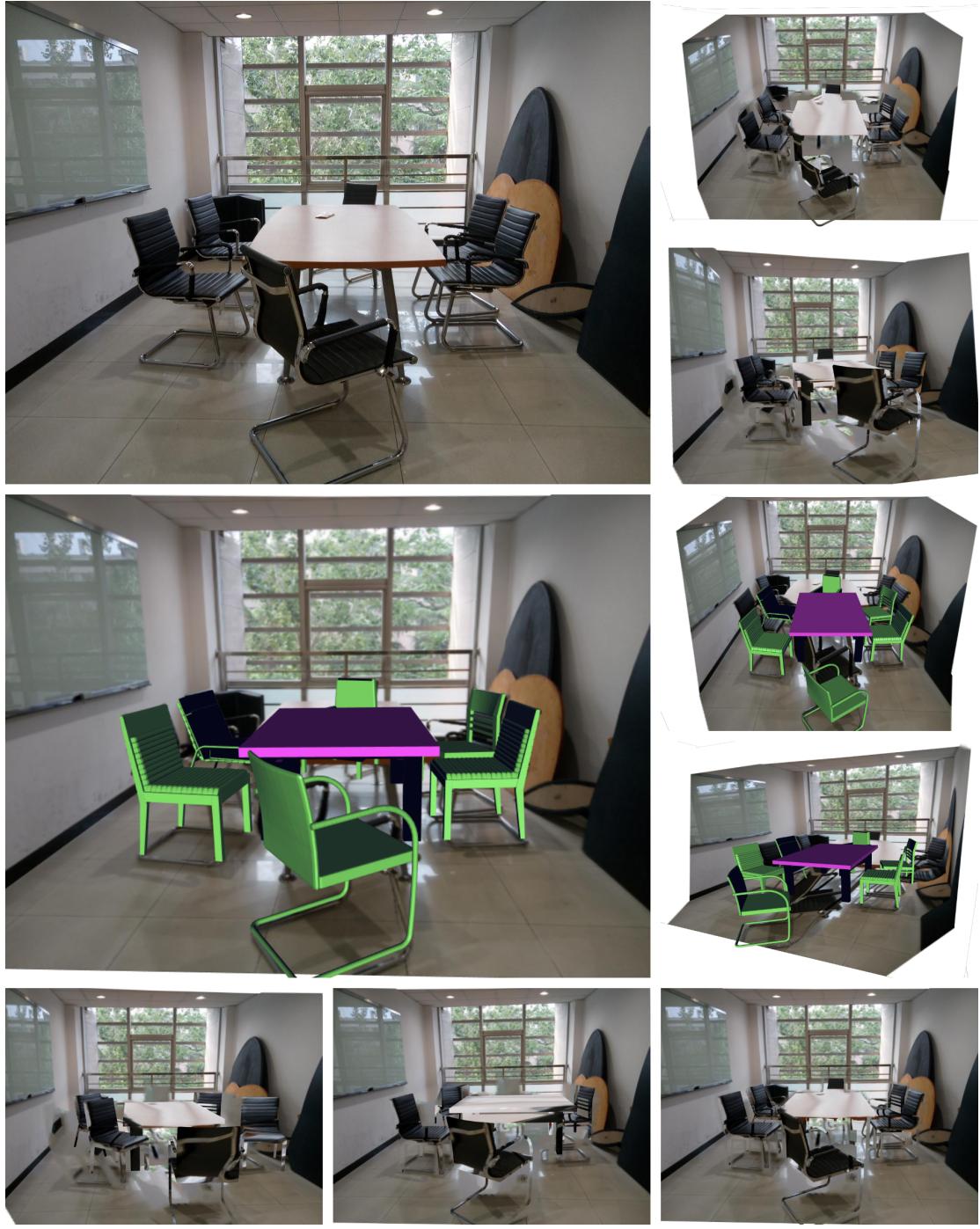


图 5.17 实验室讨论区的恢复，场景摄制自校内 FIT 楼，包含了 6 个椅子和一个桌子，在优化摆放的时候用户手工剔除了离相机最近的椅子从而获得更加稳定的收敛结果；图像编辑的时候用户旋转了桌子的角度，并移动了各个椅子的位置

第6章 结论

本文总结概括了现阶段三维室内场景建模的发展概况，并提出了一种基于机器视觉、从单张彩色图像恢复三维室内场景的方法。不同于常规的需要较多帧图片的建模方法，这种方法不仅考虑到了图像表层的低阶形状特征，还考虑到了深层的图像语义信息。通过二者的有机结合，在最大限度挖掘图像信息的同时，也保持了恢复的精细程度。辅助以用户的手工校准与修改，可以达到几乎和原图一样的场景恢复效果。

总体来说，本文主要进行了如下工作：

- 对图形学与机器视觉交叉领域的三维重建问题进行了详细的介绍与综述；
- 提出了一个估计房间几何布局的机器学习算法框架以及高低阶特征融合方法，准确率相比以往方法有所提升；
- 提出了一种基于图片的三维检索和姿态估计的多任务学习框架，并利用大型数据集先验对识别结果进行优化；
- 整合上述方法，对三维恢复流水线的一些细节问题（例如三维摆放、纹理贴图等）提出了具体的解决方案，使得整个系统得以运行；
- 为该方法的应用场景提供了建议，并通过实例展示所介绍的基于视觉方法的应用演示。

这些贡献对于之后的研究方向和方法都有着一定的参考价值。

当然，本文提出的方法也有其不足之处，这些缺陷包括：

- 对于不满足曼哈顿结构的房屋造型失效；
- 在物体识别和恢复的时候没有考虑到遮挡的情况。对于被严重遮挡的物体来说，其包围盒内的物体仅包含了该物体的一小部分，而在训练形状预测网络框架的时候训练集中的物体均清晰可见，为解决此问题可能还需要设计更复杂的学习算法；
- 家具识别基于单个物体，没有考虑到物体之间的关系。目前的优化步骤仅仅是根据所有数据集中的房屋先验进行的最大似然优化，而并没有更加准确的后验概率，这就使得整个框架缺失了更进一步挖掘上下文的机会；
- 整个方法是流水线作业，前一步方法的不准确会直接影响到后续的步骤。例如当房间几何形状预测出错后，可能直接影响到模型接地点的判断，最

终导致家具安放失败；当我们使用的物体检测模块失效时，后期无法通过优化的方式将缺失的家具添加回来。

上述失效情况如图6.1所示。

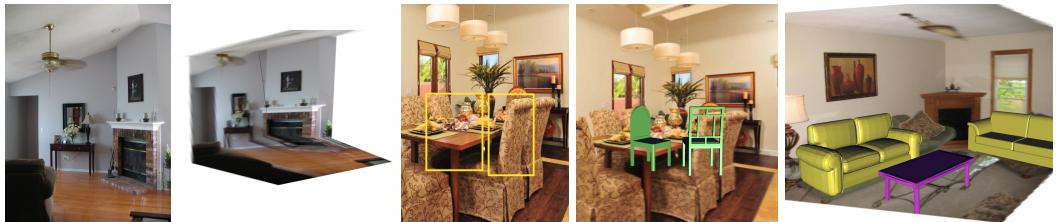


图 6.1 本方法的失效案例：(1)(2) 说明了本方法针对不满足曼哈顿假设的房间造型不起作用；(3)(4) 说明了被遮挡物体的识别效果并不鲁棒（注意右侧的包围盒实际上圈出的是最里面的椅子，而不是靠在桌子右面并排的两把椅子）；(5) 说明了优化算法进考虑到了数据集先验概率，没有针对图片更深层的考虑

在之后的工作中，可以尝试对现有的检索-姿态估计框架进行扩展，使得算法在预测的时候即对物体上下文以及图像整体特征有所考虑；另外的一个研究方向是如何进行多帧的融合：即当用户有条件采集更多的房屋数据时，如何将新加入的结果融合到已有检测结果中来提高重建效果。

最后需要指出的是，本工作仅仅是图形与视觉交叉领域的工作之一：基于机器视觉和深度学习技术的三维重建方法有效利用了图像信息，相比传统方法能够获得更优的重建效果。但是二者的交叉结合远不止三维重建这一项应用，在图形学经典领域例如渲染、流体模拟等在现阶段都已经逐步开始尝试跨领域的新方法和新技巧，而这种研究在未来还应当为继续重视和挖掘，需要靠研究者们的不断努力。

插图索引

图 1.1 现有的三维重建算法及系统效果展示	2
图 1.2 大小恒常性错觉实验	3
图 3.1 本方法处理的共 6 种房间几何类型	13
图 3.2 房屋布局预测网络整体架构	14
图 3.3 利用低阶特征对预测结果进行视角修正的例子	16
图 3.4 投影相机模型	17
图 3.5 房屋布局修正算法示例	19
图 4.1 两种三维形状编码示例图	22
图 4.2 多任务学习架构示意图	23
图 4.3 三维模型位置缩放优化算法演示	26
图 4.4 三维物体贴图算法效果	28
图 4.5 映射纹理贴图时不同变换方式的对比	28
图 5.1 使用本文方法在 LSUN2015 数据集上的结果	31
图 5.2 使用分割网络模型在 LSUN2015 数据集上的结果	31
图 5.3 分割网络模型和本文方法在 Hedau 数据集上结果对比	32
图 5.4 输入图像种类不在可预测范围内时的结果	32
图 5.5 房屋几何恢复结果	33
图 5.6 使用自编码解码器进行 UDF 学习与重建的效果	34
图 5.7 两种编码方式的对比	35
图 5.8 从 SUNCG 数据集中提取的各类转角分布	36
图 5.9 相同输入图片使用不同的形状编码检索的效果	37
图 5.10 不同类别的图像检索效果	38
图 5.11 $O_{i,j}$ 数据可视化	39
图 5.12 \mathcal{W}_i 数据可视化	39
图 5.13 优化算法迭代步骤演示	40
图 5.14 示例一：中亚风格住宅的恢复	43
图 5.15 示例二：欧美家庭客厅的恢复	44
图 5.16 示例三：儿童房间的恢复	45

图 5.17 示例四：实验室讨论区的恢复	46
图 6.1 本方法的失效案例	48

表格索引

表 5.1	房间布局估计网络在 Hedau 数据集上的测试结果	30
表 5.2	房间布局估计网络在 LSUN2015 数据集上的测试结果	30
表 5.3	三维模型检索训练及测试数据	37

参考文献

- [1] Newcombe R A, Izadi S, Hilliges O, et al. Kinectfusion: Real-time dense surface mapping and tracking[C]//IEEE International Symposium on Mixed and Augmented Reality. [S.l.: s.n.], 2012: 127-136.
- [2] Dai A, Izadi S, Theobalt C. Bundlefusion: real-time globally consistent 3d reconstruction using on-the-fly surface re-integration[J]. Acm Transactions on Graphics, 2017, 36(4): 76a.
- [3] Goldstein T, Hand P, Lee C, et al. Shapefit and shapekick for robust, scalable structure from motion[C]//European Conference on Computer Vision. [S.l.: s.n.], 2016: 289-304.
- [4] Ozysil O, Singer A. Robust camera location estimation by convex programming[C]//Computer Vision and Pattern Recognition. [S.l.: s.n.], 2015: 2674-2683.
- [5] Engel J, Schöps T, Cremers D. Lsd-slam: Large-scale direct monocular slam[C]//European Conference on Computer Vision. [S.l.]: Springer, 2014: 834-849.
- [6] Mur-Artal R, Tardós J D. Orb-slam2: An open-source slam system for monocular, stereo, and rgbd cameras[J]. IEEE Transactions on Robotics, 2017, 33(5): 1255-1262.
- [7] He K, Gkioxari G, Dollár P, et al. Mask r-cnn[C]//IEEE International Conference on Computer Vision. [S.l.: s.n.], 2017: 2980-2988.
- [8] Russakovsky O, Deng J, Su H, et al. ImageNet Large Scale Visual Recognition Challenge[J]. International Journal of Computer Vision (IJCV), 2015, 115(3): 211-252.
- [9] Wei S E, Ramakrishna V, Kanade T, et al. Convolutional pose machines[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. [S.l.: s.n.], 2016: 4724-4732.
- [10] Lin G, Milan A, Shen C, et al. Refinenet: Multi-path refinement networks for high-resolution semantic segmentation[C]//IEEE Conference on Computer Vision and Pattern Recognition. [S.l.: s.n.], 2017: 5168-5177.
- [11] Zhao H, Shi J, Qi X, et al. Pyramid scene parsing network[C]//IEEE Conference on Computer Vision and Pattern Recognition. [S.l.: s.n.], 2017: 6230-6239.
- [12] V B, A K, R C. Segnet: A deep convolutional encoder-decoder architecture for scene segmentation.[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, PP(99): 1-1.
- [13] Long J, Shelhamer E, Darrell T. Fully convolutional networks for semantic segmentation[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 39(4): 640-651.
- [14] Su H, Qi C R, Li Y, et al. Render for cnn: Viewpoint estimation in images using cnns trained with rendered 3d model views[C]//IEEE International Conference on Computer Vision. [S.l.: s.n.], 2016: 2686-2694.

- [15] Chang A X, Funkhouser T, Guibas L, et al. Shapenet: An information-rich 3d model repository [J]. Computer Science, 2015.
- [16] Jackson A S, Bulat A, Argyriou V, et al. Large pose 3d face reconstruction from a single image via direct volumetric cnn regression[C]//Computer Vision (ICCV), 2017 IEEE International Conference on. [S.l.]: IEEE, 2017: 1031-1039.
- [17] Liu J, Yu F, Funkhouser T. Interactive 3d modeling with a generative adversarial network[J]. arXiv preprint arXiv:1706.05170, 2017.
- [18] Curless B, Levoy M. A volumetric method for building complex models from range images[C]// Proceedings of the 23rd annual conference on Computer graphics and interactive techniques. [S.l.]: ACM, 1996: 303-312.
- [19] Nießner M, Zollhöfer M, Izadi S, et al. Real-time 3d reconstruction at scale using voxel hashing [J]. ACM Transactions on Graphics (ToG), 2013, 32(6): 169.
- [20] Keller M, Lefloch D, Lambers M, et al. Real-time 3d reconstruction in dynamic scenes using point-based fusion[C]//3DV 2013, 2013 International Conference on 3D Vision. [S.l.]: IEEE, 2013: 1-8.
- [21] Whelan T, Leutenegger S, Salas-Moreno R, et al. Elasticfusion: Dense slam without a pose graph[C]//[S.l.]: Robotics: Science and Systems, 2015.
- [22] Chen K, Lai Y, Wu Y X, et al. Automatic semantic modeling of indoor scenes from low-quality rgb-d data using contextual information[J]. ACM Transactions on Graphics, 2014, 33(6).
- [23] Guo R, Zou C, Hoiem D. Predicting complete 3d models of indoor scenes[J]. arXiv preprint arXiv:1504.02437, 2015.
- [24] Song S, Zeng A, Chang A X, et al. Im2pano3d: Extrapolating 360 structure and semantics beyond the field of view[J]. arXiv preprint arXiv:1712.04569, 2017.
- [25] Mur-Artal R, Montiel J M M, Tardos J D. ORB-SLAM: a versatile and accurate monocular SLAM system[J]. IEEE Transactions on Robotics, 2015, 31(5): 1147-1163.
- [26] Zheng Y, Chen X, Cheng M M, et al. Interactive images: cuboid proxies for smart image manipulation.[J]. ACM Trans. Graph., 2012, 31(4): 99-1.
- [27] Liu Z, Zhang Y, Wu W, et al. Model-driven indoor scenes modeling from a single image [C]//Proceedings of the 41st Graphics Interface Conference. [S.l.]: Canadian Information Processing Society, 2015: 25-32.
- [28] Izadinia H, Qi S, Seitz S M. Im2cad[C]//IEEE Conference on Computer Vision and Pattern Recognition. [S.l.: s.n.], 2017: 2422-2431.
- [29] Chang A, Dai A, Funkhouser T, et al. Matterport3d: Learning from rgb-d data in indoor environments[J]. arXiv preprint arXiv:1709.06158, 2017.
- [30] Dai A, Chang A X, Savva M, et al. Scannet: Richly-annotated 3d reconstructions of indoor scenes[C]//Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR): volume 1. [S.l.: s.n.], 2017.

- [31] Song S, Yu F, Zeng A, et al. Semantic scene completion from a single depth image[J]. IEEE Conference on Computer Vision and Pattern Recognition, 2017.
- [32] Coughlan J M, Yuille A L. Manhattan world: Compass direction from a single image by bayesian inference[C]//The Proceedings of the Seventh IEEE International Conference on Computer Vision: volume 2. [S.l.]: IEEE, 1999: 941-947.
- [33] Lee D C, Hebert M, Kanade T. Geometric reasoning for single image structure recovery[C]// Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on. [S.l.: s.n.], 2009: 2136-2143.
- [34] Hedau V, Hoiem D, Forsyth D. Recovering the spatial layout of cluttered rooms[C]//IEEE International Conference on Computer Vision. [S.l.: s.n.], 2010: 1849-1856.
- [35] Lee D, Gupta A, Hebert M, et al. Estimating spatial layout of rooms using volumetric reasoning about objects and surfaces[J]. Advances in Neural Information Processing Systems, 2010: 1288-1296.
- [36] Zhao Y, Zhu S C. Scene parsing by integrating function, geometry and appearance models[C]// IEEE Conference on Computer Vision and Pattern Recognition. [S.l.: s.n.], 2013: 3119-3126.
- [37] Ramalingam S, Pillai J K, Jain A, et al. Manhattan junction catalogue for spatial reasoning of indoor scenes[C]//Computer Vision and Pattern Recognition. [S.l.: s.n.], 2013: 3065-3072.
- [38] Wang H, Gould S, Koller D. Discriminative learning with latent variables for cluttered indoor scene understanding[C]//European Conference on Computer Vision. [S.l.: s.n.], 2010: 497-510.
- [39] Mallya A, Lazebnik S. Learning informative edge maps for indoor scene layout prediction [C]//IEEE International Conference on Computer Vision. [S.l.: s.n.], 2015: 936-944.
- [40] Dasgupta S, Fang K, Chen K, et al. Delay: Robust spatial layout estimation for cluttered indoor scenes[C]//Computer Vision and Pattern Recognition. [S.l.: s.n.], 2016: 616-624.
- [41] Lee C Y, Badrinarayanan V, Malisiewicz T, et al. Roomnet: End-to-end room layout estimation [J]. arXiv preprint arXiv:1703.06241, 2017.
- [42] Zou C, Colburn A, Shan Q, et al. Layoutnet: Reconstructing the 3d room layout from a single rgb image[J]. arXiv preprint arXiv:1803.08999, 2018.
- [43] Girshick R, Donahue J, Darrell T, et al. Rich feature hierarchies for accurate object detection and semantic segmentation[C]//IEEE Conference on Computer Vision and Pattern Recognition. [S.l.: s.n.], 2014: 580-587.
- [44] Ren S, He K, Girshick R, et al. Faster r-cnn: Towards real-time object detection with region proposal networks[C]//Advances in neural information processing systems. [S.l.: s.n.], 2015: 91-99.
- [45] Redmon J, Divvala S, Girshick R, et al. You only look once: Unified, real-time object detection [C]//Computer Vision and Pattern Recognition. [S.l.: s.n.], 2016: 779-788.

- [46] Redmon J, Farhadi A. Yolo9000: Better, faster, stronger[C]//IEEE Conference on Computer Vision and Pattern Recognition. [S.l.: s.n.], 2017: 6517-6525.
- [47] Liu W, Anguelov D, Erhan D, et al. Ssd: Single shot multibox detector[C]//European conference on computer vision. [S.l.]: Springer, 2016: 21-37.
- [48] Girshick R, Radosavovic I, Gkioxari G, et al. Detectron[Z]. [S.l.: s.n.], 2018.
- [49] Lepetit V, Moreno-Noguer F, Fua P. Epnp: An accurate o (n) solution to the pnp problem[J]. International journal of computer vision, 2009, 81(2): 155.
- [50] Pavlakos G, Zhou X, Chan A, et al. 6-dof object pose from semantic keypoints[C]//Robotics and Automation (ICRA), 2017 IEEE International Conference on. [S.l.]: IEEE, 2017: 2011-2018.
- [51] Grabner A, Roth P M, Lepetit V. 3d pose estimation and 3d model retrieval for objects in the wild[J]. arXiv preprint arXiv:1803.11493, 2018.
- [52] Mottaghi R, Xiang Y, Savarese S. A coarse-to-fine model for 3d pose estimation and sub-category recognition[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. [S.l.: s.n.], 2015: 418-426.
- [53] Chen D Y, Tian X P, Shen Y T, et al. On visual similarity based 3d model retrieval[C]//Computer graphics forum: volume 22. [S.l.]: Wiley Online Library, 2003: 223-232.
- [54] Li Y, Su H, Qi C R, et al. Joint embeddings of shapes and images via cnn image purification. [J]. ACM Trans. Graph., 2015, 34(6): 234-1.
- [55] Tasse F P, Dodgson N. Shape2vec: semantic-based descriptors for 3d shapes, sketches and images[J]. ACM Transactions on Graphics (TOG), 2016, 35(6): 208.
- [56] 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. [S.l.: s.n.], 2016: 82-90.
- [57] Wu J, Wang Y, Xue T, et al. Marrnet: 3d shape reconstruction via 2.5 d sketches[C]//Advances In Neural Information Processing Systems. [S.l.: s.n.], 2017: 540-550.
- [58] Han X, Li Z, Huang H, et al. High-resolution shape completion using deep neural networks for global structure and local geometry inference[J]. arXiv preprint arXiv:1709.07599, 2017.
- [59] Fan H, Su H, Guibas L. A point set generation network for 3d object reconstruction from a single image[C]//Conference on Computer Vision and Pattern Recognition (CVPR): volume 38. [S.l.: s.n.], 2017.
- [60] Groueix T, Fisher M, Kim V G, et al. Atlasnet: A papier-m\`{a}ch\'{e} approach to learning 3d surface generation[J]. arXiv preprint arXiv:1802.05384, 2018.
- [61] Zhang Y, Yu F, Song S, et al. Large-scale scene understanding challenge: Room layout estimation[J]. accessed on Sep, 2015, 15.
- [62] Gers F A, Schraudolph N N. Learning precise timing with lstm recurrent networks[M]. [S.l.]: JMLR.org, 2003: 115-143

- [63] Caprile B, Torre V. Using vanishing points for camera calibration[J]. International Journal of Computer Vision, 1990, 4(2): 127-139.
- [64] Rother C. A new approach for vanishing point detection in architectural environments[C]// British Machine Vision Conference. [S.l.: s.n.], 2000: 40.1-40.10.
- [65] Xiang Y, Kim W, Chen W, et al. Objectnet3d: A large scale database for 3d object recognition [C]//European Conference on Computer Vision. [S.l.]: Springer, 2016: 160-176.
- [66] Elhoseiny M, El-Gaaly T, Bakry A, et al. A comparative analysis and study of multiview cnn models for joint object categorization and pose estimation[C]//International Conference on Machine learning. [S.l.: s.n.], 2016: 888-897.
- [67] Xu K, Chen K, Fu H, et al. Sketch2scene: sketch-based co-retrieval and co-placement of 3d models[J]. ACM Transactions on Graphics (TOG), 2013, 32(4): 123.
- [68] Barnes C, Shechtman E, Finkelstein A, et al. Patchmatch: A randomized correspondence algorithm for structural image editing[J]. ACM Transactions on Graphics-TOG, 2009, 28(3): 24.
- [69] Paszke A, Gross S, Chintala S, et al. Automatic differentiation in pytorch[C]//NIPS-W. [S.l.: s.n.], 2017.
- [70] 薛瑞尼. THUThESIS: 清华大学学位论文模板[EB/OL]. 2017. <https://github.com/xueruini/thuthesis>.

致 谢

衷心感谢胡事民教授对本人的精心指导，作为我的导师，他在整个本科毕业论文的选题、实验上都给了我很大的启发和帮助，并且时刻关心我的进展，并在我遇到困难的时候给予我信心与动力，花费了很多时间和经历来培养我，非常感谢您的教诲、也感谢您愿意收留我继续在您门下攻读博士学位。

同时，也感谢图形学实验室里的曹炎培、汪淼、梁缘、范若琛、梁盾、李瑞龙等师兄以及同级兄弟雷凯翔、方晓楠的给以的帮助和讨论，他们为我研究课题中许多具体的细节给出了很多有用且行之有效的建议。

感谢我的父母，他们无时无刻不在给予我精神上的支持和鼓励，特别感谢我的父亲对我毕业设计选题方面的关心和建议。

感谢毕设期间一直陪伴着我的同学们，他们中有我大学期间的挚友柳荫(JBN)；从高中时期就认识的冯岩、王博雅和张元鑫；在社工圈里相互鼓励的宣传中心和新闻中心的同学们；同住一个寝室关照对方的室友……是你们为我的生活增添了别样的色彩。

最后照例感谢 L^AT_EX 和 THU^THESIS^[70]，帮我节省了不少排版时间和添加引用的时间。

声 明

本人郑重声明：所呈交的学位论文，是本人在导师指导下，独立进行研究工作所取得的成果。尽我所知，除文中已经注明引用的内容外，本学位论文的研究成果不包含任何他人享有著作权的内容。对本论文所涉及的研究工作做出贡献的其他个人和集体，均已在文中以明确方式标明。

签 名: 黄家晖 日 期: 2018.6.19

附录 A 外文资料的调研阅读报告或书面翻译

3DLite: 消费者级别的 3D 扫描仪也能进行内容创造

摘要: 本文中我们介绍了 3DLite，一种能够使用消费者级别 RGB-D 传感器来重建三维场景的新方法，向直接将采集的三维数据应用于诸如游戏、VR 或是 AR 之类的图形学应用迈进了一步。和一笔一画地精确重建真实世界不同，我们的方法将输出一个轻量级的、由简单几何体组成的三维表示来近似扫描到的几何形状。我们认为，对于大多数图形应用而言，获取高质量的物体表面贴图比获得高精度的几何形状要重要得多。为了解决这个问题，我们通过对低质量的 RGB 图片区块进行变形与拼接来补偿运动模糊、自动曝光调整和相机姿态的不对齐所带来的效果缺陷，从而得到高清晰度、锋利的表面纹理。除了场景中观察到的区域之外，我们还对场景的几何形状和纹理进行外推延伸，来获得环境的完整三维模型。我们证明了用简单的平面来抽象场景的几何形状非常适合补全任务，能够让 3DLite 输出完整、轻量并且看起来很真实的三维场景模型。我们相信这种类似 CAD 的重建是迈向使用 RGB-D 扫描器进行内容创作的重要一步。

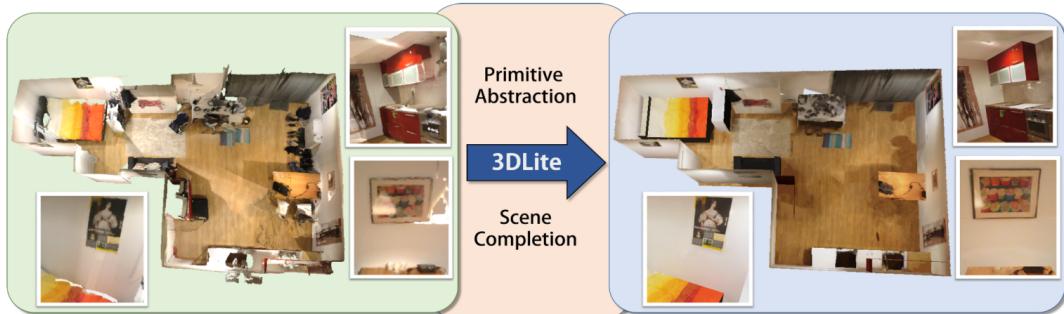


图 1 3DLite 的输入是消费者级别 RGB-D 传感器得到的三维重建模型。初始的三维重建（左）是由现有的最新算法得到的。但是模型可能存在漏洞、几何噪声以及模糊的贴图等问题。3DLite 首先计算场景的几何抽象进行简化，基于这个我们发明了一种算法来生成高质量的纹理贴图。另外，我们还能够对场景进行补全，并使用图像填充的方式补全未观察到的部分的纹理。最后，我们就能获得一个轻量级的、仅由少量多边形组成的网格模型。右侧展示了完整了房屋模型以及其对应的高质量贴图。

A.1 简介

随着诸如微软 Kinect、英特尔 RealSense 和谷歌 Tango 等消费者级别的传感器逐步进入市场，RGB-D 扫描技术在近年来也在飞速发展。最新的在线或离线的三维重建技术已经可以高精度捕捉真实环境中形形色色的物体并将其电子化。虽然这些方法潜在的应用很广泛（例如游戏、虚拟现实、增强现实等等），但他们最终输出的三维模型的质量都很难达到真实建模师的水准。事实上，现有的重建方法依然存在噪声干扰、过渡平滑或是存在缺口的问题，这使得他们还不足以应用于生产环境。

在许多图形应用中，我们发现表面纹理相比几何形状对于人的视觉来讲更重要；例如，许多电视游戏使用看板技术（Billboarding）或是凹凸贴图技术来达到小计算量、高视觉细节的效果，和使用准确三维模型进行渲染的效果几乎分辨不出。不幸的是，现有三维重建方法的颜色质量通常受到运动模糊或是 Rolling shutter 现象的影响。有时还会由于流行重建技术导致的过度平滑和相机位置微小的不对齐进一步损失颜色质量。例如，奠基性的 Volumetric fusion 工作常常被用于从连续的 RGB-D 帧生成三维模型。这种算法在投影深度和观测颜色二者之间计算加权平均，虽然有效控制了噪声，这也导致了几何形状和颜色的过度平滑。另外，由于相机角度是从不精确的颜色值和带有噪声的低清深度数值计算而来的，导致相机姿态的微小偏移，进一步降低了重建质量。不仅如此，如果需要在游戏中实际应用，重建的环境不能有洞，但在真实的扫描时，由于物体的遮挡，漏洞在所难免。因此，我们的目标就是使用相对简单的场景几何表示来提高纹理获取的质量，同时也对被遮挡的部分进行几何内容与纹理的修补。

在本工作中，我们提出了 3DLite，一种能够产生轻量级、完整、类似 CAD 模型的、拥有高质量贴图的室内场景几何表示。我们从消费者级别的手持传感器获取 RGB-D 视频序列，并首先通过已有的三维重建方法重建场景。由于我们希望生成完整的场景以及锋利且干净的贴图，我们使用简单几何体作为扫描到的环境的抽象。具体来说，我们使用平面作为几何体，这是因为平面的几何结构有助于纹理映射以及场景补全，以此能够生成去噪声的场景。我们首先基于曼哈顿假设对这些几何体进行优化，这是因为人造场景通常高度地结构化。为了补全遮挡部分，我们提出了一种新的补洞方法，通过相机轨迹得到未知空间，并以此对平面进行延伸。也就是说，我们尊重场景中已知的空白区域，仅仅修补相机没有捕捉到的区域的洞。这样一来就能生成场景完整的几何表示。接着我

们执行一个纹理优化算法，将纹理从输入的 RGB 数据映射到场景上。为了对整个场景生成完整的纹理，我们对未观测到的空洞区域进行纹理补全。从 RGB-D 数据估测的相机位置经常有微小的偏移，我们的纹理优化步骤将求解刚性和非刚性的图像对齐问题，并根据稀疏的颜色信息和稠密的几何信息来优化使得恢复出的场景与图片尽量接近。为了减轻运动模糊和自动曝光的问题，我们进行了曝光矫正，并仅将输入图像最锋利的区域黏贴在一起。最后，对于因为遮挡而丢失颜色的区域，我们使用场景中拥有相似纹理的部分的颜色数据进行填充。

3DLite 是一个全自动、端对端的框架，它被设计用来产生高质量的贴图和清晰完整轻量级的几何形状，这一切都只需要一个消费者级别的 RGB-D 传感器。我们在许多真实世界的场景展现了该方法的有效性，证明即使使用了简化的几何结构，也能从视觉上和真实结构非常接近。

总的来说，我们主要贡献了一个全自动的三维重建系统，拥有以下特性：

1. 对场景的轻量级几何抽象；
2. 表面贴图拥有很高的质量；
3. 无论是形状还是贴图都能生成完整的表示。

A.2 相关工作

三维重建一直都是重点研究话题。在这一部分，我们介绍一些当前最好最流行的通过扫描场景产生三维模型的方法，同时也介绍通过平面先验来指导重建、优化纹理映射的类似工作。

三维重建 在许多最新的在线或离线三维重建方法中，最常用的计算表面颜色的方法就是计算输入图像中对应颜色的滑动平均值。Volumetric fusion [Curless and Levoy 1996] 方法，被目前许多最新的在线和离线方法使用，这是一个生成场景隐式表面表示并减小输入噪声的行之有效的方法。但不幸的是，这种方法也会导致几何形状和颜色的过度平滑，这是因为深度信息和颜色信息被不同视角的帧通过滑动平均的方法融合了。相似地，基于表面小片（Surfel）的重建方法，能够生成场景的点阵表示，也同样倾向于使用滑动平均来计算表面点的位置和颜色。这些平均方法所带来的过度平滑问题也经常夹杂着对相机姿态估计的微小误差问题（这是因为相机姿态是由带噪声、模糊的深度信息和颜色计算而来）。

平面近似与基于图像的渲染 使用二维图像来代替距离观测者较远的三维几何体（通常这个二维图像叫做 *impostors*, 顶替者）经常在基于图像的渲染领域被讨论。这些图像实际上是带有纹理的平面，当他们被正确地安放之后，从许多角度看产生的视觉效果是大致正确的。基于图像的渲染最近也常被用于自定义的渲染器中，使用简单的几何体来代替，在使用稀疏的 DSLR 输入图片作为输入之后，能产生高质量的结果。

使用平面近似进行室内或室外三维场景重建 由于人造环境通常高度结构化，拥有大量的正交和平行结构，人们便利用曼哈顿平面假设来帮助室内场景的跟踪，尤其是在 RGB-D 扫描的应用场景下，因为在这种情况下深度信息直接就可以获取得到。许多方法都集成了平面信息来提高 3D 扫描场景中相机轨迹跟踪的效果。Dou et al. [2012] 和 Taguchi et al. [2013] 都将各式各样的带有特征点的平面应用于增强轨迹跟踪鲁棒性。Zhang et al. [2015] 同时检测平面结构以及重复物体来减轻相机漂移误差，在 KinectFusion 框架下获得了更好的重建效果。为了减少对初始检测到的结构对应误差的敏感性，Halber and Funkhouser [2017] 使用了一种层级式的优化方式，这种方法混合了平面关系约束和稀疏特征，使得超长时间的 RGB-D 扫描序列的配准（registration）变得可能。

除了增强跟踪的鲁棒性之外，Dzitsiuk et al. [2017] 将平面先验直接应用到场景的隐式距离场表示中，能够进一步对重建结果进行去噪和补全，即使在实时更新的情况下，也能产出很稳定的几何形状。在检测到的平面附近的表面度量被换成了平面度量来减少噪声，并对平面进行外推来修补重建漏洞。我们的方法和这个方法很相近，也是使用了平面作为三维场景的干净几何表示方式，且利用平面特征来补全场景；但是在 Dzitsiuk 的工作中，仅仅注重几何形状上的噪声去除和补全，而我们的工作 3DLite 的目标是通过生成高质量的纹理贴图来创造一个从视觉上看起来很真实的完整三维模型。

A.2.1 三维重建的颜色优化

现在有许多不同的方法来从多张输入图片生成向几何模型上的颜色映射。许多方法使用人工选择的点对应来提供从图片到模型之间的配准对应。当给定了精确配准的颜色和深度信息对之后，就可以对相机姿态进行优化，来最大化和原图之间的一致性。对于使用消费者级别的传感器进行的三维扫描，通常存在噪声与不精确，Zhou and Koltun [2014] 考虑到了这些因素并通过优化相机的刚性变换，同时求解图片的非刚性变形来最大化稠密的照片一致性。Bi et al [2017]

基于上述工作继续改进：他们合成了一系列照片度量上一致的对齐的彩色图像来提供高质量的纹理映射，即使在几何结果很不准确的时候也能工作的很好。我们同样基于 Zhou et al. 的方法；但是在我们的室内应用场景下，单纯地优化稠密能量项依然会导致算法对初始相机姿态很敏感，且优化容易陷入局部极小值。我们使用一种由稀疏到稠密的优化方式，使用稀疏的颜色特征以及几何约束来帮助稠密能量的收敛。

A.3 方法概览

对于输入的 RGB-D 视频，3DLite 首先对场景计算一个基于几何体的抽象，然后利用这个抽象表示来优化获得高质量的纹理映射，同时将场景中的洞进行补全，详见图 2。我们使用一个手持的消费者级别 RGB-D 传感器来获取输入 RGB-D 流。借助现有的 RGB-D 重建算法（例如 Bundle-Fusion），我们计算每一帧的初始相机姿态以及场景的 TSDF（截断符号距离场）表示，从中我们能提取一个初始的网格模型。为了计算几何抽象，对于每一帧我们检测平面，并根据相机的姿态将这些平面融合到场景中。（详见章 4）我们接着对整体的几何结构进行优化，使得所有的平面都符合曼哈顿世界假设，即鼓励使用正交和平行的结构逼近原场景。

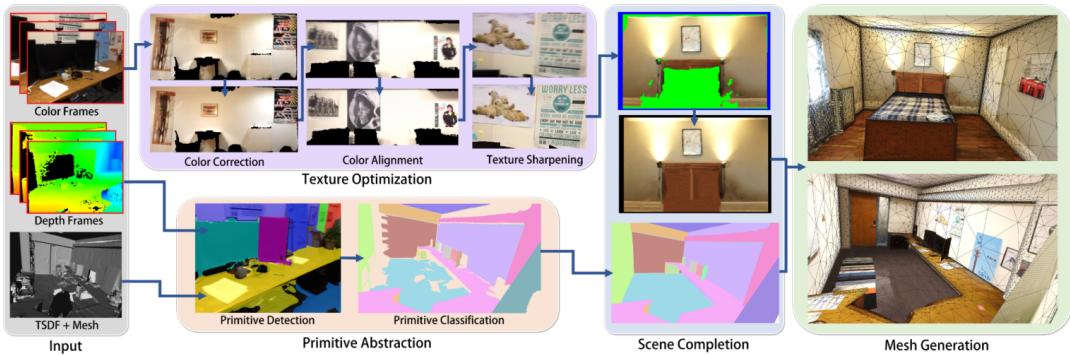


图 2 框架概览：我们的方法输入为 RGB-D 连续帧，从该输入我们计算几何抽象，从中我们能进行模型补全和锐利纹理合成。最终我们得到的是轻量级三维重建的网格模型。

从这种场景的轻量级表示中，我们就能对纹理映射进行优化，直接解决了运动模糊和相机微移的问题。我们使用曝光矫正技术使得不同的输入图片拥有相似的颜色（参见章 5.2）。我们必须修正相机姿态并精确地将颜色与模型对齐。为了解决这个问题，我们基于 Zhou and Koltun[2014] 的方法来优化相机姿态，并

对图像进行非刚体矫正。对于我们这种大型扫描的应用场景，我们引入了稀疏颜色特征以及几何约束来帮助优化算法收敛于稠密光度一致性局部最优解（参见章 5.3）。为了解决输入图像有运动模糊的问题，我们引入了一种方法来锐化从输入帧映射到模型上的图像。我们不采用从整个视频中选取最锐利的帧的方法，因为那样仍然会选到相对模糊的帧，或是由于过滤掉太多的帧而损失颜色信息。我们查找的是锐利的图片区域，之后我们设计了一个基于 Graph-cut 的优化算法来优化图像的锐利性和连贯性（参见章 5.4）。这样以来我们就为已知形状的区域生成了高质量的纹理贴图。

接下来我们需要填补模型缺口。虽然一般的高清场景补全是一个很有挑战性的工作，我们提出的基于几何体的抽象方法能够在场景补全的时候提供有效的补洞方法。为了补洞，我们根据相机的轨迹来延伸基本几何体，直到每个几何体要么接触到另一个几何体，要么接触到确定为空的区域（参见章 6.1）。我们接着使用 Image Melding 的方法对没有颜色的区域进行画布填充（参见章 6.2）。经过了上面这些步骤，一个干净、完整、轻量级、具有高质量锐利贴图的模型就完成了。

A.4 基于基本几何体的抽象

为了计算扫描场景的基本几何体抽象，我们首先对每个输入帧进行平面检测，然后将检测到的平面进行全局统一的融合，最后再进行结构优化，在平行和正交的约束下进行优化。

A.4.1 基于单帧的平面检测

对于一个输入的 RGB-D 视频序列，可以看作深度和颜色帧的集合 $f_i = (C_i, D_i)$ ，我们首先使用现有最新的 RGB-D 重建方法来获得每帧的初始相机姿态 T_i ，以及一个初始的表面 S_0 。对于每帧，我们使用 Feng et al.[2014] 的快速平面提取方法来检测平面。我们不使用每帧的深度信息直接进行平面检测，那样噪声很大且会有很大的空洞。相反地我们使用从 S_0 渲染得到的深度信息。这个渲染得到的深度包含了之前帧累加起来的数据，能够减小噪声、相比于单帧整合更多的信息，并且同时与模型的几何形状 S_0 相符。对于每一个从帧 f_i 中检测出的平面 P_k ，我们附加地存储两项为之后的几何体融合（参见章 4.2）和结构优化（参见章 5.3）使用。

1. 平面参数: $p = (n_x, n_y, n_z, w)$, 其中 $\mathbf{n} = (n_x, n_y, n_z)$ 是单位法向。它代表了帧 f_i 的相机空间中的平面 $\mathbf{n} \cdot \mathbf{x} + w = 0$ 。
2. 距离矩阵: $\mathbf{D} = \frac{1}{N} \sum_q x_q x_q^T$, 其中 x_q 是深度图中的第 q 个像素在世界空间的坐标。这样我们就能轻松计算出 P_k 到 P_l 的逐像素点到平面的平均距离平方 (简称 ASD 距离): $(p_l^j T_j^{-1}) \cdot \mathbf{D} \cdot (p_l^j T_j^{-1})^T$ 。

A.4.2 基本体分类

我们需要聚合每一帧的平面区域, 形成能够代表 3D 场景的一系列平面几何体。来自帧 f_i 的 P_k 以及来自帧 f_j 的 P_l 这两个平面区域, 如果他们很接近, 且有足够的重合部分, 我们就能判断这两个平面区域实际属于一个几何体。关于接近我们使用的判据是: 如果 $\max(ASD_{k \rightarrow l}, ASD_{l \rightarrow k}) < \tau_c$ (实际我们取 $\tau_c = 0.05m$), 那么平面 P_k 和 P_l 就是足够接近的。关于重合, 我们首先将 P_k 转换到 f_j 的相机空间, 然后计算重合像素与属于 P_k 和 P_l 的像素最小值的百分比。如果重合的比例大于一定阈值 τ_o (我们取 $\tau_o = 0.3$), 这样两个区域就被认定为是重合的。

我们层级式地进行这一聚合步骤。对于每隔 $n = 10$ 帧取得的关键帧, 我们首先对这 n 帧进行融合, 然后再在所有的帧之间进行融合。这样, 每帧都包含了一个平面的集合, 集合中的每个平面都根据全局的融合结果打上了匹配的标签。通过将这些标签从每一帧投影到 S_0 上, 我们就能获得场景的几何分解。注意对于有着多于一个标签的区域 (通常出现在平面交汇处) 会被后处理移除。图 3 展示了一个办公室扫描结果的平面分解, 不同的颜色代表不同的平面, 灰色则代表没有平面对应。

为了过滤出上述的灰色区域, 我们首先过滤出平面面积小于 $0.2m^2$ 的区域, 再用 RANSAC 方法过滤出较大的平面的离群值: 对于一个平面 P_k , 为了检查离群的点, 我们随机选择 S_0 中被打上 P_k 标签的 3 个顶点, 并用一个平面 P'_k 来拟合这三个点。如果一个点到 P'_k 的距离小于 τ_i , 那么这个点就不是离群的点。由于 S_0 中的大片平面区域经常出现噪声造成扭曲变形, 所以我们保守地将 τ_i 设置为 $0.1m$ 。我们重复上述过程 128 次, 并用包含最多非离群点的平面作为新的 P_k 。

A.4.3 结构优化

由于我们用了一个相对保守的阈值 τ_i 来进行 RANSAC 更新过程, 原本应该正交或平行的平面现在有了一定角度的偏移。因此我们进行结构优化步骤来鼓励这些平面满足曼哈顿世界约束。与 Halber and Funkhouser[2017] 的方法相似,

我们使用平行和正交的约束来构建一个能量最小化问题：

$$E_s = E_d + \lambda E_a E_d = \sum_i^{\#planes} \sum_j^{\#planeverts} D(P_i, v_{ij})^2 E_a = \sum_{i,j \in \Omega} |A(P_i, P_j) - 90 \cdot n_{ij}|^2 \quad (123)$$

对于每个平面 P , E_d 为平面拟合误差, $D(P, v_i)$ 是非离群顶点 v_i 到平面 P 的距离。 E_a 度量了正交平面的角度误差, 其中 $A(P_i, P_j)$ 是两个平面的夹角, Ω 是平行和正交平面对的集合。为了构建集合 Ω , 我们测试每一对平面, 如果他们之间法向的角度差在 $[90n_{ij} - 10, 90n_{ij} + 10]$, $n_{ij} \in \{0, 1, 2, 3\}$, 我们就将这对平面加入 Ω 中。最终的能量函数 E_s 是 E_d 和 E_a 的线性和, 系数 $\lambda = \frac{E_a^0}{E_d^0}$ 用来平衡平面拟合误差和结构误差。

优化通常在 10 步迭代之内收敛, 最终的角度误差不会超过 1° , 平面拟合误差增大约 1.1 倍。最终, 我们能够在损失很小拟合精度的前提下取得结构上的精确性。从这个优化的结果, 我们将 S_0 的顶点进行投影, 就能产出一个轻量级、干净的网格模型 S_p , 如图 7(a) 所示。

A.5 纹理优化

我们希望为 S_p 映射上锐利、干净的纹理, 从而产生一个从视觉上看起来很真实的 3D 模型。由于运动模糊和相机位置估计不准, 直接将每一帧的像素投射到模型上是不合适的。在我们的纹理优化步骤中, 我们通过求解纹理优化问题来修正相机角度和图像变形问题, 同时对纹理进行锐化, 该锐化步骤考虑到了每个像素的锐利性, 最终能够求解出全局最清晰的颜色。

A.5.1 基于颜色的几何优化

由于我们的平面是一开始是通过渲染深度来进行拟合得到的 (参见章 4.1), 有可能出现几何体边界和输入 RGB-D 帧不吻合的现象。因此, 临近平面边缘的像素常常被误分类, 因此我们必须重新对他们进行分类, 让他们对输入的彩色图像相匹配。为了保证在图像域上平面标签是一致的, 且要使得边界和输入的图像帧保持一致, 我们定义了一个基于 Graph-cut [Boykov et al. 2001] 的能量最小化问题:

$$E_l = \sum_p D(p, l_p) + \sum_{p,q} V_{pq} \delta(l_p, l_q) \quad (4)$$

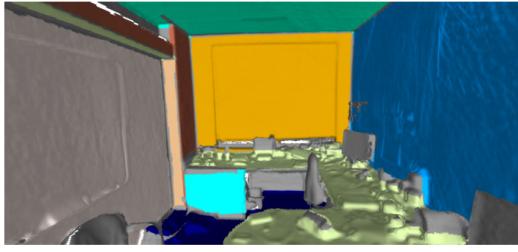


图 3 基本几何体分解：每一种颜色代表不同的平面，灰色区域没有平面关联

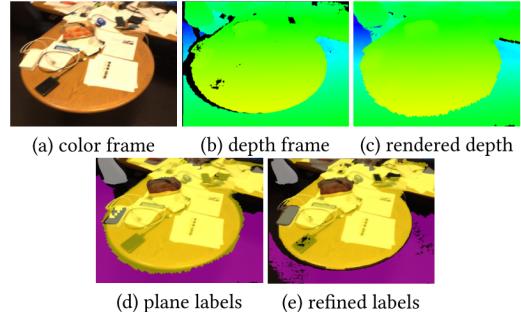


图 4 基于颜色的优化示例

这里 $D(p, l_p)$ 代表像素 p 被分类为 l_p 的惩罚。对于置信度足够大的区域，我们不会更改其标签 l_p^0 。因此当 $l_p = l_p^0$ 的时候，惩罚为 1，否则惩罚为无穷大。对于需要重新分类的区域，对于所有可能的标签 l_p ，惩罚均为 1。我们又另外增加了在平面边界的惩罚，即 $\delta(l_p, l_q)$: p 和 q 是相邻的像素，如果 $l_p \neq l_q$ ，那么 $\delta = 1$ ，否则就是 0。 V_{pq} 度量了从 p 和 q 中间切开有多少收益，这取决于这两个像素颜色的不同。我们设 $V_{pq} = e^{(||C(p)-C(q)||^2)/\sigma^2}$ 。在我们的实验中，对于一个 640×480 的图片，我们针对边缘部分 10 像素宽度的区域重新分类，颜色值 $C(x) \in [0, 255]$ ，且 $\sigma = 30$ 。图 4 展示了我们优化之后的结果。

A.5.2 颜色迁移优化

为了求解纹理映射问题，我们需要确保不同视角下颜色是统一的。我们可以使用一个映射函数来迁移颜色从而解决这一问题。例如，NRDC [HaCohen et al. 2011] 在图像之间寻找对应关系，并依据对应关系建立映射函数，从而实现整体颜色的迁移。在 RGB-D 扫描的背景下，这种对应关系能够轻松地从相机角度和几何中获取。Zhang et al. [2016] 通过给定曝光值后，最小化不同颜色帧中相同顶点的辐射度误差来解决问题。为了求解白平衡和曝光的变化，我们使用 NRDC 模型，使用三条样条曲线（即对第 j 帧的 rgb 通道有映射 $B_j(r, g, b) = (B_j^1(r), B_j^2(g), B_j^3(b))$ ）来迁移颜色。接下来对一个三维点 p_i 和其在帧 $\{j\}$ 对应的像素集合 $\{q_{ij}\}$ ，我们希望三维点的真实颜色 $C(p_i)$ 和 $B_j(q_{ij})$ 足够接近：

$$E_t = \sum_i \sum_j ||C(p_i) - B_j(q_{ij})||^2 + \lambda \sum_i \sum_j (B'_j(x_i) - 1)^2 \quad (5)$$

在 [Zhang et al. 2016] 中，使用的是第一帧作为参考曝光值，我们不需要这种约束。与之不同，我们通过限制迁移函数的导数 $B'_j(x_i) \approx 1$ ，这样能够保证颜色的多样性。我们选择参数 $\lambda = 0.1$ ，且 x_i 为从 0 到 250，间距为 10 的 25 个整数序

列。由于这个优化取决于相机变换的质量，我们将这一步优化和下面介绍的图像对齐优化一同进行。

A.5.3 纹理映射对齐优化

获得高质量贴图的关键是获得精确对齐的颜色帧。这是一个有挑战性的工作，因为消费者级别的传感器总是会采到不精确的数据。Zhou and Koltun [2014]介绍的颜色映射优化方法假设几何形状固定，并共同优化相机角度和图片变形参数。这样简化了问题的可变参数量，使用图片变形的技术来减少来自几何误差的可能的颜色不对齐问题，能够得到较好的结果。然而，他们的优化取决于稠密的光测误差，很难收敛，所以对初始姿态的估计也很敏感。对于我们这种大型扫描的应用场景，这种能量对于初始相机位置的误差并不鲁棒。

在我们的解决方案中，我们针对每一帧分别求解相机参数和矫正参数，这在实际应用中更加有效。我们采用了 EM 优化的方法，借以来自 Zhou and Koltun 提出的稠密颜色误差项，来最大化和照片的一致性。为了帮助收敛，我们附加地向优化中引入了稀疏的特征以及基于平面的限制。我们的能量函数被定义为：

$$E(\mathbf{T}) = E_c(\mathbf{T}) + \lambda_s E_s(\mathbf{T}) + \lambda_p E_p(\mathbf{T}) \quad (\text{A-1})$$

其中 E_c 代表稠密光测误差， E_s 是稀疏特征项， E_p 是平面关系约束项，我们求解的变量为每一帧的相机姿态 $\mathbf{T} = \{T_i\}$ 。和 Zhou and Koltun 相似，我们也同时优化一个代理变量 $C(p_i)$ 代表三维空间点 p_i 的颜色。我们层级式地进行优化，由粗糙到精细，以此来帮助最后的收敛。层级式优化主要分为 3 级， p_i 分别以 0.032m, 0.008m 和 0.002m 的采样间距从 S_p 中采样。参数 λ_s 和 λ_p 按照 $E_c(\mathbf{T}_0) = \lambda_s E_s(\mathbf{T}_0) + \lambda_p E_p(\mathbf{T}_0)$ 给定。

稀疏项 在我们的稀疏匹配项中，我们最小化图像空间中找到的匹配特征对在转换到世界坐标系之后的距离和。对于帧 f_i 中的平面 P_k ，我们使用单映（homography）变换矩阵 H_k^i 来变换点 p_i ，并依赖平面 P_k 和相机姿态 T_i ，将该点变换到 P_k 的空间内。接下来对于来自帧 f_i 和帧 f_j 的像素对应 $\{p_{im}, p_{jm}\}$ ，对应 P_k 和 P_l 的平面，我们优化下面的能量函数：

$$E_s(\mathbf{T}) = \sum_m^{\#corr} ||H_k^i p_{im} - H_l^j p_{jm}||^2 \quad (6)$$

我们将每个颜色帧投影到平面几何体上，并在投影出的图像中检测 SIFT 特征来确定稀疏特征。为了找到稀疏特征的匹配，我们在 SIFT 匹配之后又加入了一步，来核实的确存在一个合理的二维刚性变换能够将一张图象中的所有匹配映射到另一张对应的图像上去。合理的已经匹配的特征紧接着被反投射到原有的颜色帧上来整合得到稀疏的特征集合。这种方法能够产生鲁棒的特征匹配集合，因为上述投影操作不依赖于带噪声和扰动的深度图，而且在变形空间内进行匹配在缩放和仿射不稳定性上都要少于原来的彩色图片。

平面关系约束项 针对每一帧，我们约束章 4.1 计算出的平面区域和 S_p 对齐：

$$E_g(\mathbf{T}) = \sum_i \sum_j (P_j T_j^{-1} p_i)^T (P_j T_j^{-1} p_i) \quad (7)$$

这里， p_i 是从 S_p 中采样到的点的齐次坐标。 P_j 是第 j 帧的相机空间中第 i 个平面的参数。

稠密项 最后，稠密项度量了从彩色帧投影到 S_p 上之后的照片度量误差。

$$E_c(\mathbf{T}) = \sum_i \|C(p_i) - \sum_j I_j(\pi(T_j^{-1} p_i))\|^2 \quad (8)$$

其中 p_i 是从 S_p 上采样的三维点， π 代表投影变换。

我们通过同时求解每帧的变换 \mathbf{T} 和几何形状 $C(p_i)$ 上的代理颜色变量来最终解出 $E(\mathbf{T})$ 。前三次使用最为粗糙的清晰度进行迭代之后，我们就能算出颜色迁移矫正匹配得比较好的集合对；在颜色矫正之后，我们进一步在中清晰度迭代 5 轮和高清晰度下迭代 2 轮，优化求解出更优的相机角度。

在相机姿态求解完成之后，我们接着对图像进行变形优化，并依照前述方法来减少由于可能的几何误差造成颜色不对齐现象。

A.5.4 纹理锐化

当使用手持的商品级 RGB-D 传感器来获取视频时，运动模糊是一个比较常见的问题。即使相机的位置对齐得很完美，模糊的图片也会显著降低模型颜色的质量。大多数现有的方法对所有相对应的帧的颜色取平均值，抑或是使用加权组合在线更新模型颜色。为了减轻这个问题，Zhou and Koltun [2014] 的颜色映射优化方法首先每 1-5 秒选择一个最锐利的帧，其中是否锐利由 Crete et al.[2007] 的方法来判断。但是，当我们把他们的方法直接应用于我们的场景时，并非所有

的模糊帧都被过滤掉了，而且，如果我们选择关键帧的间隔大于 1 秒，一些区域的颜色信息就会丢失。实际上，在房间级别大小的这种设定下，这种方法倾向于保留相机离场景更远的帧的颜色，因为这样的帧通常包含更多的物体，也就包含更大的颜色变化。然而，我们希望保留相机离场景物体更近时的帧中的颜色，因为更近意味着相机能捕捉到更细节的纹理。因此，我们求其中和，希望能将图片中最锐利的区域映射到几何模型上；也就是说，对于 S_p 中的每一个平面，我们将任何一张图像中最锐利的部分映射上去。为了达到这个目的，我们需要同时考虑图像像素的锐利性以及图像区域的锐利性，从而能以 [10,20] 帧的间隔进行关键帧选取。对于图像像素的锐利性，我们使用 [Vu et al. 2012] 的度量方法；对图像区域的锐利性，我们同时考虑像素锐利性和视觉稠密度，视觉稠密度度量从特定视角观察的一个像素的质量（考虑远距离和观察角），他被定义为：

$$r_j(p_i) = ((T_j^{-1}\pi^{-1}(p_i))|_z)^{-2} \cdot \cos(\mathbf{r} \cdot \mathbf{n}) \quad (9)$$

其中 p_i 是第 j 帧的一个像素， \mathbf{r} 是从相机射向 p_i 的射线， \mathbf{n} 是相机空间 p_i 处的法向。对于一个像素而言，整个区域的锐利性就被定义为： $S_j^{reg}(p_i) = S^{pix}(p_i) \cdot r_j(p_i)$ 。

由于我们希望避免因为给相近的 3D 位置采样了很多不同的帧导致的可能噪声，我们提出了一个基于 Graph-Cut 的能量优化函数，能够更进一步地促进帧的连续性：

$$E = \sum_p |S_{l(p)}(p) - S^{max}(p)| + \lambda \sum_{(p,q)} ||C_{l(p)}(\pi(T_{l(p)}^{-1}p)) - C_{l(q)}(\pi(T_{l(q)}^{-1}q))|| \cdot \delta(l_p, l_q) \quad (10)$$

求解的最终结果是 S_p 内的每一个点 p 的帧序号 $l(p)$ 。上式第一项鼓励图像的锐利性， $S^{max}(p)$ 代表 p 最大可能的锐利值。第二项鼓励图像连续性，方法是惩罚被分到不同帧（即 $l(p) \neq l(q)$ ）的相邻点 p, q ，惩罚的大小与各自的颜色差成正比，因此能得到无视觉裂缝的图分割。

图 5 展示了一个有挑战性的例子：一个桌子的纹理由来自 3000 张图像精选出的 147 个最锐利的帧的投影。我们能够看出直接选择拥有最高区域锐利值的帧 ((c),(d)) 已经比 Zhou and Koltun 方法得到的纹理 ((a),(b)) 更好，但是也由于在三维空间上没有帧的连续性（导致出现了截断的电线）依然拥有一些噪声。我们最终的纹理锐化结果 ((e),(f)) 平衡了锐利性和连贯性，在保留锐利细节的同时还减小了噪声。

直接对原有帧的颜色进行采样依然会导致不自然的纹理效果，在过渡的区域尤为明显，参见图 6(b)。因此我们在锐利步骤最后再增加一步来减轻这个效

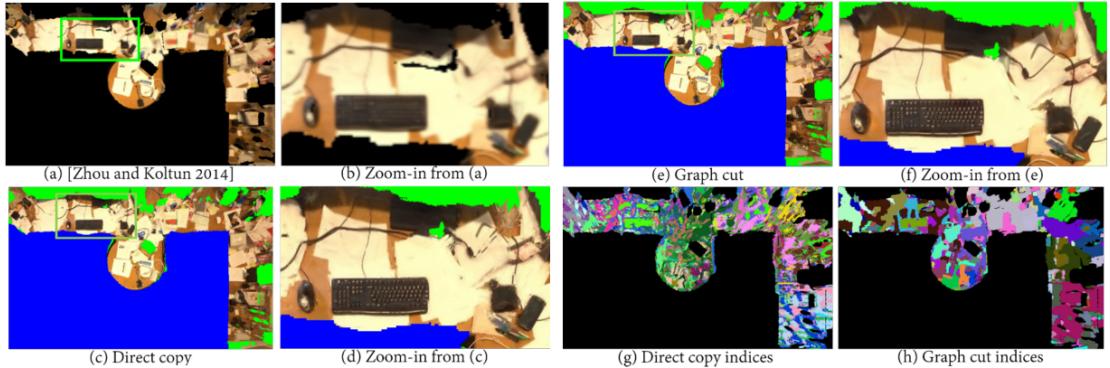


图 5 从多帧进行纹理生成: (a) 颜色映射优化; (b) 是 (a) 的放大版; (c) 直接从最锐利的区域拷贝; (d) 是 (c) 的放大版; (e) 平衡了锐度和区域连贯性; (f) 为 (e) 的放大版; (g) 最锐利区域的帧序号; (h) 优化后的帧序号

果。我们结合求解等式 (10) 得到的颜色图的散度 $\Delta C(p)$ ([Perez et al. 2003] 证明了这种散度能够代表纹理信息) 和经过平均化之后的颜色图 $\bar{C}(p)$, 并用下述等式的最小二乘解来求解最终的纹理图 $F(p)$:

$$E(F) = \sum_p ||F(p) - \bar{C}(p)||^2 + \lambda ||\Delta F(p) - \Delta C(p)||^2 \quad (11)$$

从图 6 中能够看出, 这种方法既保留了图像锐利性, 又去掉了不自然的边界。

A.6 场景补全

虽然我们目前的三维模型已经有了真实的纹理和轻量级的表示, 但依然相对不完整, 这是由于物体遮挡以及有限的扫描覆盖率引起的。三维补全是一个有挑战性的工作, 因为一般来说这需要基于示例的学习方法 [Dai et al. 2017c], 而且对于更高清的分辨率和更大的场景, 问题的难度成立方级别上升。我们因此利用前一步提取出的平面几何体来从几何和颜色两个方面来简化三维场景补全问题。

A.6.1 几何补全

我们可以通过延长平面来对未看见的区域进行补全 [Dzitsiuk et al. 2017]。我们使用了多种规则来指导补全过程, 这些规则如图 8 所示:

1. 两个平面需要被各自延长, 直到他们的交汇处, 如果延长的区域未被观察到;

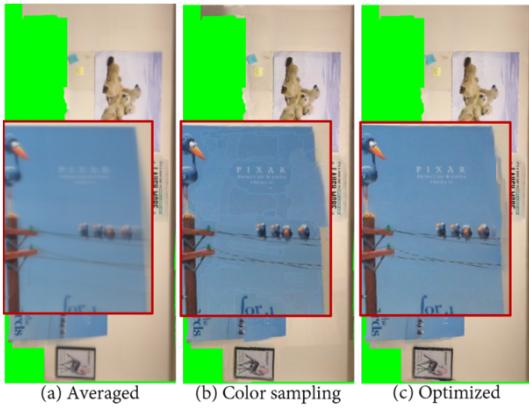


图 6 使用平均颜色和使用散度图进行纹理优化的结果对比: (a) 对帧求平均; (b) 直接从帧中拷贝; (c) 拷贝后进行优化

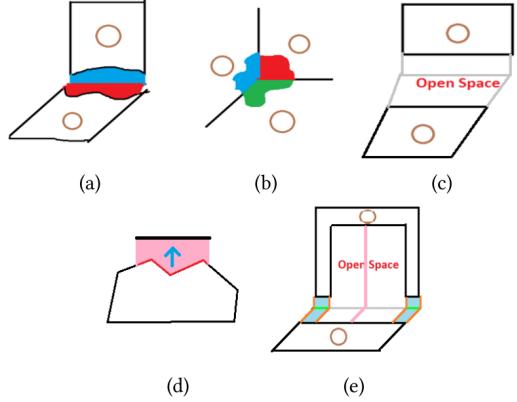


图 8 平面外插规则。其中 (e) 为较复杂的示例: 由于已知为空的区域, 平面不能全部进行外插, 只有蓝色部分才能被进行延长并相交于平面的交线。

2. 如果三个平面两两相交, 延长他们使得他们交汇于一个角落, 如果延长的区域未被观察到;
3. 不应该将平面延伸到开放区域 (有观察且为空)。注意我们使用初始的 TSDF 根据相机的轨迹来确定已知被占用、已知为空和未被观测到的区域;
4. 一个平面内的洞, 如果未被观测到, 需要被填补。

我们首先找到满足第一条规则的所有平面对, 然后进行延长。接着, 我们找到所有满足第二条规则的平面, 并进行延长。我们的算法最终结果与延长的顺序无关, 这是因为最终所有的平面都应该在未观测的区域交汇。对于每一对平面, 我们尝试检测可能可以外插的区域, 接着对所有的可能相交的三平面集合重复同样的操作。最后, 我们填补单个平面内部的洞, 注意, 虽然这个算法在所有主要的平面都已经被观测到的情形下很成功, 我们不能在大块区域都没有被观测到的情形下无中生有 (参见章 8.1 的详细讨论)。

图 7 和图 9 展示了我们的外插填洞与补全算法。

A.6.2 纹理补全

我们基于平面的算法将三维纹理补全问题简化到了二维上。由于这种二维的纹理合成问题是经典问题, 且已经被研究地较为透彻 [参见 Barnes 的一系列工作], 我们使用了目前最新的图像补全技术来对外插的几何区域进行纹理补全。

虽然最近基于深度学习的方法向我们展现了许多令人影响深刻的结果, 这

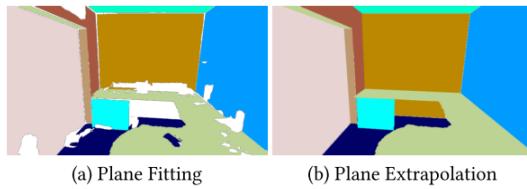


图 7 平面拟合与外插: (a) 进行了平面拟合和结构优化之后, 我们将模型的顶点投射到平面上, 得到干净的网格模型; (b) 我们延长平面来补全被遮挡的部分。

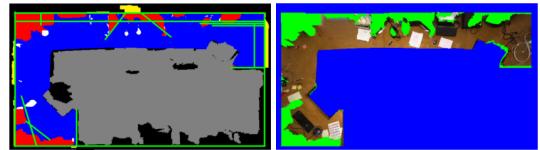


图 9 外插和补洞的示例。左: 绿色线代表了可能的平面相交点, 蓝色区域代表了原有模型的形状, 红色代表了外插的区域, 黄色区域为超出交线而被移除的区域; 灰色区域代表已知为空的区域; 白色区域代表了平面内部洞口的填补。右: 带有纹理的已补全模型。

些方法都需要大量的训练集, 且大多被限制到了固定的图像分辨率上。因此我们使用 Image Melding[Darabi et al. 2012], 一个综合考虑了图像变形和梯度的基于小块的方法, 基于 l_0 和 l_2 距离的混合优化使得它能够得到高质量的图像填充效果。我们发现这种方法在大多数情况下工作得很好, 除了前后背景差距过大的情况、或是背景上有阴影的情况之外 (如图 10(b) 所示)。在这种情况下, 我们使用图像分割的方法检测背景, 并将其使用拉普拉斯平滑的方法延伸到整个图片 (如图 10(c) 所示)。我们接着将合成的背景和前景 (如图 10(d) 所示) 进行结合, 并使用 Image Melding 技术合成距离前景 10 像素以内的区域 (如图 10(e) 所示)。

A.7 网格模型生成

本部分我们讨论生成最终网格模型的过程。我们首先对平面的边缘进行降噪, 接着将平面转化成网格, 输出一个比原有网格模型小约 25 倍的最终模型。

边界优化 平面的边界通常有噪声, 这是由传感器的不精确以及最初的重建造成的。为了对边界进行降噪, 我们通过直线和 B 样条拟合结合的方法进行边界平滑。对于每个在边界上的顶点, 我们用线段来拟合 (每 8mm 采样一次) 其相邻的 51 个邻居顶点。如果这些顶点协方差的条件数大于 50, 我们就认为该顶点属于一条线。我们接着迭代式地将关联同一条线的顶点进行融合, 如果他们对应的线夹角不超过 5° 。对于不属于任何线的连续顶点, 我们用 B 样条进行拟合。

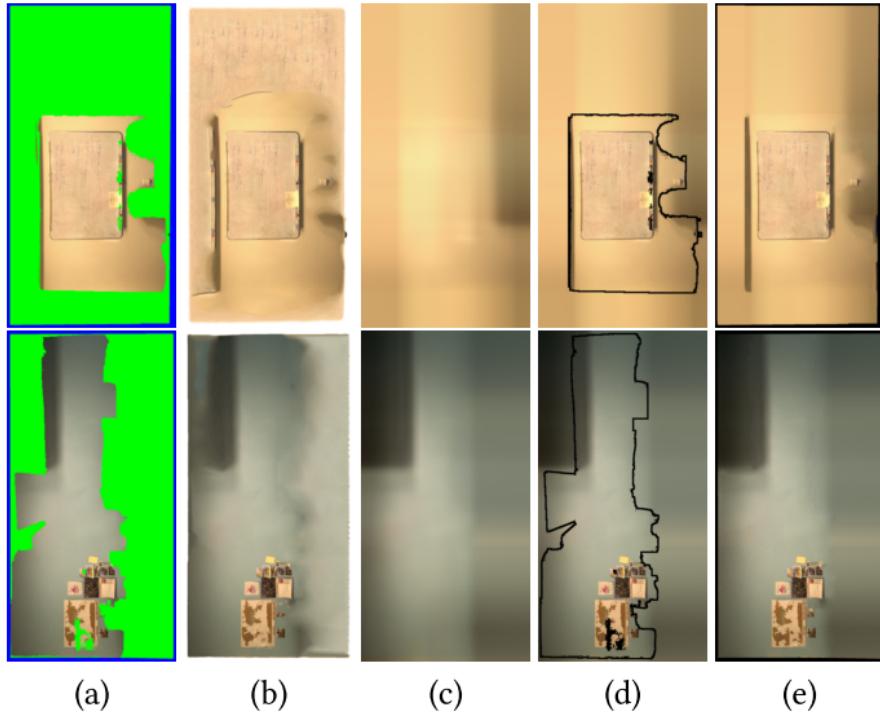


图 10 纹理补全。(a) 原始纹理, 绿色是需要生成的部分; (b) 直接使用 Image Melding 的结果; (c) 背景颜色估计; (d) 背景和原有前景融合; (e) 剩下的像素使用 Image Melding 进行填充。

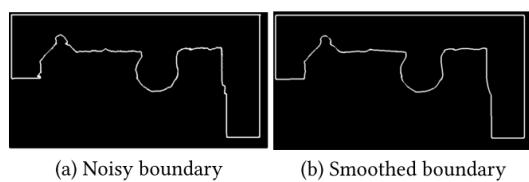


图 11 边界平滑。最初的几何体边界带有噪声 (左); 我们使用线段和 B 样条进行平滑 (右)

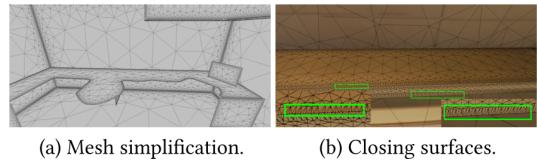


图 12 网格简化以及顶点投射。(a) 我们使用带约束的 Delaunay 三角化来简化平面; (b) 我们使用顶点投射技术来对本应闭合的两个平面进行连接。

平面重建模 为了生成最终的模型, 我们使用受限 Delaunay 三角化 [Chew 1987] 的方法来对平面几何体进行三角化。图 12(a) 展示了算法输出的简化过的三角网格模型, 这种模型精确保留了几何平面的边界。因为我们使用了离散化的像素生成原有的网格边界, 所以在几何平面交界处可能会有微小的缝隙。为了填补这些小缝隙, 我们将相交处附近的顶点投影到相交的几何体上, 如图 12(b) 所示。

A.8 结果

我们在许多室内场景的 RGB-D 扫描数据上测试了我们的方法。所有的扫描数据都是使用挂载到 iPad Air 上的 *Structure Sensor* 获取的。还有另外一些序列来自 BundleFusion 的数据 [Dai et al. 2017b] 和 ScanNet [Dai et al. 2017a] 数据集。我们使用 640×480 的颜色和深度数据，时钟同步为 30Hz。我们的算法对采集设备并不知情。对于每一份扫描序列，我们使用 BundleFusion 计算初始相机姿态。所有的 3DLite 实验在 Intel Core i7 2.6GHz 的 CPU 上完成，每一份扫描需要的处理时间平均约为 5 小时。具体的时间详见原论文，不涉及翻译部分。

定性比较 我们首先在 10 个场景上，将我们的轻量级网格模型与使用 BundleFusion 和 VoxelHashing 的三维重建模型进行对比。注意由于 BundleFusion 的在线重配准步骤使用离散到 byte 的更新，在最终的重建效果中可能会有微小的颜色不自然现象，所以我们首先运行 BundleFusion，获得相机姿态，然后使用 VoxelHashing 技术进行表面重建。如图 13 和图 14 所示，即使使用了简化的几何形状，我们的高质量纹理能够提供视觉上完整的效果，大大减轻了各种颜色过度平滑问题。

几何抽象 在图 15 中我们展示了一些重建的例子，和原有重建效果相比我们能够从中生成降噪且完整的形状，且更加轻量级（同一分辨率下体积减小大于 300 倍）。场景的面数详见原论文，不涉及翻译部分。

和传统的使用 RANSAC 方法进行平面拟合的方法相比，我们的方法在检测相对小的平面时更加鲁棒，如图 16 所示。

颜色对齐 在分析了我们的颜色对齐方法之后，我们发现我们提出的稀疏项和几何约束项很有用，如图 17 所示。如果没有颜色优化的步骤，结果依然会带有过渡平滑和鬼影的效果。使用 Zhou and Koltun[2014] 的方法进行相机姿态优化之后，纹理能够得到锐化，但依然有一些不自然的地方，这是因为初始的相机角度和最终收敛的角度相差太远，很难收敛。我们提出的新一项函数能够准确将彩色帧和几何形状进行对齐。

颜色迁移矫正 在图 18 中我们展示了颜色迁移优化的效果，这种优化不仅能够很大程度上减少由于自动曝光和自动白平衡带来的不自然现象，还能抑制各种

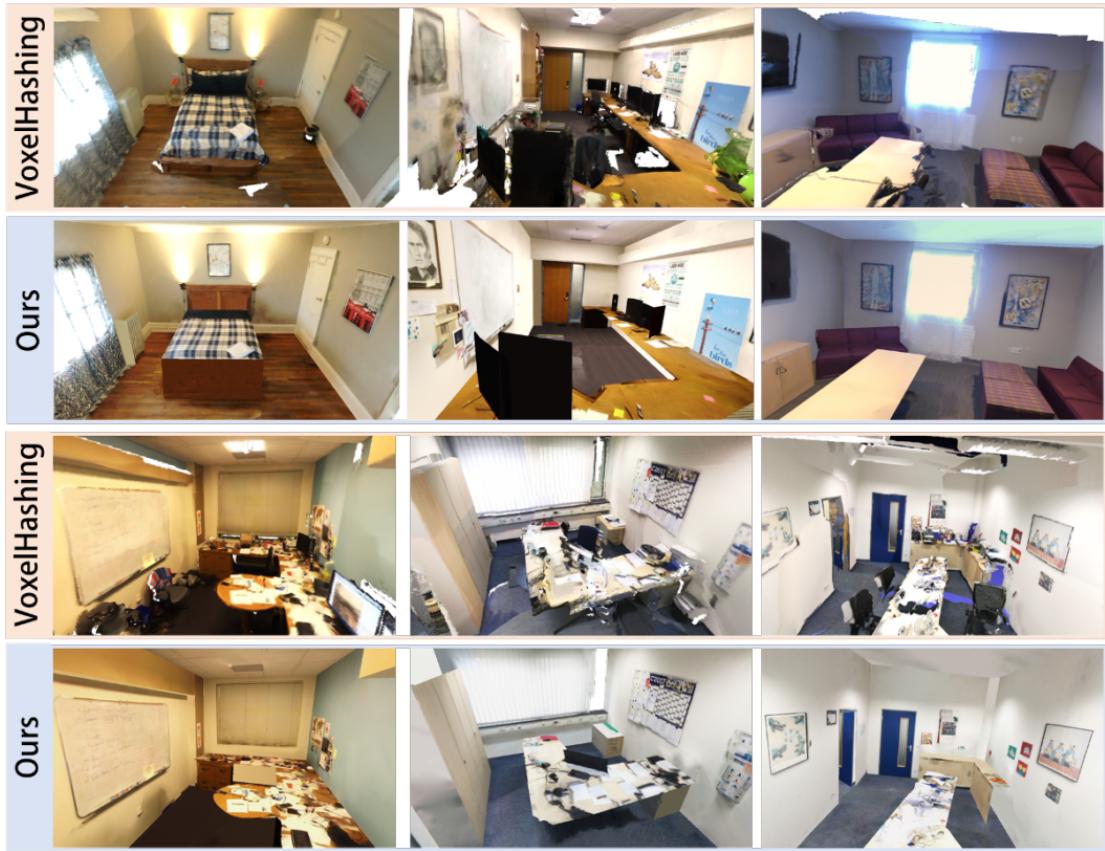


图 13 带有高质量纹理的网格模型。和 BundleFusion 与 VoxelHashing 技术相比，3DLite 能够生成完整的场景，并能让颜色更自然。

颜色不一致现象。

纹理锐化 在图 19 中，我们展现了我们纹理锐化方法的有效性。由于我们利用不同帧最锐利的数据来生成纹理，所以我们的模型渲染得到到图像能够比某些原始的 RGB 图像更加锐利（原始图像经常包含运动模糊），同时也比使用了高级去模糊算法 [Su et al. 2016] 处理过的 RGB 图像更加锐利。这样的去模糊算法，虽然明显减轻了图像中某些部分的模糊，却在图像边缘附近或是在有剧烈运动的部分不能有效去除模糊。

纹理合成 图 20 展示了许多使用 Image Melding 和背景填充进行纹理合成的结果。我们能够产生带纹理的几何补全模型。

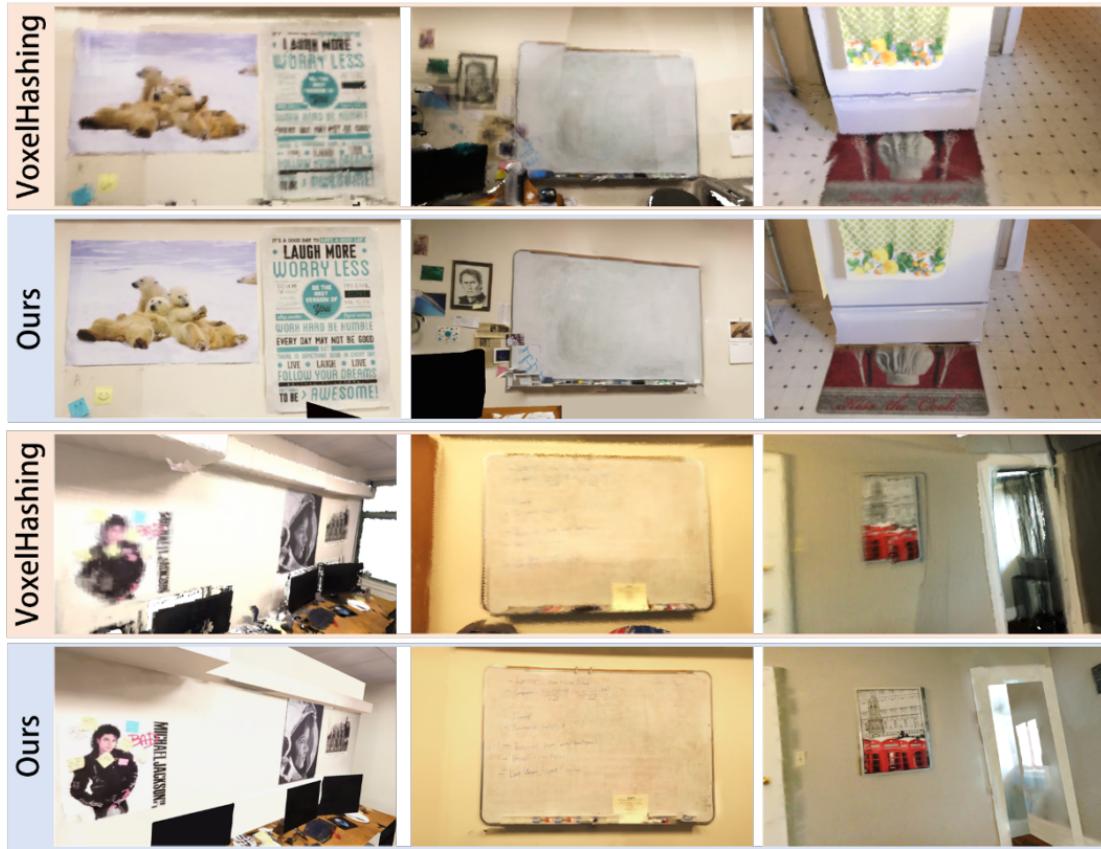


图 14 我们的方法和 BundleFusion 与 VoxelHashing 技术在放大局部细节之后的对比

A.8.1 方法局限

虽然 3DLite 能够鲁棒地生成轻量级、带有锐利纹理的模型，它依然有一些局限性，如图 21 所示。如果在原始的扫描中有一些部分完全没有观测到，我们不能无中生有来补全这些区域。另外，一些小物体（例如键盘、鼠标）常常被投影到了更大的平面几何体上，这和使用 Volumetric Fusion 方法产生的重建结果差的不多，粗糙的分辨率、噪声以及商品级深度传感器的扰动也会使得小物体很难从几何上分辨出来。另外，虽然最新的纹理合成方法能产出令人印象深刻的效果，他们也并不完美，而且也很难合成较大的确实区域的纹理。最重要的是，我们目前的几何体抽象基于平面，所以相对而言非平面的物体（例如某些椅子）并没有被我们的几何抽象步骤纳入进最终模型。我们希望能够进一步扩展我们的方法，和 CAD 模型检索、对齐和贴纹理相结合，从而保留所有场景内的几何形状和纹理。

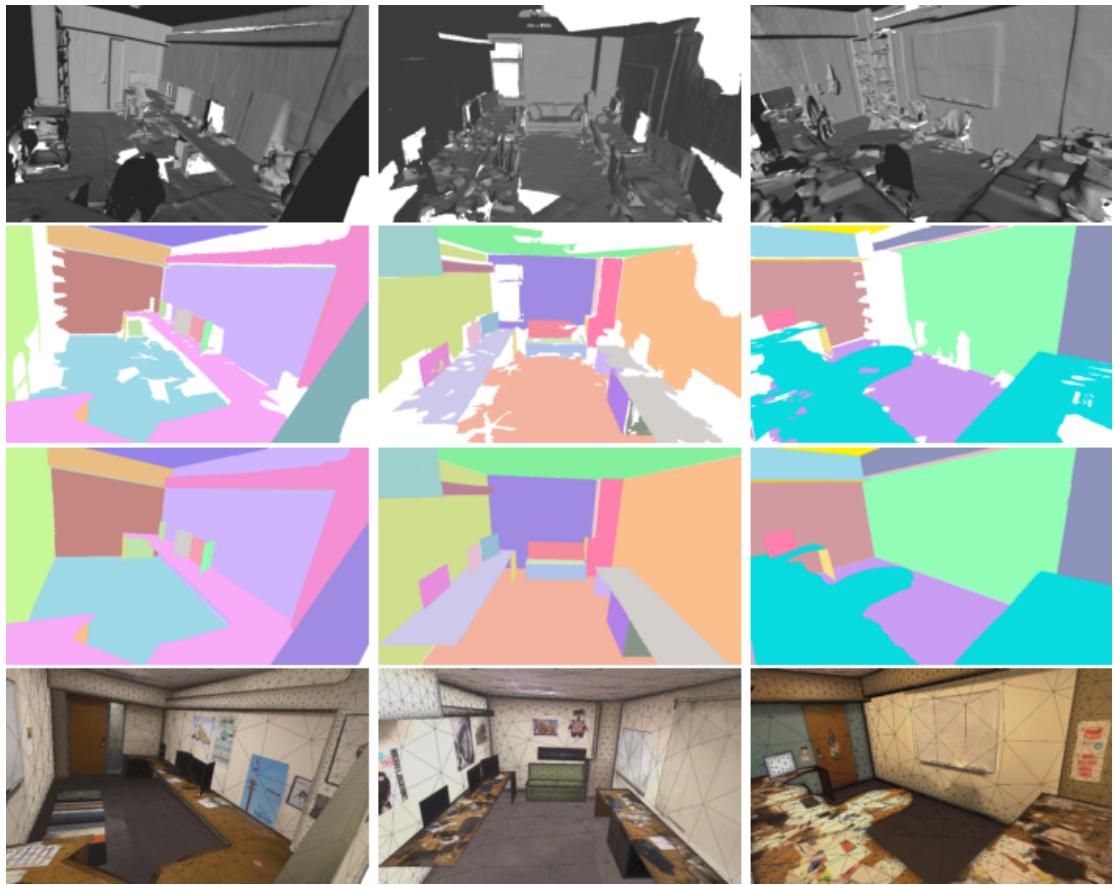


图 15 基于平面的抽象。第一列：原始模型；第二列：平面适配之后；第三列：平面补全之后；最后一列：加上纹理贴图后生成的最终网格模型。

A.9 总结

我们提出了一种新的方法，来产生视觉上真实的三维重建模型，通过解决过度平滑、颜色质量低和模型不完整的问题，向最终能生产出产品级的模型更进一步。3DLite 通过纹理优化和锐化的技术手段，在抽象的几何形状上贴上高质量的纹理，这些纹理甚至比原有的 RGB 图像更加锐利。我们又进一步利用这种基于平面的表示来补全整个场景以及未观测部分的颜色信息。我们相信这是消费者级别的 3D 扫描仪进行内容创造的第一步，而且我们也希望本工作能够为手持 3D 扫描数据能最终用于产品生产线铺平道路。

本文重要参考文献

请注意，在此仅列出本文比较重要的参考文献，其他参考文献请见原文章。



图 16 对比我们的方法和 RANSAC 平面拟合方法：我们的方法在检测较小平面时更加鲁棒。

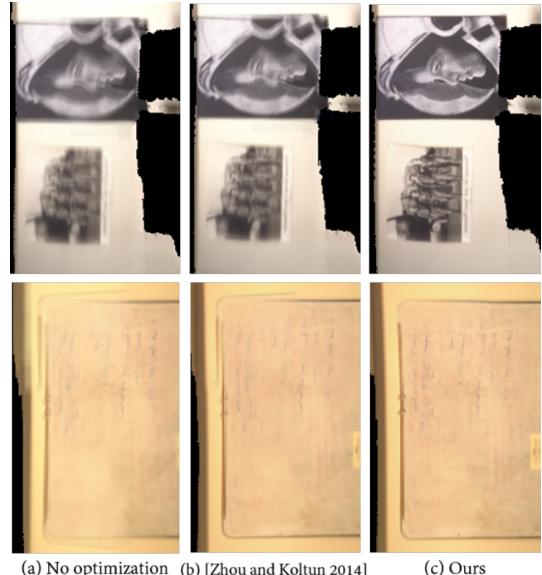


图 17 颜色对齐。(a) 取平均颜色; (b) 仅使用稠密项优化相机姿态; (c) 使用我们的方法优化的结果。

- [1] Darabi, S., Shechtman, E., Barnes, C., Goldman, D. B., & Sen, P. (n.d.). Image Melding: Combining Inconsistent Images using Patch-based Synthesis.
- [2] Zhou, Q.-Y., & Koltun, V. (2014). Color map optimization for 3D reconstruction with consumer depth cameras. ACM Transactions on Graphics, 33(4), 1–10.
- [3] Dzitsiuk, M., Sturm, J., Maier, R., Ma, L., & Cremers, D. (2017). De-noising, stabilizing and completing 3D reconstructions on-the-go using plane priors. In Proceedings - IEEE International Conference on Robotics and Automation (pp. 3976–3983).
- [4] Curless, B., & Levoy, M. (1996). A volumetric method for building complex models from range images. Proceedings of the 23rd Annual Conference on ..., 303–312.
- [5] Izadi, S., Kim, D., Hilliges, O., Molyneaux, D., Newcombe, R., Kohli, P., ...Fitzgibbon, A. (2011). KinectFusion: Real-time 3D Reconstruction and Interaction Using a Moving Depth Camera.

外文资料原文索引

Huang, J., Dai, A., Guibas, L., & Niessner, M. (2017). 3DLite: Towards Commodity 3D Scanning for Content Creation. ACM Transactions on Graphics, 36(6), 1–14. <https://doi.org/10.1145/3130800.3130824>

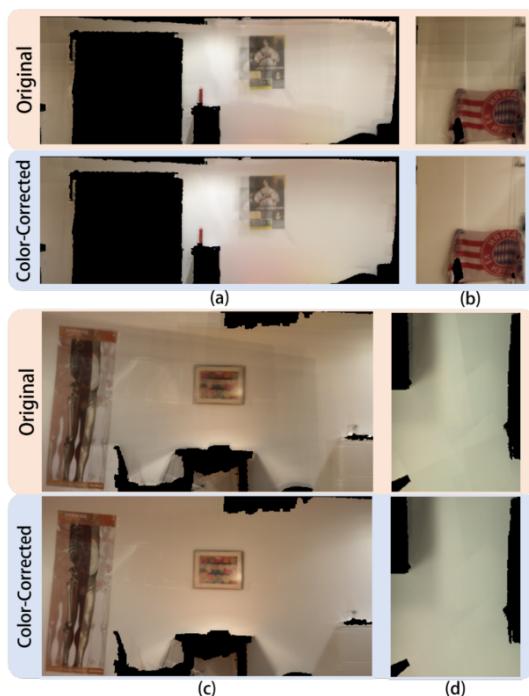


图 18 颜色迁移优化。(a)(b)(d) 为自动曝光和白平衡输入, (c) 为固定白平衡输入; 我们的方法在两种情况下都能修复颜色不一致的问题。

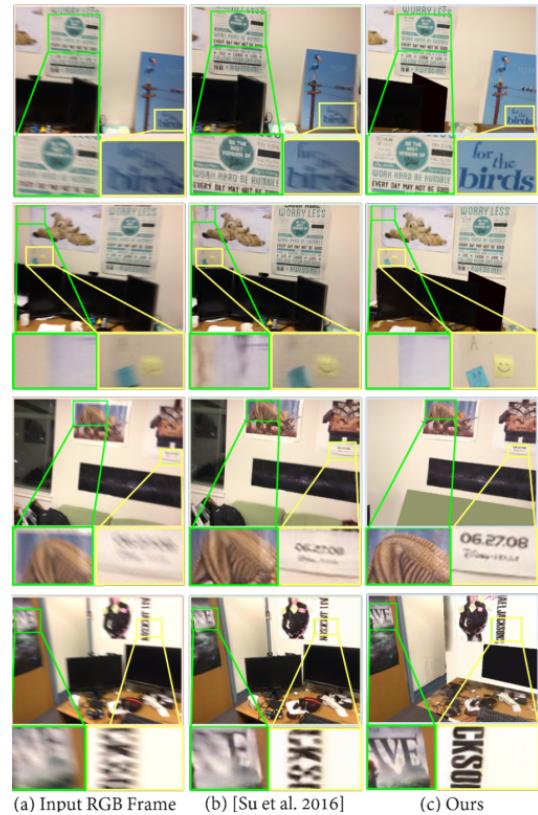


图 19 纹理锐化对比。(a) 大多数 RGB 帧是模糊的; (b)Deep Video Deblurring 能够减小一些模糊; (c) 我们的方法能够得到一致的锐化效果。

在学期间参加课题的研究成果

个人简历

黄家晖，1997年3月8日出生于江苏省徐州市。

2014年9月考入清华大学计算机科学与技术系攻读工学学士学位至今。

参与课题研究成果

- [1] Hu, W., **Huang, J.**, Wang, Z., Wang, P., Yi, K., Wen, Y., ... & Sun, L. (2017, June). MUSA: Wi-Fi AP-assisted video prefetching via Tensor Learning. In Quality of Service (IWQoS), 2017 IEEE/ACM 25th International Symposium on (pp. 1-6). IEEE.
- [2] Ren, B., **Huang, J.**, Lin, M. C., & Hu, S. M. (2018, May). Controllable Dendritic Crystal Simulation Using Orientation Field. In Computer Graphics Forum (Vol. 37, No. 2, pp. 485-495).

综合论文训练记录表

学生姓名	黄家晖	学号	2014011330	班级	计 43
论文题目	基于视觉的三维室内场景建模研究				
主要内容以及进度安排	<p>三维室内场景建模与虚拟现实、室内导航等应用联系密切，在图形学中有重要应用。近几年，学界重点研究了如何通过 RGB-D 相机数据恢复三维场景，并取得了较好的成果。以 Matterport 为首的业界已经提出了不错的解决方案。</p> <p>而本工作则希望与计算机视觉的手段结合，从单张 RGB 图片中挖掘信息，最终输出室内三维 CAD 模型。与传统的 RGBD 三维重建的工作相比，本工作致力于使用模型库中已有的模型组合形成三维场景（而并非使用精确刻画的点云）。另外，由于不需要深度信息，用户无需复杂设备，仅通过手机等普通拍照设备即可轻松完成建模。</p> <p>本研究初步计划将由若干步骤组成，这些步骤包括：①房间几何形状估计；②室内物体识别和三维模型检索；③模型布局优化。</p> <p>具体时间安排如下：</p> <ul style="list-style-type: none"> 1月-2月寒假期间：继续调研文献，对模型数据库进行远程分析； 3月：主要完成室内几何估计和算法设计和代码编写调试； 4月：设计三维模型检索算法，并初步从 SUNCG 数据集中提取模型部件关系； 5月：完善模型检索框架，将已有图像识别模型融入系统，并编写交互式环游界面； 6月：补充上述未完成工作，并撰写结题报告。 				
	<p style="text-align: right;">指导教师签字: <u>胡春民</u></p> <p style="text-align: right;">考核组组长签字: <u>徐军</u></p> <p style="text-align: right;">2018年1月15日</p>				
中期考核意见	<p><u>中期进展顺利。</u></p>				
	<p style="text-align: right;">考核组组长签字: <u>徐军</u></p> <p style="text-align: right;">2018年4月12日</p>				

指导教师评语	<p>三维室内场景建模是计算机图形学领域的热点问题，本文提出一种基于机器视觉的三维室内场景快速建模方法，在效率和效果上均有改进，是一部优秀的本科论文，同意进行答辩。</p> <p>指导教师签字：胡军伟</p>
评阅教师评语	<p>三维室内场景建模是计算机图形学领域的重点问题。论文选题具有理论意义和实用价值。论文仅使用单张图像，利用机器视觉方法自动估计空间几何形状、识别家具位置，可进一步用于场景编辑的三维模型。结果与原图近似，方法新颖。是一篇优秀的本科论文</p> <p>评阅教师签字：王海屏</p>
答辩小组评语	<p>答辩中表述清楚，回答问题正确，论文达到了综合论文训练的要求，是一篇优秀的本科毕业论文。</p> <p>答辩小组组长签字：徐昆</p>

总成绩： 94

教学负责人签字： 张虹海

2018年6月22日