中图分类号：TP391
密　　级：公开

石家莊鐵道大學
SHIJIAZHUANG TIEDAO UNIVERSITY

专业学位硕士论文

（全日制）

基于深度学习的单视图重建算法研究

Research on Single View Reconstruction Algorithm Based on Deep Learning

培 养 院 系：信息科学与技术学院
作 者 姓 名：白景禧
学　　号：1202110035
专业代码名称：085404 计算机技术
校 内 导 师：王学军　教授
校内副导师：
校 外 导 师：赵卫平　高级工程师

二〇二四年六月

摘 要

单视图三维重建技术是一种基于计算机视觉的方法，通过单张图像还原物体的三维结构。它在许多领域具有广泛的应用，如虚拟现实、增强现实、机器人导航等。然而，由于从单张图像中获得的信息有限，单视图三维重建面临着许多挑战，如深度信息的缺失、视角变化的影响以及噪声和遮挡等问题。

为了克服这些挑战，本文提出了以下两种算法：

(1)基于分组卷积编码的单视图三维重建网络。该研究改进了特征提取和注意力机制，能够生成更丰富、精细的三维体素模型。通过引入多尺度特征和上下文信息，可以更准确地还原物体的形状，并且具有较强的去噪和细节保留能力。网络的编码器通过改进特征提取网络，获取更加丰富完整、深层次的二维特征。在精炼器的网络架构中引入注意力机制，进一步细化三维特征，使其生成更加精细的三维体素模型。另外在网络中添加阈值调整模块，来弥补不同种类图像之间的差异，以达到更好的重建效果。实验结果表明，在公共数据集 ShapeNet 上三维重建的 IoU 值达到 0.675，在单图像重建方面取得了更好的效果。

(2)单视图三维重建中的 Transformer 架构设计。该研究结合了 Transformer 和 CNN 的优势，能够有效地捕捉不同尺度下的关键特征，并准确地还原三维物体的空间结构。利用 Transformer 的自注意力机制来建立全局关联，并通过卷积操作来提取局部特征，从而在学习全局结构和局部细节方面具有出色的性能。网络使用 Transformer 进行物体的重建工作，结合 CNN 细化重建物体。使得三维重建网络可以用于自适应地学习不同尺度下的重要特征，并更好地处理尺度变化带来的挑战。实验结果显示，在公共数据集 ShapeNet 上三维重建的 IoU 值达到 0.699，与纯卷积神经网络相比，该网络在精度上取得了显著提升，呈现出了更出色的效果。

为了验证算法的有效性，本文在进行了相关实验对比及分析的基础上，设计了一个单视图三维重建演示系统。该系统能够生成二维体素图像和三维重建模型，并提供直观、便捷的交互方式。用户可以通过系统对重建后的模型进行旋转、缩放等操作，以更好地观察物体的各个角度和细节。

关键词：三维重建；单视图；CNN；注意力机制；Transformer

Abstract

Single view 3D reconstruction technique is a method based on computer vision to restore the 3D structure of an object through a single image. It has a wide range of applications in many fields, such as virtual reality, augmented reality, robot navigation and so on. However, due to the limited information obtained from a single image, single-view 3D reconstruction faces many challenges, such as the loss of depth information, the influence of Angle changes, and the problems of noise and occlusion.

To overcome these challenges, this thesis proposes two algorithms:

(1)Single-view 3D reconstruction network based on packet convolution coding. The research improves the feature extraction and attention mechanisms to generate richer, finer 3D voxel models. By introducing multi-scale features and context information, the shape of the object can be restored more accurately, and it has strong de-noise and detail retention capabilities. By improving the feature extraction network, the encoder of the network can obtain more rich, complete and deep two-dimensional features. The attention mechanism is introduced into the network architecture of the refiner to further refine the 3D features and generate a more refined 3D voxel model. In addition, threshold adjustment module is added in the network to make up for the difference between different kinds of images to achieve better reconstruction effect. The experimental results show that the IoU value of 3D reconstruction on the public data set ShapeNet reaches 0.675, which achieves better results in single image reconstruction.

(2)Transformer Architecture Design in Single-View 3D Reconstruction. The research combines the strengths of Transformer and CNN to effectively capture key features at different scales and accurately restore the spatial structure of three-dimensional objects. Make use of Transformer's self-attention mechanism to establish global associations and extract local features through convolution operations for excellent performance in learning global structure and local details. The network

uses Transformer for object reconstruction and refines objects in combination with CNN. The 3D reconstruction network can be used to learn important features at different scales adaptively and deal with challenges brought by scale changes better. The experimental results show that the IoU value of 3D reconstruction on the public data set ShapeNet reaches 0.699. Compared with the pure convolutional neural network, the accuracy of this network has been significantly improved, showing better results.

To verify the effectiveness of the algorithm, a single view 3D reconstruction demonstration system is designed based on the comparison and analysis of relevant experiments. The system can generate 2D voxel images and 3D reconstruction models, and provide intuitive and convenient interaction. Users can rotate and scale the reconstructed model through the system to better observe the various angles and details of the object.

Key words: Three-dimensional reconstruction, Single view, CNN, Attention mechanism, Transformer

目　录

第一章 绪 论

1.1 研究背景

三维重建技术是近年来机器视觉领域的一个热门研究方向，其主要应用于从二维图像或视频中推断出三维信息。在过去的几十年中，已经研究出了许多基于多视角几何、双目视觉、深度传感器等方法进行三维重建的技术，同时也出现了一些基于深度学习的三维重建技术。早在 20 世纪 60 年代就有学者提出了利用多张照片恢复三维物体形态的想法。随着计算机技术和数字成像技术的发展，越来越多的三维重建技术被提出并得到应用。例如，基于三维扫描的重建方法可以通过激光、相机等传感器获取物体表面的三维坐标，并基于此生成三维模型。基于多视角几何的重建方法可以通过多个视角的图像计算出物体的深度信息，并根据深度信息构建三维模型[1]。另外，还有一些基于立体相机、结构光等技术实现三维重建的方法。随着人们对三维信息需求的日益增长，三维重建技术在许多领域得到广泛应用。例如，在文化遗产保护领域，三维重建技术可以帮助保存和恢复文物的形态信息，防止文物损毁或丢失。在建筑领域，三维重建技术可以用于建筑模型的设计、仿真和可视化。在医学领域，三维重建技术可以帮助医生更好地理解疾病的三维结构，提供更准确的诊断和治疗方案[2]。此外，三维重建技术还可以应用于虚拟现实、增强现实、自动驾驶等领域。

传统的三维重建技术主要基于多视角几何、双目视觉、结构光等原理，具有高精度和高可靠性等优点，但需要较长的计算时间和大量的数据处理[3]，对计算资源和算法效率提出了挑战。其中，基于多视角几何的方法可以通过将多个视角的图像进行匹配重建出三维模型，该法广泛应用于文物恢复、建筑建模等领域，为文化遗产的保护和数字化展示提供了重要支持。基于双目视觉的方法[4]则通过两个相机对同一场景进行拍摄并计算深度信息，这种方法在 3D 扫描、虚拟现实、机器人导航等领域得到广泛应用，为智能系统和自动化领域提供了关键支持。此外，基于结构光的方法通过投射特定光斑并结合相机拍摄进行三维重建，已成功地应用于工业检测、人脸识别等领域，这种技术的发展推动了生产效率的提升和安全性的加强，对工业制造和人机交互起到了重要作用。

随着深度学习技术的发展，越来越多的研究者开始尝试将其应用于三维重建领域。基于深度学习的三维重建技术具有计算速度快、可扩展性强等特点，可以直接从二维图像中提取三维信息，并可将其分为多视图三维重建以及单视图三维重建[5]。在多视图三维重建方面，深度学习网络能够有效地整合和利用多个视角的图像信息，通过端到端的训练和优化，实现对场景的高质量三维重建。相较之下，在单视图三维重建方面，深度学习网络则能够通过学习大量的图像数据，自动地从单张二维图像中提取出场景的三维信息，包括深度、表面法向等关键信息。这种方法使得单视图三维重建更加便捷和高效，极大地降低了获取三维信息的成本和门槛，而无需复杂的多视角数据。与此同时，深度学习的强大特征提取能力也使得单视图三维重建的结果更加准确和真实。这种基于深度学习的单视图三维重建方法为从单个图像中获取物体的丰富三维信息带来了新的可能性，尤其适用于需要快速获取物体三维信息或无法获取多个视角图像的情况，具有重要的实际意义和应用前景。

现如今，人们希望能从单张真实照片中快速、便捷地进行三维重建。然而，传统的三维重建技术效果不尽如人意，存在着重建结果不准确、需要大量的计算资源等问题[6]，这些问题部分源于单张图片无法提供完整的多视角信息，可能导致重建结果缺乏准确性和深度信息不足。此外，遮挡问题可能使得部分区域无法被充分重建，光照和阴影效应也会对重建结果产生影响，因而可能降低重建的精度和可靠性，限制了其在实际应用中的广泛应用。为了解决这些问题，近年来基于深度学习的图像三维重建算法备受关注，尤其是针对单幅真实图像的物体三维重建问题。因此，开展基于深度学习的单视图三维重建算法研究并设计实现适用于真实场景的三维重建系统具有重要的实际应用和学术研究价值。该项研究不仅能够推动三维重建技术在实际应用中的发展，同时也有助于深入理解深度学习在计算机视觉领域的潜力与局限性，为开拓图像三维重建领域的新可能性提供了重要的理论和实践基础。

1.2 国内外研究现状

三维重建作为计算机视觉领域的重要研究方向，旨在从二维图像中恢复出三维场景的几何形状、纹理、光照等信息[7]。然而，三维重建任务是一个复杂且具有挑战性的问题，因为它涉及到从二维图像中推断出三维空间中的信息，这

需要解决如遮挡、视角、光照等诸多难题[8]。近年来，随着深度学习技术的快速发展，三维重建技术也取得了显著的进步。深度学习技术为三维重建任务提供了强大的工具，它可以自动学习和提取图像中的特征，从而更好地恢复出三维场景的信息。国内外许多研究机构和学者都在致力于研究三维重建技术，并取得了一系列重要的研究成果。

2016 年，Choy[9]等人提出了一种用于三维物体重建的深度学习模型 3D-R2N2，它通过训练将 2D RGB 图像映射到 3D 体素空间中的概率分布，与传统的 3D 扫描技术相比，这种基于深度学习的方法只需要一个单独的图像即可实现三维重建，方便且高效。Wu[10]等人提出了一种基于生成对抗网络（GAN）的三维物体生成模型 3D-GAN，该模型采用了两个主要组件：一个生成器网络和一个判别器网络。生成器网络负责从随机噪声中生成逼真的三维物体模型，而判别器网络则负责评估生成器网络生成的模型的真实性。与传统的 GAN 模型专注于生成二维图像不同，3D-GAN 旨在生成逼真的三维物体模型。

2017 年，Wu[11]等人根据 David Marr 感知理论提出了一种端到端的三维重建网络 MarrNet，实现了从图像到 2.5D 特征以及三维模型的转换。Wang[12]等人提出了基于八叉树的卷积神经网络 O-CNN，结合了卷积神经网络（CNN）和八叉树（Octree）的数据结构，实现了高分辨率的三维重建。Gadelha[13]等人提出了一种专门用于超分辨率图像处理的生成对抗网络 PrGANs，采用了生成器和判别器之间的对抗性训练方式，以及使用残差连接和跳跃连接来提高网络性能。

2018 年，Tatarchenko[14]等人提出了一种基于八叉树的三维卷积神经网络 OGN，通过逐层遍历八叉树，O-CNN 可以自动学习并提取三维点云数据的特征和上下文信息，从而生成高质量的三维模型。Yao[15]等人提出一种基于多视图图像的深度估计网络 MVSNet，采用多视图几何约束和光学一致性来构造匹配代价，进行匹配代价累积，并估计深度值，旨在从具有一定重叠度的多视图视角中恢复场景的稠密结构。

2019 年，Mescheder[16]等人提出了 OccNet 网络，用于恢复物体在任意分辨率下的连续形状。Gkioxari[17]等人提出了 Mesh R-CNN，基于 mask R-CNN 框架来有效地检测图像中的目标，并为每个实例生成高质量的分割掩码。Wang[18]等人提出了 Pixel2Mesh，将输入图像中的每个像素映射到潜在空间中的一个向量，并将潜在空间中的向量映射回图像空间，生成三维网格的顶点坐标。Xie[19]等人提出了 Pix2Vox，采用一种名为条件生成对抗网络的神经网络结构构建了一个生

成器和一个判别器，并通过对抗损失函数来优化模型，使其可以处理不同视角下的二维图像，生成高质量的三维模型。

2020 年，Wen[20]等人对 Pixel2Mesh 进行了改进，提出了 Pixel2Mesh++，在保持 Pixel2Mesh 优点的基础上，通过引入轻量级 CNN 和注意力机制等技术，进一步提高了三维重建的精度和效率。Gu[21]等人提出了 Cascade-MVSNet 网络，通过级联的方式进行深度估计，逐步细化深度估计结果，提高准确性。Xie[22]等人提出了 Pix2Vox++，采用深度表面几何网络结构来进一步提高生成三维模型的精度和准确性，通过逐层编码和解码的方式，将 2D 图像转化为具有高精度的 3D 网格结构。

2021 年，Zhao[23]等人提出了一种从单个图像中增强三维点云重建的方法，旨在通过关注图像中的边界和角落点，更精确地重建三维对象。Shi[24]等人提出了一种基于 Transformer 的三维重建网络 3D-RETR，证明了 Transformer 在三维重建方面的潜能。Wang[25]等人提出了一种面向目标的隐式曲面重建方法与 Transformer 模型相结合的模型 EVoiT，通过端到端的训练实现了高效的三维重建。

2022 年，Leslie[26]等人提出了一种轻量级的网络结构 3D-C2FT，引入了一种动态特征选择机制，根据输入图像的特征分布来动态选择合适的特征进行重建。Zhu[27]等人首次提出全局感知的多视图三维重建网络 GARNet，采用全局先验网络捕捉全局上下文信息，局部细节重建网络以 U-Net 结构为基础，结合全局上下文信息与局部特征，实现准确的三维重建。

尽管近年来提出的三维重建网络技术在该领域展示出了显著的发展潜力，但它们在处理具有复杂几何形状和高细节密度的对象时仍遇到了显著的挑战。对于具有高度复杂几何结构的对象，现有方法往往难以准确捕捉其整体形状和拓扑结构，尤其是在仅依赖于从有限视角获得的二维图像信息进行重建时。同时，许多网络对于重建物体的局部细节的捕捉能力不足，无法精确重建出局部的细节特征，如锐利的边缘、细小的纹理。

1.3 主要研究内容

本文主要研究基于深度学习的单视图重建算法，目的是使用单张二维图像重建出物体的三位体素特征，通过融合当前的先进算法，提升传统三维重建算

法对物体重建的准确度。

本文的主要内容如下：

(1)基于分组卷积编码的单视图三维重建网络研究

搭建分组注意力增强特征提取模块。针对在特征提取阶段中对重要特征的提取和利用能力不足的问题，通过将分组注意力增强特征提取模块嵌入到在三维重建网络编码器中，有效增强了特征表达的能力，有助于模型更准确地理解和表达数据中的高级语义信息。

构造三维模型细节增强器。为改善二维特征在经过解码器后生成的三维模型比较粗糙问题，设计三维模型细节增强器，提升特征建模和泛化能力，使其适应复杂任务并获得更好效果。

提出阈值自适应策略。动态调整三维重建任务中的阈值，将检测到的物体信息纳入重建过程，可以有效降低误检率，实现更准确、完整的三维模型重建。

(2)单视图三维重建中的 Transformer 架构设计

基于 Transformer 的二维图像特征提取网络。通过位置编码，模型能够将位置信息融入输入特征中，从而更好地处理序列数据，同时引入的 HiLo 注意力机制有效解耦了高频和低频特征，提高了模型对不同频率特征的建模能力，Transformer 驱动三维重建网络在二维图像特征提取中展现出更强的全局上下文理解和适应性。

三维模型重建细节的关键技术优化。通过设计全新的精炼器模块，采用了残差卷积结构、密集连接以及通道、空间注意力机制等技术，能够更好地提取三维模型的细节信息，生成更具精确度的三维模型。

Dice 损失函数。其在处理重叠区域时表现优异，能够更准确地评估重建结果的准确性，尤其对于需要考虑目标边界和空间位置关系的三维重建任务具有显著优势。

(3)单视图三维重建系统的设计与实现

该系统能够将单图像转化为具有立体感的三维模型，并提供交互操作界面，使用户可以自由地旋转、放大、缩小模型，并获取相关信息。这样的系统不仅能够使普通用户直观观察和理解三维重建结果，还有助于促进该技术在实际应用中的推广与应用。

1.4 论文组织结构

本文的结构安排如下：

第一章：绪论。该章节的主要内容包括三个方面：研究背景及意义，国内外相关研究成果的总结以及本章研究内容。在研究背景及意义方面，章节首先介绍了本研究所涉及的领域、研究的重要性和实际应用意义。接着总结了国内外相关研究成果的过往及现状，旨在为读者提供一个研究领域的概览和背景知识，以便更好地理解本研究的价值和创新点。在本章节研究内容方面，章节详细阐述了本研究的研究内容，包括研究的对象、研究的方法和技术路线等方面。

第二章：基于深度学习的三维重建相关理论与技术。该章主要分析了与本文研究内容相关的注意力机制以及三维重建相关理论技术，为后续的研究打下基础。

第三章：基于分组卷积编码的单视图三维重建算法研究。该章节详细阐明了所提出算法的整体框架，包括网络结构、损失函数、阈值调整模块、数据集以及算法的实现细节等内容，使读者能够清晰地理解算法的设计思路和关键技术。最后通过展示实验结果并进行详细分析和讨论，呈现了本算法在单视图三维重建方面的实际效果和潜在优势。

第四章：单视图三维重建中的 Transformer 架构设计。本章首先分析了 Transfomer 架构的基本原理和在三维重建任务中的应用潜力，随后设计了结合 CNN 和 Transfomer 的混合网络结构，同时在该网络中引入双通道注意力机制的创新点。此外，采用 Dice 损失函数替代 BCE 损失函数，最终通过实验验证了该架构在任务性能上的优势。

第五章：单视图三维重建可视化系统的设计与实现。本章依据前文研究成果，设计了一个三维重建的可视化系统，系统通过图像来生成重建后的二维体素图像以及三维重建模型。并对系统的需求、系统设计、系统实现以及系统展示等内容做了详细介绍。

第六章：总结与展望。本章对全文进行归纳总结，概括论文的核心结论和研究成果，同时提出后续可改进的思路和可能的发展方向，以供后续相关研究参考。

第二章　基于深度学习的三维重建相关理论与技术

基于深度学习的三维重建是计算机视觉领域的一项关键技术，它能够从二维图像中恢复出物体的三维结构信息。随着深度学习技术的快速发展，这一领域已经取得了显著的进步，极大地推动了三维视觉研究和应用的前沿发展。该技术依托于强大的深度神经网络，通过学习大量的图像与其对应的三维模型之间的关系，实现了从简单的二维图像到复杂的三维结构的转换。这一过程不仅涉及到复杂的模型训练和优化策略，还包括针对不同应用场景的定制化设计。基于深度学习的三维重建技术已经被广泛应用于自动驾驶、虚拟现实、文化遗产保护等多个领域，展现出巨大的应用潜力和价值。

本章旨在探讨基于深度学习的三维重建技术的相关理论和方法，并深入剖析其在计算机视觉领域的重要意义和应用前景。首先，重点分析了注意力机制的相关理论技术，如自注意力机制、通道注意力机制、空间注意力机制和多头注意力机制，这些机制在三维重建中起着至关重要的作用。其次，详细讨论了基于深度学习的三维重建相关理论技术，包括基于体素、基于网格和基于点云的方法。通过对这些内容的深入探讨，揭示了深度学习在三维重建领域的应用前景和挑战，为该领域的发展提供了全面的理论基础和实践指导。

2.1　注意力机制相关理论技术

注意力机制（Attention Mechanism）[28]是一种在深度学习中广泛应用的技术，旨在模拟人类的视觉或听觉系统中的注意机制。它通过对输入的不同部分分配不同的权重，将重点放在与当前任务相关的信息上。注意力机制的核心思想是根据输入的上下文信息，计算每个输入位置的重要性，并将这些重要性用于加权求和或者生成对应的权重向量。这样可以使模型更加专注于与当前任务相关的信息，提高模型的表达能力和泛化性能。

在自然语言处理领域，注意力机制被广泛应用于机器翻译、文本摘要、问答系统等任务中。其中最为经典的是 Transformer 模型[29]，它引入了多头注意力机制，这一创新极大地提升了模型对输入文本的理解能力。通过多头注意力机

制，Transformer 能够同时关注输入序列中不同位置的信息，捕捉到不同粒度的语义信息，从而提高了模型在处理长距离依赖和建模全局上下文信息时的效果。这种机制使得 Transformer 模型能够更好地处理输入序列中的关键信息，并实现不同长度句子之间的有效信息交互。

在计算机视觉领域，注意力机制被用于图像分类、目标检测、图像生成等任务中。在图像分类任务中，注意力机制使模型能够专注于图像中的关键区域，提高分类的准确性。通过引入注意力机制，模型可以学习识别和关注与类别判定密切相关的局部区域或特征，而非简单地平均考虑整个图像。

2.1.1　自注意力机制

自注意力机制（Self-Attention Mechanism,SA）[30]是一种用于序列建模和自然语言处理等任务的常见注意力机制。用于在序列建模和自然语言处理等任务中对输入进行编码。它通过将输入序列中的每个位置视为查询、键和值，并计算它们之间的相关性来为每个位置生成一个上下文向量，如图 2-1 所示。

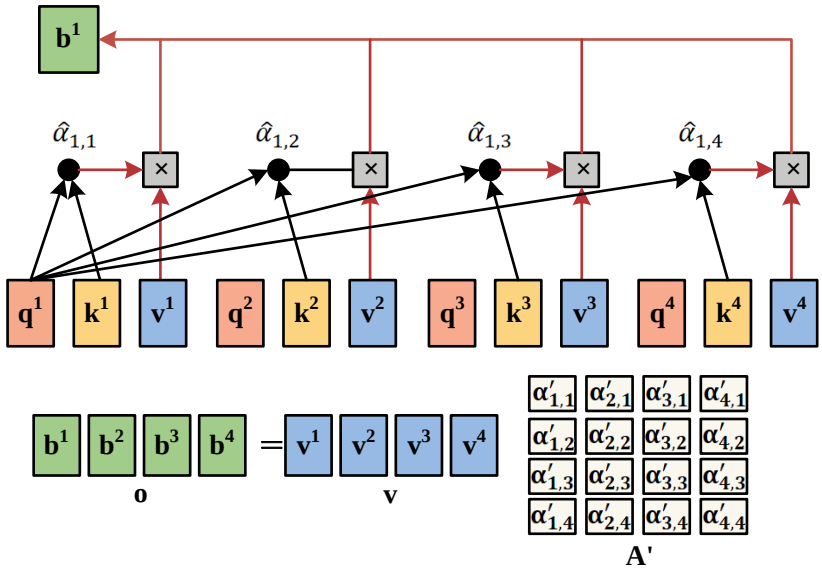


图 2-1　自注意力机制

具体而言，自注意力机制可以分为映射、相关性计算、归一化和加权求和、逆映射四个步骤。自注意力机制具有捕捉序列中不同位置之间依赖关系的优势，并且不受序列长度的限制，能够有效地处理长序列。它还可以自适应地学习不同的输入表示，提高了模型的表达能力。给定输入序列 $X=[x_1,x_2,...,x_n]$，其中 n 是序列的长度。自注意力机制通过线性变换将输入序列映射到查询(Q)、键(K)

和值(V)空间，如公式（2-1）所示。

$$Q = XW_Q, K = XW_K, V = XW_V \tag{2-1}$$

其中W_Q、W_K和W_V是可学习的参数矩阵。然后，计算每个查询向量q_i与所有键向量k_j的相关性得分s_{ij}如公式（2-2）所示。

$$s_{ij} = q_i \cdot k_j \tag{2-2}$$

接下来，对相关性分数进行归一化，并按照归一化后的权重对值向量进行加权求和，得到每个查询向量的上下文向量如公式（2-3）所示。

$$a_{ij} = \frac{e^{s_{ij}}}{\sum_{k=1}^{n} e^{s_{ik}}}$$

$$c_i = \sum_{j=1}^{n} a_{ij} v_j \tag{2-3}$$

其中a_{ij}是归一化后的权重，v_j是第j个值向量。最后，通过逆映射将上下文向量映射回原始空间如公式（2-4）所示。

$$Y = cW_O \tag{2-4}$$

其中W_O是可学习的参数矩阵，Y是最终的输出序列。

自注意力机制在自然语言处理和图像处理领域都有广泛的应用。在自然语言处理领域，它被广泛应用于机器翻译、文本摘要和语言模型等任务中。在图像处理领域，自注意力机制也被用于图像描述生成和图像分类等任务。自注意力机制的出现解决了传统循环神经网络在处理长序列时出现的梯度消失和梯度爆炸问题，成为序列建模的重要技术。

2.1.2　多头注意力机制

多头注意力机制（Multi-Head Attention,MHA）[31]是在自注意力机制的基础上发展起来的，是自注意力机制的变体，旨在增强模型的表达能力和泛化能力如图 2-2 所示。

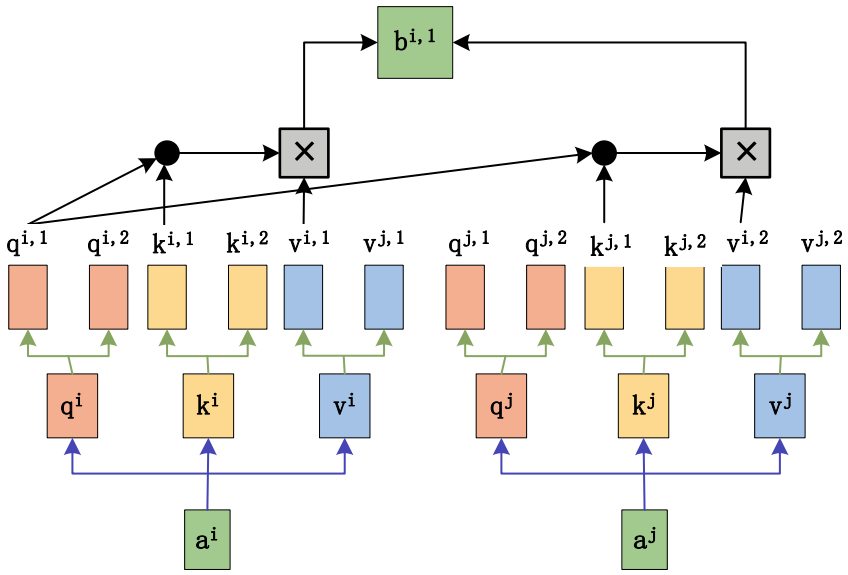


图 2-2　多头注意力机制

通过使用多个独立的注意力头，分别计算注意力权重，并将它们的结果进行拼接或加权求和，从而获得更丰富的表示。给定输入序列 $X=[x_1,x_2,...,x_n]$，其中 x_i 表示序列中第 i 个元素，如公式（2-5）所示。

$$Q=XW_Q,K=XW_K,V=XW_V \tag{2-5}$$

其中 W_Q、W_K 和 W_V 是可学习的参数矩阵，用于将输入序列映射到查询、键和数值向量空间。接着使用 Q 和 K 的点积作为注意力权重的分数，通过 Softmax 函数进行归一化，然后将注意力权重应用于数值向量 V，如公式（2-6）所示。

$$Attention(Q,K,V)=Softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{2-6}$$

然后通过并行计算多个头部的注意力权重，得到多组注意力输出。其中 Q_i、K_i 和 V_i 分别表示第 i 个头部的查询、键和数值向量，如公式（2-7）所示。

$$Head_i=Attention(Q_i,K_i,V_i) \tag{2-7}$$

最后使用 concatenate 函数将多个头部的输出连接起来，W_O 是输出映射的权重矩阵，用于将多头输出映射回原始向量空间，如公式（2-8）所示。

$$MultiHeadOutput=concatenate(Head_1,Head_2,...,Head_h)W_O \tag{2-8}$$

多头注意力机制在自然语言处理领域中具有重要意义。其优势在于可以让模型同时关注输入的不同部分，并从多个角度捕捉输入之间的关联信息。这种并行计算多组注意力权重的方法使得模型能够更好地处理长距离依赖关系，有效地减轻了传统注意力机制中存在的注意力瓶颈问题。通过多头注意力，模型

可以在不同“头”上进行并行计算，从而实现对不同维度的语义信息进行捕捉和表征。这样的设计使得模型能够更全面地理解输入序列，提高了模型对输入文本的建模能力，并且在一定程度上提升了模型的泛化性能。因此，多头注意力机制为处理自然语言处理任务提供了更为灵活和强大的工具，使得模型能够更好地捕捉文本中的重要信息，从而在机器翻译、文本摘要、问答系统等任务中取得更好的性能表现。

2.1.3　通道注意力机制

通道注意力机制（Channel Attention,CA）[32]是一种用于增强卷积神经网络性能的注意力机制。它通过对输入特征图的通道维度进行注意力加权，使网络能够自适应地关注不同通道之间的相关性和重要性。给定输入特征图 $X \in \mathbb{R}^{C\times H\times W}$，其中 C 是通道数，H 和 W 分别是特征图的高度和宽度。如公式（2-9）所示。

$$Avgpool = \frac{1}{H\times W}\sum_{i=1}^{H}\sum_{j=1}^{W} X[:,i,j] \tag{2-9}$$

这里对每个通道的特征图进行了全局平均池化操作，得到一个维度为 C 的向量 $Avgpool \in \mathbb{R}^{C}$。接着，通过两个并行的线性变换将全局平均池化后的结果映射为查询 (Q)、键 (K) 和值 (V)，如公式（2-10）所示。

$$Q = W_Q \cdot Avgpool, K = W_K \cdot Avgpool, V = W_V \cdot Avgpool \tag{2-10}$$

其中 W_Q、W_K 和 W_V 是可学习的参数矩阵。最后，通过 Softmax 函数计算查询向量 Q 与键向量 K 的相似性得分 S，然后对值向量 V 进行加权求和得到最终的特征表示 $output$，如公式（2-11）所示。

$$\begin{aligned} S &= Softmax\left(Q\cdot K^{T}\right) \\ output &= S\cdot V \end{aligned} \tag{2-11}$$

这种注意力机制可以自适应地学习不同通道之间的相关性和重要性，从而有效地提升了网络的特征表达能力。通过通道注意力，网络可以动态地调整不同通道的权重，突出对当前任务更为重要的特征通道，抑制对任务无关或冗余的特征通道，进而提高了网络对输入数据的表征能力和泛化能力。这种机制使得网络更加灵活和智能，能够根据不同的输入数据动态地调整特征表达，从而更好地适应复杂多变的任务需求。

2.1.4　空间注意力机制

空间注意力机制（Spatial Attention,SA）[33]和通道注意力机制具有异曲同工之妙，通道注意力机制旨在捕捉通道的重要性的程度，空间注意力机制旨在通过引入注意力模块，使模型能够自适应地学习不同区域的注意力权重。这样，模型可以更加关注重要的图像区域，而忽略不重要的区域。给定输入特征图 $X \in \mathbb{R}^{C\times H\times W}$，其中 C 是通道数，H 和 W 分别是特征图的高度和宽度。如公式（2-12）所示。

$$Z = AvgPool2d(X) \tag{2-12}$$

这里，$AvgPool2d$ 表示在空间维度上对输入特征图进行全局平均池化操作，得到一个长度为 C 的向量。然后通过 Softmax 函数用于对全连接层的输出进行归一化，得到空间注意力权重，如公式（2-13）所示。

$$Y_{i,j} = X_{i,j} \cdot Softmax\left(FC\left(Z\right)_{i,j}\right) \tag{2-13}$$

其中，FC 是一个全连接层，用于学习空间注意力权重，Y 表示加权后的输出特征图，i 和 j 分别表示特征图的竖直和水平位置。

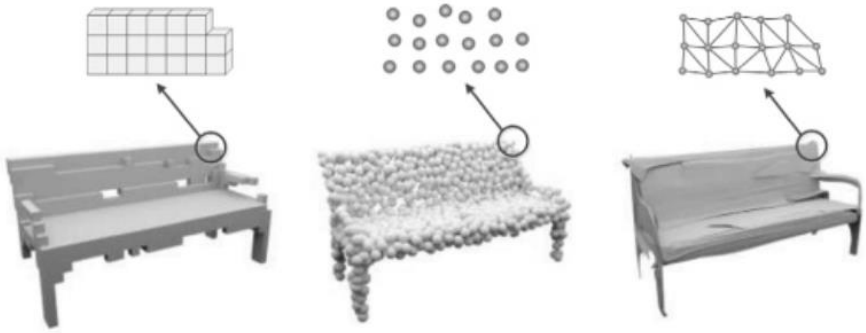
通过以上步骤，输入特征图 X 中的每个像素都会乘以相应的注意力权重，从而突出重要的空间位置信息。这样的注意力机制可以帮助网络自动学习不同空间位置之间的相关性和重要性，并提升特征表示的能力。

通过空间注意力，模型可以有效地捕捉输入数据中不同位置之间的关系，从而提高了对空间结构的建模能力。这种机制使得模型能够更加准确地处理图像、视频等具有空间特征的数据，从而在各种计算机视觉任务中取得更好的性能表现。空间注意力机制的引入也有助于模型对局部细节和全局上下文进行有效的整合，进而提升了模型对输入数据的表示学习能力。

2.2　基于深度学习的三维重建相关理论技术

三维重建的目标是从单张二维图像或多张二维图像中重建出物体和场景的三维模型，并对三维模型进行纹理映射。三维重建是计算机视觉领域的一个重要研究方向，利用计算机重建出物体的三维模型，已经成为众多领域进行深入研究前不可或缺的一部分。三维模型的表示形式有三种：体素模型、点云模型和网格模型[34]。体素是三维空间中的正方体，类似于三维空间中的像素，用于

表示物体的体积和密度分布；点云是坐标系中的点的集合，包含了三维坐标、颜色、分类值等信息，适用于从激光扫描等设备获取的三维数据；网格是由多个三角形组成的多面体结构，可以表示复杂物体的表面形状，常用于渲染和建模。三维模型的表示形式如图 2-3 所示。



(a)体素模型　　(b)点云模型　　(c)网格模型

图 2-3　三维模型的表示形式

根据三维模型的表示形式，图像三维重建方法可以分为基于体素的三维重建、基于点云的三维重建和基于网格的三维重建。在基于网格的三维重建方法中，又可以细分为单一颜色的网格三维重建和具有色彩纹理的网格三维重建。根据输入图像的类型，图像三维重建方法可以分为单张图像三维重建和多张图像三维重建。如图 2-4 所示，展示了图像三维重建方法的分类。

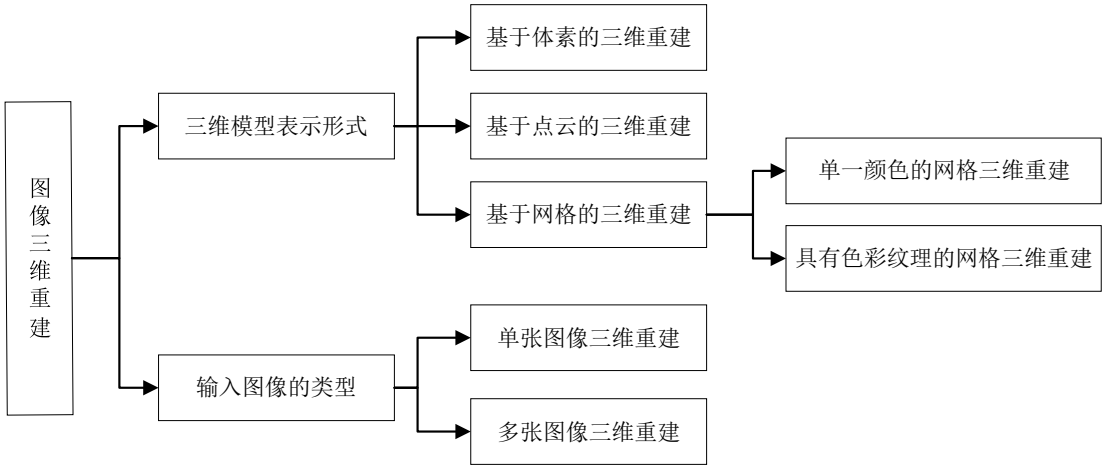


图 2-4　图像三维重建方法的分类

2.2.1　基于体素的单张图像三维重建

基于体素的单张图像三维重建是一种常见且有效的方法，它通过对单张图像进行分析和处理，生成包含对象的三维体素表示。该方法的原理是将三维空

间划分为规则的小立方体单元，称为体素[35]。体素表示了物体在三维空间中的位置和形状，为实现三维重建提供了重要的基础。

在使用体素作为重建结果的网络中较为经典的是 3D-R2N2 网络，其结构如图 2-5 所示。3D-R2N2 网络的基本结构是一个卷积神经网络，它由两个主要部分组成：编码器和解码器。编码器负责将输入的二维图像转换为低维度的特征向量，而解码器则将这些特征向量转换为对应的三维体素网格。

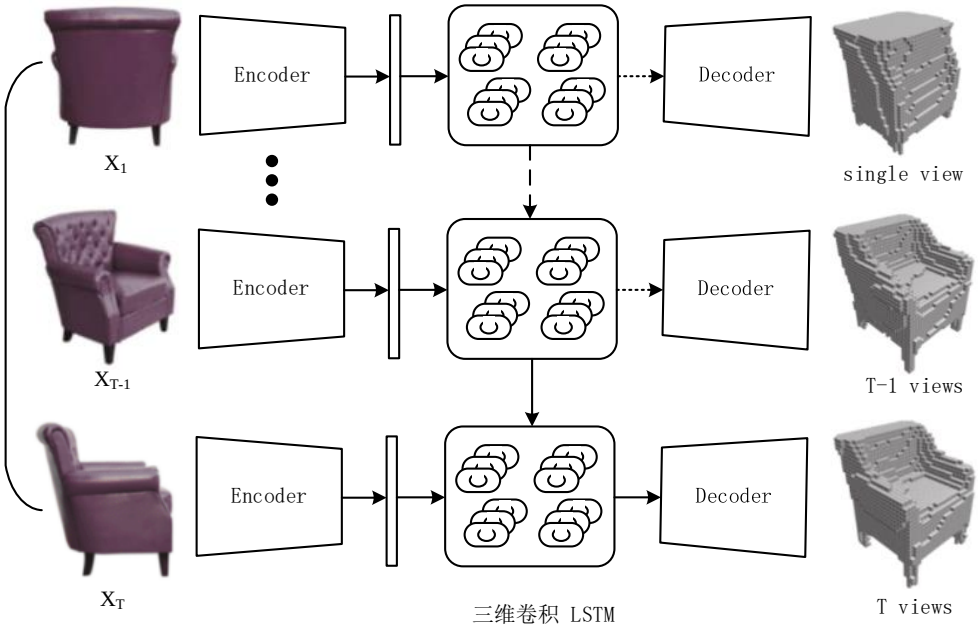


图 2-5　3D-R2N2 网络架构

3D-R2N2 网络的优点在于可以从单个或多个二维图像中生成三维物体，而无需任何先验知识。此外，它还可以处理不同形状和大小的物体，并且可以在多个领域进行应用。

2.2.2　基于点云的单张图像三维重建

基于点云的单张图像三维重建是一种先进的技术，它能够从单张图像中推断出物体的三维结构。相比于传统的多视图或立体视觉方法，基于点云的重建方法更加灵活和高效。该方法在许多领域有广泛应用。例如，在计算机视觉领域，它可以用于物体识别和姿态估计，帮助机器理解和感知环境[36]。在计算机图形学领域，它可以用于虚拟现实和游戏开发，创建逼真的虚拟场景。此外，基于点云的三维重建还可以应用于文化遗产保护、工业制造和医学影像等领域，为相关应用提供精确的三维信息。

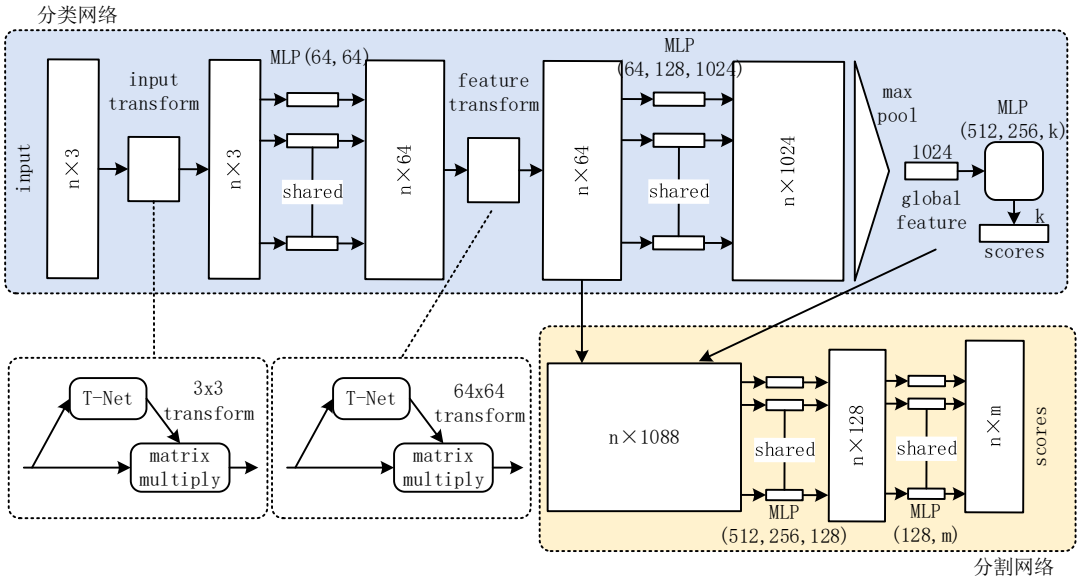


图 2-6　PointNet 网络架构

在使用点云作为重建结果的网络中较为经典的是 PointNet 网络[37]。其结构如图 2-6 所示。PointNet 的主要创新之处在于其直接处理点云数据的能力，以及对点云内点排列顺序不变性的处理方式。PointNet 网络通过其独特的架构设计，有效地解决了点云数据在三维空间中固有的无序性问题。这一突破性的创新不仅提高了点云处理的效率和准确性，为后续的研究提供了新的视角和方法。

2.2.3　基于网格的单张图像三维重建

基于网格的单张图像三维重建是一种常见的方法，用于从单张二维图像中恢复出物体的三维结构。这种方法结合了计算机视觉和计算机图形学的技术，能够将平面图像转换为具有空间信息的三维网格模型[38]。该方法广泛应用于计算机视觉和计算机图形学领域。在计算机视觉方面，它可以用于物体识别、姿态估计和目标跟踪等任务。通过恢复物体的三维结构，可以提供更多的几何信息，从而改善这些任务的效果。在计算机图形学方面，基于网格的单张图像三维重建可以用于虚拟现实、游戏开发和电影制作等领域。通过将真实世界中的物体转化为具有纹理和细节的三维模型，可以构建逼真的虚拟环境，并提供更加沉浸式的用户体验。

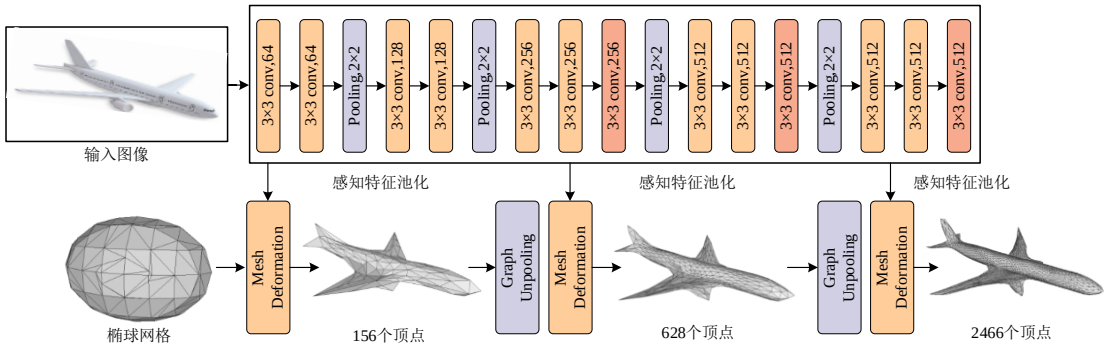


图 2-7　Pixel2Mesh 网络架构

在使用网格作为重建结果的网络中较为经典的是 Pixel2Mesh 网络。其结构如图 2-7 所示。Pixel2Mesh 网络是一种基于深度学习的三维重建模型，它可以从单张二维图像生成高质量的三维网格模型。相比传统的基于体素的方法，Pixel2Mesh 网络更加适用于处理复杂形状和表面细节。

Pixel2Mesh 网络的优点在于它可以生成高质量的三维网格模型，能够处理复杂的形状和表面细节。此外，通过迭代优化的方式，它可以逐步改善生成的网格，提高重建结果的准确性和细节保真度。

2.3　本章小结

深度学习作为人工智能领域的重要分支，其相关理论技术在本章中得到了深入分析和探讨。

本章详细分析了注意力机制的不同类型，包括自注意力机制、多头自注意力机制、通道注意力机制和空间注意力机制。这些机制在图像处理、自然语言处理等领域广泛应用，能够帮助模型更好地理解和关注输入数据的相关部分，从而提升模型的表现和效果。通过深入学习这些注意力机制，可以为解决复杂的任务和提升模型性能提供新的思路和方法。

另外，本章深入探讨了基于深度学习的三维重建技术，包括体素重建、点云重建和网格重建等方法。这些技术可应用于计算机视觉、机器人视觉和虚拟现实等领域，为实现精准的三维重建提供了重要技术支持。

第三章　分组卷积编码的单视图三维重建算法研究

3.1　引言

在计算机视觉领域，三维重建是一项重要的任务，它可以从二维图像中恢复出物体的三维形状和结构信息。传统的三维重建方法通常需要多个视角的图像或大量的标注数据，限制了其在实际应用中的可行性。随着深度学习的快速发展，基于单个视角的三维重建方法逐渐受到关注。这种方法借助于卷积神经网络（CNN）等深度学习技术，通过从单张图像中学习到的特征来进行三维重建，这使得三维重建更加灵活和高效，并在许多应用领域取得了显著的成果。然而，基于单视图的三维重建仍然面临一些挑战。例如，由于单个视角的信息受限，重建结果可能存在模糊或不完整的问题。

为了克服这些挑战，本研究提出了一种分组卷积编码的单视图三维重建算法[39]。通过改进编码器、解码器以及精炼器网络结构，并优化重建过程中的阈值选择策略，使得网络获取的特征更精确、恢复的三维模型细节更丰富，来达到增强模型表达能力和性能的目的，从而提升三维重建任务的精确性。

本章研究内容如下：

(1)对编码器的特征提取网络进行改进，设计了分组注意力增强特征提取模块并将其嵌入编码器中，有效提升特征表达的判别性、可区分性和泛化能力，从而增强模型对数据中的高级语义信息的抽象和表达能力。

(2)优化解码器生成的三维模型，构造三维模型细节增强器，有效地提升模型在特征建模、特征整合和泛化能力等方面的性能，以解决二维特征在经过解码器后生成的三维模型比较粗糙问题。

(3)设计了阈值自适应策略，根据输入数据的特点和复杂度，灵活地控制保留物体细节的程度，有助于避免过度平滑或过度保留细节的问题，提高重建结果的质量和真实感，并具备更好的适应性和鲁棒性。

3.2　算法设计

3.2.1　分组卷积编码的多尺寸融合网络

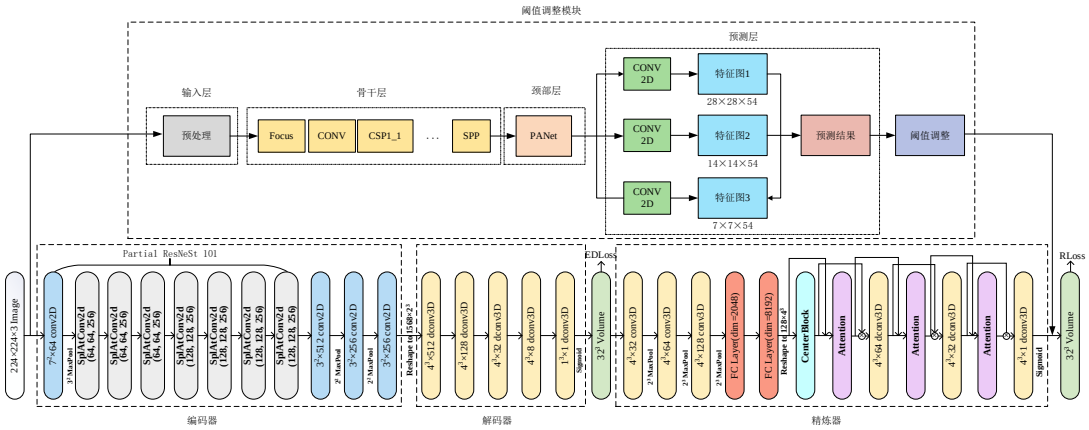


图 3-1　3D-PCCF 网络结构

本章网络命名为 3D-PCCF（Three-Dimensional Progressive Convolutional Coding Fusion），结构如图 3-1 所示，网络编码器引入分组注意力机制来提高特征提取和分类的性能；精炼器中利用注意力机制以及特征拼接的方式来更好地反映对象的形状、纹理和空间关系，从而提升三维模型的重建质量和细节还原能力；阈值调整模块则使用 YOLO 对输入图像的信息进行识别，根据识别结果选择性地调整重建时的阈值，以达到最优的重建效果。

在考虑整个网络的正确性和复杂性时，首先需要验证网络结构的合理性。这涉及到确保编码器能够有效地提取输入图像的关键特征，同时解码器可以准确地利用这些特征重建三维模型。正确性证明可以通过对比实验来完成，比如检验模型在标准数据集上的性能表现，以及观察加入注意力机制前后的性能差异。这些实验结果通常显示，通过精心设计的网络结构和优化的模块配置，模型能够有效地完成三维重建任务，并且在处理细节方面有明显提升。

在复杂性分析方面，需要综合考虑计算和存储需求。计算复杂度主要涉及到整个网络在前向传播和反向传播过程中的计算负荷，包括各层的卷积计算、激活函数处理和梯度下降等。存储复杂度则关注于网络参数的总量和中间输出的存储需求。

3.2.1.1　基于层级表示学习的图像特征提取

传统的注意力机制是根据输入的权重来调整每个位置或特征在模型中的贡献，使模型能够在处理输入时更加集中地关注那些与任务相关的信息。通常使用软注意力的形式计算每个位置或特征的注意力权重并给予它们不同的重要

性。相比之下，使用分组注意力机制通过对输入特征图的通道维度进行注意力加权，自适应地调整不同通道的权重，从而增强对重要特征的提取和利用能力，并有效改善模型的表达能力。通过共享注意力机制的方式，避免了为每个通道学习独立的权重参数，极大地减少模型的参数量，提升模型的轻量化和计算效率。除此之外，SplAtConv2d 能够更加灵活地对特征进行非线性调整，使得模型更好地理解数据中的关联和特征重要性，并将其有效应用于建模过程中，来实现更准确的预测和推断。模型结构如图 3-2 所示，结构如公式（3-1）所示。

$$
\begin{gathered}
\mathrm{x1}, x2, \ldots, xr = split(x, r) \\
m_i = AvgPool\left(x_i\right) \\
a = FC_2\left(FC_1\left(m_1\right); FC_1\left(m_2\right); \ldots; FC_1\left(m_r\right)\right) \\
a_1, a_2, \ldots, a_r = split\left(a, r\right) \\
y = \sum_{i=1}^{n}\left(Conv2D\left(x_i, a_i\right) + b_i\right)
\end{gathered}
\tag{3-1}
$$

其中 $x \in \mathbb{R}^{H \times W \times C}$ 代表输入特征，r 为特征划分的数量，$AvgPool$ 表示平均池化操作，$x_i \in \mathbb{R}^{H \times W \times \frac{C}{r}}$ 划分后的特征，$m_i \in \mathbb{R}^{\frac{C}{r}}$ 为经过平均池化后的特征，$FC_1 \in \mathbb{R}^{C'}$、$FC_2 \in \mathbb{R}^{H \times W \times in\_channels \times r}$ 代表全连接操作，$a_i \in \mathbb{R}^{H \times W \times in\_channels}$ 代表划分后的特征，$in\_channels$ 表示通道数，$Conv2D$ 代表 2D 卷积操作，b_i 表示第 i 组 2D 卷积层的偏置项，$y \in \mathbb{R}^{H \times W \times C}$ 为最终得到与输入特征维度相同的特征。

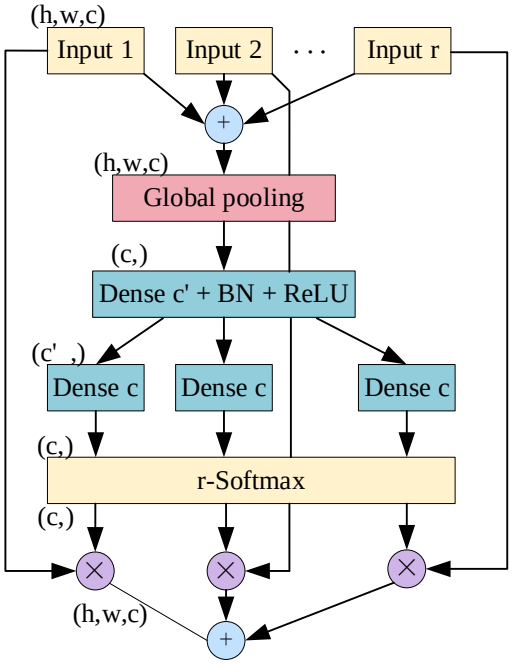


图 3-2　SplAtConv2d 结构

此模块用于分组处理输入数据，并计算每个分组的平均值。然后通过全连接层对输入的平均池化结果进行线性变换，可以理解为对输入特征进行降维或提取更高级别的特征表示，从而帮助网络更好地理解输入数据。其次将输出进行分割，有助于将不同来源的特征或信息进行组合和整合。最后通过卷积操作可以在局部区域上提取特征，并且参数共享机制使得卷积具有平移不变性，可以进一步处理输入数据，捕捉更复杂的特征。

3.2.1.2　注意力驱动的三维模型细化

由于经过解码器生成的三维模型比较粗糙，部分物体的边缘细节得不到很好的处理，因此加入了注意力机制用以改善三维模型的质量和细节，通过在神经网络中添加一个中心块，用于进行特征提取和上采样操作。具体来说，它在网络中间位置对输入的特征图进行处理，以提取更高级的抽象特征，并进行上采样以恢复细节。这对于处理具有复杂结构和丰富纹理的数据，如三维体积数据或图像数据，特别有帮助。结构如公式（3-2）所示。

$$\begin{aligned} y_1 &= ReLU\left(BatchNorm\left(Conv3D\left(x\right)\right)\right) \\ y_2 &= ReLU\left(BatchNorm\left(Conv3D\left(y_1\right)\right)\right) \\ y_3 &= ReLU\left(BatchNorm\left(Conv3DTranspose\left(y_2\right)\right)\right) \\ y &= MaxPool\left(y_3\right) \end{aligned} \tag{3-2}$$

其中 $x \in \mathbb{R}^{D' \times H' \times W' \times C''}$ 代表输入特征，$y_1 \in \mathbb{R}^{D' \times H' \times W' \times C}$ 代表经过 3D 卷积操作后的特征，C、C'、C'' 均表示输出通道数，*Conv3D* 表示 3D 卷积操作，*BatchNorm* 表示批处理操作，*ReLU* 表示 ReLU 激活函数，*Conv3DTranspose* 表示 3D 转置卷积操作，*MaxPool* 表示最大池化操作，$y \in \mathbb{R}^{D' \times H' \times W' \times C''}$ 为最终得到与输入特征维度相同的特征。这一部分用于对输入特征进行处理并生成与其维度相同的输出特征。它通过 3D 卷积操作提取输入数据的局部特征，然后通过批处理、ReLU 激活函数、转置卷积和最大池化等操作进行特征加工和维度变换，最终得到丰富、高维度的输出特征。这个网络结构能够有效地捕捉输入数据的空间关系和层级特征，并通过融合和整合操作生成更全面和表达力强的特征表示。模块结构如图 3-3 所示

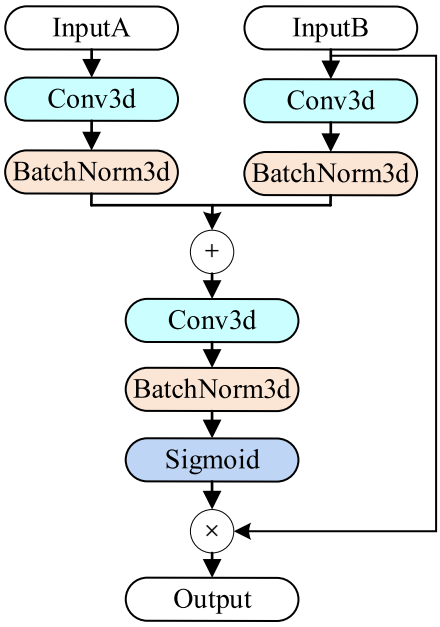


图 3-3　注意力机制模块图

经过强化后的特征将通过不同的卷积和归一化操作进行融合，然后通过 Sigmoid 函数生成权重，最后与原始特征相乘得到增强的特征。这样可以使模型更关注对当前任务而言重要的特征，提高三维模型对边缘细节部分的细化能力。同时在网络进行输入特征融合的过程中，通过 ReLU 激活函数保留了对应位置上两个分支的共同关注的特征，并使用 Sigmoid 函数生成的权重来动态调整不同位置的重要性，从而增强对不同种类物体三维重建的自适应性，提高重建物体各个部分的关联性和准确性。

3.2.1.3　三维模型重建中的阈值自适应策略

本章提出的阈值调整模块可以动态调整三维重建任务过程中的阈值，以适应不同图像和任务的需求。模块选用 YOLO 作为识别网络可以很好的提高网络的适应性和鲁棒性，并保证物体识别的准确性，使其在不同场景和条件下都能保持较好的性能。通过将检测到的物体信息纳入三维重建过程中，可以更好地还原物体的特征和细节，提供更准确、完整的三维模型。同时可以有效降低系统的误检率，使其在生成三维模型时能够更准确地还原图像的真实形状和结构。它拥有输入层、骨干层、颈部层和预测层四个部分。

输入层完成对数据的预处理，如公式（3-3）所示。

$$a = f_{Anchor}(x) \in \mathbb{R}^{9\times4}$$
$$x' = f_{Mosaic}(x) \in \mathbb{R}^{672\times672\times3} \tag{3-3}$$

$x \in \mathbb{R}^{224\times224\times3}$ 表示输入图像，f_{Anchor} 表示锚框尺寸计算，f_{Mosaic} 表示数据增强操作。这个预处理过程的作用在于，对输入数据进行有效的预处理和数据增强操作，以增强模型对不同类型物体的检测和识别能力。锚框可以帮助模型根据输入图像中物体的位置和大小进行预测，而数据增强操作可以扩展数据集，提供更多实例来训练模型，同时增加模型的鲁棒性和泛化能力。

骨干层对预处理后的图像进行特征提取获得不同尺寸的特征图，如公式（3-4）所示。

$$y_3 = f_{Backbone}(x') \in \mathbb{R}^{168\times168\times64}, \mathbb{R}^{84\times84\times128}, \mathbb{R}^{42\times42\times256}$$
$$z_1, z_2, z_3 = f_{Down}(y_1), f_{Down}(y_2), f_{Down}(y_3) \in \mathbb{R}^{28\times28\times256}, \mathbb{R}^{14\times14\times512}, \mathbb{R}^{7\times7\times1024} \tag{3-4}$$

$f_{Backbone}$ 表示骨干层函数，它可以将经过数据增强处理后的输入图像 x' 转换为三个不同尺寸的特征图，f_{Down} 表示下采样操作。目的是对预处理后的图像进行特征提取，得到多尺度、多层次的特征表示。

颈部层完成不同尺寸的特征拼接工作，如公式（3-5）所示。

$$p = f_{PANet}(z_1, z_2, z_3) \in \mathbb{R}^{28\times28\times512} \tag{3-5}$$

其中 f_{PANet} 表示 PANet 网络函数，它可以将不同尺寸的特征图拼接并进行相关操作，得到新的特征图 p。其作用在于整合不同尺度的特征信息，以提供更全面和丰富的特征表示，提高模型对目标的理解能力。

$$y_{pred} = f_{Predict}(p, a) \in [0,1]^{3\times S\times S\times(C+5)} \tag{3-6}$$

预测层对图像结果进行预测，其中 a 代表锚框尺寸，S 表示特征图的大小，C 表示检测中物体类别数量，如公式（3-6）所示。它将根据锚框尺寸和特征图大小生成一组锚框，然后为每个锚框预测它所属的物体类别概率。

3.2.2　损失函数

BCE Loss（Binary Cross Entropy Loss）作为二元交叉熵损失函数是通过计算模型输出的概率与真实标签之间的交叉熵来衡量预测结果与实际情况的差异。这种直观性使得 BCE Loss 易于理解和实现，同时也方便在训练过程中进行调试和优化，其定义如公式（3-7）所示。

$$Loss = -\frac{1}{N}\sum_{i=1}^{N} y_i \cdot log\left(p(y_i)\right) + (1-y_i)\,log(1-p(y_i)) \tag{3-7}$$

其中 y 是二元标签 0 或 1，$p(y)$ 是输出属于标签的概率，N 表示模型预测对象的组数。即对于标签 y 为 1 的情况，如果预测值 $p(y)$ 趋近于 1，那么 *Loss*

值应当趋近于 0。

在三维体素重建中，目标物体和背景的分布通常不平衡，即一个类别的样本数量明显多于另一个类别。BCE Loss 能够有效地处理这种情况，因为它将每个类别的预测概率与真实标签对应起来，并根据标签进行加权处理，从而确保在训练过程中对两个类别都给予适当的关注。此外，BCE Loss 具有计算效率高的优势，相对于其他复杂的损失函数，BCE Loss 的梯度计算相对简单，这使得在反向传播过程中更容易进行参数更新，这种高效性使得 BCE Loss 在大规模三维体素重建任务中具有较好的可扩展性和实用性。通过对以上优势进行分析，本文将使用二元交叉熵作为损失函数来提高模型的鲁棒性。

3.2.3　阈值调整模块数据集

在本实验中，采用了基于 YOLO 的目标检测网络作为阈值调整模块的识别网络，并利用 ShapeNet 数据集中的 13 种不同物体进行了训练和评估。为了训练模型，随机选取了 1000 张图片进行标注，共计 13000 张图像样本。标注的过程包括对每个图像中的目标进行边界框标注以及类别标签的标注。这一过程通过使用 LabelImg 数据集标注工具进行，旨在准确地确定目标的位置和类别，为模型提供高质量的训练数据。

在训练阶段，将标注的图像样本输入到 YOLO 模型中，并通过反向传播算法来优化模型参数，使得模型能够逐渐减小预测结果与真实标签之间的差距。通过迭代训练，YOLO 模型逐渐学习到了识别和定位图像中目标的能力，为后续的阈值调整任务打下了良好的基础。

在实验评估阶段，使用未标注的图像样本对训练好的模型进行测试，通过计算模型的准确率、召回率和 F1 得分等指标来评估其性能表现。这些指标可以帮助全面了解模型在目标检测任务中的表现，并根据需要进行进一步的优化和调整，以确保模型在实际应用中能够达到理想的效果。

本次实验建立了一个基于有监督学习的 YOLO 目标检测网络，并将其应用于阈值调整模块中。该模型能够准确地识别和定位图像中的目标，为后续的阈值调整提供可靠的输入数据，从而提高整体系统的性能和准确性。这一成果为相关领域的研究和应用提供了有益的参考，具有重要的理论和实践意义。

3.3　实验结果与分析

3.3.1　实验环境

本章实验以 PyTorch 框架为支撑，在 NVIDIA RTX A4000GPU 显卡上运行。实验采用 ShapeNet 3D 公共数据集，该数据集包含 13 个类别、48235 个模型数据，每个数据包括 1 个三维模型和 24 张图像[40]。在实验中，训练集、测试集和验证集之间的比例为 8:2:1，遵循 Pix2Vox++网络的数据分割策略。网络输入数据是三通道 RGB 图像，通过预处理后调整为 224×224 像素尺寸，同时以大小为 32×32×32 的三维体素模型作为输出数据。将交并比(Intersection over Union，IoU)作为评价指标，以 250 次迭代训练作为最终结果，选用二元交叉熵作为损失函数，使用 Adam 优化器更新模型参数，并采用与 Pix2Vox++网络一致的优化器参数设置。其中权重衰减因子设置为 0.0005，超参数 β1 和 β2 分别设置为 0.900 和 0.999。初始学习率为 0.0010，并在 150 个 Epoch 后衰减至 0.0001。初始阈值设定为 0.3。

3.3.2　编码器模块消融实验

为分析改进后的编码器模块在三维重建任务过程中的影响，设计了编码器模块消融实验，实验以 5 种模型的参数量、图像特征提取所需时间以及 IoU 值为参照对比，实验结果如表 3-1 所示。

表 3-1　不同特征提取网络实验对比

模型	参数量/Mb	时间/s	IoU
MobileViT	11.36	1692	0.656
MaxViT	16.16	2916	0.659
VGG16	3.58	324	0.661
ResNet50	5.58	684	0.670
3D-PCCF	**5.82**	**1152**	**0.672**

根据实验结果分析，本文所提出 3D-PCCF 网络在参数量方面明显优于 MobileVit、MaxVit 模型，接近 ResNet50 模型的水平。相较于 ResNet50 模型，本文模型的参数量略高，因此导致输入图像到特征提取完成所花费的时间略长，但在经过 250 个 Epochs 的训练后，本文模型的 IoU 值相较于 ResNet50 模型提升了 0.002，能够有效改善网络的重建效果。

3.3.3　精炼器模块消融实验

为探究精炼器加入注意力机制对重建任务的提升效果，进行不同网络精炼器模块的消融实验，对其差异进行比较。实验以参数量、三维模型精细化所需时间以及 IoU 值为参照，实验结果如表 3-2 所示。

表 3-2　精炼器模块消融实验对比

网络	参数量/Mb	时间/s	IoU
Pix2Vox	90.73	239	0.661
Pix2Vox 与改进精炼器	96.16	370	0.662
Pix2Vox++	92.73	239	0.670
Pix2Vox++与改进精炼器	98.16	368	0.672
3D-PCCF	**106.97**	**240**	**0.672**
3D-PCCF 与改进精炼器	**113.98**	**381**	**0.674**

在本文网络的改进方法中，三维重建任务的性能得到了显著提升，同时在保持参数量不明显增长的前提下，通过对 Pix2Vox 和 Pix2Vox++网络结合改进的精炼器模块进行测试，实验结果显示，尽管重建时间相对较长，但整体而言，该模块对重建任务的提升显著，尤其在 3D-PCCF 网络中添加改进精炼器模块后，重建精度得到明显提升，整体 IoU 值达到了 0.674，相比于 Pix2Vox++模型提升了 0.002。

3.3.4　阈值调整模块消融实验

在实验过程中，选择 YOLOv7m.pt 为初始训练模型，经过 100 次迭代训练生成最终结果。实验以 5 种网络的参数量、训练时间、准确率(Acc)以及 IoU 值为参照，对比结果如表 3-3 所示。

表 3-3　不同网络阈值调整模块消融实验对比

网络	参数量/Mb	训练时长/h	准确率/%	IoU
LeNet-5	0.06	4.3	63	0.673
AlexNet	0.61	5.7	69	0.672
VGG16	3.58	7.1	72	0.673
ResNet50	5.58	7.8	77	0.674
YOLOv7m	7.19	9.2	88	0.675

结果证明，不同阈值对于不同物体的重建效果存在显著影响。准确率指标较低表示网络的识别准确性较差，这会导致阈值向着错误的方向调整，从而影

响三维重建的效果。使用 YOLO 进行训练通常需要较长时间，但其识别准确率远高于其他网络。在 3D-PCCF 网络引入该模块后，整体 IoU 值达到了 0.675，三维重建效果得到明显优化。

3.3.5　实验结果分析

本章实验中与多种三维重建网络进行对比实验，对比结果如表 3-4 所示。本章所提出的网络 3D-PCCF 相比于 Pix2Vox++网络在整体评估中取得了 0.005 的 IoU 值提升，将 IoU 值提高至 0.675。

表 3-4　7 种网络在不同重建物体上的 IoU 值对比

物体名称	3D-R2N2	OccNet	Matryoshka	IM-Net	Pix2Vox	Pix2Vox++	3D-PCCF
飞机	0.513	0.532	0.647	**0.702**	0.684	0.674	0.666
长椅	0.421	0.597	0.577	0.564	**0.616**	0.608	0.606
衣柜	0.716	0.674	0.776	0.680	0.792	0.799	**0.804**
汽车	0.798	0.671	0.850	0.756	0.854	0.858	**0.862**
椅子	0.466	0.583	0.547	**0.644**	0.567	0.581	0.581
显示器	0.468	**0.651**	0.532	0.585	0.537	0.548	0.557
台灯	0.381	**0.474**	0.408	0.433	0.443	0.457	0.462
音响	0.662	0.655	0.701	0.683	0.714	0.721	**0.726**
步枪	0.544	0.656	0.616	**0.723**	0.615	0.617	0.629
沙发	0.628	0.669	0.681	0.694	0.709	0.725	**0.731**
桌子	0.513	**0.659**	0.573	0.621	0.601	0.620	0.632
电话	0.661	0.794	0.756	0.762	0.776	0.809	**0.813**
船	0.513	0.579	0.591	0.607	0.594	0.603	**0.617**
总计	0.560	0.626	0.635	0.659	0.661	0.670	**0.675**

注：加粗表示最优值

本文网络的重建效果在细节方面特别是在枪托部分表现出了更精细的特点。在汽车部分的重建中，特别是在车轮处可以看到更流畅、笔直的线条。对沙发的重建效果也更符合原始模型，整体沙发更整洁，没有多余的凸出方块。在轮船的重建效果方面，船身整体更协调，船顶也更平滑。整体效果明显优于改进前的 Pix2Vox、Pix2Vox++网络。并在衣柜、汽车等物品的三维重建任务中达到最高的 IoU 值，重建细节得到进一步提升，重建效果如图 3-4 所示，重建效

果图中物体的边缘细节部分有了显著增强，视觉效果更为接近真实模型。因此可以得出结论：本章所提出的网络能够生成更加精细的 3D 网络模型，证明其在提高 3D 模型精度方面的有效性。

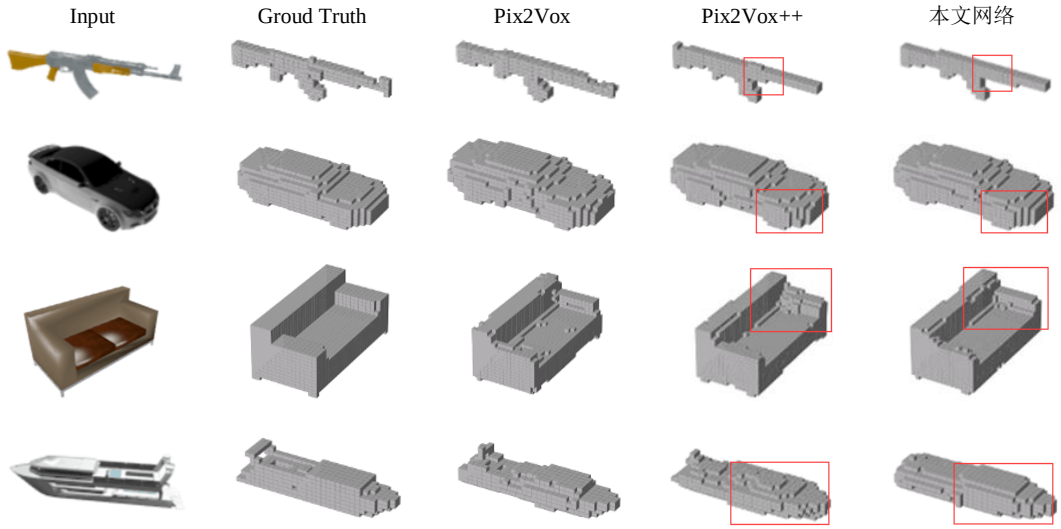


图 3-4　不同网络重建效果对比

3.4　本章小结

为了进一步提升单视图图像三维重建的精度，通过对当前的先进算法进行研究，提出了一种改进的单视图三维重建网络 3D-PCCF。该网络的编码器通过改进特征提取网络，获取更加丰富完整、深层次的二维特征。在精炼器的网络架构中引入注意力机制，进一步细化三维特征，使其生成更加精细的三维体素模型。另外在网络中添加阈值调整模块，来弥补不同种类图像之间的差异，以达到更好的重建效果。实验结果表明，本文提出的单视图三维重建网络在公共数据集 ShapeNet 上取得了不错的重建效果，有效的提升了三维重建的质量和细节还原能力，为实现更精确、高质量的三维模型重建提供了有价值的思路和方法。

第四章　单视图三维重建中的 Transformer 架构设计

4.1　引言

在单视图三维重建任务中，传统的卷积神经网络（CNN）已被广泛应用于提取图像特征和生成三维重建结果。然而，随着注意力机制的兴起，Transformer 架构作为一种基于自注意力机制的模型，逐渐在计算机视觉领域展示出强大的表示学习和序列建模能力。在这个背景下，本章探索了将 Transformer 架构应用于单视图三维重建任务的可能性。

为了使得模型能够更好地捕捉图像中不同区域的重要性，并适应不同尺度的场景，提升特征的表达能力。本文设计了一种基于 Transformer 架构的单视图三维重建模型，并评估其在公共数据集 ShapeNet 上的性能。结合了 Transformer 与 CNN 的三维重建融合网络 3D-TFCR（Three-Dimensional Transformer and Convolutional Refinement），通过使用 Transformer 二维图像特征提取网络结构，通过加入位置编码以及 HiLo[41]注意力机制，神经网络能够更好地利用输入序列中的空间信息和不同频率的特征，提高模型的表达能力和性能。同时设计了一个更为高校的精炼器网络，实现对三维模型数据的高效特征提取和重建，更好地还原输入数据的细节和结构信息。最后使用 DICE 损失函数来替换传统的二分类交叉熵损失函数，使得三维重建网络在训练过程中能够更好地优化模型参数，提高对于不同目标形状和大小变化的适应能力，从而获得更加精确和鲁棒的重建结果。

本章研究内容如下：

(1)将 Transformer 应用于三维重建网络的特征提取模块，利用其自注意力机制来捕捉三维数据中不同位置之间的依赖关系，并获得更加丰富和准确的特征表示。并通过引入当前高效注意力机制 HiLo 实现将图像数据转换为更具表征能力和可用性的特征表示。

(2)设计了一个新的三维模型精炼器，创新之处在于融合了多个关键技术，引入了密集连接和双注意力机制的网络结构，使得模型可以更好地捕捉和利用三维数据中的重要信息，并在各种三维物体重建任务中展现出了卓越的性能。

(3)使用了 DICE 损失函数作为三维重建网络的损失函数，在评估重建结果与目标之间的相似度时，更加关注它们的重叠部分，而不是只考虑整体的匹配程度。这使得网络能够更灵活地处理各种形状和大小的目标，从而提高了重建的鲁棒性和适应性。

4.2　算法设计

4.2.1　Transformer-CNN 三维重建融合网络

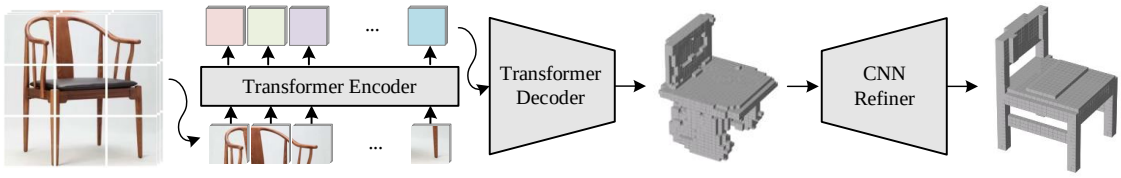


图 4-1　3D-TFCR 网络结构

相比于传统的卷积神经网络在处理序列数据时存在一些不足之处，本章引入了 Transformer 模型作为一种替代方案，以改进序列建模的性能和表达能力，网络称之为 3D-TFCR，网络结构如图 4-1 所示。网络编码器使用 Transformer 编码器代替传统的 CNN 进行图像特征提取，对三维空间中的全局关系进行建模，捕获物体之间的依赖关系。特征融合模块用于提取更具表现力和丰富性的特征表示，增强特征表达能力。解码器将 Transformer 解码器与 CNN 相结合，来获得更全面、准确的三维体素模型重建结果。这些结构设计的合理性和实验结果的验证共同证明了 3D-TFCR 网络在序列建模任务上的正确性。

复杂性主要体现在计算和存储两个方面。首先，由于引入了 Transformer 模型，网络的计算复杂性相比传统的 CNN 可能会增加。Transformer 模型的自注意力机制使得每个位置的输出都受到输入序列中所有位置的影响，因此其计算复杂度较高。另外，特征融合模块也需要一定的计算资源来提取更具表现力的特征表示。其次，网络的存储复杂性主要取决于网络的参数量和中间特征的存储需求。由于 3D-TFCR 网络结构复杂，参数量可能较大，需要考虑在训练和推理阶段的存储开销。

4.2.1.1　基于 Transformer 的二维图像特征提取网络

本章中的二维图像特征提取网络使用了 Transformer 代替了传统的卷积神经

网络，它由自注意力层和前馈神经网络层组成，主要分为图像块的嵌入和 Transformer 编码器两个部分，其网络结构如图 4-2 所示。

图像块的嵌入是指将图像划分为固定大小的图像块后，每个图像块需要被转换为一个嵌入向量，以便能够作为输入传递给 Transformer 模型。具体实现是将一张图片 $(H\times W\times C)$ 分割为 N 个像素点个数为 $(P\times P\times C)$ 的图像块，图像块长宽为 P，那么 $N=\frac{H\times W}{P\times P}$，再把每个图像块展平后连接得到一个 $N\times(P^2\bullet C)$ 的二维矩阵，即为 Transformer 编码器的输入。

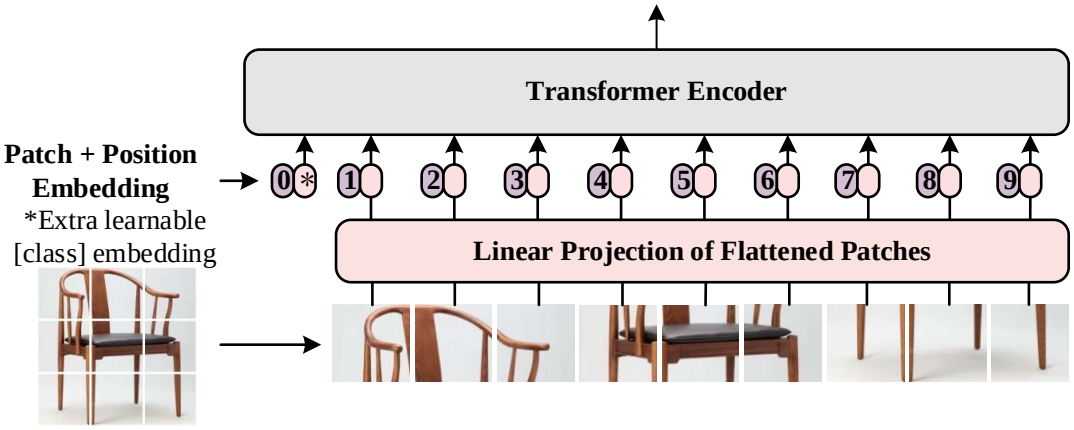


图 4-2　二维图像特征提取网络结构

Transformer 编码器是一种基于自注意力机制（self-attention）的神经网络结构，用于处理序列数据。在这里，图像块的嵌入向量被看作是一个序列，每个向量对应一个图像块。通过计算每个图像块向量与其他图像块向量之间的关系，以及它们在整个图像中的重要性。这样可以捕捉到图像块之间的相互作用和全局信息，并有效地编码图像的特征。其网络结构如图 4-3 所示。

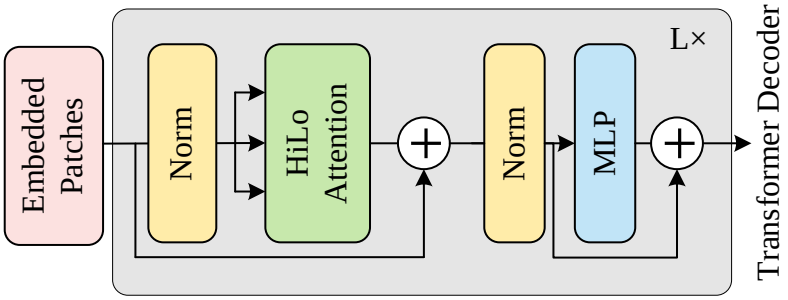


图 4-3　Transformer 编码器网络结构

为了将位置信息融入模型中，在网络中引入了位置编码(Positional Encoding)[42]。位置编码是一种固定的向量，其维度与输入序列的维度相同，通过将位置编码与输入进行逐元素相加，使得模型能够区分不同位置的特征。这

种设计允许模型直接感知到序列的顺序，从而更好地处理序列数据。

通过对输入特征与位置编码矩阵的相加操作，将位置信息融入输入特征中，以便在后续处理中能够考虑到序列中元素的相对位置。其位置编码使用正余弦编码，在矩阵中给定一个长度为 n 的输入序列，让 t 表示词在序列中的位置，$\overrightarrow{p_t} \in \mathbb{R}^d$ 表示 t 位置的对应向量，d 是向量的维度。$f:\mathbb{N} \to \mathbb{R}^d$ 是生成位置向量 $\overrightarrow{p_t}$ 的函数，定义如公式（4-1）所示。

$$\overrightarrow{p_t}^{(i)} = f(t)^{(i)} := \begin{cases} \sin(\omega_k . t), & if\ i=2k \\ \cos(\omega_k . t), & if\ i=2k+1 \end{cases} \tag{4-1}$$

其中，频率 ω_k 定义如公式（4-2）所示

$$\omega_k = \frac{1}{10000^{2k/d}} \tag{4-2}$$

通过对查询向量和键向量进行点积运算，并除以头维度的平方根得到，这样的计算方式能够强调重要的特征之间的关系，并减轻无关特征的影响。其注意力得分计算如公式（4-3）所示

$$\frac{Q \bullet K^T}{\sqrt{head\_dim}} \tag{4-3}$$

在这个公式中，Q 表示查询张量，K^T 表示键张量的转置。$head\_\dim$ 是指注意力机制中的头部维度。通过将查询张量 Q 与键张量 K 的转置相乘，并除以 $head\_\dim$ 的平方根来计算注意力得分。

同时本文网络引入了一种新的高效注意力机制 HiLo 连接多层感知机（MLP，Multilayer Perceptron）。在每个块之前应用 Layernorm（LN），在每个块之后应用残差连接。

HiLo 注意力机制是一种用于解耦高频率和低频率特征的有效注意力方法，如图 4-4 所示。在该机制中，通过两种不同的注意力方式分别处理高频（Hi-Fi）和低频（Lo-Fi）信号。对于高频部分，采用了 Local Window Self-Attention 来捕获细粒度的高频信息。这种方法比传统的多头自注意力（MSA）更有效，通过将一些注意力头（Head）专门分配给高频注意力，可以更好地捕捉到局部区域的细节信息。而对于低频部分，首先对输入特征图的每个窗口进行平均池化，得到了低频信号。然后，剩余的注意力头被分配给 Lo-Fi，用于建模每个查询位置与每个窗口的平均池化低频键（key）之间的关系。由于低频部分的 key 和 value 长度减少，Lo-Fi 的复杂度大幅降低。最后，将经过细化的高频和低频特征连接

起来，并将结果输出到后续层进行进一步处理。这样的设计能够有效地分离处理高频和低频信息，提高了模型对不同频率特征的建模能力，从而在各种任务中表现出更好的性能。

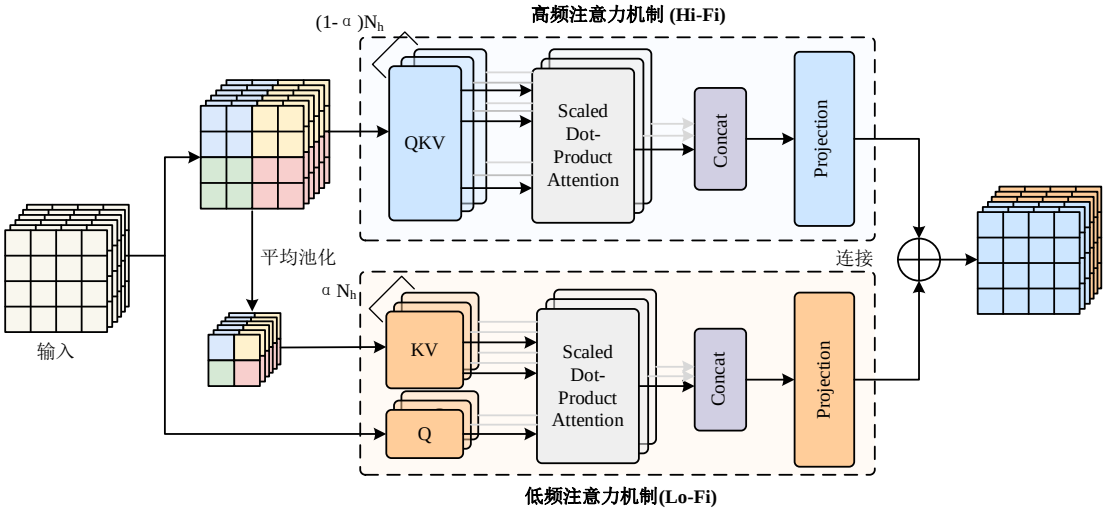


图 4-4　HiLo 注意力机制结构

MLP 包含具有 GELU 非线性的两个全连接层。其结构如公式（4-4）所示。

$$z_0 = [x_{class}; x_p^1 E; x_p^2 E; ...; x_p^N E] + E_{pos},\quad E \in \mathbb{R}^{(P^2 \cdot C)\times D},\ E_{\text{pos}} \in \mathbb{R}^{(N+1)\times D}$$

$$y = LN(z_L^0) \tag{4-4}$$

Encoder 第一层的输入 z_0 是通过第一个公式得到的，其中 $X_p^1,...,X_p^N$ 即未线性映射的图像块嵌入，其维度大小为 P^2C，右乘 $P^2C\times D$ 维的矩阵 E 表示线性映射，得到的 $X_p^1E,...,X_p^NE$ 都是 D 维向量;这 N 个 D 维向量与同样是 D 维向量的 X_{class} 拼接就得到了 $(N+1)\times D$ 维矩阵。加上 N＋1 个 D 维位置嵌入拼成的 $(N+1)\times D$ 维矩阵 E_{pos}，即得到了 Encoder 的原始输入 z_0。

对于 Encoder 的第 l 层，记其输入为 z_{l-1}，输出为 z_l，计算过程如公式（4-5）所示。

$$\begin{aligned} z'_l &= HiLo(LN(z_{l-1})) + z_{l-1} \\ z_l &= MLP(LN(z'_l)) + z'_l \end{aligned} \tag{4-5}$$

相比于上一章所提出的纯 CNN 的三维重建网络结构，使用 Transformer 进行三维重建任务提供了一种不同的方法来提取图像的二维特征。Transformer 网络中的自注意力机制允许模型在处理序列数据时能够同时关注到序列中的不同位置，从而捕获全局的上下文信息并理解序列中不同元素之间的关系。这种注意力机制使得 Transformer 在处理长距离依赖关系时更加有效。另外，使用 Transformer 网络可以避免参数共享的限制。在 CNN 中，卷积操作通常使用相同

的权重参数对输入进行处理，这种参数共享机制可以减少模型的参数量，但也可能限制了模型的灵活性。而在 Transformer 中，每个位置都有自己的权重矩阵进行特征提取，这使得模型可以更自由地学习不同位置的特征表示。此外，由于没有像卷积操作那样的局部感知野限制，Transformer 能够更灵活地建模不同类型的输入序列以及处理变长序列，而无需进行填充或截断操作。这使得模型在处理多样性的图像数据时具有更强的适应性，能够充分利用图像中的信息，从而准确、高效的实现对图像二维特征的提取。

4.2.1.2　三维模型重建细节的关键技术优化

在 3D-TFCR 中，Transformer 解码器的主要目标是进行三维特征的重建。但是直接通过 Transformer 解码器生成的三维模型较为粗糙，边缘细节处理的不够细腻，无法达到很高的精确度。为了解决这一问题，对上一章的精炼器模块进行了深入研究，并设计了一个全新的精炼器模块。该模块完全由卷积神经网络构成，旨在从初始的、可能较为粗糙的三维模型中提取并学习更多的细节信息，并生成更具有精确度的三维模型。

在精炼器模块中，采用了 CNN 来对三维模型进行精炼。与传统的精炼器不同，全新的精炼器模块采用了新型的残差卷积（Residual Convolution）[43]结构，这种结构能够更好地保留原始模型的细节信息，并且能够有效地提高精炼后的三维模型的精确度。具体来说，全新的精炼器模块包括多个卷积层，每个卷积层都会对输入的三维模型进行一次卷积操作，并对其输出进行一次残差连接操作。这些卷积层的设计能够让网络更好地学习到原始模型中的细节信息，并将其融入到精炼后的三维模型中，生成更加细腻的边缘细节。精炼器模块如图 4-5 所示。

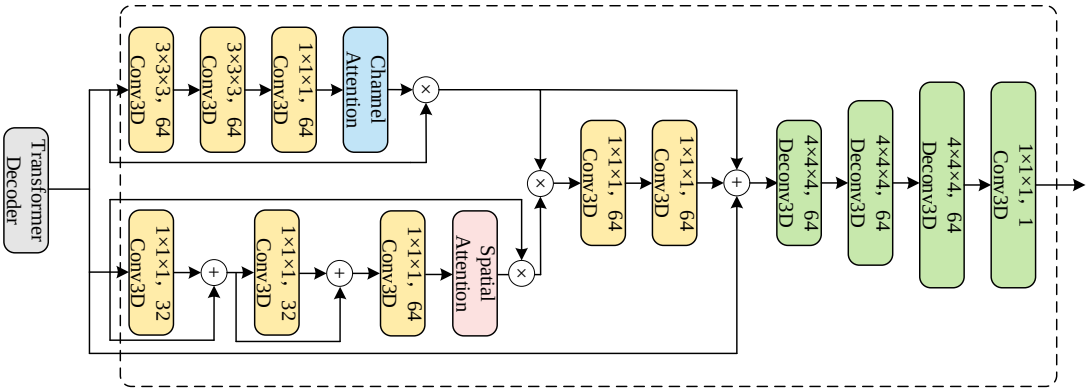


图 4-5　精炼器网络结构

在精炼器网络中，通过引入密集连接（Dense Connection）[44]的概念，将每个残差块的输入特征反向连接到该块之前所有残差块的输出特征上。这样的设计不仅使得网络更加深层，还增强了特征的传递和累积，提高了特征的重用性和信息流动性。这有助于更好地利用历史信息，提高对特征的记忆能力，从而更好地解决深度学习中的难题，例如梯度消失和梯度爆炸等问题。同时，由于每个节点都具有更强的特征表达能力，网络的可解释性和鲁棒性也得到了增强，从而更好地适应不同类型的数据和任务，如公式（4-6）所示

$$H(i)=\left[H(1),H(2),...,H(i-1),F(i)\left(H(i-1)\right)\right] \tag{4-6}$$

该公式表示了密集连接块的操作，$H(i-1)$ 表示前一个密集连接块的输出特征，$H(i)$ 是由经过 $F(i)$ 操作得到新的特征与之前所有的输出特征 $\left[H(1),H(2),...,H(i-1)\right]$ 进行拼接得到，即第 i 个密集连接块的输出特征。通过这种方式，每个密集连接都将前面所有密集连接的输出特征图与当前残差块的输入特征进行拼接，并经过卷积操作后得到输出特征图。这样可以更充分地利用之前的特征信息，并加强特征的重用性和信息的传递性，在整个网络中形成了密集的连接和信息流动。

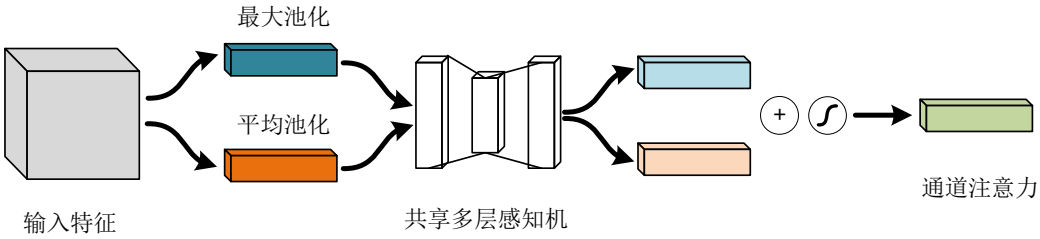


图 4-6　通道注意力机制模块

其次，在网络中混合引入了通道注意力机制以及空间注意力机制。通道注意力机制是一种用于卷积神经网络中的特征选择方法，如图 4-6 所示。对输入特征图进行全局最大池化和全局平均池化操作，分别计算每个通道上的最大特征值和平均特征值。并将全局最大特征向量和平均特征向量输入到一个共享的全连接层中，该全连接层用于学习每个通道的注意力权重，通过学习，网络可以自适应地决定哪些通道对于当前任务更加重要。为了确保注意力权重在 0 到 1 之间，应用 Sigmoid 激活函数来产生通道注意力权重。这样可以使得每个通道的权重落在 0 到 1 的范围内。最后，使用得到的注意力权重，将其与原始特征图的每个通道相乘，得到注意力加权后的通道特征图。这将强调对当前任务有帮助的通道，并抑制无关的通道，从而提高模型性能。总的来说，通道注意力机

制通过全局池化、全连接层和 Sigmoid 激活来学习每个通道的注意力权重，并将其应用于原始特征图，以实现特征选择和加权。这种方法可以提高模型的表达能力和性能，使其能够更好地适应不同的任务需求。

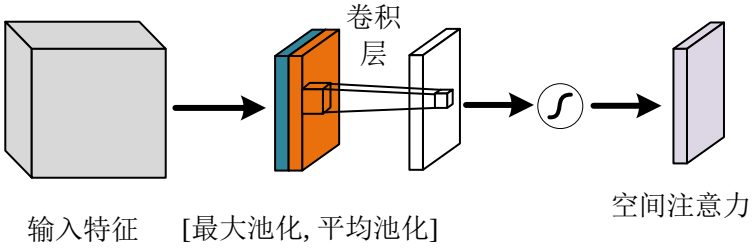


图 4-7　空间注意力机制模块

空间注意力机制是一种用于图像处理的技术，如图 4-7 所示。通过全局最大池化和全局平均池化操作，生成了不同上下文尺度的特征[45]。随后，将这些特征沿着通道维度进行连接，形成了一个具有不同尺度上下文信息的特征图。接着利用卷积层处理这个特征图，生成了空间注意力权重。这些权重经过 Sigmoid 激活函数的处理，被限制在了 0 到 1 之间。最后，利用这些注意力权重对原始特征图进行加权处理，突出了重要的图像区域，减少了不重要区域的影响。这种空间注意力机制能够帮助模型更有效地理解图像内容，集中精力关注重要区域，从而提高了图像处理和理解的性能。

网络对通道注意力模块和空间注意力模块进行结合，使它们可以共同作用，通道注意力模块通过整合所有通道映射之间的相关特征，能够选择性地强调存在相互依赖的通道映射，从而更精确地分割结果，为三维重建提供更准确的特征表示[46]。空间注意力模块则关注于计算每个像素在空间上的重要性，能够将更广泛的上下文信息编码为局部特征，从而增强三维重建模型对局部特征的捕捉能力，更好地重建模型的细节部分。这将构建出一种强大的注意力机制，这种机制不仅可以提高三维重建模型的性能，还可以更好地聚焦于图像的重要区域，同时忽略不相关的信息，从而提升模型的性能。通过这种方式，结合通道注意力模块和空间注意力模块可以为三维重建提供更丰富、更准确的细节信息，从而获得更精确的三维重建结果。

4.2.2　损失函数

Dice 损失函数是一种用于图像分割任务的评价指标和优化目标。它在医学影像分析、计算机视觉和深度学习等领域得到广泛应用。Dice 损失函数的主要

目标是衡量预测结果与真实标签之间的相似度，以便优化模型的训练过程。Dice 损失函数的定义如公式（4-7）所示。

$$Loss = 1 - \frac{2\sum_{i=1}^{N} y_i y_i}{\sum_{i=1}^{N} y_i + \sum_{i=1}^{N} y_i} \tag{4-7}$$

其中，y_i 与 y_i 分别表示像素 i 的标签值与预测值，N 为像素点总个数，等于单张图像的像素个数乘以 batchsize。

在三维重建中，数据分布可能存在不同类别或不同区域之间的不均衡情况。Dice 损失函数考虑了预测结果与真实标签之间的相似度，能够给予少数类别或区域足够的关注，提高模型对不平衡数据的适应能力[47]。并且由于数据噪音、分辨率等因素，三维重建中可能会出现边界模糊的情况。Dice 损失函数使用了平滑项，可以减少由于边界模糊而导致的不稳定性，提高模型对边界信息的处理能力。同时在三维重建中，有时需要同时考虑像素级别的预测结果和整体结构的一致性，Dice 损失函数能够有效地衡量重建结果与真实标签的相似度，帮助模型更好地学习和优化，从而进一步提高三维重建模型的性能和稳定性。

4.3　实验结果与分析

4.3.1　编码器模块消融实验

为了深入分析改进后的编码器模块在三维重建任务中的影响，进行了编码器模块的消融实验。该实验设计包括 5 种不同的模型配置，分别考察参数量、训练时间和 IoU 值等指标的变化情况，以便进行全面的对比和评估。实验结果如表 4-1 所示。

表 4-1　不同特征提取网络实验对比

模型	参数量/Mb	时间/h	IoU
Pix2Vox++	5.58	8.4	0.670
3D-PCCF	5.82	10.5	0.672
3D-RETR	163.54	35.5	0.674
Transformer-Encoder	133.21	21.3	0.680
3D-TFCR	**135.42**	**27.5**	**0.685**

根据实验结果的分析，本文提出的 3D-TFCR 网络中，使用 Transformer 作

为图像特征提取器，使得模型在重建任务上取得了卓越的成果。具体而言，该网络在准确度方面达到了 0.685 的水平，这是一个显著的提升。这一结果表明，通过利用 Transformer 的强大的自注意力机制和全局感知能力，3D-TFCR 网络能够更好地捕捉图像的复杂特征，并且在模型重建任务中取得更高的准确度。此外，尽管使用 Transformer 作为特征提取器导致网络参数的上升和训练时间的增加，但这个额外的计算开销是值得的，给模型性能方面带来了显著的提升。

总的来说，本文的实验结果表明，在图像处理任务中，使用 Transformer 作为特征提取器可以有效地提高模型的准确度，为图像重建等应用领域带来了巨大的潜力。

4.3.2　精炼器模块消融实验

为了深入探究精炼器加入注意力机制对重建任务的提升效果，进行了一系列不同网络精炼器模块的消融实验，并对其差异进行了细致比较。实验结果如表 4-2 所示，主要以参数量、模型整体训练时间以及 IoU 值作为参照指标。在实验中，考察了精炼器模块加入注意力机制前后的网络性能变化，通过比较不同配置下的参数量、训练时间和 IoU 值，以评估注意力机制对模型性能的影响。

表 4-2　精炼器模块消融实验对比

网络	参数量/Mb	时间/h	IoU
3D-TFCR	135.42	27.5	0.685
3D-TFCR 与通道注意力机制	136.11	28.6	0.689
3D-TFCR 与空间注意力机制	135.76	28.3	0.687
3D-TFCR 与改进精炼器	**138.51**	**28.7**	**0.693**

通过消融实验的详细结果分析，对不同精炼器模块配置所导致的参数量和训练时间的变化进行了深入研究，并针对模型在重建任务中的 IoU 值表现进行了比较。这些数据揭示了精炼器加入注意力机制对模型性能的显著影响，提供了更加全面的实验结果和结论。

实验证实，将注意力机制集成到精炼器中明显提升了三维重建网络的性能。特别是在引入改进的精炼器后，整体 IoU 值达到了 0.693，这表明注意力机制的引入对于提高三维重建网络的准确性和稳定性起到了关键作用。这一发现强调了注意力机制在优化模型学习和提升性能方面的重要性，为进一步研究和改进三维重建网络提供了有力支持。

4.3.3　损失函数消融实验

损失函数的选择对于模型训练和性能表现具有重要影响。为了深入研究不同损失函数对模型性能的影响，进行了损失函数的消融实验，并使用了 4 种不同的损失函数进行对比。

在这个消融实验中，选择了 4 个常用的损失函数进行对比分析，以评估它们在三维重建任务中的效果差异。通过比较这些损失函数在训练过程中的准确率、召回率、稳定性、收敛速度以及重建效果，可以深入了解它们在模型优化中的作用和影响。

表 4-3　损失函数消融实验对比

损失函数	准确率/%	召回率/%	稳定性	收敛速度/Epochs	IoU
BCE	0.85	0.82	高	117	0.693
Focal	0.87	0.84	低	125	0.686
CD	0.82	0.79	低	132	0.694
DICE	**0.88**	**0.85**	**高**	**129**	**0.699**

实验结果显示，DICE 损失函数在本次三维重建任务中表现出色，DICE 损失函数在准确率、召回率、IoU 值上都表现的更好，但是收敛速度可能略慢，其三维重建的 IoU 值达到了 0.699，突出了在模型训练中对于正负样本不均衡情况的处理能力。相比之下，BCE、Focal Loss 和 CD 损失函数在本次实验中未能达到与 DICE 相匹敌的性能水平。

综上所述，选取合适的损失函数对于三维重建任务的成功至关重要，而 DICE 损失函数在本次实验中表现突出，为未来的三维重建研究和实践提供了参考。

4.3.4　实验结果分析

经实验结果分析，本章提出 3D-TFCR 网络在多个物体的重建任务中均能达到最优值，整体 IoU 值达到了 0.699。与其它基于 Transformer 的三维重建网络在单视图重建方面的 IoU 值相比，本章所提出的网络也展现出了明显的优势。实验对比结果如表 4-4 所示，证明了 3D-TFCR 网络在单视图三维重建任务中的有效性。

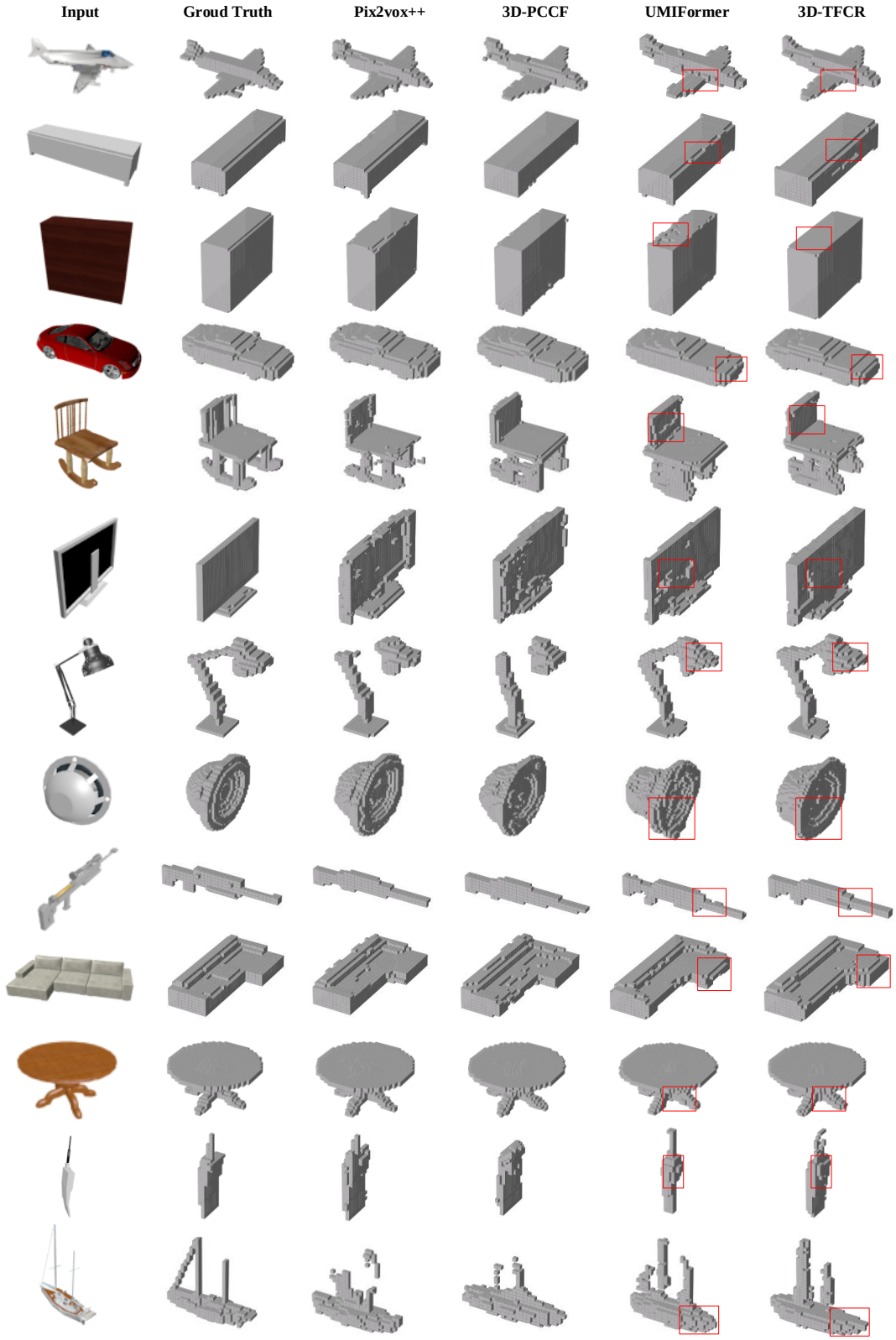


图 4-8　不同网络重建效果对比

表 4-4　7 种网络在不同重建物体上的 IoU 值对比

物体名称	3D-R2N2	Pix2Vox	Pix2Vox++	3D-PCCF	3D-RETR	UMIFormer	3D-TFCR
飞机	0.513	0.684	0.674	0.666	**0.704**	0.683	0.702
长椅	0.421	0.616	0.608	0.606	**0.650**	0.620	0.648
衣柜	0.716	0.792	0.799	0.804	0.802	0.802	**0.811**
汽车	0.798	0.854	0.858	0.862	0.861	0.850	**0.864**
椅子	0.466	0.567	0.581	0.581	0.592	0.590	**0.609**
显示器	0.468	0.537	0.548	0.557	0.574	**0.603**	0.597
台灯	0.381	0.443	0.457	0.462	0.483	**0.511**	0.509
音响	0.662	0.714	0.721	0.726	0.668	**0.745**	0.743
步枪	0.544	0.615	0.617	0.629	**0.735**	0.669	0.677
沙发	0.628	0.709	0.725	0.731	0.724	0.733	**0.744**
桌子	0.513	0.601	0.620	0.632	0.633	0.645	**0.658**
电话	0.661	0.776	0.809	0.813	0.781	0.791	**0.819**
船	0.513	0.594	0.603	0.617	0.636	0.626	**0.647**
总计	0.560	0.661	0.670	0.675	0.680	0.685	**0.699**

注：加粗表示最优值

根据实验结果分析，本文所提出网络模型在以 IoU 为评价指标的前提下，与其他几种网络对比取得了更好的重建效果。对此，选取了重建效果较好的 5 种网络进行三维重建可视化对比，对比效果如图 4-8 所示。可以明显看出，本章网络能够更好地保留图像的纹理、轮廓和细微变化，使重建结果更接近真实体素图像。在网络重建的过程中，特别是在飞机的机翼、汽车的机壳、椅子的靠背、步枪的瞄准镜以及桌子的桌腿上，表现出了更贴近实际、更美观的效果。这些部分的重建更加精细，细节处理更为细致，使得物体看起来更加逼真。同时，在长椅、衣柜、电视、台灯、音响、沙发、手机以及轮船等物体的整体重建中，也能够观察到与原始模型更相似的效果，证明了本章网络在三维重建任务的优越性和有效性。

4.4　本章小结

为了探究 Transformer 网络对单视图图像三维重建的影响，本章的研究工作

集中在利用 Transformer 和 HiLo 注意力机制来提高神经网络在三维重建任务中的性能。首先，研究通过将 Transformer 应用于三维重建网络的特征提取模块，利用其自注意力机制来捕捉三维数据中不同位置之间的依赖关系，从而获得更加丰富和准确的特征表示。同时，引入了 HiLo 注意力机制，有效将图像数据转换为更具表征能力和可用性的特征表示，进一步提升了模型的性能。

其次，研究设计了一个新的三维模型精炼器，该模型融合了密集连接和双注意力机制的网络结构，可以更好地捕捉和利用三维数据中的重要信息，并在各种三维物体重建任务中展现出了卓越的性能。

最后，研究采用了 DICE 损失函数作为三维重建网络的损失函数，更加关注重建结果与目标之间的重叠部分，使得网络能够更灵活地处理各种形状和大小的目标，从而提高了重建的鲁棒性和适应性。

综上所述，本章的研究工作通过引入先进的注意力机制和损失函数，为三维重建任务带来了性能的提升，也为相关领域的研究和应用提供了有益的启示与借鉴。

第五章　单视图三维重建可视化系统的设计与实现

5.1　引言

随着计算机视觉和深度学习的快速发展，单视图三维重建成为了一个备受关注的研究领域。传统的三维重建方法通常需要多个视角的图像或使用复杂的传感器设备，限制了其在实际应用中的可行性和便利性[48]。然而，单视图三维重建技术通过从单张图像中恢复三维结构，克服了这些限制，并在许多领域展现出巨大的潜力。

单视图三维重建可视化系统的设计与实现是使得普通用户能够直观地观察和理解三维重建结果的关键。该系统能够将单张图像转化为具有立体感的三维模型，并提供交互式的操作界面，使用户能够自由地旋转、放大和缩小三维模型，以及获取模型的其他相关信息。设计并实现一个高效可靠的单视图三维重建可视化系统对于研究人员和相关行业具有重要意义。这样的系统可以帮助用户直观地理解三维重建的结果，加快数据处理和分析的速度，同时也有助于推动该技术在实际应用中的推广和应用。这种系统的设计不仅对学术研究具有重要意义，还在工业、医疗、娱乐等领域具有广泛的应用前景。

通过深入探讨系统的架构设计、关键算法实现以及交互式界面开发等方面，将提供给读者一个清晰的系统总体认识，并帮助他们理解如何利用现代计算机视觉和深度学习技术来构建这样一个系统。期望能为单视图三维重建可视化系统的设计与实现提供有益的参考，并为相关领域的研究和应用做出贡献。同时，也希望能够促进单视图三维重建技术在实际应用中的推广和应用，为用户提供更加直观、便捷的三维视觉体验。

5.2　需求分析

5.2.1　功能需求分析

进行系统功能性需求分析是为了确定单视图三维重建系统应该具备的完整

功能，这些功能应该包括以下功能：

(1)清晰明了的系统界面操作

系统设计了直观清晰的用户界面，功能模块完备且易于理解。与传统算法不同，用户无需具备专业知识储备，系统操作简单流畅，用户能够轻松上手，清晰明确的操作流程让用户能够快速完成任务。

(2)优秀的三维模型视觉效果

系统生成的三维模型视觉效果出色，这确保了其后续应用的实际意义。系统不过度依赖输入图像的标准，能够适用于不同分辨率、视角或场景的图片，具有很强的通用性和适应性。

(3)交互性和反馈

系统应具备良好的交互性，能够及时给出反馈，帮助用户理解当前操作的效果，并在必要时提供指导。这样能够增强用户对系统操作的掌控感和满意度。

(4)用户友好的提示和帮助功能

系统应该提供用户友好的提示和帮助功能，当用户遇到问题或不清楚如何操作时，能够直接获取相关帮助信息，减少用户的困惑和学习成本。

(5)实时渲染展示

在前端页面使用 Three.js[49]技术，通过引入图形库来实现实时的三维模型渲染展示。这样用户可以在浏览器中即时查看推理生成的三维模型，并进行交互操作，如旋转、放缩等。

5.2.2　性能需求分析

进行系统性能需求分析是为了确定单视图三维重建系统应该具备的完整功能，这些功能应该包括以下功能：

(1)重建速度

基于深度学习的图像三维重建算法使系统能够以较快的速度生成三维模型，与传统算法相比，不会过度消耗时间成本。系统的高效率满足了对实时性的应用场景要求，为用户提供了更好的体验和便利。

(2)系统稳定性和可靠性

系统应具备良好的稳定性，能在长时间运行或处理复杂任务时不易崩溃或出现错误。且应对异常情况具备容错机制，如处理无效输入或不完整数据时能给出合理的反馈和处理方式。

(3)资源利用

系统在计算资源方面应合理利用，尽量减少资源占用，提高系统的效率和性能。并且应该支持多种硬件配置，以满足不同用户的资源限制和需求。

(4)可扩展性和兼容性

系统应具备一定的可扩展性，能够适应未来可能的功能扩展和新技术的集成。同时应支持常见的输入数据格式，并与其他相关软件或系统进行良好的互操作。

5.3　系统设计

本章系统的设计主要分为用户操作模块、页面展示模块以及算法实现模块。这三个子模块共同构成了系统的核心功能，保障了用户操作的顺利进行和三维重建结果的有效展示。系统总体框架图如 5-1 所示。

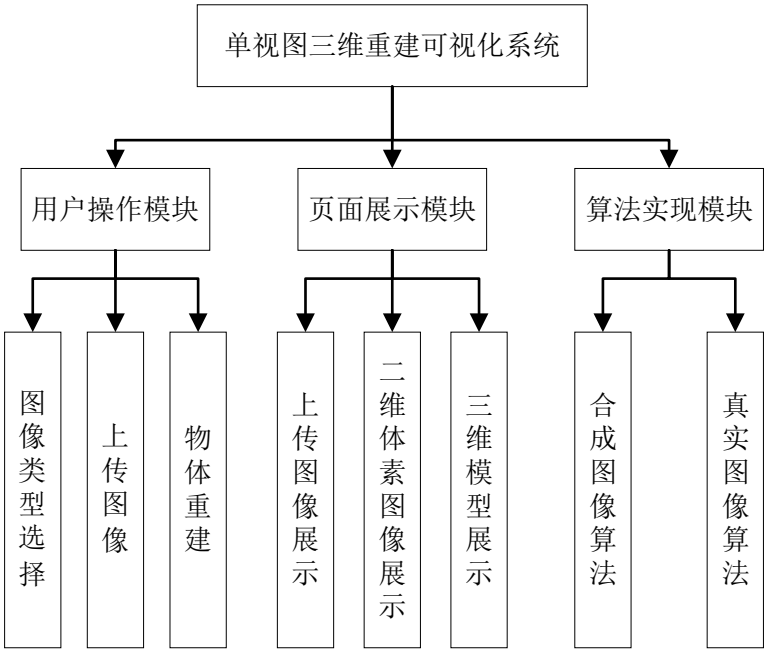


图 5-1　系统总体框架

5.3.1　用户操作模块

用户可以选择重建的物体图像是基于虚拟合成的图像还是真实的图像，系统会根据用户的选择来动态改变重建模型。这种灵活性使用户能够根据实际需求选择不同的图像来源，以满足其特定的应用场景。系统会相应地处理所选图

像，并生成相应的三维模型，以提供个性化的重建结果。这种可选性提高了系统的适用性和用户体验，使用户能够根据需要进行定制化的图像重建。

用户通过点击上传图像按钮，在本地计算机上选择要重建的图像。系统支持多种常见的图像格式，包括 JPG、PNG 等。用户可以方便地从本地文件系统中选择符合需求的图像文件进行上传，以供后续的图像重建处理。这种灵活性和兼容性使用户能够使用各种常见的图像格式，满足不同场景下的需求，并提升了系统的易用性和适应性。

用户点击重建按钮后，系统将执行单图像的三维重建算法。该算法会对所选图像进行处理，提取其中的特征信息，并通过计算和分析这些特征来还原图像所代表的物体的三维结构。重建过程中，系统会生成对应的二维体素图像，即将三维空间划分为离散的体素，并将每个体素与图像中的像素进行关联。同时，系统还会生成具有几何形状和纹理信息的三维立体模型，以呈现物体的真实外观。通过这种专业的处理和生成过程，用户可以获得准确、细致的三维重建结果，用于进一步的分析、可视化或其他应用领域。

5.3.2　页面展示模块

“上传图片展示”模块用于呈现用户上传的图片，提供了一个界面组件或功能模块，使用户能够方便地查看其所上传的图片内容。

“图像重建结果展示”模块用于展示经过三维重建算法处理后得到的体素数据，并将其转化为二维体素图像进行展示。

“三维模型展示与交互”模块是一个用于展示重建后的三维模型，并提供用户交互功能的模块。在该模块中，使用 Three.js 等技术进行开发，通过将三维模型转化为网格形式进行展示，用户可以自由旋转、缩放和平移三维模型，以便更好地查看和操作。

5.3.3　算法实现模块

首先，使用了一种基于分组卷积编码的算法。经过了对 Pix3D 真实数据集的充分训练。利用卷积神经网络从单一视角的输入图像中重建出高质量的三维模型。这一方法的关键在于其能够将输入图像中的信息有效地编码并转换为三维空间中的结构。通过该算法能够实现三维模型的重建，而且还能够生成与输

入图像相对应的三维模型，为后续任务提供了重要的基础。

其次，算法采用了 Transformer 架构并与 CNN 结合。经过对 ShapeNet 虚拟数据集的训练。在这个算法中，将 Transformer 应用于三维重建任务，并结合了 CNN 的特征提取能力。通过这种结合，能够有效地捕获输入图像中的特征信息，还能够实现高效的三维模型重建。与分组卷积编码的算法相似，这种方法也能够生成与输入图像相对应的三维模型，为单视图三维重建任务提供了另一种有效的解决方案。

同时，为了进一步提高系统重建真实物体的重建精度和稳定性，我们引入了阈值调整模块。该模块利用 U-Net 网络对待重建的物体图像进行分割，并通过检测算法识别图像中的物体。根据识别结果，动态调整系统中的阈值参数，以更好地适应不同场景和物体的重建需求。通过这种方式，系统能够在保证重建准确性的同时，更好地应对复杂场景和不同类型物体的重建挑战。效果图如图 5-2 所示。

图 5-2　阈值调整模块效果图

5.4　系统实现

系统的前端开发采用了流行的 Vue.js 框架，结合 HTML、CSS 和 JavaScript 等 Web 前端技术和工具进行开发，以实现用户界面的构建和交互功能。Vue.js 作为一种轻量级的 JavaScript 框架，利用其响应式数据绑定和组件化的特点，实现了高效的页面渲染和交互逻辑的管理。同时，HTML 用于定义页面结构，CSS 用于布局和样式设计，JavaScript 用于处理页面逻辑和用户交互，这些技术的结合使得系统能够提供流畅、动态且具有吸引力的用户界面，从而实现更好的用户体验。

系统的算法实现可以采用深度学习框架 PyTorch 进行开发，并利用 CPU 或 GPU 等硬件加速进行计算。PyTorch 是一种基于 Python 的深度学习框架，具有易用性和灵活性，能够帮助开发人员快速构建和训练各种深度学习模型。同时，PyTorch 支持使用 CPU 或 GPU 等硬件来进行加速计算，从而提升模型的计算速度和效率。通过采用 PyTorch 进行算法实现，系统能够更好地应对复杂的数据处理和分析任务，进而提升系统的性能和可靠性。

系统的模型展示功能借助 Three.js 等技术进行开发，以提供丰富的三维模型展示和交互功能。Three.js 是一种基于 WebGL[50]的 JavaScript 库，专门用于在网页上呈现和交互 3D 图形。通过利用 Three.js，开发人员能够创建逼真的三维场景，并在其中展示复杂的模型和效果。同时，Three.js 还提供了丰富的交互功能，如平移、旋转、缩放等，使用户能够与模型进行直观的操作和探索。通过利用 Three.js 等技术，系统能够为用户呈现令人印象深刻的三维模型展示，提供更加沉浸式和交互式的用户体验。

5.5　系统界面

系统的首页实现效果如图 5-3 所示。在首页上，用户可以通过操作选择“合成图像”或者“真实图像”的选项。此外，用户还可以通过点击“选择文件”按钮来上传他们想要进行三维重建的图像。最后，用户可以通过点击“重建”按钮来触发系统对图像的三维重建处理。这样的交互设计能够使用户方便地选择操作，并上传自己的图像进行处理，从而提升用户体验。

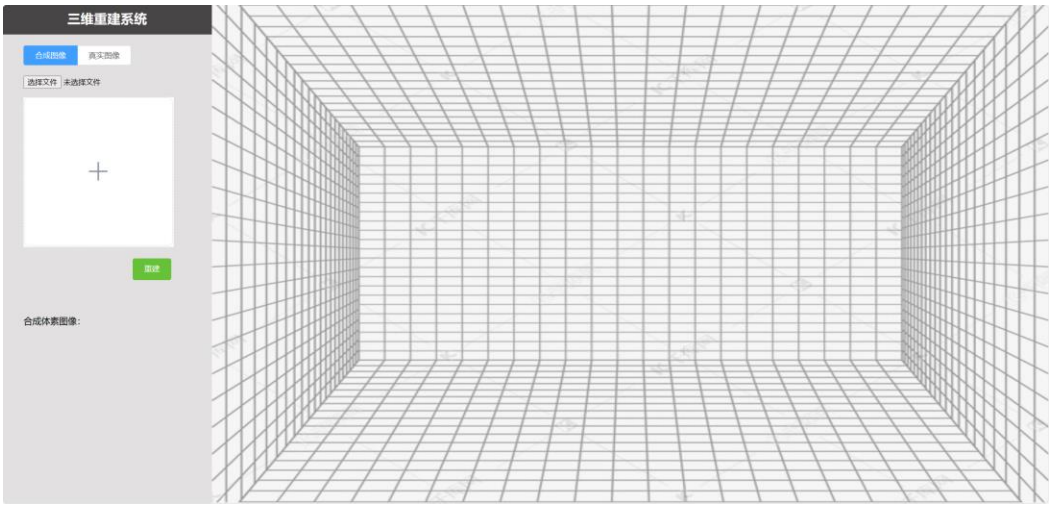


图 5-3　系统首页

用户上传并触发三维重建后，系统将启动相应的算法和模型进行图像处理和分析，通过在后端对图像进行剪裁、缩放等操作，使其达到图像的重建标准。在处理完成后，系统会呈现出类似图 5-4 所示的三维重建效果，让用户可以通过界面进行交互和观察。这种视觉化的展示方式能够帮助用户更直观地理解图像的三维重建效果，从而增强用户对系统功能的认知和满意度。

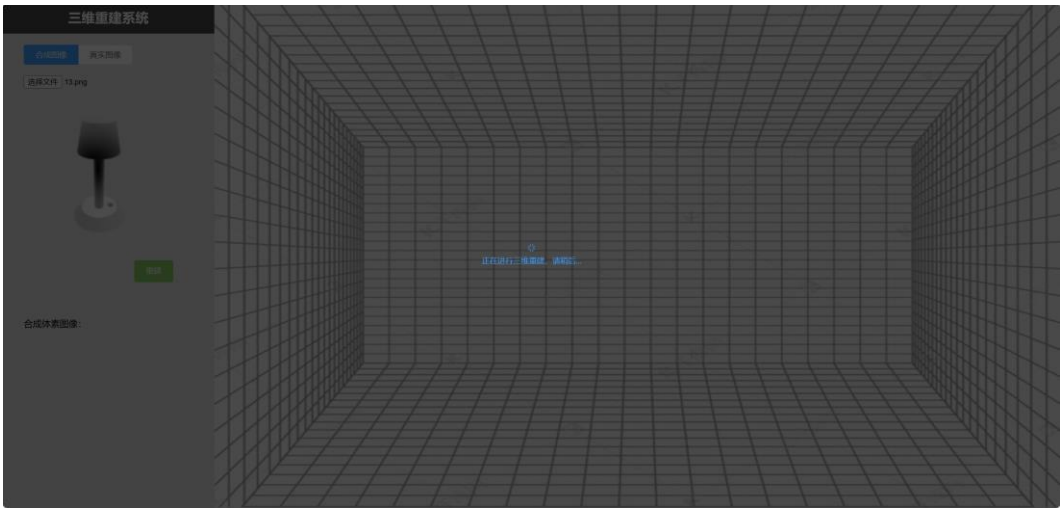
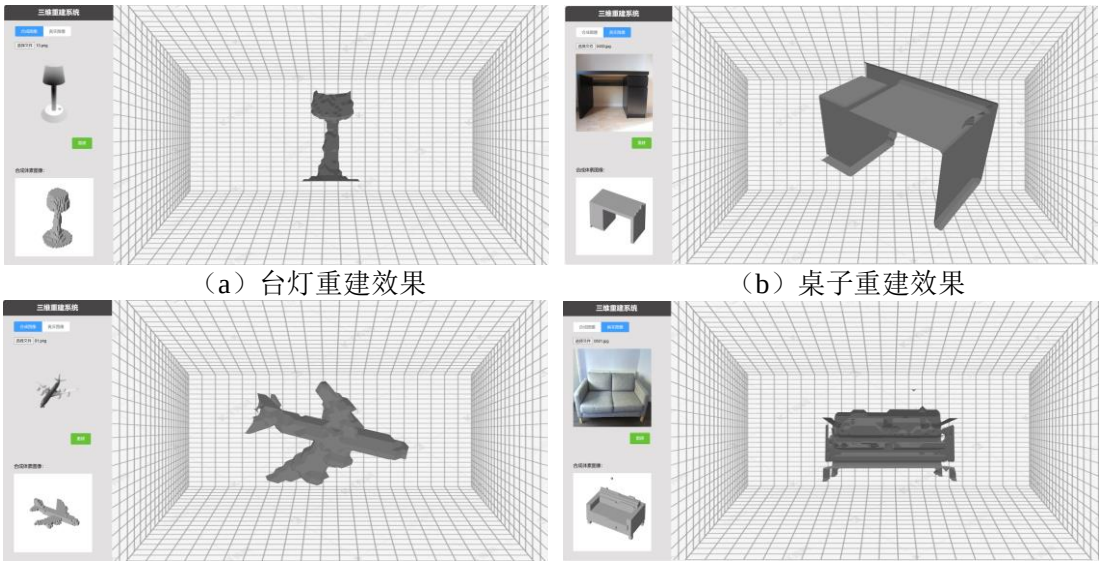


图 5-4　重建过程

在完成三维重建后，系统将呈现出类似于图 5-5 所示的效果图。其中，合成体素图像部分展示的是经过处理生成的三维重建模型的二维体素图像，通过这个图像，可以观察到模型的空间结构和细节信息。而模型显示部分则展示了完整的三维重建模型，并通过界面进行交互，来更好地查看和理解重建后的模型。

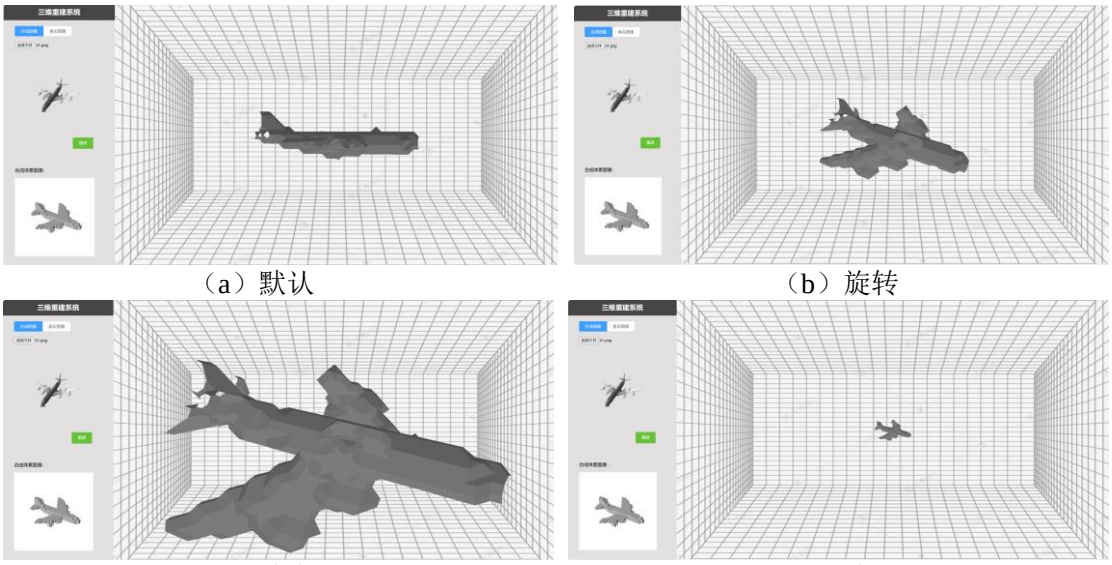


（a）台灯重建效果　（b）桌子重建效果

（c）飞机重建效果　（d）沙发重建效果

图 5-5　三维重建效果

用户可以通过系统提供的交互功能，对三维模型进行灵活的操作和探索。其中，旋转功能允许用户以不同的视角观察模型，并体验模型在不同角度下的外观和细节。用户可以通过鼠标或触摸屏手势来旋转模型，从而获得全方位的视图。缩放功能则可以让用户调整模型的大小，使其适应当前展示环境或满足特定需求。用户可以通过拉伸或收缩模型来改变其尺寸，以便更好地观察细节或在特定场景中使用。如图 5-6 所示。



（a）默认　（b）旋转

（c）放大　（d）缩小

图 5-6　模型交互效果

5.6　本章小结

本章介绍了三维重建系统的设计与实现。首先，描述了系统首页的设计，包括用户选择“合成图像”或“真实图像”，通过“选择文件”按钮上传图像，并通过“重建”按钮进行三维重建的操作流程。接着，阐述了用户上传并触发三维重建后，系统进行图像重建的工作过程。最后，介绍了重建后的效果图展示，包括合成体素图像和完整的三维重建模型，以及用户可以进行的交互操作。

第六章　总结与展望

本研究致力于探索单视图三维重建领域的前沿技术，提出了两种基于深度学习的算法，并结合算法设计了一个全新的单视图三维重建系统。这些算法和系统的核心目标在于融合最先进的深度学习技术，以提升传统三维重建方法在准确性和可视化效果上的表现。主要结论为：

(1)本文提出了 3D-PCCF 网络，它采用基于分组卷积编码的网络架构，在编码器中改进特征提取网络，获取更加丰富、深层次的二维特征。同时，在精炼器中引入注意力机制，进一步细化三维特征，生成更加精细的三维体素模型。另外，在网络中添加阈值调整模块，来弥补不同种类图像之间的差异，以达到更好的重建效果。该算法的优点是提取的特征更加全面，可以更好地表达物体的形态和特征，从而提高重建的准确性。

(2)本文提出了 3D-TFCR 网络，它结合了 Transformer 和 CNN，使用 Transformer 进行物体重建，并结合 CNN 进行物体细化。在 Transformer 编码器中引入 HiLo 注意力机制进行更全面、更准确的注意力计算，提高模型在处理长序列数据和各种任务中的表现，进一步提高模型的泛化能力和稳定性。设计了一个纯 CNN 的精炼器模块，通过引入双通道注意力机制，使得三维重建网络可以用于自适应地学习不同尺度下的重要特征，并更好地处理尺度变化带来的挑战。它还有助于更好地理解三维物体中的全局结构和局部细节，并从中准确捕捉关键的空间信息。该算法的优点是使用 Transformer 进行物体重建，可以有效地处理长序列数据和复杂任务，同时通过引入双通道注意力机制，可以更好地适应不同尺度下的重要特征，提高重建准确性。

(3)本文设计了一个单视图三维重建系统，可以生成物体重建后的二维体素图像和三维重建模型。二维体素图像帮助用户在平面上直接观察物体的三维结构，方便理解物体的形态和特征。同时，用户可以对重建模型进行旋转、缩放等操作，以更好地观察模型的各个角度和细节。该系统的优点是直观、易用，可以帮助用户更好地理解物体的三维结构和特征。有望提高传统方法的准确性和可视化效果，对于三维重建技术的发展具有积极的推动作用。

基于体素的三维重建技术在计算机视觉和图形学领域具有广泛的应用，可

用于建模、渲染、虚拟现实等方面。然而，目前存在一些挑战和局限性，需要进一步研究和改进。以下是对基于体素的三维重建技术未来发展方向的展望：

首先，随着深度学习技术的不断发展和普及，未来基于体素的三维重建技术可以更多地与深度学习相结合，利用深度神经网络提高重建的准确性和效率。例如，可以探索更加有效的网络结构和训练方法，以提高对三维结构的识别和重建能力。

其次，当前基于体素的三维重建技术在处理大规模数据时存在计算和存储成本较高的问题，未来可以研究如何优化算法和数据结构，以降低计算复杂度和提高系统的实时性。可能的方向包括基于 GPU 加速的算法设计、稀疏表示方法等。同时，基于体素的三维重建技术在处理复杂场景和动态物体时往往面临挑战，容易出现模糊或失真现象。未来的研究可以致力于改进算法，提高对复杂场景和动态变化的适应能力，实现更加精确和稳定的三维重建结果。其中，基于体素的三维重建技术还可以与传感器技术、机器学习等领域相结合，开发更加智能和自适应的三维重建系统。通过引入更多的传感器信息、环境信息和上下文信息，可以提高重建的准确性和鲁棒性，适用于更广泛的场景和应用领域。

综上所述，基于体素的三维重建技术在未来仍然有许多发展的空间和潜力，需要进一步探索更加高效、准确和智能的重建方法，以应对复杂场景和不断增长的应用需求。通过不断创新和改进，基于体素的三维重建技术将在计算机视觉和图形学领域发挥更加重要的作用，推动相关技术的发展和应用。

参 考 文 献

[1] Feng D, Haase-Schütz C, Rosenbaum L, et al. Deep multi-modal object detection and semantic segmentation for autonomous driving: Datasets, methods, and challenges[J]. IEEE Transactions on Intelligent Transportation Systems, 2020, 22(3): 1341-1360.

[2] Huang C, Vaska P, Gao Y, et al. Proceedings of the 17th Virtual International Meeting on Fully 3D Image Reconstruction in Radiology and Nuclear Medicine[J]. arXiv preprint arXiv:2310.16846, 2023.

[3] Liu F, Tran L, Liu X. 3d face modeling from diverse raw scan data[C]//Proceedings of the IEEE/CVF international conference on computer vision. 2019: 9408-9418.

[4] Kundu J N, Rahul M V, Ganeshan A, et al. Object pose estimation from monocular image using multi-view keypoint correspondence[C]//Computer Vision–ECCV 2018 Workshops: Munich, Germany, September 8-14, 2018, Proceedings, Part III 15. Springer International Publishing, 2019: 298-313.

[5] 张彦雯, 胡凯, 王鹏盛. 三维重建算法研究综述[J].南京信息工程大学学报(自然科学版), 2020,12(5):591-602.

[6] 杨硕,谢晓尧,刘嵩.多视图几何轻量级三维重建算法[J].重庆邮电大学学报(自然科学版),2022,34(06):1005-1012.

[7] 陈加, 张玉麒, 宋鹏, 等. 深度学习在基于单幅图像的物体三维重建中的应用[J]. 自动化学报, 2019, 45(4): 657-668.

[8] 李阳, 陈秀万, 王媛, 等. 基于深度学习的单目图像深度估计的研究进展[J]. Laser & Optoelectronics Progress, 2019, 56(19): 190001.

[9] Choy C B, Xu D, Gwak J Y, et al. 3d-r2n2: A unified approach for single and multi-view 3d object reconstruction[C]//Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part VIII 14. Springer International Publishing, 2016: 628-644.

[10] Wu J, Zhang C, Xue T, et al. Learning a probabilistic latent space of object shapes via 3d generative-adversarial modeling[J]. Advances in neural information processing systems, 2016:82-90.

[11] Wu J, Wang Y, Xue T, et al. Marrnet: 3d shape reconstruction via 2.5 d sketches[J]. Advances in neural information processing systems, 2017: 540-550.

[12] Wang P S, Liu Y, Guo Y X, et al. O-cnn: Octree-based convolutional neural networks for 3d

shape analysis[J]. ACM Transactions On Graphics (TOG), 2017, 36(4): 1-11.

[13] Gadelha M, Maji S, Wang R. 3d shape induction from 2d views of multiple objects[C]//2017 International Conference on 3D Vision (3DV). IEEE, 2017: 402-411.

[14] Tatarchenko M, Dosovitskiy A, Brox T. Octree generating networks: Efficient convolutional architectures for high-resolution 3d outputs[C]//Proceedings of the IEEE international conference on computer vision. 2017: 2088-2096.

[15] Yao Y, Luo Z, Li S, et al. Mvsnet: Depth inference for unstructured multi-view stereo[C]//Proceedings of the European conference on computer vision (ECCV). 2018: 767-783.

[16] Mescheder L, Oechsle M, Niemeyer M, et al. Occupancy networks: Learning 3d reconstruction in function space[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2019: 4460-4470.

[17] Gkioxari G, Malik J, Johnson J. Mesh r-cnn[C]//Proceedings of the IEEE/CVF international conference on computer vision. 2019: 9785-9795.

[18] Wang N, Zhang Y, Li Z, et al. Pixel2mesh: Generating 3d mesh models from single rgb images[C]//Proceedings of the European conference on computer vision (ECCV). 2018: 52-67.

[19] Xie H, Yao H, Sun X, et al. Pix2vox: Context-aware 3d reconstruction from single and multi-view images[C]//Proceedings of the IEEE/CVF international conference on computer vision. 2019: 2690-2698.

[20] Wen C, Zhang Y, Li Z, et al. Pixel2mesh++: Multi-view 3d mesh generation via deformation[C]//Proceedings of the IEEE/CVF international conference on computer vision. 2019: 1042-1051.

[21] Gu X, Fan Z, Zhu S, et al. Cascade cost volume for high-resolution multi-view stereo and stereo matching[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2020: 2495-2504.

[22] Xie H, Yao H, Zhang S, et al. Pix2Vox++: Multi-scale context-aware 3D object reconstruction from single and multiple images[J]. International Journal of Computer Vision, 2020, 128(12): 2919-2935.

[23] Zhao M, Xiong G, Zhou M C, et al. 3D-RVP: A method for 3D object reconstruction from a single depth view using voxel and point[J]. Neurocomputing, 2021, 430: 94-103.

[24] Shi Z, Meng Z, Xing Y, et al. 3D-RETR: end-to-end single and multi-view 3D reconstruction with transformers[J]. arXiv preprint arXiv:2110.08861, 2021.

[25] Wang D, Cui X, et al. Multi-view 3d reconstruction with transformers[C]//Proceedings of the

IEEE/CVF international conference on computer vision. 2021: 5722-5731.

[26] Tiong L C O, Sigmund D, Teoh A B J. 3D-C2FT: Coarse-to-fine Transformer for Multi-view 3D Reconstruction[C]//Proceedings of the Asian Conference on Computer Vision. 2022: 1438-1454.

[27] Zhu Z, Yang L, Lin X, et al. GARNet: Global-aware multi-view 3D reconstruction network and the cost-performance tradeoff[J]. Pattern Recognition, 2023, 142: 109674.

[28] Bahdanau D, Cho K, Bengio Y. Neural machine translation by jointly learning to align and translate[J]. arXiv preprint arXiv:1409.0473, 2014.

[29] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[J]. Advances in neural information processing systems, 2017, 30.

[30] Shaw P, Uszkoreit J, Vaswani A. Self-attention with relative position representations[J]. arXiv preprint arXiv:1803.02155, 2018.

[31] Cordonnier J B, Loukas A, Jaggi M. Multi-head attention: Collaborate instead of concatenate[J]. arXiv preprint arXiv:2006.16362, 2020.

[32] Hu J, Shen L, Sun G. Squeeze-and-excitation networks[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2018: 7132-7141.

[33] Tootell R B H, Hadjikhani N, Hall E K, et al. The retinotopy of visual spatial attention[J]. Neuron, 1998, 21(6): 1409-1422.

[34] 郑太雄, 黄帅, 李永福, 等. 基于视觉的三维重建关键技术研究综述[J]. 自动化学报, 2020, 46(4): 631-652.

[35] 刘圣然.基于体素化识别算法的三维重建技术研究[D].西安电子科技大学,2022.DOI:10.27389/d.cnki.gxadu.2022.002577.

[36] 刘洲岐.基于点云数据处理的三维重建技术研究及平台设计[D].山东理工大学,2023.DOI:10.27276/d.cnki.gsdgc.2023.000215.

[37] YQi C R, Su H, Mo K, et al. Pointnet: Deep learning on point sets for 3d classification and segmentation[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2017: 652-660.

[38] 徐枫.基于深度学习的单视图三维网格重建算法研究[D].南京信息工程大学,2023.DOI:10.27248/d.cnki.gnjqc.2023.000930.

[39] 白景禧,刘春宇,王学军.分组卷积编码的单视图三维重建算法研究[J].石家庄铁道大学学报(自然科学版),2024,37(01):107-113.DOI:10.13319/j.cnki.sjztddxxbzrb.20230216.

[40] 何鑫睿,李秀梅,孙军梅,李美玲,袁珑.基于改进Pix2Vox的单图像三维重建网络[J].计算机辅助设计与图形学学报,2022,34(03):364-372.

[41] Liu S, Cao J, Chen W, et al. HILONet: Hierarchical Imitation Learning from Non-Aligned

Observations[J]. arXiv preprint arXiv:2011.02671, 2020.

[42] Chu X, Tian Z, Zhang B, et al. Conditional positional encodings for vision transformers[J]. arXiv preprint arXiv:2102.10882, 2021.

[43] He K, Zhang X, Ren S, et al. Deep residual learning for image recognition[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 770-778.

[44] Huang G, Liu Z, Van Der Maaten L, et al. Densely connected convolutional networks[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2017: 4700-4708.

[45] 张连超, 乔瑞萍, 党祺玮, 等. 具有全局特征的空间注意力机制[J]. 西安交通大学学报, 2020, 54(11): 129-138.

[46] Woo S, Park J, Lee J Y, et al. Cbam: Convolutional block attention module[C]//Proceedings of the European conference on computer vision (ECCV). 2018: 3-19.

[47] 英陈, 伟张, 洪平林, 等. 医学图像分割算法的损失函数综述[J]. Sheng Wu Yi Xue Gong Cheng Xue Za Zhi= Journal of Biomedical Engineering, 2023, 40(2): 392.

[48] Yin W, Zhang J, Wang O, et al. Learning to recover 3d scene shape from a single image[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021: 204-213.

[49] Danchilla B, Danchilla B. Three. js framework[J]. Beginning WebGL for HTML5, 2012: 173-203.

[50] Rego N, Koes D. 3Dmol. js: molecular visualization with WebGL[J]. Bioinformatics, 2015, 31(8): 1322-1324.