

東 南 大 學

毕业设计(论文)报告

题 目: 基于深度学习的三维重建和点云生成

学 号: 04216747

姓 名: 周烨凡

学 院: 信息科学与工程学院

专 业: 信息工程

指导教师: 杨绿溪

起止日期: 2020 年 1 月 -2020 年 5 月

东南大学毕业（设计）论文独创性声明

本人声明所呈交的毕业（设计）论文是我个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，除了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或撰写过的研究成果，也不包含为获得东南大学或其它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了谢意。

论文作者签名：_____ 日期：____年____月____日

东南大学毕业（设计）论文使用授权声明

东南大学有权保留本人所送交毕业（设计）论文的复印件和电子文档，可以采用影印、缩印或其他复制手段保存论文。本人电子文档的内容和纸质论文的内容相一致。除在保密期内的保密论文外，允许论文被查阅和借阅，可以公布（包括刊登）论文的全部或部分内容。论文的公布（包括刊登）授权东南大学教务处办理。

论文作者签名：_____ 导师签名：_____

日期：____年____月____日 日期：____年____月____日

摘要

单视角三维物体形状重建 (Single-view 3D Reconstruction) 是三维视觉领域一直以来的核心问题之一。由于拟合非线性方程和学习模式的有效性，深度神经网络被期望在重建三维非线性形状的任务中表现出色。本课题探究了深度神经网络在三维重建任务中的有效性和内在机制，主要的工作和创新如下：

1. 为了验证基于自编码器架构的深度神经网络在三维重建任务中的有效性，提出采用弱标签信息监督对编码器进行初始化，以及用残差层改进解码器这两种方法优化网络架构。在 ShapeNet 公开数据集上进行测试，取得了超过基于识别机制的非深度学习方法的理论极限的水平，并证明了神经网络在此公开数据集上倾向于执行重建机制而非识别机制。
2. 为了衡量三维数据集的聚类程度，提出了基于亲和传播 (affinity propagation) 和轮廓系数 (silhouette score) 的度量指标，该指标得到的量化结果与数据集降维后的可视化结果吻合。
3. 提出并通过实验证明了影响三维重建任务中深度神经网络内在机制的主要因素是数据集的本质特征：只有当训练集中的三维形状集有相比于图片集更高的聚类程度的结构时，在这个数据集上训练的深度神经网络才会执行识别机制而非重建机制。

最后，本文分析了上述结论对三维重建任务中的数据收集和神经网络训练的指导意义。

关键词：三维重建，点云，深度学习，数据挖掘

ABSTRACT

Single-view 3D reconstruction is one of the elementary tasks in the field of 3D vision. Due to the effectiveness in learning and approximating non-linear function, deep neural networks are expected to perform well on the task of reconstructing 3D non-linear shapes. This paper investigates the effectiveness and internal mechanism of deep neural network in 3D reconstruction task. The main contribution and innovation are summarized as follows:

1. In order to verify the effectiveness of the deep neural network based on auto-encoder architecture in 3D reconstruction task, it proposes to leverage supervision of weak class information to initiate the encoder and improve the decoder with residual layer in order to optimize network architecture. These methods are trained and tested on ShapeNet public dataset and they outperform theoretical limit of the non-learning method based on recognition, and it proves that neural network tend to do reconstruction task instead of recognition task in this dataset.
2. Define novel way to measure clustering coefficient of 3D reconstruction dataset based on affinity propagation and silhouette score, the quantitative results of this metric is corresponding to qualitative results of visualization of dataset in low dimension.
3. It claims and proves by experiments that the bias of internal mechanism of network is mainly affected by the intrinsic properties of dataset: only when the training set of 3D shapes has a more clustered structure than images, the deep neural networks trained on this dataset become more likely to perform recognition than reconstruction.

Finally, this paper analyzes the significance of the above conclusions for data collection and neural network training.

KEY WORDS: 3D Reconstruction, Point Cloud, Deep Learning, Data mining

目 录

摘要	I
ABSTRACT	II
目录	III
第一章 绪论	1
1.1. 课题的背景和意义	1
1.2. 研究现状	2
1.2.1. 相关模型	2
1.2.2. 相关数据集	3
1.2.3. 神经网络学习机制	3
1.3. 本文研究内容	4
1.3.1. 课题关键问题以及难点	4
1.4. 论文组织结构	5
第二章 问题与定义	6
2.1. 问题描述	6
2.2. 识别机制和重建机制	7
2.3. 数据集的聚类趋势	8
2.3.1. 亲和传播	8
第三章 模型方法	10
3.1. 基准模型	10
3.1.1. 模型介绍	10
3.1.2. 实验结果与分析	12
3.2. 模型优化	14
3.2.1. 架构优化	14
3.2.2. 实验结果与分析	16
第四章 机制探究	20
4.1. 理论分析	20
4.2. 实验设计与数据集生成	21
4.3. 实验结果	23
4.3.1. 数据集指标测量	23
4.3.2. 距离矩阵的可视化	25
第五章 总结与展望	28
5.1. 工作总结	28
5.2. 工作展望	29
参考文献	30
附录 A 可视化结果	33
A.1. 各模型预测结果可视化	33

附录 B 实验数据与结果	34
B.1. 公开数据集统计	34
致 谢	35

第一章 绪论

1.1 课题的背景和意义

问题 1：在 1.1 节中要说明你研究的基于单视角二维图片进行三维物体形状重建的重要用途，可以举 1-2 个较生动的例子，并配以图形说明。

基于单视角二维图片输入进行三维物体形状重建是三维视觉领域的核心问题之一，它能有效解决现实世界中二维数据丰富而三维数据稀少的问题，满足无人驾驶，智能建造，机器人领域的需求。如图1-1，在智能建造领域的城市三维建模中，大规模应用该技术将高清航拍获取的 RGB 图片数据1-1a 重建为三维立体数据1-1b，结合雷达获取的遥感点云数据1-1c，进行高精度的城市建模1-1e。点云数据具有易处理，易存储，易获得的特点，因此成为三维重建研究中常用的三维数据表示之一。

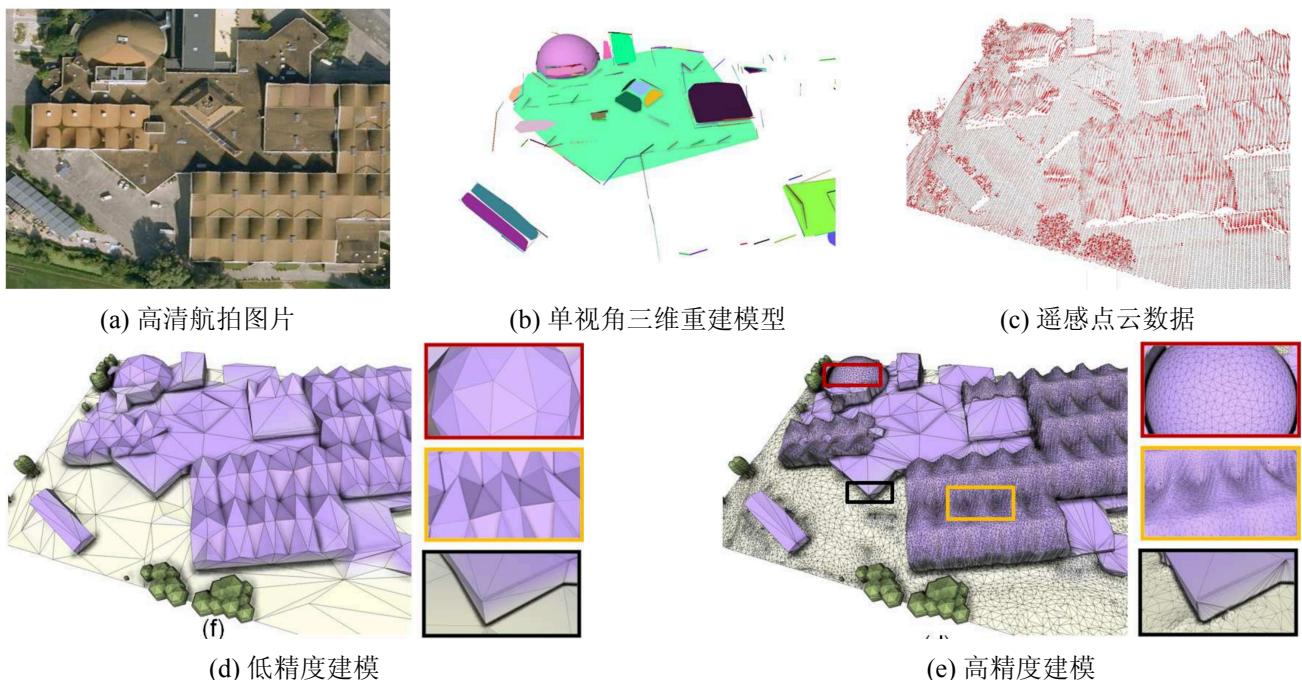


图 1-1 应用单视角三维重建技术的城市建模

使用深度学习来进行单视角三维重建持续受到人们的关注。尽管很多的文章已经展示了创新的深度学习框架来提高三维重建任务中的最高水平^[1-11]，但很少有文章来尝试探究这些任务的本质特性。不可否认的是，重建三维非常规形状的问题已经变成了一个新的机器学习范例，并且从业者使用的处理数据和训练网络的方法与在常规数据上的学习使用的方法不同（比如，Adam^[12] 的使用频率比 SGD 要高）。因此，这引发了研究者对其内在机制的思考，即神经网络在三维重建学习这个新范例上怎样进行学习？是否与传统的向量分类和回归问题不同？

最近，文章^[13] 的作者针对上述问题提出了一个让人惊讶的观点。他们尝试性的展示了当

前最先进的用于三维重建任务的深度神经网络更倾向于通过首先分类输入图片到一个特定的簇，然后生成对应簇的平均三维形状，以此来进行预测。支持这一观点的主要实验证据是这些深度神经网络的三维预测结果与纯粹基于聚类和基于检索的基准模型的三维预测结果效果相近。这是一个非常有趣的观察，因为它说明了对于三维重建的任务来说，深度神经网络倾向于记住平均形状并且将其与图片输入的语义联系起来，而不是使用几何方法来生成一个形状，比如通过融合细粒度的局部结构形成一个整体形状。如果这个观点是正确的，这说明对于三维重建任务来说，最先进的深度神经网络实际执行了一个记忆任务而不是一个泛化任务^[14]。

本课题旨在优化三维重建任务中深度神经网络的架构，证实其有效性。同时进一步探究深度神经网络在三维重建任务中的内在机制，分析导致其偏倚于识别或重建机制的因素，这一研究将会帮助当前应用于三维重建的深度学习方法避免陷入识别机制的误区，同时也探讨了深度学习的本质问题，即如何帮助神经网络进行更好的泛化。

1.2 研究现状

1.2.1 相关模型

问题 2：在 1.2 节要适当介绍传统的经典重建方法有哪些？存在什么问题？你研究的方法目前国际上已有哪些进展？还存在什么问题？所以你想解决它的哪些问题。后面就是 1.3 节，本文的研究内容。

早期进行三维重建的经典方法是基于多目几何的重建比如从运动恢复结构模型^[15] (Structure From Motion)，通过相机的移动获取三维物体各个视角的信息，也有少量基于学习的方法如 Make3D^[16] 使用了马尔科夫随机场 (Markov Random Field) 来将单目线索与图片各部分的关系进行建模，能够重建基础简单的几何机构。文章^[17] 的研究者采用了非刚性从运动恢复结构模型 (NRSfM) 并结合现有的目标检测数据集中的二维标注来重建三维模型。尽管这些工作在三维重建领域做出了从无到有的伟大进展，但是这些方法依然不能具有强鲁棒性的从单张图片重建完整和高水平的形状，他们的方法依然缺少较强的形状先验。

近年来，因为三维数据集规模的增长以及计算能力的提高，使用深度学习方法来完成单视角三维重建受到了广泛关注，且涌现了大量工作。PSGN^[3] 的作者首次提出了采用基于自编码器架构的深度神经网络解决以单张图片作为输入的三维点云重建任务，深度学习首次在重建任务上呈现出了显著的效果，并达到了当时最优水平。但这是一份偏应用的工作，使用了海量的数据进行训练以达到较高的重建水平，数据集的样本量达到了 20 万，使用了 2000 个类，且没有进行与神经网络学习机制相关的讨论。从网络模型上来看，缺少了深度学习一些先进的技巧与架构，仍有提升的空间。

FoldingNet^[6] 的作者提出了以三维点云作为输入的三维点云重建深度自编码器的架构，虽然该模型的输入为三维点云，但是文章却提出了通用的将码字与二维规则点云解码为三维点云形状的基于多层感知器的解码器，并用“折纸”的理论来支持这一网络，根据该思想实现

的解码器仅使用了全连接层解码器的 7% 的参数量，却在重建效果上超过当时的基准模型。令人惊讶的是，同年的 CVPR，还有一篇文章 AltasNet^[18]，提出了从单张图片重建三维网格的深度神经网络模型，且其核心部分解码器架构与 FoldingNet^[6] 解码器的架构思想十分相似，且该模型至今仍保持了当前单视角三维重建的最优水平。这两篇工作都证明了基于“折纸”思想的多层感知器重建三维形状的优越性。

虽然以上工作已经使得三维重建领域逐渐成熟，且在数据丰富的条件下，将三维重建水平提高到了难以超过的高度。但这些都是仅在模型架构上进行研究，以直接提出方法提高重建准确率为目的，且这些方法都缺少对于神经网络进行三维重建内在机制的理论支持。而且，深度学习的核心部分是数据，比如图像领域就有研究者关于用数据增强来提高神经网络能力进行了深入的研究，但当前三维重建领域并没有相关工作探究数据本身对三维重建神经网络的影响，比如开展少量数据训练，或数据增强的研究。

文章 What3D^[13] 中首次提出了三维数据集对重建模型的影响以及对神经网络内在机制的思考，主要观点为深度神经网络实际在任务中执行了识别机制而非重建机制，并根据这一理论提出了基于识别和检索机制的多个非深度学习方法，且这些方法的重建性能都超过了当前的深度学习网络模型。这一工作引发了研究者们对神经网络在三维重建中的有效性以及内在机制的思考与讨论。

针对文章 What3D^[13] 对三维重建神经网络重建能力的置疑，我们首先想验证这一置疑的真实性，即当前先进的深度学习架构 AltasNet^[18] 以及 FoldingNet^[6] 的重建性能是否确实差于 What3D^[13] 提出的基于识别的模型，如果这一置疑属实，是否可能采用模型优化方法来缩小这一差距。在此基础上，对三维重建神经网络的内在机制，尤其是数据对模型的影响机制，提出一些理论层面的理解并用实验证明。研究问题的具体描述展现在 1.3.1。

1.2.2 相关数据集

ShapeNet^[19] 是一个注释丰富且规模较大的三维形状数据集，涵盖 55 个常见的类别，有大约 5 万个样本，每个样本内有一个三维模型和多张从不同视角渲染的该三维模型的图片，三维模型的数据格式为网格、体素，图片格式为 PNG，在进行单视角三维重建时会从多个视角图片中选择一个视角的图片，因此这种情况下每个样本内有一张图片与一个三维模型。该数据集的创立者在 ICCV 2017 举办了基于该数据集的单视角三维重建任务的竞赛，并将当时的基准成绩发表在论文^[20] 中。论文^[13] 提供的数据集在 ShapeNet^[19] 的基础上增加了点云数据格式，每个三维形状由 9000 多个点构成。共有 52430 个样本，涵盖 55 个类。

1.2.3 神经网络学习机制

深度神经网络是执行记忆还是泛化一直是现代机器学习中的主要问题。与我们将要介绍的类似，众所周知的猜测是优化过程是“内容感知”的，并且取决于数据本身的属性^[14]。该论文^[14] 中还显示，训练期间的某些正则化技术可帮助深度神经网络泛化而不是记住任务。对于三维形状重建，^[13] 表明神经网络倾向于记忆平均形状，而不是在几何意义上进行重建。确

实，许多作品还显示了平均形状和识别信息在提高三维重建效果中的有效性^[21-23]。

相比之下，也有很多作品利用三维形状的连续潜在空间中的分布信息^[1, 24-29]来提高三维重建效果，这超出了基于识别的范围。值得注意的是，相关工作^[30, 31]表明形状算术可以在三维形状的潜在空间中进行，从而排除了神经网络仅在此问题设置中执行识别的可能性（因为执行算术需要的不是均值信息离散簇的形状）。其他一些工作建议通过将每个形状分解为部分^[32-34]或通过连续过程^[35, 36]来生成三维形状，这也超出了简单的识别任务。但是，相关论文^[6, 30, 37]的作者将基于三维重建的自动编码视为点云上无监督分类的基础。尽管这些作品中的输入数据是三维形状，而不是二维图像，但是形状信息有助于分类的事实似乎确实增强了这样的概念，即三维重建更着重于识别而不是重建。尽管如此，形状信息有助于识别的事实不能成为判断三维重建网络所学知识的主要理由。有意义的未来方向是研究分别用于无监督分类和受监督三维形状重建的神经网络学习之间的差异，因为这两个方向的主要目标并不完全相同。

注意三维重建问题可以被看作是一个更普适的分布学习^[38, 39]的特殊情况。但是，与分布学习中的理论工作如拓展核方法到回归分布不同，我们的工作集中于深度学习。即便如此，使人感兴趣的是能看到分布学习的传统工作把输出分布当做一个所有训练分布样本的连续线性组合直接处理，而不是使用两步法，预测簇索引后再预测平均分布。

1.3 本文研究内容

1.3.1 课题关键问题以及难点

首先，正式定义本课题的关键问题：一、优化深度神经网络框架，使其在单视角三维重建任务中超过基于识别机制的非深度学习方法，证明其有效性。二、讨论深度神经网络在单视角三维重建任务中执行的是识别机制还是重建机制。初步猜想是影响其在两者之间偏倚的因素为训练数据集的整体特征：聚类程度。因此本课题需要考虑以下几个难点：

1. 改进当前作为基准的神经网络架构以获得更好的重建性能。我们希望能对当前作为基准的深度神经网络进行架构优化，以期望其在公认的标准数据集上能接近并超越基于识别机制的非深度学习方法。因此我们考虑借鉴图像领域成熟且有效的神经网络架构优化方法和训练技巧。
2. 定义神经网络的识别与重建这两种机制的数学表达，并用具体的实验结果来描述。
3. 设计并产生具有量化特征的三维重建数据集。为了探究神经网络训练集和网络的性能之间是否有强相关性，需要能定量的操控数据集的某些整体特征，如聚类系数。可以考虑的方式就是生成自定义的数据集，同时在生成过程中通过采样改变整体特征。为此可以考虑使用计算机图形学的相关软件来合成三维模型，并探究一些能进行不同三维形状之间插值形变的算法。
4. 定义衡量数据集指标的度量标准。在解决问题三后，需要构建一个度量标准来衡量数据

集的聚类趋势，得到量化评分。

1.4 论文组织结构

本论文主要有六个章节，各章的内容安排如下：

第一章 绪论，简要介绍了单视角三维重建任务的研究背景和意义，并对现行的相关理论，模型和数据集进行简单概述与分析。介绍本论文的研究思路、关键问题以及主要贡献。

第二章 问题与定义，介绍了本论文探究的损失函数和算法的理论基础。给出了重建机制与识别机制的数学定义。

第三章 模型方法，详细介绍了基准模型，以及优化模型的架构，并给出了网络在 ShapeNet 大型数据集上训练的过程信息以及训练结果的分析比较。

第四章 机制分析，给出了机制探究的理论基础与实验设计，介绍了自定义数据集的制作，网络在自定义数据集上训练的过程信息以及训练结果的定性与定量分析。

第五章 总结与展望，分析并总结本课题研究成果的总体优缺点，并提出未来的研究方向。

第二章 问题与定义

2.1 问题描述

问题 3：第 2 章的问题描述，不能马上就写一串数学公式，2.1 节就要给出一个从二维到三维重建的形象示意图，或框图，生动地说明你要解决的问题。后面再写数学公式，读者就容易理解了。

我们探究的问题是基于单张图片输入重建一个三维形状。输入 I 是一张二维 RGB 图片，输出 S 是一个点云，如图 2-1。我们在本文中不考虑基于体素的体积表达。对于基于点云的表达，每个形状 S 是一个包含三维点的点集。

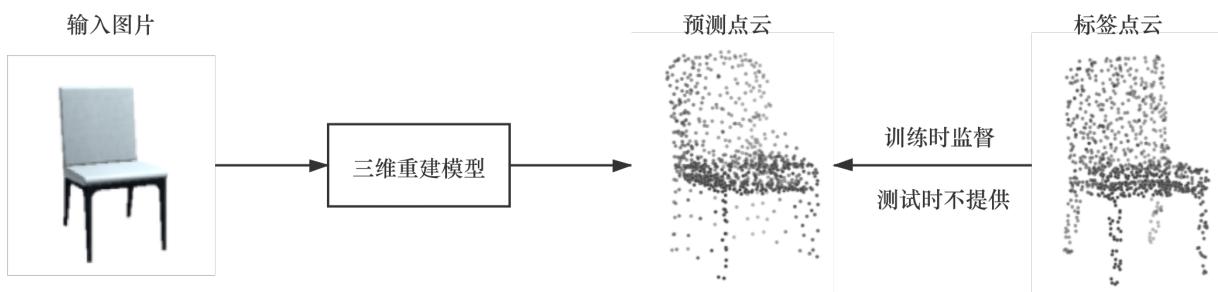


图 2-1 三维重建任务描述

一个神经网络模型从输入图片 I 预测以点云表示的形状 \hat{S} ，在训练中通过减小特定的损失函数定义的经验损失 l ，使得 \hat{S} 的形状接近标签点云 S 。该过程的数学定义为：

$$\min_f \sum_{i=0}^{n-1} l(f(I_i), S_i). \quad (2.1)$$

我们想通过优化模型 f ，使其在训练数据集上训练后，在测试数据集上取得更低的平均损失。同时，我们还想学习 f 的特性来了解它执行的是一个重建任务还是识别任务。两个指标被用来测量预测结果 $\hat{S} = f(I)$ 与标签点云 S 之间的差异，称为 Chamfer Distance 和 F-score。

Chamfer Distance. Chamfer Distane 通过搜索另一个点集中最近点来测试一个点集到另一个点集的整体距离。

$$d_{CH}(S, \hat{S}) = \frac{1}{|S|} \sum_{\mathbf{x} \in S} \min_{\hat{\mathbf{x}} \in \hat{S}} \|\mathbf{x} - \hat{\mathbf{x}}\|_2 + \frac{1}{|\hat{S}|} \sum_{\hat{\mathbf{x}} \in \hat{S}} \min_{\mathbf{x} \in S} \|\hat{\mathbf{x}} - \mathbf{x}\|_2. \quad (2.2)$$

尽管 Chamfer Distance 是运算高效且方便的，它受到很小一部分异常点的影响强烈。所以我们也要如文章^[13]建议的那样采用 F-score 来预测点集。

F-score. 另外一个衡量形状重建的指标是 F-score，该指标是精确度和回忆度的调和平均数。

在给定相应的标签点云 S , 且在固定的距离阈值 d 内, 重建点云 \hat{S} 的精确度 $Prec$ 被定义为:

$$Prec(d, S, \hat{S}) = \frac{1}{|\hat{S}|} \sum_{r \in \hat{S}} [\min_{s \in S} \|r - s\| < d], \quad (2.3)$$

其中 $[\cdot]$ 是艾佛森括号。

相似的, 标签点云 S 对重建点云 \hat{S} 的回忆度 Rec 是:

$$Rec(d, S, \hat{S}) = \frac{1}{|S|} \sum_{s \in S} [\min_{r \in \hat{S}} \|s - r\| < d]. \quad (2.4)$$

其中 $[\cdot]$ 是艾佛森括号。

由 $Prec$ 和 Rec , 如下计算 F-score:

$$F(d, S, \hat{S}) = \frac{2 \times Prec(d, S, \hat{S}) \times Rec(d, S, \hat{S})}{Prec(d, S, \hat{S}) + Rec(d, S, \hat{S})}. \quad (2.5)$$

重建的准确度被精确度量化, 用以测量重建点集和标签点集的距离。重建的完整度被回忆度量化, 用来测量标签点云多少部分被重建点云覆盖。所以, 一个高的 F-score 显示了重建是准确且完整的^[40]。

2.2 识别机制和重建机制

在这个部分, 我们正式定义本文中研究的两种学习范式, 被称为识别 (*recognition*) 和重建 (*reconstruction*)

定义 1 (识别) 一个基于识别机制的神经网络 f 用两步预测形状重建。神经网络重建方程可以写成:

$$\hat{S} = f(I) = f_1(f_2(I)), \quad (2.6)$$

方程中的 $f_2(\cdot)$ 将输入图片映射到一个标量索引, 并且 $f_1(f_2(I))$ 将这个标量索引映射到索引 $f_2(I)$ 对应的特定簇的平均形状。

定义 2 (重建) 一个基于重建机制的神经网络直接进行三维重建。即

$$\hat{S} = f(I), \quad (2.7)$$

并且重建不会明显的使用图片簇的任何信息。因为当前最常见的单视角三维重建经常使用一个“编码器-解码器”结构, 所以一个相似的概念是由编码器获得的码字不会形成簇。

主要问题

在这篇论文中, 我们研究神经网络执行识别机制还是重建机制。我们展示了在两者之间偏倚的趋势由数据集特性决定。

值得注意的是论文^[13]的主要结论是单视角三维重建中的深度神经网络主要执行识别工作。换句话说，该论文的观点是神经网络的方程更接近于定义1而不是定义2。

2.3 数据集的聚类趋势

在这个部分，我们定义用来测量数据集聚类趋势的指标。具体来说，我们使用 silhouette score^[41] 来测量聚类趋势。给定一个数据集 $D = \{x_i\}_{i=0}^{N-1}$ 和一个随机距离 $d(x, y)$ 方程¹。我们能通过明确一个聚类特性方程 $C(\cdot)$ 来确定一个数据集的簇。对于每个样本 x_i ，聚类方程会给出一个聚类标签 $C(x_i)$ 。我们使用 $C(x_i)$ 来指示包含有 x_i 的簇，比如说，聚类标签 C_i 等价于 $C(x_i)$ 。有时候，数据集已经包含了标签聚簇划分。更多情况下，聚簇划分需要通过一个算法来获得。然后，第 i 个样本的轮廓系数（silhouette score）被定义为：

$$s(i) = \frac{b(i) - a(i)}{\max(a(i), b(i))}, \quad (2.8)$$

其中

$$a(i) = \frac{1}{|C_i| - 1} \sum_{j \in C_i, j \neq i} d(i, j), \quad b(i) = \min_{k \neq i} \frac{1}{|C_k|} \sum_{j \in C_k} d(i, j). \quad (2.9)$$

聚类趋势，以轮廓系数定义，由以下方程得出：

$$\text{Clustering-tendency} = \frac{1}{N} \sum_i s(i). \quad (2.10)$$

对于没有标签聚类的数据集，我们需要找到聚类方程 $C(\cdot)$ 。在这篇论文中，我们使用亲和传播（affinity propagation）^[42] 来进行聚类。亲和传播非常适合我们的设定，因为它不需要预先确定簇的数量，并且它使用一个预设的距离矩阵来分配数据点 x_i 到一个簇 C_i 。我们使用 Chamfer Distance(2.2)作为距离方程 $d(x, y)$ 来处理数据集中的无序点云。

2.3.1 亲和传播

亲和传播（Affinity propagation）是一个基于数据点消息传播理论的聚类算法。亲和传播在运行算法之前不需要预先设定簇的数量，而是通过反复迭代从输入数据集中选出“示例（exemplars）”来代表一个簇。

给定一组数据点从 x_1 到 x_n ，先根据预先定义的距离方程计算距离矩阵，再将距离矩阵取负后转化为相似度矩阵（Similarity matrix）。在这个基础定义处，我们使用欧式距离作为距离方程，相似度矩阵可以被写为：

$$s(i, k) = -\|x_i - x_k\|^2 \quad (2.11)$$

亲和传播以相似度矩阵作为输入进行无监督训练，在运行过程中我们需要考虑这几个矩

¹ 我们考虑满足下面三个属性的任何距离方程， $d(x_1, x_2) = d(x_2, x_1)$, $d(x_1, x_2) \geq 0$, 和 $d(x, x) = 0$ 。在我们的实验中，我们使用 Chamfer Distance 作为点云的距离指标。我们使用 ℓ_1 距离作为图片之间的距离。

阵：责任矩阵（Responsibility matrix），获取度矩阵（Availability matrix），规则矩阵（Criterion matrix）。算法在运行过程中对这几个矩阵进行如下更新：

首先责任矩阵 $r(i, k)$ 量化了第 k 个元素适合成为第 i 个元素的示例的程度，并考虑了最近的竞争者第 k' 个元素可能成为第 i 个元素的示例这个因素。

$$r(i, k) \leftarrow s(i, k) - \max_{k' \neq k} \{a(i, k') + s(i, k')\} \quad (2.12)$$

接着获取度矩阵的非对角线元素 $a(i, k)$ 量化了第 i 个元素选择第 k 个元素作为示例的合适程度，考虑了其他元素支持 k 成为一个示例的程度。

$$a(i, k) \leftarrow \min \left(0, r(k, k) + \sum_{i' \notin \{i, k\}} \max(0, r(i', k)) \right) \text{ for } i \neq k \quad (2.13)$$

获取度矩阵的对角线元素 $a(k, k)$ 反映了第 k 个元素适合成为一个示例的累积证据，基于 k 对于其他元素的积极责任。

$$a(k, k) \leftarrow \sum_{i' \neq k} \max(0, r(i', k)) \quad (2.14)$$

责任矩阵与获取度矩阵被运算 2.12, 2.13, 2.14 迭代更新，这个过程也许在一定的迭代步骤之后停止，也许在这两个矩阵的值收敛后停止。停止后，我们计算规则矩阵 $c(i, k)$ ：

$$c(i, k) \leftarrow r(i, k) + a(i, k) \quad (2.15)$$

每行中具有最高标准值的元素将被指定为示例。与该示例的标准值相同的行对应的元素被划分为同一集群。

第三章 模型方法

在这个部分，我们介绍了本文研究的三维重建任务中深度神经网络基准模型和非深度学习基准模型，并结合在 ShapeNet 上的实验结果对其优缺点进行分析。基于这些分析，对深度神经网络基准模型进行优化，使其超过基于检索机制的非深度学习基准模型，证实了深度学习的优越性，我们使用了这两种方法：

1. 引入弱标签信息进行多任务训练
2. 使用残差层改进解码器结构

我们还展示了在真实数据集 ShapeNet 上，即使是一个标准的三维重建深度神经网络也倾向于执行重建任务而不是识别任务。我们的结论基于两点观察：

1. 标准的三维重建网络获得的重建结果在量化指标如 Chamfer Distance 和 F-score 上超过了基于识别机制的分类器的理论极限 (Oracle-Nearest-Neighbor)。
2. 由自编码训练获得的高维码字的二维 T-SNE 可视化，和使用标签信息进行训练相比，并没有展示出明显的形成簇的趋势。而且，没有标签的重建效果比有标签的重建效果要更好。

在这个章节，我们使用的数据集均为 ShapeNet，共计 55 个类，52430 个样本，训练集 / 验证集 / 测试集按照 70% / 10% / 20% 随机采样划分。图片大小为 224×224 ，每个三维点云有 1024 个三维点构成。

3.1 基准模型

3.1.1 模型介绍

3.1.1.1 非深度学习方法

论文^[13] 中提出了以下三个基于识别机制的非深度学习模型：

- *Clustering*: 通过对样本点云集使用聚类算法，将一组点云集划分为多个集群，每个集群内部计算一个平均三维形状，接着训练一个基于输入图片预测特定集群的分类器，将集群的平均三维形状作为预测结果。
- *Retrieval*: 借鉴了现有的基于图片检索对应物体的三维形状的方法^[43]，在测试时，根据输入图片检索训练集中对应的三维点云，直接提取出来作为预测结果。
- *Oracle-Nearest-Neighbor*¹: 该模型在预测时，在训练集中直接搜索与标签点云损失最小的训练点云，将其作为预测结果。因为该模型在进行搜索时需要提供不应该获得的标签点云，而实际测试时只有对应的二维图片作为输入，所以在实践中是不可能应用的。

¹下文缩写为 Oracle-NN

因为论文^[13]的实验结果表明 Oracle-NN 的水平超过了所有当前最先进的深度学习模型，而且它是表征任何实际基于识别（检索）机制的非深度学习方法的性能极限的理论基准。所以我们将其当做本课题的非深度学习基准模型，并进行如算法1的实现。我们发现在将该模型应用到 ShapeNet 时，因为测试集有 10000 个样本，训练集有 35000 个样本，如果直接应用，复杂度很高，需要在 GPU 上消耗大概一周时间，占用了大量的计算资源。于是考虑利用数据集自带的类标签，共计 55 个类，如车，床，椅子等，在每个类内数据集中运行该算法1，大约 10 小时完成。

算法 1 Oracle Nearest Neighbor

输入: $T_{i=0}^n$ 测试集, $D_{j=0}^t$ 训练集

输出: $R_{k=0}^n$ 预测结果

```

1:  $i \leftarrow 0$ 
2:  $k \leftarrow 0$ 
3: while  $i < n$  do
4:    $index \leftarrow 0$ 
5:    $mindis \leftarrow inf$ 
6:    $j \leftarrow 0$ 
7:   while  $j < t$  do
8:      $dis \leftarrow Distance(T_i, D_j)$ 
9:     if  $dis < mindis$  then
10:       $index \leftarrow j$ 
11:       $mindis \leftarrow dis$ 
12:    end if
13:     $j \leftarrow j + 1$ 
14:  end while
15:   $R_{k=i} \leftarrow D_{j=index}$ 
16:   $i \leftarrow i + 1$ 
17: end while
```

3.1.1.2 深度学习方法

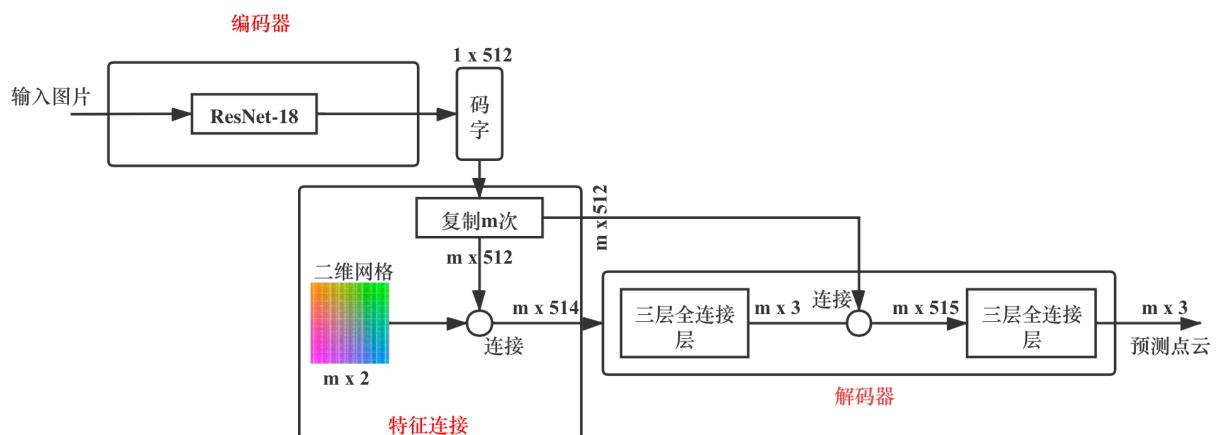


图 3-1 基于自编码器架构的深度神经网络基准模型

根据论文 PSGN^[3] 提出的基于自编码器架构¹的深度神经网络模型，我们进行了如图3-1的实现。网络的主要架构如下：

- 编码器：采用 ResNet-18^[44] 的卷积层架构作为接受二维图像输入并提取图像特征的编码器，具体来说，移除 ResNet-18 最后一层用来分类的全连接层，将一维特征层直接输出作为码字。
- 特征连接：模型中存储有一个预先初始化好的二维正方形网格点云，长宽在 [-1,1] 之间，由 $m \times 2$ 的矩阵表示。在将码字输入解码器之前，我们先将其复制 m 次，然后将 $m \times 512$ 矩阵（码字）与 $m \times 2$ 矩阵（二维网格）连接起来，连接的结果是大小为 $m \times 514$ 的矩阵。
- 解码器：采用 FoldingNet 解码器^[6] 作为从码字恢复到点云结构的解码器，具体来说，这个解码器由两个感知器构成，每个感知器由三层全连接层构成。

该基准模型的关键是采用了 FoldingNet^[6] 提出的解码器架构，使用两个连续的 3 层感知器将固定的 2D 网格扭曲为输入点云的形状。这个过程称为折叠操作，实质上形成了通用的二维到三维映射。为了直观地解释为什么这种折叠操作是通用的二维到三维的映射，用矩阵 U 表示输入的二维网格点。 U 的每一行都是一个二维网格点。用 U_i 表示 U 的第 i 行，用 θ 表示编码器输出的码字。然后，在特征连接之后，解码器的输入矩阵的第 i 行为 $[u_i, \theta]$ ，由于感知器并行应用于输入矩阵的每一行，因此输出矩阵的第 i 行可以写成 $f([u_i, \theta])$ ，其中 f 表示感知器进行的功能。可以将该函数视为带有码字 θ 的参数化高维函数，码字 θ 是指导功能结构（折叠操作）的参数。由于多层感知器擅长拟合非线性函数，因此它们可以在二维网格上执行精细的折叠操作。本文使用的解码器架构包含有两个感知器模块，第一个模块将二维网格折叠到三维空间，第二个模块在三维空间内继续折叠网格，使其接近标签形状。

可以考虑增加深度神经基准模型的预测结果可视化

3.1.2 实验结果与分析

在这个部分，我们展示了上述的深度学习基准模型和非深度学习基准模型在 ShapeNet 上的实验效果，并指出深度学习基准模型的问题。

对于非深度学习基准模型 Oracle-NN，我们展示了所有重建样本的损失平均值作为量化的评估结果。对于基于自编码器架构的深度学习基准模型，我们以 Chamfer Distance² 作为损失函数，使用 Adam 优化器，训练了 100 个周期，使其达到收敛，并对超参数进行调参，总共进行了 8 次训练，最终展示了由所有重建样本的损失平均值表示的最优重建效果。两个模型的最优结果展示在表3.1。从这个结果中我们可以看出，在最优平均损失方面，Oracle-NN 优于深度神经网络，这一结果与论文^[13] 中展示的两者之间的差距相吻合，而且正如前文所展示的，Oracle-NN 是一个理论上的基于识别机制的基准模型，无法应用于实际任务，所以这一

¹实际为“编码器-解码器”架构，因为输入输出分别是同一个物体的图片与三维点云形状所以简称为自编码器架构

²Chamfer Distance 数值越小说明重建效果越好

优越结果是被预计的。但是，在训练的过程中，我们发现了神经网络基准模型暴露出的问题，这意味着有可能通过一些架构优化或者训练技巧来继续缩小这一差距。结合图3-2展示的在不同学习率下深度神经网络的两次训练过程的损失变化¹，对这些问题的分析如下：

- 神经网络训练状态不稳定，对初始化要求较高。首先可以看出，在使用 Adam 优化器的情况下，学习率从 0.003 到 0.01 仅增大了三倍，就使得网络陷入了与最优效果相差较大的局部最优，且 Adam 优化器也没有能使得网络跳出这个局部最优点，如图3-2蓝色线。在正常情况下，网络不应该对超参数有如此高的敏感程度。这也说明重建任务对神经网络参数初始化有比较高的要求。
- 当前深度神经网络架构学习能力弱，一方面解码器较浅的网络深度可能不足以完成当前复杂的任务。另一方面可能出现了梯度消失的问题，如图3-2b橙色线所示，网络在第 20 周期到第 40 个周期之间停止了损失更新，且收敛后的损失仍处于较高水平，这说明当前的网络可能出现了训练过程中梯度消失的问题，导致网络权重难以更新。
- 神经网络的编码器缺少显性的训练指导。从理论角度来说，网络损失函数仅描述两个三维点云的相似程度，因此有理由怀疑该函数传播的梯度并不能给予对二维图片进行处理的编码器足够的优化指导，使其提取需要的图片特征。从实验角度来说，我们将编码器输出的码字进行了二维 T-SNE 降维可视化，如图3-8b，可以看出编码器提取的码字并不具有区分图像类别的特征，虽然这不能证明编码器完全没有提取有助于三维重建的图像信息。

基于上述对神经网络基准模型的问题分析，我们将在下文对网络初始化和梯度消失等问题提出优化方案。

表 3.1 基准模型的评估结果

模型	Chamfer Distance
Oracle-NN	0.0719
自编码器（基准模型）	0.1179

¹其他所有超参数如权重损失，动量等均相同

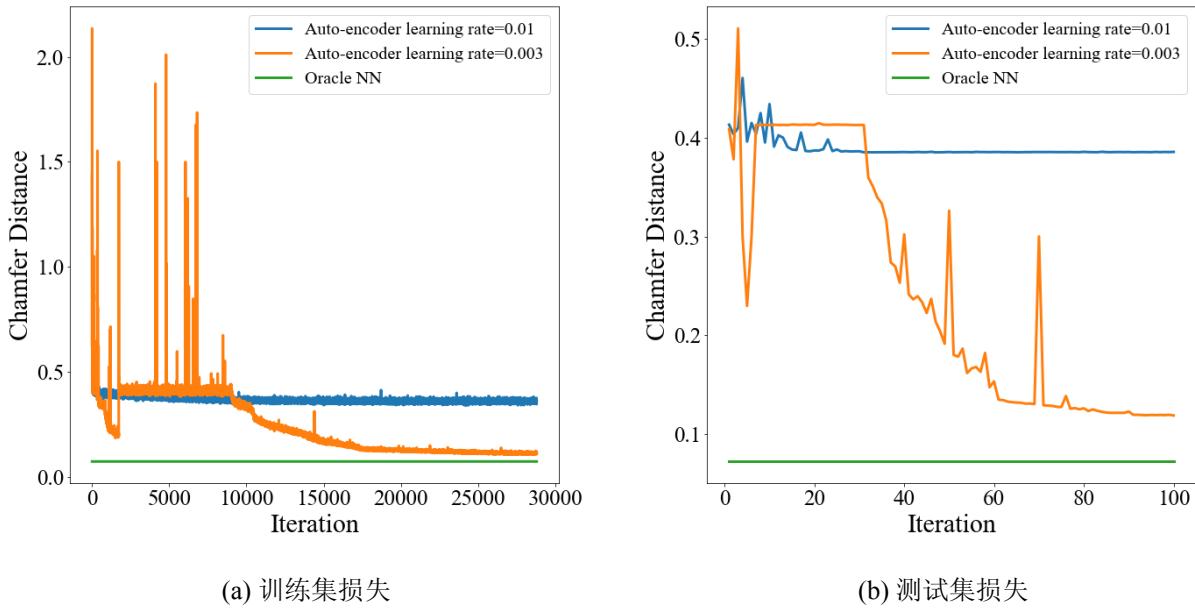


图 3-2 深度神经网络基准模型的两次训练过程。蓝色与橙色指同一网络在不同优化器学习率下的训练，绿色指非深度学习基准模型的量化结果

3.2 模型优化

3.2.1 架构优化

在这个部分，我们介绍用残差层改进解码器以及引入弱标签信息进行多任务训练这两种方法的实现细节以及使用这两种方法的原因。

引入弱标签信息进行多任务训练

针对如何给网络参数提供良好的初始化，以及给编码器显性优化指导的问题，首先考虑图像深度学习领域常用的方法，即使用预训练的网络参数进行网络初始化，具体来说，某任务需要使用神经网络进行图片分类前，会先把网络在 ImageNet 上训练到收敛后，再把网络在任务数据集上训练进行微调 (fine-tuning)，我们没有采用这种预训练参数进行初始化的方法。一是因为 ShapeNet 的图片是由三维网格渲染而成的合成图片，ImageNet 等图片分类数据集绝大部分图片来自现实世界图片，这两个数据域的分布不同，因此需要考虑域适应的问题。二是如果在 ShapeNet 上预先训练图像分类网络，增加了三维重建的所需的步骤（变成了训练两个网络）和所需的计算资源。

基于上述的常用方法的启发，以及需要对编码器进行明显的训练指导的需求，我们提出加入弱标签信息进行多任务训练以达到初始化的目的。在训练的前 5 个周期，提供类标签给网络，在预测重建点云的同时将码字从中间层抽出进行图像分类训练，如图3-3。在第 6 个周期以及之后取消这图片分类训练，只进行三维点云重建训练。关于多任务训练，我们原本有两个选择，一是图片分类，二是图片重建，增加这两种训练损失的任一种从理论上来说都能指导编码器提取图片的某一种特征，比如分类任务会更关注于全局特征，而重建任务会关注

局部特征，但是经过实验表明增加分类损失对三维重建的效果提高比增加图片重建损失更加明显，所以我们选择加入分类任务来进行编码器初始化。

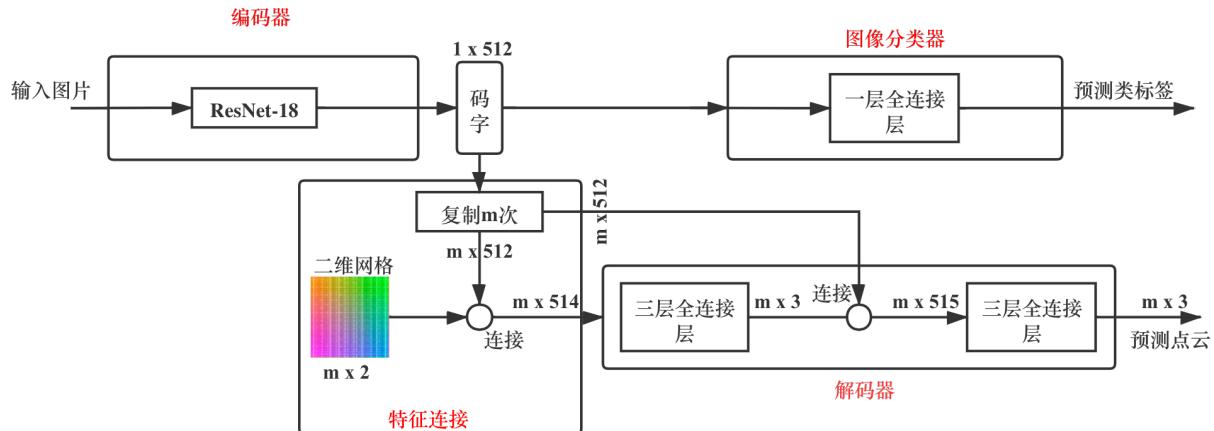


图 3-3 引入弱标签信息进行初始化

使用残差层改进解码器结构

针对深度神经网络梯度消失的问题，一般的解决方法有：1) 使用其他激活函数，如 ReLU 等；2) 批归一化；3) 优化网络权重的初始化方式，如 xavier 的参数初始化方法。4) 调整网络结构，使用残差网络结构。

因为我们在基准模型中已经使用了 ReLU 作为激活函数，使用 xavier 初始化全连接层，所以我们主要使用第四种方法，即调整网络结构，辅助以批归一化。

具体的实现方法为用残差层改进基准模型3-1的译码器，将三层全连接层感知器改为由三个残差块组成的感知器，如图3-4。且残差块内部全连接层之间进行批处理归一化，改进后译码器的参数量是基准模型译码器的两倍，保留译码器之前编码器和特征连接结构不变。残差层对于解码器的优化在于，通过引入“捷径连接”的方式，一方面在反向传播时，增大梯度的绝对值，另一方面改变了学习目标，不再学习一个完整的输出，而是学习一个恒等映射，将网络部分的残差结果逼近于 0，降低了网络学习的难度。批处理归一化通过将每个隐层神经元的输入分布强制拉回到均值为 0 方差为 1 的标准正态分布，使得非线性变换函数的输入值落入对输入比较敏感的区域，以此避免梯度消失的问题。

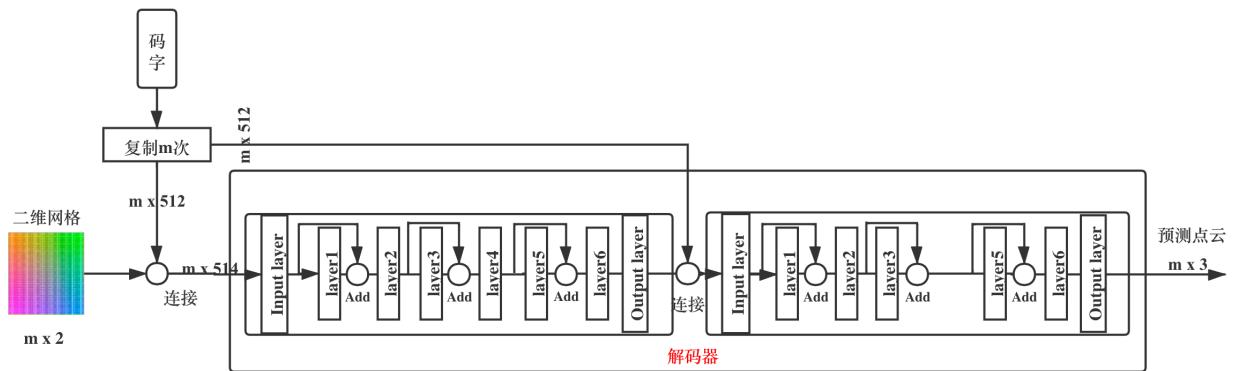


图 3-4 用残差层改进译码器

3.2.2 实验结果与分析

在这个部分，我们展示了将上述两种方法分别应用于优化网络，会在 Chamfer Distance 和 F-score 两个指标上取得接近甚至超过 Oracle-NN 的重建效果。同时，我们进一步对深度神经网络的内在机制进行探究，展示了一个标准的三维重建深度神经网络在公开数据集 ShapeNet 上倾向于执行重建任务而不是识别任务。

重建效果

表3.2汇报了基准模型，两种优化模型以及 Oracle-NN 的最优平均重建损失 (Chamfer Distance)，数值越小说明重建效果越好，我们可以看到加入弱标签信息进行初始化后，基于自编码器的深度神经网络的重建水平 (0.0716) 优于 Oracle-NN (0.0719)，使用残差层解码器后，神经网络的水平 (0.0768) 也得到了显著提高。特别的，我们将两个优化模型和 Oracle-NN 的 Chamfer Distance 表示在图3-5a，从该图可以直观的看出，标准的基于自编码器的深度神经网络模型与 Oracle-NN 呈现出近乎相同的重建效果，且 Oracle-NN (基于识别机制的非学习模型的理论极限) 在测试时需要不应获得的标签信息。

除了 Chamfer Distance 这个指标外，我们还使用了 F-score 这个指标，将测试结果展现 在3-5b，这个指标下的模型比较结果与 Chamfer Distance 的比较一致。因此，我们可以确定的得出结论：弱标签初始化与残差层改进都成功提高了深度神经网络在 ShapeNet 整体数据集的平均重建效果，前者上超过了基于识别机制的非学习模型的理论极限，后者接近了这一理论极限。

表 3.2 优化模型的评估结果

模型	Chamfer Distance
自编码器 (基准模型)	0.1179
自编码器 (弱标签信息)	0.0716
自编码器 (残差层译码器)	0.0768
Oracle-NN	0.0719

除了平均重建效果，我们还考虑了每个类的重建结果的统计特性，如图3-6。我们可以看到基于自编码器的三维重建神经网络在不同的类的重建效果上下波动较小，近似一致。而

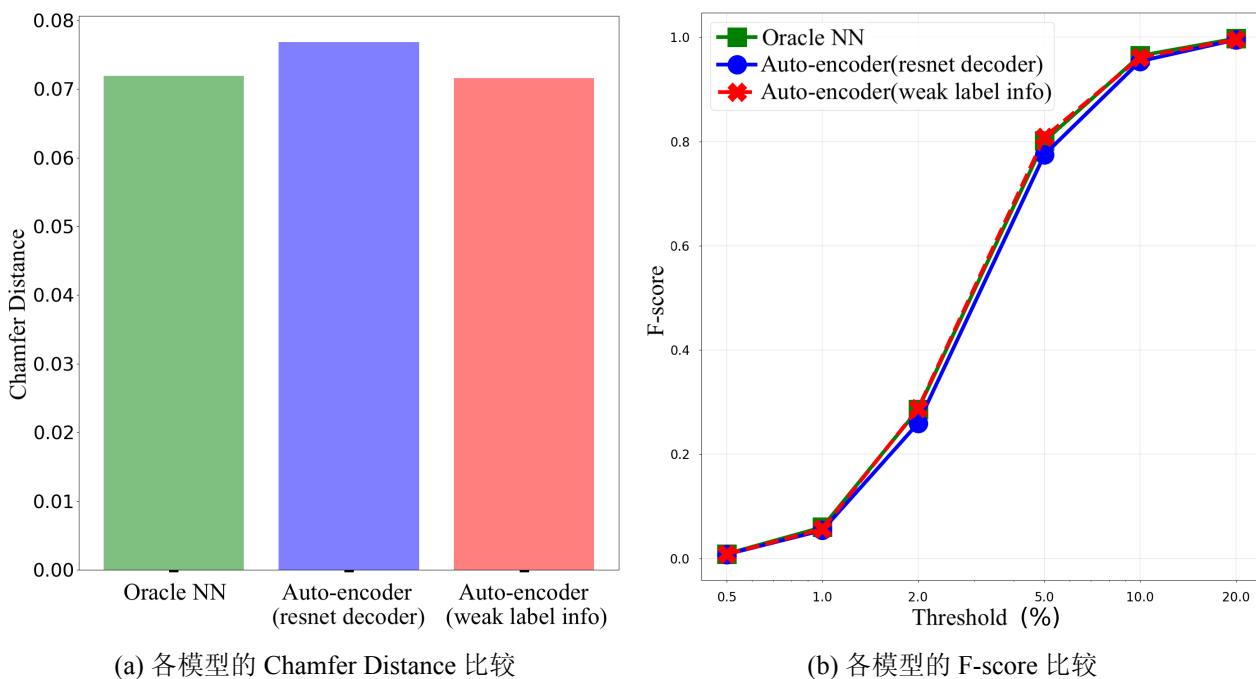


图 3-5 优化深度学习模型与非深度学习模型各指标比较

Oracle-NN 在一些类别如 tower，呈现出非常差的重建效果，出现了各个类重建效果不一致的情况。这个问题同样也在模型预测结果可视化中呈现出来，如图 3-7第四行，我们可以看到 Oracle-NN 预测出了在三维形状层面上非常不合理的点云。结合附录B-1展示的 ShapeNet 数据集的分布，我们可以看到有一些类的样本量很少，Oracle-NN 在样本量少的类内进行搜索时，有很大概率无法找到与预测点云相近的训练点云，因此会在这些类出现比较差的重建效果。

图3-7展示了各个模型预测结果的可视化，从这些定性结果可以看出，在不同物体类别之间，标准自动编码器的重建性能比 Oracle-NN 更稳定。更多的可视化结果提供在附录A-1中。

综上所述，我们可以看出基于识别机制的非深度学习模型的局限性，此类模型缺少对训练数据集分布的鲁棒性，当训练集无法在低维度层面解释测试集需要的形状特征时（即当训练集中没有和测试样本在损失距离上相近的点云时），该类模型的方法效果远差于深度学习模型。

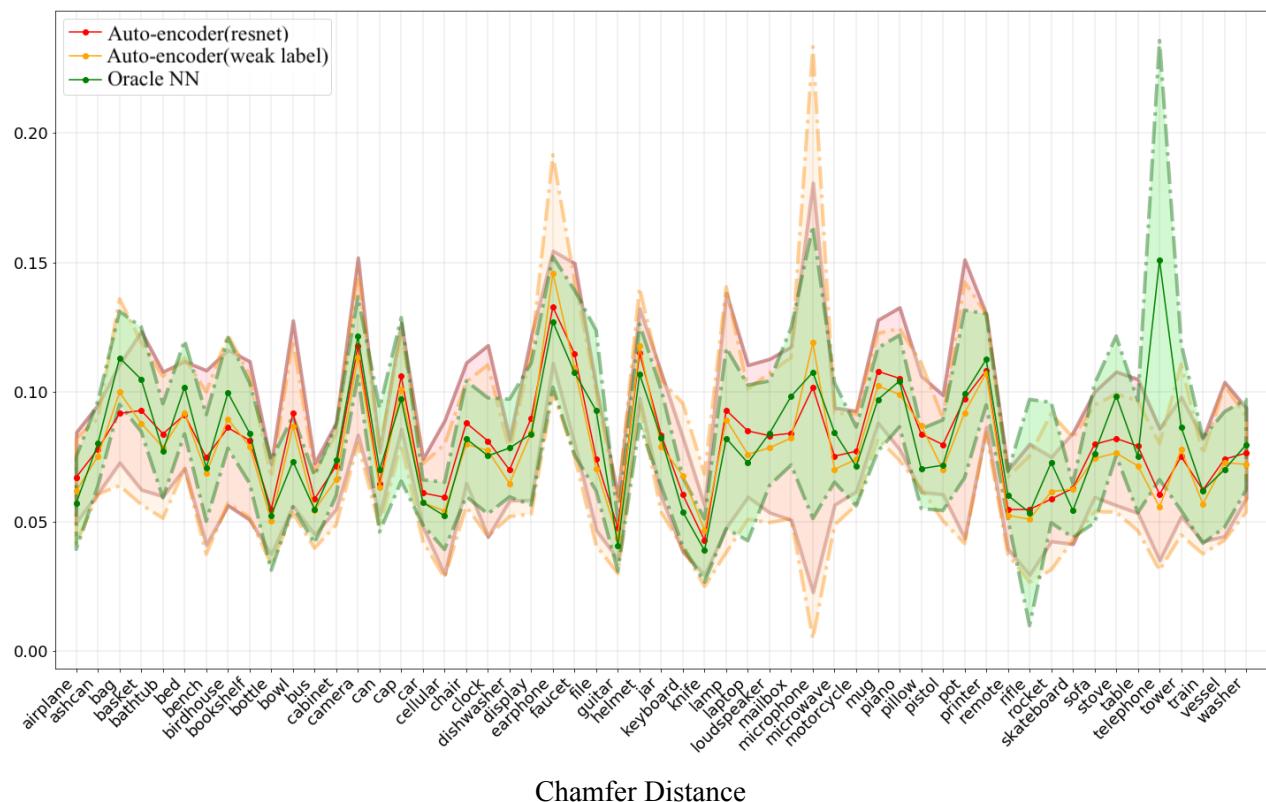


图 3-6 每个类的重建评价指标，X 轴每一列代表一个类的 Chamfer Distance 统计。实心点代表类内所有样本指标平均值，上边沿虚线代表类内所有样本指标最大值，下边沿虚线代表类内所有样本指标最小值。

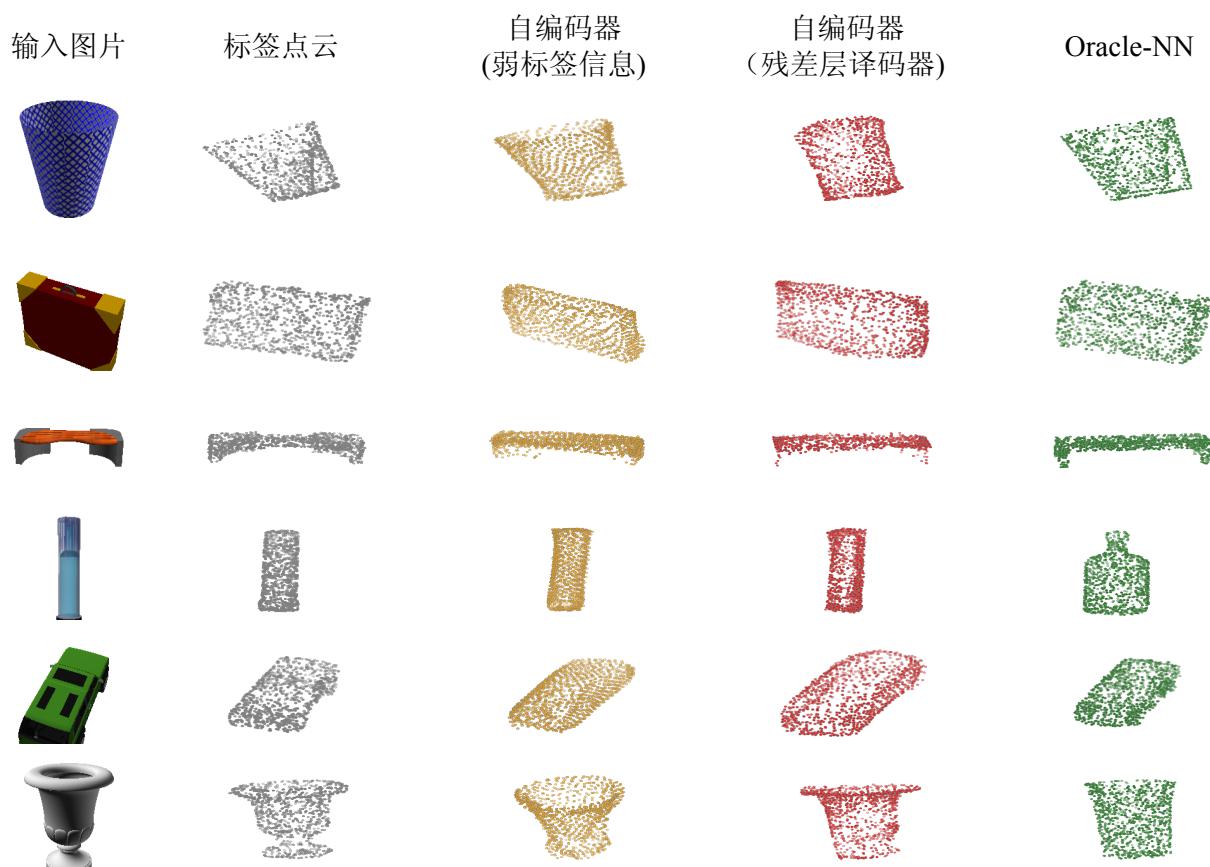


图 3-7 各模型预测结果的可视化

机制探究

在这个部分，我们证明了在 ShapeNet 上训练的标准的神经网络执行一个重建任务而不是识别任务。论证理由如下：

- 1) 根据上文部分3.2.2展示的训练结果，由标准的三维重建网络获得的重建结果在量化指标如 Chamfer Distance 和 F-score 上超过了基于识别机制的非深度学习方法 Oracle-NN 的重建结果。且 Oracle-NN 是基于识别机制的非深度学习方法的理论极限，因此可以推理出在我们展示的实验中，三维重建神经网络并没有执行识别机制，且基于重建机制的模型在 ShapeNet 上的性能超越了基于识别机制的模型。
- 2) 我们展示了二维 T-SNE (见图3-8)。图3-8a和图3-8b分别展示了在训练时有类标签监督和没有类标签监督的两个神经网络的可视化结果 (每个模型都训练了 100 个周期)。T-SNE 图上的每个二维点都表示在测试时编码器输出的一个码字。我们只选择了 9 个样本最多的类，因此如果存在聚类效果的话那么这一效果会更明显。这两个神经网络有同样的结构，都使用的是上文优化模型中使用残差层改进解码器后的基准模型3.2.1，并且他们使用同样的流程进行训练 (100 个周期，学习率 0.003，批处理大小 32，权值衰减 1E-6，并且都使用同样的学习率衰减步骤)。

在图3-8a，神经网络使用类标签监督训练，我们看到清楚的聚类特征同时网络有比较高的 Chamfer Distance，也意味着比较差的重建效果。相反图3-8b并没有展现出清楚的聚类特征但有比较低的 Chamfer Distance。这个比较说明了分类信息并不总是对三维重建有提升作用。T-SNE 比较也是一个能够证明标准训练不会导致神经网络执行识别任务的有力证据。

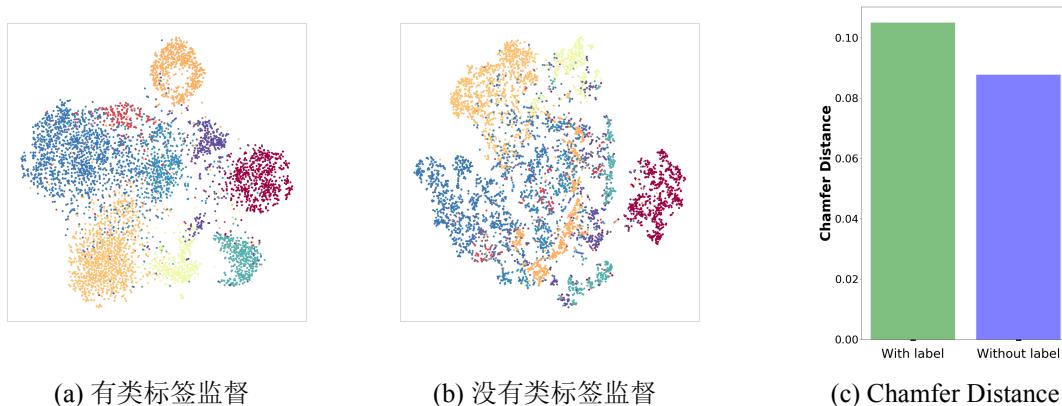


图 3-8 两种训练模式（有类标签监督和没有类标签监督）下的神经网络的 T-SNE and Chamfer Distance

第四章 机制探究

在这个章节，我们首先理论分析了训练数据的聚类特性与训练后的神经网络在识别机制与重建机制之间偏倚的关联性（可以参考前文章节2.2的定义）。然后，我们使用在合成数据集上得出的实验结果来支持理论分析（参考章节4.2）。

4.1 理论分析

在这个子章节，我们分析了在三维重建任务中观察到的识别现象背后的直观理解。我们展示了，当训练点云集的聚类趋势比训练图片集的聚类趋势大时，训练后的模型更倾向于执行识别机制而不是重建机制。参考图4-1。

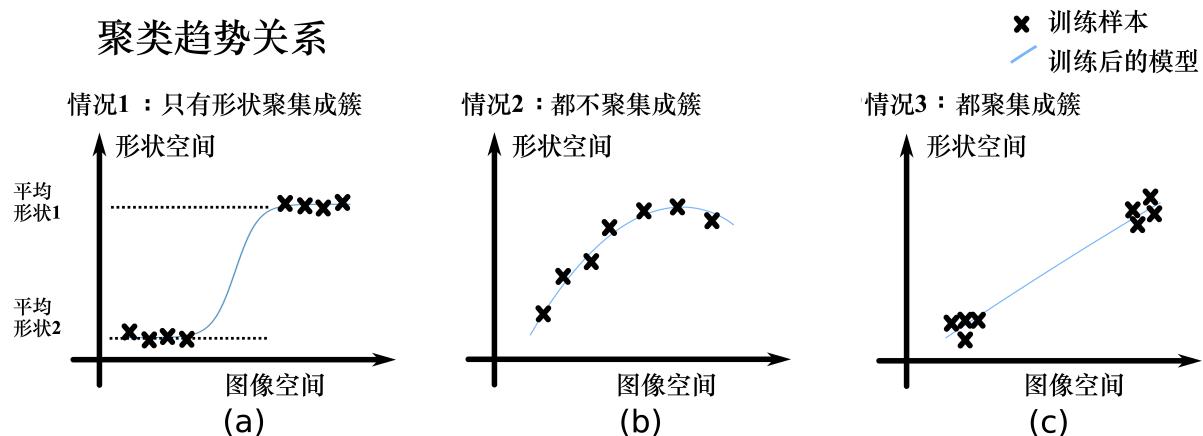


图 4-1 数据集聚类趋势关系

在图4-1，我们展示了三种不同类别的训练数据。在子图4-1a，训练图片比训练形状聚类程度更小。在这种类型数据上训练的模型展示出一个更明显的倾向去执行识别机制，比如形状预测集中在平均形状上，表示为虚线与y轴的交界处。在子图4-1b，训练图片和训练形状都没有展示出明显的强烈的聚类趋势。所以在这种类型的数据上训练的模型不会执行识别机制。在子图4-1c，我们展示了另外一种情况，当图片和形状都成簇时，训练后的模型预测的结果不会呈现一个强的聚类趋势。这也许违背直观认知，但是它确实符合我们在实验中观察的现象。

总结一下，列举我们在单视角三维重建实验（从图片到点云形状）中探索的三类数据情况。

- 情况 1：训练形状集聚集成簇而训练图片集不成簇
- 情况 2：训练图片集与训练形状集都不聚集成簇
- 情况 3：训练图片集与训练形状集都聚集成簇

我们关于神经网络向识别机制还是重建机制偏倚的观点如下：

主要观点：基于深度学习的单视角三维重建神经网络倾向于只在情况 1 进行识别，在情况 2 和情况 3 进行重建。

4.2 实验设计与数据集生成

为了证明上一子章节4.1理论分析中提出的观点，我们需要实验数据集具备不同程度的聚类特征，因此我们需要制造自定义数据集。在这个子章节，我们提供了合成形状数据集的细节。我们使用软件 Blender 生成网格形式的基础形状。然后，我们使用“最近顶点”模式的 Shrinkwrap modifier 来定义两个基础形状之间的形变，并且使用 Blender Shape Keys 面板控制这个插值过程。由此获得的基础数据集中的所有样本的三维形状在数据域中近似均匀分布且连续变化，这样方便我们采样出具有不同聚类特征的子数据集。我们渲染训练图片并且从网格上采样点云形成样本对（图片，点云）以完成基础数据集的创建。之后，我们对基础数据集进行二次采样获得聚类程度不同的小数据集，以覆盖子章节4.1提到的情况 1, 2, 3。

基础数据集 #1（正立方体-球）第一个基础数据集通过在一个正立方体和一个有相似大小的球之间插值获得。参考图4-4。我们在两端之间插值 1000 个过渡形状。对于每个过渡形状，我们从等距视角渲染一张图片并且采样获得一个包含有 1024 个三维点的点云表示。1000 个过渡形状被随机划分为 700 个训练形状，200 个测试形状以及 100 个验证形状。



图 4-2 基础数据集 #1

二次采样数据集 1.1 描述章节4.1的情况 2 和情况 3 创建二次采样数据集 1.1 目的是覆盖从情况 2 到情况 3 的转化。这个数据集是一组 7 个二次采样数据集的集合 1.1.n ($n=1,2,\dots,7$)。每个二次采样数据集 1.1.n 包含了 5 个形状簇和相应的图片。从二次采样数据集 1.1.1 到二次采样数据集 1.1.7，聚集趋势逐渐变低。更具体的是，二次采样数据集 1.1.1 五个分离的簇（参考图4-3a第一列），同时二次采样数据集 1.1.7 几乎连续覆盖了从球到正立方体的整个变化（参考4-3b）。需要注意的是全部的七个二次采样数据集包含了相同数量的训练图片，比如每个数据集包含五个簇，每个簇包含 20 个样本。但是每个二次采样数据集的测试集保持与基础数据集 1 的测试集一样。换句话说，我们只对训练集进行二次采样，同时保持固定的测试集。

二次采样数据集 1.2 描述章节4.1的情况 1 和情况 3 二次采样数据集 1.2 是基础数据集 1 的另一个二次采样版本。这个数据集的目的是覆盖从情况 1 到情况 3 的转变。这个数据集是一个 7 个二次采样数据集 1.2.n ($n=1,2,\dots,7$) 的集合。每个数据集 1.2.n 包含了两个样本簇。但是，与二次采样数据集 1.1 不同，尽管二次采样数据集 1.2 中的每个训练对（图片，三维形状）有一个独特的二维图片，相对应的三维形状只能为两个端点形状中的其中一个（例如两个基础形状正立方体和球），并且没有过渡形状存在。所以在这种情况下，训练形状的聚类趋势总是非常高，同时从二次采样数据集 1.2.1 到 1.2.7 的训练图片的聚类趋势逐渐变低。与之

前一样，二次采样数据集 1.2.7 的训练图片集几乎覆盖了从球到正立方体的整个变化。所以，二次采样数据集 1.2.1 表示了情况 3 的一个案例，同时 1.2.7 表示了情况 1 的一个案例并且数据集 1.2.n ($n = 1, 2, \dots, 7$) 表示了从情况 3 到情况 1 的转变。在每个数据集，训练图片的数量总是 2×10 , 2 是簇的数量，10 是每个簇的样本数量。与二次采样数据集 1.1 相似，二次采样数据集 1.2 的测试集依然是基础数据集 1 的测试集。

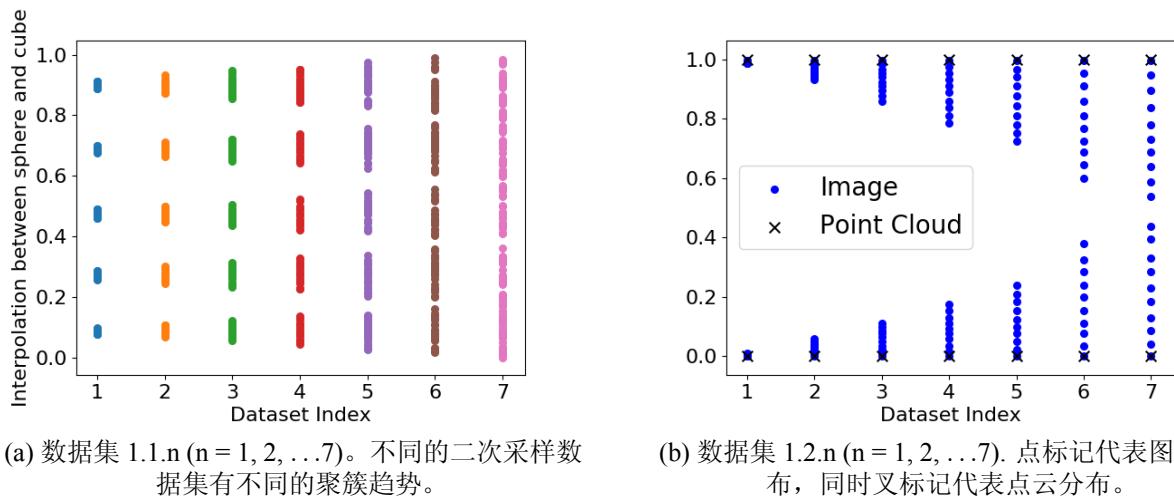


图 4-3 二个二次采样的数据集 1.1.n 和 1.2.n 在基础数据集上的分布

基础数据集 #2 (正立方体-正四面体-球) 为了增加合成数据集的复杂度，增加实验验证的有效性，我们生成了一个比基础数据集 #1 更复杂的数据集，在三个基础形状之间插值，分别是正立方体，正四面体，球。这三个形状有相似的大小。在这三个基础形状之间，我们插值了 10000 个过渡形状。与基础数据集 #1 相似，我们从基础数据集 #2 中创造了二次采样数据集来表现在章节4.1中讨论的不同情况。具体来说，我们首先使用了正四面体和正立方体作为两个基础形状并生成了 100 个过渡形状 S_1, S_2, \dots, S_{100} ，然后，我们选择每个基础过渡形状 S_i ，并且在 S_i 和基础形状球之间线性插值 100 个形状。这个二维插值的过程总共给了我们 $100 \times 100 = 10000$ 个形状。这 10000 个形状再次被分为 7000 个训练样本，1000 个验证样本和 2000 个测试样本。

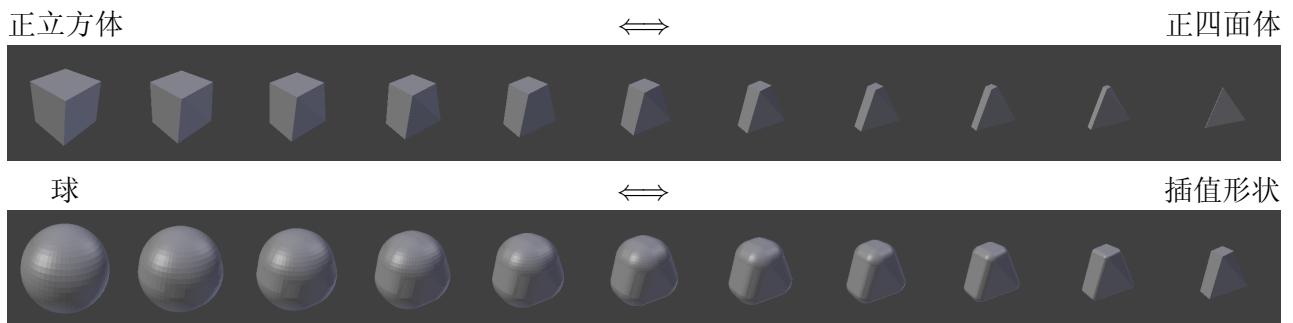


图 4-4 基础数据集 #2

二次采样数据集 2.1 描述章节4.1的情况 2 和情况 3 二次采样数据集 2.1 的目的是覆盖从情况 2 到情况 3 的变化。这个数据集是一组 6 个二次采样数据集的集合 2.1.n ($n=1, 2, \dots, 6$)。

从二次采样数据集 2.1.1 到 2.1.6，聚集趋势逐渐变低。更具体的是，二次采样数据集 2.1.1 有三个明显分离的簇，由在三个基础形状周围采样获得，同时二次采样数据集 2.1.6 几乎连续覆盖了从 2.1.1 到 2.1.6 的所有变化。每个二次采样数据集有三个簇，每个簇有 30 个样本。与正立方体-球数据集 #1 相似，测试集保持固定（是基础数据集 #2 的初始测试集）。但是，因为测试集里的形状数量非常多，我们对测试集通过随机选择 300 个测试样本进行了二次采样。当训练神经网络时，我们使用了验证数据集来选择最优的模型。

二次采样数据集 2.2 描述章节4.1的情况 1 和情况 3 二次采样数据集 2.2 的目的是覆盖从情况 3 到情况 1 的变化。它包含了六个数据集 2.2.n ($n=1,2,\dots,6$)。从二次采样数据集 2.2.1 到 2.2.6，训练图片集的聚类趋势逐渐降低，同时训练点云集的聚类趋势保持在高水平。我们固定训练点云，比如对于每个样本对 (I, S) ，我们使得与每个训练图片 I 相关的训练点云 S 都是三个基础形状中的一个。所以，从二次采样数据集 2.2.1 到 2.2.6 的转变表示了从情况 3 到情况 1 的转变。在每个二次采样数据集中，我们生成接近三个基础形状的簇，并且每个簇有 16 个样本。当训练神经网络时，我们不使用验证集，并且简单地选择最后一个周期的模型参数。原因是在当前的设定下，从训练集（只有三个基础形状点云）到验证集（插值形状点云）存在一个域的转变。所以，验证数据集在训练时不能用于选择最优模型参数。

4.3 实验结果

在这个子章节，我们展示了实验结果来论证4.1中的主要论点：识别机制只会在训练图片聚簇程度小于训练形状聚簇程度时才会发生，例如，情况 1。为了证明这个论点，我们已经在章节4.2中创建了不同的数据集来覆盖情况 1, 2 和 3 之间的过渡转变。关于神经网络模型，我们采用了在章节3.2.1中使用残差层优化过的自编码器架构3-4，采用 Chamfer Distance 作为损失函数。对于二次采样数据集 1.1，我们为从 1.1.1 到 1.1.7 的每个数据集都分别训练了一个单独的神经网络，进行了 1500 个周期的训练，以到达保证他们的学习能力饱和。对于子数据集 1.2，我们也分别在每个子数据集 1.2.n 上训练一个单独的神经网络，但训练周期更多，有 3000 个周期，因为 1.2.n 每个数据集的样本量更小。在训练期间，我们使用 Adam 优化器，初始化学习率为 0.003，学习率衰减为 0.1。我们使用验证集来选择最好的模型用于预测。

4.3.1 数据集指标测量

在这个子章节，我们使用2.3提出的衡量数据集聚类程度的指标来对我们机制探究实验中所有数据集以及网络预测结果的聚类程度进行测量。

基础数据集 #1 的实验结果 在这个部分，我们展示了基础数据集 #1 的量化实验结果来支持章节4.1中的主要观点。对于每个二次采样数据集，我们画出模型输入数据的聚类趋势与输出数据的聚类趋势，由2.3我们提出的轮廓系数（silhouette score）表示。因为我们有两类输入数据——训练图片与训练形状，我们分别画出“输入图片-预测结果”和“输入形状-预测结果”这两个关系。结果展示在图4-5。第一行的结果关于二次采样数据集 1.1，即从情况 3 到情况 2 的转变，同时第二行的结果关于二次采样数据集 1.2，即从情况 3 到情况 1 的转变。对

于二次采样数据集 1.1 的结果，即使是输入的聚类趋势变化，输出的聚类趋势保持不变。这一结果证明了我们在4-1b 和4-1c 的理论分析，即在情况 2 和情况 3 下输出的聚类趋势比较低，神经网络执行重建机制。但是，在图4-5的第二行，我们可以看到当输入图片聚类趋势降低时，输出聚类趋势明显提升。这正是与我们在图4-1a 中的理论分析一致，即重建形状的聚簇只会在训练图片不聚簇且训练形状聚簇的情况发生、注意在第二行，训练形状只包含两个端点形状（球和正立方体）。所以，训练形状的聚类趋势总是非常高（注意图4-5第二行右边的图片）。

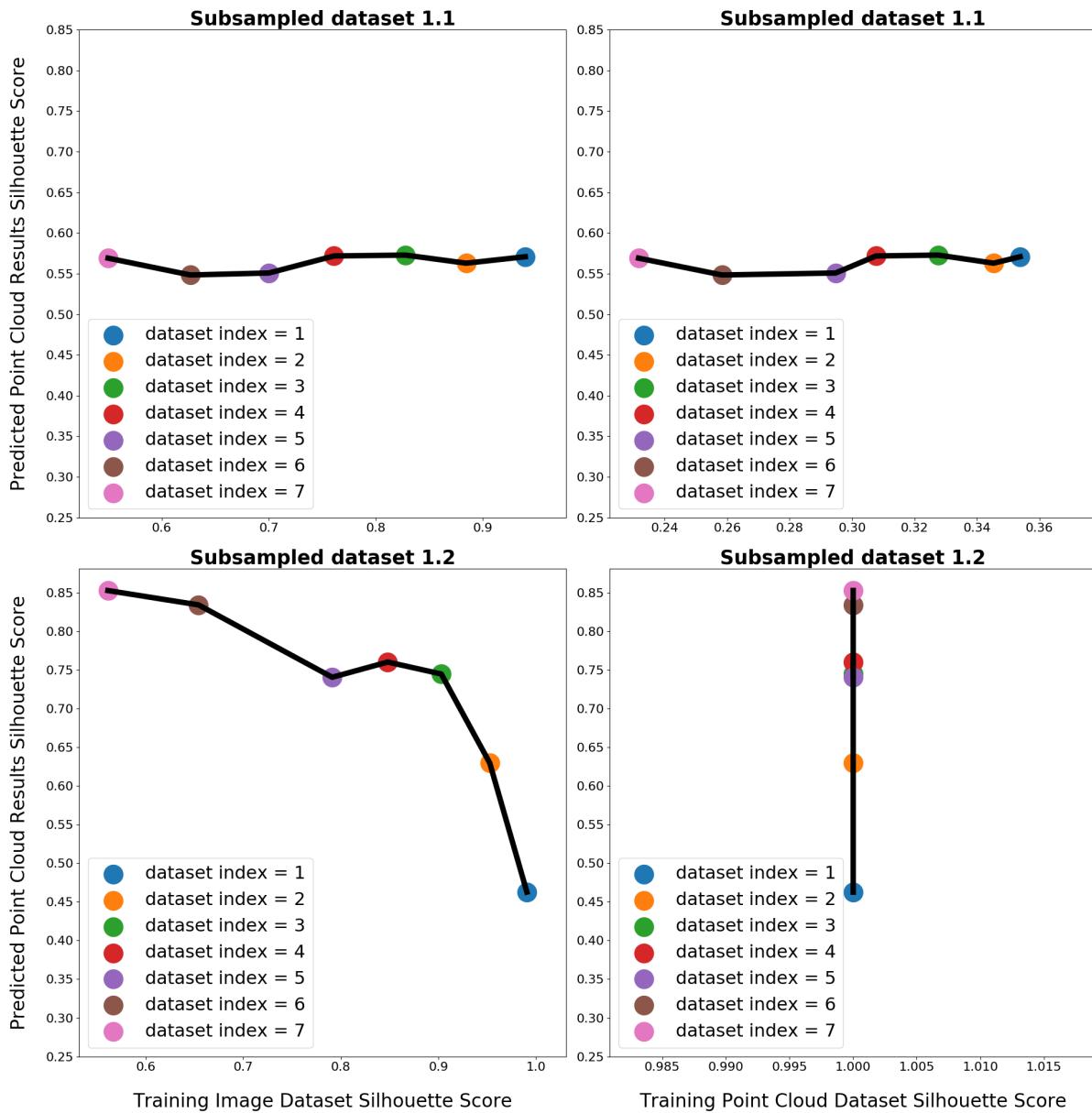


图 4-5 “基础数据集 #1 的训练图片/形状点云-输出形状”的轮廓系数 (silhouette score)。只有情况 1 (左下角子图中较大的数据集索引指向的点) 展示出高的聚类趋势

基础数据集 #2 的实验结果 在这个部分，我们展示了基础数据集 #2 的量化实验结果，该实验结果由在每个二次采样数据集上 (2.1.n 以及 2.2.n) 进行训练和测试获得。具体来说，该量化结果为输入与输出的轮廓系数 (silhouette score)，展示在图4-6。从该图的第二行左侧的图可以看出只有在情况 1 预测结果才会呈现出明显的聚类趋势，与章节4.1的结论一致。该图

第一行的两张子图展示了在情况 2 和情况 3 下预测结果都不会呈现出高的聚类趋势。

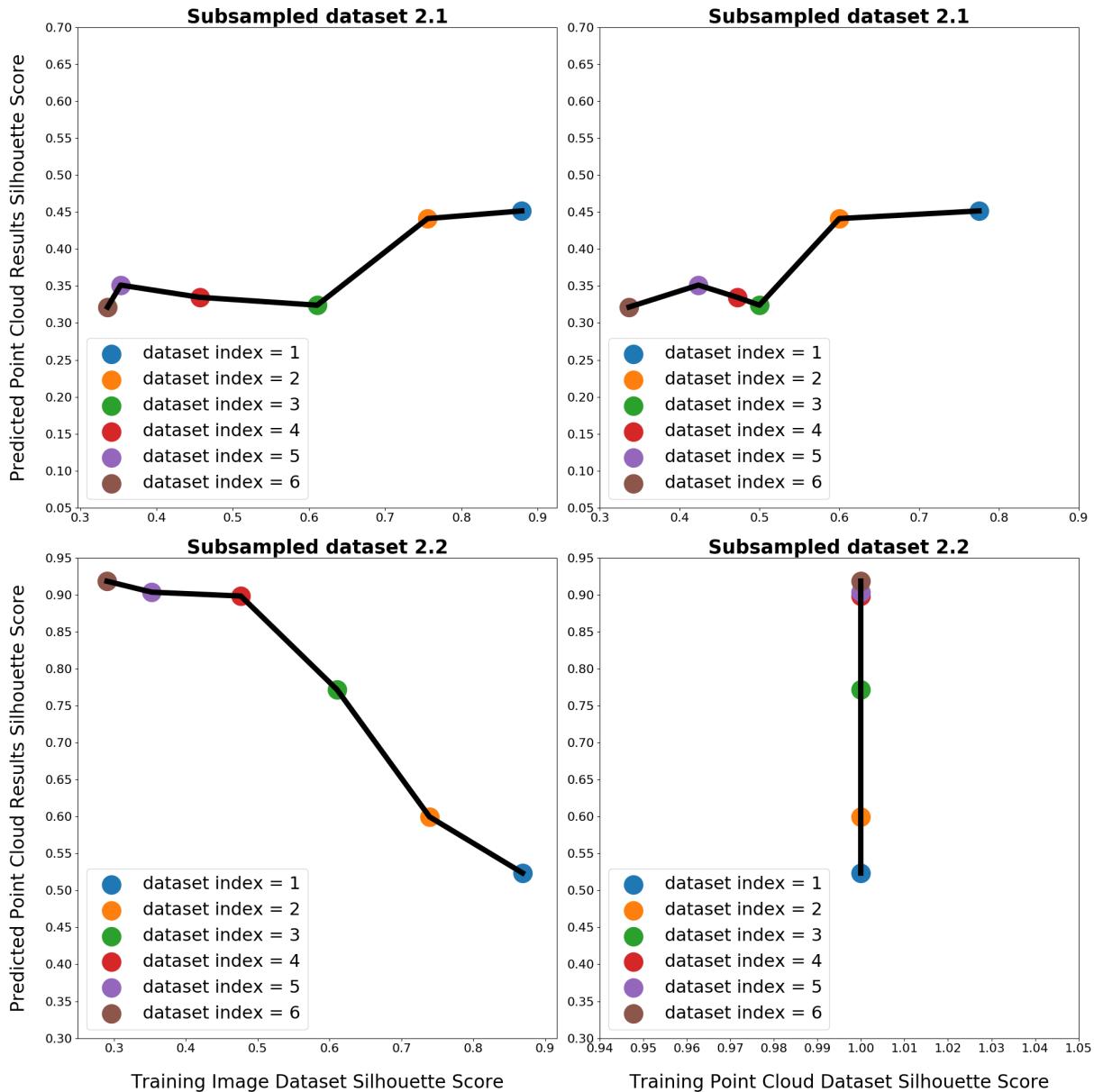


图 4-6 “基础数据集 #2 的训练图片/形状点云-输出形状”的 (silhouette score)。只有情况 1 (左下角子图中较大的数据集索引指向的点) 展示出高的聚类趋势

4.3.2 距离矩阵的可视化

在这个子章节，我们可视化了在不同情况的数据集上训练的神经网络的预测结果，一方面验证我们在2.3提出的衡量数据集聚类程度的指标在上文展示4.3.1的结果的准确性，另一方面进一步支持我们在章节4.1中的主要观点。我们使用降维方法，对一个网络的一组数据（点云或图片）计算一个距离矩阵。

基础数据集 #1 的实验结果 在这个部分，我们可视化了在基础数据集 #1 上训练的神经网络得到的预测结果的距离矩阵。参考图4-7。在左侧子图，我们可视化了二次采样数据集 1.1 的结果，覆盖了情况 3 和情况 2。在右侧子图，我们可视化了二次采样数据集 1.2 的结果，覆盖了情况 3 和情况 1。左侧子图的七行对应二次采样数据集 1.1.1 到 1.1.7。右侧子图的七行对

应二次采样数据集 1.2.1 到 1.2.7。在每个子图，左侧列表示输入图片的距离矩阵，中间列表示训练点云的距离矩阵以及右侧列表示测试时的预测点云的距离矩阵。对于训练图片，距离由图片的 ℓ_1 -距离。对于点云，距离为 Chamfer Distance。

在右侧子图，我们看到预测点云只在下方几行形成簇。这意味着，训练完的神经网络只在训练数据集接近情况 1 的时候执行识别机制。在其他两种情况（情况 2 和情况 3），训练完的神经网络倾向于插值而不是集中在平均形状上，这导致了相似度矩阵连续的变化。

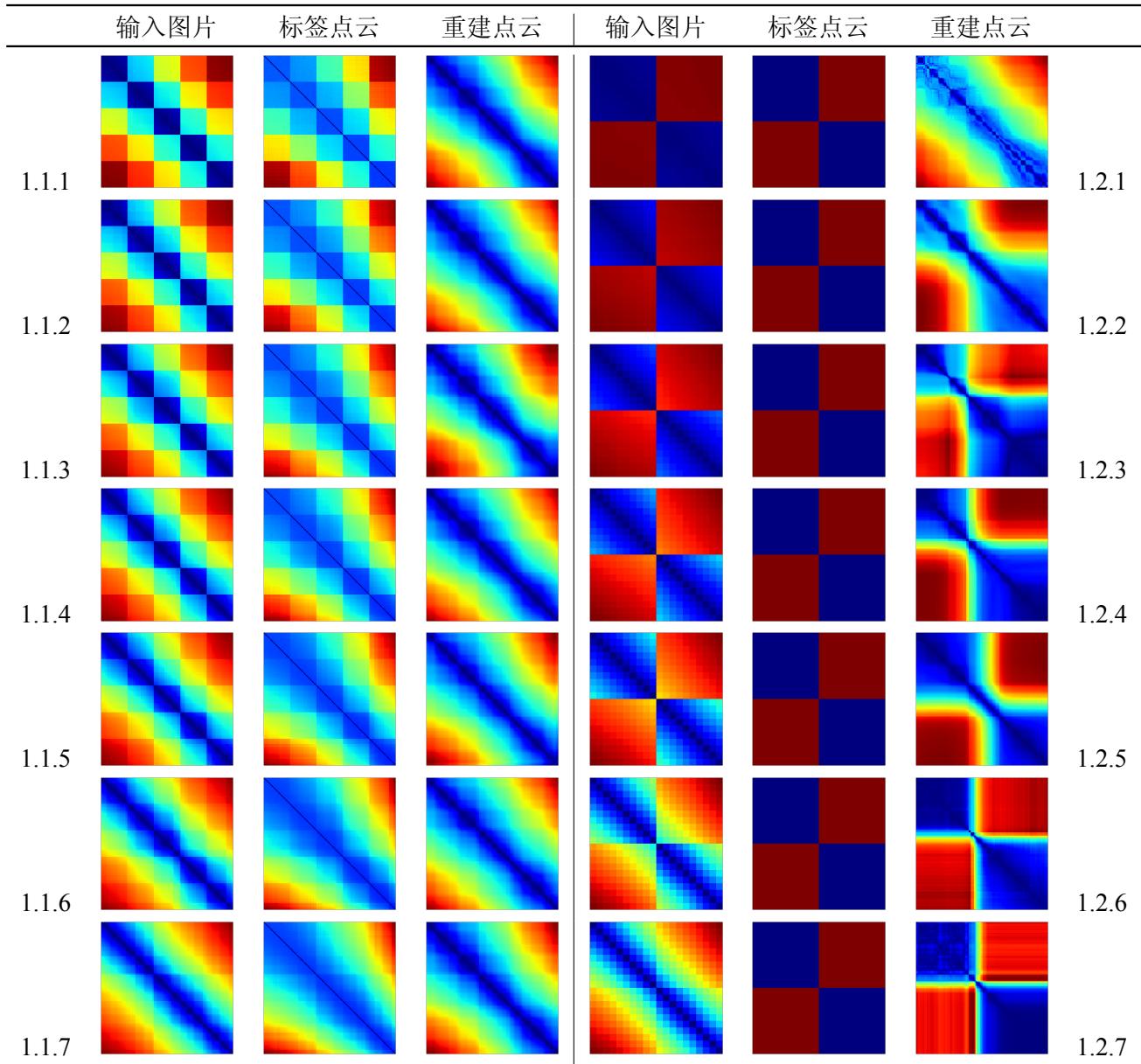


图 4-7 基础数据集 #1 的距离矩阵。蓝色表示小距离以及红色代表大距离。（左侧三列）二次采样数据集 1.1 的预测形状不成簇。（右侧三列）如果数据集情况符合 4.1 的情况 1，二次采样数据集 1.2 的预测形状成簇。从上到下：数据集索引 1 到 7。

基础数据集 #2 的实验结果 这里，我们展示了数据集 #2 的距离矩阵可视化结果。参考图4-8。我们可以看到该图左起第六列的预测结果距离矩阵逐渐成簇，这意味着当数据集从情况 3 变化到情况 1 时，点云预测逐渐从重建转变为识别。这一定性结果与图4-6左下角的子图展示定量结果 silhouette score 的变化趋势一致，即预测点云结果从 2.2.1（情况 3）到 2.2.6

(情况 1) 有一个持续增大的 silhouette score。相反, 左起第三列并没有明显的变化, 意味着情况 2 和情况 3 下预测结果都没有高的聚类趋势。这也与图4-6第一行展示的 silhouette score 没有明显的变化趋势一致。

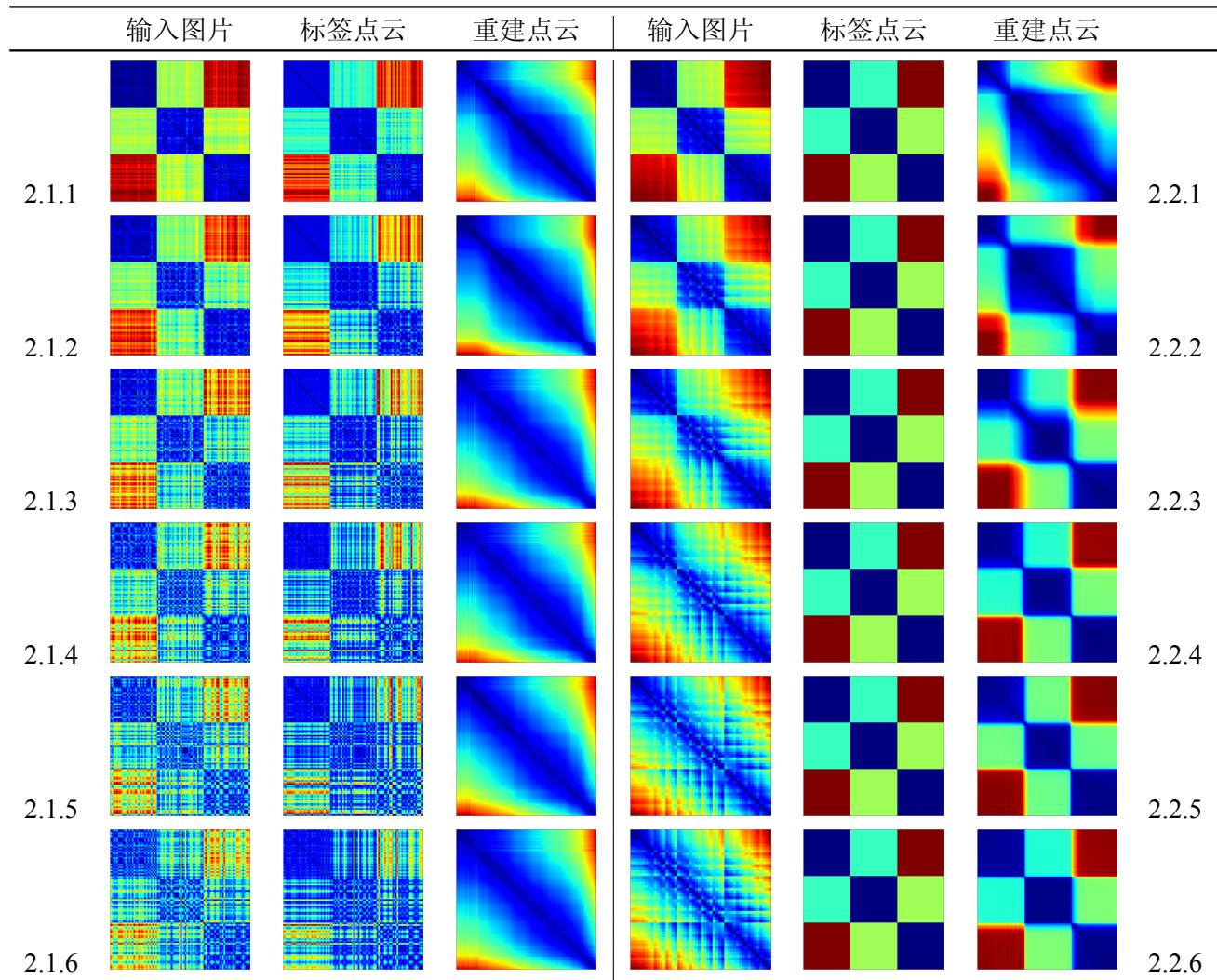


图 4-8 基础数据集 #2 的距离矩阵。蓝色表示小距离以及红色代表大距离。(左侧三列) 二次采样数据集 2.1 的预测形状不成簇。(右侧三列) 如果数据集情况符合4.1的情况 1, 二次采样数据集 2.2 的预测形状成簇。从上到下: 数据集索引 1 到 6。

第五章 总结与展望

5.1 工作总结

在这篇文章中，我们研究了深度神经网络在单视角三维重建任务中的应用。针对深度学习基准模型在训练中的问题，提出了引入弱标签信息监督以及用残差层改进解码器的模型优化方案，使其在公开数据集 ShapeNet 上的测试结果超过了基于识别机制的非深度学习方法的理论极限。在模型优化的基础上，我们进一步探究了神经网络在重建任务中的内在机制，即神经网络是否倾向于执行识别（检索）机制或者重建（插值）机制。我们的结论是这一倾向取决于数据集的特征——训练样本的聚类趋势。我们提出了基于亲和传播（affinity propagation）和轮廓系数（silhouette score）的数据集聚类趋势度量指标，并结合实验结果总结规律：只有当训练形状集的聚类程度高于训练图片集的聚类程度时，神经网络才会执行识别机制，在其他正常情况下，训练后的神经网络更倾向于重建而不是识别。概括而言，识别现象是不合适的数据集收集和数据使用导致的数据集复杂特征引发的，而不是源于深度神经网络模型。因此，我们建议训练集中的三维形状之间应该比训练集中的图片之间呈现出更多的差异，以此来避免产生一个基于识别的机器学习模型，这为三维数据集的收集以及三维重建训练提供了指导。

问题 4：第 5.1 节的工作总结，各项工作再指明自己创新点之后可以再展开描述一下，类似于摘要中的各条描述，还可以更详细些，因为你已经有实验和验证结果了。

本课题具有如下创新点：

1. 提出了单视角三维重建任务中行之有效的深度神经网络优化方案。提出两种创新的优化方法，一是创造性的使用残差层改进基于“折纸”思想的解码器的架构，二是引入弱标签监督来进行编码器初始化。从整个数据集平均重建指标上来看，用残差层改进解码器的方法使得基准模型获得了更稳定的学习能力以及对超参数调参的鲁棒性，缩小了与基于识别机制的模型的理论极限（Oracle-NN）的差距。使用弱标签监督的方法使得基准模型超过了 Oracle-NN。从类重建指标统计特性上来看，两种方法优化后的基准模型对于类间数据量不平衡的鲁棒性明显优于 Oracle-NN。这两种方法都可以被推广到三维重建领域除点云外其他三维表示的重建，如体素，网格等。
2. 提出了测量三维重建数据集聚类程度的度量指标。该度量指标借鉴了亲和传播（affinity propagation）以及轮廓系数（silhouette score）的方法，衡量得出的数据集聚类程度与数据集降维可视化结果吻合，有希望成为三维重建数据集公开权威的度量指标。
3. 提出了对深度神经网络内在机制的丰富理解，并设计了实验流程进行了成功的验证。提出的主要观点为：影响三维重建深度神经网络在识别机制与重建机制之间偏倚的主要因

素为数据集的本质特征——聚类程度。并且基于计算机图形学算法创造性的创建了有不同聚类特征的形状插值数据集，并在此基础上进行了大规模的神经网络训练与测试，并利用自定义的度量指标以及数据集降维可视化分别定量与定性的验证了猜想。

当前研究存在如下不足之处：

1. 优化后的模型对算法需求过高，单个模型训练所需时间过长。
2. 优化后的模型重建的三维形状缺少复杂的局部细节（如飞机的机翼，椅子背上的凹槽等）。
3. 在自定义合成数据集上的机制探究实验尚未在大型公开数据集上进一步验证，因为没有解决计算大规模距离矩阵所需要的算力问题。

5.2 工作展望

本课题仍有一些问题没有得到合理解决，未来将会在如下方面进行改进：

1. 应用并行计算等加速方法计算距离矩阵和训练神经网络，以提高验证猜想的效率并满足在大型数据集上验证的要求。
2. 采用神经网络剪枝和模型压缩等方法优化模型，减少网络训练所需的时间。

基于本课题的成果与结论，决定从如下方面进一步拓展工作：

未来有意义的工作方向是从数据收集，网络架构，优化过程等角度研究深度学习技术，使得神经网络即使是在数据集特征偏倚于识别时依然能执行重建任务。

我们考虑使用数据增强的方法在数据簇的周围增加数据点，增大簇的半径，以达到降低数据集聚类程度的目的。我们已经尝试使用了 mixup 以及 manifold mixup 的方法，但是因为暂时缺少方法对两个三维点云进行算术运算，因此只对图片和码字进行了 mixup，而这种残缺的增强方法并没有起到提高重建效果的作用。与 mixup 的作者联系后，我们了解到 mixup 的方法需要作用于输入与标签，只在一端进行增强并不能达到目的。因此我们接下来需要考虑如何对两个三维点云进行算术运算，获取中间的过渡形状以实现数据增强，正在尝试的方案是训练一个基于三维点云输入的自编码器，用训练完的编码器输出的码字代表一个三维形状，并对码字进行算术运算，运算后的码字结果再经过解码器生成两个形状之间的过渡形状。

目前还有一种正在探索的数据增强方法，源于图像领域 AutoAugment 以及 Population-based Augmentation，即使用强化学习或者进化算法来学习数据增强的策略。我们的方案是，先设计对三维形状进行增强的基础策略，如旋转，不规则缩放，切断等，将这些基础策略应用于三维形状后，直接渲染出对应的单视角图片，将图片与形状输入神经网络进行训练，在此基础上结合策略学习的方法。

参考文献

- [1] Li C L, Zaheer M, Zhang Y, et al. Point cloud gan[J]. arXiv preprint arXiv:1810.05795, 2018..
- [2] Park J J, Florence P, Straub J, et al. Deepsdf: Learning continuous signed distance functions for shape representation. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019. 165–174.
- [3] Fan H, Su H, Guibas L J. A point set generation network for 3d object reconstruction from a single image. Proceedings of the IEEE conference on computer vision and pattern recognition, 2017. 605–613.
- [4] Tatarchenko M, Dosovitskiy A, Brox T. Octree generating networks: Efficient convolutional architectures for high-resolution 3d outputs. Proceedings of the IEEE International Conference on Computer Vision, 2017. 2088–2096.
- [5] Groueix T, Fisher M, Kim V G, et al. A papier-mâché approach to learning 3d surface generation. Proceedings of the IEEE conference on computer vision and pattern recognition, 2018. 216–224.
- [6] Yang Y, Feng C, Shen Y, et al. Foldingnet: Point cloud auto-encoder via deep grid deformation. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018. 206–215.
- [7] Wang N, Zhang Y, Li Z, et al. Pixel2mesh: Generating 3d mesh models from single rgb images. Proceedings of the European Conference on Computer Vision (ECCV), 2018. 52–67.
- [8] Sun X, Wu J, Zhang X, et al. Pix3d: Dataset and methods for single-image 3d shape modeling. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018. 2974–2983.
- [9] Tulsiani S, Zhou T, Efros A A, et al. Multi-view supervision for single-view reconstruction via differentiable ray consistency. Proceedings of the IEEE conference on computer vision and pattern recognition, 2017. 2626–2634.
- [10] Wu J, Wang Y, Xue T, et al. Marrnet: 3d shape reconstruction via 2.5 d sketches. Advances in neural information processing systems, 2017. 540–550.
- [11] Yan X, Yang J, Yumer E, et al. Perspective transformer nets: Learning single-view 3d object reconstruction without 3d supervision. Advances in neural information processing systems, 2016. 1696–1704.
- [12] Kingma D P, Ba J. Adam: A method for stochastic optimization[J]. arXiv preprint arXiv:1412.6980, 2014..
- [13] Tatarchenko M, Richter S R, Ranftl R, et al. What do single-view 3d reconstruction networks learn? Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019. 3405–3414.
- [14] Arpit D, Jastrz̄bski S, Ballas N, et al. A closer look at memorization in deep networks. Proceedings of the 34th International Conference on Machine Learning-Volume 70. JMLR. org, 2017. 233–242.

- [15] Häming K, Peters G. The structure-from-motion reconstruction pipeline - a survey with focus on short image sequences[J]. *Kybernetika*, 2010, 46:926–937.
- [16] Saxena A, Sun M, Ng A Y. Make3d: Learning 3d scene structure from a single still image[J]. *IEEE Trans. Pattern Anal. Mach. Intell.*, 2009, 31(5):824–840.
- [17] Kar A, Tulsiani S, Carreira J, et al. Category-specific object reconstruction from a single image[J]. 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015..
- [18] Groueix T, Fisher M, Kim V G, et al. A papier-mache approach to learning 3d surface generation[J]. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2018..
- [19] Chang A X, Funkhouser T, Guibas L, et al. Shapenet: An information-rich 3d model repository[J]. arXiv preprint arXiv:1512.03012, 2015..
- [20] Yi L, Shao L, Savva M, et al. Large-scale 3d shape reconstruction and segmentation from shapenet core55[J]. ArXiv, 2017, abs/1710.06104.
- [21] Kanazawa A, Tulsiani S, Efros A A, et al. Learning category-specific mesh reconstruction from image collections. Proceedings of the European Conference on Computer Vision (ECCV), 2018. 371–386.
- [22] Pontes J K, Kong C, Sridharan S, et al. Image2mesh: A learning framework for single image 3d reconstruction. Asian Conference on Computer Vision. Springer, 2018. 365–381.
- [23] Kurenkov A, Ji J, Garg A, et al. Deformnet: Free-form deformation network for 3d shape reconstruction from a single image. 2018 IEEE Winter Conference on Applications of Computer Vision (WACV). IEEE, 2018. 858–866.
- [24] Li K, Garg R, Cai M, et al. Single-view object shape reconstruction using deep shape prior and silhouette[J]. arXiv preprint arXiv:1811.11921, 2018..
- [25] Gwak J, Choy C B, Chandraker M, et al. Weakly supervised 3d reconstruction with adversarial constraint. 2017 International Conference on 3D Vision (3DV). IEEE, 2017. 263–272.
- [26] Sinha A, Unmesh A, Huang Q, et al. Surfnet: Generating 3d shape surfaces using deep residual networks. Proceedings of the IEEE conference on computer vision and pattern recognition, 2017. 6040–6049.
- [27] Wu J, Zhang C, Zhang X, et al. Learning shape priors for single-view 3d completion and reconstruction. Proceedings of the European Conference on Computer Vision (ECCV), 2018. 646–662.
- [28] Yang B, Wen H, Wang S, et al. 3d object reconstruction from a single depth view with adversarial learning. Proceedings of the IEEE International Conference on Computer Vision Workshops, 2017. 679–688.
- [29] Yang G, Huang X, Hao Z, et al. Pointflow: 3d point cloud generation with continuous normalizing flows. Proceedings of the IEEE International Conference on Computer Vision, 2019. 4541–4550.

- [30] Wu J, Zhang C, Xue T, et al. Learning a probabilistic latent space of object shapes via 3d generative-adversarial modeling. *Advances in neural information processing systems*, 2016. 82–90.
- [31] Girdhar R, Fouhey D F, Rodriguez M, et al. Learning a predictable and generative vector representation for objects. *European Conference on Computer Vision*. Springer, 2016. 484–499.
- [32] Tulsiani S, Su H, Guibas L J, et al. Learning shape abstractions by assembling volumetric primitives. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017. 2635–2643.
- [33] Niu C, Li J, Xu K. Im2struct: Recovering 3d shape structure from a single rgb image. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018. 4521–4529.
- [34] Sung M, Su H, Kim V G, et al. Complementme: weakly-supervised component suggestions for 3d modeling[J]. *ACM Transactions on Graphics (TOG)*, 2017, 36(6):1–12.
- [35] Sun Y, Wang Y, Liu Z, et al. Pointgrow: Autoregressively learned point cloud generation with self-attention[J]. *arXiv preprint arXiv:1810.05591*, 2018..
- [36] Choy C B, Xu D, Gwak J, et al. 3d-r2n2: A unified approach for single and multi-view 3d object reconstruction. *European conference on computer vision*. Springer, 2016. 628–644.
- [37] Achlioptas P, Diamanti O, Mitliagkas I, et al. Learning representations and generative models for 3d point clouds[J]. *arXiv preprint arXiv:1707.02392*, 2017.
- [38] Oliva J, Póczos B, Schneider J. Distribution to distribution regression. *International Conference on Machine Learning*, 2013. 1049–1057.
- [39] Póczos B, Rinaldo A, Singh A, et al. Distribution-free distribution regression[J]. 2013..
- [40] Knapitsch A, Park J, Zhou Q Y, et al. Tanks and temples: Benchmarking large-scale scene reconstruction[J]. *ACM Transactions on Graphics (ToG)*, 2017, 36(4):1–13.
- [41] Van Craenendonck T, Blockeel H. Using internal validity measures to compare clustering algorithms[J]. *Benelearn 2015 Poster presentations (online)*, 2015. 1–8.
- [42] Wang K, Zhang J, Li D, et al. Adaptive affinity propagation clustering[J]. *arXiv preprint arXiv:0805.1096*, 2008..
- [43] 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).
- [44] He K, Zhang X, Ren S, et al. Deep residual learning for image recognition[J]. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016..

附录 A 可视化结果

A.1 各模型预测结果可视化

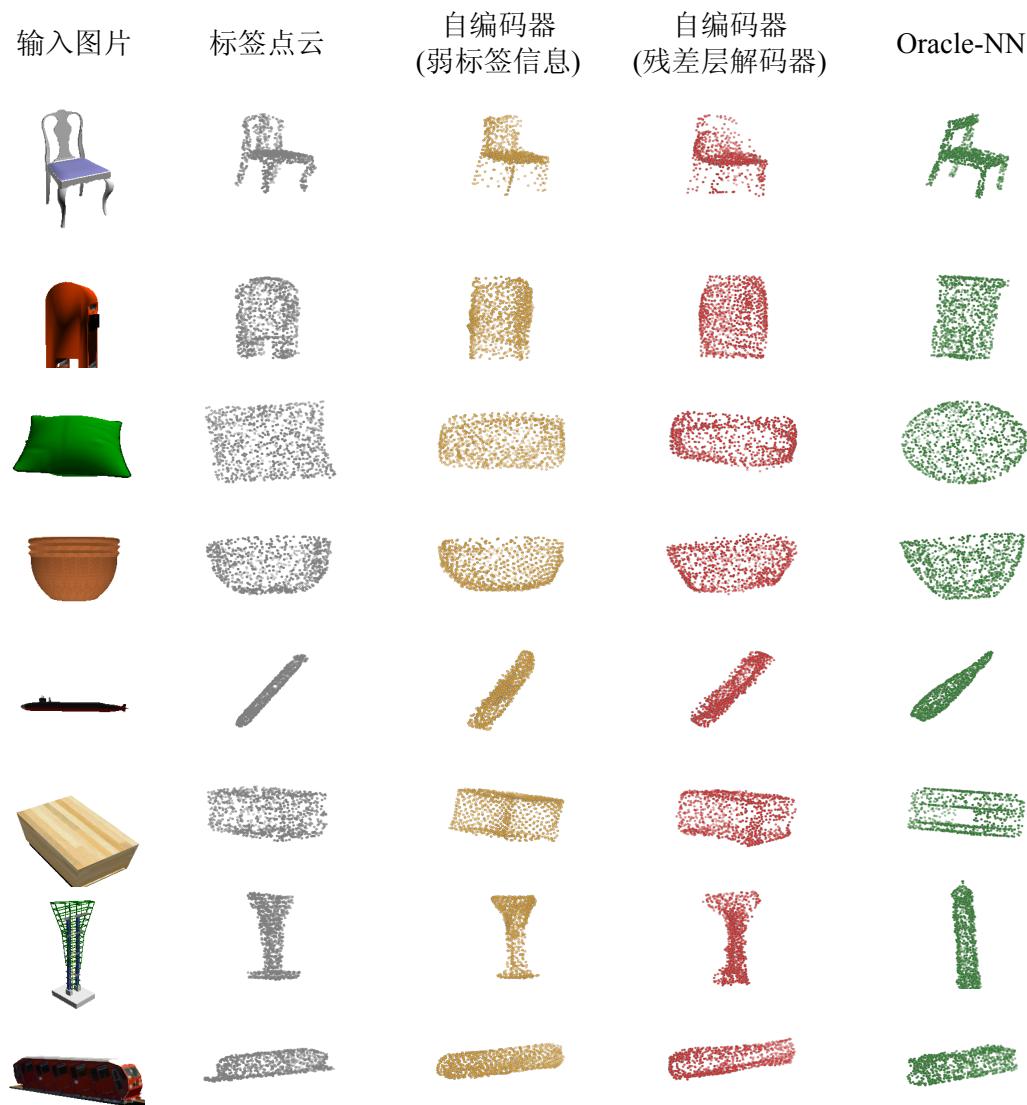


图 A-1 更多的可视化. 从上到下的物体类别: 椅子, 邮箱, 枕头, 壶, 火箭, 桌子, 火车

附录 B 实验数据与结果

B.1 公开数据集统计

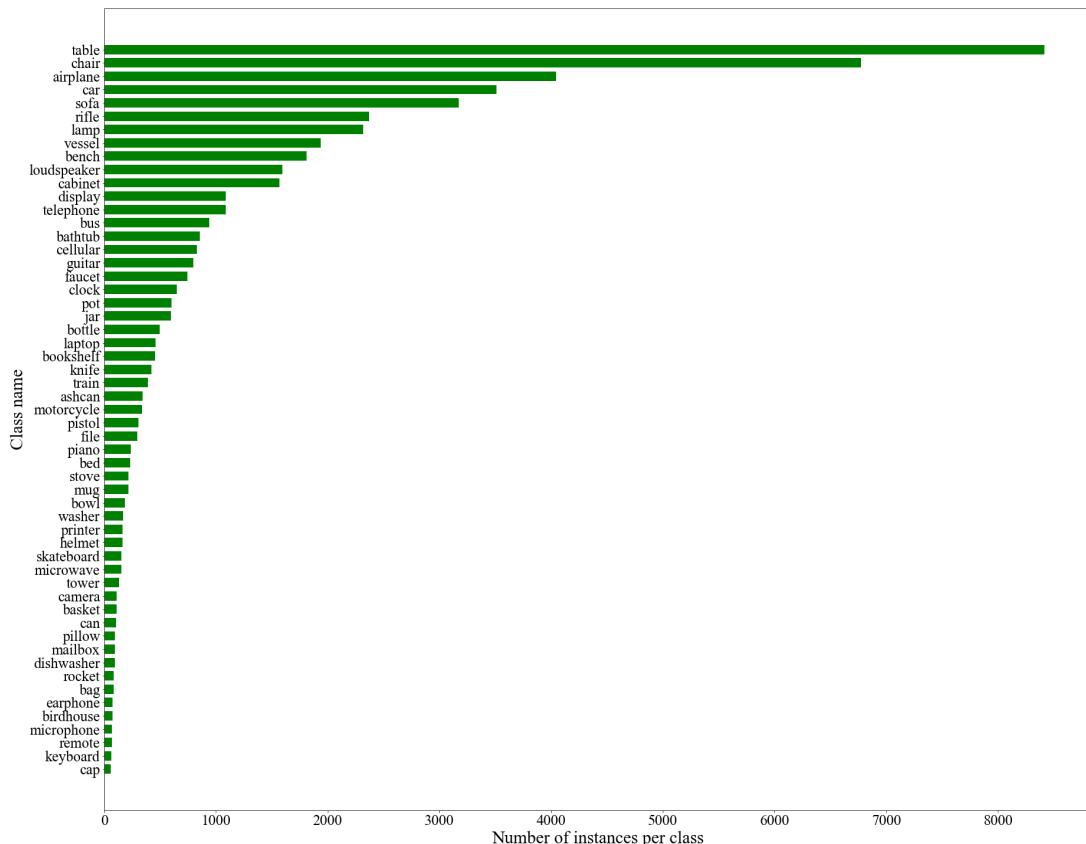


图 B-1 ShapeNet 数据集每个类的样本量分布

致 谢

这次的毕业论文设计总结是在我的指导老师 xxx 老师亲切关怀和悉心指导下完成的。从毕业设计选题到设计完成，x 老师给予了我耐心指导与细心关怀，有了莫老师耐心指导与细心关怀我才不会在设计的过程中迷失方向，失去前进动力。x 老师有严肃的科学态度，严谨的治学精神和精益求精的工作作风，这些都是我所需要学习的，感谢 x 老师给予了我这样一个学习机会，谢谢！

感谢与我并肩作战的舍友与同学们，感谢关心我支持我的朋友们，感谢学校领导、老师们，感谢你们给予我的帮助与关怀；感谢肇庆学院，特别感谢计算机科学与软件学院四年来为我提供的良好学习环境，谢谢！