分类号\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ 密级\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

ＵＤＣ\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ 编号\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

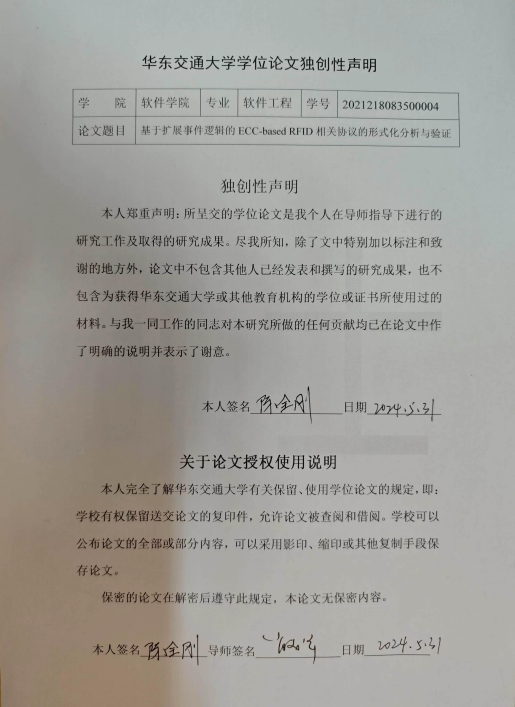
****

**硕士学位论文**

|  |  |
| --- | --- |
| **基于双注意力的单视图体素三维重建研究** | |
|  | |
| **学位申请人：** | **李承欢** |
| **专业领域：** | **软件工程** |
| **校内导师：** | **肖美华 教授** |
|  |  |
|  | |
|  | |
|  | |

**答辩日期：2025.5.26**

|  |
| --- |
| **华东交通大学2025届学术硕士学位论文**  **书脊** |
| **基于双注意力的单视图体素三维重建研究** |
| **信息与软件工程学院** |
| **李承欢** |



**基于双注意力的单视图三维重建研究**

**摘要**

三维重建是计算机视觉领域一个重要研究方向，是从2D的图像数据中恢复3D数据的过程。这些3D数据可以是被重建对象的几何拓扑数据，纹理颜色数据等。从宏观上可以将三维重建分为传统三维重建方法，基于深度学习的三维重建方法。传统三维重建方法已经有几十年的历史，相关理论和技术已经非常成熟，但是严重依赖几何原理和数学优化，并且需要被重建对象的多视角图像数据，面对具有无纹理区域、重复图案的重建对象鲁棒性较差。最近基于深度学习的三维重建方法越来越火热，通过利用深度神经网络来学习 2D 图像和 3D 结构之间的复杂关系，这种方法在面对具有无纹理区域、重复图案的重建对象时具有较好的鲁棒性。但是目前大多数此类重建方法是在3D模型渲染图像数据集上训练的，在面对真实世界拍摄的图像时鲁棒性大大降低，尽管已经有人提出了将被重建对象从复杂背景中分割出再重建的方法，但是这并没有从本质上提高模型的鲁棒性。我们从图像分类，语义分割等其他视觉任务获得灵感，提出了一个基于双注意力机制的体素三维重建模型R3Davit，该模型在面对真实 图像输入可以不依赖前景分割直接重建对象为体素模型，在公开合成数据集ShapetNet和真实数据集Pix3D与最近的先进工作比较，我们在IOU和F1分数指标领先%1。最后为了验证该三维重建系统的可应用性，我们基于所提出的模型设计了一个单视图三维重建软件系统。

本文的主要工作如下：

1. 基于双注意力机制设计了一个体素三维重建模型，其中编码器基于双注意力机制设计，分别利用了空间窗口注意力建立不同补丁间空间维度相关性，通道分组注意力建立不同补丁通道维度相关性，解码器基于选择性状态空间模型设计以替代传统的注意力层，同时额外设计了一个非线性增强块增强解码器的非线性表达能力。

2. 我们同时在公开合成数据集和真实数据集进行了实验，最近的先进工作比较，我们在IOU和F1分数指标领先%1，消融实验证明了我们设计的编码器，解码器的有效性。为了使模型对真实图像重建更有鲁棒性，我们将SUN数据集作为背景与ShapeNet数据集作为前景进行图像组合生成了一个新的具有随机复杂背景的数据集ShapeNetRFC。

3. 基于所提出的三维重建模型，我们设计了一个单视图三维重建系统，通过前端后端分离设计，提高了不同设备平台的用户交互体验感。

**关键词**：三维重建，深度学习，双注意力机制，体素模型，ShapeNet，Pix3D

**Research on Single View 3D Reconstruction Based on Dual Attention**

**ABSTRACT**

3D reconstruction is an important research direction in the field of computer vision. It is the process of recovering 3D data from 2D image data. These 3D data can be geometric topological data, texture color data, etc. of the reconstructed object. From a macro perspective, 3D reconstruction can be divided into traditional 3D reconstruction methods and 3D reconstruction methods based on deep learning. Traditional 3D reconstruction methods have a history of several decades, and the relevant theories and technologies are very mature, but they rely heavily on geometric principles and mathematical optimization, and require multi-view image data of the reconstructed object. They have poor robustness in the face of reconstructed objects with textureless areas and repeated patterns. Recently, 3D reconstruction methods based on deep learning have become increasingly popular. By using deep neural networks to learn the complex relationship between 2D images and 3D structures, this method has good robustness when facing reconstructed objects with textureless areas and repeated patterns. However, most of these reconstruction methods are currently trained on 3D model rendering image datasets, and their robustness is greatly reduced when facing images taken in the real world. Although some people have proposed methods to segment the reconstructed object from the complex background and then reconstruct it, this does not fundamentally improve the robustness of the model. We got inspiration from other visual tasks such as image classification and semantic segmentation, and proposed a voxel 3D reconstruction model R3Davit based on a dual attention mechanism. This model can directly reconstruct objects as voxel models without relying on foreground segmentation when facing real image input. Compared with recent advanced work on the public synthetic dataset ShapetNet and the real dataset Pix3D, we lead by 1% in IOU and F1 score indicators. Finally, in order to verify the applicability of the 3D reconstruction system, we designed a single-view 3D reconstruction software system based on the proposed model.

The main work of this paper is as follows:

1. A voxel 3D reconstruction model is designed based on the dual attention mechanism. The encoder is designed based on the dual attention mechanism, and the spatial window attention is used to establish the spatial dimension correlation between different patches. The channel grouping attention is used to establish the channel dimension correlation between different patches. The decoder is designed based on the selective state space model to replace the traditional attention layer. At the same time, a nonlinear enhancement block is additionally designed to enhance the nonlinear expression ability of the decoder.
2. We conducted experiments on both public synthetic datasets and real datasets. Compared with the latest advanced work, we lead by 1% in IOU and F1 score indicators. Ablation experiments have proved the effectiveness and scalability of the encoder and decoder we designed. In order to make the model more robust to real image reconstruction, we combined the SUN dataset as the background and the ShapeNet dataset as the foreground to generate a new dataset ShapeNetRFC with random complex background.

3. Based on the proposed 3D reconstruction model, we designed a single-view 3D reconstruction system. Through the front-end and back-end separation design, the user interaction experience on different device platforms is improved.

**Key Words:** 3D Reconstruction, Deep Learning, Dual Attention Mechanism, Voxel Model, ShapeNet, Pix3D

**目录**

[主要符号说明 VI](#_Toc12524)

[第一章 绪论 1](#_Toc20225)

[1.1 研究背景及意义 1](#_Toc21317)

[1.2 国内外研究现状 1](#_Toc15070)

[1.3 本文主要内容 4](#_Toc14096)

[1.4 论文结构安排 5](#_Toc14579)

[第二章 深度学习相关理论基础 6](#_Toc24977)

[2.1 深度学习相关网络 6](#_Toc26580)

[2.1.1 卷积神经网络 6](#_Toc29666)

[2.1.2 循环神经网络 7](#_Toc21927)

[2.1.3 双注意力视觉转换器 8](#_Toc14142)

[2.1.4 选择性状态空间模型 11](#_Toc29627)

[2.2 三维重建相关技术 12](#_Toc30894)

[2.2.1 重建模型表达形式 12](#_Toc24580)

[2.2.2 基于深度学习的三维重建相关理论 12](#_Toc5923)

[2.2.3 三维重建相关数据集 12](#_Toc31575)

[2.3 本章小结 13](#_Toc32567)

[第三章 基于双注意力的体素重建算法研究 14](#_Toc14997)

[3.1 引言 14](#_Toc27185)

[3.2 算法设计 14](#_Toc21034)

[3.2.1 网络结构 14](#_Toc8709)

[3.2.2 损失函数 14](#_Toc5772)

[3.2.3 数据集 14](#_Toc23136)

[3.2.4 评价指标 14](#_Toc409)

[3.2.5 实现细节 14](#_Toc28416)

[3.3 实验结果与分析 14](#_Toc3079)

[3.3.1 无背景的合成数据集 14](#_Toc6634)

[3.3.2 具有随机背景的合成数据集 14](#_Toc17227)

[3.3.3 具有复杂背景的真实数据集 14](#_Toc2110)

[3.3.4 消融实验 14](#_Toc1939)

[3.4 本章小结 14](#_Toc10936)

[第四章 单视图三维重建软件系统设计与实现 15](#_Toc31442)

[4.1 引言 15](#_Toc18202)

[4.2 需求分析 15](#_Toc12526)

[4.2.1 功能性需求 15](#_Toc14873)

[4.2.2 非功能性需求 15](#_Toc29325)

[4.2.3 设计约束 15](#_Toc5510)

[4.3 系统设计 15](#_Toc12050)

[4.3.1 系统概要设计 15](#_Toc25786)

[4.3.2 系统总体框架设计 15](#_Toc24524)

[4.3.3 系统流程设计 15](#_Toc29090)

[4.4 系统实现 15](#_Toc28014)

[4.4.1 系统开发环境 15](#_Toc30419)

[4.4.2 关键技术实现 15](#_Toc29238)

[4.5 系统界面展示 15](#_Toc20365)

[4.5.1 主界面 15](#_Toc18039)

[4.5.2 针对合成图像的体素重建系统 15](#_Toc31902)

[4.5.3 针对真实图像的体素重建系统 15](#_Toc29288)

[4.6 系统测试 15](#_Toc19834)

[4.6.1 界面测试 16](#_Toc14175)

[4.6.2 功能测试 16](#_Toc14093)

[4.6.3 性能测试 16](#_Toc13279)

[4.7 本章小结 16](#_Toc28167)

[第五章 总结与展望 17](#_Toc292)

[5.1 全文工作总结 17](#_Toc18620)

[5.2 展望 17](#_Toc24088)

[参考文献 18](#_Toc19659)

[个人简介 在读期间发表学术论文 21](#_Toc23245)

[致谢 22](#_Toc12069)

# 主要符号说明

 输入图像的宽

 输入图像的高

 步长

 填充

 为输出的特征图尺度

 为输入的特征图尺度，

 为卷积核大小

 为填充值

 为步长

 时间步

 时间步时刻的输入

 时间步时刻的隐藏状态

 函数

 查询变量，键向量，值向量

 第个线性变换权重矩阵

 拼接后的权重矩阵

 S4模型的参数

 指数计算

# 绪论

## 研究背景及意义

在计算机视觉领域中，三维重建是指根据被重建对象的图像特征恢复对象的三维数据的过程，这些三维数据可以是对象的形状，颜色，材质等信息。三维重建在很多 重要领域具有重要意义，例如自动驾驶中的下游建图任务，传统文物的数字化保护，游戏资产的自动化制作，工业三维产品设计等。

根据三维重建发展的历史轨迹，可以将其划分成传统三维重建方法，基于深度学习的三维重建方法。传统三维重建方法中以运动恢复结构[1]，多视图立体视觉为代表[2]，这种数学优化的方法需要通过提取图像序列中具有代表性的特征点并进行匹配，并进行多视图几何关系估计相机运动参数等一系列复杂步骤来进行三维重建。并且面对具有平坦均匀纹理的对象重建时鲁棒性不高。

随着大型三维重建数据集的提出，如ShapNet[3]，Pix3d[4]，基于深度学习的三维重建方法逐渐热门，根据输入的图像数量，基于深度学习的三维重建可以分为单视图三维重建[5][6][7][13][16][17][18][19][21][23][24][25][29][30]，多视图三维重建[8][9][10][11][12][14][15][20][22] 。模型表示形式又分为基于体素的三维重建[5][6][7][13][16][17][18][19][21][23][24][8][9][10][11][12][14][15][20][22]，基于点云的三维重建[28][29][30][31][32][33][34]，基于多面体几何的三维重建[25][26][35][36]，以及基于隐式表达的三维重建[26][37][38]。在本文中，我们专注研究以体素为表达形式的单视图三维重建。顾名思义，单视图三维重建可以定义为：仅以重建对象的单个图像信息为输入，输出重建结果。然而人们使用移动手机，等设备采集的图像往往是充满了复杂背景的，这些复杂背景带来的嘈杂信息对三维重建模型来说是一个极具挑战性的因素，并且由于在单视图中存在重建对象的部分自我遮挡区域，如何提高模型对被遮挡区域重建效果也是一个值得研究的问题。

因此，不论从实际应用层面还是学术研究层面出发，如何设计一个能面对复杂背景下的单视图三维重建模型系统是一个非常具有学术，应用价值的课题。

## 国内外研究现状

随着深度学习的发展，以及大型三维重建数据集的提出，以体素为表达形式的三维重建方法也出现了一系列代表性工作，这些工作随着更先进的深度学习通用主干网络的提出而不断迭代。

2016年，Choy[24]等人提出了一种新的循环神经网络架构，称之为3D循环重建神经网络3D-R2N2。该网络从大量合成数据中学习从物体图像到其底层3D形状的映射。从任意视点获取对象实例的一张或多张图像，并以体素的形式输出对象的重建。与之前的大多数工作不同，3D-R2N2不需要任何图像注释或对象类标签来进行训练或测试。

2017年，Kar[22]等人提出了一个学习的多视角立体视觉系统LSM。与最近基于学习的3D重建方法不同，LSM通过特征投影和沿观察光线的反投影来利用问题的潜在3D几何。通过以可微分的方式表述这些操作，LSM能够学习系统端到端的度量三维重建任务。端到端学习使LSM能够在符合几何约束的情况下共同推理形状先验，使LSM能够从比经典方法所需的更少的图像(甚至单个图像)进行重建，并完成看不见的表面。

2017年，Tatarchenko[23]等人提出了一种深度卷积解码器架构OGN，OGN可以通过使用八叉树表示以计算和内存效率的方式生成体积3D输出。网络学习预测八叉树的结构和单个细胞的占用值。这使得它成为生成3D形状的特别有价值的技术。与作用于规则体素网格的标准解码器相比，该架构没有立方复杂度。这允许在有限的内存预算下表示更高分辨率的输出。OGN在几个应用领域展示了这一点，包括3D卷积自编码器，从高级表示生成对象和整个场景，以及从单个图像生成形状。

2018年，Wang[25]等人提出了一种端到端深度学习架构Pixel2Mesh，可以从单色图像中生成三角形网格中的3D形状。受深度神经网络特性的限制，以前的方法通常是用体积或点云表示三维形状，将它们转换为更易于使用的网格模型是很困难的。与现有方法不同，Pixel2Mesh在基于图的卷积神经网络中表示3D网格，并通过逐步变形椭球来产生正确的几何形状，利用从输入图像中提取的感知特征。Pixel2Mesh采用了从粗到精的策略，使整个变形过程稳定，并定义了各种网格相关的损失来捕捉不同层次的属性，以保证视觉上的吸引力和物理上的精确3D几何。

2019年，Xie[18]等人提出了一个新的单视图和多视图三维重建框架，命名为Pix2Vox。通过使用精心设计的编码器-解码器，它从每个输入图像生成粗3D体。然后，引入上下文感知融合模块，自适应地从不同粗三维体中选择每个部件(如桌腿)的高质量重建，以获得融合的三维体。最后，细化器进一步细化融合的3D体以生成最终输出。

2020年，Xie[17]等人对已经提出的Pix2Vox进行了升级，并命名为Pix2vox++。使用了新的主干网络ResNet[39]。

2021年，Yang[19]等人提出了一种名为Mem3D的新方法，该方法明确地构建形状先验来补充图像中缺失的信息。具体而言，形状先验在记忆网络中以“图像-体素”对的形式存在，在训练过程中通过精心设计的书写策略进行存储。Mem3D还提出了一个体素三重损失函数，它有助于从形状先验中检索与输入图像高度相关的精确3D形状。还引入了基于LSTM[40]的形状编码器，从检索到的三维形状中提取信息，可用于在严重遮挡或复杂环境中恢复物体的三维形状。Shi[16]等人提出3D- RETR来填补Vision Transformer[41]能否用于3D重建这一空白，它能够使用Vision Transformer[41]执行端到端的3D重建。3D-RETR首先使用预训练的Vision Transformer[41]从2D输入图像中提取视觉特征。3D-RETR然后使用另一个Transformer[42]解码器来获得体素特征。然后，CNN解码器将体素特征作为输入来获得重建的对象。Yagubbayli[14]等人提出了LegoFormer，这是一个基于体素的3D重建的Transformer[42]模型，它在所有计算阶段使用注意层在视图之间共享信息。此外，LegoFormer使用一系列低秩分解因子来参数化输出，而不是独立预测每个体素，类似乐高搭积木的方法将物体的不同部分独立重建并累积。这种重新表述允许将对象预测为一组独立的规则结构，然后聚合以获得最终重建。Wang[15]等人提出了EVoiT，将三维重建重新表述为序列到序列的预测问题，并提出了一个名为EVoiT的框架。不同于以往使用独立设计的基于CNN[43]的方法，EVoiT将特征提取和视图融合统一在一个Transformer[42]网络中。EVoiT设计的一个天然优势在于利用多个无序输入之间的自注意力机制来探索视图与视图之间的关系。

2022年，Tiong[11]等人提出了3D-C2FT网络，引入了一种新颖的粗到精(C2F)注意机制来编码多视图特征并校正基于体素的缺陷三维物体，提出了一种新的模型3D粗到精变压器(3D-C2FT)。C2F注意机制使模型能够学习多视图信息流，并以粗粒度到细粒度的方式综合三维曲面校正。

2023年，Zhu[10]等人提出了一种全局感知的基于注意力的融合方法GarNet，该方法建立了每个分支与全局特征之间的相关性，为权重推理提供了全面的基础。在此基础上，GarNet设计了一个完整的重构算法。GarNet提出了一种基于多样性最大化的视图缩减方法。Zhu[8]等人还提出了一种新的非结构化多图像Transformer[42]三维重建模型UMIFormer。UMIFormer利用转换块来解耦视图内编码，并设计了token校正块来挖掘不同视图中相似token之间的相关性，以实现解耦访谈编码。然后，从各个分支获得的所有令牌被压缩成固定大小的紧凑表示，同时通过利用令牌之间的相似性来保留丰富的信息以供重建。Yang[9]等人提出了基于分而治之原则的远程分组注意(LGA)。来自所有视图的令牌被分组以进行单独的注意操作。每个组中的令牌从所有视图中采样，可以为驻留视图提供宏表示。不同群体之间的多样性保证了特征学习的丰富性。利用LGA连接视图间特征，利用标准自注意力机制层提取视图内特征，建立了一种高效的编码器。此外，还设计了一种新颖的逐进上采样解码器，以实现较高分辨率的体素生成。在此基础上，构建了一个强大的基于变压器的网络，称为LRGT。Gao[7]等人提出了提出了一个深度三维重建网络DASI来学习领域自适应形状印象，以便从任意视图图像中进行三维重建。DASI包括两个模块:形状重构模块和形状细化模块。前者通过学习域自适应形状印象作为图像重构的嵌入，重构粗体;DASI首先利用3D对象来学习与3D对象的先验知识相关的形状印象。为了从二维图像中获得一致的形状印象，DASI将三维形状和二维图像视为两个不同的域。通过适应这两个域，将从3d物体中学习到的形状印象转移到2D图像中，并指导基于图像的重建。后一个模块通过将整个3D体建模为局部3D块并探索其内在几何关系来细化对象。

2024年，Xiong[6]等人引入了一种新的单视图3D体素重建模型，该模型利用跨领域特征融合，称为SV3D-CDFF。SV3D-CDFF利用跨域特征（图像特征与体素特征）聚类来减轻数据分布差异的影响，并采用特征监督方法学习体素特征。此外，它还引入了注意机制来融合图像和体素特征，并利用残差网络进行三维体素重建。

自2016年以来，已经有多个先进工作被提出，但是他们并不是全部支持真实图像重建，尽管有的模型支持真实图像重建，却依然需要对图像进行去背景处理，并且面对被遮挡区域时的鲁棒性不佳，有的工作则无法支持单视图输入。

窗体顶端

## 本文主要内容

本文主要研究基于深度学习的单视图三维重建，以体素为表达形式，我们从其他视觉任务中获取灵感，设计更高效的，鲁棒的单视图三维重建模型。

本文的主要内容如下：

（1）基于双注意力机制的单视图三维重建网络研究

我们提出了一个利用空间维度注意和通道维度注意进行单视图三维重建的网络，命名为R3Davit[5]。具体来说，R3Davit[5]由一个编码器和一个解码器组成。

其中编码器来自Davit[44]骨干网。与以往的变压器骨干网络不同，Davit[44]侧重于空间维度和通道维度，在保持线性复杂性的同时，充分构建全局上下文信息和局部细粒度信息。

为了有效地从双重注意中学习特征并保持网络的整体推理速度，我们没有在解码器中使用自注意力机制层，而是设计了一个具有非线性增强块、选择状态空间模型块和上采样残差块的解码器。非线性增强块用于增强网络的非线性学习能力。选择状态空间模型块取代了自注意力机制层的作用，保持了线性复杂性。上采样残差块将体素特征转换为体素模型，同时保留该层的体素。特征用于下一层的上采样块。

1. 在公开合成数据集和真实数据集进行大量实验

在纯背景图像数据集ShapeNet[3]中，我们与最近的先进工作比较，在IOU和F1分数指标领先%1，消融实验证明了我们设计的编码器，解码器的有效性。同时，为了使模型对真实图像重建更有鲁棒性，我们将SUN[45]数据集作为背景与ShapeNet[3]数据集作为前景进行图像组合生成了一个新的具有随机复杂背景的数据集ShapeNetRFC。在Pix3D[4]数据集上的实验表明，我们提出的R3Davit[5]网络在面对真实图像时具有更好的鲁棒性。

1. 单视图三维重建软件系统设计与实现

为了验证提出的三维重建模型可应用性，我们基于该模型设计了一个单视图三维重建软件，该软件可以将输入的图像进行规范裁剪加工，并重建为体素模型进行可视化交互。

## 论文结构安排

本文的结构安排如下：

第一章：绪论。本章节包括研究背景及意义，国内外研究现状，本文主要内容。研究背景及意义主要介绍了三维重建的定义和不同方法的三维重建，以及三维重建的实际应用领域，国内外研究现状总结了自2016年以来的先进工作及这些先进工作的方法，旨在让读者对整个领域的历年先进工作有个大致印象以区别本文的工作。本文主要内容详细描述了本文的工作要点，及创新点。

第二章：深度学习相关理论基础。该章节不仅介绍了经典的深度学习相关网络，还紧跟最新技术，更加详细描述了视觉转换器（Vision Transformer[41]），选择性状态空间模型（Mamba[46]）的相关理论及原理。同时介绍了不同的三维重建结果表达形式，以及相关数据集。

第三章：基于双注意力的体素重建算法研究。该章节详细描述了所提出的R3Davit[5]整体架构及各个模块的实现细节，还介绍了使用的损失函数、数据集、评价指标等内容。最后我们展示了大量的实验及可视化比较，其中包括无背景的合成数据集ShapeNet[3]，具有随机背景的合成数据集ShapeNetRFC，具有复杂背景的真实数据集Pix3D[4]，以及消融实验。

第四章：单视图三维重建软件系统设计与实现。该章节根据第三章提出的R3Davit[5]单视图三维重建模型设计了一个软件系统，并详细介绍了该软件系统的需求分析，系统设计，实现，等内容。

第五章：总结与展望。该章节总结本文的研究结论及成果，并提出未来可能的研究方向，供后续学者参考。

# 深度学习相关理论基础

## 深度学习相关网络

深度学习通过构建具有很多层的神经网络模型，让计算机自动从大量数据中学习特征和模式，以实现对数据的分类、预测、生成等任务。这些神经网络能够自动提取数据中的高级特征，无需人工手动设计特征工程，从而大大提高了模型对复杂数据的处理能力。经典的深度学习网络有卷积神经网络[47]、循环神经网络[48]，除此之外也有更先进的网络架构被提出，如转换器（Transformer[42]）、视觉转换器（Vision Transformer[41]）、选择性状态空间模型（Mamba[46]）。下文将详细介绍研究过程中涉及的相关网络技术。

### 卷积神经网络

在生物学和神经科学领域，对动物视觉系统的研究为卷积的提出提供了重要的灵感来源。研究发现，动物的视觉神经元具有局部感受野的特性，即每个神经元只对视觉场景中的一个局部区域敏感。例如，在猫的视觉皮层中，一些神经元对特定方向的边缘或线条有强烈的响应，这种局部感知的特性启发了人们在图像处理和计算机视觉中引入类似的局部操作，即卷积操作。卷积操作共有两种：常规卷积、转置卷积。以步长大于1的卷积为例，其中常规卷积用于提取特征，随着常规卷积层数的增加特征图的大小会变得更小，属于下采样操作；而转置卷积用于上采样特征，随着转置卷积层得增加特征图的大小会变得更大。被提取特征的输入可能是2D数据例如图像，也可能是3D数据例如体素特征，连续视频帧，因此卷积又可分为2D卷积 与3D卷积。2D卷积如图2-1，3D卷积如图2-2：

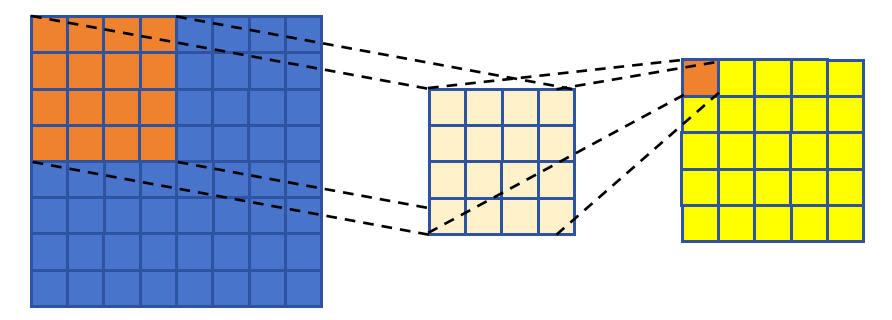


图2-1 2D卷积

Fig.2-1 2D Convolution

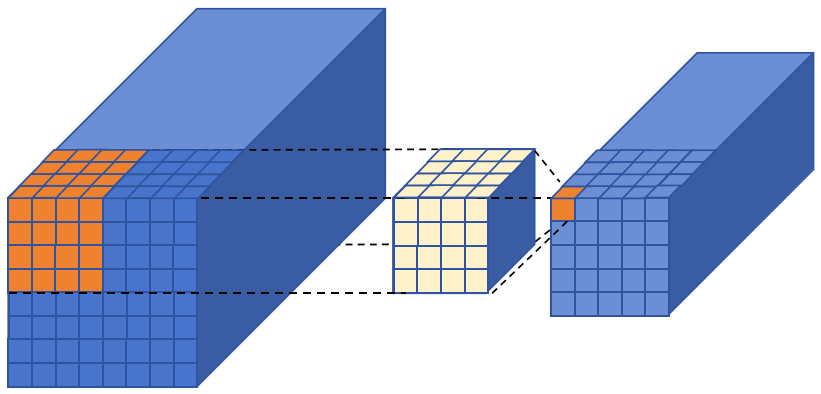


图2-2 2D卷积

Fig.2-2 3D Convolution

以2D卷积为例，设输入图像为单通道灰度图像，其宽为，高为，卷积核为一个二维矩阵，该矩阵中的每个元素都是可学习的参数，会在神经网络训练过程中不断更新。将卷积核放置在输入图像的左上角，使其与输入图像的对应位置元素相乘，然后将所有乘积相加得到卷积结果的一个元素，卷积顺序按照从左到右，从上到下的顺序执行，每次可在输入图像上滑动一个步长值，重复上述操作直到卷积核遍历整个输入图像。除进行滑动步长操作外，还可以设置一个填充值，进行填充处理。填充处理可以控制卷积层输出的特征图尺寸并且可以防止边缘信息在卷积过程中过度丢失。输出的特征图尺度计算公式：

 （2-1）

式（2-1）中为输出的特征图尺度，为输入的特征图尺度，为卷积核大小，为填充值，为步长。

而3D卷积则是在2D卷积基础上增加了一个维度，其余计算方式与2D卷积相同。

### 循环神经网络

循环神经网络[48]RNN是一类用于处理序列数据的神经网络，在自然语言处理，语音识别，时间序列预测等领域有广泛应用，在视觉领域中，研究者们常常将图像的像素数据摊平后再喂给循环神经网络进行训练。与传统的前馈神经网络不同，RNN具有循环连接，允许信息在神经元之间的传递形成一个闭环，从而使得网络能够对之前的输入信息进行记忆和利用。具体来说，在每个时间步,RNN不仅接收当前的输入，还接收来自上个时间步的隐藏状态，并通过一个函数来更新当前的隐藏状态，如公式（2-2）所示：

 （2-2）

隐藏状态可以看作是RNN对截止当前时间步的所有输入信息的一种记忆表示，它综合了当前输入和之前输入的隐藏状态信息，通过不断更新和传递，使得RNN能够捕捉到序列数据中的长期依赖关系。这种记忆与反馈机制是RNN处理序列数据的核心。RNN结构如图2-3所示：

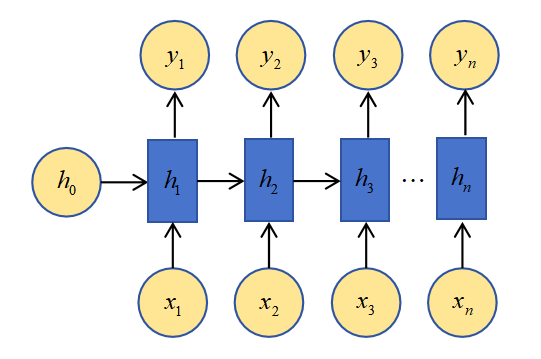


图2-3 循环神经网络

Fig.2-3 RNN

其中为输入序列的一个元素，为输出结果，为隐藏状态。最基本的循环神经网络由输入层，隐藏层，输出层组成。其中输入层负责接收外部输入序列中的每个元素，将其转换为网络能够处理的向量形式，并传递给隐藏层进行处理，隐藏层包含多个神经元，这些神经元通过循环连接来处理序列中的每个时间步的信息，隐藏层中的神经元会根据当前的输入和上一时刻的隐藏状态更新自己的状态，并将更新后的隐藏状态传递到下个时间步，同时也会生成一个输出，该输出可能会作为下一层的输入或作为网络最终输出的一部分。输出层根据隐藏层的输出产生最终预测结果。

### 双注意力视觉转换器

双注意力视觉转换器是基于初始版本的视觉转换器（Vision Transformer[41]）的一个变体版本。称为Davit[44]骨干网络。下面我们先描述初始版本的Vision Transformer[41]，再更加详细地描述双注意力视觉转换器Davit[44]骨干网络。

Vision Transformer[41]在传统Transformer[42]基础上增加了图像分块操作使得它更加适合视觉类任务。如图2-4所示：

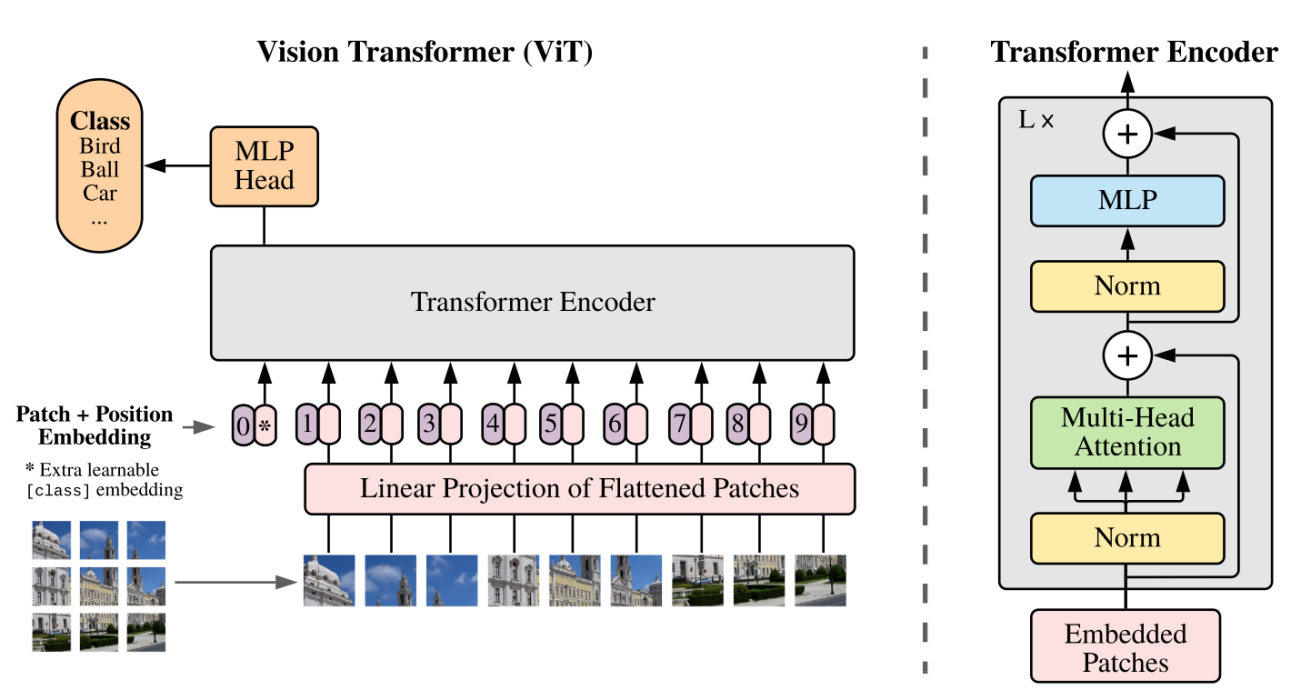


图2-4 视觉转换器[41]

Fig.2-4 Vision Transformer[41]

Vision Transformer[41]先将图像分割为多个小块（Patch），然后经过线性变换映射为类似序列形式的向量，再添加位置编码来保留空间信息。这些可学习的位置编码能够适应不同图像的空间特征，让模型学习到图像中不同小块（Patch）之间的空间关系。之后数据进入Transformer[42]进行多头自注意力（Multi-Head Attention）计算用于捕捉各个小块（Patch）之间的依赖关系，以此获取图像的全局信息。多头自注意力机制公式（2-3）：

 （2-3）

其中为单头注意力，为多头注意力，为查询向量表示需要关注的信息，为键向量表示内容的特征，为值向量表示要传递的信息，是键向量的维度，是为了防止数值过大导致梯度不稳定而设置的缩放值。分别是对应的第线性变换权重矩阵，为拼接操作，是拼接后的权重矩阵。多头自注意力能够自动学习序列中的不同特征，每个头可以关注到不同方面的信息。函数用于将得分转换为概率分布，得到注意力权重，再与值向量相乘得到自注意力的输出。

Davit[44]骨干网络从一个正交的角度来保持计算效率的同时捕获全局上下文，利用空间令牌（Spatial Tokens）和通道令牌（Channel Tokens）的自注意力机制机制。其中Token由Patch经过线性映射得到，是Patch的特征表示形式。对于空间令牌，空间维度定义令牌范围，通道维度定义令牌特征维度。对于通道令牌，Davit[44]有相反的情况：通道维度定义令牌范围，空间维度定义令牌特征维度。Davit[44]进一步将空间和通道令牌沿着序列方向分组，以保持整个模型的线性复杂性。这两种自注意力相互补充：由于每个通道令牌包含整个图像的抽象表示，因此在计算通道之间的注意力得分时，通道注意通过考虑所有空间位置来自然地捕获全局交互和表示；空间注意力通过跨空间位置执行细粒度交互来细化局部表征，这反过来又有助于通道注意力的全局信息建模。Davit[44]骨干网络架构如图2-5所示：

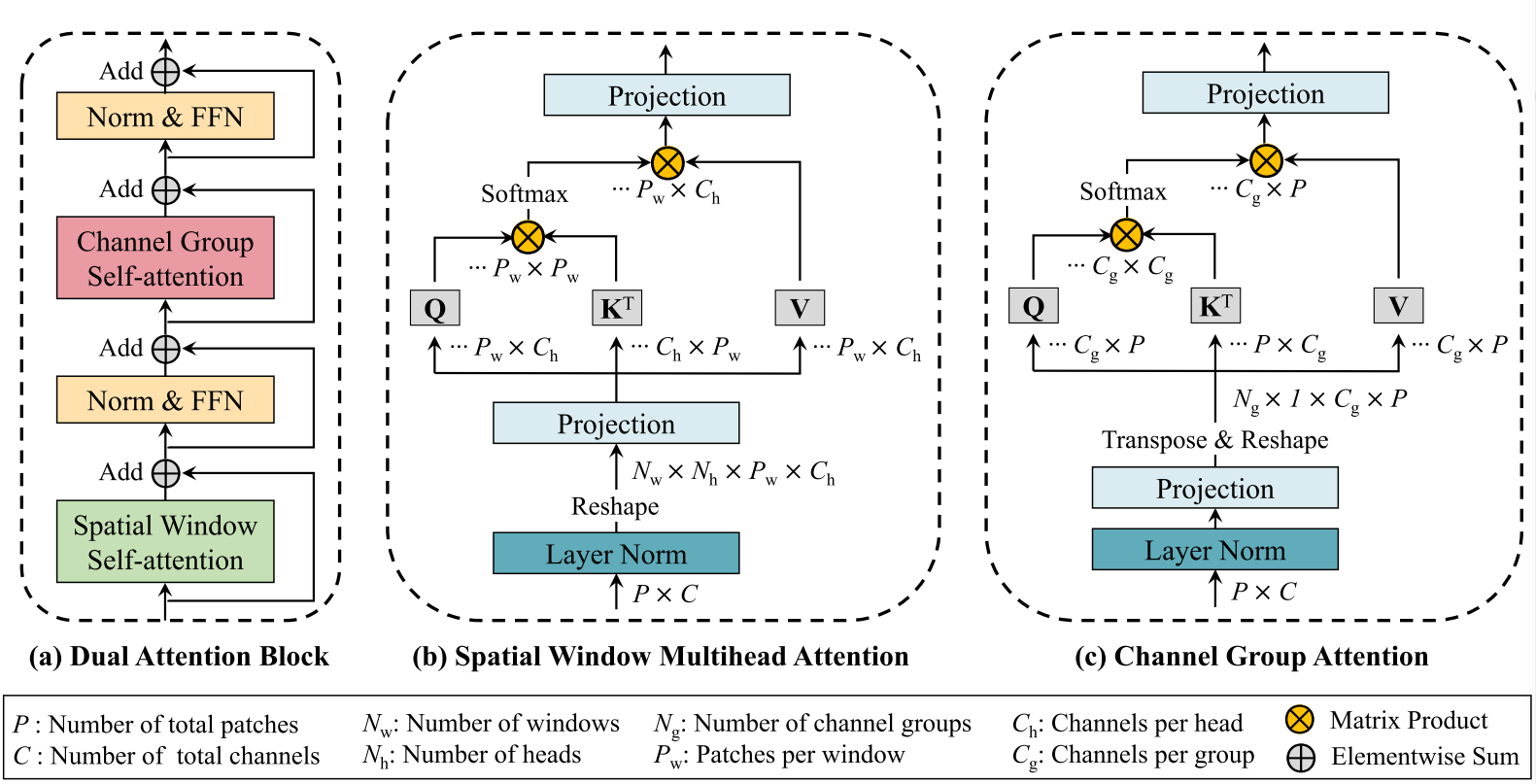


图2-5 Davit[44]骨干网络

Fig.2-5 Dual Vision Transformer[44]

在单个双注意力块（Dual Attention Block）包含一个空间窗口多头注意力（Spatial Window Multihead Attention），一个通道分组注意力（Channel Group Attention）。空间窗口多头注意力计算机局部窗口内的自注意力，所述窗口被安排为以不重叠的方式对图像进行分区，假设这里有个不同窗口，每个窗口包括个Patches。则。

如式（2-4）所示：

 （2-4）

式中为局部窗口的查询、键、值。空间窗口多头注意力是在Patch层级执行注意力计算，而通道分组注意力是在Patch层级的转置上计算注意力。简单将特征转置再计算注意力将造成复杂度，为了进一步降低计算复杂度，Davit[44]将通道分成多个组合，再对每个组合分别进行自注意力计算。

### 选择性状态空间模型

目前已有的体素三维重建先进工作[7][8][9]都是以Vision Transformer[41]或者其变体[49]为基础模型，虽然Vision Transformer[41]的图像分块操作能够在一定程度上缓解传统Transformer[42]的长序列计算复杂度增加问题，但是注意力计算依旧是二次时间复杂度的。而最近提出的选择性状态空间模型（Linear-Time Sequence Modeling with Selective State Spaces[50]，简称S6模型）在拥有了Transformer[42]的长序列建模能力同时，还保持了线性时间复杂度。

选择性状态空间模型（Linear-Time Sequence Modeling with Selective State Spaces[50]）是在初始版本SSM[51]（简称S4模型）基础上发展而来的，初始版本SSM[51]无法执行基于内容的推理。而S6[50]模型对S4[51]进行了改进，首先将S4[51]参数成为输入的函数，通过离散模态使其适应各种离散信息，如文本，语音，图像。并且提出了选择性机制来允许模型可以根据当前令牌（token）有选择地沿序列长度维度传播或忘记信息。

状态空间模型S4[51]与循环神经网络[48]RNN，经典状态空间模型息息相关。状态空间模型S4[51]的灵感来自一个将一维序列通过隐藏状态映射到另一维序列的持续模型。S4[51]用四个参数来定义序列到序列转换的两个阶段。如式（2-5）、（2-6）、（2-7）所示：

 （2-5）

 （2-6）

 （2-7）

第一阶段将连续变量通过零阶保持规则（Zero-Order Hold（ZOH））离散化为，其中，，被称为离散化规则。其定义（2-8）所示：

 （2-8）

第二阶段为计算，在上一阶段已经将连续变量离散化为，计算方式可以为线性递归如式（2-6），全局卷积如式（2-7），通常，该模型使用卷积模式进行有效的并行训练，这样可以提前看到整个输入序列，并切换到线性递归模式进行有效的自回归推理，每次看到一个时间步。变量，，。为了操作具有通道的批量大小和长度的输入序列，SSM独立地应用于每个通道。每个输入的总隐藏状态的维度为，并且在序列长度上计算它需要时间复杂度。S4[51]，S6[50]算法流程如图2-6所示：

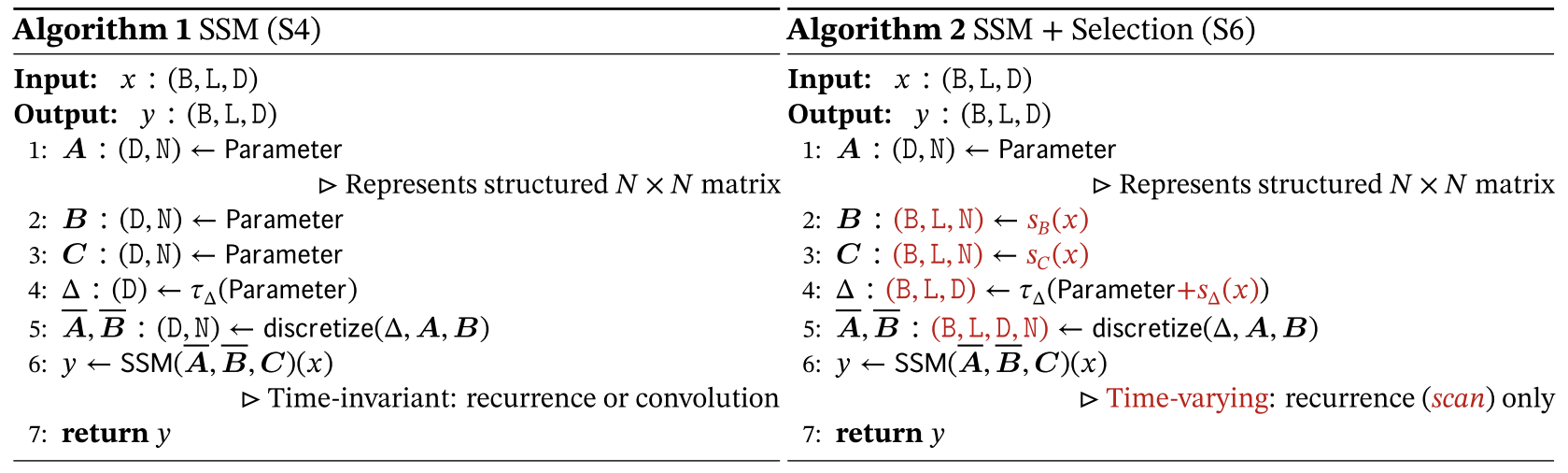


图2-6 S4[51]，S6[50]

Fig.2-6 S4[51]，S6[50]

其中，，，，是参数化投影到维度。

## 三维重建相关技术

### 重建模型表达形式

重建模型表达形式即被重建的物体在计算机中的存储方式，截止目前常用的表达方式有点云[28][29]，体素[8][9][10][11][12][14][15][20][22]，多面体几何[25][26][35][36]，以Nerf[37]和3DGS[38]为代表的隐式表达[26][37][38]。其中体素模型如图2-7所示：

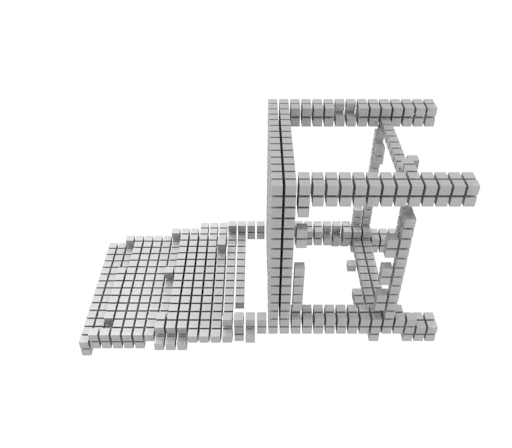


图2-7 来自R3Davit[5]的体素重建结果

Fig.2-7 Result From R3Davit[5]

### 基于深度学习的三维重建相关理论

在基于深度学习的三维重建中[17][16][15][10][9][8][7]，研究者常常基于编码器（Encoder）、融合器（Merger）、解码器（Decoder）这种架构设计三维重建模型。对于单视图三维重建模型，无须设计融合器，而多视图三维重建则需要设计融合器来融合来自多个不同视图的特征到固定特征大小来达到端到端的目的。如图2-8，图2-9，所示：

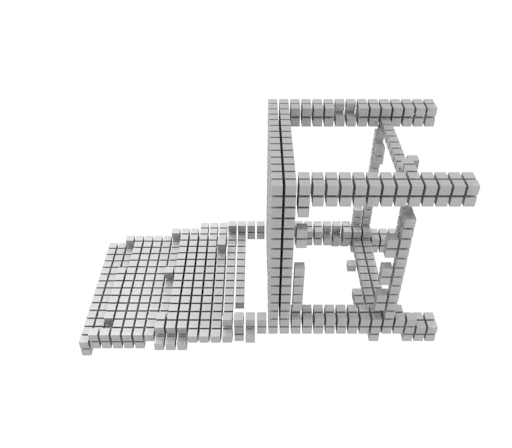


图2-8 来自R3Davit[5]的体素重建结果

Fig.2-8 Result From R3Davit[5]

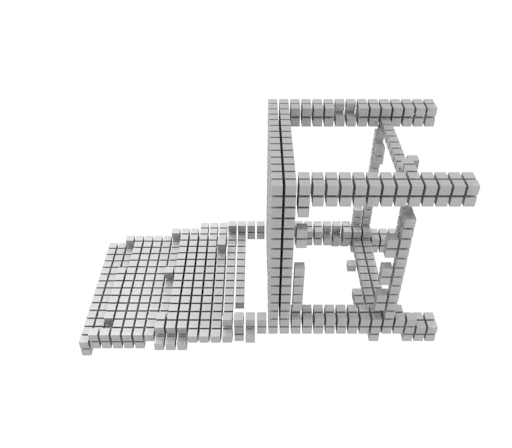


图2-9 来自R3Davit[5]的体素重建结果

Fig.2-9 Result From R3Davit[5]

### 三维重建相关数据集

## 本章小结

# 基于双注意力的体素重建算法研究

## 引言

## 算法设计

### 网络结构

### 损失函数

### 数据集

### 评价指标

### 实现细节

## 实验结果与分析

### 无背景的合成数据集

### 具有随机背景的合成数据集

### 具有复杂背景的真实数据集

### 消融实验

## 本章小结

# 单视图三维重建软件系统设计与实现

## 引言

## 需求分析

### 功能性需求

### 非功能性需求

### 设计约束

## 系统设计

### 系统概要设计

### 系统总体框架设计

### 系统流程设计

## 系统实现

### 系统开发环境

### 关键技术实现

## 系统界面展示

### 主界面

### 针对合成图像的体素重建系统

### 针对真实图像的体素重建系统

## 系统测试

### 界面测试

### 功能测试

### 性能测试

## 本章小结

# 总结与展望

## 全文工作总结

## 展望

# 参考文献

1. Schonberger J L, Frahm J M. Structure-from-motion revisited[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 4104-4113.
2. Seitz S M, Curless B, Diebel J, et al. A comparison and evaluation of multi-view stereo reconstruction algorithms[C]//2006 IEEE computer society conference on computer vision and pattern recognition (CVPR'06). IEEE, 2006, 1: 519-528.
3. Chang A X, Funkhouser T, Guibas L, et al. Shapenet: An information-rich 3d model repository[J]. arXiv preprint arXiv:1512.03012, 2015.
4. Sun X, Wu J, Zhang X, et al. Pix3d: Dataset and methods for single-image 3d shape modeling[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2018: 2974-2983.
5. Li C, Xiao M, Li Z, et al. Single-view 3D reconstruction via dual attention[J]. PeerJ Computer Science, 2024, 10: e2403.
6. Xiong W, Huang F, Zhang H, et al. 3D voxel reconstruction from single-view image based on cross-domain feature fusion[J]. Expert Systems with Applications, 2024, 256: 124957.
7. Gao J, Kong D, Wang S, et al. DASI: Learning domain adaptive shape impression for 3D object reconstruction[J]. IEEE Transactions on Multimedia, 2022, 25: 5248-5262.
8. Zhu Z, Yang L, Li N, et al. Umiformer: Mining the correlations between similar tokens for multi-view 3d reconstruction[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2023: 18226-18235.
9. Yang L, Zhu Z, Lin X, et al. Long-Range Grouping Transformer for Multi-View 3D Reconstruction[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2023: 18257-18267.
10. 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.
11. 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.
12. Peng K, Islam R, Quarles J, et al. Tmvnet: Using transformers for multi-view voxel-based 3d reconstruction[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2022: 222-230.
13. Arshad M S, Beksi W J. LIST: Learning Implicitly from Spatial Transformers for Single-View 3D Reconstruction[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2023: 9321-9330.
14. Yagubbayli F, Wang Y, Tonioni A, et al. Legoformer: Transformers for block-by-block multi-view 3d reconstruction[J]. arXiv preprint arXiv:2106.12102, 2021.
15. Wang D, Cui X, Chen X, et al. Multi-view 3d reconstruction with transformers[C]//Proceedings of the IEEE/CVF international conference on computer vision. 2021: 5722-5731.
16. 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.
17. 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.
18. 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.
19. Yang S, Xu M, Xie H, et al. Single-view 3d object reconstruction from shape priors in memory[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021: 3152-3161.
20. Yang B, Wang S, Markham A, et al. Robust attentional aggregation of deep feature sets for multi-view 3D reconstruction[J]. International Journal of Computer Vision, 2020, 128(1): 53-73.
21. 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.
22. Kar A, Häne C, Malik J. Learning a multi-view stereo machine[J]. Advances in neural information processing systems, 2017, 30.
23. 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.
24. 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.
25. 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.
26. Chen Z, Zhang H. Learning implicit fields for generative shape modeling[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2019: 5939-5948.
27. Goueix T, Fisher M, Kim V G, et al. Atlasnet: A papier-mch approach to learning 3d surface generation[C]. CVPR, 2018.
28. Huang Z, Wen Y, Wang Z, et al. Surface reconstruction from point clouds: A survey and a benchmark[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2024.
29. Yu Q, Yang C, Wei H. Part-wise AtlasNet for 3D point cloud reconstruction from a single image[J]. Knowledge-Based Systems, 2022, 242: 108395.
30. Melas-Kyriazi L, Rupprecht C, Vedaldi A. Pc2: Projection-conditioned point cloud diffusion for single-image 3d reconstruction[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2023: 12923-12932.
31. Li Y, Zhao Z, Fan J, et al. ADR-MVSNet: A cascade network for 3D point cloud reconstruction with pixel occlusion[J]. Pattern recognition, 2022, 125: 108516.
32. Huang J, Stoter J, Peters R, et al. City3D: Large-scale building reconstruction from airborne LiDAR point clouds[J]. Remote Sensing, 2022, 14(9): 2254.
33. Lin S, Xiao D, Shi Z, et al. Surface reconstruction from point clouds without normals by parametrizing the gauss formula[J]. ACM Transactions on Graphics, 2022, 42(2): 1-19.
34. Li J, Ma W, Li Q, et al. Multi-view real-time acquisition and 3D reconstruction of point clouds for beef cattle[J]. Computers and Electronics in Agriculture, 2022, 197: 106987.
35. Shen T, Gao J, Yin K, et al. Deep marching tetrahedra: a hybrid representation for high-resolution 3d shape synthesis[J]. Advances in Neural Information Processing Systems, 2021, 34: 6087-6101.
36. Hasselgren J, Hofmann N, Munkberg J. Shape, light, and material decomposition from images using monte carlo rendering and denoising[J]. Advances in Neural Information Processing Systems, 2022, 35: 22856-22869.
37. Mildenhall B, Srinivasan P P, Tancik M, et al. Nerf: Representing scenes as neural radiance fields for view synthesis[J]. Communications of the ACM, 2021, 65(1): 99-106.
38. Kerbl B, Kopanas G, Leimkühler T, et al. 3D Gaussian splatting for real-time radiance field rendering[J]. ACM Trans. Graph., 2023, 42(4): 139:1-139:14.
39. 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.
40. Graves A, Graves A. Long short-term memory[J]. Supervised sequence labelling with recurrent neural networks, 2012: 37-45.
41. Alexey D. An image is worth 16x16 words: Transformers for image recognition at scale[J]. arXiv preprint arXiv: 2010.11929, 2020.
42. Vaswani A. Attention is all you need[J]. Advances in Neural Information Processing Systems, 2017.
43. Krizhevsky A, Sutskever I, Hinton G E. Imagenet classification with deep convolutional neural networks[J]. Advances in neural information processing systems, 2012, 25.
44. Ding M, Xiao B, Codella N, et al. Davit: Dual attention vision transformers[C]//European conference on computer vision. Cham: Springer Nature Switzerland, 2022: 74-92.
45. Xiao J, Hays J, Ehinger K A, et al. Sun database: Large-scale scene recognition from abbey to zoo[C]//2010 IEEE computer society conference on computer vision and pattern recognition. IEEE, 2010: 3485-3492.
46. Gu A, Dao T. Mamba: Linear-time sequence modeling with selective state spaces[J]. arXiv preprint arXiv:2312.00752, 2023.
47. LeCun Y, Bengio Y. Convolutional networks for images, speech, and time series[J]. The handbook of brain theory and neural networks, 1995, 3361(10): 1995.
48. Medsker L R, Jain L. Recurrent neural networks[J]. Design and Applications, 2001, 5(64-67): 2.
49. Touvron H, Cord M, Douze M, et al. Training data-efficient image transformers & distillation through attention[C]//International conference on machine learning. PMLR, 2021: 10347-10357.
50. Gu A, Dao T. Mamba: Linear-time sequence modeling with selective state spaces[J]. arXiv preprint arXiv:2312.00752, 2023.
51. Gu A, Goel K, Ré C. Efficiently modeling long sequences with structured state spaces[J]. arXiv preprint arXiv:2111.00396, 2021.

# 个人简介 在读期间发表学术论文

**个人简历：**

**李承欢**，男，2000年3月生。

2022年7月毕业于江西工程学院人工智能与数据科学学院软件工程专业，获学士学位。

2022年9月入华东交通大学读硕士研究生。

**已发表论文：**

1. **Li C**, Xiao M, Li Z, et al. Single-view 3D reconstruction via dual attention[J]. PeerJ Computer Science, 2024, 10: e2403. doi: 10.7717/peerj-cs.2403

**发明专利：**

1. ZL 2023 1 1084904.2发明名称：一种三维重建方法、装置、系统以及存储介质;发明人：肖美华，**李承欢**，李泽寰，徐锐涵

**参与项目：**

[1] 基于事件逻辑理论的安全协议实施安全性形式化分析与验证，国家自然科学基金(62362033)

[2] 基于事件逻辑的云计算环境下数据完整性验证模型及协议研究，国家自然科学基金(61962020)

[3] 基于事件逻辑理论的安全协议实施安全性形式化分析若干关键技术研究, 江西省“双千”人才项目(JXSQ2023201009)

# 致谢

光阴似箭，日月如梭，转瞬之间，三年研究生生涯即将结束。回忆在校园度过的每个日夜，感慨良多，收获丰厚。三年的读研生活，有成功的喜悦、有精彩的故事、也有失败、挫折的痛苦，无论如何，这将促使我继续前行。在此感谢读研期间老师对我的谆谆教诲，家人默默无闻的支持以及师兄弟之间的互相帮助。

首先，我要感谢我的导师肖美华教授，在肖老师的悉心栽培和指导下，才有了今天的我。在学习上，他教会我如何做科研、撰写论文、做演讲，增强我的学习能力；在生活上，他更加关心我们的日常生活，也会让我们多运动锻炼身体，同时注重对我们的道德品行上的教育，注重德智体全面发展；当我在生活和研究中遇到困惑时，他会化身为慈爱的长者，耐心地指导我、帮助我。肖老师高尚的人格、宽广的胸怀使我受益终生。他不仅教会了我专业知识和科研技能，更教会了我人生的学问和做人的道理。

感谢父母和家人们。你们呵护我的成长，无私奉献、对我毫无保留的支持，鼓励我追求更高的学术理想。

感谢305、307实验室的各位师兄师姐以及师弟师妹们，在实验室里的生活有了你们的帮助和支持，使得我在研究生阶段能够过得充实并且丰富多彩，也希望我们在未来的日子里能更上一层楼！

最后，向参加论文评审、答辩以及提出宝贵意见的各位老师、专家们表示衷心的感谢，愿您们工作顺利、幸福安康！