分类号 TP391.4 学号 16060275

U D C 004.92 密级 公开

工学硕士学位论文

RGB 图像的三维重建深度学习方法研究

硕 士 生 姓 名 牛成洁

学 科 专 业 计算机科学与技术

研 究 方 向 计算机图形学

指 导 教 师 熊岳山 教授

协助指导教师 徐凯 副教授

国防科技大学研究生院

二〇一八年十一月

The research of 3D reconstruction from a single RGB image based on deep learning

Candidate: Niu Chengjie

Supervisor: Prof. Xiong Yueshan

Associate Supervisor: Associate Prof. Xu Kai

A dissertation

Submitted in partial fulfillment of the requirements

for the degree of Master of Engineering

in Computer Science and Technology

Graduate School of National University of Defense Technology

Changsha, Hunan, P. R. China

November, 2018

独 创 性 声 明

本人声明所呈交的学位论文是我本人在导师指导下进行的研究工作及取得的研究成果。尽我所知，除文中特别加以标注和致谢的地方外，论文中不包含其他人已经发表和撰写过的研究成果，也不包含为获得国防科技大学或其他教育机构的学位或证书而使用过的材料。与我一同工作的同志对本研究所做的任何贡献均已在论文中作了明确的说明并表示谢意。

学位论文题目：RGB 图像的三维重建深度学习方法研究

学位论文作者签名：_____ 日期：　　年　　月　　日

学位论文版权使用授权书

本人完全了解国防科技大学有关保留、使用学位论文的规定。本人授权国防科技大学可以保留并向国家有关部门或机构送交论文的复印件和电子文档，允许论文被查阅和借阅；可以将学位论文的全部或部分内容编入有关数据库进行检索，可以采用影印、缩印或扫描等复制手段保存、汇编学位论文。

（保密学位论文在解密后适用本授权书。）

学位论文题目：RGB 图像的三维重建深度学习方法研究

学位论文作者签名：_____ 日期：　　年　　月　　日

作者指导教师签名：_____ 日期：　　年　　月　　日

目 录

图 目 录

表 目 录

摘　要

近年来，基于单张单视角 RGB 图像的三维重建研究得到了广泛关注。由于基于 RGB 图像的深度神经网络获得了巨大成功，三维重建的质量也得到大幅提升。目前的深度学习模型大都是重建体素信息来表示三维模型，换言之，这些深度模型是将 RGB 图像映射到三维图像（体素类似于像素）。即使体素重建质量很高，但是却丢失了三维物体一些重要的信息，例如形状拓扑和部件之间的关系等。

为了重建出三维物体的结构信息，从 RGB 图像中还原出更完整更细节性的三维信息，本文提出一种基于 RGB 图像三维重建的深度学习模型，是一个卷积递归自编码器，由结构掩膜网络和结构重建网络两个子网络组成。首先，给定一张具有目标对象的 RGB 图像作为输入，由结构掩膜网络进行 RGB 图像轮廓特征和结构特征提取。然后利用结构重建网络将特征解码，并利用长方体和树状层次结构分别表示三维物体的每一个部件和部件之间的相互关系，包括连接关系和对称关系（即旋转对称关系、平行对称关系以及镜面对称关系），从而实现自动重建 RGB 图像中目标对象的三维结构信息。

其中，结构掩膜子网络的目的是对 RGB 图像进行解析，它是一个多尺度卷积神经网络，通过学习在各种尺度和环境下目标物体的特征，以识别 RGB 图像中目标物体的轮廓信息和结构特征。而结构重建子网络的目的是为了解码 RGB 图像特征，以获得三维物体的结构信息，是一个递归结构解码器。解码器融合结构掩膜网络提取的特征和原始图像的特征，递归地解码长方体的层次结构。由于解码网络可以恢复三维物体各部件之间的连接性和对称关系，因此本文基于 RGB 图像三维重建的深度学习模型可以保证重建的三维物体的合理性和通用性。

本文采用轮廓 - 掩膜和立方体 - 结构训练数据联合对深度学习模型进行训练，同时采取了很多机制防止产生过拟合。通过实验结果可以看出本文的研究取得了非常成功的结果，高质量地从 RGB 图像恢复出了细节性的目标物体的三维部件结构信息。而且在与其他前沿的研究做对比后，充分说明了本文具有很强的创新性和广泛的应用性。本文的研究可以应用到对三维体素重建的补全和优化研究中，利用结构信息将缺失的体素进行对称填充可以产生良好的结果，同时本文的研究也可以应用于高级图像编辑领域，通过对 RGB 图像对应的三维物体进行编辑，将编辑后的效果重新体现在 RGB 图像中。

关键词: **三维重建**; **图像处理**; **深度学习**; **神经网络**

ABSTRACT

The last few years have witnessed a continued interest in 3D reconstruction based on single-view RGB images. The performance of 3D reconstruction has improved dramatically, due to the great success in RGB image-based deep neural networks. However, the outputs of existing learning models are mostly volumetric representations of 3D shapes. In other words, these deep models map RGB images to 3D images(voxels are similar to pixels). Even if the quality of 3D volumetric representation is high, it does lose some important information about the 3D shapes, such as the shape topology and the relations between the parts.

In order to recover the structure of the 3D model with more completed and detailed information, we propose a 3D reconstruction method based on a single-view RGB image, which is a convolutional-recursive auto-encoder consisting of a process of parsing the structure of a RGB image and a process of recovering the cuboid hierarchy. Firstly, given a single RGB image, the structure mask network extracts the contour information and structure features of the interesting object in the RGB image. Secondly, the structure reconstruction network decodes the features and uses the cuboid and tree hierarchy to represent each part and the relations between the parts respectively. The relations include connection relations and symmetry relations (the symmetry relations include connection, symmetry and parallelism, etc.). Finally, the thesis realizes the process of automatically recovering the cuboid representation of the parts of the target object and the part relations.

The purpose of the structure mask network is to parse the RGB image. It is a multi-scale convolutional neural network that can be used to estimate the contour information and structural features of the target object at various scales and environments. The goal of the structural reconstruction network is to decode the RGB image features to obtain the structural information of the 3D shape. The decoder fuses the features extracted by the mask network and the features of the original image, then recursively decodes into the hierarchy structure of the cuboid. Since the decoding network can recover the connectivity and symmetry between the parts of the 3D shape, the deep learning model we proposed can ensure the rationality and versatility of the reconstructed 3D shape.

The deep learning model we proposed is jointly trained by contour-mask and cube-structure training data. During the training stage, several mechanisms are devised to avoid

over-fitting. Through the experimental results of this thesis, we can see that the method we proposed has achieved very successful results. The method of this thesis recovers the detailed 3D shape structure of the target object from the RGB image with very high quality. After comparing with other state-of-the-art work, it fully demonstrates that the method we proposed is full of innovation and value. There are two applications of this thesis including structure-guided completion of 3D volumes recovered from a single RGB image and the structure-aware interactive editing of RGB images.

Key Words: 3D Reconstruction; Image Processing; Deep Learning; Neural Networks

第一章　绪论

三维重建是计算机图形学、计算机视觉领域的一个重要研究内容。三维重建是指对 RGB 图像或者 RGBD 信息进行三维模型重建的过程。从单张 RGB 图像估计三维几何信息是基于若干张 RGB 图像进行三维重建的特殊情况，但是由于不能直接从图像像素估计对应的深度信息，因此相较于基于 RGBD 信息的三维重建，单纯基于 RGB 信息进行三维重建则更具有挑战性。基于 RGB 图像进行三维重建的相关研究成本低廉，自动化程度高，具有广阔的应用前景。本文则利用数据驱动的方法，基于深度学习，实现了从单张 RGB 图像进行三维重建，获取对应的三维物体的结构信息的过程，本文的研究方法在三维重建这个领域具有重要意义。

1.1　研究背景

基于单视角 RGB 图像的三维重建是计算机图形学领域一个重要的研究内容，它通过输入一张 RGB 图像，获得一个具有三维信息的与之对应的物体，为机器人视觉、自动驾驶和增强现实等诸多研究领域提供了夯实的基础和强大的支撑。

1.1.1　三维物体的表示方式

为了描述物体的三维空间信息，三维物体的表示方式多种多样，包括点云 (Point Cloud)、体素 (Volume)、多边形网格 (Polygonal Mesh)、多视角深度图 (Multi-View RGB(D)) 以及有向包围盒 (Oriented Bounding Box) 等，如图 1.1所示。各种表示方式都有各自的优缺点，而本文是基于有向包围盒的方式来表示重建物体的三维信息。

(a)点云　(b) 体素　(c)网格　(d) 多视角RGB(D)图像　(e)有向包围盒

图 1.1　三维物体的表示方式

1.1.1.1　点云 (Point Cloud)

随着三维采集设备的不断发展，获取三维数据的方式越来越多，而大多数采集设备获取的三维数据是以点云的形式进行存储的。点云是分布在空间中的坐标点的集合，有的还带有颜色、法线等信息。虽然点云的获取方式简单快捷，但是

它在数据处理方面面临诸多挑战。首先是点云的无序性，一个点云可以由多个完全不同的矩阵来表示，在几何上，点的顺序不影响它在底层矩阵结构中的表示方式。其次是扫描的模型通常被遮挡，部分数据有丢失，加之传感器的嘈杂性，这就导致一个点有一定的概率应该位于它被采样的地方附近的某一半径范围内，或者它可能出现在空间中的任意位置[1]。这就都导致了在数据方面对点云的处理有一定的困难性，而且点云是非结构化的数据，很难用神经网络进行处理。

1.1.1.2 体素 (Volume)

体素是一个体积元素，类似于像素是一个图像元素，体素是三维物体表示的基本单元。由于体素的表示简单、明了，用一个规则的矩阵就可以获得三维物体的信息，从而可以用神经网络进行学习处理，因此在基于深度学习的三维重建方面得到了广泛的应用，大多数点云数据也是转化为规则的体素数据进行神经网络的学习。然而，体素的表示是离散化的，当重建一个三维物体后，体素表示不能很好的恢复该物体的结构信息以及一些细粒度的信息。因此并不能总是获得研究者想要的预期效果。

1.1.1.3 多边形网格 (Polygonal Mesh)

多边形网格是一个多边形列表，可以用来模拟三维物体的表面信息。而多边形网格中存储的三维点一般是顶点或者相连的点，通常都包含三维坐标以及法线，根据不同的需求，有的还需要保存纹理和颜色等相关信息[2]，而三角形网格在传统的图形分析与处理方面使用最为广泛。但是多边形网格的不规则性限制了其在深度学习领域的快速发展。

1.1.1.4 多视角深度图 (Multi-View RGB(D))

多视角深度图在基于深度学习的三维重建方面应用广泛。多视角深度图是由多个视角的深度图共同组成用来描述一个三维物体的各个角度。随着基于二维图像的神经网络的发展，用多视角深度图来描述三维物体，可以很好的适应神经网络的学习训练。

1.1.1.5 有向包围盒 (Oriented Bounding Box)

本文所采用的有向包围盒是利用一些有向的立方体来表示三维物体的各个部件，同时存储三维物体的对称信息、连接信息以及旋转信息等。相较于体素的表示方式，在对物体进行重建后，有向包围盒可以完整恢复出三维物体的对称信息，不会造成单个部件的缺失。而且由于各个部件都是单一的立方体，可以很好地应用于三维物体的分割和合成更加多样化的三维物体。

1.1.2 三维重建方法概述

经典的三维建模方法自计算机图形学学科的开端就开始研究，三维重建主要是通过深度数据获取、数据预处理、配准与融合点云、最后进行表面生成四个过程，把真实场景转换为数学模型方便计算机读取和展示。而三维重建的重点在于如何获取目标物体或者场景的深度信息。在深度信息已知的情况下，再对点云数据进行配准与融合就可以恢复出物体的三维形状。而获取物体深度信息的方法又可以依据测量方式的不同，可以分为主动式重建和被动式重建，如图 1.2所示。

主动式是指利用光源或能量源发射至目标物体，如声波、电磁波、激光等，通过接收返回的光波的时间和光波的速度，通过计算路程等于时间乘速度来获取物体的深度信息。主动式测量方法有四种方法，如图 1.2所示。常见的三维数据采集设备也大都是主动式测量。被动式则主要是利用环境中的自然光反射等因素，利用相机获取图像，然后再通过使用一些算法来计算该物体的三维信息，主要有纹理恢复形状法、阴影恢复形状法和立体世界法等三种方法，如图 1.2所示。被动式测量法在实际使用中较少，主要是由于需要假设的条件与实际空间相差较远，同时运算量较大。

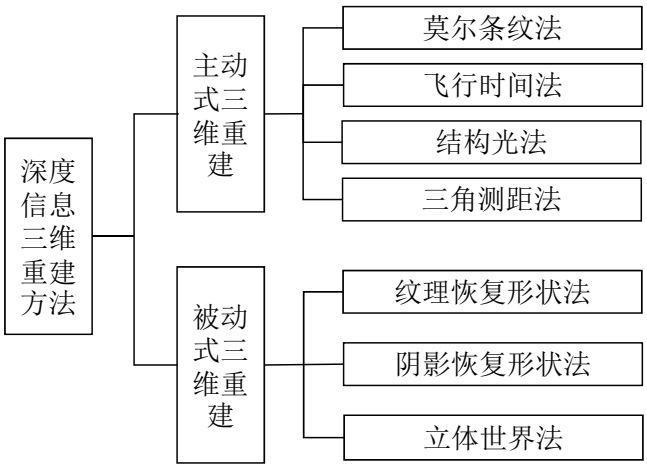

图 1.2 深度信息三维重建方法分类

随着科学技术的飞速发展，计算机视觉和计算机图形学领域的研究人员开始进行更加具有挑战性的研究，深入思考研究从二维图像恢复三维结构的方法。其中最为经典方法是 Marr 时间信息处理过程，他将该过程概括为图像获取，2.5 维结构、形状与位置的恢复，三维物体的识别、分析、理解与描述这六个步骤，如图 1.3所示。

深度学习技术的飞速发展，也为三维重建领域注入了新的血液。图形学领域的部分研究者开始尝试使用深度学习的技术来解决三维重建过程中的难题，并且

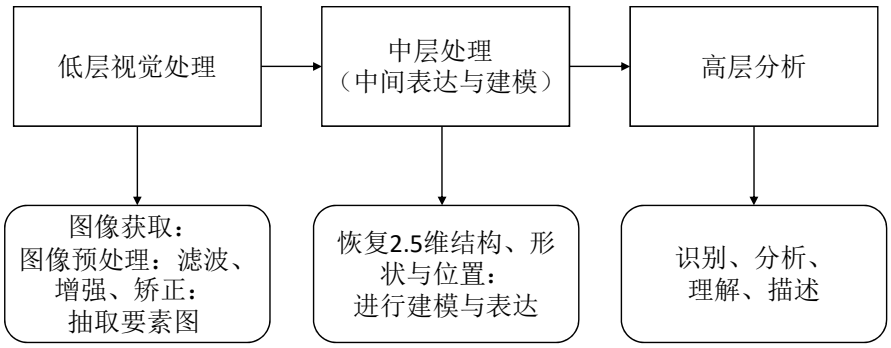

图 1.3　Marr 视觉信息处理过程

取得了很好的效果。本文也是采用深度学习的技术，利用数据驱动的方式实现了 RGB 图像的三维重建，并且取得了非常好的效果。利用深度学习的方法进行三维重建主要分为四个步骤：数据准备及预处理、深度模型的设计与搭建、训练深度模型、重建结果的优化与改进，如图 1.4所示，其中每个步骤都需要根据具体的三维重建内容进行针对性的思考与设计。

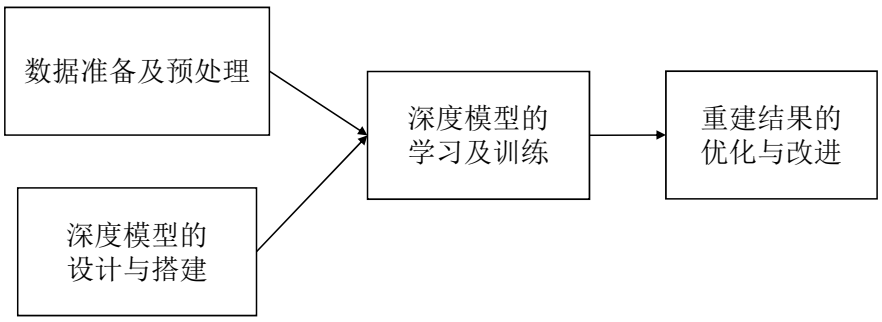

图 1.4　基于深度学习的三维重建步骤

1.1.3　基于 RGB 图像的深度学习方法

传统的机器学习技术有很多限制，往往需要采用原始的形式来处理自然数据，导致机器学习的模型不能拥有强大的学习能力，而深度学习则通过对大量数据的学习，可以具有自动提取特征的能力。换言之，深度学习可以自动发现需要从原始数据中检测的特征。深度学习方法拥有多个学习层次，每个层次都可以利用从上一层次学习到的特征变换为更加抽象的特征，即将底层更具体的特征转换为高层次更抽象的特征。所以，即使是非常复杂的数据，只要深度学习拥有足够多的层次，它也可以自动学习。

深度学习最广泛的应用是对 RGB 图像进行特征提取，实现分类、识别、生成图像等诸多任务。而本文则采用深度学习的方法，自动剔除 RGB 图像中不相关的

部分，比如背景的颜色、目标物体的位置等信息，同时将 RGB 图像中目标物体的三维信息自动进行提取，比如目标物体的对称性、各个小部件的连接信息等。

广泛用于 RGB 图像的神经网络有卷积神经网络 (Convolution Neural Networks, CNN)、递归神经网络 (Recurrent Neural Networks, RNN)、深度信念网络 (Deep Belief Networks, DBN) 和生成对抗网络 (Generative Adversarial Networks, GAN) 等类型，并且都取得很好的效果。本文则采用了一种 CNN 从 RGB 图像中提取目标物体，去除背景，并且在不同层次之间加入了连接层，可以在获得整体信息的同时，保留目标物体的细节信息，同时还在与之并行的另一条线采用 CNN 进行目标物体进行处理，进行目标物体的特征提取，最后再将两条 CNN 线结合一起，用全连接层进行信息融合，以获得更加清晰的目标物体的特征，从而用于对其进行三维模型重建。

1.2 国内外研究现状

从单视角 RGB 图像进行三维模型重建的工作一直是计算机视觉和图形学领域的热点话题。由于它的不适定性，之前的研究大都需要一定的假设和先验知识，直到大容量的深度神经网络的出现和大规模应用才可以不用进行假设，独立进行三维重建。本文则重点关注基于深度学习的模型，并且将近年来相关的一些文献从三个维度进行分类。

1.2.1 基于深度估计的三维重建

深度估计可能是从单个图像恢复三维信息的最直接的解决方案。深度学习已被证明对深度估计非常有效[3, 4]。与深度估计相比，由于需要对三维模型看不见的部分进行推理，重建完整的三维模型更具挑战性。后者必须采用形状或结构的先验知识。在快速增长的大型三维模型库的背景下[5]，使用深度神经网络对特定类别模型的先验知识进行编码得到越来越多的关注。Choy 等研究人员[6] 在单个图像作为输入的情况下，设计了三维递归重建神经网络，并且以体素表示生成三维模型。Fan 等研究者[7] 提出了一种点集生成网络，用于从一张 2D 图像生成点云表示的三维模型。目前为止，还没有任何直接从单个图像生成基于模型部件且感知模型结构的相关研究成果。

1.2.2 生成对抗式三维重建

对于从 RGB 图像进行三维建模的任务，辨别模型通过学习一个从输入 RGB 图像到 3D 表示的映射。辨别模型一般是通过一个深度卷积神经网络进行一次生成操作，或者使用递归神经网络模型进行渐进循环生成操作[6]。辨别模型进行映射操作的主要优势是易于训练，而且能够产生高质量的结果。近期深度生成神经

网络迅猛发展，出现了变分自动编码器 (Variational Auto-Encoder, VAE)[8]，生成对抗网络 (Generative Adversarial Nets, GAN)[9] 以及它们的变种。学习三维形状生成的生成模型已经获得了广泛的研究[10–13]。对于生成模型，输入的图像可以用于调整在预定义的参数空间或者学到的嵌入空间进行的采样操作[11, 14]。众所周知，生成模型很难训练。对于跨模态映射的任务，我们选择使用中等大小的数据集去训练判别模型。

1.2.3 几何结构信息的恢复重建

现有的基于深度学习模型的研究大都是利用体素对三维模型进行表示的[11, 12]。当然也有一些值得关注的工作不是利用体素表示，包括使用点云[7]、立方体基元[15]、流形表面[16] 等方式来表示生成的三维模型。然而，这些表示方式都不包含三维模型各个部件之间的连接关系和部件的结构信息。相反，之前做结构恢复的相关研究工作都没有使用深度学习的方法进行研究，主要原因在于很难找到合适的深度神经网络对三维模型的结构进行表示。近期，Li 等研究者[10] 提出了使用递归神经网络 (Recursive Neural Networks, RvNN) 用于对三维模型结构表示的学习训练，很好地解决了对任意数目的三维模型部件以及多种多样的部件之间的关系进行编码和解码的操作。本文则很好地利用了这个优点，并将其集成到跨模态映射架构中，以实现从单张 RGB 图像恢复出目标物体的结构信息。

1.3 本文工作

1.3.1 研究内容

利用深度学习的方法，对单一的 RGB 图像进行三维结构重建是计算机视觉以及计算机图形学领域的热点问题，该研究为增强现实、自动导航、机器人等领域中的视觉研究工作提供了支持，也为图形学领域进行结构感知的研究提供了新思路，新想法。本文基于 RGB 图像的三维结构感知的研究现状，结合国内外先进的研究内容，抽丝剥茧，从基础的对 RGB 图像进行背景剔除，获取目标物体开始，层层递进，用深度学习的方法设计合适的神经网络提取 RGB 图像中目标物体的特征信息，为将该特征恢复成三维结构做铺垫。于此同时，研究三维结构感知模型的重建方法，设计合适的深度学习模型和三维结构的表示方式恢复三维物体各个部件之间的结构信息和连接关系。最后，在确保能够获得准确的 RGB 图像中目标物体的特征信息后，结合能够恢复三维结构信息的深度学习模型，搭建一个完整的神经网络框架，实现输入一张 RGB 图像，映射到该图像中对应目标物体的三维信息，尤其是三维结构信息以及三维物体中各个部件的连接关系等。

本文的研究工作主要从以下几个方面展开：

1.3.1.1 RGB 图像特征提取

本文对 RGB 图像进行特征提取的深度学习模型称为结构掩膜网络，该网络为多尺度卷积神经网络，采用两条线并行，一条线是为了剔除 RGB 图像中的背景信息以获得目标物体轮廓特征。为了能够在复杂背景中找到目标物体，获得准确的重建结果，本文首先设计了一个 RGB 图像背景剔除网络，可以得到目标物体的轮廓信息以及目标物体在图片中的像素位置，本文规定在每个像素的位置用二值进行表示，可以突出目标物体，为进行 RGB 图像的特征提取提供基础。另一条线采用卷积神经网络直接对 RGB 图像进行卷积操作，然后连接两条线，将所提取的特征进行结合，再进行卷积和全连接操作，从而获得 RGB 图像中目标物体的特征信息。

1.3.1.2 三维结构感知重建

三维结构感知重建能够恢复出三维物体的结构信息以及各部件直接的关系。该部分的顺利实现主要包括具有结构信息的三维模型表示方式（即有向包围盒）以及针对由有向包围盒组成的三维模型设计的结构恢复网络。

在三维模型表示方式方面，本文采用了一种树状结构，该树状结构的每一个叶节点代表三维模型的一个部件（如一个椅子的一条椅腿），而其他节点则代表有一定关系的部件组合（如一个椅子中具有对称关系的两个扶手），每个叶节点由一个有向包围盒表示，所以叶节点都绘制出来就是一个由有向包围盒组成的三维模型。这种表示方式创新了传统的三维模型表示，可以更好地保留三维模型中各个部件的结构信息和连接信息等关系。

在结构重建网络方面，由于本文用一种树状结构来表示三维模型，所以在网络的设计方面，采用了一种类似于对树状结构进行处理的网络，即递归神经网络(Recursive Neural Networks, RvNN)。依据每个节点所包含的各部件直接的信息对其进行解码。

结构重建网络有针对性地对有向包围盒进行解码，从而可以重建结构感知的三维模型。

1.3.1.3 RGB 图像的三维结构重建

结合结构掩膜网络和三维结构重建网络，将其作为一个整体共同进行训练，实现单一 RGB 图像映射到由有向包围盒表示的三维模型，并且对该网络的性能进行了评估，将本文的重建结果与其他领先的工作进行比较，最后对本文的工作进行了总结和展望。

1.3.2 本文贡献

本文对基于 RGB 图像的三维重建研究主要贡献如下:

1. 本文提出一个新颖结构掩膜网络，包含一个 RGB 图像背景剔除网络和一个 RGB 图像特征提取网络，前者以二值矩阵的方式进行存储，将目标物体的轮廓与背景区别开，为 RGB 图像特征提取排除背景干扰。
   该部分内容对应本文第三章。
2. 本文采用一种利用树状结构表示三维模型的方式，该树状结构的树枝代表各个三维部件之间的关系，每个叶节点代表不同的三维部件。该树状结构很好地保持了三维模型各个部件的结构信息和连接关系等。而该树状结构的每个叶节点使用有向包围盒进行表示，三维模型的信息清晰明确，易于分析理解。
   该部分内容对应本文第四章。
3. 本文采用一种递归神经网络 (Recursive Neural Networks, RvNN) 对结构感知的三维模型进行解码。递归神经网络可以理解为一种树状结构的神经网络，和本文所采用的用树状结构表示三维模型有异曲同工之处，实现了对结构感知三维模型的解码重建。
   该部分内容对应本文第四章。
4. 本文提出一种从 RGB 图像直接恢复三维模型形状结构信息的网络架构。本文的工作创新性很强，之前还没有研究者使用深度学习的方法来进行从 RGB 图像到三维模型结构信息重建的相关工作。同时本文提出的网络架构整合了卷积结构掩膜网络以及递归结构重建网络，具有创新性，并且取得了非常好的效果。
   该部分内容对应本文第五章。

1.4 论文组织结构

本文共分为六章，各章节之间的关系如图 1.5所示。

第一章，绪论。介绍本文研究课题基于 RGB 图像的三维重建的研究背景、国内外相关研究的现状，并且对本文的工作进行简单介绍。

第二章，RGB 图像特征提取相关方法。分析 RGB 图像特征的定义以及特征提取的相关方法。并就常见的用于进行特征提取的卷积神经网络进行介绍，说明本文进行 RGB 图像特征提取的相关方法。

第三章，RGB 图像目标对象掩膜提取。分析目标物体提取的相关方法，深入理解利用深度学习进行 RGB 图像背景去除的相关研究。结合本文提出的背景去除，提取目标物体的方法，详细描述本文提出的 RGB 图像卷积结构掩膜网络，并

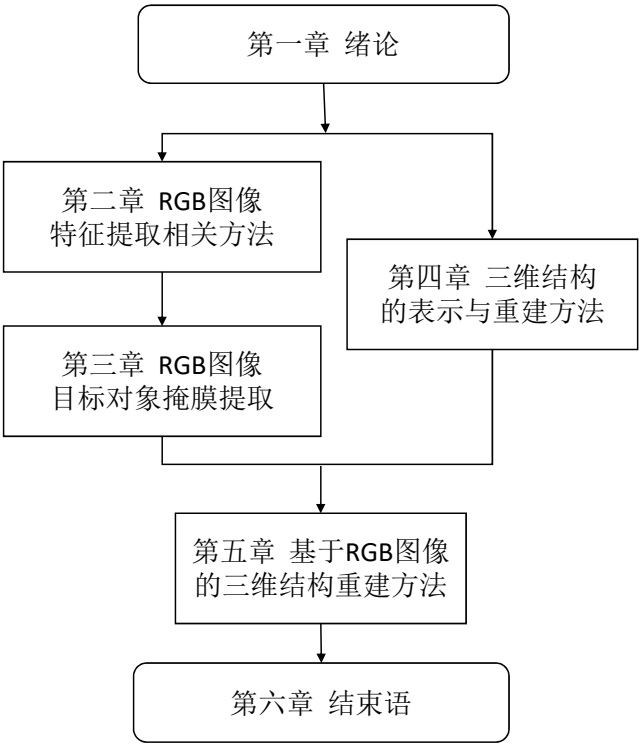

图 1.5　论文组织结构图

且对该网络的实验效果进行展示分析。

第四章，三维结构的表示与重建方法。基于目前三维重建的相关研究以及对三维物体结构的相关表示方法，详细介绍本文提出的利用树状结构以及有向包围盒表示三维模型的结构，并且详细阐述本文提出的递归结构重建网络，以及其在每个层次的工作原理。

第五章，基于 RGB 图像的三维结构重建方法。提出基于 RGB 图像重建三维结构的深度学习模型，结合卷积结构掩膜网络和递归结构重建网络，实现 RGB 图像到三维结构的映射。充分对实验结果进行展示和评估，并且呈现与其他先进的实验进行比较的结果。

第六章，结束语。对硕士期间的研究工作进行总结，并对下一阶段的工作进行展望。

1.5　小结

本章介绍了 RGB 图像三维重建深度学习研究的背景以及国内外相关研究的现状，概述了本文的重点研究内容以及主要贡献，最后展示了论文的组织结构。本章是对全文的概括性介绍。

第二章　RGB 图像特征提取相关方法

特征提取在图像处理领域起着非常重要的作用。在获取图像特征之前，先要对采样图像进行预处理，比如二值化、阈值化、调整大小以及归一化等操作。图像特征提取技术可以获得对图像进行分类和识别等有用的特征，可以应用于各种各样的图形处理研究，比如字符识别、手势识别以及人脸识别等。本章则重点介绍通过获取 RGB 图像的相关方法，即从原始图像中获得尽可能多的相关信息，并以低维的向量进行表示。

2.1　RGB 图像特征的定义和属性

RGB 图像的三维重建，必然少不了对 RGB 图像特征提取的部分，通过提取到的 RGB 图像的结构特征，再进行操作将该特征进行解码，以三维结构信息进行展示，从而实现三维重建。然而图像特征有多种表示方法和分类方法，本节则重点介绍图像的全局特征和局部特征，并就图像特征所具有的一些性质进行分析。

2.1.1　RGB 图像特征的定义

在图像处理和计算机视觉领域，我们需要用图像特征去代表图像。人类眼睛可以从原始图片中直接获取所有信息，然而用计算机算法是不能实现的。一般来说，存在两种方法来表示图像，即全局特征和局部特征[17–19]。如图 2.1所示，其中 (a) 表示全局特征，(b) 表示局部特征。在全局特征表示方面，图像使用一个多维的特征向量来描述整个图片的信息。换言之，全局特征表示法可以生成单个向量，其值可以用来测量图像的各个方面，例如颜色、纹理或者形状。实际上，提取来自每张图像的单个向量，然后通过比较它们的特征向量来比较两个图像的差别。比如，当我们想要区别大海的图像（蓝色）和森林的图像（绿色），一个基于颜色信息的全局特征能够为每个种类生成各种不同的向量。在本章中，全将特征可以理解为包含所有像素的图片的特定属性。这个属性可以是颜色直方图、纹理、边缘，甚至是从应用于图像的某些过滤器中提取的特征描述符。另一方面，局部特征表示的主要目的是基于一些显著区域区别地对图像进行表示，同时保持视点和光照变化的不变性。因此，通过从感兴趣的区域（即关键点）的一组图像区域提取的一组局部特征描述符，基于其局部结构来表示图像，如图 2.1中 (b) 所示。大多数局部特征表示图像块内的纹理。

通常，具体使用哪种特征很大程度上取决于具体的应用任务。而开发人员更喜欢选择具有辨别力的特征。比如，一个人是大鼻子小眼睛，而另一个人是小鼻子大眼睛，但是他们在直方图或者强度分别方面具有相似的表示，这个时候局部

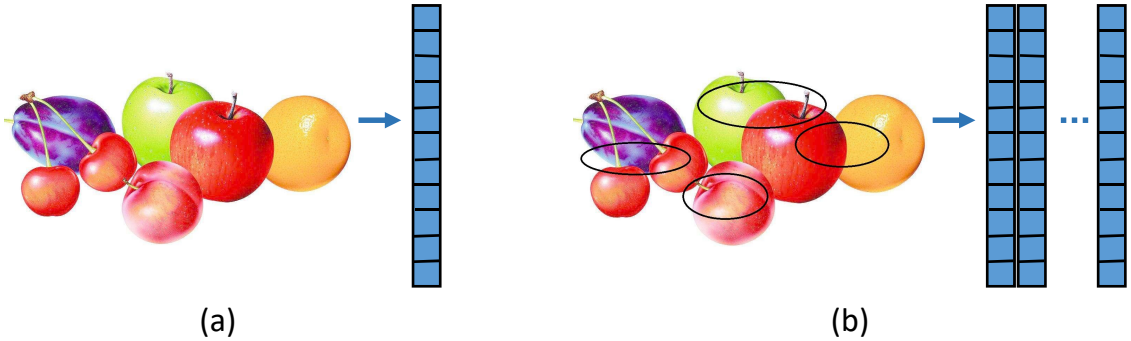

图 2.1　全局特征与局部特征

特征或者从局部特征簇中提取的全局特征更具有辨别力和区分性。此外，全局特征在对目标对象进行粗略分割方面是十分有用的。全局对象的优点除了易于计算，速度快，占用内存小之外，它的表示也更加紧凑简洁。然而，众所周知全局特征有一些局限性，特别是它会随着一些明显的变换而改变，而且对杂波和遮挡都十分敏感。在某些应用方面，比如拷贝检测，大多数非法拷贝和原件非常相似，它们仅仅进行了压缩、缩放或者有限的裁剪等相关操作。相对而言，局部特征的优势就显现出来，局部特征具有卓越的性能[20]，同时，相对于全局特征，局部特征对大规模图片搜索表现出更优的性能[21]。此外，由于局部特征比光滑区域的其他结构的更具有可区别性和稳定性，因此，在图像匹配和目标识别方面，局部特征更是首选特征。但是局部特征通常需要更大的内存空间，因为一张图像可能具有数百个局部特征。作为这个问题的解决方案，研究人员建议将局部图像特征描述符集合成为一个非常紧凑的向量表示，并优化这个向量维度，使用尽可能少维度的向量来表示这些局部特征。而本文则是通过结构掩膜网络将 RGB 图像中的背景去除，然后通过神经网络的方法将目标物体的全局特征进行表示。

2.1.2　RGB 图像特征的属性

Tuytelaars 和 Mikolajczyk[22] 将局部特征定义为“它是一种与其邻近区域不同的图像模式”。因此，他们认为局部不变特征的目的是为了提供一个特征表示方式，可以有效地匹配图像之间的局部结构。换言之，我们希望获得一组稀疏的局部测量值，这些测量值能够获得输入图像的潜在特征，并且能对其目标物体进行编码。为了实现这一目标，特征检测器和提取器必须具有某些属性，当然这些属性的重要性取决于实际的应用任务设置和需要作出的一些妥协。以下属性对于在计算机视觉应用中使用特征检测器非常重要：

- **鲁棒性**。特征检测算法应该能够检测到相同的特征位置，而不会受一些无关因素影响，比如缩放、旋转、平移、光照变化、图像压缩以及噪声等。

- **可重复性**。特征检测算法应该能够在各种变换视角下，重复对同一个场景或者物体检测出相同的特征。
- **精确性**。特征检测算法应准确定位图像特征（相同像素位置），尤其是图像匹配任务，其中需要精确对应来估计极线几何。
- **通用性**。特征检测算法应该能够在不同的应用场景下进行特征检测。
- **效率**。特征检测算法应该能够快速地应用到新图像中以支持实时应用的需求。
- **数量**。特征检测算法应该能够检测出图像中全部或者大多数特征。其中，检测到的特征密度应该能够反映图像中的信息内容，目的是提供一个紧凑的图像表示方式。

2.2 RGB 图像特征提取方法

图像特征提取的过程[23, 24]，其实就是一个维度不断递减的过程，将原图像数据用一个低维度的矩阵或向量来表示特征，但是依旧可以准确和全面地描述原始图像数据中的信息。当输入数据对一个算法而言过大以至于难以处理的话，这个数据就会被认为有大量的冗余信息，比如表示图像信息的像素。然而它们可以被表示为低维的特征或者称为特征向量[25]。其中决定初始图像中关键信息的数据或特征称为特征选择，这部分选择的特征会包含与初始图片相关的信息，所以用这个低维的特征来进行图片处理的一些操作完全可以替换使用输入图片进行。

图形特征提取方法中，使用广泛的几个方法是尺度不变特征变换 (Scale-Invariant Feature Transform, SIFT)、SIFT 方法的变形加速稳健特征 (Speeded-Up Robust Features, SURF)、梯度位置方向直方图 (Gradient Location-Orientation Histogram, GLOH) 以及方向梯度直方图 (Histogram of Oriented Gradient, HOG)。本节重点介绍这四种经典的特征提取方法。

2.2.1 尺度不变特征变换

尺度不变特征变换 (Scale-Invariant Feature Transform, SIFT) 是特征提取方面一个经典的算法[26]。SIFT 通过在尺度空间中寻找图像中的极值点，提取出其旋转不变量、位置、尺度等重要信息，不受目标物体在 RGB 图像中的颜色、位置、方向等变化的影响。SIFT 提取的信息是一种非常稳定的局部特征，比如亮区的暗点和暗区、边缘点、角点等特征信息，可以保持对旋转缩放、亮度变化的不变性。因此在图像处理方面具有非常广泛的应用[27]。

SIFT 进行特征提取一共有四个主要步骤:

1. 在多尺度空间中，检测极值点，即对关键点的探查。该过程主要利用高斯微分函数对在不同尺度中的稳定关键点进行识别。
2. 对稳定关键点的精确定位。主要通过利用一个拟合精细的模型来确定稳定关键点的位置和尺度。
3. 对稳定关键点的方向信息分配进行计算，确定其主方向，实现旋转不变性。由于图像局部具有不同的梯度方向，因此给每个关键点位置会分配一个或者多个方向。
4. 对稳定关键点进行描述。主要将关键点周围领域内图像局部的梯度变换为一种表示，该表示允许较大的局部变化包括形变和光照变化。

SIFT 算法找到不同尺度空间中关键点后，然后对特征点进行对比匹配，就可以找到多幅图像中同一个目标物体的特征和所在的位置，具体流程如图 2.2所示。

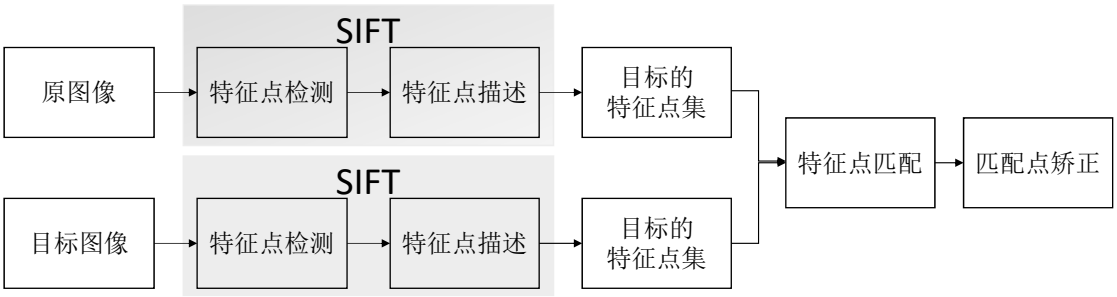

图 2.2　SIFT 算法提取特征

2.2.2　加速稳健特征

Bay 等人提出了加速稳健特征 (Speeded-Up Robust Features,SURF)[28] 检测器—描述符方法，被描述为 SIFT 算法的高效替代品。相对于 SIFT，SURF 更加高效和鲁棒。在目标点的检测阶段，SURF 计算是基于简单的二维盒滤波器，而不是依赖于理想的高斯导数，其中它是使用 Hessian 矩阵行列式的尺度不变斑点检测器来进行尺度选择和定位[29]。

与 SIFT 算法类似，SURF 算法也是先后经过局部特征点提取，特征点描述以及特征点匹配这三步。它的基本思想是在使用一组盒式滤波器的积分图像的帮助下，以有效的方式近似二阶高斯导数。图 2.3中描绘的 9×9 盒式滤波器是具有 $\sigma = 1.2$ 的高斯近似值，并且表示用于计算斑点响应图的最低比例。这些近似值由 D_{xx}，D_{yy} 和 D_{xy} 表示。因此，Hessian 的近似行列式可以表示为公式 2.1。

$$det(H_{approx}) = D_{xx}D_{yy} - (xD_{xy})^2 \tag{2.1}$$

其中，w 是过滤器响应的相对权重，它用于平衡 Hessian 行列式的表达式。Hessian 的近似行列式表示图像中的斑点响应。这些响应存储在斑点响应图中，并使用二次插值检测和细化局部最大值，与 DoG 做法类似。最后，在 3×3 领域中进行非最大值抑制以获得稳定的兴趣点和值的规模。

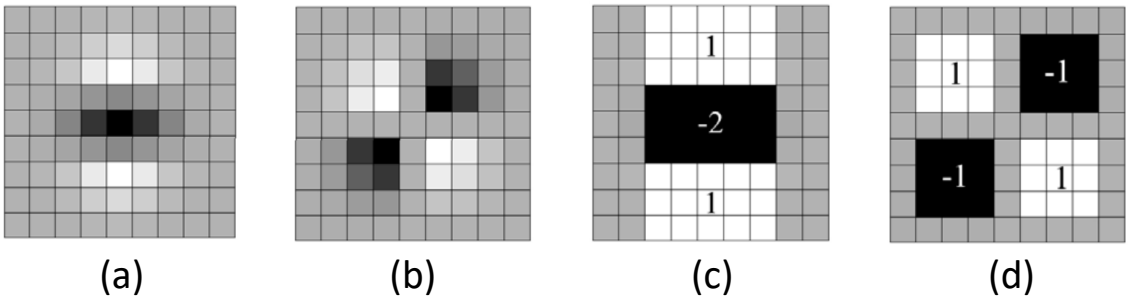

图 2.3　近似二阶高斯导数的过程[28]

SURF 描述子首先构造一个以检测到的兴趣点为中心的方形区域，并沿其主方向定向。这个窗口大小为 $20s$，其中 s 是检测兴趣点的比例。然后，将感兴趣的区域进一步划分为 4×4 的子区域，并且对于每个子区域，在 5×5 的采样点处计算垂直和水平方向上的 Harr 小波响应（分别表示为 d_x 和 d_y），如图 2.4所示。这些响应在以兴趣点为中心的的高斯窗口被加权，以增加对几何变形和定位误差的鲁棒性。小波响应 $d_x d_y$ 对每个子区域求和并作为输入传入特征向量 v 中，如公式 2.2所示。

$$v = (\sum d_x, \sum |d_x|, \sum d_y, \sum |d_y|) \tag{2.2}$$

其中，为所有 4×4 的子区域进行计算，产生一个长度为 $4 \times 4 \times 4$ 维的特征描述子。最后，这个特征描述子被规范化为一个单位向量以减少光照影响。

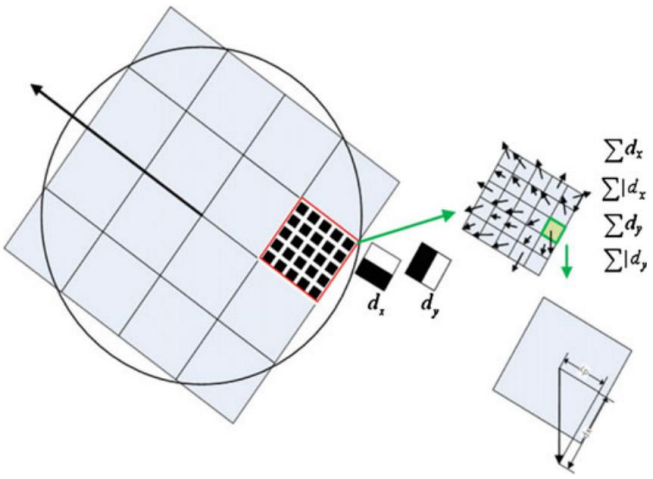

图 2.4　将感兴趣的区域划分为 4×4 的子区域进行 SURT 描述子的计算[28]

相比于 SIFT 描述子，SURF 的优点主要是处理速度快[30]，因为它使用 64 维的特征向量来描述局部特征，而 SIFT 是使用的 128 维。然而，SIFT 描述子更适合描述受平移、旋转、缩放或者其他照明变化影响的图像。尽管 SURF 在各种计算机视觉应用中显示出了巨大的潜力，但是它也存在一些缺点，比如在比较二维或三维对象时，当旋转角度过大或者视角进行变化，SURF 都会失败，同时 SURF 并不是完全仿射不变的[31]。

2.2.3 梯度位置方向直方图

由 Mikolajczyk 和 Schmid[32] 提出的梯度位置方向直方图 (Gradient Location-Orientation Histogram, GLOH) 也是 SIFT 描述符的扩展。GLOH 与 SIFT 描述子非常相似，它仅将 SIFT 使用的笛卡尔定位网格替换为对数极坐标网格，并应用 PCA 来减小描述子的大小。GLOH 使用一个对数极坐标定位网格，其径向有 3 个区间（半径设置为 6,11 和 15），角度方向为 8 个，产生 17 个位置区，如图 2.5所示。GLOH 描述符构建一组直方图使用 16 个区间中的梯度方向，为每个兴趣点产生 $17 \times 16 = 272$ 个元素的特征向量。通过计算 PCA 的协方差矩阵将 272 维描述符简化为 128 维描述符，并且选择最高 128 个特征向量用于描述。根据 [32] 中进行的实验评估，已经证明了 GLOH 优于原始 SIFT 描述子并且提供更优的性能，尤其是在光照条件变化的情况下下。此外，它已被证明比其对应的 SIFT 更有特色但也更昂贵[33]。

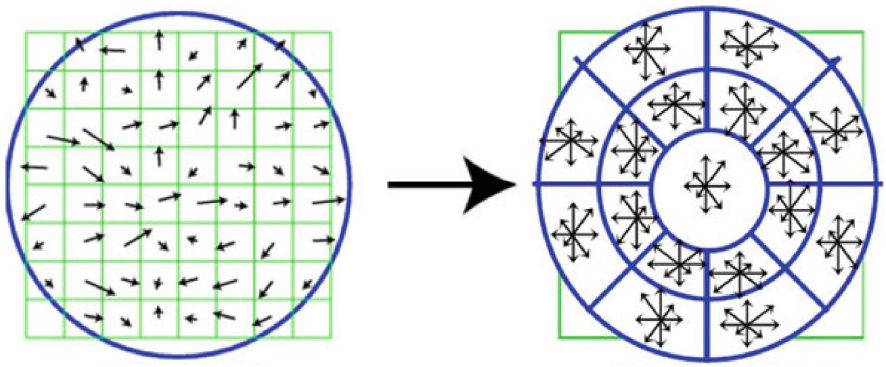

图 2.5　使用对数极坐标定位网格的 GLOH 算法示意图[32]

2.2.4 方向梯度直方图

方向梯度直方图 (Histogram of Oriented Gradient, HOG) 主要通过计算局部图像梯度方向直方图来表示图像特征[34]。HOG 和支持向量机 (Support Vector Machine, SVM) 的结合广泛应用在图像识别等诸多领域，并取得了极大的成功[35, 36]。

HOG 的主要思想是在一幅 RGB 图像中，利用梯度或者边缘方向的密度信息

对局部目标的形状或者轮廓进行描述，即符合梯度主要存在于边缘地方的原则。HOG 方法的实现首先是将图像划分成一个个小的连通区域，然后采集该区域中像素点的梯度或者边缘的方向直方图，最后将所有直方图组合在一起构成该图像的特征。

HOG 特征提取算法的主要过程有:

1. 灰度化（将 RGB 图像用灰度值表示)。
2. 归一化。主要通过 Gamma 校正法对 RGB 图像进行颜色空间的标准化，以降低因光照变化对图像局部信息造成影响，比如阴影等，同时可以调节图像对比度，抑制噪音干扰。
3. 获取图像每个像素的梯度，捕获图像中物体的轮廓信息，进一步降低光照的干扰。
4. 将图像划分成一个个小的连通区域（例如 6*6 像素)；
5. 统计每个连通区域的梯度直方图，形成每个连通区域的特征；
6. 将每几个连通区域组成一个大的连通区域（例如 3*3 个小连通区域)，一个大连通区域内包含所有小连通区域的特征，串联起来便得到大连通区域的 HOG 特征。
7. 将 RGB 图像中所有大连通区域的 HOG 特征串联起来就可以得到该图像的 HOG 特征。

2.3 RGB 图像特征提取深度模型分析

随着 RGB 图像采集设备的不断发展，RGB 图像数据也越来越多，比如 ImageNet [37]。大量数据的获得，为深度学习在图像方面的应用提供前提。同时计算机处理速度不断提高，可以满足更深层次的深度神经网络。基于 RGB 图像的深度神经网络不断发展，各种各样的神经网络架构被提出，其中 1998 年提出的 LeNet-5 [38] 是较早的一种卷积神经网络，但是它的整体设计还是比较小，总参数大约 6 万个。因此，随着时代发展，更深层次的神经网络在图像特征提取方面的应用得到了广泛发展，其中 AlexNet [39] 和 VGGNet [40] 是应用比较广泛的两个利用 RGB 图像特征进行图像分类任务的深度神经网络模型。

2.3.1 AlexNet

AlexNet 属于较早期的一个卷积神经网络，于 2012 年在 ImageNet 比赛中大放异彩，掀起了一阵深度神经网络的热潮。相比较 LeNet-5，AlexNet 则更复杂，学习参数 6 千万个，神经元约有 65 万个，一共 8 层，5 个卷积层和 3 个全连接层, 如图 2.6所示。

AlexNet 在每个卷积层和全连接层之后都采用 ReLU 激活函数，如公式 2.3所

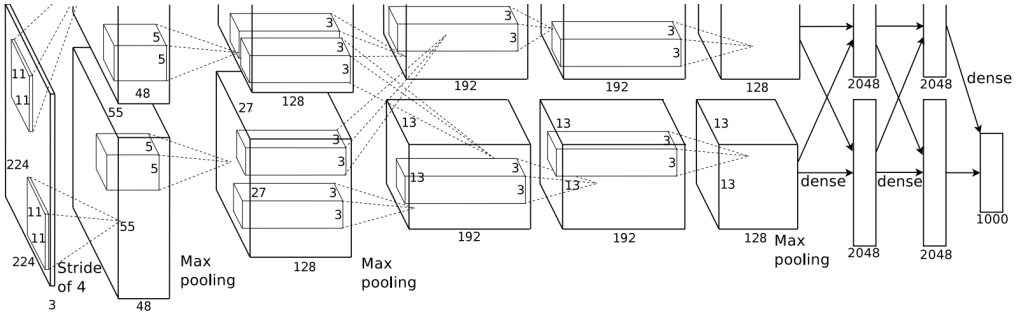

图 2.6　AlexNet 结构框架 [39]

示。当 $x > 0$ 的时候输出为 x，斜率恒为 1。ReLU 函数的使用，使神经网络的收敛速度更快，效果更好，十分适用于大型神经网络模型的训练，而且不需要对数据进行标准化处理。本文所设计的 RGB 图像特征提取网络中借鉴该函数的优势，采用了 ReLU 激活函数。

$$f(x) = max(0, x) \tag{2.3}$$

AlexNet 在全连接层之后加入 Dropout 函数 [41]，随机关闭部分神经元，可以减少过拟合现象，同时使训练的迭代次数增加。这也是本文所设计的 RGB 图像特征提取网络中所采用的用来减少过拟合现象的一种有效方法。

2.3.2　VGGNet

VGG 模型是在 AlexNet 网络的基础上发展而来的，虽然在 2014 年的 ILSVRC 竞赛中未获得第一名，但是它在诸多迁移学习任务中表现优于第一名，成为从图像中提取特征的首选算法。本文也通过多个深度卷积神经网络进行图像特征提取，最后采用了 VGG-16 这个网络框架。VGG-16 的各个层次如表 2.1 所示。其主要贡献在于采用了非常小的卷积核，以 3*3 的卷积核替换其他神经网络中较大的卷积核，以获得更大的特征图，有更多的非线性变换，提取更多的特征，同时拥有更小的计算量。比如用 3 个 $3 * 3$ 的卷积层代替 $7 * 7$ 的卷积层可以有效地减少参数的数量，其中 3 个 $3 * 3$ 有 $3(3^2C^2)$ 个参数，而 $7 * 7$ 有 7^2C^2 个参数。VGG-16 同时增加网络深度到 16—17 层，明显提升了模型的效果，而且对其他数据集具有很好的泛化能力。

[40] 提到网络权重的初始化非常重要，如果没有选好合适的初始值可能会造成学习停止。VGGNet 为了克服这个问题，首先对 VGG 中最简单的网络版本的参数进行随机初始化开始进行训练，然后，用训练好的该网络的参数对后边深层网络的前 4 层以及全连接层进行初始化，最后用均值为 0，方差为 0.01 的正态分布对剩下的层进行随机初始化，同时将 biases 的初始值置为 0。并且 [40] 里提到，

表 2.1　VGG-16 结构表

VGG-16 Configuration
16 weight layers
input (224 × 224 RGB image)
conv3-64 conv3-64
maxpool
conv3-128 conv3-128
maxpool
conv3-256 conv3-256 conv3-256
maxpool
conv3-512 conv3-512 conv3-512
maxpool
conv3-512 conv3-512 conv3-512
maxpool
FC-4096
FC-4096
FC-1000
soft-max

可能存在不需要预训练参数直接进行初始化参数的方法，但是该方法需要进一步探索。

为了获得归一化的 224 × 224 的输入图像，论文里首先对原图进行各向同性放缩，然后采取随机裁剪的方式得到规定尺寸的图像大小，使用 S 表示放缩后的 RGB 图像的最短边，规定 $S >= 224$。同时为了获得更大的数据集，论文将裁剪后的图像进行了随机的 RGB 颜色转换以及随机的水平翻转。

同时，VGGNet 还采用了多尺度的训练方式，换言之，就是将原始图片采用不同的 S 放缩后进行裁剪训练，比如让 $S = 256$，$S = 384$ 分别训练，还有一种动态改变 S 的方式, 就是给定 S 的范围 $[Smin, Smax]$，让每张训练图片随机选取 S 进行训练，这样相当于通过抖动尺度来增强数据。

本文采用了 VGG-16 的网络框架，对输入的 RGB 图片进行特征提取可以获得很好的效果。同时本文所提出的结构掩膜网络也是采用了 VGG-16 的框架，并使用其在 ImageNet 训练好的参数对本文网络进行初始化。通过本文的实验结果（第 3.3.3节和第 5.3.2节）可以看出 VGG-16 在提取 RGB 图像的结构特征方面效果十分显著，同时可以通过与 VGG-19 框架的对比可以观察到，VGG-19 相对 VGG-16 效果更胜一筹，但同时由于网络层数的增加，网络参数的扩大，该网络为实验硬件平台的要求也更高。故本文主体采用 VGG-16 进行实验，用 VGG-19 进行实验比较，肯定 VGG-19 效果更好的事实。

2.4 小结

本章首先围绕 RGB 图像特征提取的相关方法进行介绍。首先阐述了 RGB 图像特征的定义以及应该具备的属性，然后对传统 RGB 图像特征提取的方法进行详细刻画，最后针对深度神经网络发展的现状，重点讨论了 AlexNet 和 VGG-16 这两个神经网络，而本文也是采用了 VGG-16 这个网络框架进行 RGB 图像特征提取以及在此基础上搭建了结构掩膜网络，并取得了很好的效果。

第三章　RGB 图像目标对象掩膜提取

通过第一章的论述，实现基于 RGB 图像的三维重建，首先需要对 RGB 图像进行目标提取，识别图像中目标物体的区域，排除图片中背景的干扰，才能实现高质量的重建结果。目标提取是指单幅图像或序列图像中将感兴趣的目标与背景分割开来，从图像中识别和解译有意义的物体实体而提取不同的图像特征的操作。目标提取的应用范围很广，在计算机视觉提取人脸特征和指纹等，在摄影测量与遥感中，用于特征点线的提取来进行影像匹配和三维建模等。本章将分析传统的目标物体提取方法和基于深度学习的目标物体提取方法，进而详细描述本章所设计的深度学习模型用来实现目标提取的目的，并且对设计方法的实验效果进行展示。

3.1　目标物体提取相关方法

近年来，基于图像的目标物体提取被看做是图像处理方面一个关键且具有挑战性的领域。目前的工作进行目标物体提取操作有一个特定的流程，大多数提出的方法都使用这一共同的流程，并改进了流程中一些步骤。图 3.1展示了基于图像的目标物体提取的一般流程。

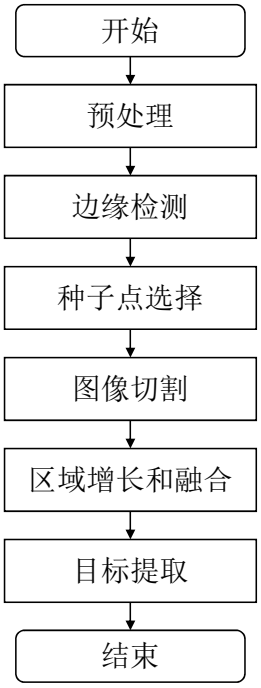

图 3.1　基于图像的目标物体提取流程图

3.1.1 图像预处理

图像预处理是任何图像处理技术的基础部分，它依赖于具体的技术需求。在进行图形分析中，图像质量的好坏直接影响到实验的效果以及算法的准确性，因此进行图像预处理是进行图像处理必不可少的一步。图像预处理主要是将图像中的无关信息或者较大的噪声信息进行消除，恢复图像中真实有用的信息，增强图像信息的可检测性，最大限度地监护数据，从而尽可能提高图像处理的效果，恢复实验的准确性。一般预处理流程需要经过三步：灰度化、几何变换以及图像增强。

对图像进行预处理，首先需要进行灰度化操作，在对 RGB 图像进行处理的过程中，实验者往往需要对图像的三个通道依次进行处理，内存开销和时间开销较大。而采用灰度化的方法后，可以减少需要处理的数据，大大提高数据处理效率，加快处理速度。

在进行灰度化操作后，还需要对图像进行几何变换。为了修正图像采集过程中造成的误差，需要通过平移、旋转、对称、缩放等几何变换对采集到的图像进行处理，从而尽可能恢复最真实的图像。同时在进行几何变换的过程中，为了修正因为几何变换导致的坐标映射到非整数坐标上的问题，通常采用灰度插值算法。

图像增强主要是为了增强图像中的有用信息[42]。针对给定图像的应用场合，有目的地强调图像的整体或局部特性，将原来不清晰的图像变得清晰或强调某些感兴趣的特征，扩大图像中不同物体特征之间的差别，抑制不感兴趣的特征，使之改善图像质量、丰富信息量，加强图像判读和识别效果，满足某些特殊分析的需要。

3.1.2 边缘检测

为了从图像中提取目标物体，首先需要确定目标物体的范围，这就需要在静止的图像中找到对象的边缘[43]。而边缘检测是为了找到图像中亮度变化剧烈的像素点构成的集合，而一般这个变化剧烈的地方往往就是对象的轮廓。

3.1.3 图像分割

图像分割，顾名思义，就是需要根据图像不同部分的类型对图像进行分区，让区域间显差异性，区域内呈相似性[44]。这一操作主要是将图像分割成不断区域，方便对图像进行检测。分区和背景分离主要适用于图像分割。

对图像进行分区是目标提取的一部分。为了达到从图像中提取任何不同对象的目的，首先需要提取不同的区域或者区分每个区域以进一步进行操作[45]。

背景分离：背景分离是简单地用该图像减去背景，使图像中剩下的唯一的东西是对象，然后再将前景像素与背景像素区分开来[46]。

3.1.4 种子点选择

种子点选择是对象提取的第一步，该过程可以通过两种方法完成。第一种是在用户交互类别过程中，用户需要在种子点选取的运行时间内进行交互，被用户选中的部分或者点将作为种子点[45]。第二种是采用种子点自动选取技术，该方法在基于窥视点及其变种中自动选取种子点[47]。

3.1.5 区域增长和融合

区域增长和融合操作是根据相似性约束迭代地合并一组初始小区域。首先需选择一个任意点作为种子像素然后将其与相邻像素进行比较。从种子点开始不断增加与其相似邻居像素点，该种子区域就会不断增长。区域合并操作通过不断合并属于同一对象的邻接区域来消除错误边界以及虚假区域。

3.1.6 目标提取

根据不同的任务需求，有的需要从图像中检测所有的对象，有的只需要找到某一种类的物体或者中心对象或者最大的对象等。通过以上步骤，可以获得目标物体所在大致区域，最后再对边缘进行优化就可以实验目标物体的提取过程。

3.2 深度学习模型分析

深度学习是机器学习领域，尤其是在计算机视觉领域真正改变游戏规则的玩家。深度学习已经击败在图像分类任务中的其他经典模型，与此类似，深度学习目前也是目标检测方面最先进的计算方式。

3.2.1 神经网络概述

深度学习模型的架构一般是由一些相对简单的模块多层堆叠起来，并且每个模块将会计算从输入到输出的非线性映射[48]。每个模块都拥有对于输入的选择性和不变性。一个具有多个非线性层的神经网络通常具有 5 到 20 的深度，它将可以选择性地针对某些微小的细节非常敏感，同时针对某些细节并不敏感，例如图像中的背景。

深度学习也就是自动地学习特征，它的另一个别名是特征学习。而深度神经网络是深度学习一个具有代表性的实例。David Hubel 和 Torsten Wiesel 发现了视觉系统的信息处理本质[49]：可视皮层是分级的，而神经、中枢、大脑的工作过程，或许就是一个不断迭代、不断抽象的过程，如图 3.2所示。

1. 从原始信号摄入开始，例如瞳孔摄入像素 (pixels)。
2. 接着做初步处理，例如大脑皮层某些细胞发现方向和边缘 (edges)。

3. 然后抽象，例如大脑判定眼前的物体的形状是圆形的 (object parts)。
4. 然后进一步抽象，例如大脑进一步判定该物体 (object models)。

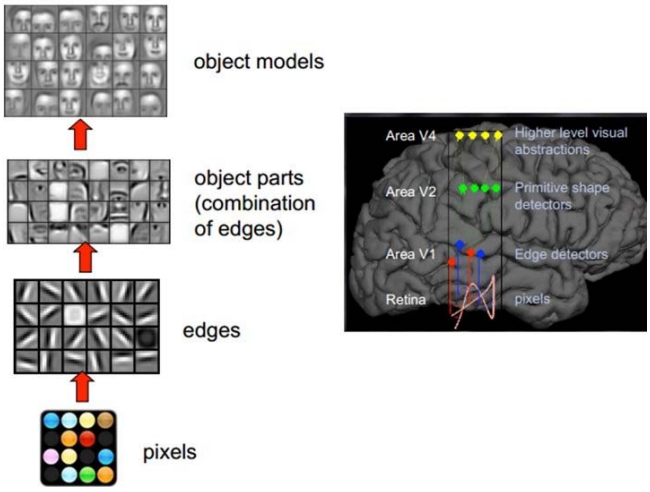

图 3.2　人类视觉信息处理系统

人的视觉系统的信息处理是分级的。从低级的 V1 区提取边缘特征，再到 V2 区的形状或者目标部分等，再到更高层，整个目标、目标的行为等，高层特征是低层特征的组合，从低层到高层特征表示越来越抽象，越来越能表现语义或者意图。抽象层面越高，存在的可能猜测就越少，就越利于分类。比如，单词集合和句子的对应是多对一的，句子和语义的对应又是多对一的，语义和意图的对应还是多对一的。

神经网络对图片信息的处理过程，就类似与人类的视觉信息处理系统的工作原理，一步一步不断迭代，获得特征，识别物体。而神经网络根据输入神经元和输出神经元之间是否有反馈分为前馈神经网络和反馈神经网络。

1. 前馈神经网络是一种最简单的神经网络，各神经元分层排列。每个神经元只与前一层的神经元相连。接收前一层的输出，并输出给下一层。各层间没有反馈，可用一个有向无环图表示，是目前应用最广泛、发展最迅速的人工神经网络之一。前馈型神经网络的学习主要采用误差修正法（如 BP 算法），计算过程一般比较慢，收敛速度也比较慢。卷积神经网络 (Convolution neural networks, CNN) 是前馈神经网络的一个典型代表。
2. 反馈型神经网络可以用离散变量也可以用连续取值，每个神经元同时将自身的输出信号作为输入信号反馈给其他神经元，它需要工作一段时间才能达到稳定, 需要用动态方程来描述系统的模型。Hopfield 神经网络是反馈网络中最简单且应用广泛的模型，它具有联想记忆的功能，如果将李雅普诺夫函数定义为巡游函数，Hopfield 神经网络还可以用来解决快速寻优问题。递归神

经网络 (Recurrent Neural Networks, RNN) 是反馈神经网络的一个典型代表。

3.2.2 卷积神经网络

在前馈神经网络中，各神经元分别属于不同的层。整个网络中无反馈，信号从输入层到输出层单向传播。前馈神经网络的训练过程也被称为反向传播算法，可以分为以下三步：

1. 前馈计算每一层的状态和激活值，直到最后一层；
2. 反向传播计算每一层的误差；
3. 计算每一层参数的偏导数，并更新参数。

卷积神经网络是一种典型的前馈神经网络。卷积神经网络有三个结构上的特性：局部连接，权重共享以及（空间或者时间上的）次采样。这些特性使得卷积神经网络具有一定程度上的平移、缩放和扭曲不变形。

一个典型的卷积神经网络的框架一般包括输入层、卷积层、采样层、全连接层和输出层，如图 3.3所示为一个基于图像的卷积神经网络框架。输入层一般是指输入的图像信息，比如 RGB 信息或者灰度信息等，在三维方面则是指采集到的三维数据。卷积层是用来进行特征提取。池化层是对输入的特征图进行压缩，一方面是特征图变小，简化网络计算复杂度；另一方面是进行特征压缩，提取主要特征。全连接层从特征图中学习到一个非线性组合特征，另一方面卷积层的输出代表着数据的高级特征，全连接层将高级特征扁平化并且能够直接连接到输出层。而本文就采用了卷积神经网络进行 RGB 图像方面的目标提取和特征提取。

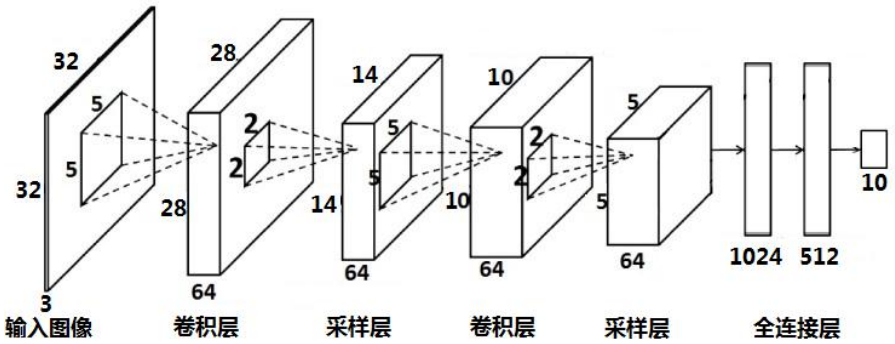

图 3.3 卷积神经网络的整体框架

3.2.3 递归神经网络

递归神经网络是两个人工神经网络的总称。一种是时间递归神经网络又称为循环神经网络（Recurrent Neural Network），另一种是结构递归神经网络（Recursive Neural Network）。递归神经网络引入“记忆”的概念，递归是指其每一个元素都

执行相同的任务，但是输出依赖于输入和“记忆”。

递归神经网络是一种典型的反馈神经网络。从网络结构上，递归神经网络会记忆之前的信息，并利用之前的信息影响后面节点的输出，还包括了上一时刻隐藏层的输出。从循环神经网络的结构特征可以很容易得出它最擅长解决的问题是与时间序列相关的。循环神经网络也是处理这类问题时最自然的神经网络。对于一个序列数据，可以将这个序列上不同时刻的数据依次传入循环神经网络的输入层，而输出可以是对序列中下一个时刻的预测，也可以是对当前时刻信息的处理结果。循环神经网络要求每一个时刻都有输入，但是不一定每个时刻都需要输出。

循环神经网络可以看做是同一个神经网络结构在时间序列上被复制多次的结果，这个被复制多次的结构被称之为循环体。在循环神经网络中，循环网络结构中的参数在不同时刻也是共享的。

3.3 结构掩膜网络的搭建

本文搭建了结构掩膜网络用来进行目标提取。类似于其他的神经网络的研究过程，本文采用的用来提取目标物体的神经网络的研究过程也是依次进行了数据准备、目标提取网络搭建、学习训练、实验结果分析。

3.3.1 数据准备

本文的结构掩膜网络可以得到 RGB 图片的前景和后景，前景是指 RGB 图像中目标物体所在占据的像素，后景是指 RGB 图像中背景所占据的像素，然后以二值矩阵的形式对该提取出目标物体区域的 RGB 进行表示。

首先，结构掩膜网络是完成 RGB 图像三维重建这个大的神经网络框架的一部分。因为基于 RGB 图像三维重建这个网络的训练数据是由 RGB 图像与 3D 模型共同组成的，因此，本文采用的训练数据中的 RGB 图像是对 3D 模型渲染获得的，并且通过对 3D 模型不同视角的渲染来达到扩大数据对的目的。而本文使用的 3D 模型是来自于 ShapeNet3D 模型数据集，该数据集包含 55 个种类，大约 51300 个互不相同的 3D 模型。同时为了使 RGB 图像更加真实，我们通过添加背景获得训练使用的 RGB 图像来模拟真实的 RGB 图像。

在训练目标提取神经网络的过程中，为了获得 RGB 图像的前景和后景，本文所采用的实验数据对需要一张 RGB 图像以及与 RGB 图像对应的二值矩阵。为了获得与 RGB 图像对应的二值矩阵，本文首先定义在互联网上下载的大量图像作为背景图，如图 3.4(a)，然后定义直接渲染 3D 模型所获得的目标物体的 RGB 为前景图，即图像中像素点值不是 255 的像素点集合，如图 3.4(b)。然后将前景图置于背景图上方，即将前景图像中像素点为 255 的值替换为背景图中相应位置的像素值，形成最后的训练图像，如图 3.4(c)。

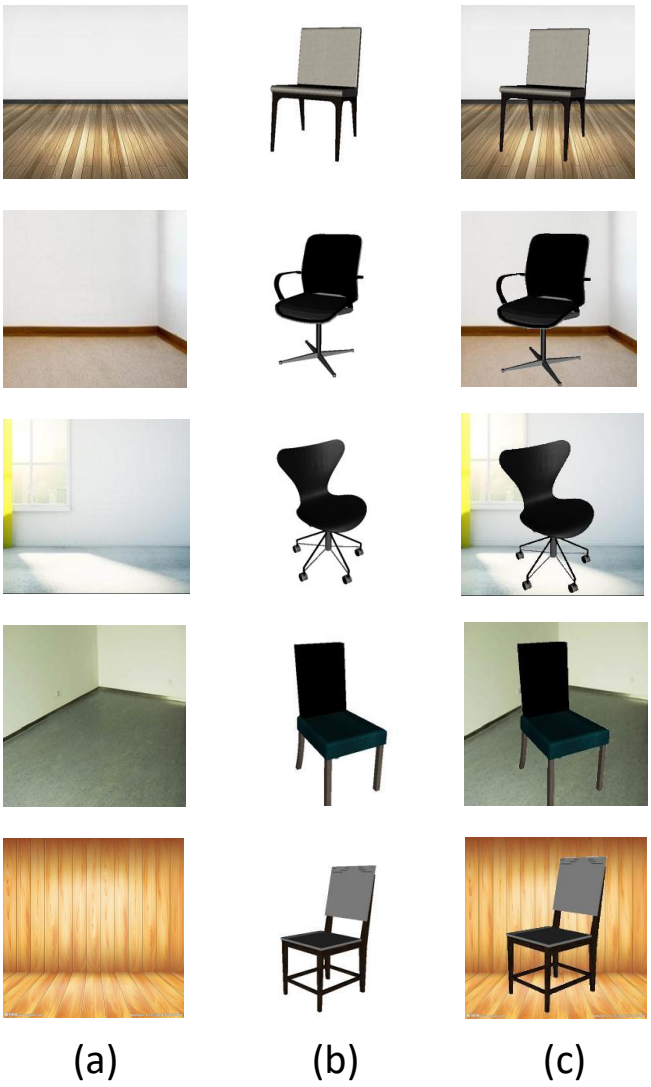

图 3.4　训练图像合成

在生成训练数据对的时候，我们将网络需要输出的矩阵定义为二值图像，即训练图片中保留前景像素点的位置置为 1，保留背景像素点的位置置为 0，这样就形成了与每一张 RGB 训练图像相对应的二值图像。由一张训练图像和与其对应的二值图像构成一个训练数据对，用来作为目标提取网络的训练数据，也是基于 RGB 图像三维重建的训练数据的一部分。

3.3.2　目标提取网络搭建

本文所采用的目标提取网络是一个深度卷积神经网络，如图 3.5，被命名为结构掩膜网络。该网络以一个 RGB 图像作为输入，同时进入该神经网络的两条网络线，并且通过该网络的学习，输出对应于该 RGB 图像的二值图像，该二值图像中为 1 的位置就是目标物体所占据的区域或者像素点集。

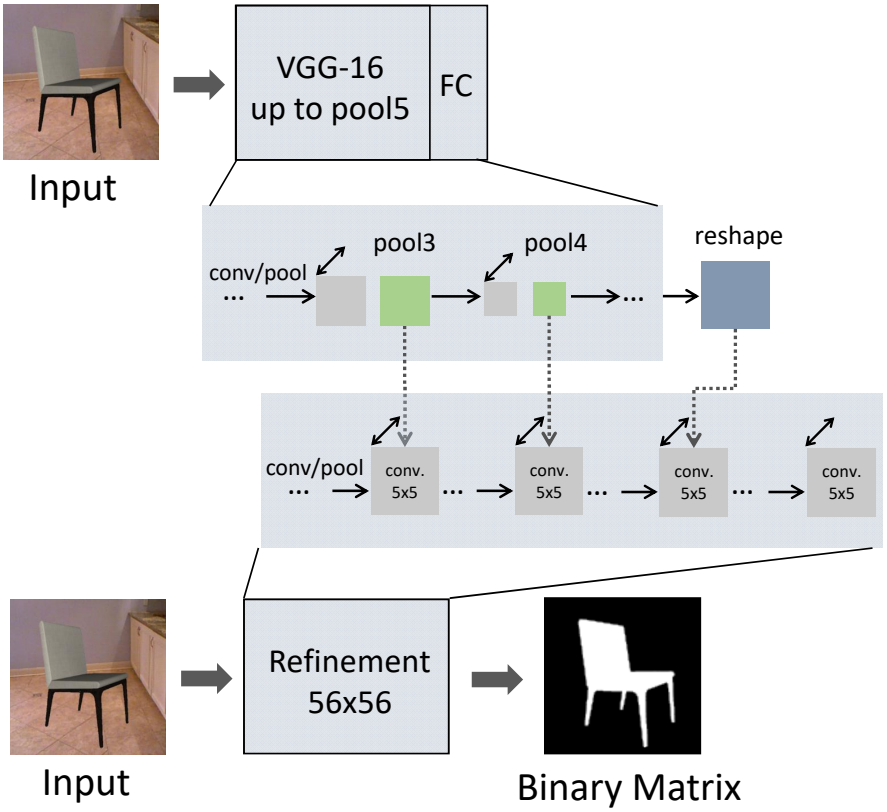

图 3.5　目标提取网络

目标提取网络是由两条网络线组成多尺度卷积神经网络，图 2.5 上方的线采用 VGG-16 的前 5 层，并在其后添加了两层全连接层，为了和第二条线融合到一起，在其最后加入了一个 reshape 函数，将特征矩阵的维度与第二条线的维度进行匹配然后连接，第一条线主要起到获取全局特征的效果。第二条线包含一个 9*9 的卷积层和一个池化层，然后是 9 个连续的 5*5 的卷积层并且没有池化层，主要起到保持 RGB 图像局部特征的作用。两条线的输入是同一个 224*224 的 RGB 图像，第一条线中特征矩阵在第三个池化层后和第四个池化层之后分别作为输入通过一个连接层到第二层的第二个和第四个卷积层，其中连接层由一个 5*5 的卷积层和一个上采样操作共同构成，可以有效地提取图片中的细节信息。第一条线的特征矩阵在 reshape 函数后直接连接到第二条线的第六个卷积层。第二条线最后通过所有的卷积层后，输出一个 56*56*2 的二值矩阵。

3.3.3　神经网络训练

本文采用的目标物体提取神经网络通过输入一个 RGB 图像，可以得到与之对应的二值矩阵，该二值矩阵可以表示 RGB 图像中目标物体所占据的区域。由于该网络的输出是一个二值矩阵，即由 0 与 1 来表示的矩阵，故本文采用 SoftMax

Loss 作为本网络的损失函数。

本文定义 $OP(OverallPixel)Error$ 为：标记错误的像素占总像素的比例，如公式 (3.1) 所示，共 k 类，P_{ij} 表示属于 i 类的像素被预测为 j 类的像素数目。在目标物体提取网络中 $k=2$, 即只有两类，前景一类，背景一类。

$$OPError = 1 - \frac{\sum_{i=1}^{k} P_{ii}}{\sum_{i=1}^{k}\sum_{j=1}^{k} P_{ij}} \tag{3.1}$$

本文对目标物体提取神经网络训练了 100 次，共训练了 $10K$ 对数据。该网络在训练和测试阶段的错误率曲线如图 3.6所示，在第 100 次的时候已经收敛。

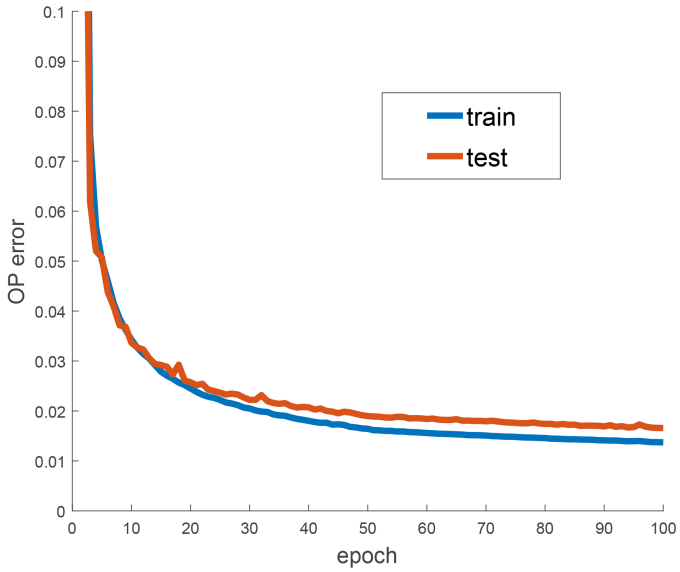

图 3.6　目标物体提取神经网络的训练曲线

3.3.4　实验结果分析

3.3.4.1　结果展示

目标物体提取神经网络在训练到第 100 次后，损失函数已经收敛，实验效果也很明显。通过二值图与原 RGB 图对比，可以看到对目标物体所属区域的提取是很好的，实验结果图如图 3.7所示。

3.3.4.2　结果评估

本文通过 OP(Overall Pixel)_Accuracy 与 PC(Per-Class)_ Accuracy 来对实验结果进行评估，具体计算方式如公式 3.2和公式 3.3 所示。该公式中共 k 类，其中 P_{ij} 表示属于 i 类的像素被预测为 j 类的像素数目。在目标物体提取网络中 $k=2$, 即

图 3.7　结果展示

表 3.1　结果评估表

Method	OP_Accuracy	PC_Accuracy
single-scale	0.953	0.917
two-scale(w/o jump)	0.982	0.964
two-scale(with jump)	0.988	0.983

只有两类。

$$OP\_Accuracy = \frac{\sum_{i=1}^{k} P_{ii}}{\sum_{i=1}^{k}\sum_{j=1}^{k} P_{ij}} \tag{3.2}$$

$$PC\_Accuracy = \frac{1}{n}\sum_{i=1}^{k}\frac{P_{ii}}{\sum_{j=1}^{k} P_{ij}} \tag{3.3}$$

同时，本文还对目标物体提取神经网络的架构进行了修改：第一种只有一条流水线，即采用第二条流水线；第二种两条流水线但是二者之间没有连接层，只在最后 reshape 操作是连接到一起；第三种两条流水线且有连接层，即本文所采用的网络的架构。表 3.1是对这三种网络在 OP(Overall Pixel)_Accuracy 和 PC(Per-Class)_Accuracy 两方面的评估，可以看出本文所采用的网络架构是效果最好的，不论是双流水线还是两条流水线之间的连接层都为提高准确率做出了贡献。

3.4　小结

本章围绕基于深度学习的目标对象掩膜提取方法，介绍了传统的目标物体提取方法，以及随着深度学习的不断发展，卷积神经网络和递归神经网络等深度神经网络在图形图像处理领域不断得到广泛应用。最后着重阐述了本章所采用的基

于卷积神经网络的目标物体提取方法，并且从数据准备、网络框架、训练细节以及实验结果分析等方面，描述了从 RGB 图像到二值图像的获取过程，是完成 RGB 图像的三维重建深度学习方法第一步。

第四章　三维结构的表示与重建方法

本章重点介绍三维模型的结构表示方式和对应的解码方法，采用有向包围盒和以其为叶节点的树状层次结构来代表三维物体中各个部件和部件之间的相互关系。这些关系包括连接性、镜面对称性、平行对称性、旋转对称性等三维物体本身所具有的结构信息。针对这种树状层次结构，本章采用递归神经网络的方法进行解码，构成结构重建网络。

4.1　RGB 图像三维重建的相关方法

近几年，我们见证了基于单视图图像三维重建的快速发展[6, 7, 11]，由于深度卷积神经网络的巨大成功和广泛应用，三维重建的性能得到了很大提升。然而，现有的深度模型主要针对三维体积表示的输出或者是多视角图片作为输出，很少能很好地恢复三维模型的结构信息。而三维模型的结构信息对三维形状的理解非常重要 [50]。通过网格模型或者体素模型推测三维形状的部件是非常困难的 [51]，即使给出了分段，对于诸如对称性等部件关系的推理仍然具有挑战性。

4.1.1　基于部件检索的图像三维重建

随着 RGB 图像采集设备的不断更新，价格也越来越低廉，使得 RGB 图像数据量飞速增长。互联网上已经有大量的 RGB 图像数据，并且可以被用于重建具有成千上万访客密集采样的地标场景[52]。随着数百万的日常生活物品的 RGB 图像已经在网上出现和存储，有研究者[53] 提出即使一个三维物体只出现在一个图像中，也可以利用物体形状的规律性重建出该三维模型。这种想法的提出也为通过挖掘网络创建大型三维模型数据库铺平了道路。

大量数据的存在，为研究者提供了新的思路，于是近几年的研究就有利用大型三维模型数据库进行检索，找到与 RGB 图像中对应的三维物体相匹配的部件，然后将这些部件组装为一个与 RGB 图像中相似的三维物体，或者尽可能相似。其中 Qixing 等人[54] 的工作就具有一定的代表性，他们实现了重建 RGB 图像的三维物体的方法，即使该三维物体仅出现在单张图像中，

使用现有的相对较少的三维数据集通过检索匹配相似三维模型部件，从而指导重建过程具有很多挑战，其中最关键的一个挑战是现有的三维模型数据可能只是对潜在的形状空间进行的稀疏采样。而且在已经存在的三维数据库中，如 3D Warehouse，高质量的三维形状数据远远少于互联网上已有的 RGB 图像上数量。对于很多物体种类而言，高质量的图像数据远远超过了高质量的三维形状模型。因此简单地为每个输入图像检索最相似的现有三维模型不会产生令人满意的结果，

即使检索可靠，最接近的预先存在的模型也是经常和所描绘的对象不同。

针对上述挑战，[54] 实现了一种基于组装的方法，通过组装数据库中预先存在的形状模型来重建对象。这种方法的关键之处是联合分析图像和三维模型。首先通过优化一个可以测量相似图片的一致性全局值来估计所有图片的相机位姿。然后再在自然图像和渲染图像直接建立像素级对应的全局密集性网络。这些对应关系可以用于联合分割 RGB 图像和三维模型，同时计算出的分割片段和对应关系也可以被用于指导构建新模型，然后对新模型进行优化。

基于检索的图像三维重建很好地保持了原有物体的对称等相关关系，但是这种方法有一个很大的弊端，必须有一个足够大的用于检索的数据库，才能满足对输入 RGB 图像中对应物体的相似检索，否则很难有很好的效果。如果数据库中的椅子全是四条腿的椅子，那么他就无法检索到转椅的椅腿，因而也就无法对转椅进行重建。在本文的第 5.3.3节，对该方法和本文的方法做了对比，可以看出由于在检索的数据库中缺少了一些三维模型部件，导致部分椅子恢复效果差，或者无法输出结果，比如转椅、交叉腿的椅子。

4.1.2 基于体元组装的图像三维重建

在计算机视觉领域，Binford 于 1971 年引入了广义圆柱体，其横截面区域沿直线或曲线轴展示，同时在此过程中可能缩小或扩展[55]。其中圆柱体简洁的描述性是最主要的动机——一个三维物体可以用相对较少的广义圆柱来描述，每个圆柱体仅仅需要几个参数来表示。体积基元在 20 世纪 90 年代仍然流行，因为它们提供了一个连贯的框架，用于解释单个图像、感知三维物体的形状结构，以及从二维图像识别三维物体。然而，将广义圆柱体拟合到图像数据需要大量的人力标记，并且随着基于物体识别的机器学习技术在 20 世纪 90 年代浮现，这种广义圆柱体的模式逐渐退出计算机视觉的舞台。

基于学习的视觉理解的核心是找到复杂现象的简约解释。事实上，机器学习是最可行的操作。即使我们的视觉世界充满了巨大的复杂性，但是它依旧具有高度结构化的性质，即视觉模式不只是发生了一次，它在各种不同的配置中不断重复。在当代计算机视觉领域，这个结构更是经常通过人为监督来建模：重复的部分被标记为三维物体或者物体的部件，同时监督学习方法多被用于在新颖的图像中查找和命名他们，然而如果可以用简单的底层结构来解释复杂结构，会取得令人更加满意的效果。

基于体元的图像三维重建，其中 [15] 重拾用体积基元来解释物体的经典问题，同时使用无监督学习的方式和卷积神经网络 (CNNs)。该研究采用最简单的体积基元（即刚性变换的长方体），并且展示了利用深度卷积神经网络来组装任意的三维

物体（或者在某种程度上近似三维物体)。该研究是利用体元组装进行三维重建的一项代表性工作，并且取得了令人满意的效果。在经典方法失败的情况下，该研究取得成功的主要原因在于他们的目标是联合解释三维模型数据集中的所有物体，这样可以直接从数据中学习到常见的三维模式。

正如我们所了解到的，网格或者体素这种三维物体表示方法通常是高维而且复杂的，而体积基元这种表示方法则十分简约，它只需要少量的参数进行表示。与本文提出的利用有向包围盒对三维物体的各个部件进行表示方法类似，极少的参数一方面加快了运行速度，减少了内存占用，另一方面对三维物体可视化的呈现直观、快捷、容易分析。在本文的第 5.3.3节，可以看到通过组装体积基元表示三维物体的效果，同时可以发现这种体积基元的表示方法丢失了三维物体的细节信息，并且缺少结构信息。对于一些难以识别的物体还会产生多余的体积基元。

4.1.3 基于体素表示的图像三维重建

类似于像素是图像的最基本的描述单位，而体素是三维物体最基本的描述单位。虽然体素具有诸多缺点，比如没有结构性，重建效果一般有残缺或者有噪声体素，导致整个三维重建方法不具有很强的说服性。但它仍然是很多新方法提出最先使用的三维物体的表示方式。比如最先将用于 RGB 图像中的生成对抗网络[9]用于进行三维物体的生成[12] 就是使用的三维体素的表示方式。其主要的原因在于体素表示的规则性，也是可以直接应用深度神经网络最简单的表示方式。而 [11]算得上是基于 RGB 图片应用深度学习的方法直接进行三维重建的一个代表性工作，该研究也是使用体素的方式进行三维物体呈现，并且该工作的实现过程易于理解，用体素表示的结果也具有前沿性。

大多数基于体素的三维重建工作都是直接利用深度神经网络对 RGB 图像进行编码，将 RGB 图像映射到一个低维的向量中，然后使用深度神经网络直接对该低维向量进行解码，将其解码为一个体素表示的三维物体，而且编码器和解码器结构相似，因为二维像素和三维体素都是规则的数据，可以直接进行卷积操作，因此对于应用于二维图像的深度神经网络都可以直接进行修改拓展到三维体素上面。同时在求损失函数时，可以直接使用类似像素比较的损失函数，比如欧几里得距离、曼哈顿距离等方法。相对于其他类型的三维表示方式，基于体素的三维重建有诸多应用。

类似于可以利用深度神经网络对二维图像的细节特征进行合理化改变，比如可以提取戴眼镜男生的眼镜特征加上不带眼镜女生的特征可以获得戴眼镜女生的结果。同样地也可以将这种方法进行迁移，使其显示到三维体素中来[11]，并且取得的了类似于二维图像的结果。可见体素和像素具有极高的相似性，很多方法都

可以进行迁移，将应有在图像中成功的方法延伸到三维体素中。

虽然体素的规则性和简洁性在各个领域都获得了极大的成果，但是利用体素进行表示的一些工作仍然有所欠缺。一方面体素表示所丢失的结构信息是我们进行三维重建需要重点重视的部分，一个具有结构信息的三维物体才能称之为一个自然的三维物体。另一方面体素在进行分割等操作时仍有一定的弊端。首先是体素无法对三维物体的各个部件进行标识，另一方面体素重建带来的噪声体素对损失函数的求解造成了很大的困扰。本文提出的深度神经网络可以恢复三维形状结构，所以如果将本文的结构信息和三维物体的体素信息相结合必然可以进行互补，产生很好的效果，这也是本文所要进行的下一步工作，用本文恢复出的三维形状结构信息优化重建的体素表示，将噪声点去除，用结构信息补全残缺或者丢失的体素信息。

4.2 三维物体结构表示法

任何一个三维物体都有结构信息，即其对称性、连接性、旋转性等方面。如果不加入结构信息进行三维信息恢复，很容易造成残缺，很难得到一个自然合理的模型。本文针对三维物体具有结构性这个特点，采用一个树状结构对三维模型进行表示，而三维物体的每一个部件用有向包围盒 (Oriented Bounding Box, OBB) 来表示，如图 4.1 所示。

4.2.1 有向包围盒

包围盒是一个简单的几何空间，可视化来看就是一个立方体，但是它包含一个形状复杂的物体，在本文中它包含了一个复杂模型中一个部件的点面信息。如图 4.2 所示，这个三维模型的每一个部件都是由一个有向包围盒表示的。包围盒的边用红色来表示。三维模型用彩色线条组成的网格表示。

本文定义有向包围盒的表示信息包括该包围盒的中心点 (x,y,z 轴的坐标) 共 3 维，然后是包围盒的方向，即包围盒围绕同一个顶点两两垂直的边的方向向量，共 9 维，最后是该包围盒在这三个方向向量上面的长度，共 3 维。所以一个有向包围盒在本文中使用 15 个数进行表示。

4.2.2 树状结构表示法

将三维模型表示为树状结构时，每个叶节点是一个单独的三维模型部件。如图 4.1 所示，红色有向包围盒是一个个部件在由下向上构成一棵树的过程中不断组合的。最底层是一个椅子的左前腿和左后腿分别作为一个叶节点，然后通过前腿和后腿的连接性，将左前腿和左后腿组合为一个具有左前腿和左后腿的父节点。因为椅子的右腿和左腿是对称的，则该父节点通过对称性变成了四条腿，而后通

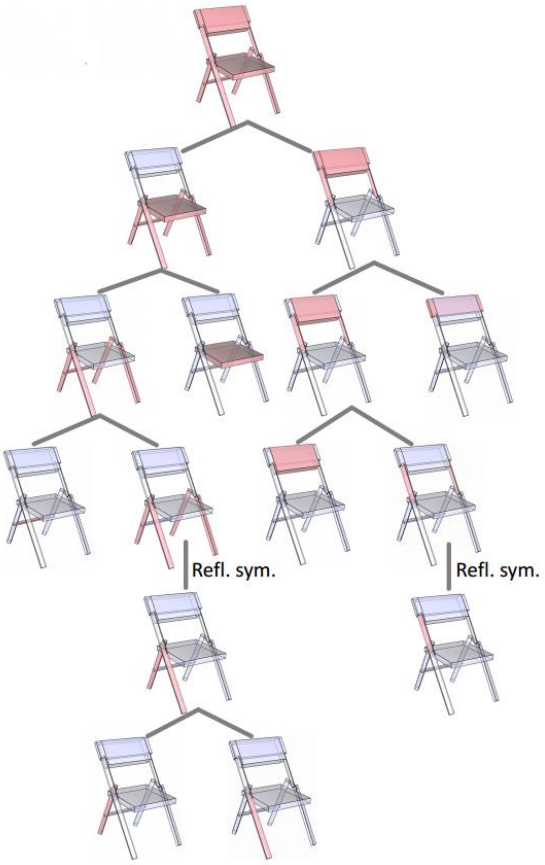

图 4.1　三维模型的树状结构[10]

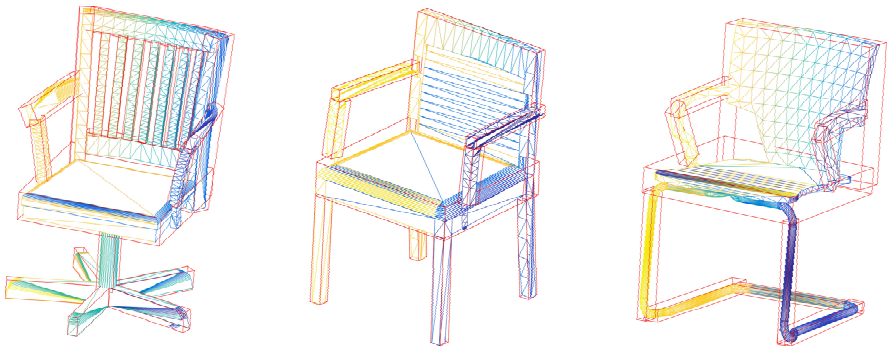

图 4.2　有向包围盒

过连接性与左右腿之间连接轴组合成为一个新的父节点。通过连接性与椅座组合，同时另一侧的椅背也以同样的规则进行组合，最后构成一个完整的椅子。

树状结构中每个叶节点代表三维模型的一个部件，而组合成为父节点的组合规则有四种：旋转性、镜面对称性、平移对称性以及连接性。如图 4.3所示，其中 (a) 代表旋转性，转椅下面的五条腿，拥有一个旋转轴和一个固定的旋转角度，通

过一个腿的旋转就可以获得其他四条腿，因此转椅的腿具有旋转性。(b) 代表镜面对称性，椅子的两个扶手拥有一个对称平面，通过一个扶手关于对称平面做对称可以得到另一个扶手，因此扶手具有镜面对称性。(c) 代表平移对称性，椅座的八个横杆，拥有一个平移距离和平移方向，其中最左边的横杆沿着平移方向移动整数倍的平移距离就可以得到其他七个横杆，因此椅座的横杆以及椅背的横杆都具有平移对称性。(d) 代表部件之间的连接性。椅座和椅腿相连接，因此具有连接性，连接性也是任何一个椅子必然具备的属性。

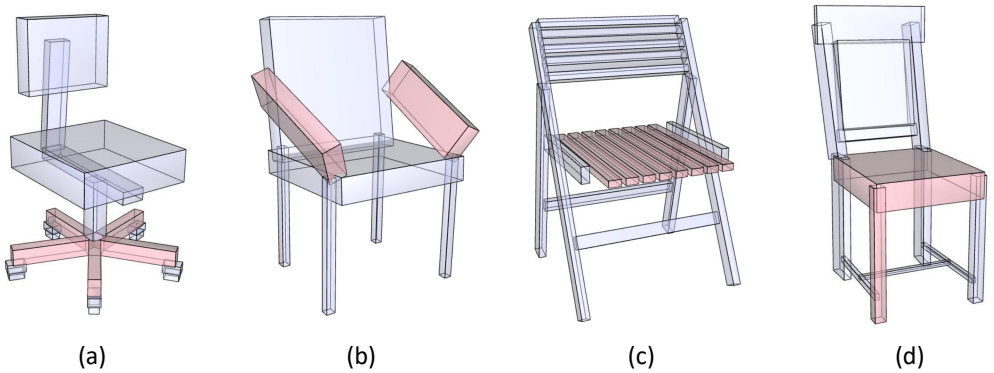

图 4.3　树状结构组合规则[10]

在构造树状结构的过程中，对于三维模型连接性和对称性（旋转性）的优先顺序，本文通过实验得出，连接性先于对称性（旋转性）进行组合。如图 4.4所示，(a) 中采用先连接组合再旋转组合，得到的椅子模型效果很好，符合实际标准。(b) 中采用先旋转组合再进行连接组合，这可以看到椅腿和轮子之间会有误差（图中画红色圈圈处)。因此本文构造树状结构时采用先进行连接组合再进行对称性或者旋转性组合。

4.3　结构重建网络的搭建

本文采用递归神经网络 (Recursive Neural Networks, RvNN) 来学习本文使用的对称层次结构的树状结构，这些层次结构由有向包围盒 (OBB) 的空间排列组成，每个 OBB 都由固定长度的向量定义来表示其几何形状，固定长度的向量对其子 OBB 的几何形状和其详细的分组机制进行编码，分组机制包括连接、旋转、平移、对称四类。为了可以从 RGB 图像中得到的固定维度的特征向量重建出具有层次结构的三维模型，本文提出一个解码器将固定维度的向量解码为三维结构。在搭建本文提出的基于 RGB 图像的三维结构重建网络时，将 RGB 图像固定维度的特征通过该解码器将其解码为三维结构。

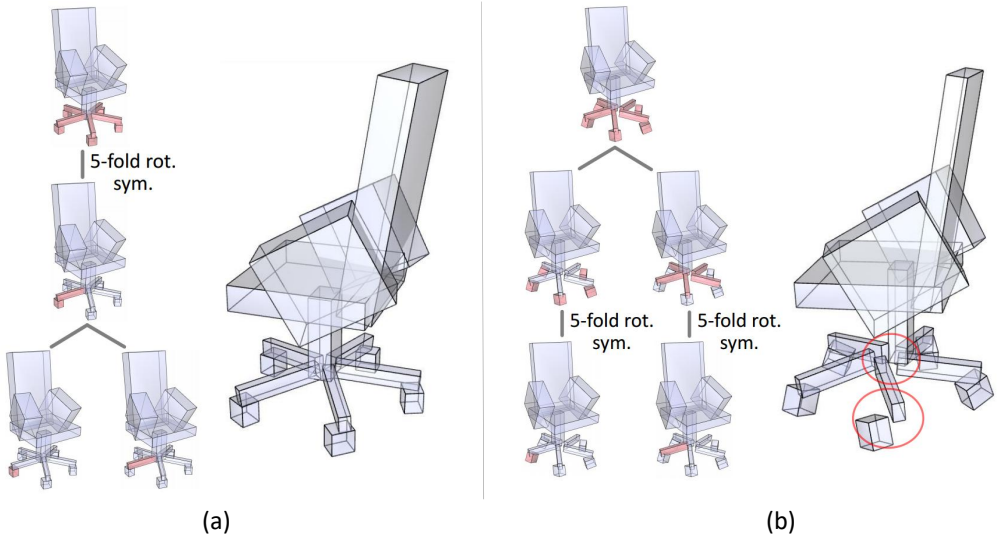

图 4.4　树状结构组合优先顺序[10]

4.3.1　递归神经网络解码

递归结构通常存在于不同模态的输入中，例如自然场景图像或自然语言句子。同时这种递归结构不仅可以帮助我们识别图像或句子所包含的单位个体，还可以识别它们之间是如何相互作用从而构成一个整体的。与之类似，三维模型的各个部件也是通过一定的相互作用构成了一个整体，可以用三维模型的各个部件类比自然语言中的单词，三维模型整体类比完整句子，如图 4.5所示，那么对三维模型的处理也可以采用递归结构进行。

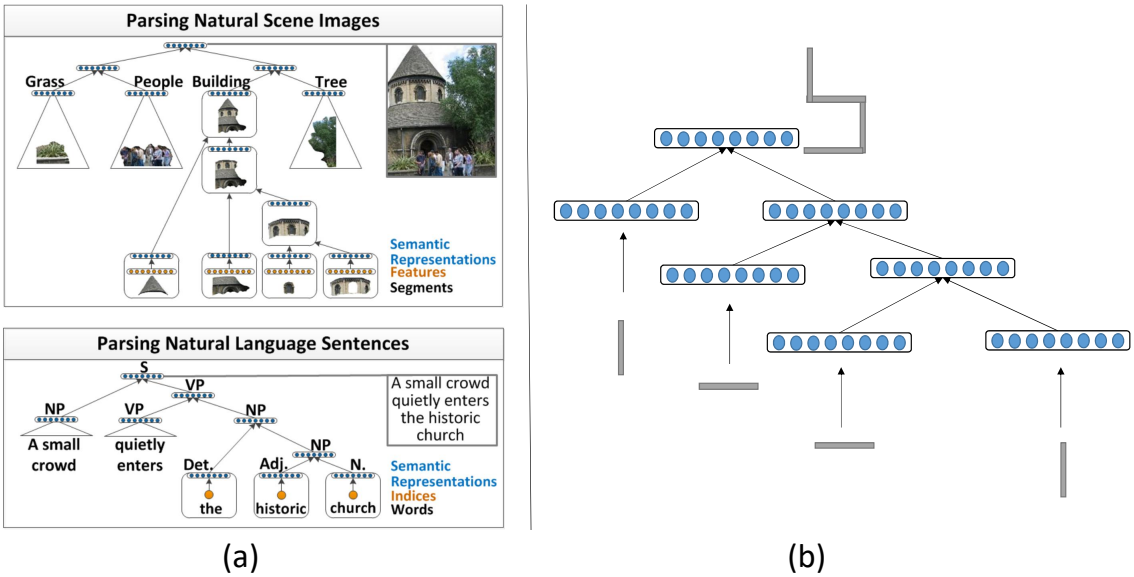

图 4.5　递归结构表示

递归结构通常以不同的形式存在，如图 4.5中 (a)[56] 所示。已知自然语言的句

法规则是递归的，其中包含相对子句的名词短语本身包含名词短语，例如，...the church which has nice windows...。类似地，人们在场景图像中找到嵌套的分层结构，包含部分和邻近关系。例如，汽车通常位于街道区域。大型汽车区域可以递归地分成较小的汽车区域，诸如轮胎和窗户之类的部件，并且这些部件可以出现在其他环境中，例如飞机下方或房屋中。类似于恢复这种结构有助于理解和分类场景图像，恢复三维模型中的这种结构可以更好地进行三维重建，如图 4.5中 (b) 所示，椅子可以细分为椅腿、椅座、椅背和扶手等部件。本文引入递归神经网络 (recursive neural networks, RvNN) 来预测三维模型中的递归结构，主要关注三维模型的结构信息。

本文采用递归神经网络 (RvNN) 作为三维结构解码器，从根特征节点开始（即从 RGB 图像中提取的特征向量），RvNN 递归地将其解码为特征层次，直到到达所有叶节点，而叶节点可以被进一步解码为一个有向包围盒参数。本文的层次结构中有三种类型的节点：叶节点、连接节点以及对称节点（包含旋转、平移、镜面对称)。在解码的过程中，内部节点的部件关系就会被解码为两类：连接关系和对称关系（包括旋转、平移、镜面对称)。因此，每个节点根据它的类型（连接节点、对称节点或者叶节点（代表部件的有向包围盒节点））会被下面三种解码器之一进行解码：

1. 连接关系解码器 (Decoder AdjDec)：将父节点分割成两个子节点 c_1 和 c_2, 使用映射函数：

$$[c_1\ c_2] = tanh(W_{ad} \cdot p + b_{ad}) \tag{4.1}$$

其中 $W_{ad} \in \mathbb{R}^{2n \times n}$，$b_{ad} \in \mathbb{R}^{2n}$。$n = 80$ 代表非叶节点的维度。

2. 对称关系解码器 (Decoder SymDec)：将对称组（包括旋转、平移、镜面对称）解码为一个对称生成器（一个节点向量 c）和一个对称参数向量 s：

$$[c\ s] = tanh(W_{sd} \cdot p + b_{sd}) \tag{4.2}$$

其中 $W_{sd} \in \mathbb{R}^{(n+m) \times n}$，$b_{sd} \in \mathbb{R}^{m+n}$。本文使用 $m = 8$ 代表对称参数，包括对称类型 (1D); 旋转或者平移过程中重复的次数 (1D)；用于镜面对称的对称平面，旋转关系的旋转轴，或者平移关系的位置和位移（6D)。

3. 有向包围盒解码器 (Decoder BoxDec)：将一个叶节点映射为 12D 的有向包围盒参数，包括包围盒中心坐标、坐标轴以及维度。

$$[x] = tanh(W_{lb} \cdot p + b_{ld}) \tag{4.3}$$

其中 $W_{ld} \in \mathbb{R}^{12 \times n}$，$b_{ld} \in \mathbb{R}^{12}$。

在递归神经网络解码期间，需要递归地使用解码器。解码过程中最关键的一点是如何确定节点的类型，从而可以在节点上使用相对应的解码器。这个得基于结构恢复网络的训练任务来学习实现节点分类器，训练过程中给定的 RGB 图像和三维模型的结构层次数据对必须是已知的。节点分类器和三类解码器联合进行训练。结构解码网络的过程如图 4.6所示，在本文网络的实现过程中，邻接节点分类器和对称节点分类器都是双层的神经网络，其隐藏层和输出层分别是 200D 和 80D，而有向包围盒解码器是一层的神经网络，将 n-D 的叶节点映射为 12D 的包围盒参数。

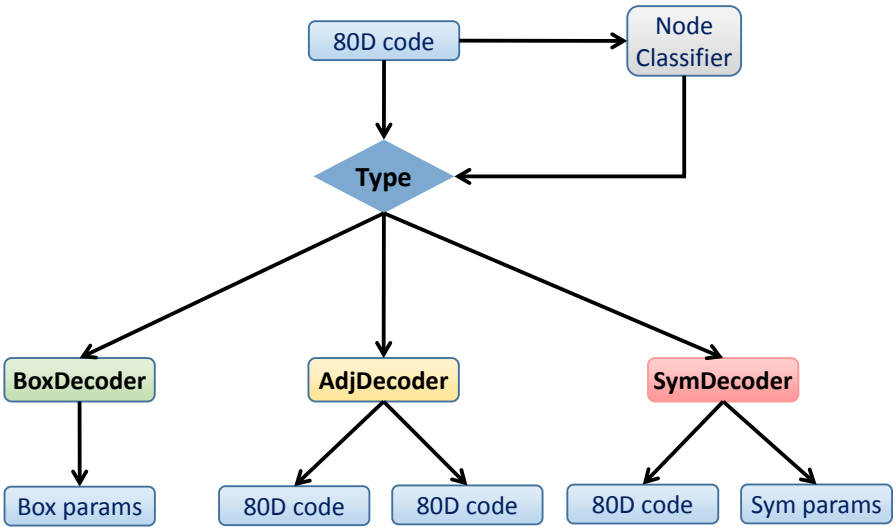

图 4.6　节点解码过程表示图

4.3.2　三维结构重建网络

本文提出的结构重建网络将图像特征提取网络（第二章已详细描述）所提取的图片特征作为特征根节点，并将其递归地解码为部件的有向包围盒的层次结构，即树状结构。

4.3.2.1　数据准备

为了满足训练三维重建递归神经网络，本文从三维模型数据库中合成大量层次结构。假设这些模型预先被分割成部件，但是没有真实的层次结构（即树状结构），因此采用迭代的策略来生成合理的层次结构来满足树状结构合并的标准。在每次迭代的过程中，两个或者多个部件合并成为一个父节点。可合并的部件子集满足连接性或者对称性、旋转性等相互关系，随机抽取一对满足这两个标准之一的节点对，直到无法进一步合并为止。在本文的实验中，这些层次结构是为了代表真实的三维模型层次结构。

输入图片的获取是通过对三维模型数据库中的三维模型进行不同角度的渲染得到的，本文使用了 20 个角度，这样在三维模型数据量确定的情况下，扩大了图像和三维模型层次结构数据对的训练数据量，而在测试的过程中，输入图片是在真实环境下采集的 RGB 图片进行测试的，本文提出的网络可以很好地重建与图像中对应目标物体的三维层次结构。

4.3.2.2 结构重建网络模型

本文提出的结构重建网络可以实现从提取的图片的特征向量转换为三维的层次结构，结构重建网络框架如图 4.7 所示。选取 VGG-16 进行 RGB 图片特征提取，获得 80D 的特征作为根节点，利用结构重建递归神经网络递归地利用节点分类器对节点进行分类，然后根据类别使用不同的节点解码器对节点进行解码，直到不能再进行解码为止，即到达树状结构的叶节点处，此处指每一个部件的有向包围盒的参数。

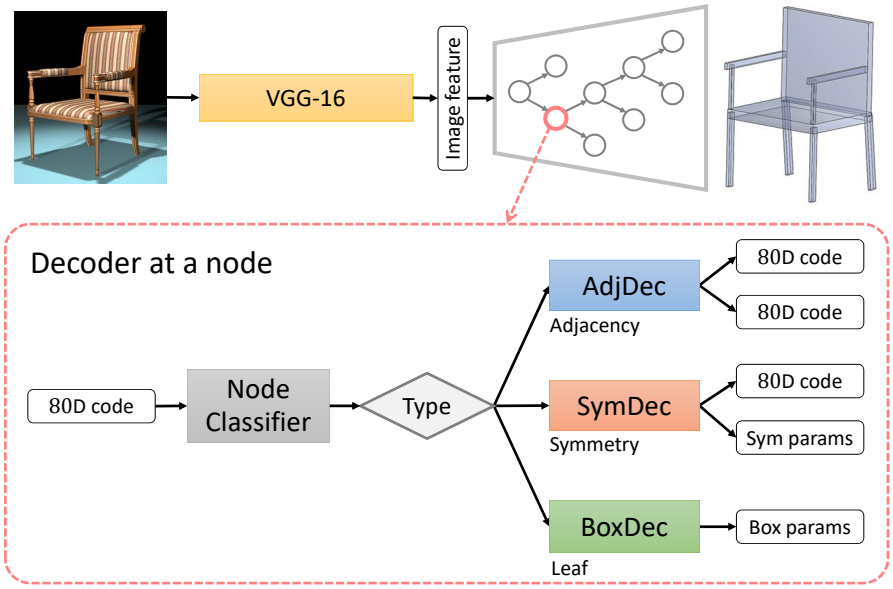

图 4.7 结构重建网络框架

4.3.2.3 训练细节

在结构重建网络的训练阶段，用 VGG-16 对输入 RGB 图像进行特征提取，其初始化参数使用在 ImageNet 上预训练的结果，然后加入两层全连接层使图像特征表示为 80D 的根节点，然后从根节点开始递归地应用相应的解码器 (AdjDec 或者 SymEnc)，直到最后应用有向包围盒解码器 (BoxDec)，损失函数用表示每个部件有向包围盒的输入参数和输出参数的平方差计算。值得注意的是，在训练阶段，需要输入三维模型的层次结构，即树状结构，来进行解码，目的是为了使不同种

类的解码器可以学到应该展开哪个节点，以及学习输入到输出有向包围盒之间的映射。本文同时训练节点分类器，使用具有交叉熵损失的三维 softmax 分类，以便在此时期间进行树形拓扑结构的重建。而在测试阶段，就不需要输入三维模型的层次结构，解码器在训练期间已经学到应该展开的节点以及输入和输出映射，故直接可以输出此时图像所对应的三维模型的层次结构。

在结构重建网络的测试阶段，有一个挑战是解码代表图像特征的根代码以恢复三维模型部件的有向包围盒。为了解码根节点（例如，代表图像特征的根代码），我们递归地调用节点分类器来决定哪个解码器应该扩展节点，相应的解码器 (AdgDec, SymDec 或者 BoxDec) 用于恢复子节点的代码，直到完整的层次结构扩展到具有相应有向包围盒参数的叶节点。

本文使用随机梯度下降 (Stochastic Gradient Descent, SGD) 来优化结构重建网络，同时通过结构的反向传播 (Backpropagation Through Time, BPTT) 进行递归神经网络解码器的训练。

4.4 小结

本章采用由浅入深的介绍方式，介绍了 RGB 图像三维重建的一般方法知识，提出本文三维重建方法的优越性。重点介绍本文采用的用有向包围盒表示三维物体的各个部件，用有向包围盒组成的树状结构代表由各个部件组成的三维物体。树状结构在解码过程中的各个节点代表了三维物体各个部件的关系。同时本章针对树状结构这种表示方式详细介绍了递归神经网络的解码过程。最后在对三维物体结构表示方式和递归神经网络解码的理解之上，本章引出本文所提出的三维结构重建网络。本章的描写循循善诱，条理地介绍了三维物体的表示方式以及针对层次结构这种表示方式提出的三维结构重建网络的框架以及具体的实现过程。

第五章　基于 RGB 图像的三维结构重建方法

本章从全局角度分析本文提出的利用深度学习的方法从单张 RGB 图像重建 3D 形状的方法，其中结构信息是指用有向包围盒表示的三维模型各个部件以及部件之间的连接性和对称性等关系。给定一个带有目标物体的 RGB 图像，目标是自动地重建物体各部件的有向包围盒信息以及它们之间的相互关系。本文提出的卷积递归自编码器，可以对 RGB 图像进行结构解析并将其重建为有层次信息的树状结构，并以有向包围盒的形式保存其结构信息。编码器是通过训练目标物体轮廓信息的任务实现的一个多尺度卷积神经网络，从而在各种各样的光照和尺度环境中获取目标物体的结构信息，第三章已详细描述并通过实验验证。解码器通过结构掩膜网络获得的输入图像特征，递归地解码立方体层次结构（即由有向包围盒组成的具有层次结构的树状表示方式），第四章有详细介绍和分析。由于解码器能够显示地恢复包括连接性和对称性的部件关系，因此可以去判断目标物体结构恢复的合理性和通用性。使用由 RGB 图像和立方体结构对组成的训练数据对结构掩膜网络（编码器）和结构重建网络（解码器）进行联合训练。这样的数据对是通过渲染三维模型数据集并且切割三维模型的各个部件组成的。本章强有力地展示了从单张 RGB 图像恢复出三维部件结构信息的结果，达到国际前沿研究水平的效果。

5.1　三维重建网络框架

本文提出一个深度神经网络框架可以实现从单张单视角 RGB 图像重建其三维形状结构的功能。这个 RGB 图像三维重建深度学习模型是一个由两部分组成的自编码器：结构掩膜网络和结构重建网络。结构掩膜网络可以从 RGB 图像中获取目标物体的结构特征，结构重建网络可以根据三维模型各个部件之间的关系递归地恢复目标物体的层次结构。

5.1.1　RGB 图像三维重建深度模型

本文学习了一个深度神经网络可以实现直接从 RGB 图像重建出目标物体的三维形状结构信息。重建后的结构信息可以用来对现有研究中重建的体素效果进行优化，使体素恢复的更完整和全面，也可以用于进行 RGB 图像的高级编辑，甚至实现对重建后的三维形状的结构感知编辑。然而直接从 RGB 图像映射到部件之间的结构是一个很大的挑战。Tulsiani 等人[15] 提出的深度模型将三维体素映射为立方体元的集合。然而，他们的方法并不适用于解决我们的问题，因为他们的体元输出并不含有任何结构信息，即体元之间的关系并没有实现重建。

本文要解决的问题不仅仅包括需要合理地重建几何信息，还有包括部件之间的高层次的组合和相互关系。这就使本文面临以下三个方面的挑战：

1. 不同于形状几何，部件分解和部件之间的关系不能明显地从 RGB 图像中获取。相对于已有研究中从像素到体素的映射学习，从像素到部件的层次结构映射具有很强的不确定性和不适定性。
2. 很多人为构建的三维 CAD 模型包含丰富的细粒度的结构信息，这些复杂的三维结构信息的重建远远困难于根据三维模型的种类对三维形状进行合成。
3. 自然图片经常包含具有干扰性的背景，而且由于不同的光照条件和物体不同的纹理信息，同一个物体在图片中的呈现也有较大的差异。

人类由于具有大量三维模型结构组成的先验信息，能合理的根据 RGB 图像对该三维模型的结构层次进行合理地猜测和推理。人类大脑在处理图像信息，获得目标物体的三维结构信息方面进行了两个阶段的处理，首先是根据看到的图片进行分析，然后根据对三维形状理解的先验知识，合理地推测高层次的三维结构信息。本文受到人脑根据单视角 RGB 图像能够推测出对应三维物体的信息的启发，设计了 RGB 图像三维重建深度学习模型，集成了图像信息处理和三维结构重建两个功能的子网络，结构掩膜网络为了理解和分析 RGB 图像中物体的多尺度结构信息，而三维结构重建网络是为了递归地重建物体的各个部件，并以立方体（有向包围盒）的方式进行呈现。

本文设计的从单视角 RGB 图像重建三维结构信息的深度模型如图 5.1所示。该深度神经网络是一个由两个部分组成的自编码器：结构掩膜网络和结构重建网络。

结构掩膜网络为 RGB 图像中的目标对象阐释多尺度的注意力掩膜，从而以多种形式和尺度来理解和辨别目标物体的形状结构信息。结构掩膜网络被设计为一个多尺度的卷积神经网络，并增加跳跃连接层以保留图像中目标物体形状结构的细节信息，同时剔除与结构无关的背景和输出掩膜图像中的纹理等干扰信息。结构重建网络融合结构掩膜网络获得的 RGB 图像的特征和从原始输入图片中提取的卷积特征，这个融合后的 RGB 图像的特征传入到结构重建网络中的递归神经网络，作为进行三维结构信息解码的根节点。递归神经网络解码器经过训练后可以明确地模拟三维物体中各个部件之间的关系，将融合后的 RGB 特征递归地解码为具有合理空间配置和相互信息的以立方体组成的树状结构呈现的三维模型。

这两个神经网络联合起来共同进行训练，训练数据是 RGB 图像 - 掩膜和有向包围盒 - 结构数据对。数据对中的 RGB 图像可以通过渲染三维 CAD 模型获得，而包围盒可以将给定的三维形状抽象为用立方体（即有向包围盒）表示的各个部件来获得。同时在对网络的训练过程中，本文采用了一些机制来避免产生过拟合

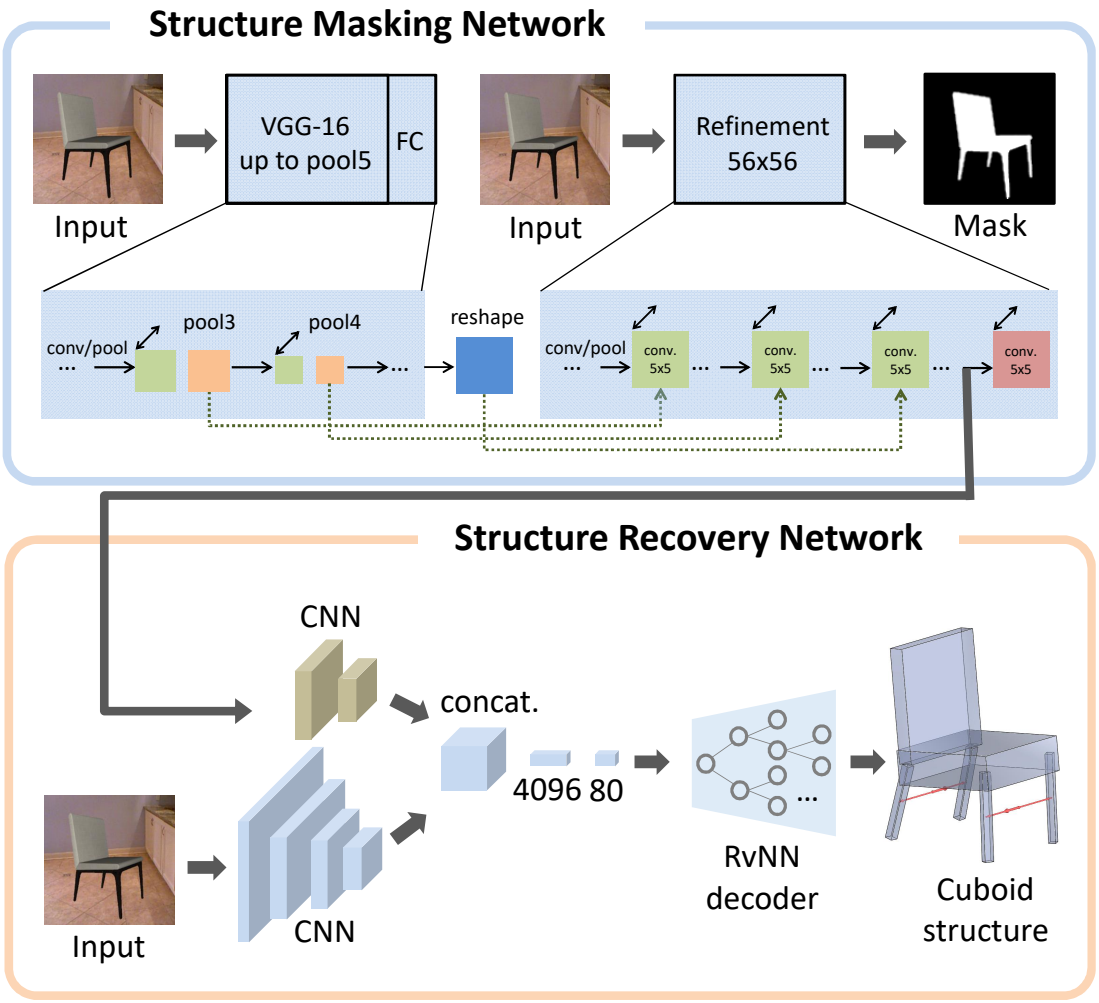

图 5.1　RGB 图像三维重建深度模型框架

的现象。通过实验结果可以看出，本文提出的方法重建效果非常好。

5.1.2　结构掩膜网络

本文提出的结构掩膜网络框架的实现是受到 Li 等人[57] 提出的用来进行细粒度深度估计的多尺度神经网络架构的启发。首先将输入的 RGB 图像统一缩放裁剪为 224×224，具有两条特征提取线的结构掩膜网络输出一个具有输入图像四分之一的二值轮廓掩膜 (56×56)。第一条特征提取线用来获取整张图像的信息，而第二条特征提取线输出一个具有输入图像四分之一细节掩膜特征图。由于结构掩膜网络的预测目标是一个二值掩膜，我们使用 SoftMax Loss 作为该网络的训练损失函数。

本文采用 VGG-16 初始化第一条特征提取线的卷积层（直到 pool5）和随后的两个全连接层，如图 5.1上半部分所示。然后将第一条特征提取下的特征图和网络输出作为输入进入到第二条流水线的不同层次中，目的是优化对输入图像的结构

感知。第二条流水线作为优化模块，首先对输入的 RGB 图像进行一个 9×9 的卷积操作和一个池化操作，然后进行 9 次连续的 5×5 卷积操作，而没有池化操作。从第一条特征提取线中的 pool3、pool4 和最后来自全连接层的输入分别进入到第二条特征提取线中的第四、第六层卷积层。所有的特征图融合操作都经过了跳跃连接层，这个跳跃连接层有一个 5×5 的卷积层和一个 2 倍或者 4 倍的上采样，来匹配第二条特征提取线中的 56×56 的特征图大小，而连接第一条特征提取线中全连接层的跳跃连接层是一个简单的合并操作。通过相关文章的实验结果，可以看出跳跃连接层可以有效地提高从 RGB 图像中获取细节性的结构信息[57]。

5.1.3 结构重建网络

结构重建网络将从结构掩膜网络中提取的特征和从输入图像中提取的特征集成到一个特征向量中，即特征融合过程，并将其递归地解码为三维模型各个部件有向包围盒的层次结构中，即特征解码，如图 5.1下半部分所示。

特征融合。结构掩膜网络融合从两个卷积通道获得的特征。一个通道是将从结构掩膜网络中获得的特征图（结构掩膜预测层的前一层获得的特征图）作为输入，然后跟随两个卷积层和池化层。另一个通道是通过 VGG-16 获取的原始输入图像的卷积特征图。从这两个通道获得的输出特征被拼接为大小为 7×7 的特征图，然后再经过两层全连接层被编码为一个 80D 的代码（或者称为向量），这个 80D 的向量代表从输入的 RGB 图像中学习到的目标物体的结构信息。本文通过实验发现，融合后的特征不仅能够提高结构重建的准确度，还能很好地适应从渲染图像到真实图像的过渡。本文认为产生这种效果的原因是结构掩膜网络提取的特征将物体的形状结构细节信息从诸如纹理变化、背景杂波、光照变化等干扰条件中提取出来。由于结构掩膜网络难以产生非常完美的掩膜预测，原始图像的 CNN 特征保留有更多的对象信息就可以作为结构掩膜网络的补充信息。

结构解码。结构掩膜网络采用递归神经网络作为三维模型部件结构解码器[10]，从根节点开始，递归神经网络将其递归地解码为特征层次，直到到达所有的叶节点，而叶节点可以被进一步解码为一个有向包围盒参数。本文定义的树状结构层次中有三种类型的节点：叶节点、连接节点以及对称节点（包含旋转、平移、镜面对称三种关系）。在解码的过程中，内部节点的部件关系就会被解码为两类：连接关系和对称关系（包括旋转、平移、镜面对称）。因此，每个节点根据它的类型（连接节点、对称节点或者叶节点（代表部件的有向包围盒节点））会被三种解码器 (AdjDec、SymDec、BoxDec) 之一进行解码。AdjDec 解码器会将节点解码为两个有连接关系节点 (80D)；SymDec 解码器会把节点解码为一个节点（80D）和与该节点对应的对称参数包括对称类型 (1D); 旋转或者平移过程中重复的次数 (1D);

用于镜面对称的对称平面，旋转关系的旋转轴，或者平移关系的位置和位移（6D)；BoxDec 解码器将节点解码为一个有向包围盒的参数（12D)。具体的解码操作，本文已在第 4.3.1节中有详细描述。

5.2 三维重建网络实验细节

本文从 ShapeNet 数据集中收集了 800 个三维模型，其中椅子 500 个，桌子 200 个以及飞机 100 个。数据中分为两个子集，用于训练的数据占 70%，用于测试的数据占 30%。本文利用这些三维模型生成 RGB 图像的掩膜图与三维模型层次结构数据，以训练网络并定量评估本文的方法。本文还通过 Google 图片搜索挑战来定性评估本文的方法。定量和定性的评估都证明了本文方法的可行度，可以准确地从单张 RGB 图像重建出三维形状结构。

5.2.1 训练数据对生成

对于每个三维模型，我们围绕该模型创建 36 个渲染视图，每 30° 旋转一次，并以 3 个不同的高度进行。另外，本文还为每个三维模型随机生成 24 个视图，我们为每个形状创建了 60 个渲染的 RGB 图像。本文随机选择 NYU v2 数据集中的背景图像渲染三维模型。对于每个 RGB 图像，可以使用深度值轻松进行渲染，提取前景和背景作为训练数据用于网络训练。

本文数据集中所有的三维模型都是基于其原始网格组件或者 [58] 中提出的对称感知分段软件进行预分割。本文使用对称层次[58] 来表示形状结构信息，同时还定义了形状在各个部件是如何递归地利用对称进行分组组装的。本文采用了 [10] 中的方法为每个类别的形状构造一致层次结构树。具体而言，本文为所有的形状训练了一个无监督的自编码器，其任务是自我重建。在测试阶段，我们使用这个自编码器，用贪婪搜索的方法为每个三维模型进行层次结构信息分组。因此，我们为每个三维模型生成了 60 个图像 - 三维结构层次对。

5.2.2 数据处理和增强

为了进一步增强本文的数据集并减轻过拟合现象，本文基于部件感知控制器[59] 对每一个训练数据进行三维形状结构感知变形[60]，以为训练数据生成一组结构合理的三维形状结构数据。这种基于结构感知的变形保持了原有三维形状部件之间的连接性和对称关系，同时还为每个部件保持了形状纹理。这一步骤的实现是完全自动的，每个变体生成的参数是在给定范围内随机设置的。在本文的实现过程中，为每个三维模型随机生成了 20 个变形，因此将本文的三维模型数据库扩大到 16K。对于输入 RGB 图像（以及与之对应的物体掩膜)，本文采用图像数据增强的常用操作[57]，例如色彩扰动、对比度调整、图像翻转和变换等。

5.2.3 训练细节

RGB 图像的三维重建深度神经网络的训练分为两个阶段。首先单独训练结构掩膜网络以估计输入 RGB 图像的二值掩膜，将结构掩膜网络的第一和第二特征提取线联合进行训练。然后再联合训练结构掩膜网络和结构重建网络，在联合训练的过程中，结构掩膜网络的学习率设置较小。结构重建网络的重建损失用每个有向包围盒重建误差以及每个节点分类器的交叉熵之和来计算。有向包围盒的重建误差用输入和对应输出的有向包围盒参数的平方差来表示。在训练之前，所有的三维模型都会调整为单位有向包围盒，目的是使重建误差在不同形状上相当。图 5.2分别绘制在训练和测试阶段有向包围盒重建损失、对称恢复重建损失以及节点种类的损失值，充分展示了本文实现从 RGB 图像到三维结构重建网络的收敛性。

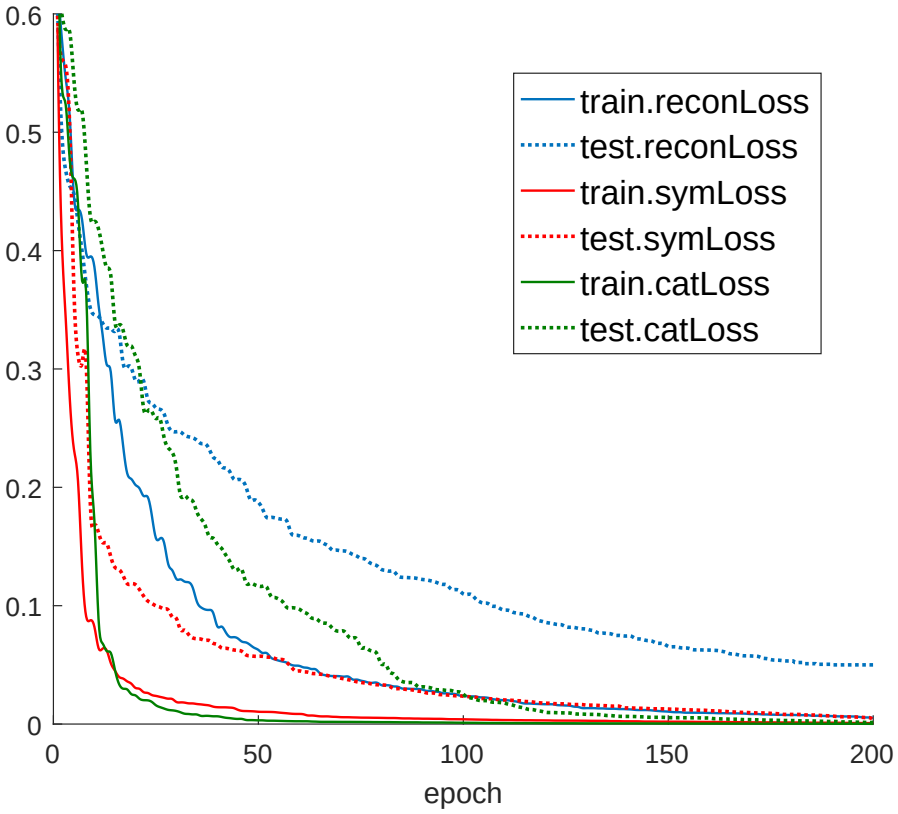

图 5.2 RGB 图像三维重建深度神经网络的收敛性

本文使用随机梯度下降 (Stochasic Gradient Descent, SGD) 方法来优化结构重建网络，通过结构的反向传播 (Back Propagation Trough Time, BPTT) 进行 RvNN 解码器的训练。VGG-16 的卷积层参数初始化为在 ImageNet 上预训练的值。而其它的卷积层、全连接层以及结构重建都是随机初始化的。结构掩膜网络在预训练

的过程中学习率设置为 10^{-4}，在优化的过程中设置为 10^{-5}。在联合训练期间，结构掩膜网络的学习率为 10^{-3}，RvNN 解码器的学习率为 0.2，RvNN 节点分类器的学习率为 0.5。神经网络每训练 50 次，学习效率为前一次的 1/10。本文 RGB 图像三维重建深度神经网络的代码是在 Matlab 的平台上编写的，使用 MatConvNet 这一学习工具进行神经网络的编写和调试。

5.3 三维重建网络实验效果

本文一方面通过不同种类的大量数据对实验进行定量评估，另一方面使用 Google 图像搜索挑战对实验进行定性评估。从结果来看，本文提出的方法可以从单张 RGB 图像中准确地恢复出三维形状的结构信息。

5.3.1 实验结果展示

第 3.3.4节中显示了我们的结构掩膜网络预测对象掩膜的一些结果。从输出中可以看出，复杂的背景信息被成功过滤掉，并捕获了一些对象的细节结构信息。

针对结构重建的 Google 图像搜索挑战。本文首先对结构重建的能力和多功能性进行定性评估。为了进行更多更客观的研究，本文选择使用 Google 图像搜索挑战[14] 进行小规模的压力测试，而不是挑选一些图像进行测试。在测试期间，我们分别使用“chair”，“table”和“airplane”作为关键字在 Google 上执行基于文本的图像搜索。对于每次搜索，本文尝试使用结构重建网络对前八个返回的图像中的每一个重建其三维结构。

结构如图 5.3所示，从结果中，我们可以看到本文的方法能够准确地从真实图像中重建出三维模型的形状结构。更重要的是，本文的方法可以从单视图 RGB 图像中恢复部件之间的对称和连接关系，通过连接和合理的结构实现高质量的结果。镜面对称结构的恢复例子有椅腿和飞机机翼，旋转对称的例子包括转椅的椅腿或者桌子的椅腿。

图 5.3中还有一些重建错误或者失败的例子（标有红框的例子）。被标记的椅子示例并不是由多个部件组成的，因此不能恢复出部件结构信息。当本文的三维形状结构训练数据集（比如被标记的椅子、桌子）中没有见过目标物体的结构时，本文的方法无法重建出合理的结构。

5.3.2 实验结果评估

本文利用测试数据集对本文的方法进行定量评估。对于结构掩膜网络，本文在第 3.3.4节中已经作过评估，而且结果显示本文的多层结构掩膜网络效果非常好。

而对三维结构重建网络，本文采用两种操作进行准确性评估：

图 5.3　针对三维形状结构重建的 Google 图像搜索挑战

1. Hausdorff Error:

$$Error_{Hausdorff} = \frac{1}{2T}\sum_{i}^{T}(D(S_i, S_i^{gt}) + D(S_i^{gt}, S_i)) \tag{5.1}$$

其中 S_i 是一个重建后的三维形状结构（由立方体集进行表示），S_i^{gt} 是 S_i 相对应的真实值。T 是测试数据集中三维模型的数目。$D(S_1, S_2) = \frac{1}{n}\sum_{B_j^1 \in S_1} min_{B_k^2 \in S_2} H(B_j^1, B_k^2)$ 测量的是在形状结构 S_1 中所有立方体到形状结构 S_2 中所有立方体的平均 Huasdorff 距离，其中 B_j^1 和 B_k^2 分别代表在 S_1 和在 S_2 中立方体结构。$H(B^1, B^2) = max_{p \in B^1} min_{q \in B^2} ||p - q||$ 是两个立方体直接的 Huasdorff 距离，p 和 q 是指立方体的顶点。因为 Hausdorff 是非对称的函数，所以需要进行两个方向的计算并且都取平均值。

2. **阈值化比例**

使公式 5.2成立的立方体 B_i 的比例

$$\delta = \frac{H(B_i, B_i^*)}{L(B_i^*)} < threshold \tag{5.2}$$

在重建的三维形状结构 S 中，B_i 是第 i 个立方体，在真实三维形状结构 S^{gt}

表 5.1　不同方法下重建形状结构的准确率比较

方法	Hausdorff Error	阈值化比例	
		δ<0.2	δ<0.1
VGG-16	0.0980	96.8%	67.8%
结构掩膜网络 (VGG-16)	0.0894	97.8%	75.3%
VGG-19	0.0922	96.4%	72.2%
结构掩膜网络 (VGG-19)	0.0846	97.6%	78.5%

中, B_i^* 是 B_i 最邻近的立方体。H 是两个立方体之间的 Hausdorff 距离，与上述定义一致。L 是一个立方体的对角线长度。

本文将一个普通的 VGG-16 作为结构掩膜网络为基准，在表 5.1中，我们基于上述两个评估指标，比较本文的方法和基准方法三维形状结构恢复的准确性，本文还比较了 VGG-16 被 VGG-19 取代的两种方法。结果证明了本文提出的结构掩膜网络在帮助三维形状结构解码方面具有显著效果。这也可以从图 5.4中绘制的重建误差中观察到。更深层次的 VGG-19 也在一定程度上提升了性能。

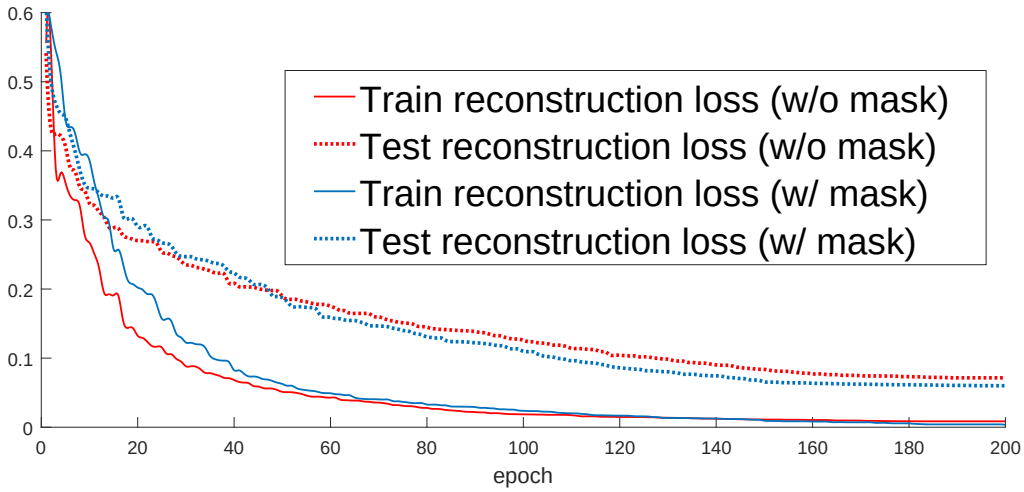

图 5.4　不同结构掩膜网络下的重建损失比较

5.3.3　实验结果对比

图 5.5展示了本文方法和另外两个前沿的方法[15, 54] 的可视化的比较结果，都是实现了从单视角的 RGB 图像重建三维形状的任务。这两种方案都可以实现基于部件的三维形状表示，使其与本文的方法具有可比较性。Huang 等人[54] 的方法通过在数据库查找部件进行组装，从而恢复三维形状，保持它们的对称关系。但是如果数据库不够大，没有包含到与输入图像中目标物体类似的三维形状，那么就会产生很奇怪的结果或者无法产生输出，如图 5.5中第二行的第二个以及第二行的

倒数两个。Tulsiani 等人[15] 的方法产生的体元类似与本文的立方体表示形式，但是并不产生对称关系，而且都是粗粒度的结果。从结果可以看出，由于集成了结构掩膜网络，本文的方法产生的三维部件结构更适应于输入的 RGB 图像，同时由于本文对部件结构关系的解码，结果在结构上也更合理。

图 5.5　不同方法重建结果比较

5.4　小结

本章从网络框架、数据生成、训练细节到实验结果的展示评估和对比等方面对本文提出的 RGB 图像的三维重建深度学习模型进行了全面细致的刻画，很好地将本文的研究成果从大局上进行介绍和分析。本章通过做足够多的实验，与研究前沿的实验结果进行对比，充分说明了本文提出的方法具有很高的研究和应用价值。不论是在计算机图形学还是计算机视觉领域，本文的研究成果都是具有前沿性和极强的应用性的。

第六章　结束语

基于 RGB 图像的三维重建一直是计算机图形学以及计算机视觉领域长期研究的问题。近年来，随着深度学习热潮的来临，利用深度神经网络来解决三维重建问题的研究也越来越多。而且随着互联网和云数据的不断发展，RGB 图像和三维模型数据库也越来越大，数据质量也越来越高，为本文采用深度学习的方式，利用监督学习的学习模式实现从 RGB 图像到三维结构的映射提供了平台。同时通过本文的实验结果可以看到本文的工作取得了非常好的效果，具有很高的应用价值，也为三维重建提供了新的思路和想法。

6.1　工作总结

作者在整个硕士阶段对计算机图形学尤其是从 RGB 图像到三维模型的映射方面抱有极高的研究热情，积极主动地学习和实践相关方面的研究论文和工作。对于本文的研究工作，在被 CVPR2018 收录后，作者整理了相关代码并在互联网上开源，随时准备和有兴趣的研究者交流和探讨，为三维重建领域不断发展贡献自己的力量。基于 RGB 图像的三维重建不仅在图形学领域，在机器人领域、自动驾驶领域以及虚拟现实领域都具有重要的研究意义和应用价值。然而目前前沿的基于 RGB 图像进行三维重建的研究多是基于体素或者点云等表示方式，没有很好地保持三维物体的结构信息。因此本文的文章具有很高的创新性，实现了对重建物体三维结构的恢复，具有很强的前沿性和创新性。

本文依据人类视觉对 RGB 图像重建三维物体的过程，重点分析了对 RGB 图像进行特征提取的过程和基于提取的 RGB 图像特征进行三维结构重建的过程。同时为了重建出三维物体具有的结构信息，本文特别介绍了本文采用的三维物体表示方式以及针对该种表示方式所采用的深度神经网络架构。最后整体介绍了本文提出的基于 RGB 图像三维重建的神经网络框架，以及本文呈现的良好的实验结果、精确的实验评估和与前沿研究对比的效果。因此本文工作可以总结为以下几个方面：

1. **结构掩膜网络**。结构掩膜网络想法是基于人类视觉对 RGB 图像的理解过程所提出的。该网络的框架是基于 VGG-16 实现的，该结构掩膜网络由双通道特征提取线组成，是一个多尺度的卷积神经网络。其中第一条特征提取线是为了获得整张 RGB 图像的信息，而第二条特征提取线输出一个具有输入图像四分之一细节的结构掩膜特征图。同时第一条流水线在一些特定的卷积层之后会将提取到的特征图作为输入传入到第二条特征提取线的不同的卷积层中，目的是为了优化整个网络对输入 RGB 图像的结构感知。

2. **结构重建网络**。结构重建网络是基于从结构掩膜网络中提取的 RGB 图像的特征，然后将该特征解码为三维结构。这部分的创新点主要有两部分，一个是用有向包围盒组成的树状层次结构来表示三维模型，另一种是针对这种树状的层次结构所采用的递归神经网络解码器。两者的配合很好地恢复了三维模型的结构信息。

   - 树状结构表示法。类似于一个句子是由多个单词根据一定的语法规则拼接而成，一个三维物体也是由多个部件根据各自的功能组装而成。根据这样一个想法，本文通过有向包围盒来表示三维物体的各个部件，通过树状结构来表示不同部件之间的组装规则，而这些规则通常是三维物体的对称性（包括平移、镜面对称、旋转）以及它们之间的连接关系，这种表示法很好地体现了三维物体的结构层次，十分具有创新性。

   - 递归神经网络解码器。针对树状结构这种三维表示形式，单纯地使用一般的卷积神经网络是很难进行训练学习的。本文则采用了递归神经网络针对三维物体的结构层次一个节点一个节点地解码，直到到达叶节点。解码过程中判断节点类型选择相对应解码方式。这种解码器很好地对三维物体的树状结构进行解释，成功地将 RGB 图像的特征向量映射为一个具有树状结构的三维模型。

3. **RGB 图像三维重建模型**。RGB 图像三维重建的深度学习模型结合 RGB 图像的特征提取（即结构掩膜网络）和由特征向量重建三维模型的层次结构（即结构重建网络）两部分组成。最后本文进行了大量的实验，包括 Google 图像搜索挑战、多种误差计算方法评估实验结果以及与各种前沿的重建结果进行对比等，充分说明了本文所提出方法的可靠性和前沿性，具有极高的应用价值。

本文工作从上述三个方面进行了详细介绍，而且本文获得了很好的实验结果，具有很强的创新性和研究前景。

6.2 工作展望

硕士毕业论文的撰写过程也是作者回归自己的硕士生涯，总结研究工作的过程，在这个过程中不论结果是否完美，总是会存在一些值得改进的地方。

1. 本文所提出的树状结构表示方法，虽然很好地保持了三维模型的结构层次，能够很好地恢复三维物体的结构信息，但是同时也限制了这种三维重建方法的可移植性。因为这种树状结构与三维模型并存，就会对研究者对部件进行修改或者优化造成困扰。当然这个问题是可以解决的，只要我们找到想要调

整的节点在树状结构中的位置就可以轻易对其进行调整，故我们只需要事先对其进行标记即可。

2. 与此同时，本文的工作针对视角很奇怪的图片不能很好进行恢复。这个方面的改进步骤可以通过进行图像姿态的预测来完成，即对于一个视角很怪的图像，先学习猜测其可能的三维形状，最后在对其进行恢复。

虽然凡事有利有弊，而且这些欠缺的地方都是可以改进和优化的。同时我们更应该学习这个方法的优点，扬长避短。因此在本文工作的基础之上，下一步在博士期间的前进方向有以下两点：

1. 本文所获得的结构信息可以很好地应用于体素或者点云的重建任务中，利用本文结果中的三维结构信息，将残缺的体素或者点云信息进行补全和优化，可以获得带有结构信息的体素或者点云结果。
2. 本文的方法可以应用到高级图像编辑方面。针对图像中三维物体的结构信息进行手动编辑，自然地反馈到图像中的二维目标物体相应的变化中去，这种应用可以在很多领域得到发展，比如虚拟现实方面等。

总而言之，本文的工作具有很强的创新性，下一步可以应用于三维体素补全和高级图像编辑等领域，有着广泛的应用前景。

致　谢

时光荏苒，白驹过隙，两年半的硕士生涯转瞬即逝。回顾这段时光，充实和快乐是它的主旋律，努力和进步是它的伴奏曲。感谢生命中认识的所有人，感谢大家给我带来快乐和知识，感谢学校让我有机会和大家在博士生涯继续共处！

首先，我要感谢我尊敬的导师熊岳山教授。作为一名优秀的教学能手，他不仅在研究中帮我指明了大方向，在生活中也给予了我许多帮助和指导。他学术态度严谨、逻辑思维缜密，他要求在求稳的基础上进行拔高，就像盖房子一定要打好地基一样，这样的人生态度也是我不断追求的高度。“代码一定要练，基础一定要扎实”，这是他经常对我们说的话，也是我需要长期践行的至理箴言。生活中遇到的问题也在熊教授的帮助下得到克服，使我能够顺利迈过生活中遇到的诸多磕绊，健康快乐的成长进步。

其次，我要感谢徐凯师兄和李俊师兄对我长期的帮助，能够容忍我的诸多问题，带领我不断进入学术研究前沿。两位师兄虽然都是学校的老师，但是一旦有代码问题，他们都会亲自帮我解决，就像值得信赖又可靠的兄长，带领我参加他们高端的讨论，让我不断以一种大局整体的角度来看待每一段研究过程和每一步项目进展。感谢他们，让我每天都在进步，每天都在增长新知识。

然后，我要感谢师门的师兄师弟们对我的关心和指导。感谢郑林涛师兄帮我解决的诸多代码问题和生活问题，感谢施逸飞师兄和刘敏师兄对我的帮助，给我传授了丰富的经验。感谢申强强、吴殿元、肖云哲、于洋给我提供了良好的实验室氛围和莫大的帮助，在朝夕相处的日子里，有他们的帮助，我才能够安心舒适地进行每天的学习，共同进步。

再者，我要感谢 505 宿舍（现在的 304 宿舍）的全体成员宋蕊、谢莹和张心言为我提供了良好的生活环境、干净的宿舍环境、安静的睡眠环境。使我能够保持充足的精力投入到每天的科研任务中去，高效地完成每天的任务。感谢好朋友林玉同学带我运动和娱乐，让我在科研间隙不断丰富自己。

最后，我要感谢家人对我的支持和理解，感谢有妹妹照顾父母，使远在异乡的我可以放心学习。感谢男友董一帆对我的支持和提供的后勤保障。有了他陪伴，我的硕士生活充满了乐趣和激情。

总之，感谢身边人的帮助和支持，感谢自己的不断坚持，今后我会继续努力，在科研生涯中做出更大的贡献！

参考文献

[1] Bay H, Ess A, Tuytelaars T, et al. Speeded-up robust features (SURF) [J]. Computer vision and image understanding. 2008, 110 (3): 346–359.

[2] Mikolajczyk K, Schmid C. A performance evaluation of local descriptors [J]. IEEE transactions on pattern analysis and machine intelligence. 2005, 27 (10): 1615–1630.

[3] Krizhevsky A, Sutskever I, Hinton G E. Imagenet classification with deep convolutional neural networks [C]. In Advances in neural information processing systems. 2012: 1097–1105.

[4] Li J, Xu K, Chaudhuri S, et al. Grass: Generative recursive autoencoders for shape structures [J]. ACM Transactions on Graphics (TOG). 2017, 36 (4): 52.

[5] Rusu R B, Cousins S. 3d is here: Point cloud library (pcl) [C]. In Robotics and automation (ICRA), 2011 IEEE International Conference on. 2011: 1–4.

[6] Kalvin A D, Taylor R H. Superfaces: Polygonal mesh simplification with bounded error [J]. IEEE Computer Graphics and Applications. 1996, 16 (3): 64–77.

[7] Eigen D, Fergus R. Predicting Depth, Surface Normals and Semantic Labels with a Common Multi-scale Convolutional Architecture [C]. In IEEE International Conference on Computer Vision. 2015: 2650–2658.

[8] Laina I, Rupprecht C, Belagiannis V, et al. Deeper Depth Prediction with Fully Convolutional Residual Networks [J]. 2016: 239–248.

[9] Chang A X, Funkhouser T, Guibas L, et al. ShapeNet: An Information-Rich 3D Model Repository [J]. Computer Science. 2015.

[10] Choy C B, Xu D, Gwak J Y, et al. 3D-R2N2: A Unified Approach for Single and Multi-view 3D Object Reconstruction [J]. 2016: 628–644.

[11] Fan H, Su H, Guibas L J. A Point Set Generation Network for 3D Object Reconstruction from a Single Image. [C]. In CVPR. 2017: 6.

[12] Kingma D P, Welling M. Auto-encoding variational bayes [J]. arXiv preprint arXiv:1312.6114. 2013.

[13] Goodfellow I, Pouget-Abadie J, Mirza M, et al. Generative adversarial nets [C]. In Advances in neural information processing systems. 2014: 2672–2680.

[14] Girdhar R, Fouhey D F, Rodriguez M, et al. Learning a Predictable and Generative Vector Representation for Objects [J]. 2016: 484–499.

[15] Wu J, Zhang C, Xue T, et al. Learning a Probabilistic Latent Space of Object Shapes via 3D Generative-Adversarial Modeling [J]. 2016.

[16] Zhao R, Wang Y, Martinez A M. A simple, fast and highly-accurate algorithm to recover 3d shape from 2d landmarks on a single image [J]. IEEE transactions on pattern analysis and machine intelligence. 2017.

[17] Xu K, Zheng H, Zhang H, et al. Photo-inspired model-driven 3D object modeling [C]. In ACM Transactions on Graphics (TOG). 2011: 80.

[18] Tulsiani S, Su H, Guibas L J, et al. Learning shape abstractions by assembling volumetric primitives [C]. In Proc. CVPR. 2017.

[19] Lun Z, Gadelha M, Kalogerakis E, et al. 3D shape reconstruction from sketches via multi-view convolutional networks [C]. In 3D Vision (3DV), 2017 International Conference on. 2017: 67–77.

[20] Hassaballah M, Abdelmgeid A A, Alshazly H A. Image features detection, description and matching [C]. In Image Feature Detectors and Descriptors. 2016: 11–45.

[21] Lisin D A, Mattar M A, Blaschko M B, et al. Combining local and global image features for object class recognition [C]. In Computer vision and pattern recognition-workshops, 2005. CVPR workshops. IEEE Computer society conference on. 2005: 47–47.

[22] Ahmed K T, Irtaza A, Iqbal M A. Fusion of local and global features for effective image extraction [J]. Applied Intelligence. 2017, 47 (2): 526–543.

[23] Bianco S, Mazzini D, Pau D P, et al. Local detectors and compact descriptors for visual search: a quantitative comparison [J]. Digital Signal Processing. 2015, 44: 1–13.

[24] Jegou H, Perronnin F, Douze M, et al. Aggregating local image descriptors into compact codes [J]. IEEE transactions on pattern analysis and machine intelligence. 2012, 34 (9): 1704–1716.

[25] Tuytelaars T, Mikolajczyk K, et al. Local invariant feature detectors: a survey [J]. Foundations and trends® in computer graphics and vision. 2008, 3 (3): 177–280.

[26] Chary R, Lakshmi D R, Sunitha K. Feature extraction methods for color image similarity [J]. arXiv preprint arXiv:1204.2336. 2012.

[27] Vojvoda J, Beran V. Feature extraction for efficient image and video segmentation [C]. In Proceedings of the 32nd Spring Conference on Computer Graphics. 2016: 75–80.

[28] Lumb M, Sethi P. Texture Feature Extraction of RGB, HSV, YIQ and Dithered Images using GLCM, Wavelet Decomposition Techniques [J]. International Journal of Computer Applications. 2013, 68 (11).

[29] Lowe D G. Distinctive image features from scale-invariant keypoints [J]. International journal of computer vision. 2004, 60 (2): 91–110.

[30] Jindal R, Vatta S. Sift: scale invariant feature transform [J]. 2010.

[31] Bay H, Tuytelaars T, Van Gool L. Surf: Speeded up robust features [C]. In European conference on computer vision. 2006: 404–417.

[32] Battjes J A. Surf similarity [J]. 1975: 466–480.

[33] Pang Y, Li W, Yuan Y, et al. Fully affine invariant SURF for image matching [J]. Neurocomputing. 2012, 85: 6–10.

[34] Liang Y, Liu L, Xu Y, et al. Multi-task gloh feature selection for human age estimation [C]. In Image Processing (ICIP), 2011 18th IEEE International Conference on. 2011: 565–568.

[35] Dalal N, Triggs B. Histograms of oriented gradients for human detection [C]. In Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on. 2005: 886–893.

[36] Zhu Q, Yeh M-C, Cheng K-T, et al. Fast human detection using a cascade of histograms of oriented gradients [C]. In Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on. 2006: 1491–1498.

[37] Déniz O, Bueno G, Salido J, et al. Face recognition using histograms of oriented gradients [J]. Pattern Recognition Letters. 2011, 32 (12): 1598–1603.

[38] Deng J, Dong W, Socher R, et al. Imagenet: A large-scale hierarchical image database [C]. In Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on. 2009: 248–255.

[39] LeCun Y, Bottou L, Bengio Y, et al. Gradient-based learning applied to document recognition [J]. Proceedings of the IEEE. 1998, 86 (11): 2278–2324.

[40] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition [J]. arXiv preprint arXiv:1409.1556. 2014.

[41] Hinton G E, Srivastava N, Krizhevsky A, et al. Improving neural networks by preventing co-adaptation of feature detectors [J]. arXiv preprint arXiv:1207.0580. 2012.

[42] Maini R, Aggarwal H. A comprehensive review of image enhancement techniques [J]. arXiv preprint arXiv:1003.4053. 2010.

[43] Perona P, Malik J. Scale-space and edge detection using anisotropic diffusion [J]. IEEE Transactions on pattern analysis and machine intelligence. 1990, 12 (7): 629–639.

[44] Morar A, Moldoveanu F, Gröller E. Image segmentation based on active contours without edges [C]. In 2012 IEEE 8th International Conference on Intelligent Computer Communication and Processing. 2012: 213–220.

[45] Zhang Y, Song H, Gu J, et al. Interactive object extraction using hierarchical graph cuts [C]. In Audio Language and Image Processing (ICALIP), 2010 International Conference on. 2010: 851–858.

[46] Cho H S, Bae K, Kyung K-M, et al. Background subtraction based object extraction for time-of-flight sensor [C]. In Consumer Electronics (GCCE), 2013 IEEE 2nd Global Conference on. 2013: 48–49.

[47] Huang C, Liu Q, Li X. Color image segmentation by seeded region growing and region merging [C]. In Fuzzy Systems and Knowledge Discovery (FSKD), 2010 Seventh International Conference on. 2010: 533–536.

[48] Szegedy C, Toshev A, Erhan D. Deep neural networks for object detection [C]. In Advances in neural information processing systems. 2013: 2553–2561.

[49] Hubel D H, Wiesel T N. Receptive fields, binocular interaction and functional architecture in the cat's visual cortex [J]. The Journal of physiology. 1962, 160 (1): 106–154.

[50] Mitra N, Wand M, Zhang H R, et al. Structure-aware shape processing [C]. In SIGGRAPH Asia 2013 Courses. 2013: 1.

[51] Kalogerakis E, Averkiou M, Maji S, et al. 3D shape segmentation with projective convolutional networks [C]. In Proc. CVPR. 2017: 8.

[52] Snavely N, Simon I, Goesele M, et al. Scene reconstruction and visualization from community photo collections [J]. Proceedings of the IEEE. 2010, 98 (8): 1370–1390.

[53] Kalogerakis E, Chaudhuri S, Koller D, et al. A probabilistic model for component-based shape synthesis [J]. ACM Transactions on Graphics (TOG). 2012, 31 (4): 55.

[54] Huang Q, Wang H, Koltun V. Single-view reconstruction via joint analysis of image and shape collections [J]. ACM Transactions on Graphics (TOG). 2015, 34 (4): 87.

[55] Binford I. Visual perception by computer [C]. In IEEE Conference of Systems and Control. 1971.

[56] Socher R, Lin C C, Manning C, et al. Parsing natural scenes and natural language with recursive neural networks [C]. In Proceedings of the 28th international conference on machine learning (ICML-11). 2011: 129–136.

[57] Li J, Klein R, Yao A. A two-streamed network for estimating fine-scaled depth maps from single rgb images [C]. In Proceedings of the 2017 IEEE International Conference on Computer Vision, Venice, Italy. 2017: 22–29.

[58] Wang Y, Xu K, Li J, et al. Symmetry hierarchy of man-made objects [C]. In Computer Graphics Forum. 2011: 287–296.

[59] Zheng Y, Fu H, Cohen-Or D, et al. Component-wise controllers for structure-preserving shape manipulation [C]. In Computer Graphics Forum. 2011: 563–572.

[60] Xu K, Zhang H, Cohen-Or D, et al. Fit and diverse: set evolution for inspiring 3D shape galleries [J]. ACM Transactions on Graphics (TOG). 2012, 31 (4): 57.

作者在学期间取得的学术成果

发表的学术论文

[1] **Niu Chengjie**, Li Jun, Xu Kai. Im2Struct: Recovering 3D Shape Structure from a Single RGB Image. **CVPR 2018**. (**CCF A 类**会议已发表. 计算机学会推荐的 A 类会议收录.)

[2] Gai Wei, Yang Chenglei, Bian Yulong, Shen Chia, Meng Xiangxu, Wang Lu, Liu Juan, Dong Mingda, **Niu Chengjie**, et al. Supporting Easy Physical-to-Virtual Creation of Mobile VR Maze. **CHI 2017**. (**CCF A 类**会议已发表. 计算机学会推荐的 A 类会议收录, 检索号:20181504988153.)

[3] Gai Wei, Yang Chenglei, Dong Mingda, Liu Juan, Dong Yifan, **Niu Chengjie**, et al. UbiMaze: a new genre of virtual reality game based on mobile devices. **MoblieHCI 2016**. (**CCF B 类**会议已发表. EI 收录, 检索号:20164202915560.)

参加的主要科研项目

[1] 国家自然科学基金课题“计算机图形学”（2017-2019，参加成员）

[2] 国家自然科学基金课题“三维场景智能感知与建模”（2017-2019，参加成员）