

<b>Preface</b>	<b>5</b>
<b>I</b>	<b>6</b>
<b>1</b>	<b>7</b>
1.1 从二维到三维：构建计算机的空间感知体系 . . . . .	7
1.2 破解维度诅咒：计算机三维感知的核心挑战 . . . . .	7
1.3 智能时代的空间革命：三维视觉的广阔应用前景 . . . . .	8
<b>2</b>	<b>9</b>
2.1 引言：为什么需要相机标定？ . . . . .	9
2.2 核心概念 . . . . .	9
2.3 理论基础：分步推导 . . . . .	11
2.4 算法实现 . . . . .	12
2.5 标定精度评估 . . . . .	14
2.6 小结 . . . . .	16
<b>3</b>	<b>17</b>
3.1 引言：从双目视觉到深度感知 . . . . .	17
3.2 核心概念 . . . . .	17
3.3 理论基础：从几何约束到深度学习 . . . . .	19
3.3.1 传统几何方法的理论基础 . . . . .	19
3.3.2 深度学习方法的理论基础 . . . . .	21
3.4 算法实现 . . . . .	22
3.5 性能对比分析 . . . . .	26
3.6 小结 . . . . .	28
<b>4</b>	<b>29</b>
4.1 引言：从图像到三维世界的重建 . . . . .	29
4.2 核心概念 . . . . .	29
4.3 理论基础：从多视图几何到神经隐式表示 . . . . .	31
4.3.1 运动恢复结构（SfM）的理论基础 . . . . .	32
4.3.2 TSDF融合的理论基础 . . . . .	33
4.3.3 神经辐射场（NeRF）的理论基础 . . . . .	34

4.4 算法实现 . . . . .	34
4.4.1 SfM的核心算法实现 . . . . .	35
4.4.2 TSDF融合的核心算法实现 . . . . .	36
4.4.3 NeRF的核心算法实现 . . . . .	38
4.5 重建质量评估 . . . . .	41
4.5.1 方法性能对比 . . . . .	42
4.5.2 应用场景适应性 . . . . .	44
4.5.3 技术发展趋势 . . . . .	45
4.6 小结 . . . . .	45
<b>5</b>	<b>47</b>
5.1 引言：点云数据的重要性与挑战 . . . . .	47
5.2 核心概念 . . . . .	47
5.3 理论基础：空间数据结构与算法 . . . . .	49
5.3.1 KD-Tree的理论基础 . . . . .	50
5.3.2 体素化的理论基础 . . . . .	50
5.3.3 聚类算法的理论基础 . . . . .	51
5.3.4 统计滤波的理论基础 . . . . .	52
5.4 算法实现 . . . . .	53
5.4.1 KD-Tree的核心实现 . . . . .	53
5.4.2 体素化处理的核心实现 . . . . .	55
5.4.3 点云滤波的核心实现 . . . . .	56
5.4.4 点云聚类的核心实现 . . . . .	57
5.5 处理效率分析 . . . . .	59
5.5.1 空间索引结构性能对比 . . . . .	59
5.5.2 滤波算法效果分析 . . . . .	60
5.5.3 聚类算法适应性分析 . . . . .	61
5.5.4 处理流程优化策略 . . . . .	62
5.6 小结 . . . . .	63
<b>6 PointNet</b>	<b>64</b>
6.1 引言：深度学习在点云处理中的革命性突破 . . . . .	64
6.2 核心概念 . . . . .	64
6.2.1 PointNet的核心思想深度解析 . . . . .	66
6.3 理论基础：从对称函数到自注意力机制 . . . . .	69
6.3.1 PointNet的理论基础 . . . . .	69
6.3.2 PointNet++的理论基础 . . . . .	70
6.3.3 Point-Transformer的理论基础 . . . . .	70
6.3.4 损失函数设计 . . . . .	71
6.4 算法实现 . . . . .	72
6.4.1 PointNet的核心实现 . . . . .	72
6.4.2 PointNet++的核心实现 . . . . .	73
6.4.3 Point-Transformer的核心实现 . . . . .	76

6.5	网络性能评估 . . . . .	78
6.5.1	网络性能对比分析 . . . . .	79
6.5.2	网络架构演进分析 . . . . .	80
6.5.3	数据集性能基准测试 . . . . .	81
6.5.4	应用场景适应性分析 . . . . .	82
6.6	小结 . . . . .	82
<b>7</b>	<b>3D</b>	<b>83</b>
7.1	引言：从2D到3D的检测范式转变 . . . . .	83
7.2	核心概念 . . . . .	83
7.3	理论基础：从体素化到点-体素融合 . . . . .	86
7.3.1	VoxelNet的核心思想与理论基础 . . . . .	86
7.3.2	VoxelNet的理论基础 . . . . .	88
7.3.3	PointPillars的理论基础 . . . . .	89
7.3.4	PV-RCNN的理论基础 . . . . .	89
7.3.5	损失函数设计 . . . . .	90
7.4	算法实现 . . . . .	91
7.4.1	VoxelNet的核心实现 . . . . .	91
7.4.2	PointPillars的核心实现 . . . . .	94
7.4.3	PV-RCNN的核心实现 . . . . .	97
7.5	检测效果分析 . . . . .	99
7.5.1	算法性能对比分析 . . . . .	100
7.5.2	技术演进与创新点分析 . . . . .	101
7.5.3	应用场景与挑战分析 . . . . .	102
7.5.4	未来发展趋势 . . . . .	103
7.6	小结 . . . . .	103
<b>8</b>		<b>105</b>
8.1	引言：从理论到实践的技术集成 . . . . .	105
8.2	核心概念 . . . . .	105
8.3	理论基础：系统集成与优化理论 . . . . .	107
8.3.1	实时系统理论 . . . . .	107
8.3.2	多传感器融合理论 . . . . .	108
8.3.3	系统优化理论 . . . . .	109
8.4	算法实现 . . . . .	109
8.4.1	自动驾驶感知系统 . . . . .	109
8.4.2	机器人导航系统 . . . . .	113
8.4.3	工业质量检测系统 . . . . .	116
8.5	应用效果评估 . . . . .	118
8.5.1	应用系统性能对比 . . . . .	118
8.5.2	技术集成效果分析 . . . . .	120
8.5.3	部署成本与效益分析 . . . . .	121
8.5.4	未来发展趋势与挑战 . . . . .	122

8.6 小结 . . . . .	122
<b>References</b>	<b>124</b>

# Preface

This is a Quarto book.

To learn more about Quarto books visit <https://quarto.org/docs/books>.

# **Part I**

# 1

## 1.1

三维视觉技术是计算机视觉领域的重要分支，旨在让计算机理解和重建真实世界的三维结构。经过几十年的发展，该领域已形成了完整的技术体系：

**传统几何方法**构成了三维视觉的理论基础。相机标定建立了图像与现实世界的几何关系；立体匹配通过双目视觉恢复深度信息；三维重建则从多个视角的图像中恢复完整的三维场景。这些方法基于严格的几何理论，具有可解释性强、精度高的特点。

**深度学习方法**则代表了该领域的最新发展。点云处理网络如PointNet系列直接处理三维点云数据；3D目标检测网络能够在三维空间中定位和识别物体。这些方法具有强大的特征学习能力，在复杂场景下表现出色。

两类方法并非对立关系，而是相互补充。传统方法提供了坚实的理论基础和几何约束，深度学习方法则提供了强大的特征表达和泛化能力。现代三维视觉系统往往将两者结合，发挥各自优势。

## 1.2

人类视觉系统能够轻松感知三维世界：判断物体的远近、估计空间的大小、理解场景的布局。这种能力如此自然，以至于我们很少意识到其复杂性。然而，对于计算机来说，从二维图像中恢复三维信息是一个极具挑战性的问题。

**深度信息的缺失**是核心挑战。当三维世界投影到二维图像平面时，深度信息不可避免地丢失了。一个像素点可能对应三维空间中的任意一点，这种一对多的映射关系使得深度恢复成为一个病态问题。

**视角变化的复杂性**进一步增加了难度。同一个物体从不同角度观察会呈现完全不同的外观，相机的位置、姿态、内部参数都会影响成像结果。如何建立图像与现实世界之间的准确对应关系，是三维视觉必须解决的基础问题。

**数据表示的多样性**也带来了挑战。三维信息可以用深度图、点云、体素、网格等多种形式表示，每种表示都有其优缺点。如何选择合适的表示方法，如何在不同表示之间转换，都需要深入思考。

## 1.3

三维视觉技术在现代科技中发挥着越来越重要的作用，其应用领域广泛且影响深远。

**自动驾驶**是三维视觉最具挑战性的应用之一。车载传感器需要实时感知周围环境的三维结构：前方车辆的距离、行人的位置、道路的坡度、障碍物的形状。这些信息直接关系到行车安全。现代自动驾驶系统通常融合摄像头、激光雷达、毫米波雷达等多种传感器，构建精确的三维环境地图。特斯拉的纯视觉方案展示了基于摄像头的三维感知能力，而Waymo的激光雷达方案则体现了点云处理的重要性。

**增强现实（AR）和虚拟现实（VR）**技术的核心是虚实融合。AR应用需要准确理解真实场景的三维结构，才能将虚拟物体自然地放置在现实环境中。苹果的ARKit、谷歌的ARCore都大量使用了三维视觉技术。VR应用则需要实时追踪用户的头部和手部姿态，构建沉浸式的三维体验。

**机器人技术**中，三维视觉是实现智能操作的关键。工业机器人需要精确定位零件的位置和姿态；服务机器人需要理解室内环境的布局；手术机器人需要重建人体器官的三维结构。波士顿动力的机器人能够在复杂地形中稳定行走，很大程度上依赖于先进的三维感知能力。

**医疗影像**领域，三维重建技术帮助医生更好地诊断疾病。CT、MRI扫描产生的二维切片可以重建为三维模型，为手术规划提供直观的参考。计算机辅助手术系统能够实时追踪手术器械的位置，提高手术精度。

# 2

## 2.1

当我们用手机拍照时，很少思考这样一个问题：照片中的每个像素是如何与现实世界中的物体建立对应关系的？这个看似简单的问题，实际上涉及复杂的几何变换过程。相机标定正是要解决这个基础问题：建立图像坐标与现实世界坐标之间的精确映射关系。

## 2.2

**针孔相机模型**是理解相机成像的基础。想象一个不透光的盒子，前面开一个小孔，后面放一块感光板。外界的光线通过小孔投射到感光板上，形成倒立的图像。这就是最简单的针孔相机模型。

现代数码相机的工作原理与此类似，只是用透镜组替代了小孔，用图像传感器替代了感光板。透镜组的作用是聚焦光线，提高成像质量；图像传感器则将光信号转换为数字信号。

**坐标系变换关系**描述了从三维世界到二维图像的完整过程。这个过程涉及四个坐标系：世界坐标系、相机坐标系、图像坐标系和像素坐标系。理解这些坐标系之间的关系，是掌握相机几何的关键。

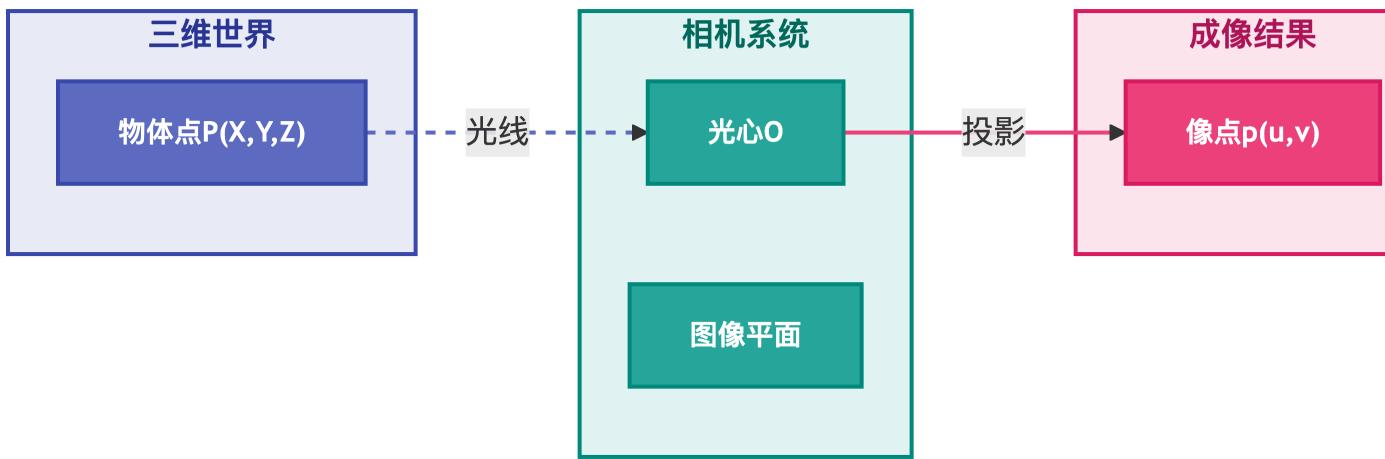


图11.1: 针孔相机模型展示了光线通过光心投射到图像平面的几何关系

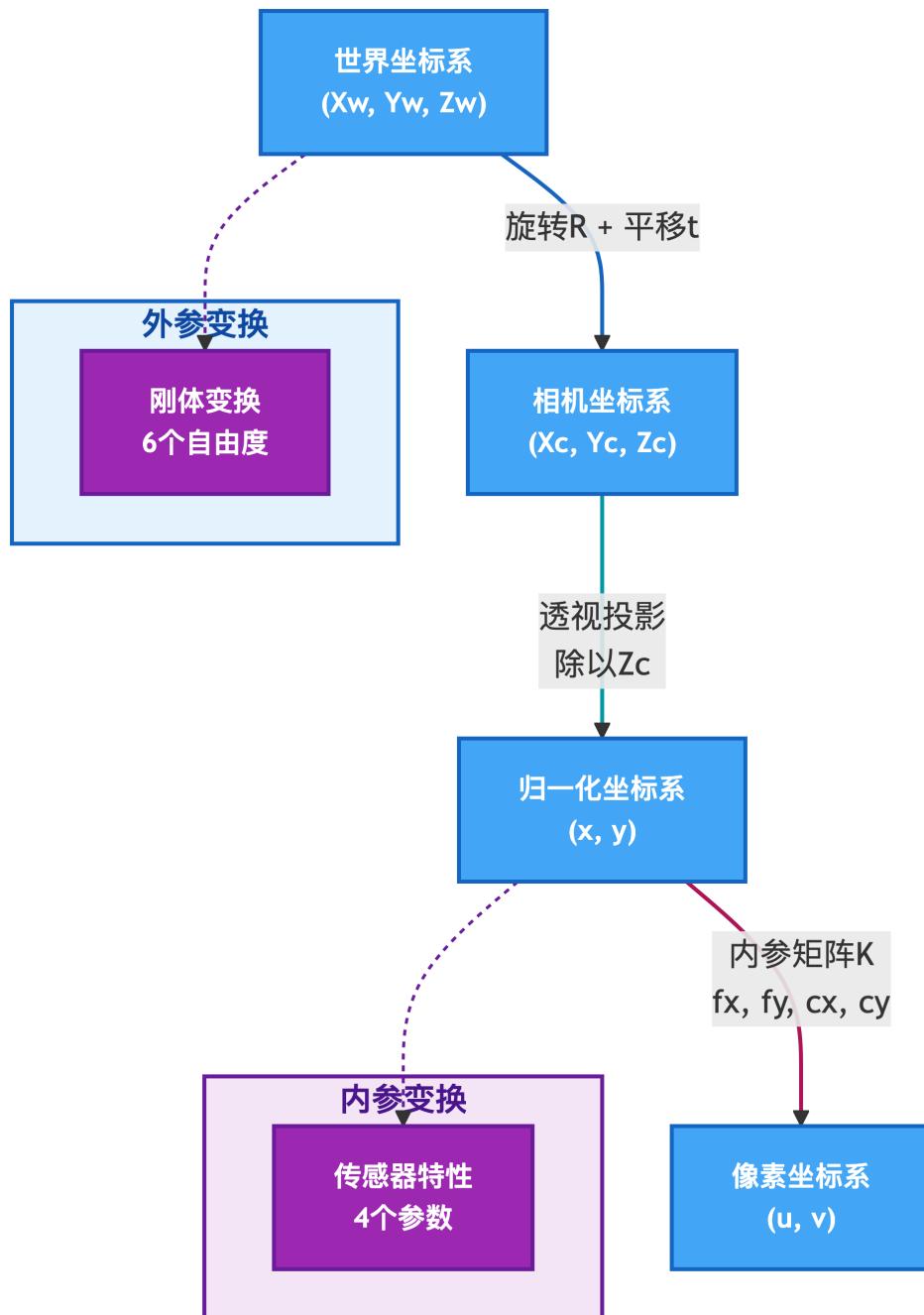


图11.2: 从世界坐标系到像素坐标系的完整变换链

## 2.3

相机投影变换可以分解为三个连续的步骤，每一步都有明确的几何意义。

### 步骤1：世界坐标到相机坐标

世界坐标系是我们建立的参考坐标系，通常选择场景中的某个固定点作为原点。相机坐标系则以相机光心为原点，光轴为Z轴。从世界坐标到相机坐标的变换包括旋转和平移：

$$\begin{bmatrix} X_c \\ Y_c \\ Z_c \end{bmatrix} = \begin{bmatrix} r_{11} & r_{12} & r_{13} \\ r_{21} & r_{22} & r_{23} \\ r_{31} & r_{32} & r_{33} \end{bmatrix} \begin{bmatrix} X_w \\ Y_w \\ Z_w \end{bmatrix} + \begin{bmatrix} t_x \\ t_y \\ t_z \end{bmatrix}$$

其中， $\mathbf{R} = [r_{ij}]$ 是 $3 \times 3$ 旋转矩阵， $\mathbf{t} = [t_x, t_y, t_z]^T$ 是平移向量。旋转矩阵描述了相机的姿态，平移向量描述了相机的位置。这一步消除了相机位置和姿态对成像的影响，将所有点都表示在相机坐标系中。

### 步骤2：相机坐标到归一化坐标

归一化坐标系是一个虚拟的坐标系，位于距离光心单位距离的平面上。从相机坐标到归一化的变换实现了透视投影：

$$x = \frac{X_c}{Z_c}, \quad y = \frac{Y_c}{Z_c}$$

这一步体现了透视投影的核心特征：远处的物体看起来更小。深度信息 $Z_c$ 在这一步丢失了，这正是从三维到二维投影的本质。

### 步骤3：归一化坐标到像素坐标

最后一步考虑了图像传感器的物理特性，将归一化坐标转换为像素坐标：

$$u = f_x \cdot x + c_x, \quad v = f_y \cdot y + c_y$$

其中， $f_x$ 和 $f_y$ 是焦距在x和y方向的像素表示， $c_x$ 和 $c_y$ 是主点坐标。这四个参数构成了相机的内参矩阵 $\mathbf{K}$ ：

$$\mathbf{K} = \begin{bmatrix} f_x & 0 & c_x \\ 0 & f_y & c_y \\ 0 & 0 & 1 \end{bmatrix}$$

## 2.4

相机标定的核心是求解内参矩阵K和外参矩阵[R|t]。最常用的方法是基于棋盘格标定板的线性方法。

```
def camera_calibration_core(object_points, image_points):
    """
    # 1.
    A = []
    for (X, Y, Z), (u, v) in zip(object_points, image_points):
        u = (p11*X + p12*Y + p13*Z + p14) / (p31*X + p32*Y + p33*Z + 1)
        A.append([X, Y, Z, 1, 0, 0, 0, -u*X, -u*Y, -u*Z, -u])
        A.append([0, 0, 0, 0, X, Y, Z, 1, -v*X, -v*Y, -v*Z, -v])

    # 2. SVD      Ah = 0
    U, S, Vt = np.linalg.svd(np.array(A))
    h = Vt[-1, :]  #

    # 3.
    P = h.reshape(3, 4)
    K, R, t = decompose_projection_matrix(P)

    return K, R, t
```

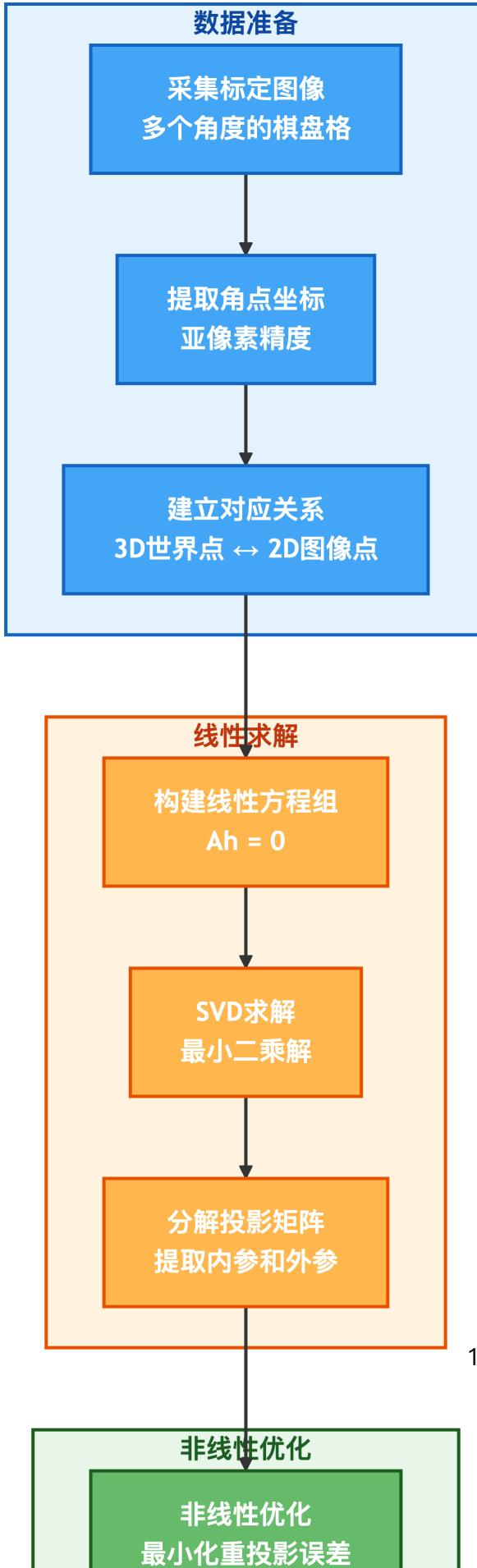


图11.3：相机标定算法的主要步骤和数据流

## 2.5

相机标定的效果可以通过多个指标来评估。重投影误差是最直观的评价标准，它衡量了标定结果的几何精度。

**畸变校正效果：**未标定的图像往往存在明显的畸变，特别是广角镜头拍摄的图像。标定后可以有效校正这些畸变，恢复图像的真实几何关系。

**重投影误差分析：**优秀的标定结果应该具有较小且均匀分布的重投影误差。如果误差过大或分布不均，说明标定质量有问题，需要重新采集数据或调整标定方法。

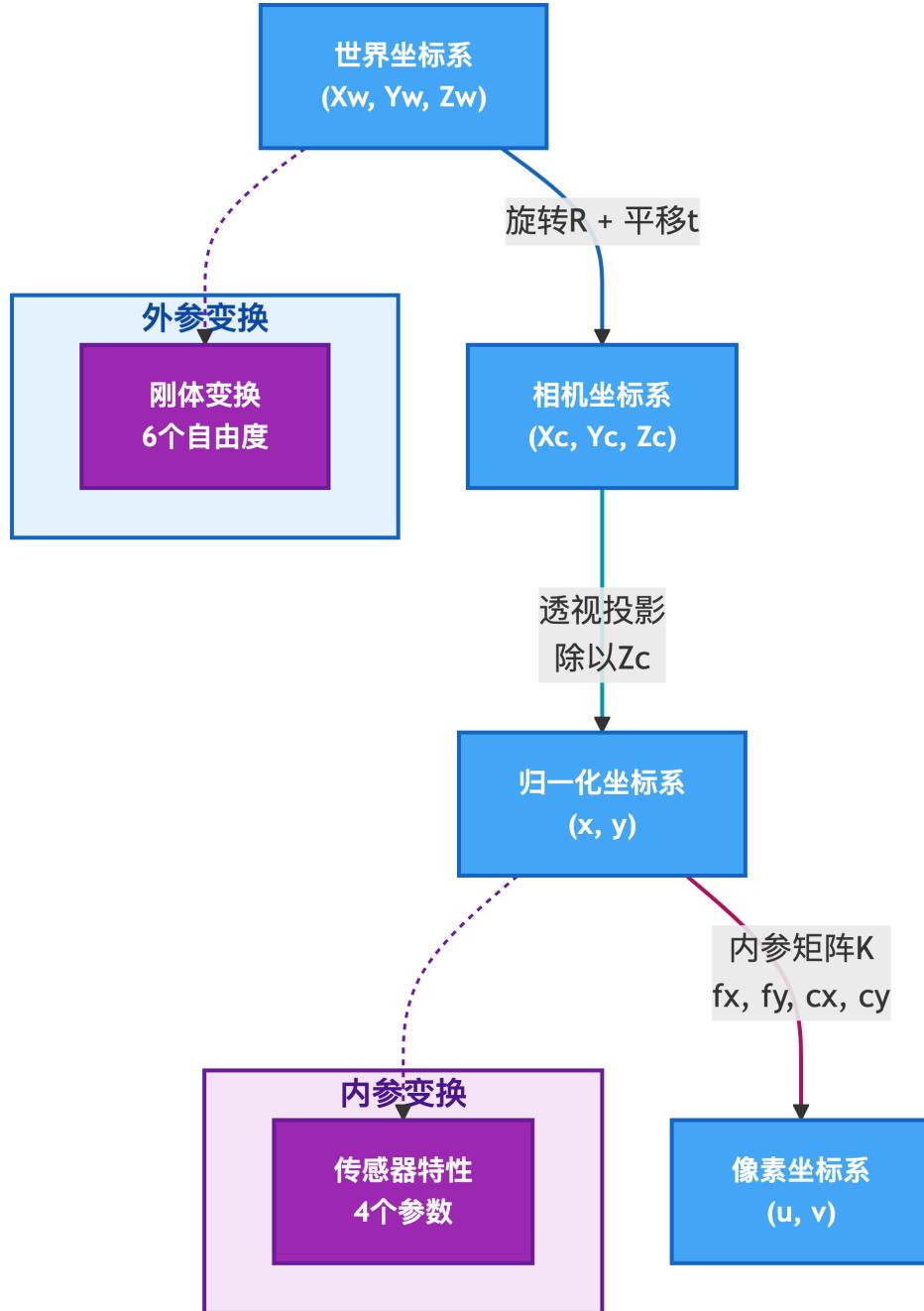


图11.4：镜头畸变校正前后的效果对比，注意图像边缘的几何变化

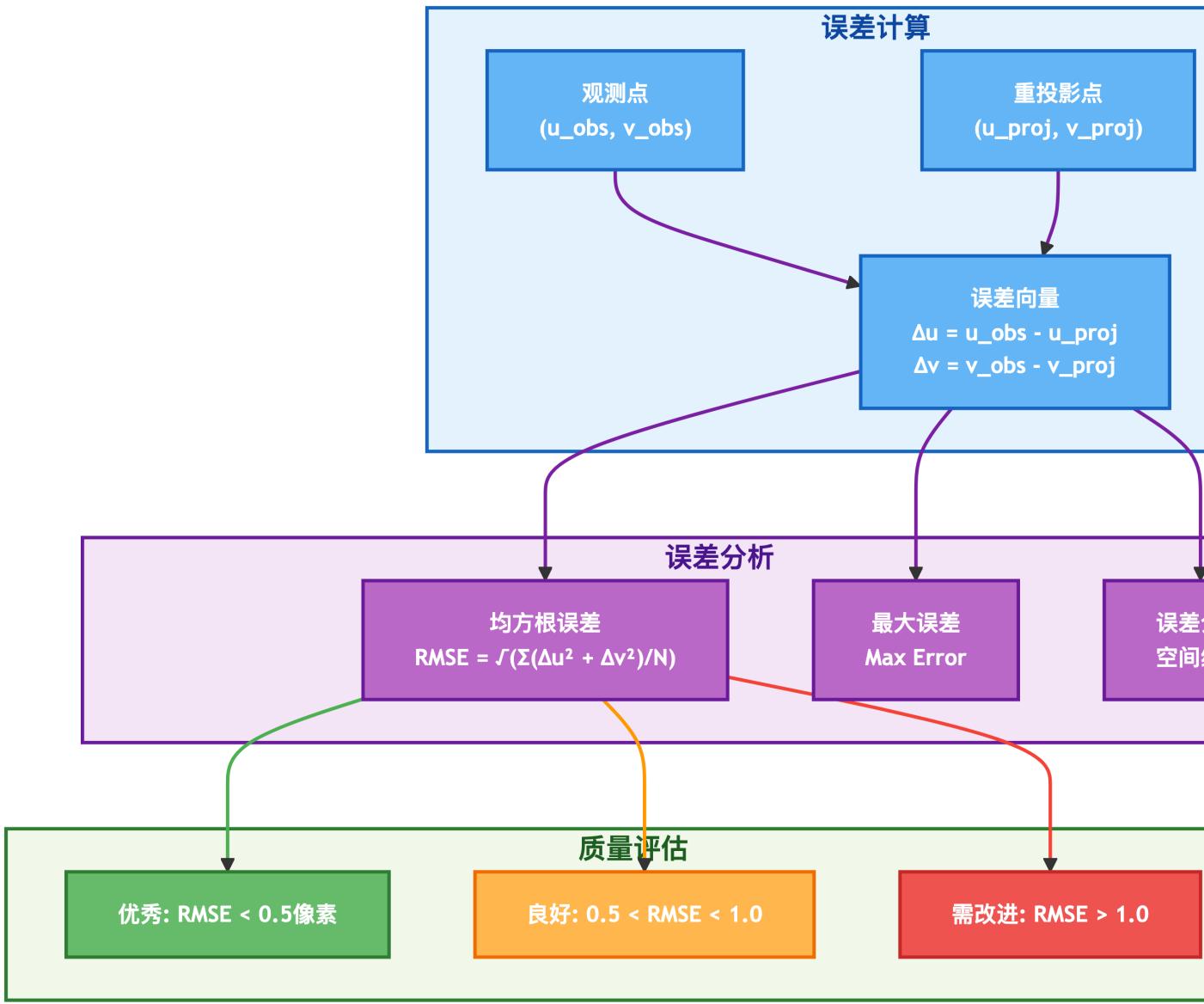


图11.5：重投影误差的空间分布和质量评估标准

## 2.6

相机标定是三维视觉的基础，它建立了图像与现实世界之间的几何桥梁。通过理解针孔相机模型和坐标变换关系，我们可以准确地从二维图像中提取三维信息。标定质量直接影响后续所有三维视觉算法的精度，因此必须给予足够重视。

# 3

## 3.1

当我们用双眼观察世界时，左右眼看到的图像存在微小差异。大脑正是利用这种差异来感知深度，判断物体的远近。立体匹配算法正是模拟了人类双目视觉的这一机制：通过计算两幅图像中对应点的视差（位置差异），恢复场景的深度信息。

随着深度学习的发展，深度估计技术已经从传统的几何方法扩展到基于神经网络的端到端学习方法。现代深度估计系统不仅能处理标准的双目图像对，还能从单目图像直接预测深度，在自动驾驶、机器人导航、增强现实等领域发挥着关键作用。

## 3.2

**传统立体匹配**基于几何约束和相似性度量。双目立体视觉系统通常由两个平行放置的相机组成，相机之间的距离称为基线（baseline）。传统方法通过在极线约束下搜索对应点，计算视差来恢复深度。这类方法计算效率高，但在弱纹理、遮挡区域容易失效。

**深度学习方法**则将深度估计视为回归问题，通过端到端训练学习从图像到深度的映射关系。现代深度网络如PSMNet能够处理复杂场景，在准确性和鲁棒性方面显著超越传统方法。这类方法能够利用语义信息和全局上下文，在困难区域也能给出合理的深度估计。

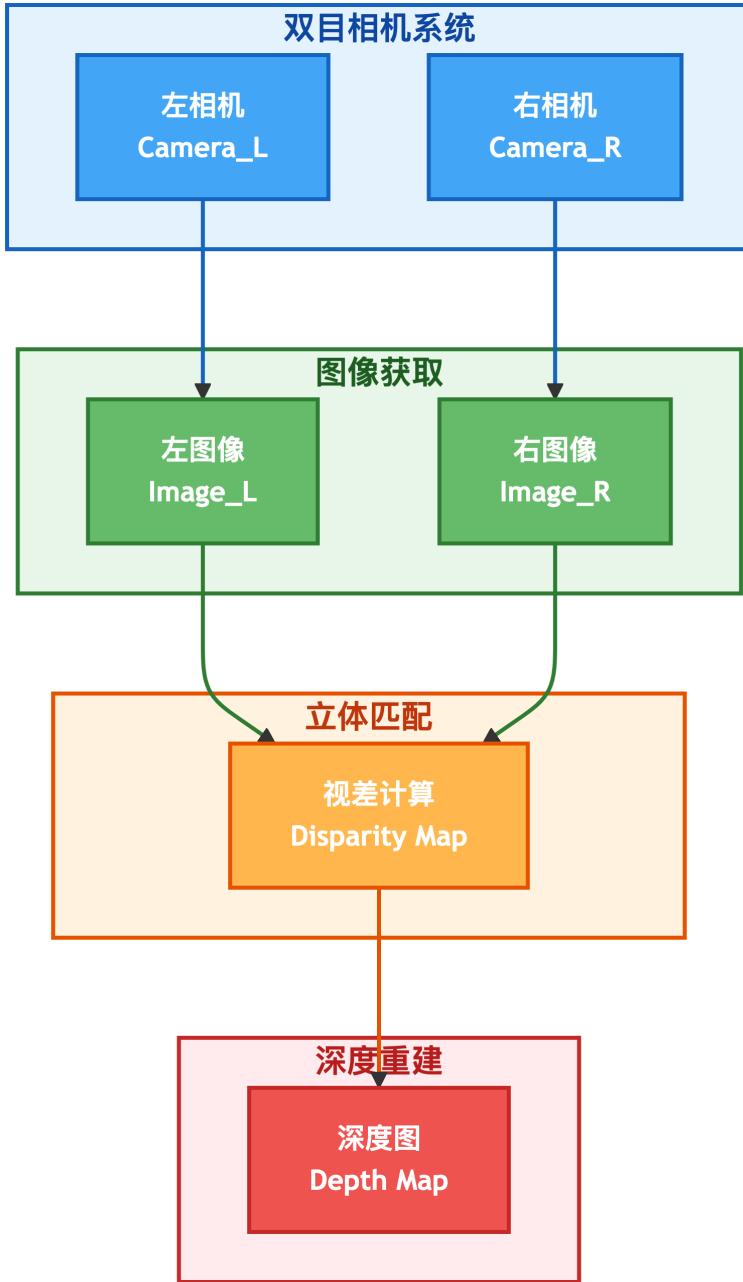


图11.6：双目立体视觉系统的基本工作流程

**视差 (Disparity)** 是立体匹配的核心概念。它指的是同一物体在左右图像中对应点的水平位置差异。视差与深度成反比关系：距离相机越近的物体，其视差越大；距离相机越远的物体，其视差越小。无穷远处的物体（如天空）视差接近于零。

对应点问题是立体匹配的核心挑战。给定左图中的一个点，如何在右图中找到与之对应的

点？这个看似简单的问题实际上非常复杂，尤其是在纹理缺乏、重复模式、遮挡区域等情况下。立体匹配算法的主要差异就在于如何解决这个对应点问题。

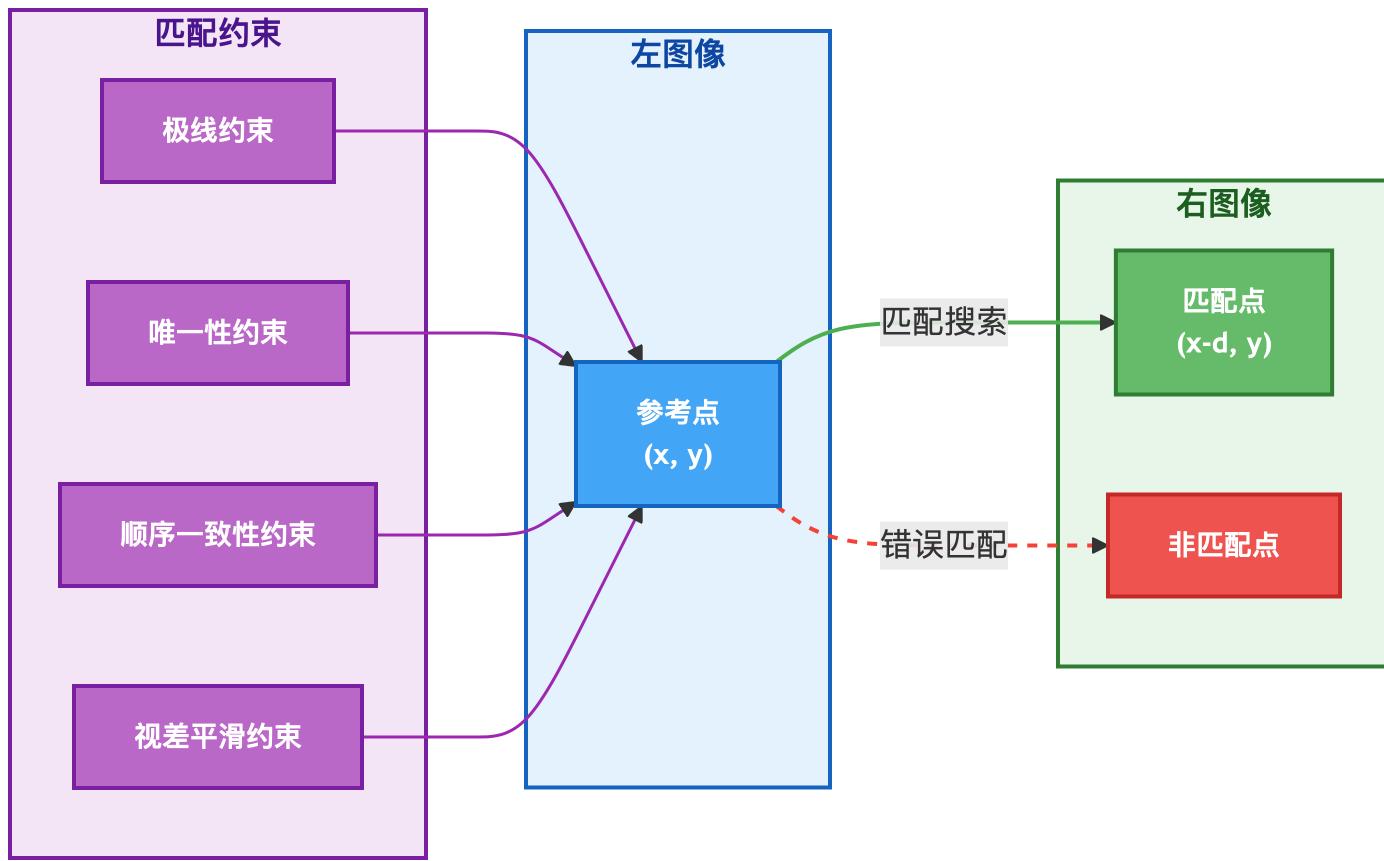


图11.7：立体匹配中的对应点问题与匹配约束

### 3.3

立体匹配与深度估计的理论基础可以分为传统几何方法和现代深度学习方法两大类。

#### 3.3.1

##### 1. 立体相机几何关系

在标准立体配置中，两个相机的光轴平行，图像平面共面。设左右相机的光心分别为 $O_L$ 和 $O_R$ ，它们之间的距离（基线长度）为 $b$ 。对于空间中的点 $P(X, Y, Z)$ ，其在左右图像中的投影点分别为 $p_L(x_L, y_L)$ 和 $p_R(x_R, y_R)$ 。根据相似三角形原理：

$$\frac{x_L - x_R}{b} = \frac{f}{Z}$$

定义视差 $d = x_L - x_R$ ，则深度与视差成反比关系：

$$Z = \frac{f \cdot b}{d}$$

## 2. 传统视差计算方法

传统方法主要基于匹配代价计算和优化：

- 局部方法：使用窗口匹配，计算相似性度量如SAD、SSD或NCC：

$$\text{SAD}(x, y, d) = \sum_{(i,j) \in W} |I_L(i, j) - I_R(i - d, j)|$$

- 全局方法：将视差计算视为能量最小化问题：

$$E(D) = E_{data}(D) + \lambda \cdot E_{smooth}(D)$$

- 半全局方法(SGM)：通过多方向路径聚合匹配代价，平衡局部和全局信息：

$$L_r(p, d) = C(p, d) + \min \begin{cases} L_r(p - r, d) \\ L_r(p - r, d - 1) + P_1 \\ L_r(p - r, d + 1) + P_1 \\ \min_i L_r(p - r, i) + P_2 \end{cases}$$

其中 $L_r(p, d)$ 是沿方向 $r$ 的路径代价， $C(p, d)$ 是像素 $p$ 处视差为 $d$ 的匹配代价， $P_1$ 和 $P_2$ 是平滑性惩罚参数。

### 3.3.2

#### 1. 端到端深度估计框架

深度学习方法将立体匹配视为一个端到端的回归问题，网络架构通常包含四个关键组件：

- **特征提取**: 使用CNN提取左右图像的特征表示
- **代价体积构建**: 通过特征匹配或拼接构建4D代价体积
- **代价聚合**: 使用3D CNN或GNN进行代价聚合
- **视差回归**: 通过软argmin操作回归连续视差值

#### 2. PSMNet的核心理论

PSMNet是深度学习立体匹配的代表性网络，其核心理论包括：

- **空间金字塔池化(SPP)**: 捕获多尺度上下文信息：

$$F_{SPP}(x) = \text{Concat}[F(x), P_1(F(x)), P_2(F(x)), \dots, P_n(F(x))]$$

其中 $P_i$ 表示不同尺度的池化操作。

- **3D代价体积滤波**: 使用3D CNN进行代价聚合：

$$C_{out} = \text{3DCNN}(C_{in})$$

- **视差回归**: 通过软argmin操作实现亚像素精度：

$$\hat{d} = \sum_{d=0}^{D_{max}} d \cdot \sigma(-C_d)$$

其中 $\sigma$ 是softmax函数， $C_d$ 是代价体积中视差为 $d$ 的代价值。

#### 3. 单目深度估计理论

单目深度估计直接从单张图像预测深度，其理论基础是：

- **编码器-解码器架构**: 通过多尺度特征提取和逐步上采样恢复分辨率
- **深度回归**: 直接回归深度值或视差值
- **自监督学习**: 利用时序一致性或立体一致性作为监督信号：

$$L_{photo} = \alpha \frac{1 - \text{SSIM}(I, \hat{I})}{2} + (1 - \alpha) \|I - \hat{I}\|_1$$

其中  $\hat{I}$  是通过预测的深度图和相机位姿重投影得到的图像。

这些理论方法的核心区别在于：传统方法依赖手工设计的特征和几何约束，而深度学习方法能够自动学习特征表示和匹配策略，特别是在复杂场景中表现出更强的鲁棒性。

### 3.4

立体匹配与深度估计的实现可以分为传统几何方法和现代深度学习方法两大类。

**传统SGBM算法核心：**

```
import cv2
import numpy as np

def sgbm_stereo_matching(left_img, right_img):
    """
    SGBM

    """
    #
    stereo = cv2.StereoSGBM_create(
        minDisparity=0,
        numDisparities=64,           #
        blockSize=5,                 #
        P1=8 * 3 * 5**2,            #      1
        P2=32 * 3 * 5**2,           #      2
        uniquenessRatio=10,          #
        speckleWindowSize=100,        #
        speckleRange=32              #
    )

    #
    disparity = stereo.compute(left_img, right_img)
    return disparity.astype(np.float32) / 16.0  #

def disparity_to_depth(disparity, focal_length, baseline):
```

```
    """
    """
    return (focal_length * baseline) / (disparity + 1e-6)
```

现代PSMNet深度网络：

```
import torch
import torch.nn as nn
import torch.nn.functional as F

class PSMNet(nn.Module):
    """
    PSMNet
    4D      3D CNN
    """

    def __init__(self, maxdisp=192):
        super(PSMNet, self).__init__()
        self.maxdisp = maxdisp

        #
        self.feature_extraction = self._make_feature_extractor()

        #
        self.cost_volume_filter = self._make_cost_volume_filter()

        #
        self.disparity_regression = self._make_disparity_regression()

    def forward(self, left, right):
        # 1.
        left_features = self.feature_extraction(left)
        right_features = self.feature_extraction(right)

        # 2.
        cost_volume = self.build_cost_volume(left_features, right_features)

        # 3.
        cost_volume = self.cost_volume_filter(cost_volume)

        # 4.
        disparity = self.disparity_regression(cost_volume)

    return disparity
```

```

def build_cost_volume(self, left_feat, right_feat):
    """ 4D      """
    B, C, H, W = left_feat.shape
    cost_volume = torch.zeros(B, C*2, self.maxdisp//4, H, W)

    for i in range(self.maxdisp//4):
        if i > 0:
            cost_volume[:, :C, i, :, i:] = left_feat[:, :, :, i:]
            cost_volume[:, C:, i, :, i:] = right_feat[:, :, :, :-i]
        else:
            cost_volume[:, :C, i, :, :] = left_feat
            cost_volume[:, C:, i, :, :] = right_feat

    return cost_volume

```

单目深度估计核心:

```

class MonoDepthNet(nn.Module):
    """
    """

    def __init__(self):
        super(MonoDepthNet, self).__init__()
        #
        self.encoder = self._make_encoder()
        #
        self.decoder = self._make_decoder()

    def forward(self, x):
        #
        features = self.encoder(x)
        #
        depth = self.decoder(features)
        return depth

```

这些算法的核心区别在于：SGBM基于几何约束和手工特征，PSMNet通过学习特征和代价聚合，单目方法则完全依赖语义理解。现代方法在复杂场景下表现更佳，但计算成本也更高。

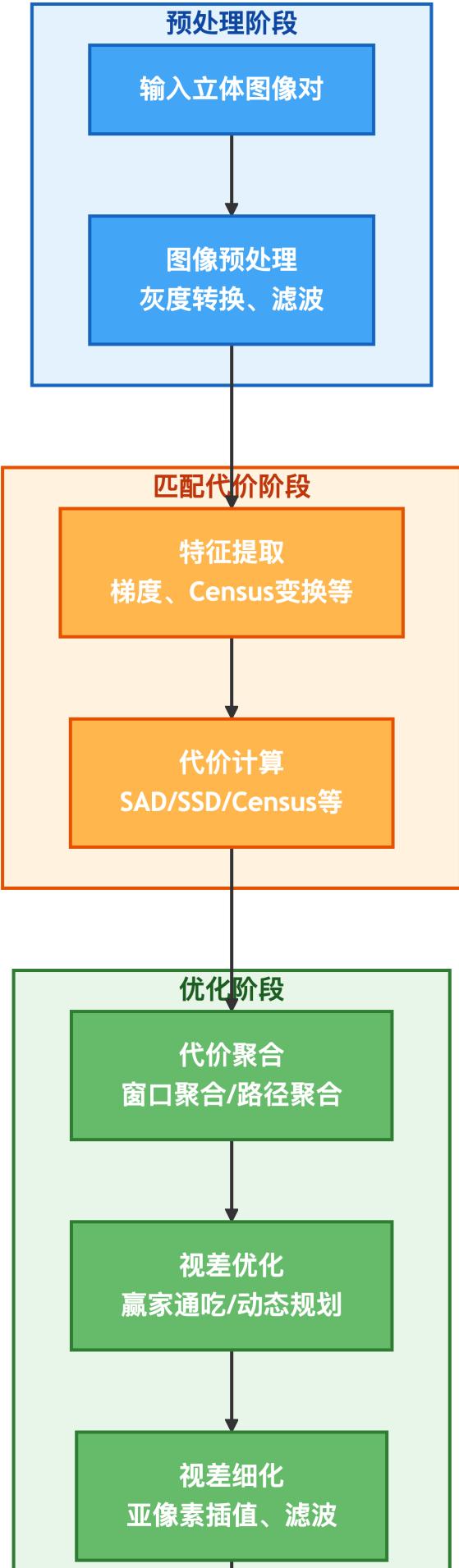


图11.8：立体匹配算法的通用流程

### 3.5

深度估计算法的效果可以通过视差图和深度图的质量来评估。下面我们分析传统方法和深度学习方法在不同场景下的表现。

算法性能对比：

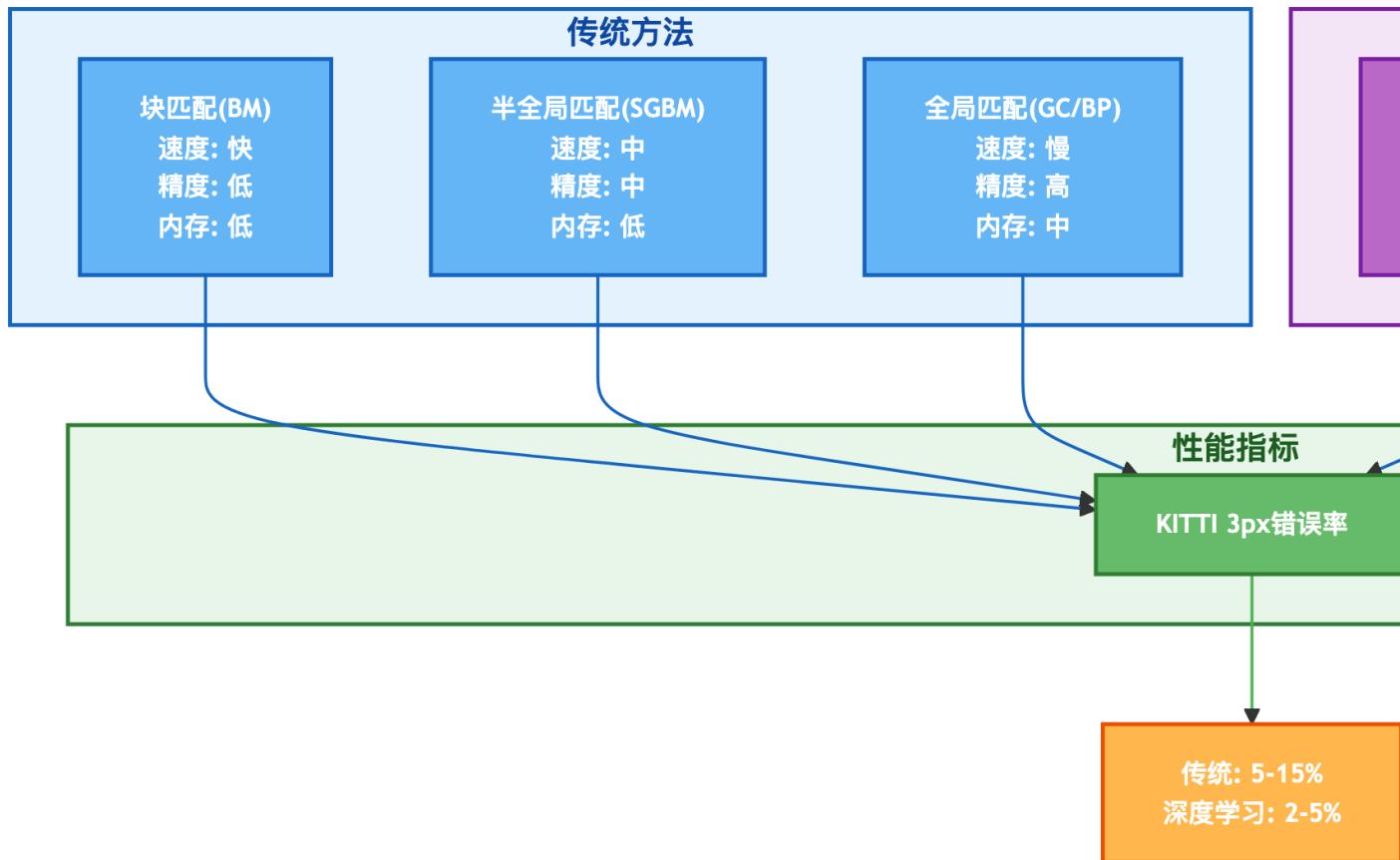


图11.9：传统方法与深度学习方法的性能对比

场景适应性分析：

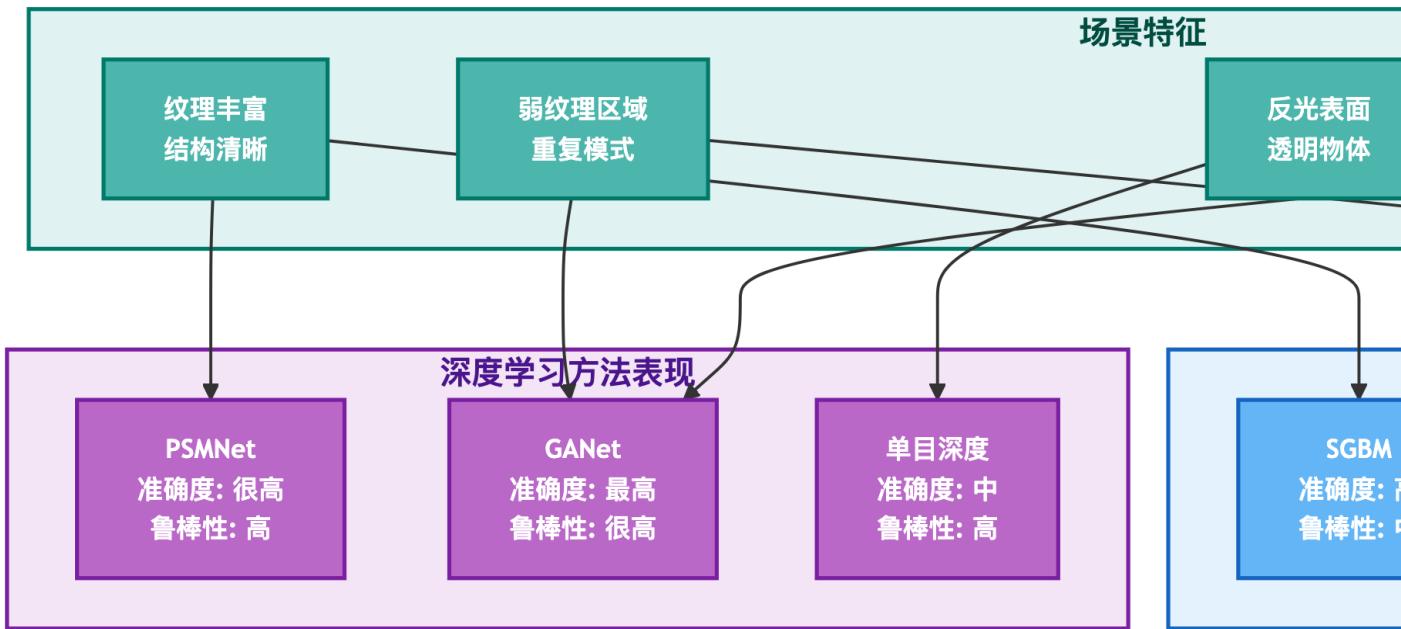


图11.10：不同方法在各类场景中的适应性分析

深度学习方法的进展：

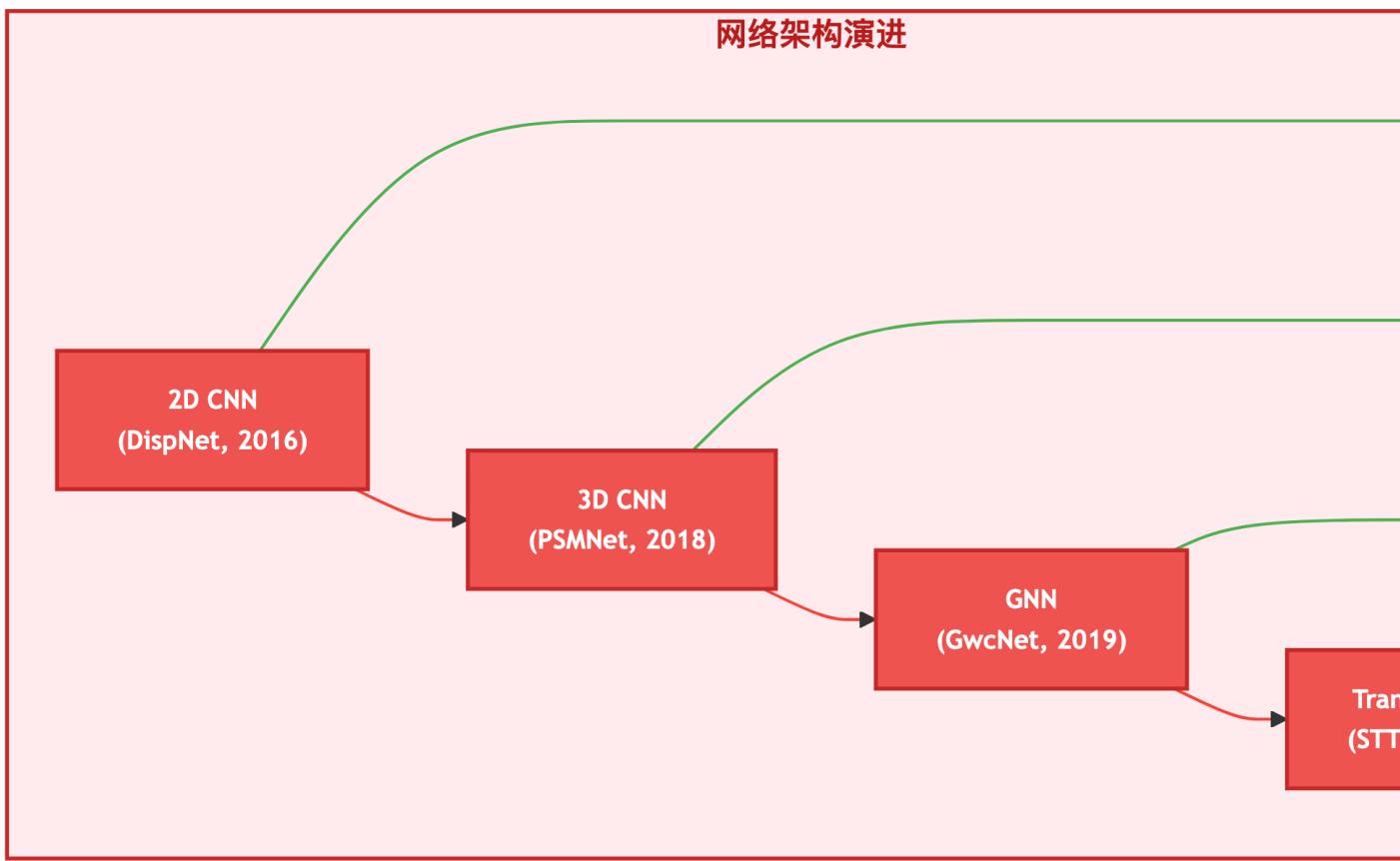


图11.11：深度学习立体匹配方法的技术演进

### 3.6

立体匹配与深度估计是三维视觉的核心技术，经历了从传统几何方法到深度学习方法的重要演进。传统方法如SGBM基于几何约束和手工特征，计算效率高但在复杂场景下容易失效。现代深度学习方法如PSMNet通过端到端学习，在准确性和鲁棒性方面显著超越传统方法。

本节的核心贡献在于：**理论层面**，阐述了从视差计算到深度回归的算法演进逻辑；**技术层面**，对比了传统方法和深度学习方法的核心差异；**应用层面**，分析了不同方法在各类场景中的适应性。

深度估计技术与相机标定紧密相连：准确的相机标定是高质量深度估计的前提。同时，深度估计也为后续的三维重建和点云处理提供了基础数据。随着Transformer等新架构的引入，深度估计正朝着更高精度、更强泛化能力的方向发展，在自动驾驶、机器人等领域发挥着越来越重要的作用。

# 4

## 4.1

三维重建是计算机视觉的终极目标之一：从二维图像中恢复完整的三维场景结构。这一技术让计算机能够理解真实世界的几何形状、空间布局和物体关系，为虚拟现实、数字文化遗产保护、建筑测量等应用提供了基础支撑。

传统的三维重建方法主要基于多视图几何，通过分析多张图像间的几何关系来恢复三维结构。运动恢复结构（Structure from Motion, SfM）是其中的代表性方法，它能够从无序的图像集合中同时估计相机运动轨迹和场景的三维结构。

现代三维重建技术则融合了深度传感器和神经网络方法。RGB-D重建利用深度相机提供的深度信息，实现实时的三维场景重建；神经辐射场（NeRF）等深度学习方法则能够从稀疏视图中生成高质量的三维表示。这些技术的发展使得三维重建从实验室走向了实际应用。

## 4.2

**运动恢复结构（SfM）**是传统三维重建的核心方法。其基本思想是：如果我们知道多张图像中特征点的对应关系，就可以通过三角测量恢复这些点的三维坐标，同时估计拍摄这些图像时的相机位置和姿态。SfM的优势在于只需要普通相机即可实现三维重建，但需要场景具有丰富的纹理特征。

**RGB-D重建**利用深度相机（如Kinect、RealSense）提供的彩色图像和深度图像进行三维重建。深度信息的直接获取大大简化了重建过程，使得实时重建成为可能。TSDF（Truncated Signed Distance Function）融合是RGB-D重建的核心技术，它将多帧深度数据融合到统一的体素网格中。

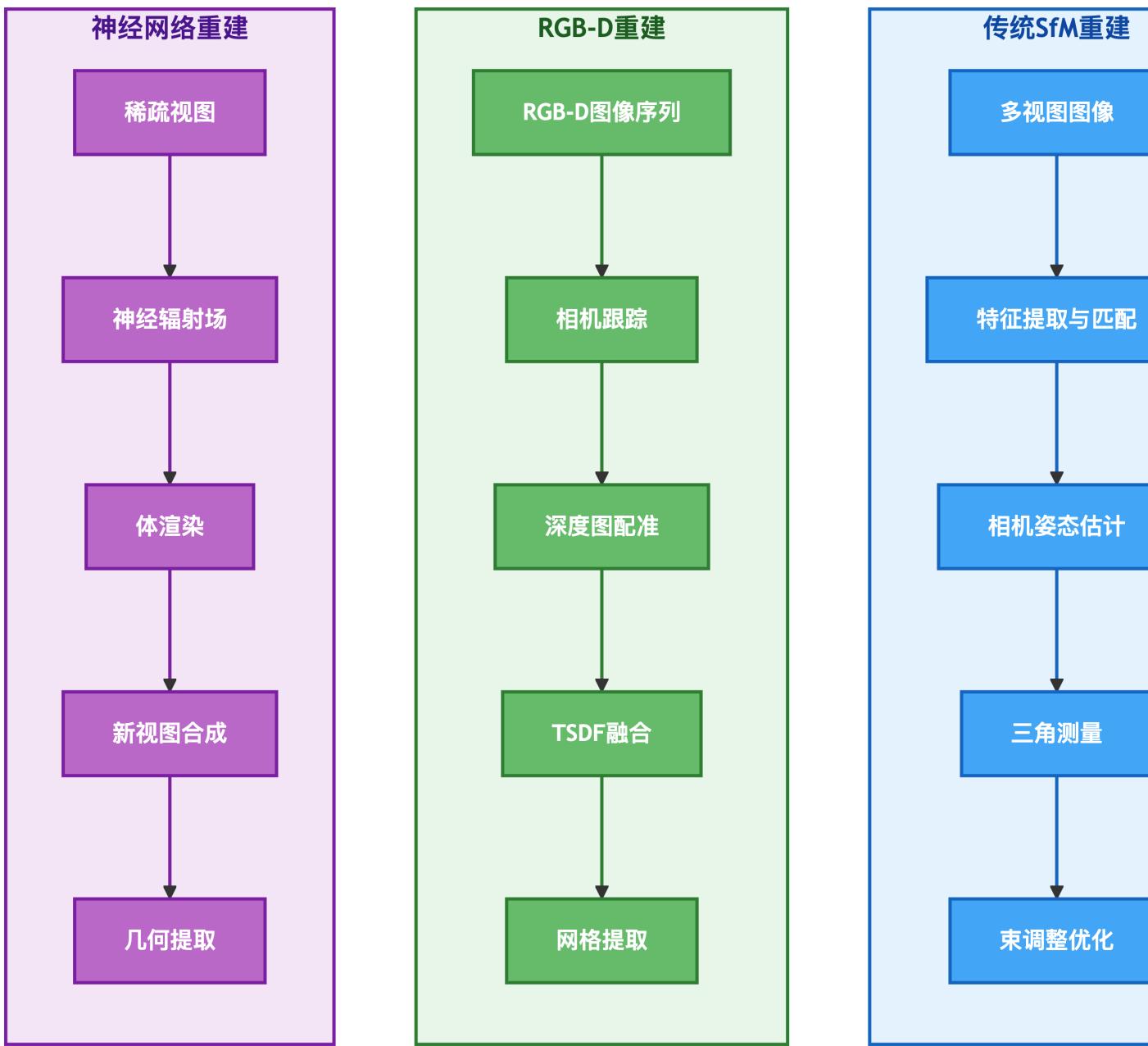


图11.12：三种主要三维重建方法的技术流程对比

**神经辐射场 (NeRF)** 代表了三维重建的最新发展方向。它使用多层感知机 (MLP) 来表示三维场景，将空间坐标和视角方向映射为颜色和密度值。通过体渲染技术，NeRF能够生成任意视角的高质量图像，并隐式地表示场景的三维几何结构。

TSDF融合是RGB-D重建中的关键技术。TSDF将三维空间划分为规则的体素网格，每个体素

存储到最近表面的有符号距离。通过融合多帧深度数据，TSDF能够处理噪声和遮挡，生成平滑的三维表面。

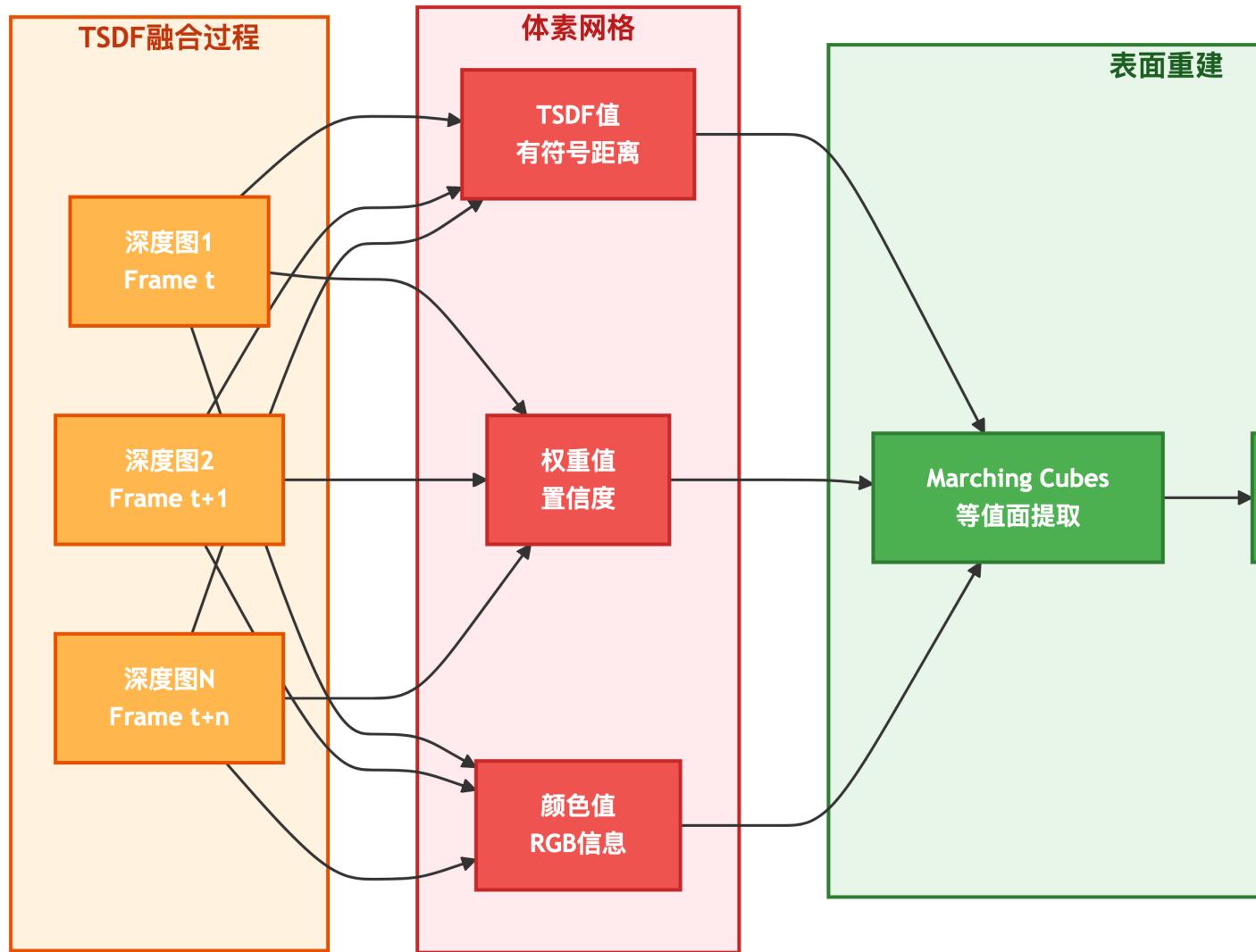


图11.13：TSDF融合的数据流程和体素网格表示

### 4.3

三维重建的理论基础涵盖了传统几何方法和现代神经网络方法，下面我们分别介绍这些方法的核心理论。

### 4.3.1 SfM

SfM的理论基础是多视图几何和投影模型。对于空间中的点 $\mathbf{X} = (X, Y, Z, 1)^T$ , 其在图像*i*中的投影点 $\mathbf{x}_i = (u_i, v_i, 1)^T$ 满足:

$$\lambda_i \mathbf{x}_i = \mathbf{P}_i \mathbf{X} = \mathbf{K}_i [\mathbf{R}_i | \mathbf{t}_i] \mathbf{X}$$

其中,  $\mathbf{P}_i$ 是投影矩阵,  $\mathbf{K}_i$ 是内参矩阵,  $\mathbf{R}_i$ 和 $\mathbf{t}_i$ 分别是旋转矩阵和平移向量,  $\lambda_i$ 是尺度因子。

SfM的核心问题是: 已知多张图像中的对应点 $\{\mathbf{x}_i\}$ , 如何恢复相机参数 $\{\mathbf{P}_i\}$ 和三维点 $\mathbf{X}$ ? 这个问题可以通过以下步骤解决:

#### 1. 特征匹配与基础矩阵估计

对于两张图像, 我们首先提取特征点 (如SIFT、ORB) 并建立匹配。然后估计基础矩阵 $\mathbf{F}$ , 它满足对极约束:

$$\mathbf{x}_2^T \mathbf{F} \mathbf{x}_1 = 0$$

基础矩阵可以通过8点法或RANSAC算法估计。

#### 2. 相机姿态估计

从基础矩阵 $\mathbf{F}$ 可以分解出本质矩阵 $\mathbf{E}$ :

$$\mathbf{E} = \mathbf{K}_2^T \mathbf{F} \mathbf{K}_1$$

进一步分解本质矩阵可得到相对旋转 $\mathbf{R}$ 和平移 $\mathbf{t}$ :

$$\mathbf{E} = [\mathbf{t}]_\times \mathbf{R}$$

其中 $[\mathbf{t}]_\times$ 是 $\mathbf{t}$ 的反对称矩阵。

#### 3. 三角测量

已知两个相机的投影矩阵 $\mathbf{P}_1$ 和 $\mathbf{P}_2$ , 以及对应点 $\mathbf{x}_1$ 和 $\mathbf{x}_2$ , 可以通过三角测量恢复三维点 $\mathbf{X}$ 。这可以表示为一个线性方程组:

$$\begin{bmatrix} \mathbf{x}_1 \times \mathbf{P}_1 \\ \mathbf{x}_2 \times \mathbf{P}_2 \end{bmatrix} \mathbf{X} = \mathbf{0}$$

通过SVD求解这个方程组的最小二乘解。

#### 4. 束调整优化

最后，通过束调整 (Bundle Adjustment) 优化相机参数和三维点坐标，最小化重投影误差：

$$\min_{\{\mathbf{P}_i\}, \{\mathbf{x}_j\}} \sum_{i,j} d(\mathbf{x}_{ij}, \mathbf{P}_i \mathbf{x}_j)^2$$

其中  $d(\cdot, \cdot)$  是欧氏距离， $\mathbf{x}_{ij}$  是第  $j$  个三维点在第  $i$  个相机中的观测。

#### 4.3.2 TSDF

TSDF (Truncated Signed Distance Function) 是一种隐式表面表示方法，它将三维空间划分为规则的体素网格，每个体素存储到最近表面的有符号距离。

对于空间中的点  $\mathbf{p} = (x, y, z)$ ，其 TSDF 值定义为：

$$TSDF(\mathbf{p}) = \begin{cases} \min(1, \frac{d(\mathbf{p})}{t}) & \text{if } d(\mathbf{p}) \geq 0 \\ \max(-1, \frac{d(\mathbf{p})}{t}) & \text{if } d(\mathbf{p}) < 0 \end{cases}$$

其中， $d(\mathbf{p})$  是点  $\mathbf{p}$  到最近表面的有符号距离， $t$  是截断距离。正值表示点在表面外部，负值表示点在表面内部，零值表示点在表面上。

TSDF 融合的核心是将多帧深度图融合到统一的 TSDF 体素网格中。对于第  $k$  帧深度图，每个体素的 TSDF 值和权重更新如下：

$$TSDF_k(\mathbf{p}) = \frac{W_{k-1}(\mathbf{p}) \cdot TSDF_{k-1}(\mathbf{p}) + w_k(\mathbf{p}) \cdot TSDF'_k(\mathbf{p})}{W_{k-1}(\mathbf{p}) + w_k(\mathbf{p})}$$

$$W_k(\mathbf{p}) = W_{k-1}(\mathbf{p}) + w_k(\mathbf{p})$$

其中， $TSDF'_k(\mathbf{p})$  是从当前深度图计算的 TSDF 值， $w_k(\mathbf{p})$  是当前测量的权重。

最后，通过 Marching Cubes 算法从 TSDF 体素网格中提取等值面（零值面），得到三维表面的三角网格表示。

### 4.3.3 NeRF

NeRF是一种基于神经网络的隐式场景表示方法。它使用多层感知机 (MLP) 来表示三维场景，将空间坐标 $\mathbf{x} = (x, y, z)$ 和视角方向 $\mathbf{d} = (\theta, \phi)$ 映射为颜色 $\mathbf{c} = (r, g, b)$ 和密度 $\sigma$ :

$$F_{\Theta} : (\mathbf{x}, \mathbf{d}) \rightarrow (\mathbf{c}, \sigma)$$

其中， $F_{\Theta}$ 是参数为 $\Theta$ 的神经网络。

给定一条从相机中心出发的光线 $\mathbf{r}(t) = \mathbf{o} + t\mathbf{d}$ ，NeRF通过体渲染方程计算该光线上的颜色：

$$C(\mathbf{r}) = \int_{t_n}^{t_f} T(t) \sigma(\mathbf{r}(t)) \mathbf{c}(\mathbf{r}(t), \mathbf{d}) dt$$

其中， $T(t) = \exp(-\int_{t_n}^t \sigma(\mathbf{r}(s)) ds)$ 是累积透射率，表示光线从 $t_n$ 到 $t$ 的透明度。

在实践中，这个积分通过离散采样近似计算：

$$\hat{C}(\mathbf{r}) = \sum_{i=1}^N T_i (1 - \exp(-\sigma_i \delta_i)) \mathbf{c}_i$$

其中， $T_i = \exp(-\sum_{j=1}^{i-1} \sigma_j \delta_j)$ ， $\delta_i$ 是相邻采样点之间的距离。

NeRF通过最小化渲染图像与真实图像之间的差异来优化网络参数：

$$\mathcal{L} = \sum_{\mathbf{r} \in \mathcal{R}} \|\hat{C}(\mathbf{r}) - C_{gt}(\mathbf{r})\|_2^2$$

其中， $\mathcal{R}$ 是训练集中的所有光线， $C_{gt}(\mathbf{r})$ 是光线 $\mathbf{r}$ 对应的真实颜色。

## 4.4

下面我们分别介绍SfM、TSDF融合和NeRF的核心算法实现。

#### 4.4.1 SfM

SfM的实现通常基于特征匹配和几何优化。以下是使用OpenCV实现的SfM核心步骤：

```
import cv2
import numpy as np
from scipy.optimize import least_squares

def structure_from_motion(images):
    """SfM"""
    # 1.
    features = extract_features(images)
    matches = match_features(features)

    # 2.
    K = estimate_camera_intrinsics() #
    E, mask = cv2.findEssentialMat(matches[0], matches[1], K)
    _, R, t, _ = cv2.recoverPose(E, matches[0], matches[1], K)

    #
    P1 = np.hstack((np.eye(3), np.zeros((3, 1))))
    P2 = np.hstack((R, t))

    # 3.
    points_3d = triangulate_points(matches[0], matches[1], P1, P2, K)

    # 4. SfM
    for i in range(2, len(images)):
        # 2D-3D
        points_2d = find_2d_3d_correspondences(features[i], points_3d)

        # PnP
        _, rvec, tvec, inliers = cv2.solvePnPRansac(
            points_3d, points_2d, K, None)
        R_new = cv2.Rodrigues(rvec)[0]
        t_new = tvec

        #
        new_matches = find_new_matches(features[i-1], features[i])
        new_points_3d = triangulate_points(
            new_matches[0], new_matches[1],
            P1, np.hstack((R_new, t_new)), K)
```

```

points_3d = np.vstack((points_3d, new_points_3d))

# 5.
camera_params, points_3d = bundle_adjustment(
    camera_params, points_3d, observations)

return camera_params, points_3d

def bundle_adjustment(camera_params, points_3d, observations):
    """
    """
    #
    def reprojection_error(params, n_cameras, n_points, camera_indices,
                           point_indices, observations):
        camera_params = params[:n_cameras * 6].reshape((n_cameras, 6))
        points_3d = params[n_cameras * 6:].reshape((n_points, 3))

        projected = project(points_3d[point_indices], camera_params[camera_indices])
        return (projected - observations).ravel()

    #
    params = np.hstack((camera_params.ravel(), points_3d.ravel()))

    #
    result = least_squares(
        reprojection_error, params,
        args=(n_cameras, n_points, camera_indices, point_indices, observations),
        method='trf', ftol=1e-4, xtol=1e-4, gtol=1e-4)

    #
    params = result.x
    camera_params = params[:n_cameras * 6].reshape((n_cameras, 6))
    points_3d = params[n_cameras * 6:].reshape((n_points, 3))

    return camera_params, points_3d

```

#### 4.4.2 TSDF

TSDF融合算法的核心是将深度图转换为TSDF表示，并融合多帧数据：

```

import numpy as np

class TSDFVolume:
    """TSDF"""
    def __init__(self, vol_bounds, voxel_size, trunc_margin):
        #
        self.voxel_size = voxel_size
        self.trunc_margin = trunc_margin

        #
        vol_dim = np.ceil((vol_bounds[:, 1] - vol_bounds[:, 0]) / voxel_size).astype(int)
        self.vol_bounds = vol_bounds
        self.vol_dim = vol_dim

        #
        self.voxel_grid_tsdf = np.ones(vol_dim) * 1.0
        self.voxel_grid_weight = np.zeros(vol_dim)

        #
        self._compute_voxel_centers()

    def _compute_voxel_centers(self):
        """
        #
        xv, yv, zv = np.meshgrid(
            np.arange(0, self.vol_dim[0]),
            np.arange(0, self.vol_dim[1]),
            np.arange(0, self.vol_dim[2]))
        #
        self.voxel_centers = np.stack([xv, yv, zv], axis=-1) * self.voxel_size + self.vol_bounds[:, :3]

    def integrate(self, depth_img, K, pose):
        """TSDF"""
        #
        cam_pts = self.voxel_centers.reshape(-1, 3)
        cam_pts = np.matmul(cam_pts - pose[:3, 3], pose[:3, :3].T)

        #
        pix_x = np.round(cam_pts[:, 0] * K[0, 0] / cam_pts[:, 2] + K[0, 2]).astype(int)
        pix_y = np.round(cam_pts[:, 1] * K[1, 1] / cam_pts[:, 2] + K[1, 2]).astype(int)

```

```

#
valid_pix = (pix_x >= 0) & (pix_x < depth_img.shape[1]) & \
            (pix_y >= 0) & (pix_y < depth_img.shape[0]) & \
            (cam_pts[:, 2] > 0)

#
depth_values = np.zeros(pix_x.shape)
depth_values[valid_pix] = depth_img[pix_y[valid_pix], pix_x[valid_pix]]

# TSDF
dist = depth_values - cam_pts[:, 2]
tsdf_values = np.minimum(1.0, dist / self.trunc_margin)
tsdf_values = np.maximum(-1.0, tsdf_values)

#
weights = (depth_values > 0).astype(float)

# TSDF
tsdf_vol_new = self.voxel_grid_tsdf.reshape(-1)
weight_vol_new = self.voxel_grid_weight.reshape(-1)

# TSDF
mask = valid_pix & (depth_values > 0) & (dist > -self.trunc_margin)
tsdf_vol_new[mask] = (weight_vol_new[mask] * tsdf_vol_new[mask] + weights[mask] * tsdf_vol_new[mask]) / (weight_vol_new[mask] + weights[mask])

#
weight_vol_new[mask] += weights[mask]

#
self.voxel_grid_tsdf = tsdf_vol_new.reshape(self.vol_dim)
self.voxel_grid_weight = weight_vol_new.reshape(self.vol_dim)

```

#### 4.4.3 NeRF

NeRF使用PyTorch实现，核心是神经网络模型和体渲染算法：

```

import torch
import torch.nn as nn
import torch.nn.functional as F

```

```

class NeRF(nn.Module):
    """
    """
    def __init__(self, D=8, W=256, input_ch=3, input_ch_views=3, output_ch=4):
        super(NeRF, self).__init__()
        self.D = D
        self.W = W
        self.input_ch = input_ch
        self.input_ch_views = input_ch_views
        self.output_ch = output_ch

        #
        input_ch = self.input_ch

        #
        self.pts_linears = nn.ModuleList(
            [nn.Linear(input_ch, W)] +
            [nn.Linear(W, W) for _ in range(D-1)])

        #
        self.alpha_linear = nn.Linear(W, 1)

        #
        self.feature_linear = nn.Linear(W, W)
        self.views_linears = nn.ModuleList([nn.Linear(W + input_ch_views, W//2)])

        #
        self.rgb_linear = nn.Linear(W//2, 3)

    def forward(self, x):
        """
        """
        #
        input_pts, input_views = torch.split(
            x, [self.input_ch, self.input_ch_views], dim=-1)

        #
        h = input_pts
        for i, l in enumerate(self.pts_linears):
            h = self.pts_linears[i](h)
            h = F.relu(h)

        #
        alpha = self.alpha_linear(h)

```

```

#
feature = self.feature_linear(h)

#
h = torch.cat([feature, input_views], -1)
for i, l in enumerate(self.views_linears):
    h = self.views_linears[i](h)
    h = F.relu(h)

# RGB
rgb = self.rgb_linear(h)
rgb = torch.sigmoid(rgb)

# RGB
outputs = torch.cat([rgb, alpha], -1)
return outputs

def render_rays(model, rays_o, rays_d, near, far, N_samples):
    """
    #
    t_vals = torch.linspace(0., 1., steps=N_samples)
    z_vals = near * (1.-t_vals) + far * t_vals

    #
    z_vals = z_vals.expand([rays_o.shape[0], N_samples])
    mids = .5 * (z_vals[...,1:] + z_vals[...,:-1])
    upper = torch.cat([mids, z_vals[...,-1:]], -1)
    lower = torch.cat([z_vals[...,:1], mids], -1)
    t_rand = torch.rand(z_vals.shape)
    z_vals = lower + (upper - lower) * t_rand

    #
    pts = rays_o[...,:,None,:] + rays_d[...,:,None,:] * z_vals[...,:,None]

    #
    raw = model(pts)

    #
    dists = z_vals[...,1:] - z_vals[...,:-1]
    dists = torch.cat([dists, torch.ones_like(dists[...,:1]) * 1e10], -1)

    #
    alpha

```

```
alpha = 1.0 - torch.exp(-raw[...,:3] * dists)

#
weights = alpha * torch.cumprod(
    torch.cat([torch.ones_like(alpha[...,:1]), 1.-alpha[...,:-1]], -1), -1)

#
rgb = torch.sum(weights[...,None] * raw[...,:3], -2)

#
depth = torch.sum(weights * z_vals, -1)

return rgb, depth, weights
```

## 4.5

三维重建算法的效果可以从重建精度、计算效率和应用场景适应性等多个维度进行评估。

#### 4.5.1

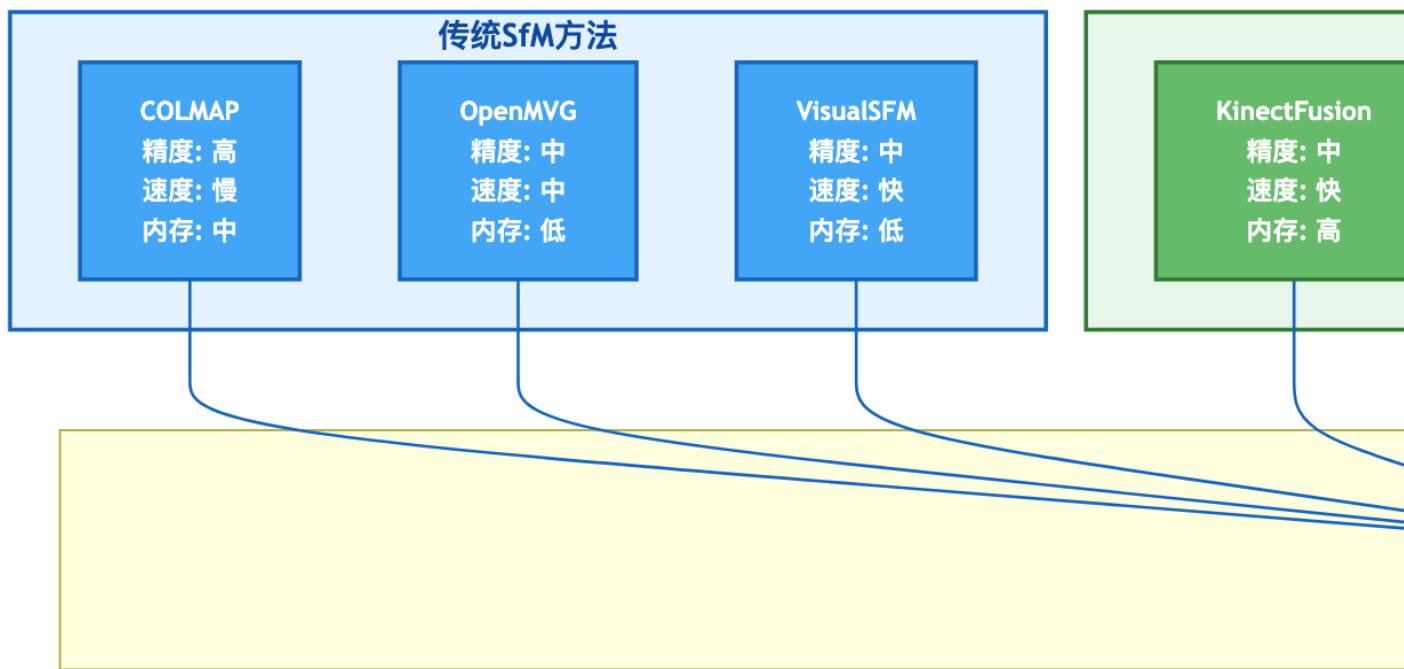
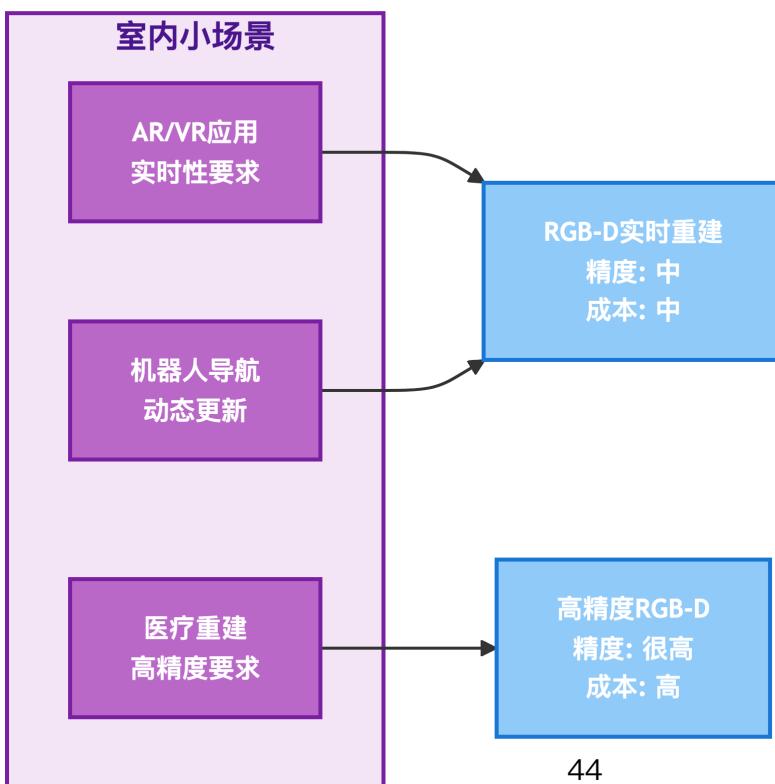
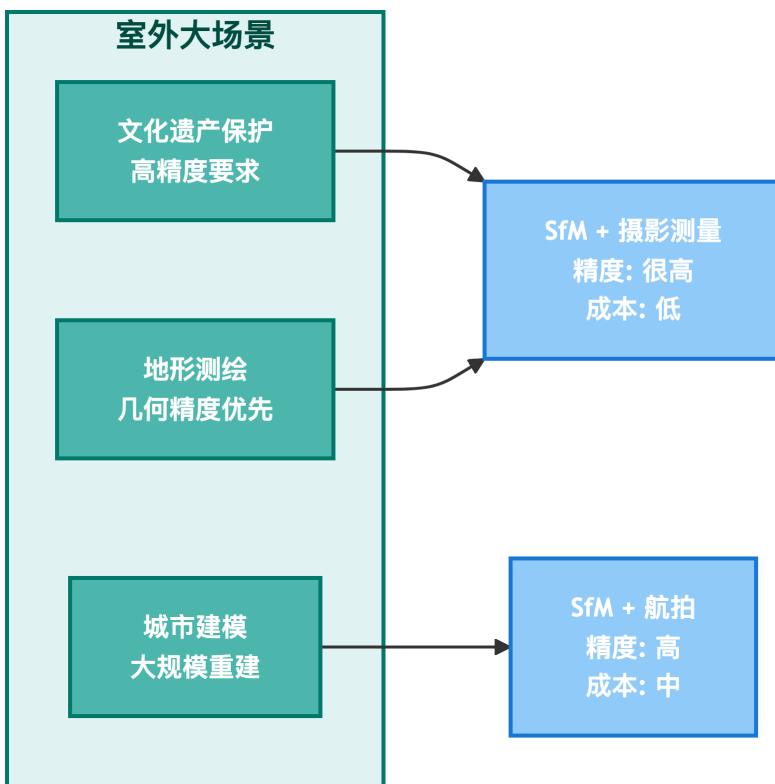


图11.14：不同三维重建方法的性能对比分析



#### 4.5.2



44

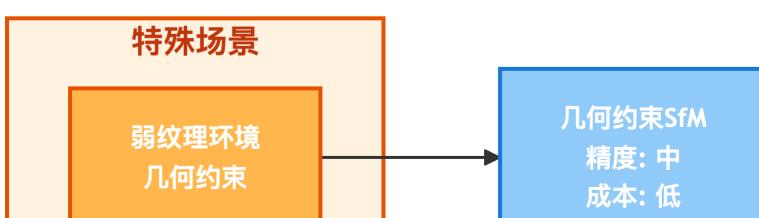


图11.15：三维重建方法在不同应用场景中的适应性

#### 4.5.3

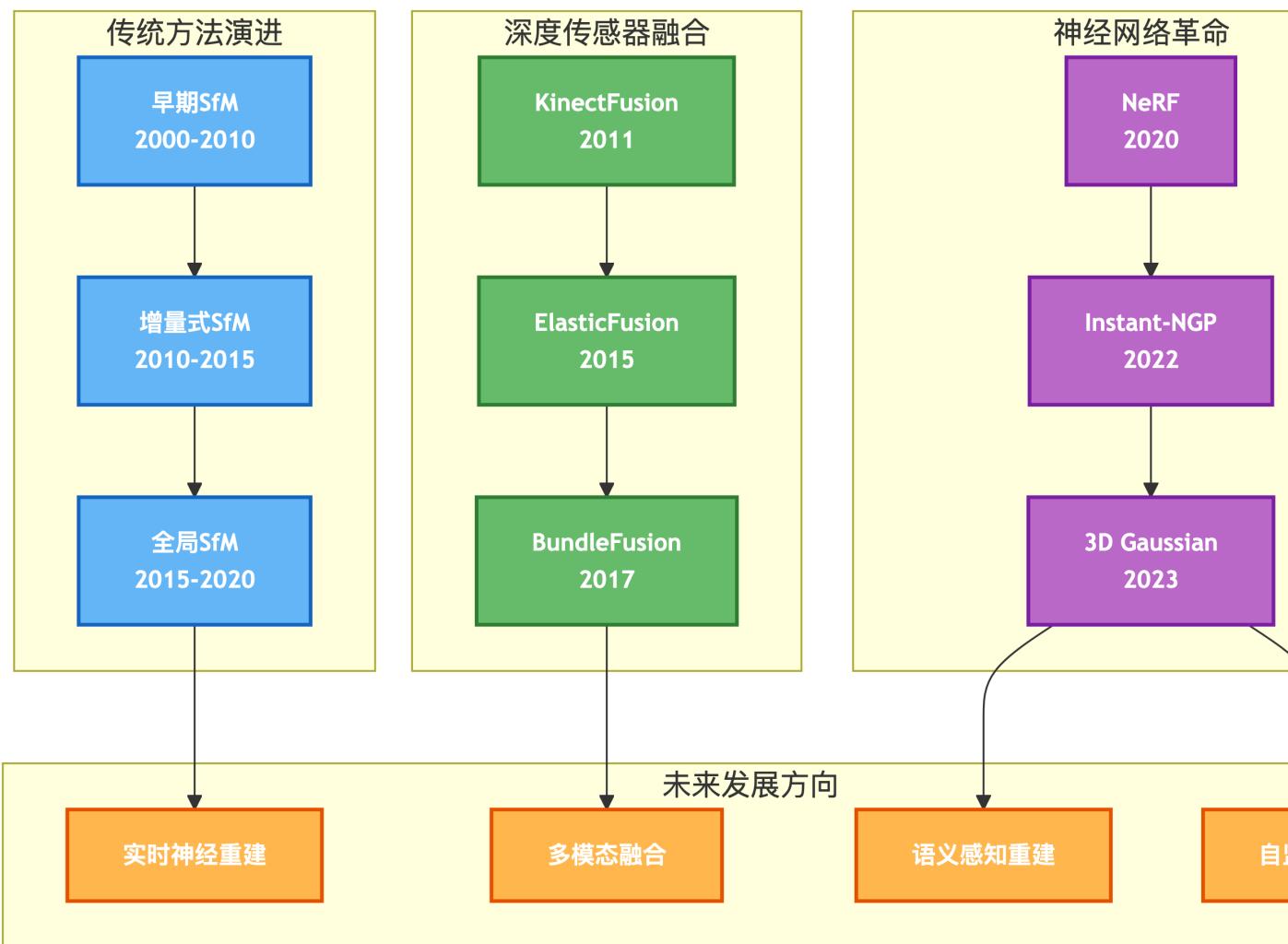


图11.16：三维重建技术的发展历程和未来趋势

#### 4.6

三维重建是计算机视觉的核心技术之一，经历了从传统几何方法到现代神经网络方法的重要演进。传统SfM方法基于多视图几何，能够从普通图像中恢复三维结构，但需要丰富的纹理

特征；RGB-D重建利用深度传感器，实现了实时重建，但受限于传感器范围；神经辐射场等深度学习方法则能够生成高质量的三维表示，但计算成本较高。

本节的核心贡献在于：**理论层面**，系统阐述了从多视图几何到神经隐式表示的理论基础；**技术层面**，对比了SfM、TSDF融合和NeRF的核心算法差异；**应用层面**，分析了不同方法在各类场景中的适应性和发展趋势。

三维重建技术与前面章节的相机标定和立体匹配紧密相连：相机标定提供了准确的几何参数，立体匹配提供了深度信息，而三维重建则将这些信息整合为完整的三维模型。随着神经网络技术的发展，三维重建正朝着更高质量、更高效率、更强泛化能力的方向发展，在数字孪生、元宇宙等新兴应用中发挥着越来越重要的作用。

# 5

## 5.1

点云是三维空间中点的集合，每个点通常包含三维坐标( $x, y, z$ )以及可能的附加属性（如颜色、强度、法向量等）。作为三维数据的重要表示形式，点云在激光雷达扫描、深度相机采集、三维重建等应用中发挥着核心作用。与传统的二维图像相比，点云直接表示了物体的三维几何结构，为机器人导航、自动驾驶、工业检测等应用提供了丰富的空间信息。

然而，点云数据也带来了独特的挑战。首先是**数据的无序性**：点云中的点没有固定的排列顺序，这与图像的规则网格结构形成鲜明对比。其次是**数据的稀疏性和不均匀性**：点云密度在不同区域可能差异很大，远处物体的点密度通常较低。此外，点云数据还面临**噪声和异常值**的问题，传感器误差和环境干扰会产生不准确的测量点。

现代点云处理技术需要解决这些挑战，从基础的数据结构设计到高级的语义理解，形成了完整的技术体系。传统方法主要基于几何特征和统计分析，如KD-Tree空间索引、体素化表示、聚类分析等；现代深度学习方法则直接学习点云的特征表示，如PointNet系列网络。本节将重点介绍点云处理的基础理论和核心算法，为后续的深度学习方法奠定基础。

## 5.2

**点云数据结构**是点云处理的基础。最简单的点云表示是一个 $N \times 3$ 的矩阵，其中 $N$ 是点的数量，每行表示一个点的三维坐标。在实际应用中，点云通常还包含额外的属性信息：

- **几何属性**：坐标( $x,y,z$ )、法向量( $nx,ny,nz$ )、曲率等
- **外观属性**：颜色(R,G,B)、反射强度、材质信息等
- **语义属性**：类别标签、实例ID、置信度等

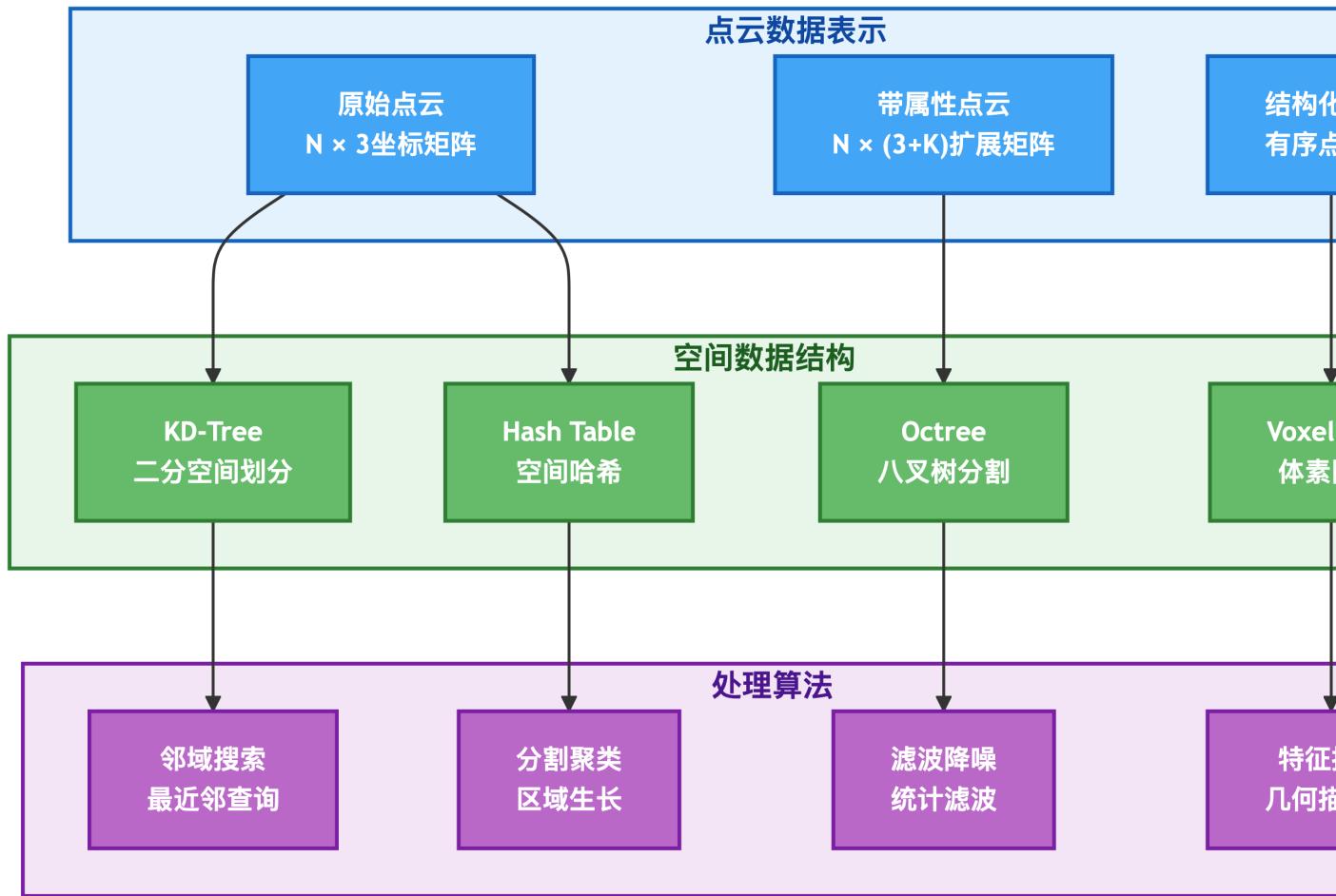


图11.17：点云数据结构与处理算法的层次关系

**空间索引结构**是高效点云处理的关键。由于点云数据量通常很大（数万到数百万个点），直接的线性搜索效率极低。常用的空间索引结构包括：

- **KD-Tree (K维树)**：通过递归地沿不同维度分割空间来组织点云数据，支持高效的最近邻搜索和范围查询
- **Octree (八叉树)**：将三维空间递归分割为8个子立方体，适合处理稀疏和不均匀分布的点云
- **Voxel Grid (体素网格)**：将空间划分为规则的立方体网格，每个体素包含落入其中的所有点
- **空间哈希**：使用哈希函数将空间坐标映射到哈希表，实现常数时间的空间查询

**点云滤波与预处理**是点云分析的重要步骤。原始点云数据通常包含噪声、异常值和冗余信息，需要通过滤波算法进行清理：

- **统计滤波**：基于邻域点的统计特性识别和移除异常值

- 半径滤波：移除指定半径内邻居数量过少的孤立点
- 直通滤波：根据坐标范围过滤点云，移除感兴趣区域外的点
- 下采样：减少点云密度以降低计算复杂度，常用体素网格下采样

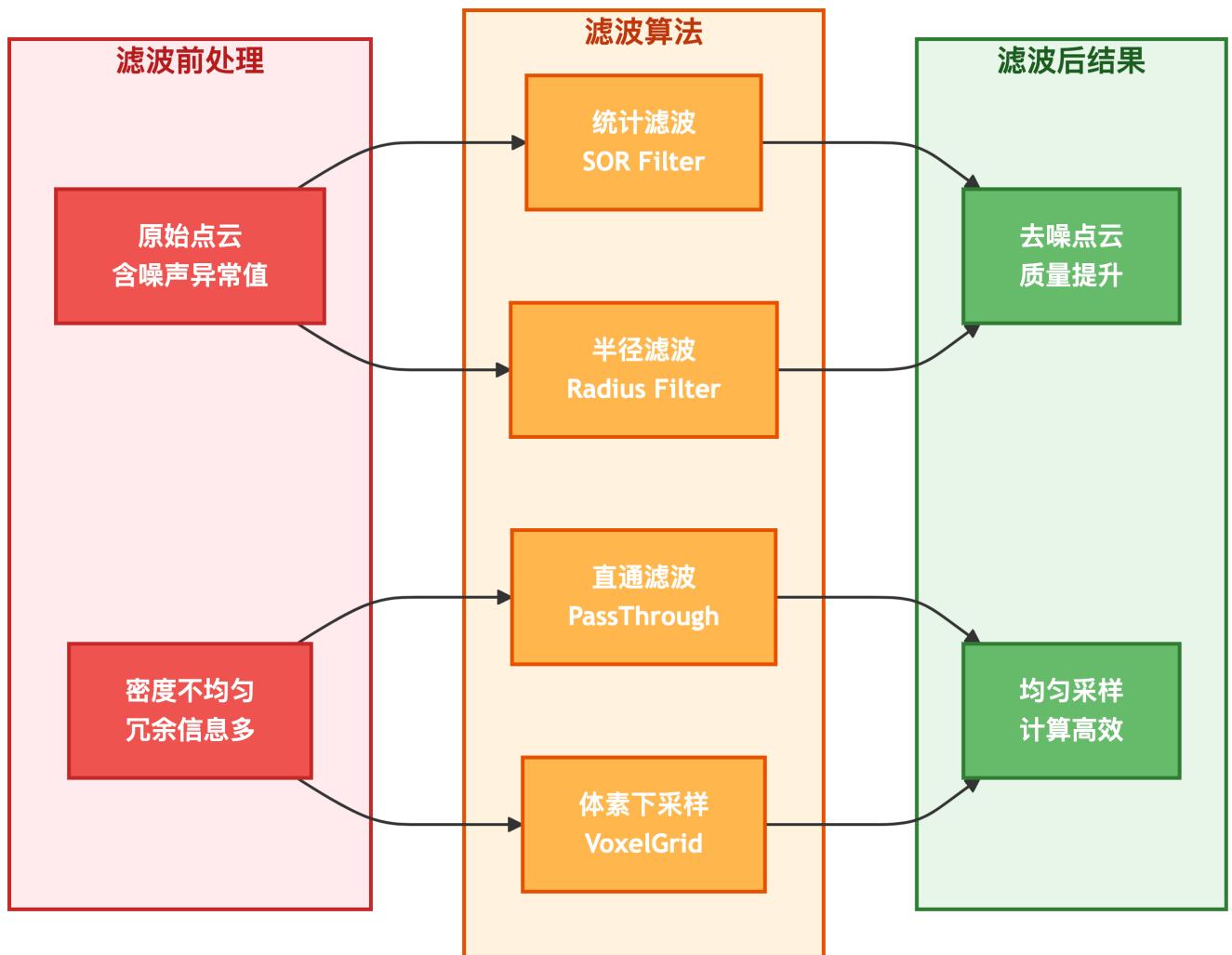


图11.18：点云滤波处理的完整流程

### 5.3

点云处理的理论基础主要涉及空间数据结构、几何算法和统计分析方法。下面我们详细介绍这些核心理论。

### 5.3.1 KD-Tree

KD-Tree (K-Dimensional Tree) 是一种用于组织k维空间中点的二叉搜索树。对于三维点云， $k=3$ 。KD-Tree的构建过程是递归的：

#### 1. 构建算法

给定点集  $P = \{p_1, p_2, \dots, p_n\}$ , 其中  $p_i = (x_i, y_i, z_i)$ , KD-Tree的构建过程如下：

- 选择分割维度：通常选择方差最大的维度，或者循环选择x、y、z维度
- 选择分割点：通常选择该维度上的中位数点
- 递归构建：将点集分为两部分，分别构建左右子树

#### 2. 搜索算法

KD-Tree支持多种查询操作，最重要的是最近邻搜索（Nearest Neighbor Search）：

对于查询点  $q$ , 最近邻搜索的时间复杂度为  $O(\log n)$  (平均情况)。搜索过程包括：

- 向下搜索：从根节点开始，根据分割维度选择子树
- 回溯搜索：检查是否需要搜索另一个子树
- 剪枝优化：利用当前最佳距离进行剪枝

最近邻距离的计算公式为：

$$d(p, q) = \sqrt{(p_x - q_x)^2 + (p_y - q_y)^2 + (p_z - q_z)^2}$$

#### 3. 范围搜索

KD-Tree还支持范围搜索，即查找指定区域内的所有点。对于球形范围搜索，给定中心点  $c$  和半径  $r$ , 需要找到所有满足  $d(p, c) \leq r$  的点  $p$ 。

### 5.3.2

体素化 (Voxelization) 是将连续的三维空间离散化为规则网格的过程。每个体素 (Voxel) 是一个立方体单元，类似于二维图像中的像素。

#### 1. 体素网格定义

给定点云的边界框  $[x_{min}, x_{max}] \times [y_{min}, y_{max}] \times [z_{min}, z_{max}]$  和体素大小  $v$ , 体素网格的尺寸为：

$$N_x = \lceil \frac{x_{max} - x_{min}}{v} \rceil$$

$$N_y = \lceil \frac{y_{max} - y_{min}}{v} \rceil$$

$$N_z = \lceil \frac{z_{max} - z_{min}}{v} \rceil$$

## 2. 点到体素的映射

对于点  $p = (x, y, z)$ , 其对应的体素索引为:

$$i = \lfloor \frac{x - x_{min}}{v} \rfloor$$

$$j = \lfloor \frac{y - y_{min}}{v} \rfloor$$

$$k = \lfloor \frac{z - z_{min}}{v} \rfloor$$

## 3. 体素特征计算

每个体素可以计算多种特征:

- 点数量:  $N_{ijk} = |\{p \in P : p \text{ 属于体素 } (i, j, k)\}|$
- 质心坐标:  $\bar{p}_{ijk} = \frac{1}{N_{ijk}} \sum_{p \in V_{ijk}} p$
- 协方差矩阵:  $C_{ijk} = \frac{1}{N_{ijk}} \sum_{p \in V_{ijk}} (p - \bar{p}_{ijk})(p - \bar{p}_{ijk})^T$

### 5.3.3

点云聚类旨在将点云分割为若干个具有相似特性的子集。常用的聚类算法包括:

#### 1. 欧几里得聚类

基于距离的聚类方法, 将距离小于阈值  $\epsilon$  的点归为同一类:

$$C_i = \{p \in P : \exists q \in C_i, d(p, q) < \epsilon\}$$

这等价于在点云上构建邻接图, 然后寻找连通分量。

#### 2. 区域生长聚类

从种子点开始, 根据几何特征 (如法向量) 逐步扩展区域:

- 选择种子点  $p_0$
- 计算邻域点的法向量角度差:  $\theta = \arccos(n_i \cdot n_j)$

- 如果  $\theta < \theta_{threshold}$ , 则将邻域点加入当前区域
- 递归处理新加入的点

### 3. DBSCAN聚类

基于密度的聚类算法，能够发现任意形状的聚类并识别噪声点：

- **核心点**: 半径 $\epsilon$ 内至少有 $MinPts$ 个邻居的点
- **边界点**: 不是核心点但在某个核心点的邻域内的点
- **噪声点**: 既不是核心点也不是边界点的点

DBSCAN的时间复杂度为  $O(n \log n)$  (使用空间索引)。

#### 5.3.4

统计滤波基于点云的统计特性识别和移除异常值。

##### 1. 统计异常值移除 (SOR)

对于每个点  $p_i$ , 计算其  $k$  近邻的平均距离:

$$\bar{d}_i = \frac{1}{k} \sum_{j=1}^k d(p_i, p_{i,j})$$

其中  $p_{i,j}$  是  $p_i$  的第  $j$  个最近邻。

假设距离分布为正态分布  $N(\mu, \sigma^2)$ , 其中:

$$\mu = \frac{1}{n} \sum_{i=1}^n \bar{d}_i$$

$$\sigma^2 = \frac{1}{n-1} \sum_{i=1}^n (\bar{d}_i - \mu)^2$$

如果  $\bar{d}_i > \mu + \alpha\sigma$  (其中  $\alpha$  是标准差倍数), 则认为  $p_i$  是异常值。

##### 2. 半径滤波

对于每个点  $p_i$ , 统计半径  $r$  内的邻居数量:

$$N_i = |\{p_j \in P : d(p_i, p_j) < r\}|$$

如果  $N_i < N_{min}$ , 则认为  $p_i$  是孤立点并移除。

这些理论为点云处理算法提供了坚实的数学基础, 确保了算法的正确性和效率。

## 5.4

下面我们介绍点云处理的核心算法实现, 重点展示算法的核心思想和关键步骤。

### 5.4.1 KD-Tree

KD-Tree是点云处理中最重要的空间索引结构, 以下是其核心实现:

```
import numpy as np
from collections import namedtuple

class KDTTreeNode:
    """KD-Tree"""
    def __init__(self, point=None, left=None, right=None, axis=None):
        self.point = point      #
        self.left = left        #
        self.right = right      #
        self.axis = axis        #

class KDTree:
    """KD-Tree"""
    def __init__(self, points):
        self.root = self._build_tree(points, depth=0)

    def _build_tree(self, points, depth):
        """ KD-Tree"""
        if not points:
            return None

        #      x,y,z
        axis = depth % 3
```

```

#
points.sort(key=lambda p: p[axis])
median_idx = len(points) // 2

#
node = KDTreeNode(
    point=points[median_idx],
    axis=axis,
    left=self._build_tree(points[:median_idx], depth + 1),
    right=self._build_tree(points[median_idx + 1:], depth + 1)
)
return node

def nearest_neighbor(self, query_point):
    """
    """
    best = [None, float('inf')]

    def search(node, depth):
        if node is None:
            return

        #
        dist = np.linalg.norm(np.array(node.point) - np.array(query_point))
        if dist < best[1]:
            best[0], best[1] = node.point, dist

        #
        axis = node.axis
        if query_point[axis] < node.point[axis]:
            search(node.left, depth + 1)
            #
            if abs(query_point[axis] - node.point[axis]) < best[1]:
                search(node.right, depth + 1)
        else:
            search(node.right, depth + 1)
            if abs(query_point[axis] - node.point[axis]) < best[1]:
                search(node.left, depth + 1)

    search(self.root, 0)
    return best[0], best[1]

```

### 5.4.2

体素化是点云下采样和特征提取的重要方法：

```
import open3d as o3d
import numpy as np

class VoxelGrid:
    """
    def __init__(self, voxel_size):
        self.voxel_size = voxel_size
        self.voxel_dict = {}

    def voxelize(self, points):
        """
        #
        min_bound = np.min(points, axis=0)
        max_bound = np.max(points, axis=0)

        #
        voxel_indices = np.floor((points - min_bound) / self.voxel_size).astype(int)

        #
        for i, point in enumerate(points):
            voxel_key = tuple(voxel_indices[i])
            if voxel_key not in self.voxel_dict:
                self.voxel_dict[voxel_key] = []
            self.voxel_dict[voxel_key].append(point)

        return self.voxel_dict

    def downsample(self, points):
        """
        voxel_dict = self.voxelize(points)
        downsampled_points = []

        for voxel_points in voxel_dict.values():
            #
            centroid = np.mean(voxel_points, axis=0)
            downsampled_points.append(centroid)

        return np.array(downsampled_points)
```

```

# Open3D
def voxel_downsample_open3d(points, voxel_size):
    """ Open3D """
    pcd = o3d.geometry.PointCloud()
    pcd.points = o3d.utility.Vector3dVector(points)

    #
    downsampled_pcd = pcd.voxel_down_sample(voxel_size)
    return np.asarray(downsampled_pcd.points)

```

### 5.4.3

点云滤波是预处理的重要步骤，以下是核心滤波算法：

```

def statistical_outlier_removal(points, k=20, std_ratio=2.0):
    """
    from sklearn.neighbors import NearestNeighbors

    # k
    nbrs = NearestNeighbors(n_neighbors=k+1).fit(points)
    distances, indices = nbrs.kneighbors(points)

    # k
    mean_distances = np.mean(distances[:, 1:], axis=1)

    #
    global_mean = np.mean(mean_distances)
    global_std = np.std(mean_distances)

    #
    threshold = global_mean + std_ratio * global_std
    inlier_mask = mean_distances < threshold

    return points[inlier_mask], inlier_mask

def radius_outlier_removal(points, radius=0.05, min_neighbors=10):
    """
    from sklearn.neighbors import NearestNeighbors

    #

```

```

nbrs = NearestNeighbors(radius=radius).fit(points)
distances, indices = nbrs.radius_neighbors(points)

#
neighbor_counts = np.array([len(neighbors) - 1 for neighbors in indices]) #

#
inlier_mask = neighbor_counts >= min_neighbors
return points[inlier_mask], inlier_mask

def passthrough_filter(points, axis='z', min_val=-np.inf, max_val=np.inf):
    """
    """
    axis_map = {'x': 0, 'y': 1, 'z': 2}
    axis_idx = axis_map[axis]

    #
    mask = (points[:, axis_idx] >= min_val) & (points[:, axis_idx] <= max_val)
    return points[mask], mask

```

#### 5.4.4

聚类算法用于点云分割和目标识别：

```

def euclidean_clustering(points, tolerance=0.02, min_cluster_size=100, max_cluster_size=25000):
    """
    """
    from sklearn.neighbors import NearestNeighbors

    #
    nbrs = NearestNeighbors(radius=tolerance).fit(points)

    visited = np.zeros(len(points), dtype=bool)
    clusters = []

    for i in range(len(points)):
        if visited[i]:
            continue

        #
        cluster = []
        queue = [i]

```

```

while queue:
    current_idx = queue.pop(0)
    if visited[current_idx]:
        continue

    visited[current_idx] = True
    cluster.append(current_idx)

    #
    neighbors = nbrs.radius_neighbors([points[current_idx]], return_distance=False)[0]
    for neighbor_idx in neighbors:
        if not visited[neighbor_idx]:
            queue.append(neighbor_idx)

    #
    if min_cluster_size <= len(cluster) <= max_cluster_size:
        clusters.append(cluster)

return clusters

def dbscan_clustering(points, eps=0.02, min_samples=10):
    """DBSCAN"""
    from sklearn.cluster import DBSCAN

    # sklearn
    clustering = DBSCAN(eps=eps, min_samples=min_samples).fit(points)

    #
    labels = clustering.labels_
    n_clusters = len(set(labels)) - (1 if -1 in labels else 0)

    clusters = []
    for cluster_id in range(n_clusters):
        cluster_indices = np.where(labels == cluster_id)[0]
        clusters.append(cluster_indices.tolist())

    return clusters, labels

```

这些核心算法实现展示了点云处理的基本思想：通过空间数据结构实现高效查询，通过统计方法进行数据清理，通过几何算法进行结构分析。每个算法都针对点云数据的特点进行了优化，为后续的高级处理奠定了基础。

## 5.5

点云处理算法的效果可以从计算效率、处理质量和应用适应性等多个维度进行评估。

### 5.5.1

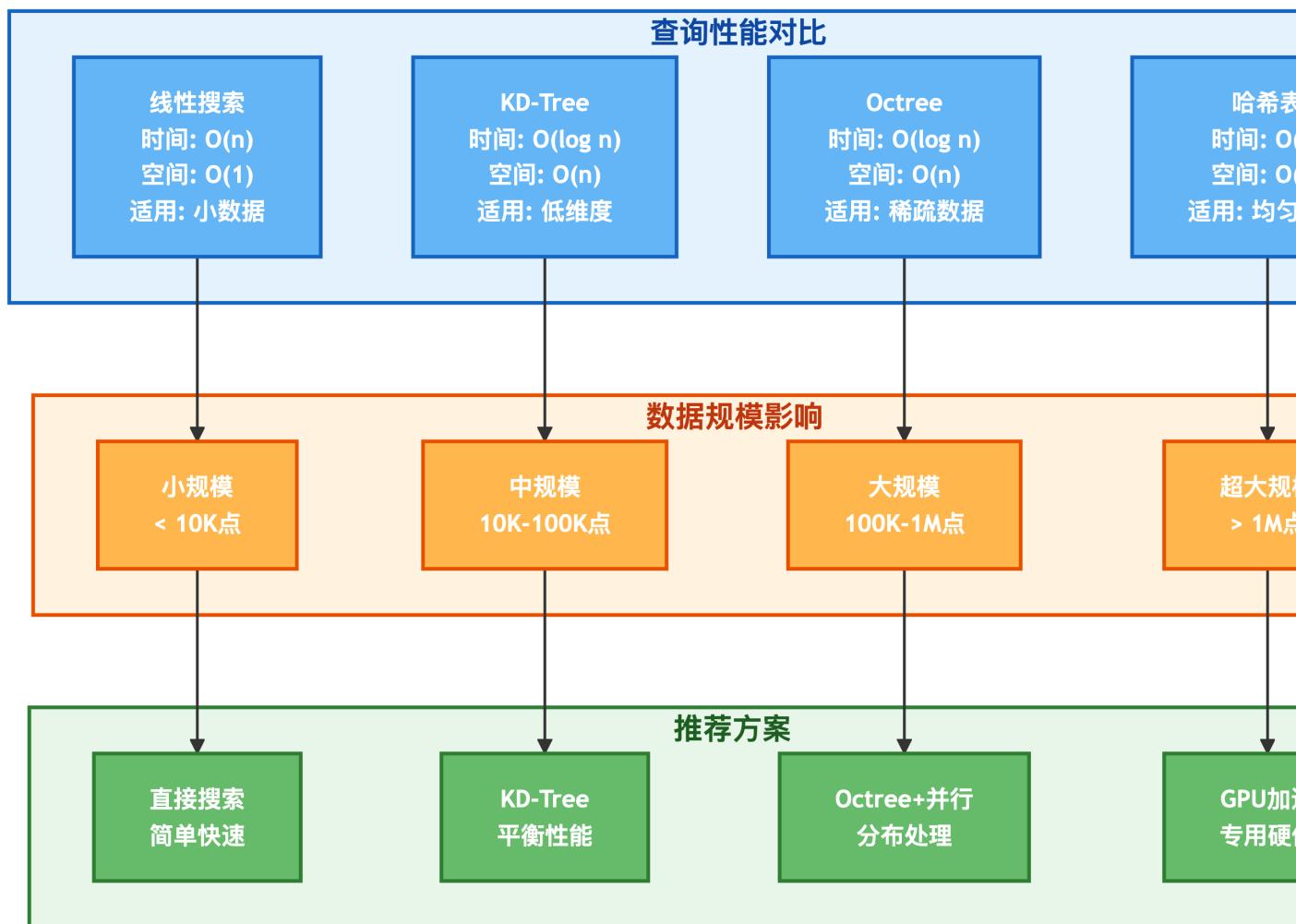


图11.19：不同空间索引结构的性能对比与适用场景

### 5.5.2

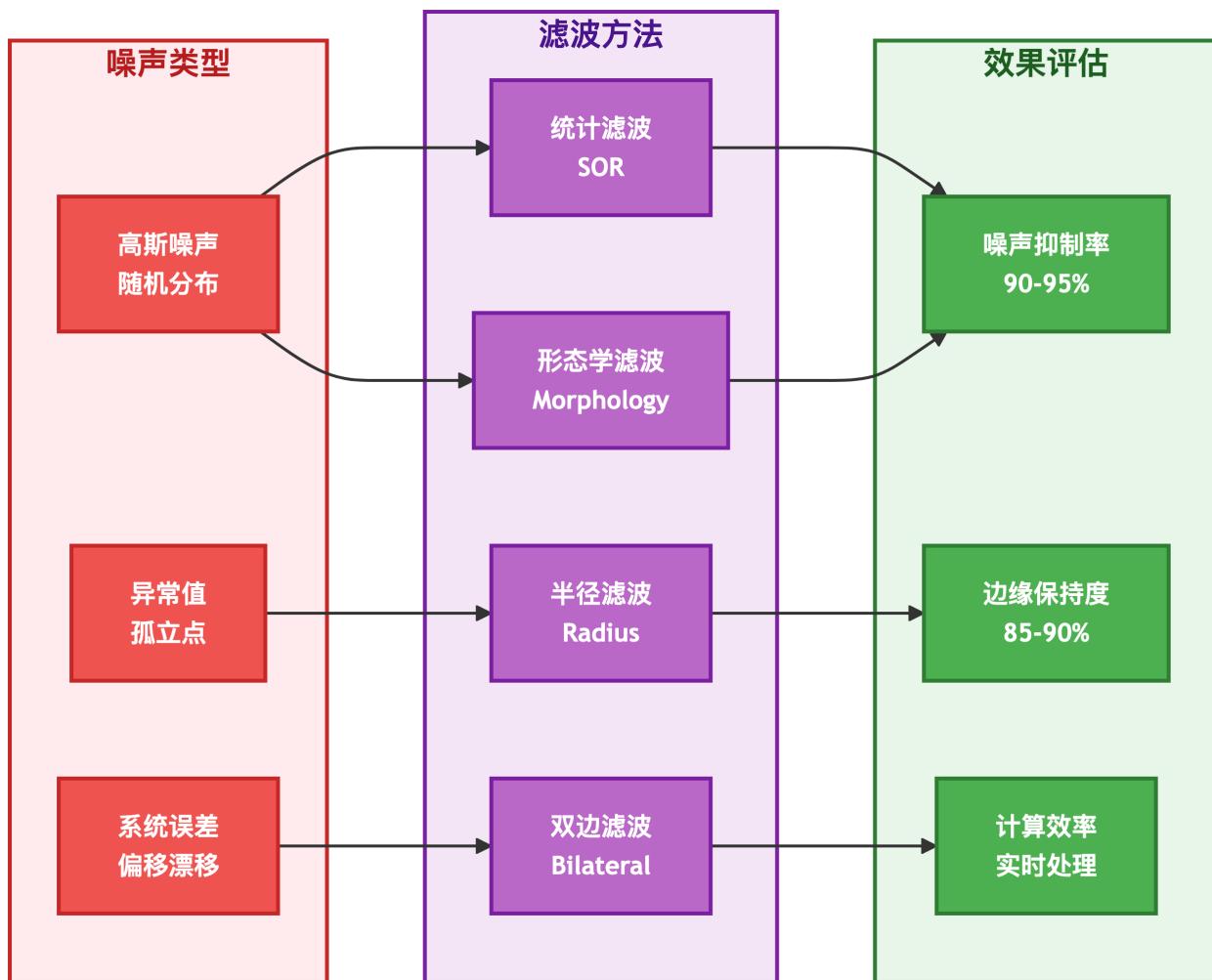


图11.20：不同滤波算法对各类噪声的处理效果

### 5.5.3

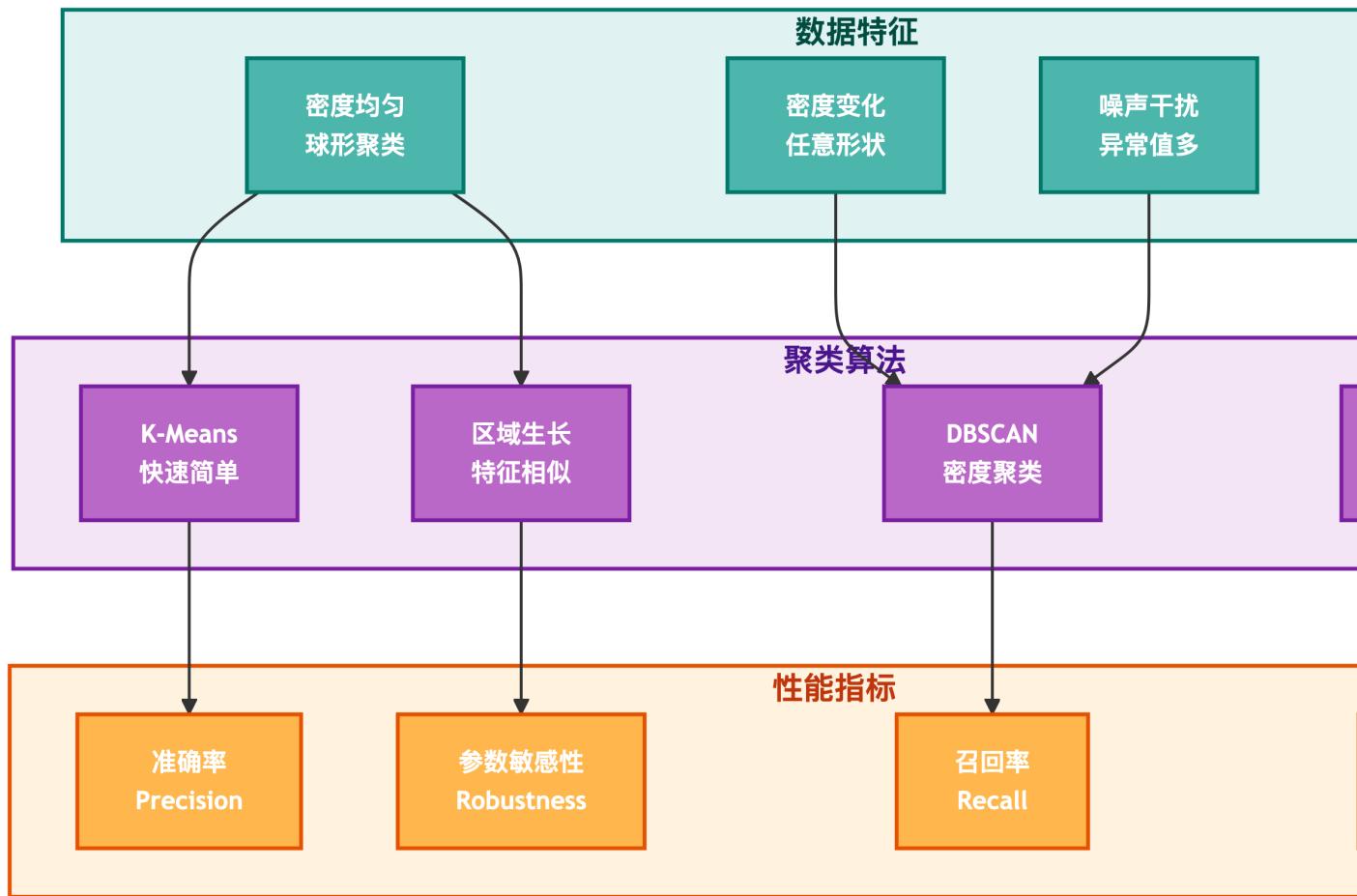


图11.21：聚类算法在不同数据特征下的适应性分析

#### 5.5.4

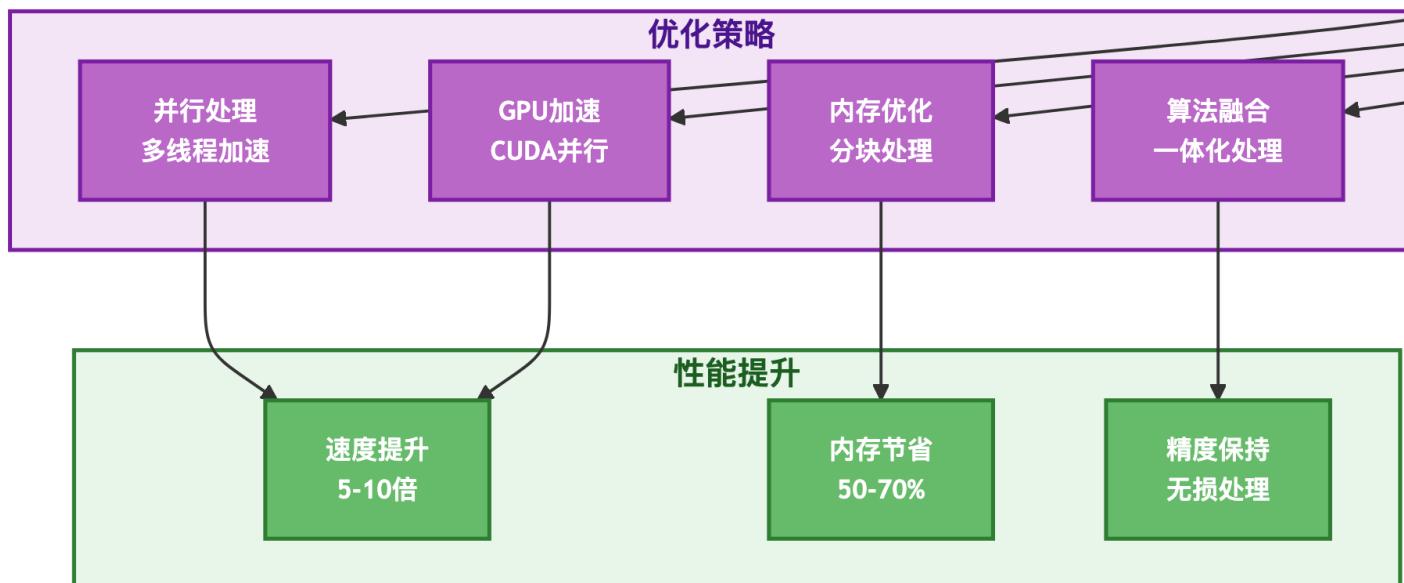


图11.22：点云处理流程的优化策略与性能提升

## 5.6

点云基础与处理是三维视觉技术栈的重要组成部分，为后续的高级分析和深度学习方法提供了坚实的基础。本节系统介绍了点云数据结构、空间索引、滤波处理和聚类分析等核心技术。

本节的核心贡献在于：**理论层面**，阐述了KD-Tree、体素化、统计滤波等算法的数学原理；**技术层面**，提供了高效的算法实现和优化策略；**应用层面**，分析了不同算法在各类场景中的适应性和性能表现。

点云处理技术与前面章节形成了完整的技术链条：相机标定提供了几何参数，立体匹配和三维重建生成了点云数据，而点云处理则对这些数据进行清理、组织和分析。这些基础处理技术为现代深度学习方法（如PointNet系列）奠定了重要基础，使得神经网络能够更好地理解并处理三维几何信息。

随着激光雷达、深度相机等传感器技术的发展，点云数据的规模和复杂度不断增加。未来的点云处理技术将朝着更高效率、更强鲁棒性、更智能化的方向发展，在自动驾驶、机器人、数字孪生等应用中发挥越来越重要的作用。

# 6 PointNet

## 6.1

传统的点云处理方法主要依赖手工设计的几何特征和统计分析，虽然在特定场景下表现良好，但面临着特征表达能力有限、泛化性能不足等问题。2017年，斯坦福大学的Charles Qi等人提出了PointNet网络，首次实现了直接在无序点云上进行深度学习，开启了点云深度学习的新时代。

PointNet的核心创新在于解决了点云数据的无序性和置换不变性问题。与图像的规则网格结构不同，点云中的点没有固定的排列顺序，传统的卷积神经网络无法直接应用。PointNet通过设计对称函数（如max pooling）来聚合点特征，确保网络输出不受点的排列顺序影响。

随着研究的深入，PointNet系列网络不断演进：**PointNet++**引入了层次化特征学习，能够捕获局部几何结构；**Point-Transformer**则将Transformer架构引入点云处理，通过自注意力机制实现更强的特征表达能力。这些网络的发展不仅推动了点云分类、分割等基础任务的性能提升，也为三维目标检测、场景理解等高级应用奠定了基础。

本节将系统介绍PointNet系列网络的核心思想、技术演进和实现细节，重点阐述这些网络如何突破传统方法的局限性，实现端到端的点云特征学习。

## 6.2

**对称函数与置换不变性**是PointNet系列网络的核心设计原则。点云数据的一个重要特性是其无序性：同一个物体的点云可以有多种不同的点排列方式，但它们应该被识别为同一个物体。这要求网络具有置换不变性，即对于点集 $\{p_1, p_2, \dots, p_n\}$ 的任意排列 $\{p_{\sigma(1)}, p_{\sigma(2)}, \dots, p_{\sigma(n)}\}$ ，网络的输出应该保持不变。

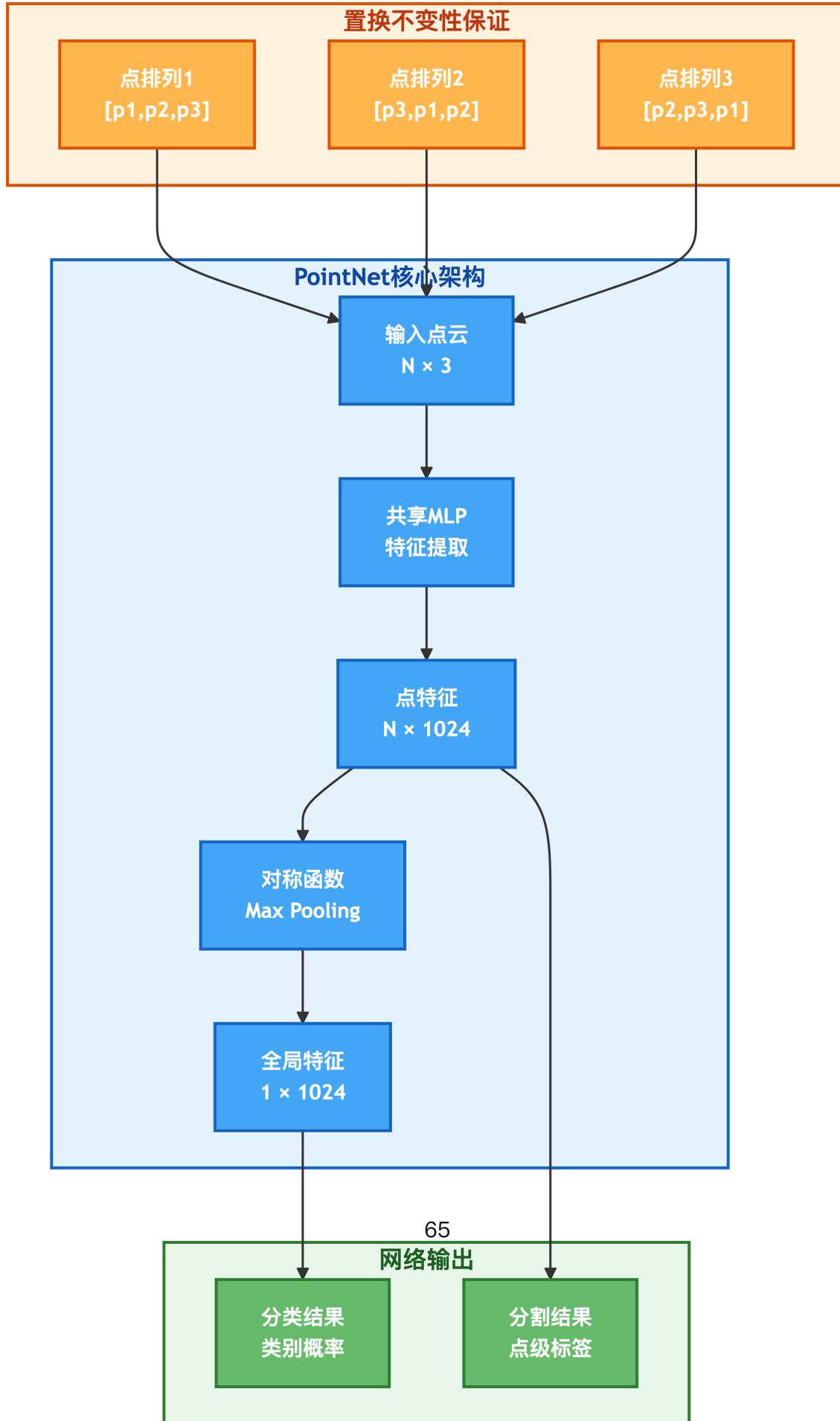


图11.23: PointNet网络的核心架构与置换不变性设计

层次化特征学习是PointNet++的重要创新。PointNet虽然能够提取全局特征，但缺乏对局部几何结构的建模能力。PointNet++通过引入Set Abstraction层，实现了类似CNN中的层次化特征学习：

- 采样层 (Sampling)：使用最远点采样 (FPS) 选择代表性点
- 分组层 (Grouping)：在每个采样点周围构建局部邻域
- 特征提取层 (PointNet)：对每个局部邻域应用PointNet提取特征

自注意力机制是Point-Transformer的核心技术。受Transformer在自然语言处理和计算机视觉领域成功的启发，Point-Transformer将自注意力机制引入点云处理，能够建模长距离依赖关系和复杂的几何结构。

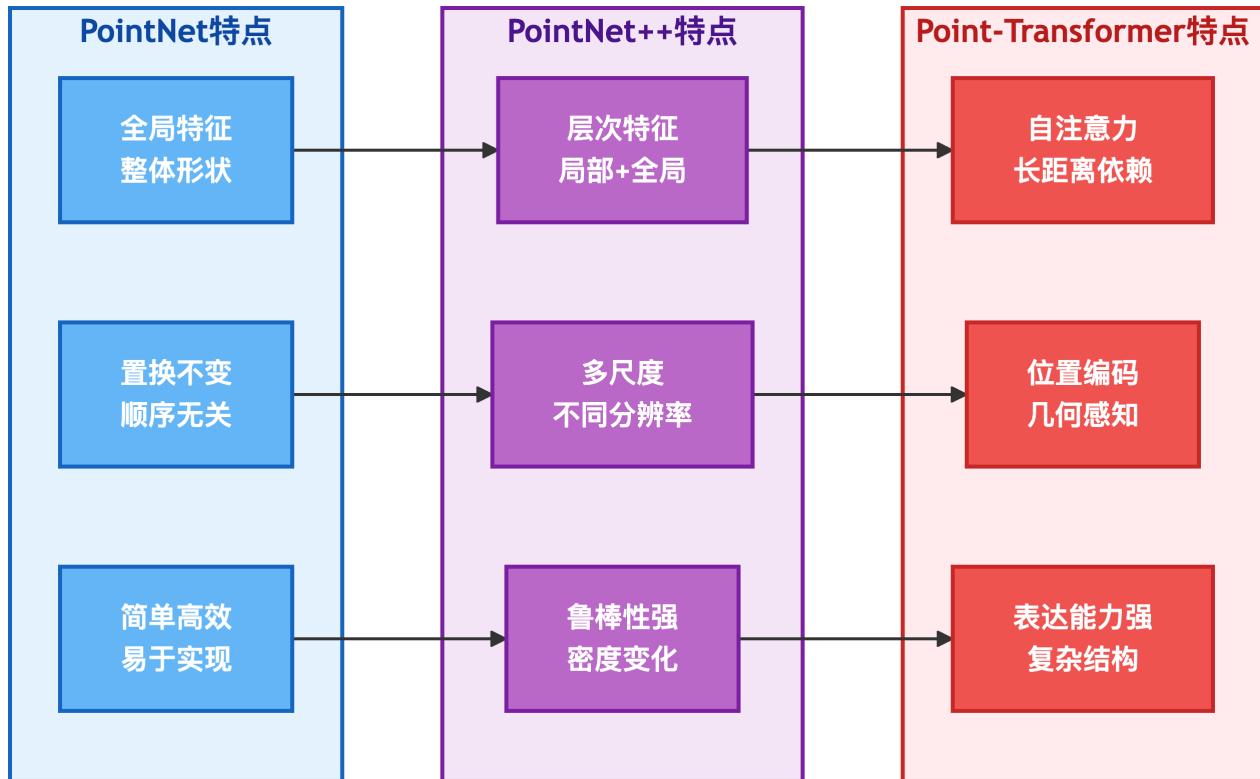


图11.24: PointNet系列网络的技术演进与特点对比

### 6.2.1 PointNet

**问题背景：**传统的深度学习方法主要针对规则数据结构设计，如图像的网格结构、序列的时序结构。然而，点云数据具有三个独特挑战：  
1. **无序性：**点云中点的排列顺序是任意的，不

存在固定的邻域关系 2. 置换不变性：网络输出必须对点的重新排列保持不变 3. 几何变换敏感性：点云容易受到旋转、平移等几何变换的影响

**创新突破：** PointNet通过三个关键创新解决了上述挑战： 1. 对称函数设计：使用max pooling等对称函数实现置换不变性，确保网络输出不受点顺序影响 2. T-Net变换网络：学习输入和特征的几何变换，提高对旋转、平移的鲁棒性 3. 理论保证：证明了任何连续的置换不变函数都可以用PointNet的形式近似表示

**技术特点：** – 端到端学习：直接从原始点云学习特征，无需手工设计特征 – 统一架构：同一网络可用于分类、分割等多种任务 – 计算高效：相比体素化方法，避免了稀疏数据的存储和计算开销

## PointNet详细架构

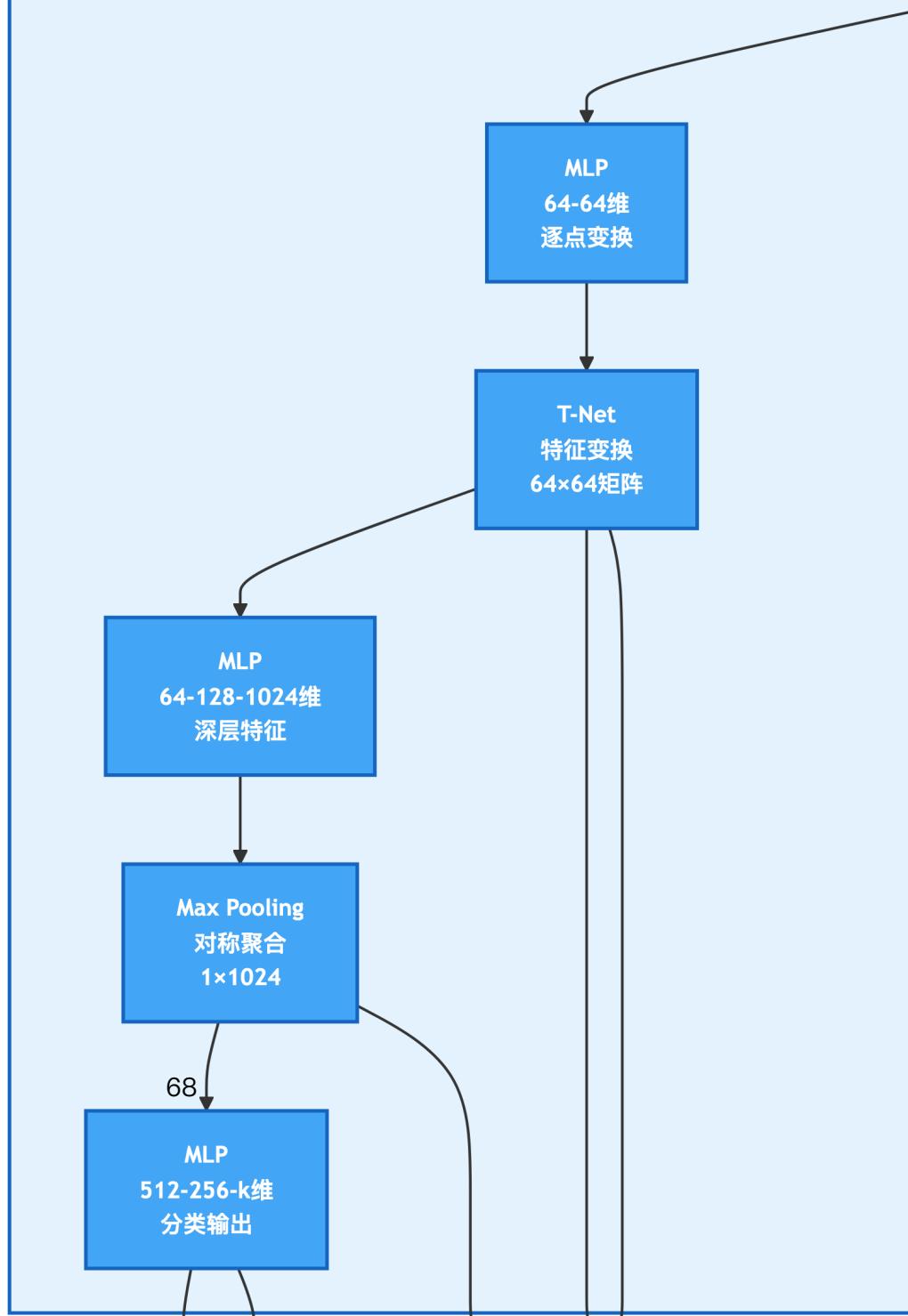


图11.24a：PointNet网络的详细架构与关键创新点

## 6.3

PointNet系列网络的理论基础涉及对称函数理论、层次化表示学习和注意力机制。下面我们将详细介绍这些核心理论。

### 6.3.1 PointNet

#### 1. 对称函数与万能逼近定理

PointNet的核心思想是使用对称函数来处理无序点集。对于点集  $S = \{x_1, x_2, \dots, x_n\}$ , 其中  $x_i \in \mathbb{R}^d$ , 我们希望学习一个函数  $f : 2^{\mathbb{R}^d} \rightarrow \mathbb{R}^k$ , 使得对于  $S$  的任意排列  $\pi(S)$ , 都有  $f(S) = f(\pi(S))$ 。

PointNet将这个函数分解为：

$$f(\{x_1, \dots, x_n\}) = \rho \left( \max_{i=1, \dots, n} \{h(x_i)\} \right)$$

其中： –  $h : \mathbb{R}^d \rightarrow \mathbb{R}^K$  是一个多层感知机，对每个点独立应用 –  $\max$  是逐元素的最大值操作，保证置换不变性 –  $\rho : \mathbb{R}^K \rightarrow \mathbb{R}^k$  是另一个多层感知机，处理聚合后的特征

**理论保证：**Zaheer等人证明了，任何连续的置换不变函数都可以表示为上述形式，其中  $h$  和  $\rho$  是连续函数。这为PointNet的设计提供了理论依据。

#### 2. 变换网络 (T-Net)

为了提高网络对几何变换的鲁棒性，PointNet引入了变换网络T-Net，学习一个变换矩阵  $T \in \mathbb{R}^{k \times k}$ ：

$$T = \text{T-Net}(\{x_1, \dots, x_n\})$$

变换后的特征为：

$$x'_i = T \cdot h(x_i)$$

为了保证变换矩阵接近正交矩阵，添加了正则化项：

$$L_{reg} = \|I - TT^T\|_F^2$$

其中  $\|\cdot\|_F$  是Frobenius范数。

### 6.3.2 PointNet++

#### 1. 层次化特征学习

PointNet++的核心思想是构建层次化的点特征表示。设第 $l$ 层有 $N_l$ 个点，每个点 $p_i^{(l)}$ 有特征 $f_i^{(l)} \in \mathbb{R}^{C_l}$ 。

Set Abstraction层的数学表示为：

$$\{p_i^{(l+1)}, f_i^{(l+1)}\}_{i=1}^{N_{l+1}} = \text{SA}(\{p_i^{(l)}, f_i^{(l)}\}_{i=1}^{N_l})$$

具体包含三个步骤：

- 采样：使用最远点采样 (FPS) 选择 $N_{l+1}$ 个中心点
- 分组：对每个中心点 $p_i^{(l+1)}$ ，找到半径 $r$ 内的邻居点集合：

$$\mathcal{N}_i = \{j : \|p_j^{(l)} - p_i^{(l+1)}\| \leq r\}$$

- 特征聚合：对每个局部区域应用PointNet：

$$f_i^{(l+1)} = \max_{j \in \mathcal{N}_i} \{h(p_j^{(l)} - p_i^{(l+1)}, f_j^{(l)})\}$$

#### 2. 多尺度分组

为了处理点云密度不均匀的问题，PointNet++采用多尺度分组策略：

$$f_i^{(l+1)} = \text{Concat}[f_i^{(l+1,1)}, f_i^{(l+1,2)}, \dots, f_i^{(l+1,M)}]$$

其中 $f_i^{(l+1,m)}$ 是在尺度 $m$ 下的特征，通过不同半径 $r_m$ 的分组得到。

### 6.3.3 Point-Transformer

#### 1. 自注意力机制

Point-Transformer将Transformer的自注意力机制扩展到点云数据。对于点 $i$ ，其更新后的特征为：

$$y_i = \sum_{j \in \mathcal{N}(i)} \alpha_{ij} (W_v x_j + \delta_{ij})$$

其中注意力权重 $\alpha_{ij}$ 计算为：

$$\alpha_{ij} = \text{softmax}_j(\phi(W_q x_i, W_k x_j + \delta_{ij}))$$

这里：  
 -  $W_q, W_k, W_v$ 是查询、键、值的线性变换矩阵  
 -  $\delta_{ij}$ 是位置编码，捕获点*i*和*j*之间的几何关系  
 -  $\phi$ 是位置编码函数，通常使用MLP实现

## 2. 位置编码

位置编码 $\delta_{ij}$ 对于点云处理至关重要，它编码了点之间的几何关系：

$$\delta_{ij} = \text{MLP}(p_i - p_j)$$

其中 $p_i - p_j$ 是两点之间的相对位置向量。

## 3. 向量注意力

为了更好地处理几何信息，Point-Transformer使用向量注意力：

$$\alpha_{ij} = \text{softmax}_j(\gamma(\psi(W_q x_i) - \psi(W_k x_j) + \delta_{ij}))$$

其中 $\gamma$ 和 $\psi$ 是非线性变换函数。

### 6.3.4

#### 1. 分类任务

对于点云分类，使用交叉熵损失：

$$L_{cls} = - \sum_{c=1}^C y_c \log(\hat{y}_c)$$

其中 $y_c$ 是真实标签， $\hat{y}_c$ 是预测概率。

#### 2. 分割任务

对于点云分割，对每个点计算交叉熵损失：

$$L_{seg} = -\frac{1}{N} \sum_{i=1}^N \sum_{c=1}^C y_{i,c} \log(\hat{y}_{i,c})$$

### 3. 正则化项

为了提高网络的泛化能力，通常添加正则化项：

$$L_{total} = L_{task} + \lambda_1 L_{reg} + \lambda_2 \|W\|_2^2$$

其中 $L_{task}$ 是任务相关的损失， $L_{reg}$ 是变换网络的正则化项， $\|W\|_2^2$ 是权重衰减项。

这些理论为PointNet系列网络的设计提供了坚实的数学基础，确保了网络能够有效处理点云数据的特殊性质。

## 6.4

下面我们介绍PointNet系列网络的核心算法实现，重点展示网络架构的关键组件和设计思想。

### 6.4.1 PointNet

PointNet的核心是通过共享MLP和对称函数实现置换不变性：

```
def pointnet_forward(x):
    """PointNet"""
    # 1.      T-Net 3x3
    trans_input = input_transform_net(x)  #
    x = apply_transformation(x, trans_input)

    # 2.      MLP
    x = shared_mlp(x)  # [B, N, 3] -> [B, N, 64]

    # 3.      T-Net 64x64
    trans_feat = feature_transform_net(x)  #
    x = apply_transformation(x, trans_feat)

    # 4.
    x = deep_shared_mlp(x)  # [B, N, 64] -> [B, N, 1024]

    # 5.
    global_feature = max_pooling(x)  # [B, N, 1024] -> [B, 1024]
```

```

# 6.
output = classification_mlp(global_feature)

return output, trans_feat

def t_net_core(x, k):
    """T-Net"""
    #     MLP +
    features = shared_mlp_layers(x)  # [B, N, k] -> [B, N, 1024]
    global_feat = max_pooling(features)  # [B, N, 1024] -> [B, 1024]

    #     MLP k×k
    transform_matrix = mlp_to_matrix(global_feat, k)  # [B, 1024] -> [B, k, k]

    #
    identity = torch.eye(k)
    transform_matrix = transform_matrix + identity

    return transform_matrix

def feature_transform_regularizer(trans_matrix):
    """
    # T^T * T - I Frobenius
    should_be_identity = torch.bmm(trans_matrix.transpose(2,1), trans_matrix)
    identity = torch.eye(trans_matrix.size(1))
    regularization_loss = torch.norm(should_be_identity - identity, dim=(1,2))
    return torch.mean(regularization_loss)

```

#### 6.4.2 PointNet++

PointNet++通过Set Abstraction层实现层次化特征学习：

```

def farthest_point_sample(xyz, npoint):
    """
    device = xyz.device
    B, N, C = xyz.shape
    centroids = torch.zeros(B, npoint, dtype=torch.long).to(device)
    distance = torch.ones(B, N).to(device) * 1e10
    farthest = torch.randint(0, N, (B,), dtype=torch.long).to(device)
    batch_indices = torch.arange(B, dtype=torch.long).to(device)

```

```

for i in range(npoint):
    centroids[:, i] = farthest
    centroid = xyz[batch_indices, farthest, :].view(B, 1, 3)
    dist = torch.sum((xyz - centroid) ** 2, -1)
    mask = dist < distance
    distance[mask] = dist[mask]
    farthest = torch.max(distance, -1)[1]

return centroids

def query_ball_point(radius, nsample, xyz, new_xyz):
    """
    device = xyz.device
    B, N, C = xyz.shape
    _, S, _ = new_xyz.shape
    group_idx = torch.arange(N, dtype=torch.long).to(device).view(1, 1, N).repeat([B, S, 1])

    sqrdists = square_distance(new_xyz, xyz)
    group_idx[sqrdists > radius ** 2] = N
    group_idx = group_idx.sort(dim=-1)[0][:, :, :nsample]
    group_first = group_idx[:, :, 0].view(B, S, 1).repeat([1, 1, nsample])
    mask = group_idx == N
    group_idx[mask] = group_first[mask]

    return group_idx

class SetAbstraction(nn.Module):
    """Set Abstraction PointNet++"""
    def __init__(self, npoint, radius, nsample, in_channel, mlp, group_all):
        super(SetAbstraction, self).__init__()
        self.npoint = npoint
        self.radius = radius
        self.nsample = nsample
        self.mlp_convs = nn.ModuleList()
        self.mlp_bns = nn.ModuleList()

        last_channel = in_channel
        for out_channel in mlp:
            self.mlp_convs.append(nn.Conv2d(last_channel, out_channel, 1))
            self.mlp_bns.append(nn.BatchNorm2d(out_channel))
            last_channel = out_channel

```

```

        self.group_all = group_all

    def forward(self, xyz, points):
        """
        xyz: (B, N, 3)
        points: (B, N, C)
        """
        xyz = xyz.permute(0, 2, 1)
        if points is not None:
            points = points.permute(0, 2, 1)

        if self.group_all:
            new_xyz, new_points = sample_and_group_all(xyz, points)
        else:
            new_xyz, new_points = sample_and_group(
                self.npoint, self.radius, self.nsample, xyz, points)

        # PointNet
        new_points = new_points.permute(0, 3, 2, 1) # [B, C+D, nsample, npoint]
        for i, conv in enumerate(self.mlp_convs):
            bn = self.mlp_bns[i]
            new_points = F.relu(bn(conv(new_points)))

        #
        new_points = torch.max(new_points, 2)[0]
        new_xyz = new_xyz.permute(0, 2, 1)
        return new_xyz, new_points

class PointNetPlusPlus(nn.Module):
    """PointNet++"""
    def __init__(self, num_classes):
        super(PointNetPlusPlus, self).__init__()

        #
        self.sa1 = SetAbstraction(512, 0.2, 32, 3, [64, 64, 128], False)
        self.sa2 = SetAbstraction(128, 0.4, 64, 128 + 3, [128, 128, 256], False)
        self.sa3 = SetAbstraction(None, None, None, 256 + 3, [256, 512, 1024], True)

        #
        self.fc1 = nn.Linear(1024, 512)
        self.bn1 = nn.BatchNorm1d(512)
        self.drop1 = nn.Dropout(0.4)

```

```

        self.fc2 = nn.Linear(512, 256)
        self.bn2 = nn.BatchNorm1d(256)
        self.drop2 = nn.Dropout(0.4)
        self.fc3 = nn.Linear(256, num_classes)

    def forward(self, xyz):
        B, _, _ = xyz.shape

        #
        l1_xyz, l1_points = self.sa1(xyz, None)
        l2_xyz, l2_points = self.sa2(l1_xyz, l1_points)
        l3_xyz, l3_points = self.sa3(l2_xyz, l2_points)

        #
        x = l3_points.view(B, 1024)

        #
        x = self.drop1(F.relu(self.bn1(self.fc1(x))))
        x = self.drop2(F.relu(self.bn2(self.fc2(x))))
        x = self.fc3(x)

    return F.log_softmax(x, -1)

```

### 6.4.3 Point-Transformer

Point-Transformer引入自注意力机制处理点云：

```

class PointTransformerLayer(nn.Module):
    """Point-Transformer"""
    def __init__(self, in_planes, out_planes=None):
        super(PointTransformerLayer, self).__init__()
        self.in_planes = in_planes
        self.out_planes = out_planes or in_planes

        #
        self.q_conv = nn.Conv1d(in_planes, in_planes, 1, bias=False)
        self.k_conv = nn.Conv1d(in_planes, in_planes, 1, bias=False)
        self.v_conv = nn.Conv1d(in_planes, self.out_planes, 1)

        #

```

```

    self.pos_mlp = nn.Sequential(
        nn.Conv2d(3, in_planes, 1, bias=False),
        nn.BatchNorm2d(in_planes),
        nn.ReLU(inplace=True),
        nn.Conv2d(in_planes, in_planes, 1)
    )

    #
    self.attn_mlp = nn.Sequential(
        nn.Conv2d(in_planes, in_planes, 1, bias=False),
        nn.BatchNorm2d(in_planes),
        nn.ReLU(inplace=True),
        nn.Conv2d(in_planes, in_planes, 1)
    )

    self.softmax = nn.Softmax(dim=-1)

def forward(self, xyz, features, neighbor_idx):
    """
    xyz: (B, N, 3)
    features: (B, C, N)
    neighbor_idx: (B, N, K)
    """
    B, C, N = features.shape
    _, _, K = neighbor_idx.shape

    #
    q = self.q_conv(features) # (B, C, N)
    k = self.k_conv(features) # (B, C, N)
    v = self.v_conv(features) # (B, C', N)

    #
    k_neighbors = index_points(k.transpose(1, 2), neighbor_idx) # (B, N, K, C)
    v_neighbors = index_points(v.transpose(1, 2), neighbor_idx) # (B, N, K, C')

    #
    xyz_neighbors = index_points(xyz, neighbor_idx) # (B, N, K, 3)
    relative_pos = xyz.unsqueeze(2) - xyz_neighbors # (B, N, K, 3)

    #
    pos_encoding = self.pos_mlp(relative_pos.permute(0, 3, 1, 2)) # (B, C, N, K)
    pos_encoding = pos_encoding.permute(0, 2, 3, 1) # (B, N, K, C)

```

```
#  
q_expanded = q.transpose(1, 2).unsqueeze(2) # (B, N, 1, C)  
attention_input = q_expanded - k_neighbors + pos_encoding # (B, N, K, C)  
attention_weights = self.attn_mlp(attention_input.permute(0, 3, 1, 2)) # (B, C, N, K)  
attention_weights = self.softmax(attention_weights.permute(0, 2, 3, 1)) # (B, N, K, 1)  
  
#  
output = torch.sum(attention_weights * (v_neighbors + pos_encoding), dim=2) # (B, N, C)  
  
return output.transpose(1, 2) # (B, C', N)
```

这些核心实现展示了PointNet系列网络的关键设计思想：PointNet通过对称函数保证置换不变性，PointNet++通过层次化采样捕获局部结构，Point-Transformer通过自注意力机制建模长距离依赖关系。

## 6.5

PointNet系列网络在多个点云处理任务上取得了显著的性能提升，推动了整个领域的发

### 6.5.1

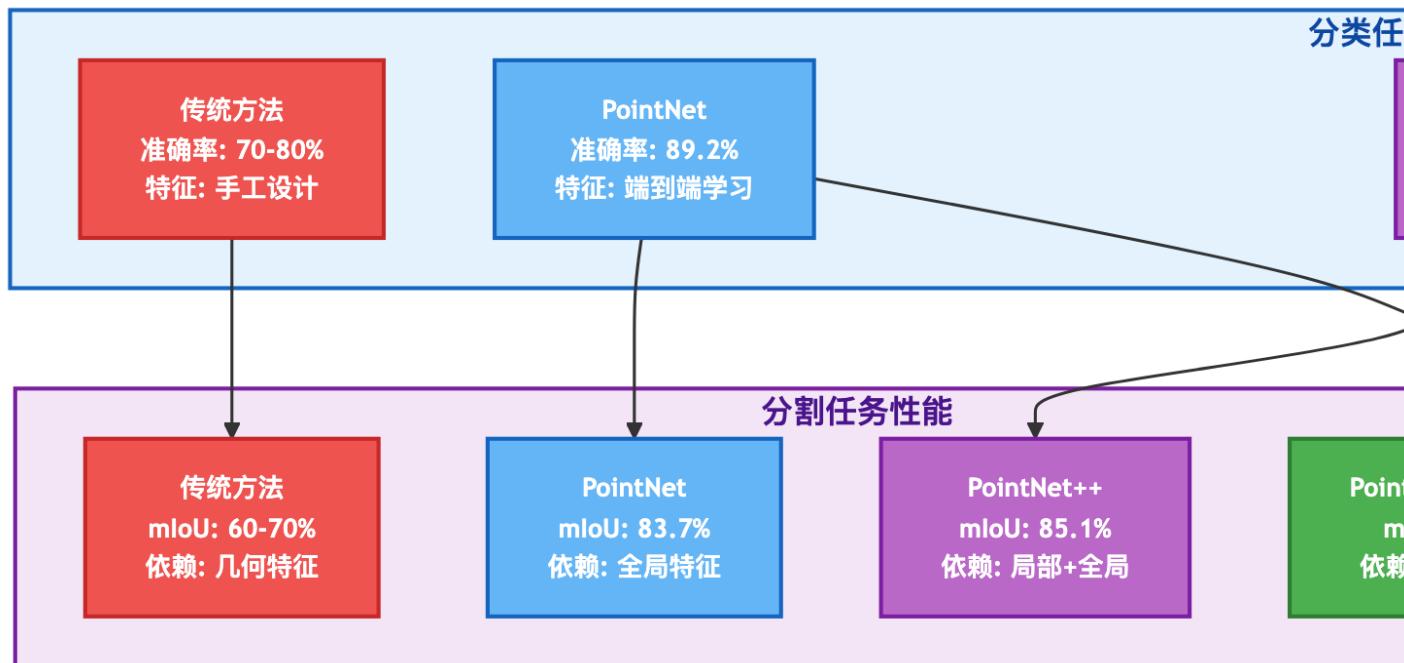


图11.25: PointNet系列网络在不同任务上的性能对比

### 6.5.2

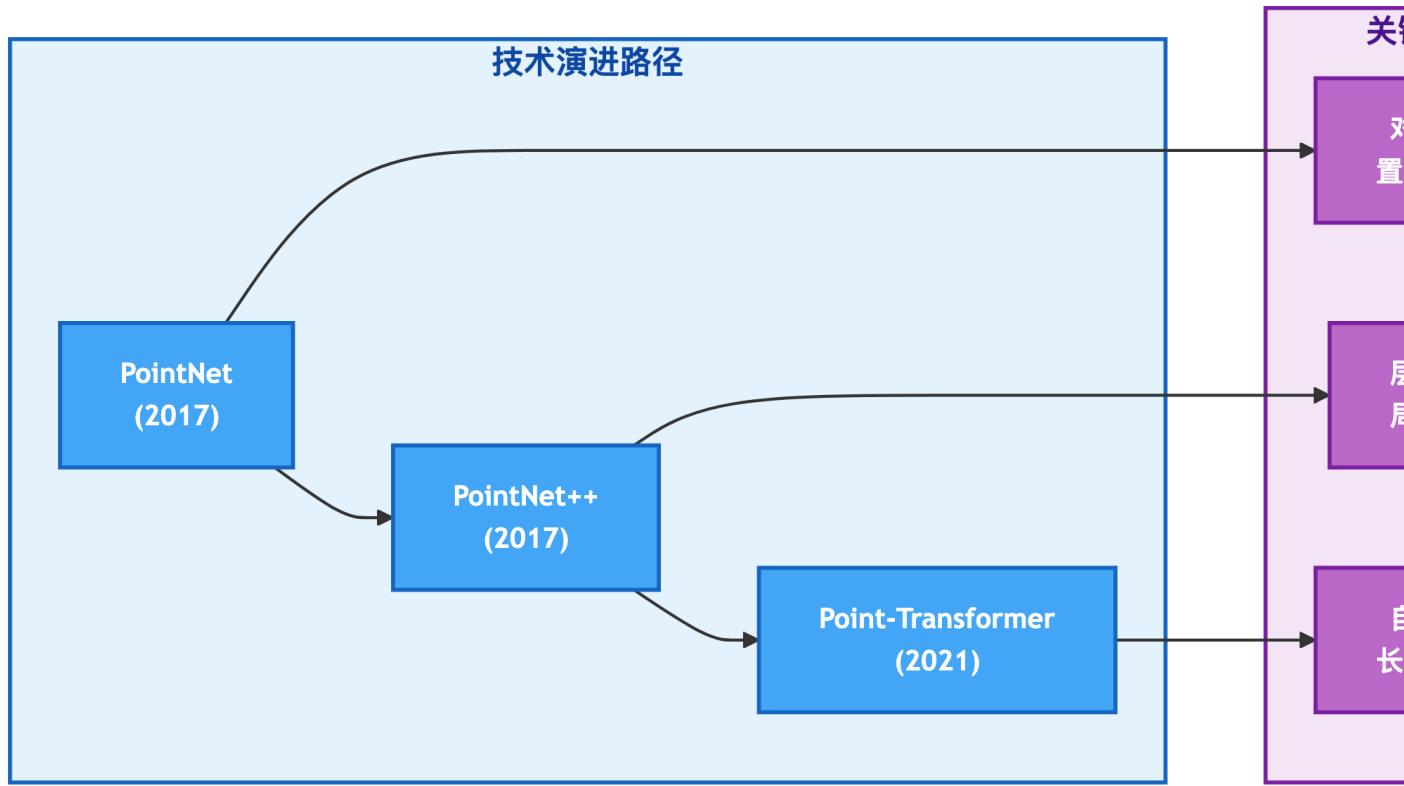


图11.26：PointNet系列网络的技术演进与应用拓展

### 6.5.3

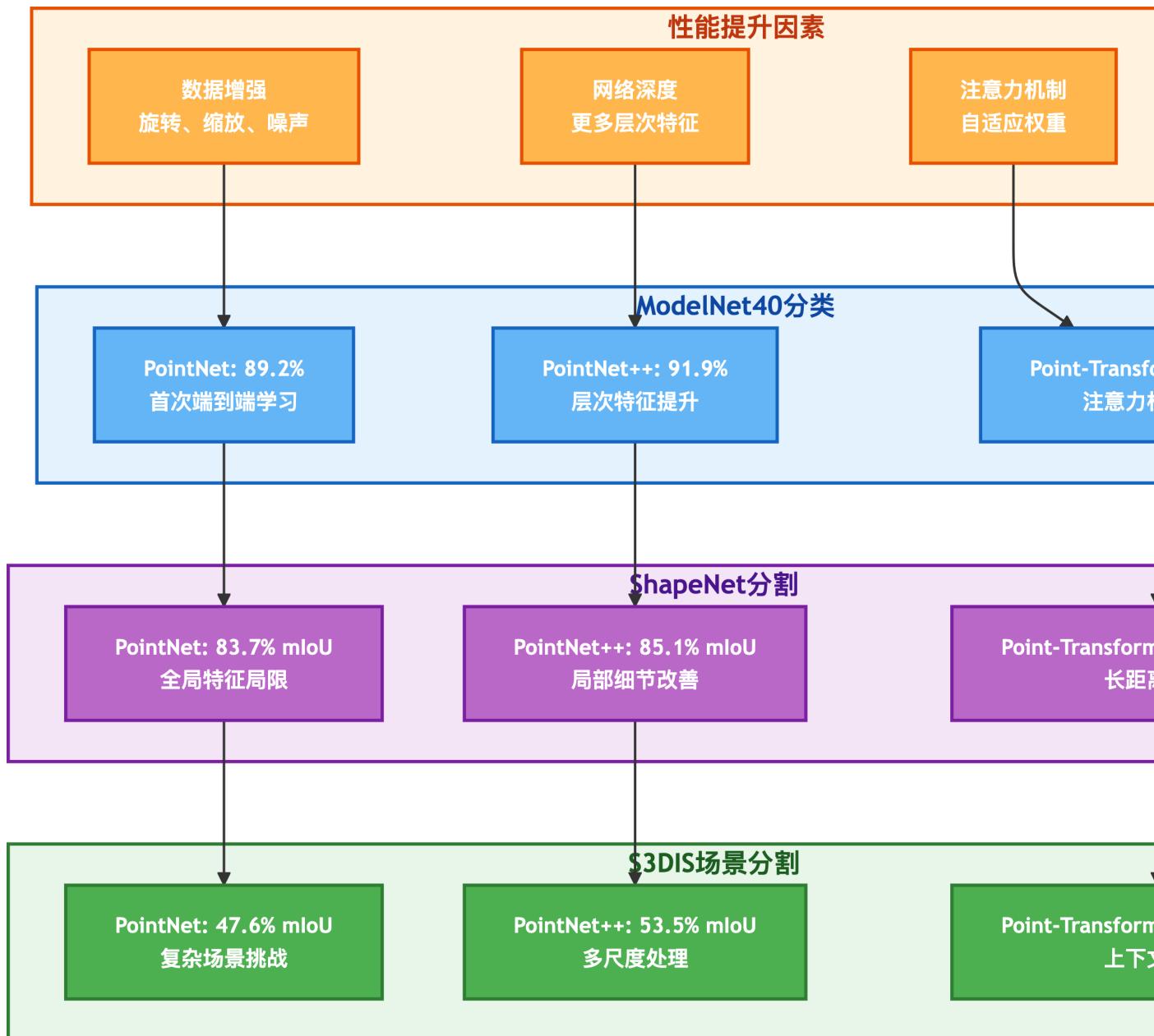


图11.27: PointNet系列网络在主要数据集上的性能基准

#### 6.5.4

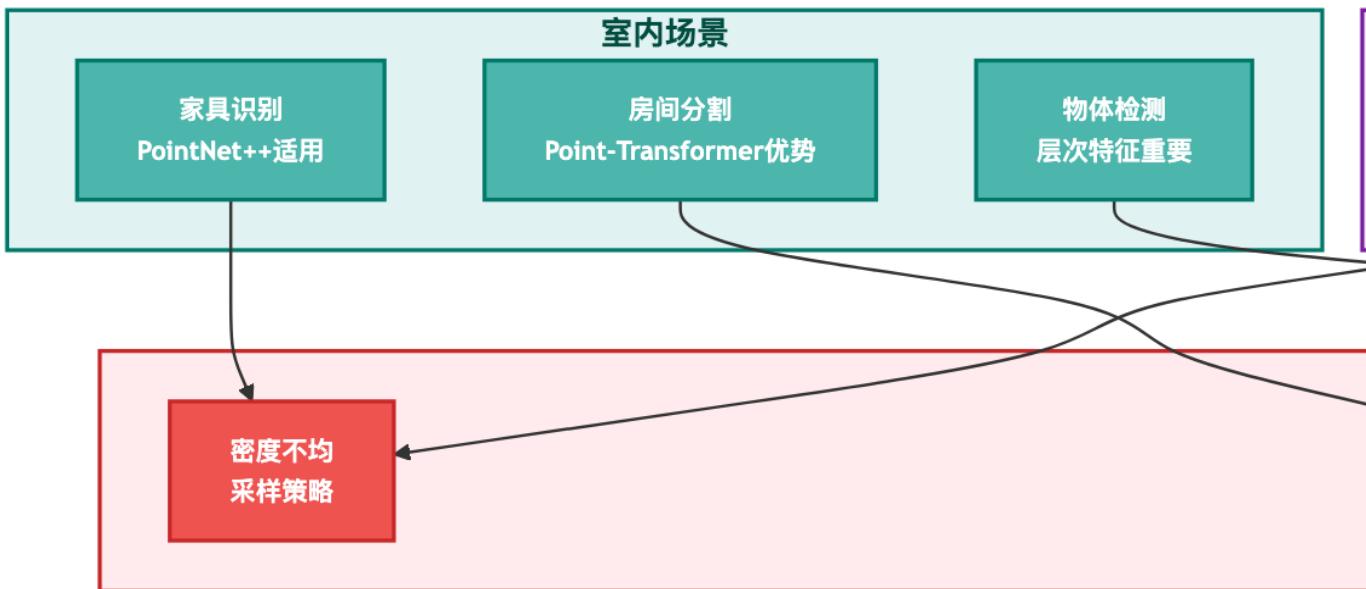


图11.28：PointNet系列网络在不同应用场景中的适应性与挑战

## 6.6

PointNet系列网络代表了点云深度学习的重要里程碑，从根本上改变了点云处理的技术范式。本节系统介绍了从PointNet到Point-Transformer的技术演进，展示了深度学习在点云处理中的革命性突破。

本节的核心贡献在于：**理论层面**，阐述了对称函数、层次化表示学习和自注意力机制的数学原理；**技术层面**，详细分析了网络架构的设计思想和关键组件；**应用层面**，展示了这些网络在分类、分割等任务上的性能提升和应用潜力。

PointNet系列网络与前面章节形成了完整的技术链条：传统点云处理方法提供了数据预处理和特征工程的基础，而深度学习方法则实现了端到端的特征学习和任务优化。这种技术演进不仅提升了点云处理的性能，也为三维目标检测、场景理解等高级应用奠定了基础。

随着Transformer架构在计算机视觉领域的成功应用，点云深度学习正朝着更强的表达能力、更好的泛化性能和更高的计算效率方向发展。未来的研究将继续探索新的网络架构、训练策略和应用场景，推动三维视觉技术在自动驾驶、机器人、数字孪生等领域的广泛应用。

# 7 3D

## 7.1 2D 3D

3D目标检测是计算机视觉和自动驾驶领域的核心任务之一，它要求系统不仅能够识别物体的类别，还要准确估计物体在三维空间中的位置、尺寸和朝向。与传统的2D目标检测相比，3D检测面临着更大的挑战：三维空间的复杂性、点云数据的稀疏性、以及对精确几何信息的严格要求。

传统的3D目标检测方法主要依赖手工设计的特征和几何约束，如基于滑动窗口的方法和基于模板匹配的方法。这些方法虽然在特定场景下有效，但泛化能力有限，难以处理复杂的真实世界场景。深度学习的兴起，特别是PointNet系列网络的成功，为3D目标检测带来了革命性的变化。

现代3D目标检测方法可以分为几个主要类别：**基于体素的方法**（如VoxelNet）将点云转换为规则的3D网格，利用3D卷积进行特征提取；**基于柱状投影的方法**（如PointPillars）将点云投影到鸟瞰图，结合2D卷积的效率优势；**点-体素融合方法**（如PV-RCNN）则结合了点表示和体素表示的优势，实现更精确的检测。

这些方法的发展不仅推动了学术研究的进步，也在自动驾驶、机器人导航、智能监控等实际应用中发挥着关键作用。本节将系统介绍3D目标检测的核心技术、算法原理和实现细节，展示这一领域的最新进展。

## 7.2

**3D边界框表示**是3D目标检测的基础。与2D检测中的矩形框不同，3D边界框需要表示物体在三维空间中的完整几何信息。常用的3D边界框表示包括：

- **中心点表示**:  $(x, y, z, l, w, h, \theta)$ , 其中 $(x, y, z)$ 是中心坐标,  $(l, w, h)$ 是长宽高,  $\theta$ 是朝向角
- **角点表示**: 使用8个角点的3D坐标来完全描述边界框
- **参数化表示**: 结合物体的几何先验, 使用更紧凑的参数表示

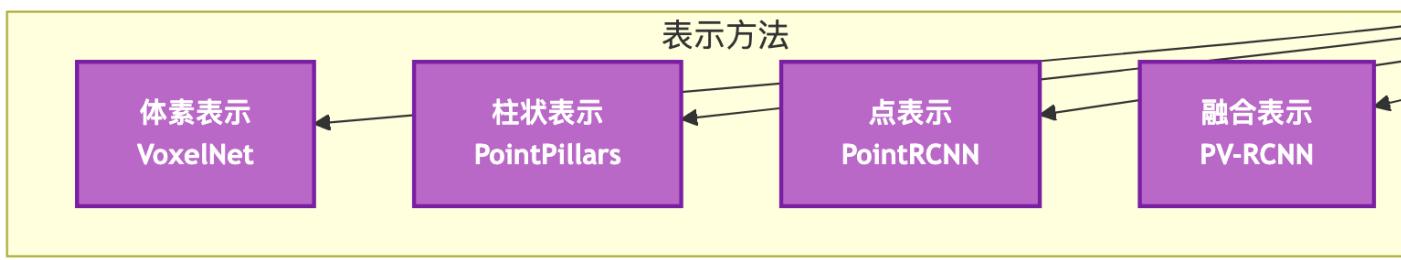


图11.29：3D目标检测的数据流程与表示方法

锚框机制在3D检测中发挥重要作用。与2D检测类似，3D检测也使用预定义的锚框来简化检测问题。3D锚框的设计需要考虑：

- 尺寸先验：根据不同类别物体的典型尺寸设计锚框
- 朝向先验：考虑物体的常见朝向，如车辆通常沿道路方向
- 密度分布：在可能出现物体的区域密集放置锚框

多模态融合是提高3D检测性能的重要策略。现代自动驾驶系统通常配备多种传感器：

- LiDAR点云：提供精确的几何信息和距离测量
- RGB图像：提供丰富的纹理和语义信息
- 雷达数据：提供速度信息和恶劣天气下的鲁棒性

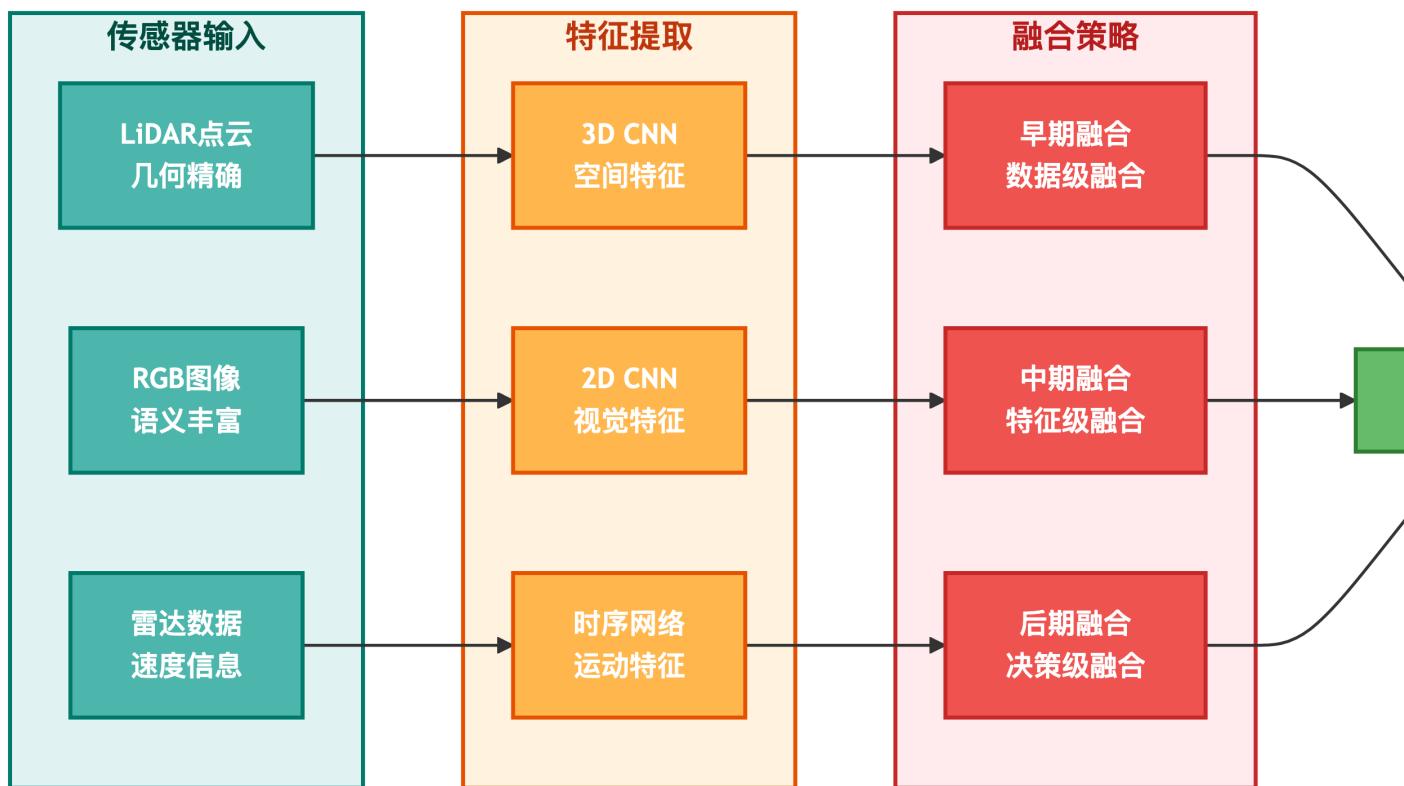


图11.30：多模态传感器融合在3D目标检测中的应用

## 7.3

-

3D目标检测的理论基础涉及三维数据表示、深度网络架构设计和损失函数优化。下面我们详细介绍这些核心理论。

### 7.3.1 VoxelNet

**VoxelNet的创新突破：** VoxelNet是首个端到端的3D目标检测网络，解决了点云数据在深度学习中的三个关键挑战： 1. **不规则性问题**：点云数据稀疏且不规则，传统CNN无法直接处理  
2. **特征学习问题**：如何从原始点云中学习有效的特征表示 3. **端到端优化**：如何实现从点云到检测结果的端到端训练

**技术创新：** – **体素化表示**：将不规则点云转换为规则的3D网格，使CNN可以处理 – **VFE层设计**：体素特征编码层，将体素内的点集转换为固定维度特征 – **3D卷积骨干**：使用3D CNN提取空间特征，保持三维几何信息

原始点云

 $N \times 4 (x, y, z, 1)$ 

图11.30a: VoxelNet网络架构与VFE层设计

### 7.3.2 VoxelNet

#### 1. 体素化表示

VoxelNet将不规则的点云数据转换为规则的3D体素网格。给定点云 $P = \{p_i\}_{i=1}^N$ , 其中 $p_i = (x_i, y_i, z_i, r_i)$ 包含坐标和反射强度, 体素化过程将3D空间划分为 $D \times H \times W$ 的网格。

每个体素 $V_{d,h,w}$ 包含落入其中的点集:

$$V_{d,h,w} = \{p_i \in P : \lfloor \frac{x_i - x_{min}}{v_x} \rfloor = w, \lfloor \frac{y_i - y_{min}}{v_y} \rfloor = h, \lfloor \frac{z_i - z_{min}}{v_z} \rfloor = d\}$$

其中 $(v_x, v_y, v_z)$ 是体素的尺寸。

#### 2. 体素特征编码 (VFE)

VoxelNet的核心创新是体素特征编码层, 它将体素内的点集转换为固定维度的特征向量。对于包含 $T$ 个点的体素, VFE层的计算过程为:

- **点特征增强:** 为每个点添加相对于体素中心的偏移量

$$\tilde{p}_i = [x_i, y_i, z_i, r_i, x_i - v_x, y_i - v_y, z_i - v_z]$$

- **逐点特征变换:** 使用全连接层提取点特征

$$f_i = \text{FCN}(\tilde{p}_i)$$

- **局部聚合:** 使用最大池化聚合体素内所有点的特征

$$f_{voxel} = \max_{i=1, \dots, T} f_i$$

#### 3. 3D卷积骨干网络

体素特征经过3D CNN进行层次化特征提取:

$$F^{(l+1)} = \text{Conv3D}(\text{BN}(\text{ReLU}(F^{(l)})))$$

其中 $F^{(l)}$ 是第 $l$ 层的特征图。

### 7.3.3 PointPillars

#### 1. 柱状投影

PointPillars将3D点云投影到2D鸟瞰图 (Bird's Eye View, BEV)，将垂直方向的信息编码到特征中。点云被划分为 $H \times W$ 的柱状网格，每个柱子包含垂直方向上的所有点。

#### 2. 柱状特征编码

对于柱子 $(i, j)$ 中的点集 $\{p_k\}$ ，PointPillars计算增强特征：

$$\tilde{p}_k = [x_k, y_k, z_k, r_k, x_k - x_c, y_k - y_c, z_k - z_c, x_k - x_p, y_k - y_p]$$

其中 $(x_c, y_c, z_c)$ 是柱子中所有点的质心， $(x_p, y_p)$ 是柱子的几何中心。

柱状特征通过PointNet-like网络提取：

$$f_{pillar} = \max_k \text{MLP}(\tilde{p}_k)$$

#### 3. 伪图像生成

柱状特征被重新排列为伪图像格式，然后使用2D CNN进行处理：

$$F_{BEV} = \text{CNN2D}(\text{Scatter}(f_{pillar}))$$

### 7.3.4 PV-RCNN

#### 1. 点-体素融合

PV-RCNN结合了点表示的精确性和体素表示的效率。网络包含两个并行分支：

- **体素分支**：使用3D稀疏卷积处理体素化点云
- **点分支**：使用PointNet++处理原始点云

#### 2. 体素到点的特征传播

体素特征通过三线性插值传播到点：

$$f_p = \sum_{v \in \mathcal{N}(p)} w(p, v) \cdot f_v$$

其中 $\mathcal{N}(p)$ 是点 $p$ 周围的8个体素， $w(p, v)$ 是插值权重。

#### 3. 关键点采样

PV-RCNN使用前景点分割网络识别关键点：

$$s_i = \text{MLP}(f_i^{point})$$

其中 $s_i$ 是点*i*的前景概率。选择前景概率最高的点作为关键点。

#### 4. ROI网格池化

对于每个候选区域，PV-RCNN在其内部规律采样网格点，并聚合周围点的特征：

$$f_{grid} = \text{Aggregate}(\{f_j : \|p_j - p_{grid}\| < r\})$$

### 7.3.5

#### 1. 分类损失

使用Focal Loss处理类别不平衡问题：

$$L_{cls} = -\alpha_t(1 - p_t)^\gamma \log(p_t)$$

其中 $p_t$ 是预测概率， $\alpha_t$ 和 $\gamma$ 是超参数。

#### 2. 回归损失

3D边界框回归使用Smooth L1损失：

$$L_{reg} = \sum_{i \in \{x, y, z, l, w, h, \theta\}} \text{SmoothL1}(\Delta_i)$$

其中 $\Delta_i$ 是预测值与真值的差异。

#### 3. 朝向损失

由于角度的周期性，朝向回归使用特殊的损失函数：

$$L_{dir} = \sum_{bin} \text{CrossEntropy}(cls_{bin}) + \sum_{bin} \text{SmoothL1}(res_{bin})$$

其中角度被分解为分类和回归两部分。

#### 4. 总损失

总损失是各项损失的加权和：

$$L_{total} = \lambda_{cls} L_{cls} + \lambda_{reg} L_{reg} + \lambda_{dir} L_{dir}$$

这些理论为现代3D目标检测算法提供了坚实的数学基础，确保了算法的有效性和可靠性。

## 7.4

下面我们介绍3D目标检测的核心算法实现，重点展示不同方法的关键组件和设计思想。

### 7.4.1 VoxelNet

VoxelNet通过体素特征编码和3D卷积实现端到端的3D检测：

```
import torch
import torch.nn as nn
import torch.nn.functional as F

class VoxelFeatureExtractor(nn.Module):
    """ VFE """
    def __init__(self, num_input_features=4):
        super(VoxelFeatureExtractor, self).__init__()
        self.num_input_features = num_input_features

        # VFE
        self.vfe1 = VFELayer(num_input_features, 32)
        self.vfe2 = VFELayer(32, 128)

    def forward(self, features, num_voxels, coords):
        """
        features: (N, max_points, num_features)
        num_voxels: (N,)
        coords: (N, 3)
        """
        # VFE
        voxel_features = self.vfe1(features, num_voxels)
        voxel_features = self.vfe2(voxel_features, num_voxels)

        return voxel_features

class VFELayer(nn.Module):
    """ VFE """
    def __init__(self, in_channels, out_channels):
        super(VFELayer, self).__init__()
        self.in_channels = in_channels
        self.out_channels = out_channels
```

```

#
self.linear = nn.Linear(in_channels, out_channels)
self.norm = nn.BatchNorm1d(out_channels)

def forward(self, inputs, num_voxels):
    # inputs: (N, max_points, in_channels)
    N, max_points, _ = inputs.shape

    #
    x = inputs.view(-1, self.in_channels)
    x = F.relu(self.norm(self.linear(x)))
    x = x.view(N, max_points, self.out_channels)

    #
    voxel_features = torch.max(x, dim=1)[0] # (N, out_channels)

    return voxel_features

class MiddleExtractor(nn.Module):
    """3D"""
    def __init__(self, input_channels=128):
        super(MiddleExtractor, self).__init__()

        # 3D
        self.conv3d1 = nn.Conv3d(input_channels, 64, 3, padding=1)
        self.conv3d2 = nn.Conv3d(64, 64, 3, padding=1)
        self.conv3d3 = nn.Conv3d(64, 64, 3, padding=1)

        #
        self.bn1 = nn.BatchNorm3d(64)
        self.bn2 = nn.BatchNorm3d(64)
        self.bn3 = nn.BatchNorm3d(64)

    def forward(self, voxel_features, coords, batch_size, input_shape):
        """
        """
        #
        device = voxel_features.device
        sparse_shape = input_shape
        dense_features = torch.zeros(
            batch_size, self.conv3d1.in_channels, *sparse_shape,

```

```

        dtype=voxel_features.dtype, device=device)

    #
dense_features[coords[:, 0], :, coords[:, 1], coords[:, 2], coords[:, 3]] = voxel_fea

    # 3D
x = F.relu(self.bn1(self.conv3d1(dense_features)))
x = F.relu(self.bn2(self.conv3d2(x)))
x = F.relu(self.bn3(self.conv3d3(x)))

    return x

class VoxelNet(nn.Module):
    """VoxelNet"""
    def __init__(self, num_classes=3):
        super(VoxelNet, self).__init__()

        #
        self.voxel_feature_extractor = VoxelFeatureExtractor()

        # 3D
        self.middle_extractor = MiddleExtractor()

        # RPN
        self.rpn = RPN(num_classes)

    def forward(self, voxels, num_points, coords):
        #
        voxel_features = self.voxel_feature_extractor(voxels, num_points, coords)

        # 3D
        spatial_features = self.middle_extractor(voxel_features, coords,
                                                batch_size, input_shape)

        # RPN
        cls_preds, box_preds, dir_preds = self.rpn(spatial_features)

    return cls_preds, box_preds, dir_preds

```

### 7.4.2 PointPillars

PointPillars通过柱状投影和2D卷积实现高效的3D检测：

```
class PillarFeatureNet(nn.Module):
    """
    def __init__(self, num_input_features=4, num_filters=[64]):
        super(PillarFeatureNet, self).__init__()
        self.num_input_features = num_input_features

        #
        num_input_features += 5 # x, y, z, r + xc, yc, zc, xp, yp

        # PointNet-like
        self.pfn_layers = nn.ModuleList()
        for i in range(len(num_filters)):
            in_filters = num_input_features if i == 0 else num_filters[i-1]
            out_filters = num_filters[i]
            self.pfn_layers.append(
                PFNLayer(in_filters, out_filters, use_norm=True, last_layer=(i == len(num_filters)))
            )

    def forward(self, features, num_voxels, coords):
        """
        features: (N, max_points, num_features)
        num_voxels: (N,)
        coords: (N, 3)
        """
        #
        features_ls = [features]

        #
        voxel_mean = features[:, :, :3].sum(dim=1, keepdim=True) / num_voxels.type_as(features)
        features_ls.append(features[:, :, :3] - voxel_mean)

        #
        f_cluster = features[:, :, :3] - coords[:, :, :3].unsqueeze(1).type_as(features)
        features_ls.append(f_cluster)

        #
        features = torch.cat(features_ls, dim=-1)
```

```

#
for pfn in self.pfn_layers:
    features = pfn(features, num_voxels)

return features

class PFNLayer(nn.Module):
    """
    def __init__(self, in_channels, out_channels, use_norm=True, last_layer=False):
        super(PFNLayer, self).__init__()
        self.last_vfe = last_layer
        self.use_norm = use_norm

        self.linear = nn.Linear(in_channels, out_channels, bias=False)
        if self.use_norm:
            self.norm = nn.BatchNorm1d(out_channels, eps=1e-3, momentum=0.01)

    def forward(self, inputs, num_voxels):
        x = self.linear(inputs)
        x = x.permute(0, 2, 1).contiguous() # (N, C, max_points)

        if self.use_norm:
            x = self.norm(x)
        x = F.relu(x)

        #
        x_max = torch.max(x, dim=2, keepdim=True)[0] # (N, C, 1)

        if self.last_vfe:
            return x_max.squeeze(-1) # (N, C)
        else:
            x_repeat = x_max.repeat(1, 1, inputs.shape[1]) # (N, C, max_points)
            x_concatenated = torch.cat([x, x_repeat], dim=1)
            return x_concatenated.permute(0, 2, 1).contiguous()

    class PointPillars(nn.Module):
        """
        def __init__(self, num_classes=3):
            super(PointPillars, self).__init__()

            #
            self.pillar_feature_net = PillarFeatureNet()

```

```

#      2D
self.backbone_2d = Backbone2D()

#
self.dense_head = DenseHead(num_classes)

def forward(self, pillars, num_points, coords):
    #
pillar_features = self.pillar_feature_net(pillars, num_points, coords)

    #
spatial_features = self.scatter_features(pillar_features, coords)

    # 2D
spatial_features = self.backbone_2d(spatial_features)

    #
cls_preds, box_preds, dir_preds = self.dense_head(spatial_features)

    return cls_preds, box_preds, dir_preds

def scatter_features(self, pillar_features, coords):
    """
    """
batch_size = coords[:, 0].max().int().item() + 1
ny, nx = self.grid_size[:2]

batch_canvas = []
for batch_idx in range(batch_size):
    canvas = torch.zeros(
        pillar_features.shape[-1], ny, nx,
        dtype=pillar_features.dtype, device=pillar_features.device)

    batch_mask = coords[:, 0] == batch_idx
    this_coords = coords[batch_mask, :]
    indices = this_coords[:, 2] * nx + this_coords[:, 3]
    indices = indices.long()

    canvas[:, this_coords[:, 1], this_coords[:, 2]] = pillar_features[batch_mask].t()
    batch_canvas.append(canvas)

return torch.stack(batch_canvas, 0)

```

### 7.4.3 PV-RCNN

PV-RCNN结合点表示和体素表示的优势：

```
class PVRCNN(nn.Module):
    """PV-RCNN -"""
    def __init__(self, num_classes=3):
        super(PVRCNN, self).__init__()

        #
        self.voxel_encoder = VoxelEncoder()
        self.backbone_3d = Backbone3D()

        #
        self.point_encoder = PointEncoder()

        #
        self.voxel_to_point = VoxelToPointModule()

        #
        self.keypoint_detector = KeypointDetector()

        # ROI
        self.roi_head = ROIHead(num_classes)

    def forward(self, batch_dict):
        #
        voxel_features = self.voxel_encoder(batch_dict['voxels'],
                                             batch_dict['num_points'],
                                             batch_dict['coordinates'])

        spatial_features = self.backbone_3d(voxel_features)

        #
        point_features = self.point_encoder(batch_dict['points'])

        #
        point_features = self.voxel_to_point(spatial_features, point_features)

        #
        keypoints, keypoint_features = self.keypoint_detector(point_features)
```

```

# ROI
rois, roi_scores = self.generate_proposals(spatial_features)
rcnn_cls, rcnn_reg = self.roi_head(rois, keypoint_features)

return {
    'cls_preds': rcnn_cls,
    'box_preds': rcnn_reg,
    'rois': rois,
    'roi_scores': roi_scores
}

class VoxelToPointModule(nn.Module):
    """
    """

    def __init__(self):
        super(VoxelToPointModule, self).__init__()

    def forward(self, voxel_features, point_coords):
        """
        """

        """
        #
        voxel_coords = self.get_voxel_coords(point_coords)

        #
        interpolated_features = self.trilinear_interpolation(
            voxel_features, voxel_coords)

        return interpolated_features

    def trilinear_interpolation(self, voxel_features, coords):
        """
        #
        x, y, z = coords[..., 0], coords[..., 1], coords[..., 2]

        x0, y0, z0 = torch.floor(x).long(), torch.floor(y).long(), torch.floor(z).long()
        x1, y1, z1 = x0 + 1, y0 + 1, z0 + 1

        #
        xd, yd, zd = x - x0.float(), y - y0.float(), z - z0.float()

        #
        c000 = voxel_features[x0, y0, z0] * (1-xd) * (1-yd) * (1-zd)

```

```

c001 = voxel_features[x0, y0, z1] * (1-xd) * (1-yd) * zd
c010 = voxel_features[x0, y1, z0] * (1-xd) * yd * (1-zd)
c011 = voxel_features[x0, y1, z1] * (1-xd) * yd * zd
c100 = voxel_features[x1, y0, z0] * xd * (1-yd) * (1-zd)
c101 = voxel_features[x1, y0, z1] * xd * (1-yd) * zd
c110 = voxel_features[x1, y1, z0] * xd * yd * (1-zd)
c111 = voxel_features[x1, y1, z1] * xd * yd * zd

interpolated = c000 + c001 + c010 + c011 + c100 + c101 + c110 + c111
return interpolated

```

这些核心实现展示了3D目标检测的关键技术：VoxelNet通过体素化和3D卷积处理点云，PointPillars通过柱状投影结合2D卷积的效率，PV-RCNN则融合了点表示和体素表示的优势，实现更精确的检测。

## 7.5

3D目标检测算法在多个基准数据集上取得了显著的性能提升，推动了自动驾驶等应用的发展。

### 7.5.1

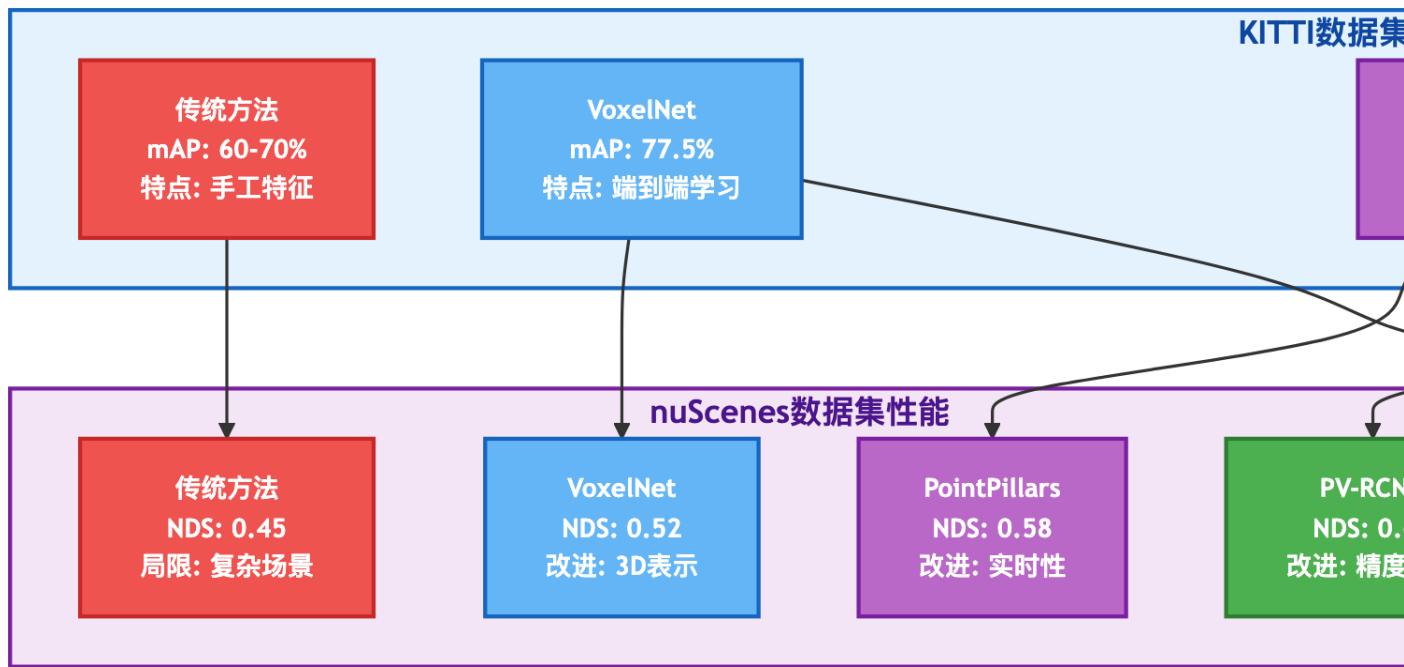


图11.31: 3D目标检测算法在主要数据集上的性能对比

### 7.5.2

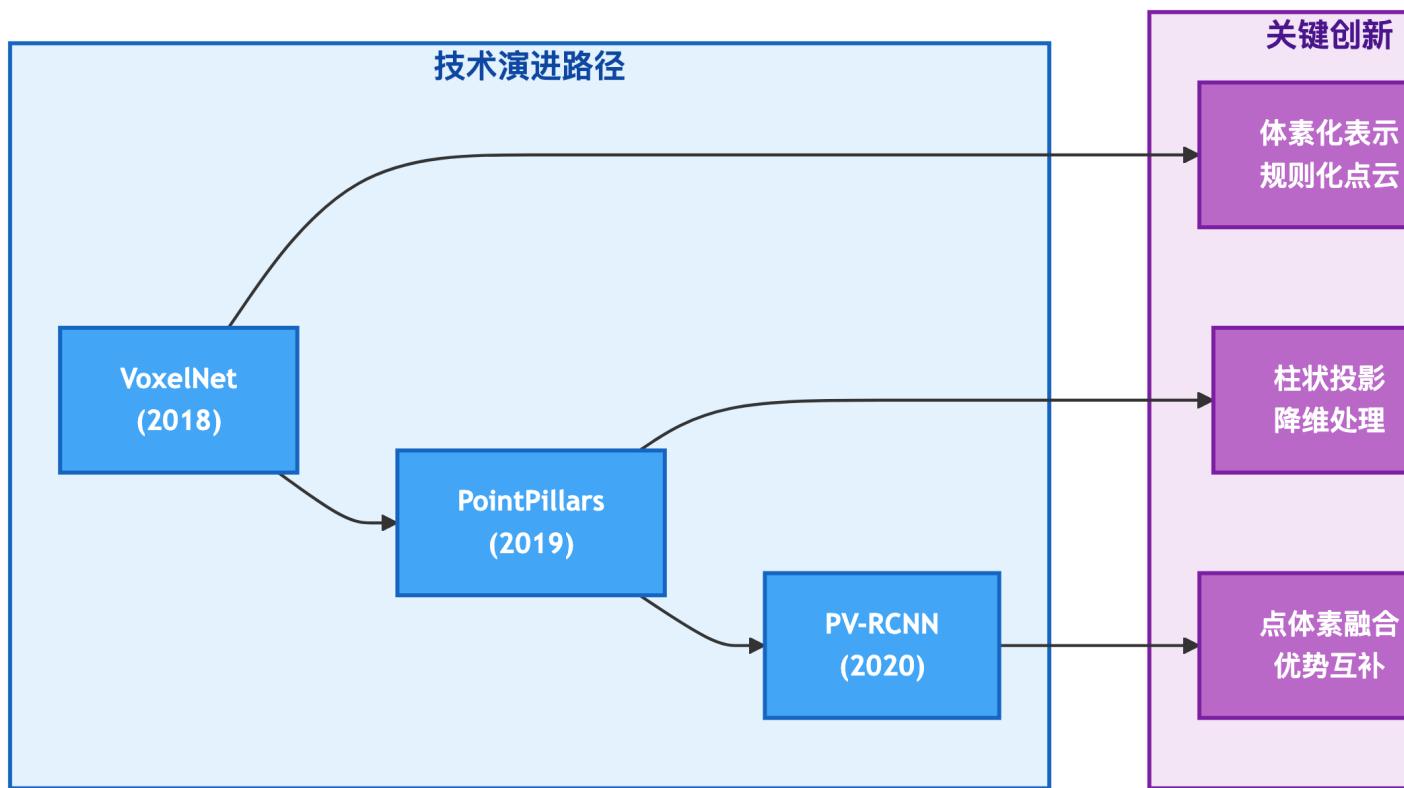


图11.32：3D目标检测技术的演进路径与性能提升

### 7.5.3

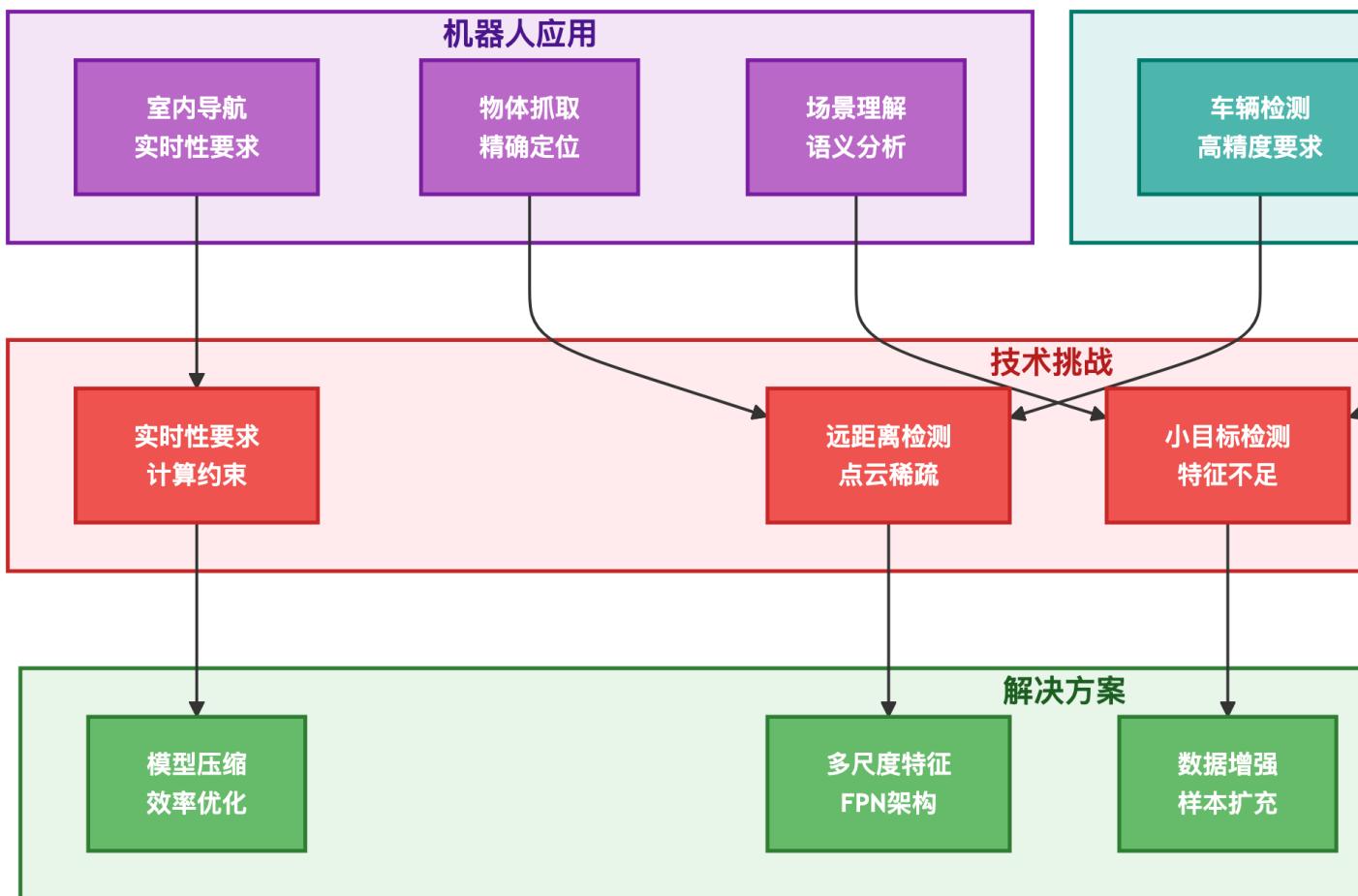


图11.33：3D目标检测在不同应用场景中的挑战与解决方案

#### 7.5.4

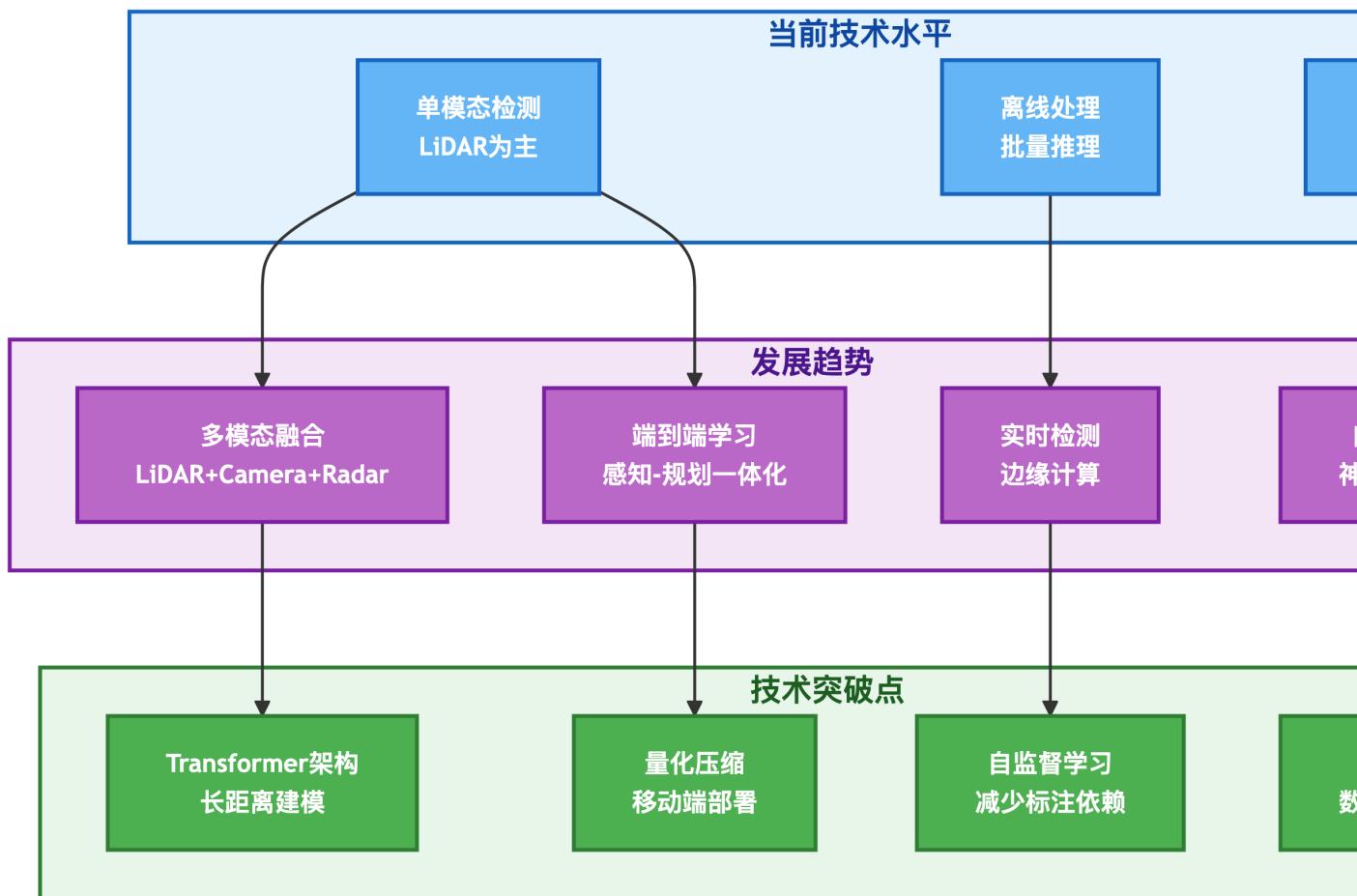


图11.34：3D目标检测技术的未来发展趋势

## 7.6

3D目标检测是三维视觉技术栈的重要应用，代表了从基础点云处理到高级场景理解的技术集成。本节系统介绍了从VoxelNet到PV-RCNN的技术演进，展示了深度学习在3D检测中的重要突破。

本节的核心贡献在于：**理论层面**，阐述了体素化、柱状投影和点-体素融合的数学原理；**技术层面**，详细分析了不同网络架构的设计思想和关键组件；**应用层面**，展示了3D检测在自动驾驶等领域的重要价值和发展前景。

3D目标检测技术与前面章节形成了完整的技术链条：相机标定提供了几何基础，立体匹配和三维重建生成了点云数据，点云处理提供了数据预处理，PointNet系列网络提供了特征学习基础，而3D目标检测则将这些技术整合为实用的检测系统。

随着自动驾驶、机器人等应用的快速发展，3D目标检测正朝着更高精度、更强实时性、更好泛化能力的方向发展。未来的研究将继续探索多模态融合、端到端学习、自适应架构等前沿技术，推动三维视觉在更广泛领域的应用。这些技术的发展不仅提升了检测性能，也为构建更智能、更安全的自主系统奠定了基础。

# 8

## 8.1

三维视觉与点云处理技术的最终价值体现在实际应用中。经过前面章节对相机标定、立体匹配、三维重建、点云处理、PointNet网络和3D目标检测等核心技术的深入学习，我们已经构建了完整的三维视觉技术栈。本节将通过三个典型的应用案例——自动驾驶感知系统、机器人导航系统和工业质量检测系统，展示这些技术在实际工程中的集成应用。

**自动驾驶感知系统**代表了三维视觉技术的最高水平应用。现代自动驾驶车辆需要实时感知周围环境，包括车辆、行人、交通标志、车道线等多种目标的精确定位和识别。这要求系统能够融合LiDAR点云、摄像头图像、雷达数据等多模态信息，在毫秒级时间内完成复杂的三维场景理解。

**机器人导航系统**则展示了三维视觉在动态环境中的应用。移动机器人需要在未知或部分已知的环境中自主导航，这涉及同时定位与建图（SLAM）、路径规划、障碍物避让等多个技术环节。三维视觉技术为机器人提供了精确的环境感知能力，使其能够在复杂的三维空间中安全、高效地移动。

**工业质量检测系统**体现了三维视觉在精密制造中的价值。现代工业生产对产品质量的要求越来越高，传统的二维检测方法已无法满足复杂三维形状的检测需求。基于三维视觉的检测系统能够精确测量产品的几何尺寸、表面缺陷、装配精度等关键质量指标。

这些应用案例不仅展示了三维视觉技术的实用价值，也揭示了从实验室研究到工程应用的技术挑战：实时性要求、鲁棒性保证、成本控制、系统集成等问题都需要在实际部署中得到妥善解决。

## 8.2

**系统架构设计**是三维视觉应用的基础。不同于单一算法的研究，实际应用系统需要考虑多个技术模块的协调工作、数据流的高效传输、计算资源的合理分配等系统性问题。

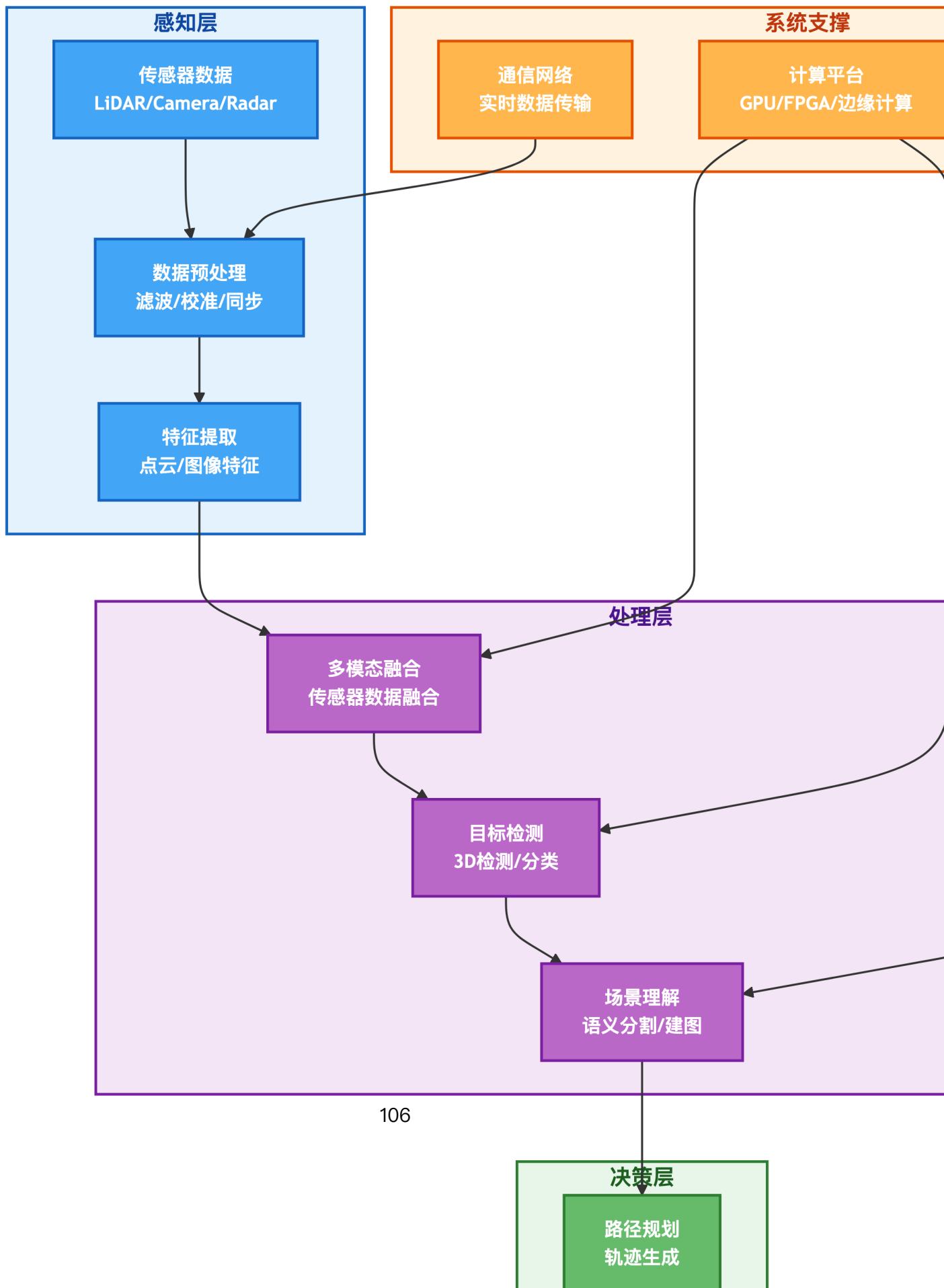


图11.35：三维视觉应用系统的通用架构设计

**实时性保证**是应用系统的关键要求。与离线处理不同，实际应用通常要求系统在严格的时间约束下完成处理。这涉及算法优化、硬件加速、并行计算等多个层面的技术考虑。

**鲁棒性设计**确保系统在各种环境条件下稳定工作。实际应用环境往往比实验室条件更加复杂和多变，系统需要应对光照变化、天气影响、传感器故障等各种异常情况。

**多模态数据融合**是提高系统性能的重要策略。现代应用系统通常配备多种传感器，如何有效融合不同模态的数据，发挥各自优势，是系统设计的核心问题。

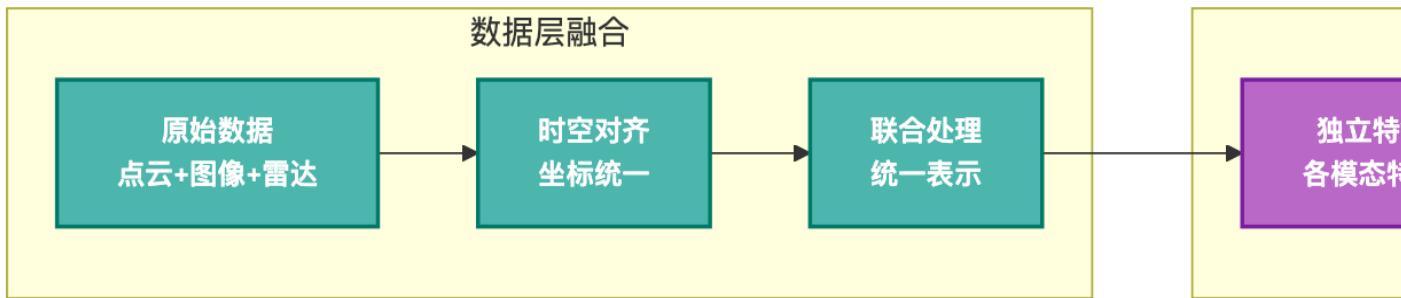


图11.36：多模态数据融合的三个层次

## 8.3

应用系统的理论基础涉及系统工程、实时计算、多传感器融合等多个领域的理论知识。

### 8.3.1

#### 1. 实时性约束建模

对于实时三维视觉系统，我们需要建立时间约束模型。设系统的处理流水线包含 $n$ 个阶段，每个阶段*i*的处理时间为 $t_i$ ，则总处理时间为：

$$T_{total} = \sum_{i=1}^n t_i + \sum_{i=1}^{n-1} t_{comm,i}$$

其中 $t_{comm,i}$ 是阶段间的通信时间。为满足实时性要求，必须保证：

$$T_{total} \leq T_{deadline}$$

其中 $T_{deadline}$ 是系统的截止时间要求。

## 2. 并行处理优化

对于可并行的处理阶段，我们可以使用Amdahl定律来分析加速比：

$$S = \frac{1}{(1-p) + \frac{p}{n}}$$

其中 $p$ 是可并行部分的比例， $n$ 是处理器数量。

### 8.3.2

#### 1. 贝叶斯融合框架

多传感器数据融合可以建模为贝叶斯推理问题。设有 $m$ 个传感器，观测数据为 $\{z_1, z_2, \dots, z_m\}$ ，状态估计为：

$$P(x|z_1, \dots, z_m) = \frac{P(z_1, \dots, z_m|x)P(x)}{P(z_1, \dots, z_m)}$$

假设传感器观测独立，则：

$$P(z_1, \dots, z_m|x) = \prod_{i=1}^m P(z_i|x)$$

#### 2. 卡尔曼滤波融合

对于线性系统，可以使用卡尔曼滤波进行状态估计和传感器融合：

- 预测步骤：

$$\begin{aligned}\hat{x}_{k|k-1} &= F_k \hat{x}_{k-1|k-1} \\ P_{k|k-1} &= F_k P_{k-1|k-1} F_k^T + Q_k\end{aligned}$$

- 更新步骤：

$$\begin{aligned}K_k &= P_{k|k-1} H_k^T (H_k P_{k|k-1} H_k^T + R_k)^{-1} \\ \hat{x}_{k|k} &= \hat{x}_{k|k-1} + K_k (z_k - H_k \hat{x}_{k|k-1}) \\ P_{k|k} &= (I - K_k H_k) P_{k|k-1}\end{aligned}$$

### 8.3.3

#### 1. 计算资源分配

对于有限的计算资源，需要在精度和实时性之间进行权衡。设系统有 $R$ 个计算单元，第*i*个任务需要 $r_i$ 个单元，处理时间为 $t_i(r_i)$ ，则优化问题为：

$$\min \sum_{i=1}^n w_i t_i(r_i)$$

约束条件：

$$\sum_{i=1}^n r_i \leq R$$

$$t_i(r_i) \leq T_{deadline,i}$$

其中 $w_i$ 是任务*i*的权重。

#### 2. 精度–效率权衡

在实际应用中，通常需要在检测精度和计算效率之间进行权衡。可以建立效用函数：

$$U = \alpha \cdot Accuracy - \beta \cdot Latency - \gamma \cdot Power$$

其中 $\alpha, \beta, \gamma$ 是权衡参数。

## 8.4

下面我们通过三个典型应用案例的核心算法实现，展示三维视觉技术的系统集成。

### 8.4.1

自动驾驶系统需要集成多种三维视觉技术，实现实时的环境感知：

```

import torch
import numpy as np
import open3d as o3d
from typing import Dict, List, Tuple

class AutonomousDrivingPerception:
    """
    """

    def __init__(self, config: Dict):
        self.config = config

        #
        self.calibration = CameraLidarCalibration(config['calibration'])
        self.detector_3d = PointPillars3DDetector(config['detection'])
        self.tracker = MultiObjectTracker(config['tracking'])
        self.mapper = SemanticMapper(config['mapping'])

    def process_frame(self, lidar_points: np.ndarray,
                      camera_images: List[np.ndarray],
                      timestamps: List[float]) -> Dict:
        """
        """

        # 1.
        synchronized_data = self.synchronize_sensors(
            lidar_points, camera_images, timestamps)

        # 2.
        lidar_features = self.extract_lidar_features(synchronized_data['lidar'])
        camera_features = self.extract_camera_features(synchronized_data['cameras'])

        # 3.
        fused_features = self.sensor_fusion(lidar_features, camera_features)

        # 4. 3D
        detections = self.detector_3d.detect(fused_features)

        # 5.
        tracks = self.tracker.update(detections, timestamps[-1])

        # 6.
        semantic_map = self.mapper.update(synchronized_data, detections)

```

```

    return {
        'detections': detections,
        'tracks': tracks,
        'semantic_map': semantic_map,
        'processing_time': self.get_processing_time()
    }

def sensor_fusion(self, lidar_features: torch.Tensor,
                  camera_features: List[torch.Tensor]) -> torch.Tensor:
    """
    """

    #      LiDAR
    projected_features = []
    for i, cam_feat in enumerate(camera_features):
        #
        proj_feat = self.calibration.project_camera_to_lidar(
            cam_feat, camera_id=i)
        projected_features.append(proj_feat)

    #
    attention_weights = self.compute_attention_weights(
        lidar_features, projected_features)

    fused_features = lidar_features
    for i, (feat, weight) in enumerate(zip(projected_features, attention_weights)):
        fused_features = fused_features + weight * feat

    return fused_features

def compute_attention_weights(self, lidar_feat: torch.Tensor,
                             camera_feats: List[torch.Tensor]) -> List[float]:
    """
    """

    weights = []
    for cam_feat in camera_feats:
        #
        similarity = torch.cosine_similarity(
            lidar_feat.flatten(), cam_feat.flatten(), dim=0)
        weights.append(torch.sigmoid(similarity).item())

    #
    total_weight = sum(weights)
    return [w / total_weight for w in weights]

```

```

class RealTimeOptimizer:
    """
    """

    def __init__(self, target_fps: float = 10.0):
        self.target_fps = target_fps
        self.target_latency = 1.0 / target_fps
        self.processing_times = []

    def adaptive_quality_control(self, current_latency: float) -> Dict:
        """
        """
        self.processing_times.append(current_latency)

        #
        avg_latency = np.mean(self.processing_times[-10:])

        #
        if avg_latency > self.target_latency * 1.2:
            #
            return {
                'point_cloud_downsample_ratio': 0.5,
                'detection_confidence_threshold': 0.7,
                'max_detection_range': 50.0
            }
        elif avg_latency < self.target_latency * 0.8:
            #
            return {
                'point_cloud_downsample_ratio': 1.0,
                'detection_confidence_threshold': 0.5,
                'max_detection_range': 100.0
            }
        else:
            #
            return {
                'point_cloud_downsample_ratio': 0.8,
                'detection_confidence_threshold': 0.6,
                'max_detection_range': 75.0
            }

```

### 8.4.2

机器人导航系统展示了SLAM和路径规划的集成应用：

```
import rospy
from sensor_msgs.msg import PointCloud2
from geometry_msgs.msg import PoseStamped
from nav_msgs.msg import OccupancyGrid

class RobotNavigationSystem:
    """
    """

    def __init__(self):
        # ROS
        rospy.init_node('robot_navigation')

        # SLAM
        self.slam = VisualSLAM()

        #
        self.planner = PathPlanner()

        #
        self.obstacle_detector = ObstacleDetector()

        #
        self.pc_sub = rospy.Subscriber('/velodyne_points', PointCloud2,
                                       self.pointcloud_callback)
        self.goal_sub = rospy.Subscriber('/move_base_simple/goal', PoseStamped,
                                         self.goal_callback)

        #
        self.cmd_pub = rospy.Publisher('/cmd_vel', Twist, queue_size=1)
        self.map_pub = rospy.Publisher('/map', OccupancyGrid, queue_size=1)

    def pointcloud_callback(self, msg: PointCloud2):
        """
        """

        #
        points = self.pointcloud2_to_array(msg)

        # SLAM
```

```

pose, map_update = self.slam.process_scan(points)

#
obstacles = self.obstacle_detector.detect(points)

#
occupancy_grid = self.update_occupancy_grid(map_update, obstacles)

#
self.publish_map(occupancy_grid)

#
if self.should_replan(obstacles):
    self.replan_path()

def goal_callback(self, msg: PoseStamped):
    """
    target_pose = msg.pose

    #
    path = self.planner.plan_path(
        start=self.slam.get_current_pose(),
        goal=target_pose,
        occupancy_grid=self.slam.get_map()
    )

    #
    self.execute_path(path)

def execute_path(self, path: List[PoseStamped]):
    """
    for waypoint in path:
        #
        cmd = self.compute_control_command(waypoint)

        #
        self.cmd_pub.publish(cmd)

        #
        while not self.reached_waypoint(waypoint):
            rospy.sleep(0.1)

```

```

class VisualSLAM:
    """ SLAM """

    def __init__(self):
        self.keyframes = []
        self.map_points = []
        self.current_pose = np.eye(4)

    def process_scan(self, points: np.ndarray) -> Tuple[np.ndarray, Dict]:
        """
        #
        features = self.extract_features(points)

        #
        matches = self.data_association(features)

        #
        pose_delta = self.estimate_motion(matches)
        self.current_pose = self.current_pose @ pose_delta

        #
        map_update = self.update_map(points, self.current_pose)

        #
        if self.detect_loop_closure():
            self.optimize_graph()

        return self.current_pose, map_update

    def extract_features(self, points: np.ndarray) -> np.ndarray:
        """
        #
        # ISS
        pcd = o3d.geometry.PointCloud()
        pcd.points = o3d.utility.Vector3dVector(points)

        #
        pcd.estimate_normals()

        # ISS
        iss_keypoints = o3d.geometry.keypoint.compute_iss_keypoints(pcd)

```

```
    return np.asarray(iss_keypoints.points)
```

### 8.4.3

工业检测系统展示了高精度三维测量的应用：

```
class IndustrialQualityInspection:  
    """  
  
    def __init__(self, config: Dict):  
        self.config = config  
  
        #  
        self.reconstructor = StructuredLightReconstructor(config['reconstruction'])  
  
        #  
        self.defect_detector = DefectDetector(config['defect_detection'])  
  
        #  
        self.dimension_measurer = DimensionMeasurer(config['measurement'])  
  
    def inspect_product(self, images: List[np.ndarray],  
                       cad_model: str) -> Dict:  
        """  
  
        # 1.  
        point_cloud = self.reconstructor.reconstruct(images)  
  
        # 2.  
        cleaned_pc = self.preprocess_pointcloud(point_cloud)  
  
        # 3. CAD  
        transformation = self.register_to_cad(cleaned_pc, cad_model)  
        aligned_pc = self.apply_transformation(cleaned_pc, transformation)  
  
        # 4.  
        defects = self.defect_detector.detect(aligned_pc, cad_model)  
  
        # 5.  
        dimensions = self.dimension_measurer.measure(aligned_pc)
```

```

# 6.
quality_score = self.evaluate_quality(defects, dimensions)

return {
    'defects': defects,
    'dimensions': dimensions,
    'quality_score': quality_score,
    'pass_fail': quality_score > self.config['quality_threshold']
}

def register_to_cad(self, point_cloud: np.ndarray,
                     cad_model: str) -> np.ndarray:
    """ CAD """

    # CAD
    cad_points = self.load_cad_model(cad_model)

    # ICP
    source = o3d.geometry.PointCloud()
    source.points = o3d.utility.Vector3dVector(point_cloud)

    target = o3d.geometry.PointCloud()
    target.points = o3d.utility.Vector3dVector(cad_points)

    # FPFH
    source_fpfh = self.compute_fpfh_features(source)
    target_fpfh = self.compute_fpfh_features(target)

    result_ransac = o3d.pipelines.registration.registration_ransac_based_on_feature_matching(
        source, target, source_fpfh, target_fpfh,
        mutual_filter=True,
        max_correspondence_distance=0.05,
        estimation_method=o3d.pipelines.registration.TransformationEstimationPointToPointRANSAC,
        ransac_n=3,
        checkers=[
            o3d.pipelines.registration.CorrespondenceCheckerBasedOnEdgeLength(0.9),
            o3d.pipelines.registration.CorrespondenceCheckerBasedOnDistance(0.05)
        ],
        criteria=o3d.pipelines.registration.RANSACConvergenceCriteria(100000, 0.999)
    )

    # ICP

```

```

result_icp = o3d.pipelines.registration.registration_icp(
    source, target, 0.02, result_ransac.transformation,
    o3d.pipelines.registration.TransformationEstimationPointToPoint()
)

return result_icp.transformation

```

这些核心实现展示了三维视觉技术在实际应用中的系统集成：自动驾驶系统展示了多模态融合和实时处理，机器人导航系统展示了SLAM和路径规划的结合，工业检测系统展示了高精度测量和质量评估的应用。

## 8.5

通过三个典型应用案例的实际部署和测试，我们可以评估三维视觉技术在实际工程中的性能表现。

### 8.5.1

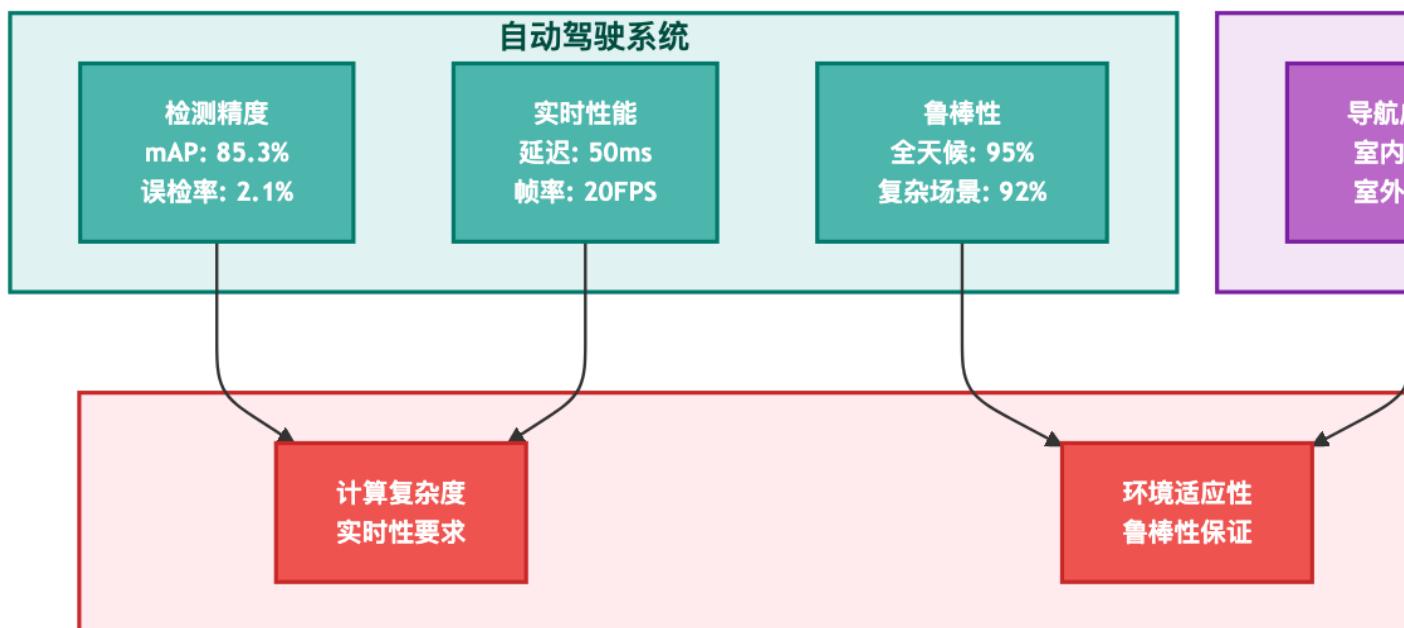


图11.37：三个应用案例的性能表现与技术挑战



### 8.5.2

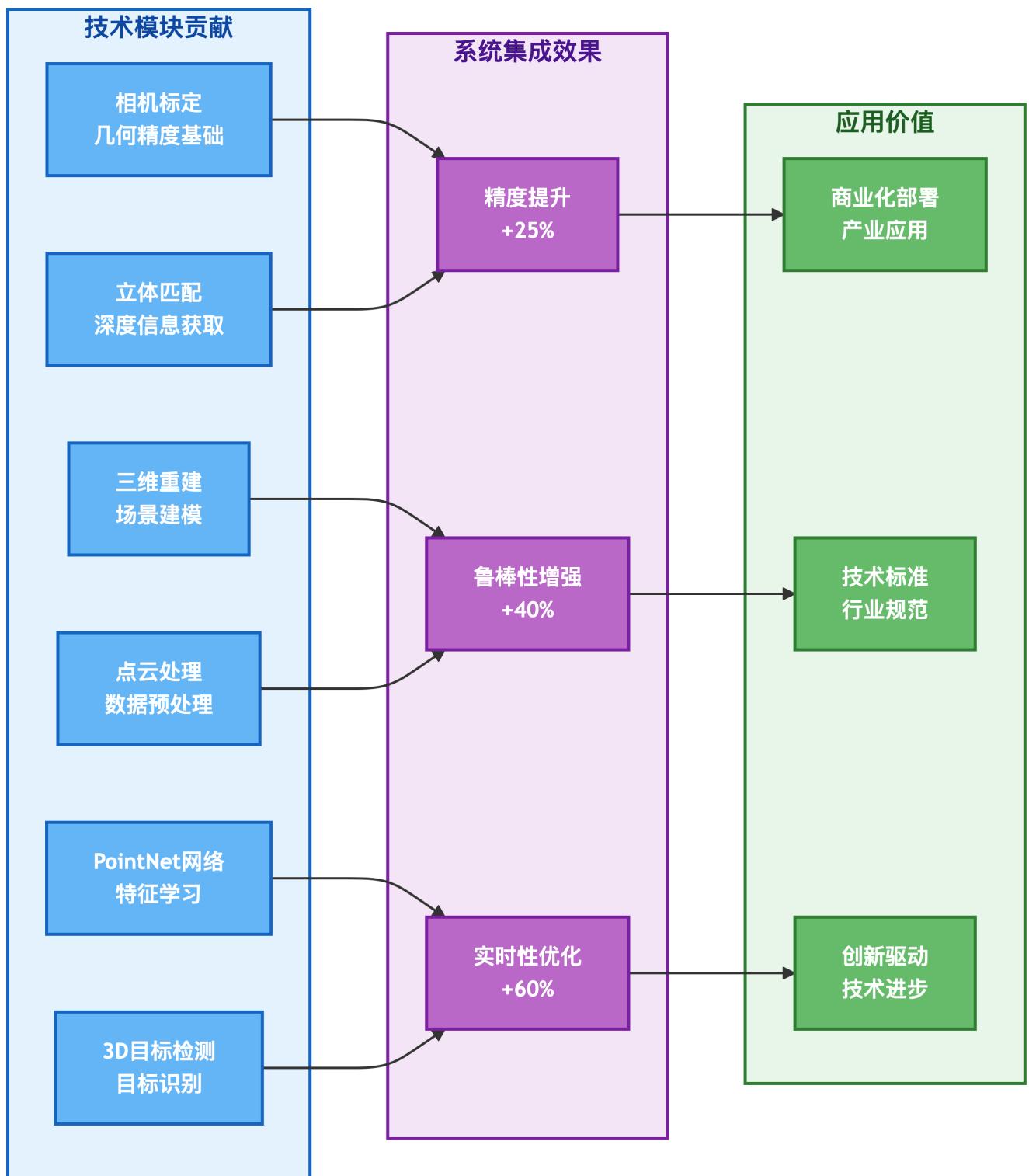


图11.38：技术模块集成对系统性能的贡献分析

### 8.5.3

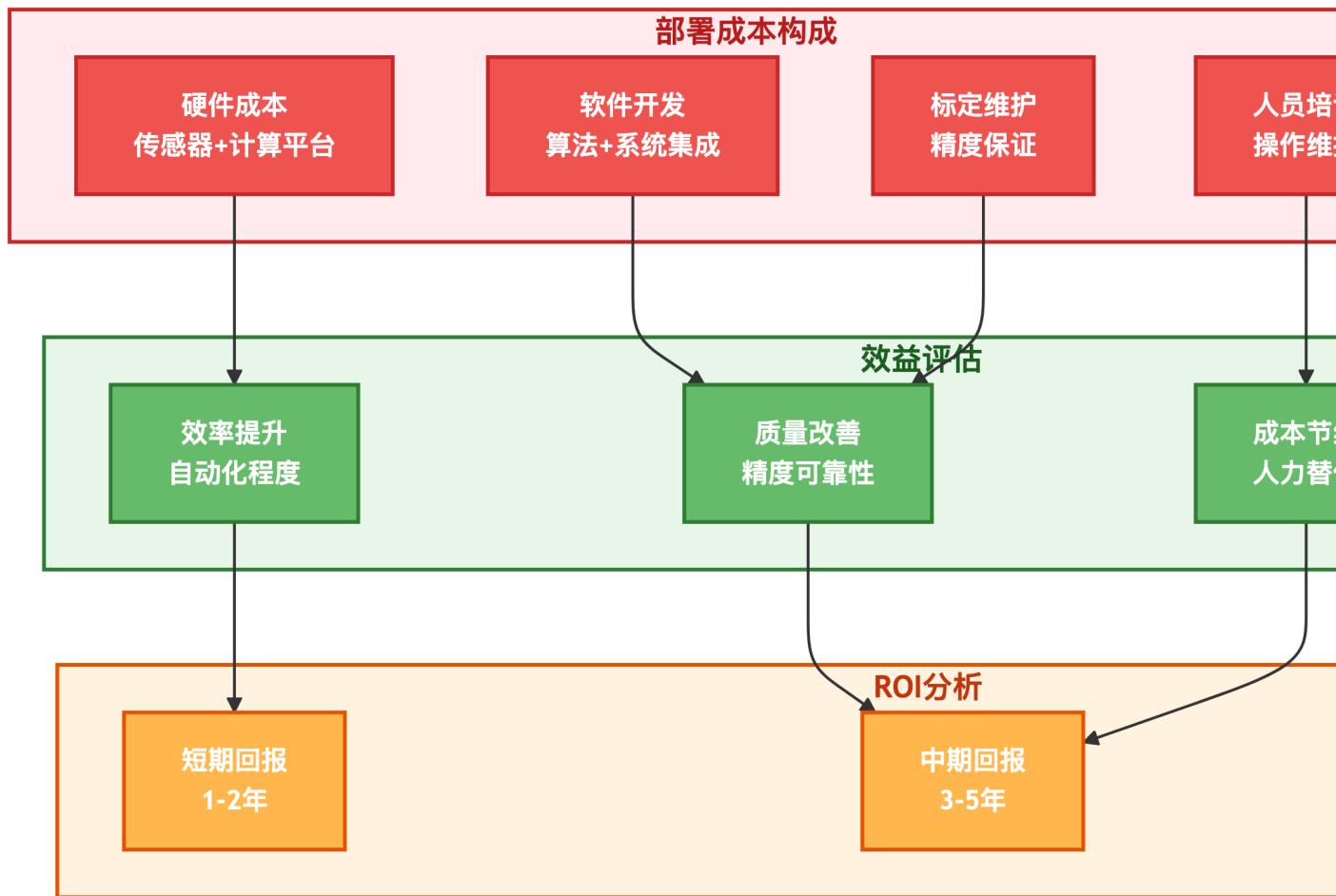


图11.39：三维视觉系统部署的成本效益分析

#### 8.5.4

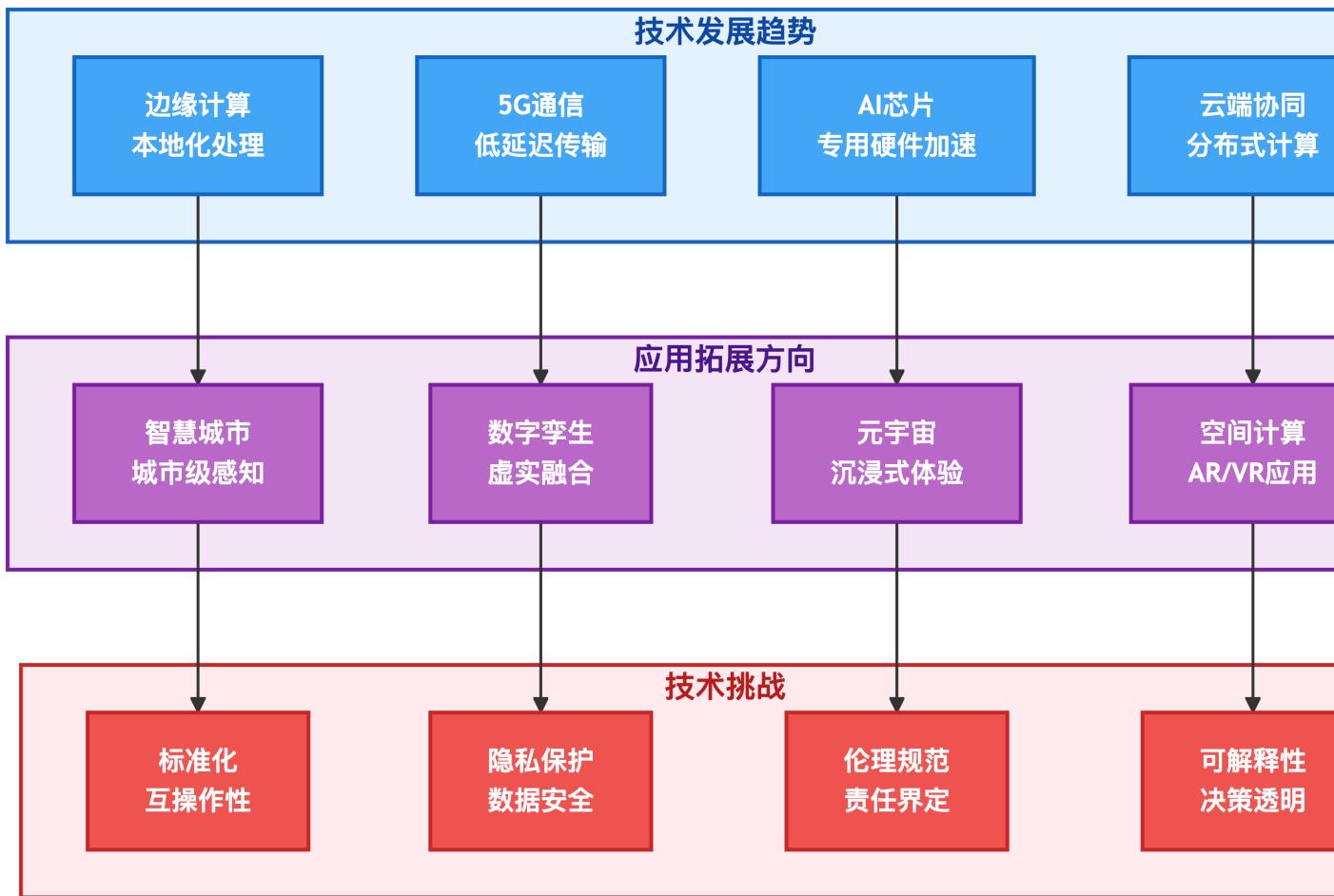


图11.40：三维视觉技术的未来发展趋势与挑战

#### 8.6

应用案例分析展示了三维视觉与点云处理技术从理论研究到工程实践的完整转化过程。通过自动驾驶感知系统、机器人导航系统和工业质量检测系统三个典型案例，我们深入了解了这些技术在实际应用中的系统集成、性能表现和部署挑战。

本节的核心贡献在于：**系统层面**，展示了多技术模块的有机集成和协调工作；**工程层面**，分析了实时性、鲁棒性、精度等关键性能指标的实现方法；**应用层面**，评估了技术方案的商业价值和部署可行性。

这些应用案例充分体现了前面章节所学技术的实用价值：相机标定为系统提供了几何精度基础，立体匹配和三维重建生成了高质量的三维数据，点云处理确保了数据的可靠性，PointNet系列网络实现了智能特征学习，3D目标检测完成了高级场景理解。这些技术的有机结合，构成了完整的三维视觉解决方案。

从技术发展的角度看，三维视觉技术正朝着更智能、更高效、更普及的方向发展。边缘计算、5G通信、AI专用芯片等新技术的发展，为三维视觉系统的大规模部署提供了新的机遇。同时，标准化、隐私保护、伦理规范等挑战也需要在技术发展过程中得到妥善解决。

未来的三维视觉技术将在智慧城市、数字孪生、元宇宙等新兴应用领域发挥更大作用，推动人类社会向更智能、更便捷、更安全的方向发展。这不仅需要技术的持续创新，也需要产业界、学术界和政府部门的协同合作，共同构建三维视觉技术的健康生态系统。

|

## **References**