

电子科技大学

UNIVERSITY OF ELECTRONIC SCIENCE AND TECHNOLOGY OF CHINA

# 硕士学位论文

MASTER THESIS



论文题目 过曝场景下的双目立体匹配技术

学科专业 信号与信息处理

学 号 201821011608

作者姓名 徐增荣

指导教师 刘光辉 教授

分类号 \_\_\_\_\_ 密级 \_\_\_\_\_

UDC <sup>注1</sup> \_\_\_\_\_

# 学 位 论 文

过曝场景下的双目立体匹配技术

(题名和副题名)

徐增荣

(作者姓名)

指导教师

刘光辉

教 授

电子科技大学

成 都

(姓名、职称、单位名称)

申请学位级别

硕士

学科专业

信号与信息处理

提交论文日期

2021.4.23

论文答辩日期

2021.5.18

学位授予单位和日期

电子科技大学

2021 年 6 月

答辩委员会主席

评阅人

注 1：注明《国际十进分类法 UDC》的类号。

# **Binocular Stereo Matching in Over-exposed Environment**

A Master Thesis Submitted to

University of Electronic Science and Technology of China

Discipline: **Signal and Information Processing**

Author: **Zengrong Xu**

Supervisor: **Prof. Guanghui Liu**

**School of Information and Communication**

School: **Engineering**

## 独创性声明

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

作者签名: 徐增荣 日期: 2021年5月25日

## 论文使用授权

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

(保密的学位论文在解密后应遵守此规定)

作者签名: 徐增荣 导师签名: 孙晓红  
日期: 2021年5月25日

## 摘要

双目立体匹配是模拟人类视觉获取深度的技术，广泛应用于路径规划、光学测量和即时定位与地图构建等领域。相比于传统方法，基于深度学习的双目立体匹配技术具有特征鲁棒性好、视差结果稠密等特点，但面临高反光物体引起的过曝现象时，仍存在误匹配问题。同时，相关数据集的缺乏限制了过曝场景下相关技术的研究。镜面反射引起的过曝现象会导致双目图像的匹配特征消失，引起误匹配导致视差估计精度降低。因此，本文从修复过曝区域丢失特征的角度出发，基于深度学习技术，结合数据渲染和图像修复理论，对过曝场景下的双目立体匹配技术开展研究，具体研究内容如下：

针对过曝场景视差标注困难和公开数据集缺乏过曝场景的问题，构建了过曝场景数据集。研究了一种基于三维软件 Blender 标注过曝场景视差的方法，建立了双目相机深度视差转换的模型，构建了过曝场景下视差数据 900 例。最后通过图像重构验证了数据标注的正确性，为过曝场景双目立体匹配的精度评估和模型预训练提供了数据支持。

针对过曝现象导致匹配特征消失引起的误匹配问题，提出了利用双目信息冗余性修复丢失特征的思路，构建了一种基于权重和特征共享的多特征提取模型。该模型基于视差注意机制和图像修复模块，从来自模型不同深度和路径的多个语义特征中提取冗余信息，重构双目特征，解决了卷积神经网络获取中长距离匹配关系困难的问题，实现了多特征融合，改善了过曝区域的视差预测精度。

针对过曝场景整体预测精度较低的问题，一方面提出了一种基于特征相关性构建匹配代价空间的方法，提高了模型度量特征匹配度的能力；另一方面建立了视差估计和图像修复的多任务模型，从网络结构和目标函数两个方面联合优化网络。最终实现了过曝场景整体视差预测精度的提高。

实验结果表明，本文提出的模型同 PSMNet、GANet、AANet 等近年主流深度学习模型相比，可将过曝场景下视差估计平均像素误差由  $10.17(\pm 0.65)$  像素降低至 5.83 像素，精度提升约 40%。本文研究成果已应用于“高铁列车双目视差估计项目”的生产环境中，同现有方法相比，本文提出的模型可显著改善列车高反光部件的视差预测精度，为后续故障检测，定位与分析提供了准确的三维信息。

**关键词：**过曝场景，立体匹配，特征融合，图像修复，特征相关性

## ABSTRACT

Binocular stereo matching is a technology that simulates human vision to obtain depth. It is widely used in simultaneous localization and mapping, path planning, Optical measurement and other fields. The binocular stereo matching technology based on deep learning has dense disparity results, and its robustness to extract features in occlusion and weak texture scenes is significantly better than traditional technologies. However, when faced with overexposure caused by highly reflective objects, there is still a problem of mismatching. At the same time, the lack of relevant data sets limits further research on overexposed scenarios. The overexposure with reflection leads to the disappearance of matching features at different positions of the binocular images, which results in the low accuracy of disparity prediction. Therefore, this thesis starts from the perspective of repairing the missing features of the overexposed area, based on deep learning technology, combined with data rendering and image restoration theory, to carry out research on the binocular stereo matching technology in the overexposed scene. The specific research content is as follows:

Aiming at the difficulty of disparity labeling of real overexposed scenes and the lack of disparity labeling of overexposed scenes in public data sets, a method based on the 3D software Blender to mark the parallax of overexposed scenes was studied, and the depth parallax conversion model of binocular camera was established, and 900 samples of the parallax of overexposed scenes was constructed. Finally, the correctness of the data annotation is verified by image reconstruction, which provides data support for the binocular stereo matching accuracy evaluation and model pre-training of the overexposed scene.

Aiming at the mismatch problem caused by the disappearance of matching features caused by overexposure, the idea of using binocular information redundancy to repair the missing features is proposed, and a multi feature extraction model based on weight and feature sharing is constructed. The model is based on the parallax attention mechanism and image repair module, extracts redundant information from multi semantic features from different depths and paths of the model, reconstructs binocular features, and solves the problem of difficulty in obtaining mid-to-long-distance matching relationships with convolutional neural networks. Multi-source feature fusion is realized, and the prediction

---

## ABSTRACT

---

accuracy of parallax in the overexposed area is improved.

Aiming at the problem of low overall prediction accuracy of overexposed scenes, on the one hand, a method of constructing a matching cost volume based on feature correlation is proposed, which improves the ability of the model to measure feature matching; on the other hand, a multi-task model for disparity estimation and image inpainting is established, jointly optimize the network from the two aspects of network structure and objective function. Finally, the overall disparity prediction accuracy of the overexposed scene is improved.

Experimental results show that compared with recent deep learning models such as PSMNet, GANet, AANet, the model proposed in this thesis can reduce the average pixel error of parallax estimation in overexposed scenes from  $10.17 (\pm 0.65)$  pixels to 5.83 pixels. The accuracy is increased by about 40%. The research results of this thesis have been applied to the production environment of the "High-speed Rail Train Binocular Parallax Estimation Project". Compared with the existing methods, the model proposed in this thesis can significantly improve the parallax prediction accuracy of high-reflective train components, which provides accurate three-dimensional information for failure detection, location and analysis.

**Keywords:** overexposed scene, stereo matching, feature fusion, image inpainting, feature correlation

## 目 录

<b>第一章 绪 论 .....</b>	1
1.1 课题研究背景与意义 .....	1
1.2 国内外研究现状 .....	2
1.2.1 基于传统算法的双目立体匹配技术 .....	2
1.2.2 基于深度学习的双目立体匹配技术 .....	4
1.3 论文主要研究内容及结构 .....	5
<b>第二章 深度学习与双目立体匹配 .....</b>	8
2.1 相机几何模型 .....	8
2.1.1 小孔成像模型 .....	8
2.1.2 仿射变换 .....	10
2.1.3 对极约束 .....	11
2.2 深度学习 .....	13
2.2.1 基础理论 .....	13
2.2.2 卷积神经网络 .....	14
2.2.3 反向传播算法 .....	19
2.3 双目立体匹配 .....	22
2.3.1 基于传统特征的双目立体匹配 .....	22
2.3.2 基于深度学习的双目立体匹配 .....	24
2.4 本章小结 .....	25
<b>第三章 过曝场景数据集构建 .....</b>	26
3.1 公开数据集分析 .....	26
3.2 数据采集方法 .....	28
3.3 数据采集场景 .....	31
3.4 数据处理和视差验证 .....	33
3.5 本章小结 .....	35
<b>第四章 基于特征一致性的双目匹配特征提取模型 .....</b>	36
4.1 多特征模型 .....	36
4.1.1 基于跳跃连接的 U 型网络 .....	37
4.1.2 由单目到双目的多特征网络 .....	38
4.2 多特征融合 .....	41

4.2.1 Attention 机制 .....	41
4.2.2 基于 Parallax-Attention 的多特征融合 .....	42
4.3 双目特征一致性修复 .....	44
4.3.1 一致性预测 .....	45
4.3.2 损失函数 .....	45
4.4 实验结果与分析 .....	46
4.5 本章小结 .....	51
<b>第五章 过曝场景下的双目立体匹配 .....</b>	<b>52</b>
5.1 基于一致性匹配特征的视差网络 .....	52
5.2 基于特征相关性的 cost-volume 构建方法 .....	54
5.2.1 双目特征相关性 .....	54
5.2.2 基于特征相关性的 cost-volume 构建 .....	57
5.3 多任务模型优化及目标函数 .....	59
5.3.1 多任务联合优化 .....	60
5.3.2 目标函数 .....	61
5.4 实验结果与分析 .....	62
5.5 本章小结 .....	66
<b>第六章 全文总结与展望 .....</b>	<b>67</b>
6.1 全文总结 .....	67
6.2 后续工作展望 .....	68
<b>致 谢 .....</b>	<b>69</b>
<b>参考文献 .....</b>	<b>70</b>
<b>攻读硕士学位期间取得的成果 .....</b>	<b>75</b>

# 第一章 绪论

## 1.1 课题研究背景与意义

双目立体匹配是模拟人类大脑通过人眼视觉差异获取深度信息的技术，作为计算机视觉中一项基础且重要的技术，广泛应用于路径规划、光学测量和即时定位与地图构建等领域。国家发改委印发的《智能汽车创新发展战略》指出，2025年要实现双目立体匹配等复杂环境感知技术的突破<sup>[1]</sup>。然而现有双目立体匹配技术在面临高反光物体导致的过曝现象时，存在误匹配，预测精度低等难题。

过曝场景下双目立体匹配技术的难点在于：1) 过曝在图像上产生的白色光斑会覆盖物体细节，导致物体纹理、几何信息丢失，增加了特征提取的难度；2) 镜面反射引起的过曝现象对相机视角敏感，导致光斑位置随相机视角改变，使得双目图像的局部匹配特征消失，同时光斑本身做作为一种强纹理特征会形成新的匹配关系，但这种匹配关系不满足双目图像像素排列的顺序性，对构建正确的特征匹配关系提出了挑战；3) 场景三维信息由局部信息构成，局部特征的消失和误匹配会影响整体三维结构，给恢复场景三维信息带来了困难。传统双目立体匹配技术依靠物体边缘信息和纹理信息构建特征，根据像素排列顺序确定匹配关系，在过曝场景下存在大量误匹配现象。深度学习通过大量数据训练网络学习适合任务的语义表征，对像素的全局匹配关系具有更好的刻画能力。近年来，基于深度学习的双目立体匹配技术在弱纹理、遮挡等挑战场景取得较大进步。目前，主要工作集中在对匹配代价聚合阶段的研究。例如，GANet<sup>[2]</sup>采用多个方向的一维加权和代替3D卷积结构，提高了网络刻画全局匹配的能力；AANet<sup>[3]</sup>提出了一种自适应的多尺度代价聚合模块，增强了特征上下文的连贯性，提高了弱纹理区域的预测精度。但现有深度学习方法在过曝场景存在以下问题：1) 双目图像的特征计算过程相对独立，在左(右)图的特征计算过程中没有引入右(左)图的颜色、纹理和几何信息，无法通过右(左)图修复左(右)图过曝区域丢失的信息，导致网络提取的物体语义特征错误；2) 双目图像的特征计算模块权重、梯度绑定，过曝光斑在双目图像的语义特征高度相似，加剧了过曝区域误匹配；3) 公开数据集缺乏过曝场景<sup>[4-12]</sup>视差标注，现有方法无法通过预训练获取过曝区域视差匹配的先验知识，难以纠正过曝光斑的错误匹配关系。

本课题基于成都铁安科技有限责任公司“高铁列车双目视差估计项目”课题，围绕高铁列车过曝场景的光学测量需求，开展了过曝场景下视差估计的研究，具有重要的理论意义和应用价值。在理论方面，本文提出的多源特征模型和多源特

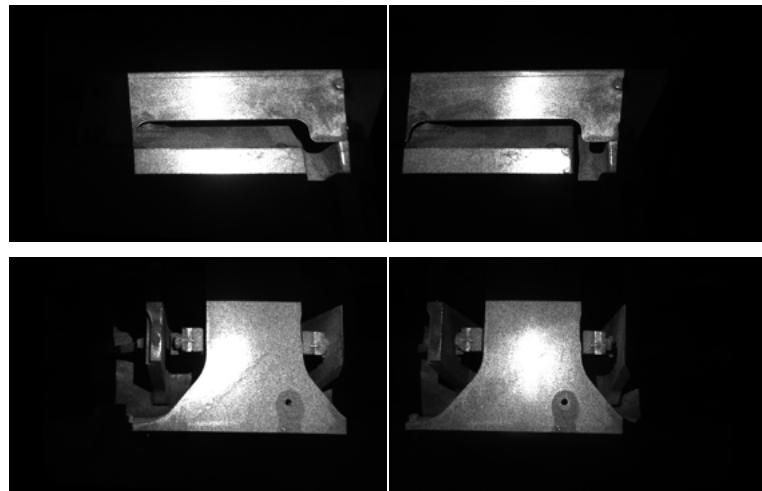


图 1-1 列车底部高反光部件

征融合方式，基于双目冗余性实现了过曝区域丢失细节的修复，为解决过曝场景误匹配提供了新思路。在应用方面，本文构建的数据集具有丰富的过曝场景，视差覆盖范围广，为过曝场景模型预训练提供了数据支持；本文提出的多源特征提取模块、特征融合模块、cost volume 模块，不仅可以实现丢失特征的修复，还可直接应用于已有的双目立体匹配模型，实现预测精度的提高。本文研究成果已应用于“高铁列车双目视差估计项目”生产环境，为后续故障检测、定位与分析铺平了道路。

## 1.2 国内外研究现状

双目立体匹配作为计算机视觉中的基础任务，自诞生以来就得到了大量的研究，其解决方案由最初基于局部像素的决策转变为各种形式匹配代价函数的优化，再到最近由数据驱动的机器学习(深度学习)方法。根据 Scharstein 和 Szeliski 等人的总结，双目立体匹配技术一般包含匹配代价计算，匹配代价聚合(优化)以及视差计算三个步骤<sup>[6]</sup>，一般来说首先需要对双目相机拍摄的图像进行校正，在校正后的图像上提取特征计算匹配代价，然后通过求和、平均或其他方式聚合代价，根据最终的匹配代价确定像素对应关系得到估计视差。随着深度学习技术的发展，双目立体匹配技术也在不断向卷积神经网络方向发展，因此本节将从传统方法和深度学习方法两个方向介绍双目立体匹配技术的研究现状。

### 1.2.1 基于传统算法的双目立体匹配技术

传统双目立体匹配可以分为局部和全局方法。局部方法又称为基于块的方法或基于窗口的方法，此类方法依赖预定义窗口内像素强度计算视差。因为仅使用

局部信息，所以此类方法计算复杂度低，常见的方法有 [13–15]。全局方法将视差计算视为所有匹配关系的全局能量函数最小化问题，此类方法同时最优化匹配项和平滑惩罚项，得到全局最优视差图，视差的平滑性更好，常见的方法有 [16–18]。下文依照双目立体匹配的技术流程介绍该领域研究现状。

双目立体匹配的第一阶段通常为匹配代价计算，该阶段在经过矫正后的双目图像平行像素间计算匹配程度。最初，学者们采用图像强度的绝对差 (Absolute Differences, AD) 或平方差 (Squared Differences, SD) 等方式度量相似度，此类方法易受到光照变化和重复纹理影响，输出视差平滑性较差。Min<sup>[19]</sup> 和 Pham<sup>[20]</sup> 等人通过在匹配点处引入图像梯度信息在一定程度上降低了匹配误差，使得该方法对光照变化具有一定鲁棒性。在单像素匹配度量的基础上，学者们进一步采用了块匹配度量方式构建匹配代价，即给定像素的匹配代价由其支持窗内所有像素绝对差值或平方差值的和决定，即绝对差值和 (Sum of Absolute Differences, SAD) 以及平方差值和 (Sum of Squared Differences, SSD) 等形式，使得方法对像素差异的鲁棒性进一步提高。Tippetts 等人<sup>[21]</sup> 首先证明了 SAD 具有低计算复杂度的特性，在资源有限的嵌入式系统上实现了该方法。Lee 等人<sup>[22]</sup> 在此基础上采用滑动窗口实现了该方法的并行处理，再利用图形处理单元 (Graphical Processing Unit, GPU) 并行加速，有效缩短了算法的运行时间。Gupta 等人<sup>[23]</sup> 则将分层思想引入匹配代价计算中，第一层尺寸较大的窗口用于计算初始匹配代价，第二层较小的窗口尺寸用于提升算法在物体边缘处的表现，保证匹配精度的同时降低了计算复杂度。在传统方法中也有部分学者对利用图像特征构建匹配代价进行了研究。Sharma 等人<sup>[24]</sup> 利用尺度不变特征变换 (Scale Invariant Feature Transform, SIFT) 中的多层特征进行匹配度计算，有效提升了自动驾驶场景的视差估计精度。基于特征的匹配代价对遮挡和弱纹理区域十分敏感，Liu 等人<sup>[25]</sup> 将图像分割和边缘检测信息引入匹配代价的计算过程中，提高了算法在上述场景的精度。基于人工设计特征的代价匹配方法对环境敏感，鲁棒性较差，因此在双目视差技术的发展历程中使用率较低。

双目立体匹配的第二阶段一般为匹配代价聚合，单个像素计算得到的匹配代价无法确定匹配关系，需要聚合邻域像素匹配代价最大程度地减少匹配不确定性。一般来说，局部方法可以通过对当前像素支持窗内的匹配代价求和或求均值的方式得到聚合代价<sup>[6]</sup>。最早学者们采用固定尺寸窗口进行计算，但 Yang 等人<sup>[26]</sup> 通过实验证明固定尺寸窗口的方法必须针对输入设置合理窗口参数，否则在物体边缘处存在视差预测模糊的问题。为解决上述问题，Hirschmüller 等人<sup>[27]</sup> 采用了多种尺寸和形状窗口的聚合匹配代价，选择具有最小匹配误差的窗口作为输出。Lu 等人<sup>[28]</sup> 在此基础上对窗口添加了限制，阻止窗口跨过物体边缘，提高了深度不连

续和平滑区域的视差预测精度。由于上述方法需要自适应调整窗口，计算复杂度较高，Chen 等人<sup>[29]</sup>尝试利用视差冗余性提高算法运算速度，将像素按深度值分类，降低了后续运算量。根据 Fang 等人<sup>[30]</sup>的试验结果，在代价聚合阶段的各种支撑窗算法中，自适应权重窗口算法鲁棒性最好。自适应权重窗口算法赋予支撑窗内每个元素不同权重，权重大小根据其与中心像素间强度差异和欧氏距离决定<sup>[31]</sup>。

双目立体匹配的第三阶段为视差计算，对局部方法而言，一般采用“赢者通吃”(Winner Takes All, WTA)的策略，对于当前像素，与其具有最小聚合匹配代价的像素被选作匹配像素，两者坐标之差即为视差，Cigla<sup>[32]</sup>、Zhang<sup>[33]</sup> 和 Lee<sup>[34]</sup> 等人均采用了该策略。局部方法的视差仅由支撑窗内像素决定，并且对噪声敏感，因此在由 WTA 策略得到初始视差值后一般需要通过多个滤波器降低误差。相对于局部方法而言，全局方法的核心在于通过全局能量函数对视差做出假设，通过最小化能量函数计算最终视差。一般来说，全局方法的能量函数包含数据匹配函数和平滑函数。匹配函数表征各个像素在当前视差下匹配程度，平滑函数表征对视差梯度的惩罚，即鼓励相邻像素具有相同视差值。全局方法首先通过局部方法或者随机初始化方式获得初始视差，在初始视差基础上进行多轮迭代获得最终视差，因此对计算资源和存储资源要求较高。Wang 等人<sup>[35]</sup>引入图割算法对能量函数进行优化，通过交换像素位置选择具有最低能量的视差，降低了全局方法计算复杂度。考虑到左右目图像同一行像素间排列顺序的一致性，Hirschmuller<sup>[36]</sup> 利用一维动态规划将复杂度降至多项式复杂度。

局部双目立体匹配方法计算效率高，但基于部分信息得到的预测视差精度较低；全局方法在预测精度和平滑性上均优于局部方法，但需要多次迭代，计算复杂度较高。传统方法需要人工设计匹配代价特征，对环境敏感，鲁棒性较差，难以应对遮挡、重复纹理、弱纹理、物体过曝反光等挑战性场景。

### 1.2.2 基于深度学习的双目立体匹配技术

随着深度学习在图像分类、语义分割和目标检测等领域的性能日益提升，研究者们开始尝试使用深度学习技术实现双目立体匹配，并在多个数据集中取得了优于传统方法的成绩。基于深度学习的双目立体匹配技术也由最初使用深度模型提取特征发展为如今端到端的双目立体匹配网络。

早期研究者主要使用学习算法代替双目立体匹配某个步骤，如 Haeusler 等人<sup>[37]</sup>最早利用随机森林分类器实现预测视差的置信度分类，判断预测视差是否为离群点，决定是否保留该点预测。将深度学习应用到匹配计算中最广为人知的算法为 LeCun<sup>[38]</sup> 等人提出的 MC-CNN，MC-CNN 利用孪生网络提取图像特征，利

用  $1 \times 1$  卷积替换全连接层，实现了两幅图像各个块间匹配程度的预测，代替了传统的匹配代价计算方法。

添加了学习机制的双目立体匹配算法的鲁棒性有所提高，但直到 Mayer 等人<sup>[11]</sup> 提出首个端到端深度学习双目立体匹配网络 DispNet 后，视差预测精度才有了较大提升。DispNet 采用了类似光流预测和图像分割任务中的编解码网络结构，将原始图像通过卷积下采样 64 倍后，利用多层反卷积和上采样逐步回到原始尺寸实现视差估计。虽然 DispNet 在某些场景上的表现不如传统方法，但证明了端到端双目立体匹配网络的可行性。Pang 等人<sup>[39]</sup> 在此思路上提出了 CRL，一种两阶段双目匹配模型。CRL 第一阶段为 DispNet，用于预测初始视差，第二阶段为 DispResNet，采用了残差结构的设计，用于矫正初始视差。两阶段模型 CRL 首次达到了传统算法在 KITTI Stereo 上的排名，Liang 等人<sup>[40]</sup> 在此基础上提出了迭代残差网络 (Iterative Residual Network, iResNet)，在第二阶段前添加了多个重复的“卷积反卷积”结构，在 2018 年 Robust Vision Challenge 中取得了第一的成绩。

相较于采用 2D 卷积的模型，具有 3D 卷积结构的模型实现了更高精度的视差预测。Kendall 等人<sup>[41]</sup> 提出的 GC-Net 首次采用了 3D 卷积结构，用于从匹配代价空间 (cost volume) 中提取像素匹配关系，并提出了一种基于概率分布的视差计算方式。采用了 3D 卷积结构的 GC-Net 首次在 KITTI Stereo 上全方位超过了传统算法的表现，其三阶段网络设计逐渐成为双目立体匹配网络基本构型。Chang<sup>[42]</sup> 在此基础上提出了 PSMNet，通过引入多尺度堆叠沙漏网络结构 (Stacked Hourglass Networks) 使得预测精度进一步提升。3D 卷积需要占用大量计算资源，为提高网络实时性，Zhang<sup>[2]</sup> 提出了半全局聚合层代替 3D 卷积层，用于提取上下文语义特征，使得网络的空间和时间复杂度都有所降低。

得益于深度学习强大的特征提取能力和上下文学习能力，基于深度学习的双目立体匹配实现了在诸如遮挡区域，弱纹理区域预测精度的显著提高，但仍然面临着诸多挑战。在工业质检和自动驾驶等场景充斥着大量金属、玻璃等反光材质和行车大灯、照明灯等复杂光源，双目相机在此类场景中不可避免的会产生过曝现象。现有算法在过曝区域无法提取有效特征，无法解决过曝光斑误匹配问题，无法获得准确的视差估计。过曝作为挑战性场景中的常见现象，是目前双目立体匹配技术的热点和难点。

### 1.3 论文主要研究内容及结构

本课题基于成都铁安科技有限责任公司“高铁列车双目视差估计项目”课题，围绕高铁列车高反光部件光学测量的需求，针对过曝场景双目立体匹配技术进行

研究。涉及深度学习中关于数据集构建、多源特征模型设计、基于视差注意力机制的多源特征融合、基于特征相关性的 cost volume 构建以及多任务联合优化的研究。本文研究内容之间的逻辑关系如图1-2所示，各章节内容简述如下：

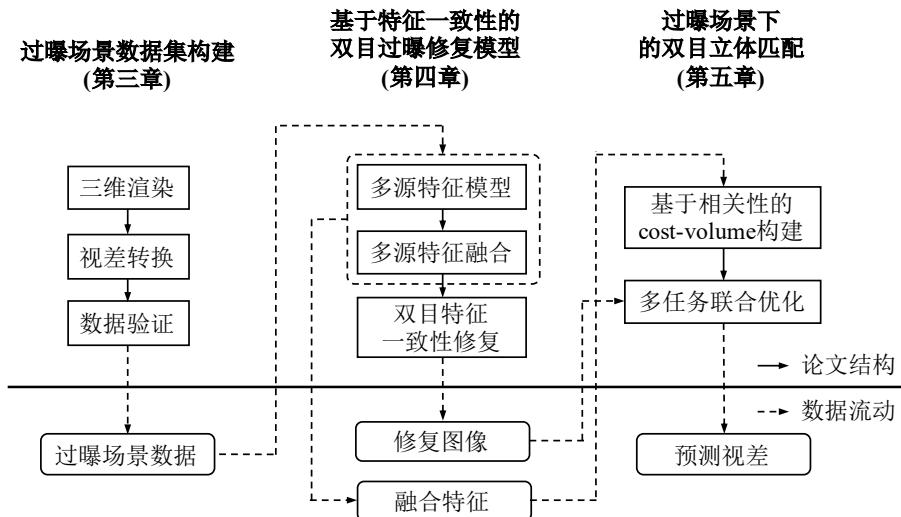


图 1-2 论文研究内容结构图

第一章阐述了课题的研究背景与意义，介绍了双目立体匹配的国内外研究现状，分析了过曝场景中双目立体匹配技术存在的难点，现有方法存在的不足，为本文后续研究打下基础。

第二章主要介绍了本文所涉及的相关技术基础。包括单目相机模型，坐标系仿射变换，双目相机中的极线约束以及深度学习基本理论，并从传统方法和深度学习两个方向对现有双目立体匹配技术进行了总结。为本文后续研究提供理论基础。

第三章研究了如何构建过曝场景双目视差数据集。针对公开数据集缺乏过曝场景视差标注的问题，首先分析了公开数据集的场景特点和采集方式，比较了物理标注和渲染标注两种方法的利弊。接着针对过曝区域视差标注困难问题，研究了如何利用三维软件 Blender 渲染过曝区域视差，并对数据的正确性进行了验证。弥补了现有数据集在过曝场景下的空白，为本文后续研究提供了数据基础，为列车过曝场景的模型预训练提供了先验知识。

第四章研究了基于特征一致性的双目过曝特征修复技术。针对过曝区域图像特征不一致问题，利用图像冗余性对丢失特征进行修复，首先提出了一种基于 U-Net 的多源特征提取模型。该模型采用对称的编解码结构将左右目特征引入到特征提取计算路径中。接着为实现冗余特征融合，提出了一种基于 Parallax-Attention 的双目特征融合方法。该方法通过特征注意力图构建特征的视差概率矩阵，利用

特征视差和冗余特征重构丢失信息实现特征融合。然后为了促进网络将右（左）图冗余信息搬到左（右）图过曝区域，引入了图像修复的网络模型和目标函数，激励网络学习特征融合，实现特征修复。最后在过曝场景进行消融实验并与基线模型进行对比实验。

第五章研究了过曝场景下的双目立体匹配技术。针对过曝场景下视差预测整体精度较差问题，一方面针对传统 cost volume 难以利用特征一致性的问题，提出了一种基于特征相关性的 cost volume 构建方法。该方法将特征相关性中点积相似性度量推广至 Hadamard 乘积形式，利用行特征相关性刻画匹配程度，不仅降低了数据体积，也提升了视差估计精度。另一方面引入特征修复和双目立体匹配的多任务模型，通过网格变换的采样方式重构图像和特征，在损失函数和网络结构两个方向实现了匹配精度的联合优化。最后在多个场景对模型性能进行测试。

第六章为全文总结，并对后续可能的改进方向进行了分析。

## 第二章 深度学习与双目立体匹配

得益于计算能力的飞速提升，在过去十几年中计算机视觉领域各个方向都逐渐由理论走向现实。路径规划、光学测量和即时定位与地图构建等应用都需要基于场景三维信息做进一步推理和决策，准确的三维重建技术已成为众多高层视觉任务的基础。随着人工智能不断发展和各种传感器精度不断提升，目前获取场景三维信息的方式主要有基于激光雷达的主动探测方式和基于相机视觉的被动计算方式。基于深度学习的双目立体匹配技术在精度上可以媲美激光雷达而在成本上几乎可以忽略不计，因而成为了近年来的研究热点。因此本章从相机模型出发，重点介绍深度学习基本理论以及双目立体匹配技术基本概念和方法。

### 2.1 相机几何模型

从十六世纪发明的第一款没有镜头的针孔照相机，到结构复杂、价格高昂的数码单镜反光相机，再到随处可见的手机内置相机，虽然外形变化多端，其成像本质都是利用针孔透视原理将物体表面发出或反射的光线投射在底板上，并记录每一小格光照强度形成图像，如下图2-1所示。从物体的世界坐标系到图像的相机平面坐标系之间有着严格的物理变换规律，双目图像包含着场景的三维信息，因此像素间也遵循着严格的对极约束。

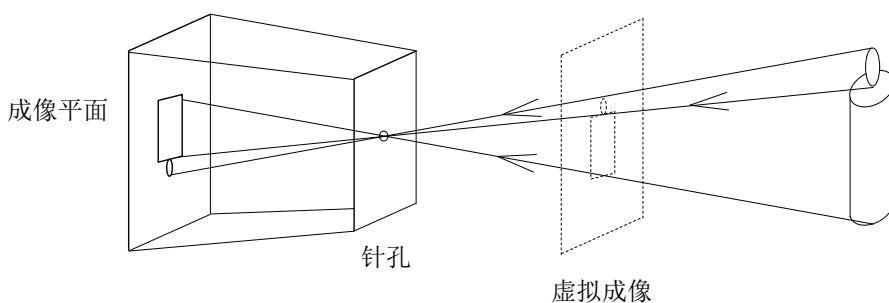


图 2-1 小孔成像模型

#### 2.1.1 小孔成像模型

严格来说，理想并简单的成像模型是不存在的，因为在现实中不论针孔多小，最终成像平面上每个点收集到的光均为带有一定角度的锥形光束。在研究中一般采用 15 世纪初 Brunelleschi 提出针孔透视投影模型（或称中心透视投影），该模型尽管简单，但可以较为全面的模拟成像过程，且数学模型易于推导<sup>[43]</sup>。

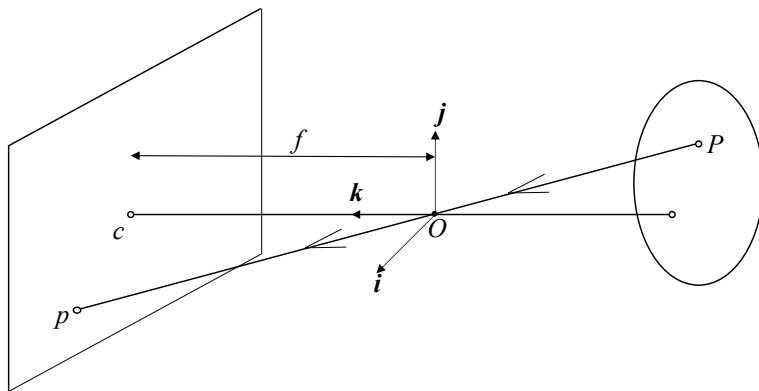


图 2-2 透视投影方程推导示意图

如图2-2所示，假设真实世界物体  $P$  在世界坐标系下坐标为  $(X, Y, Z)$ ，物体透过小孔  $O$  在相机平面上成像为  $p$ ，坐标为  $(x, y, z)$ ，世界坐标系原点位于小孔处，而物像  $p$  处在相机平面上，因此有  $z = f$ ， $f$  为相机焦距。又因为  $P, O, p$  三点共线，则存在  $\overrightarrow{Op} = \lambda \overrightarrow{OP}$ ，其中  $\lambda$  为缩放系数，故有：

$$\begin{cases} x = \lambda X \\ y = \lambda Y \\ z = \lambda Z \end{cases} \iff \lambda = \frac{x}{X} = \frac{y}{Y} = \frac{z}{Z} \quad (2-1)$$

因此有：

$$\begin{cases} x = f \frac{X}{Z} \\ y = f \frac{Y}{Z} \end{cases} \quad (2-2)$$

在齐次坐标系下物体  $P$  可以表示为  $P = (X, Y, Z, 1)^T$ ，对应物像点  $p$  在相机平面齐次坐标下可以写作  $P = (u, v, 1)^T$ ，代入公式 (2-2) 中并写作矩阵形式有：

$$Z \begin{bmatrix} u \\ v \\ 1 \end{bmatrix} = \begin{bmatrix} f & 0 & 0 & 0 \\ 0 & f & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} X \\ Y \\ Z \\ 1 \end{bmatrix} \quad (2-3)$$

由于相机成像面坐标系单位是像素并非米等单位，且像素一般为长方形，因此需要两个额外的比例因子对单位进行转换  $f_u = \frac{f}{d_u}$ ,  $f_v = \frac{f}{d_v}$ 。同时相机成像面坐标系

原点一般位于左上角而非光轴与平面交点处  $(u_0, v_0)$ , 因此我们有:

$$Z \begin{bmatrix} u \\ v \\ 1 \end{bmatrix} = \begin{bmatrix} f_u & 0 & u_0 & 0 \\ 0 & f_v & v_0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} X \\ Y \\ Z \\ 1 \end{bmatrix} \quad (2-4)$$

忽略制作工艺误差导致坐标系轴不正交的可能性,  $f_u, f_v, u_0, v_0$  称为相机内参数, 共同构成的矩阵称为相机内部矫正矩阵, 用于将相机坐标系  $C$  下坐标转换为图像平面坐标。当世界坐标系  $w$  与相机坐标系  $C$  不同时需要引入外参数进行变换, 例如将激光雷达获取的点云映射到图像时首先需要对齐坐标。令  ${}^C P$  表示相机坐标系下齐次坐标,  ${}^w P$  表示世界坐标系下齐次坐标, 由于  $C$  和  $w$  之间坐标变换为刚体变换, 因此满足下式:

$${}^C P = \begin{pmatrix} R & t \\ 0^T & 1 \end{pmatrix} {}^w P \quad (2-5)$$

其中  $R$  为旋转矩阵,  $t$  为平移量。相机内外参数共同构成任意原点世界坐标系到相机成像平面坐标系的转换矩阵如下所示。

$${}^C Z \begin{bmatrix} u \\ v \\ 1 \end{bmatrix} = \begin{bmatrix} f_u & 0 & u_0 & 0 \\ 0 & f_v & v_0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} R & t \\ 0^T & 1 \end{bmatrix} \begin{bmatrix} {}^w X \\ {}^w Y \\ {}^w Z \\ 1 \end{bmatrix} \quad (2-6)$$

### 2.1.2 仿射变换

仿射变换普遍存在于图像处理的各个领域, 如标定透视相机, 从运动中恢复结构等。仿射变化包含旋转, 平移, 缩放和错切等方式, 从几何角度来说仿射变换前后直线仍保持平行, 并且直线比例不发生改变。从数学的角度, 在二维平面上仿射变换可以表示为:

$$\begin{aligned} x_2 &= a_1 x_1 + a_2 y_1 + t_x \\ y_2 &= a_3 x_1 + a_4 y_1 + t_y \end{aligned} \quad (2-7)$$

引入齐次坐标系，可以在高维度通过线性变换表达低维度的仿射变换：

$$\begin{bmatrix} x_2 \\ y_2 \\ 1 \end{bmatrix} = \begin{bmatrix} a_1 & a_2 & t_x \\ a_3 & a_4 & t_y \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ y_1 \\ 1 \end{bmatrix} \quad (2-8)$$

将上式推广到三维则有：

$$\begin{bmatrix} x_2 \\ y_2 \\ z_2 \\ 1 \end{bmatrix} = \begin{bmatrix} a_1 & a_2 & a_3 & t_x \\ a_4 & a_5 & a_6 & t_y \\ a_7 & a_8 & a_9 & t_z \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ y_1 \\ z_1 \\ 1 \end{bmatrix} \quad (2-9)$$

### 2.1.3 对极约束

人类的双眼会在大脑中产生两幅具有重叠区域的图像，通过计算重叠区域物体的位置差异，人脑可以获得强烈的深度感。立体视觉通过融合两台(多台)相机观察到的特征，利用几何属性和对极约束模拟人脑感知深度。本节以双目相机为例介绍立体视觉相关概念，在图2-3中，物体存在于点P，左右目相机光心分别位于 $O_l$ 、 $O_r$ ，两点之间的距离称为基线距离 baseline，后续可用于视差和深度的相互转换， $O_lO_RP$ 三点共同构成对极平面。 $\Phi_l$ 、 $\Phi_r$ 分别为左右目相机的虚像面，基线 $O_lO_r$ 与 $\Phi_l$ 、 $\Phi_r$ 分别的交点 $e_l$ 、 $e_r$ 称为对极点， $e_r$ 即为右目相机观察左目相机光心 $O_l$ 的投影，反之亦然。光线 $O_lP$ 、 $O_rP$ 分别与虚像面相交于点 $p_l$ 、 $p_r$ 。因为 $p_l$ 和 $p_r$ 为同一点P的成像点，则 $p_l$ 一定在与 $p_r$ 相关联的对极线 $l_r$ 上( $l_r$ 由 $O_lO_Rp_r$ 构成的平面与虚像面 $\Phi_l$ 相交确定)。这种对极约束是立体视觉感知深度的基本原理。

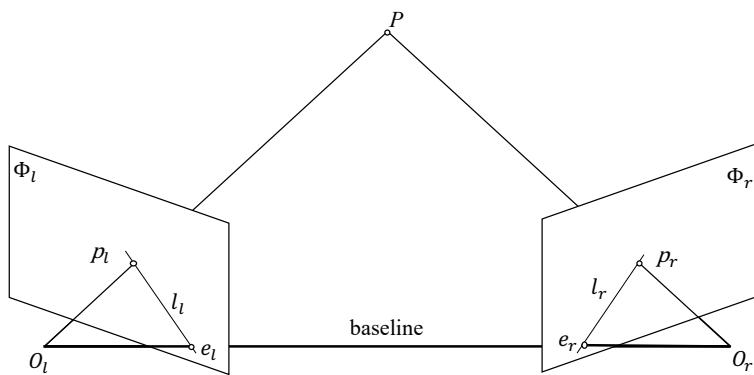


图 2-3 对极几何示意图

由上述对极约束的讨论可知，向量 $\overrightarrow{O_l p_l}$ 、 $\overrightarrow{O_r p_r}$ 和 $\overrightarrow{O_l O_r}$ 共面。等价于其中一个

向量在其他两个向量组成的平面上，即：

$$\overrightarrow{O_l p_l} \cdot [\overrightarrow{O_l O_r} \times \overrightarrow{O_r p_r}] = 0 \quad (2-10)$$

两个相机的坐标之间可通过外参数建立联系，因此式 (2-10) 可以改写如下：

$$p_l \cdot [t \times R p_r] = 0 \quad (2-11)$$

其中  $p_l, p_r$  分别是点  $p_l$  和  $p_r$  的齐次图像坐标向量。 $t$  是两坐标系间的距离  $O_l O_r$ ,  $R$  是旋转矩阵，在右目图像坐标系的自由向量  $w_r$  可以通过  $R$  变换到左目坐标系中的坐标  $Rw_r$ 。式 (2-11) 可进一步改写为：

$$p_l^T \mathcal{E} p_r = 0 \quad (2-12)$$

其中  $\mathcal{E} = [t \times]R$ ,  $[a \times]x = a \times x$  表示向量  $a$  和  $x$  叉乘。Longuet Higgins 在 [44] 中首次提出了本征矩阵的概念，即上述中的  $\mathcal{E}$ 。令  $l = \mathcal{E} p_r$ ，则  $l$  可认作左目图像中与点  $p_r$  对应的对极线  $l_l$  的自由坐标向量，因此  $p_l l = 0$  可以理解为点  $p_l$  在对极线  $l_l$  上，这一结果符合上述对对极线的讨论。本征矩阵适用于归一化图像坐标系，在原始图像坐标系中可以通过引入内参数推导两点间的关系。通常我们有：

$$\begin{aligned} {}^C p_l &= K_l p_l \\ {}^C p_r &= K_r p_r \end{aligned} \quad (2-13)$$

其中  $K_l$  和  $K_r$  分别为左右目相机内参数，通常可逆，则有：

$$\begin{aligned} p_l &= K_l^{-1} {}^C p_l \\ p_r &= K_r^{-1} {}^C p_r \end{aligned} \quad (2-14)$$

代入式 (2-12) 可得：

$${}^C p_l^T \mathcal{F} {}^C p_r = 0 \quad (2-15)$$

其中  $\mathcal{F} = K_l^{-T} \mathcal{E} K_l^{-1}$  称为基础矩阵。通常相机内外参数可以通过预先标定获取，在不知道物体世界坐标位置时，通过内外参数和对极约束可以构建两图像点间强约束关系。

## 2.2 深度学习

深度学习的概念最早可以追溯到 1943 年的 McCulloch-Pitts 神经元模型<sup>[45]</sup>，该模型通过预先设计好的系数对一组量化输入进行线性组合计算输出，并根据输出正负判断输入数据类型。随后自适应线性单元模型，多层次感知机模型和反向传播技术的提出使得模型有了一定的学习能力。但受限于当时计算能力不足和数据集匮乏，深度学习在很多重要问题上表现不如机器学习中的如支持向量机和图模型等方法。直到 2006 年 Hinton 提出可以通过“贪婪逐层预训练”方法训练深度信念网络<sup>[46]</sup>，深度学习的概念才逐渐进入人们的视野。下图反应了深度学习技术三次发展浪潮的重要节点。

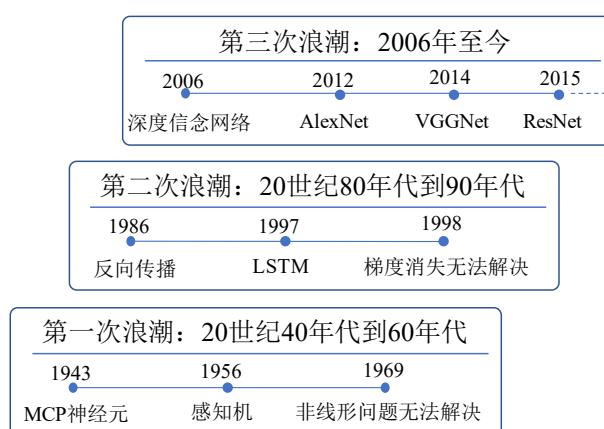


图 2-4 深度学习三次发展浪潮

### 2.2.1 基础理论

人工智能已经成为工业界和学术界的热点研究领域，在早期人工智能可以轻松解决对人类来说较为困难的某些问题，其真正挑战是利用计算机解决对人类来说容易执行，但难以规则化的问题，如识别图中汽车，理解语音中指令。最早研究者通过硬编码现实中相关知识并设立规则指导计算机学习，这种方法称为知识库方法。然而并非所有的事件都可以规则化表示，人工智能需要从原始数据中获得模式表征的能力。机器学习通过人为设置的特征模板提取数据中的高维信息进行决策。在现实生活面临的许多任务中，我们无法提前获知需要提取的特征，因此我们需要机器自身学习需要表征的特征，即表示学习。表示学习通过发掘自身表示形式获得较好的特征表示，深度学习正是通过逐层学习最优特征，利用简单表示逐渐形成复杂特征，解决了表示学习中如何发掘自身表示形式这一核心问题。

深度学习在现代社会中的含义已经超过其在机器学习领域中神经科学观点，研究的重点由最初模拟人脑学习过程转变为学习多层次特征组合表征这一更朴实

的原理。就深度学习自身发展而言，一般认为深度学习经历了三次发展浪潮：20世纪40年代到60年代，深度学习的雏形出现在控制论中；20世纪80年代到90年代，深度学习表现为联结主义；直到2006年，才真正以深度学习之名复兴<sup>[47]</sup>。

深度学习发展的第一次浪潮称为控制论，在20世纪40年代，研究者们研究如何利用简单线性模型模拟人脑学习过程，对于一组输入  $x_1, \dots, x_n$  和权重  $w_1, \dots, w_n$ ，计算加权和作为输出  $f(x, w) = x_1w_1 + \dots + x_nw_n$ 。在MCP神经元模型中通过判断输出  $f(x, w)$  正负推断输入类别，但MCP神经元模型需要人工设定正确权重系数。在1956年，Rosenblatt提出感知机模型<sup>[48]</sup>，可以通过每个类别的输入样本更新权重。简单线性模型存在较多问题，其中较为著名的是单层线性感知机无法学习异或函数，虽然两层感知机可以有效解决该类问题，但在当时却导致了深度学习第一次发展浪潮的大衰退。

在现代，神经科学被视为深度学习的重要起源，但已经不是该领域的主要指导思想。早在20世纪80年代，研究者已经从多层次特征表征的角度思考深度学习，引发了深度学习第二次发展浪潮，联结主义。联结主义核心思想是通过连接大量简单计算单元实现智能行为，系统每一个输入都应该从多个特征维度去表征，每一个特征都可能参与到多个高维特征计算中。在如今深度学习中广为流行的反向传播算法也是在深度学习第二次发展浪潮中得以推广，但基于链式法则的梯度反向传播在当时也导致了梯度消失的问题，同时以核方法为核心的支撑向量机等机器学习模型在一些重要问题上大放光彩导致了深度学习热潮的第二次衰退。

直到2006年Hinton提出了“贪婪逐层训练”策略<sup>[46]</sup>，解决了深度信念网络训练困难的问题，深度学习逐渐开启了第三次发展浪潮。由于解决了梯度消失和梯度爆炸的问题，随着层数的加深，多层感知机模型逐渐发展成了今天的深度学习模型。深度学习从人类可以理解的输入中学习复杂表示，简单来说在低层网络中学习输入的局部特征如物体的边和角，高层网络对局部特征进行综合做出判断。开始学习时，网络中的系数进行随机初始化，网络提取随机特征做出判断，通过与预定义的标签做比较并将误差反向传播得到梯度，利用误差梯度和一定的规则向误差更小的方向更新系数。经过多轮迭代，深度学习模型可以学习到正确的特征表征并对未知数据做出判断。

## 2.2.2 卷积神经网络

卷积神经网络是目前深度学习领域的研究热点，自AlexNet<sup>[49]</sup>以来研究者们提出了各种优秀的卷积神经网络模型，如ResNet<sup>[50]</sup>和DenseNet<sup>[51]</sup>等。图2-5为经典的LeNet-5<sup>[52]</sup>的网络模型结构，本节以LeNet-5为例说明卷积神经网络中常见

的卷积层，池化层，全连接层等基本计算单元。

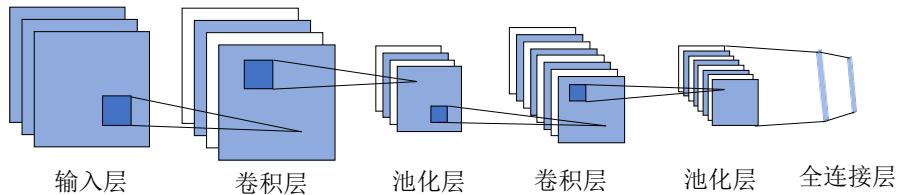


图 2-5 LeNet-5 网络结构

### 2.2.2.1 卷积层

卷积层是卷积神经网络的核心，在网络中起到提取特征，增加感知野等作用。简单来说，卷积层将卷积核（滑动窗口）在二维图像上滑动，并计算每个位置加权和作为输出，如图2-6所示。

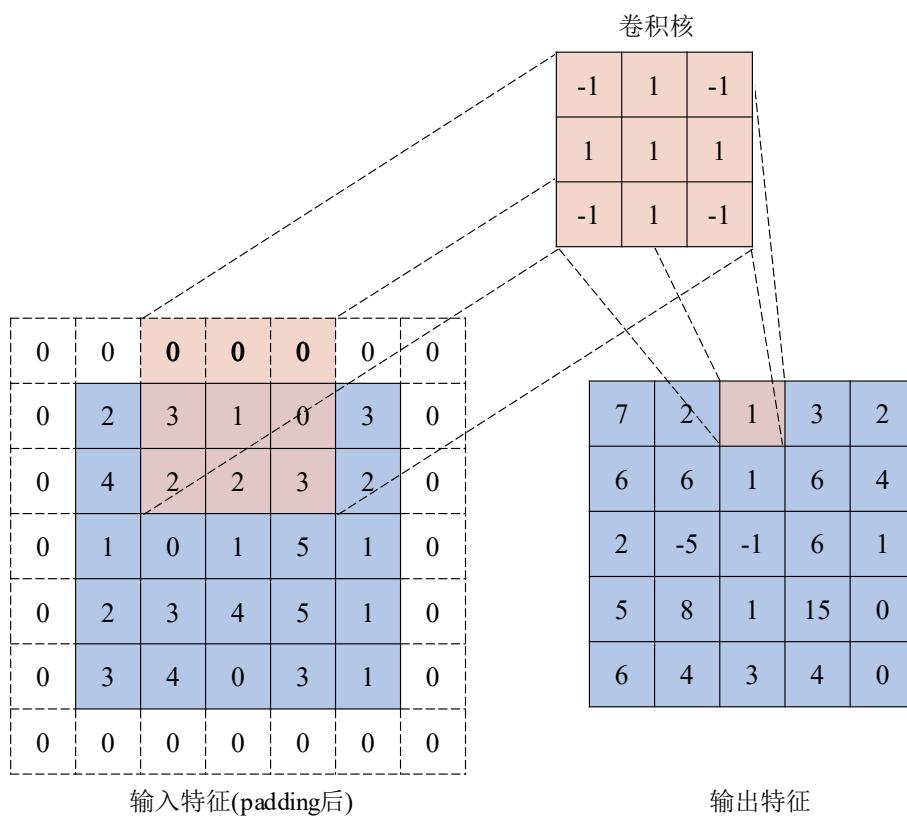


图 2-6 卷积示意图

在图2-6的示例中，输入特征的单个通道为图中左边蓝色标注的  $5 \times 5$  方格，卷积核为上方淡红色标注的  $3 \times 3$  方格。卷积过程为以卷积核中心为锚点，将卷积核在输入特征上按照一定规律逐行逐列滑动，由于输入特征边缘位置缺少足够像

素参与计算，可以通过补零的 padding 方式保证卷积后特征大小不发生改变。令  $f_{input}(u, v)$  表示输入特征中位置为  $(u, v)$  的特征值， $k(u, v)$  表示卷积核中位置为  $(u, v)$  的系数， $f_{output}(u, v)$  表示输出特征在位置  $(u, v)$  的卷积值， $M, N$  分别为卷积核的长宽，则：

$$f_{output}(u, v) = \sum_{m=-\frac{M}{2}}^{\frac{M}{2}} \sum_{n=-\frac{N}{2}}^{\frac{N}{2}} k\left(\frac{M}{2} + m, \frac{N}{2} + n\right) f_{input}(u + m, v + n) \quad (2-16)$$

为方便计算卷积核的中心位置， $M, N$  一般取相等奇数。由上式不难发现，对于输入特征每个位置，卷积核系数是一样的，这种共享参数的方式称为权重共享。权重共享允许在只保留  $MN$  个参数的情况下对整个输入计算特征，而  $MN$  一般远小于输入特征的元素总数，可以显著降低模型存储需求。

在实际应用中，出于对网络感知野和计算量的考虑（感知野一般指为了计算某层特征的某个元素值，网络输入中需要参与计算的元素个数，可通过反向逐层追溯参与卷积运算的元素求得），卷积核完成当前位置的卷积计算后需向后滑动  $S$  个像素才能进行下一次计算， $S$  称为卷积步长。对于图像边缘位置可以选择不添加额外元素或者添加  $P$  个元素进行 padding 实现对输出特征尺寸的控制，其中 padding 添加的元素可以是 0 或边缘值的复制。对于尺寸为  $H \times W$  的输入特征，采用大小为  $K \times K$  的卷积核，卷积步长为  $S$ ，边缘填充  $P$  个元素，则卷积后特征的大小为：

$$H' = \frac{H - K + 2P}{S} + 1 \quad (2-17)$$

$$W' = \frac{W - K + 2P}{S} + 1 \quad (2-18)$$

通过适当调整  $S$  和  $K$  的大小可以保证输入特征每个元素都参与计算，且显著降低计算量并增加感知野。

上述讨论仅仅针对单通道特征卷积，单通道卷积由于权重共享只能提取一种特征，这对于诸如目标识别，语义分割，立体匹配等复杂的计算机视觉任务来说是远远不够的。为了使网络具有提取复杂特征的能力，可采用多个卷积核在输入特征的多个通道上同时进行卷积，卷积结果的加权和作为输出的单通道特征。为了得到具有  $Ch_{out}$  个通道的输出特征，需将上述卷积过程重复若干次。根据输入特征通道数  $Ch_{in}$  和输出特征通道数  $Ch_{out}$ ，一个卷积层所需的卷积核数目  $N_{kernel}$  为：

$$N_{kernel} = Ch_{in} \times Ch_{out} \quad (2-19)$$

因此在不考虑偏置的情况下，一个卷积层的参数量  $N_{para}$  为：

$$N_{para} = K \times K \times Ch_{in} \times Ch_{out} \quad (2-20)$$

实际应用中， $K$  一般取 3, 5, 7 等较小的数，因此  $N_{para}$  要远远小于输入特征的元素个数。因为卷积核系数同图像位置无关，单层卷积保存  $N_{para}$  个参数即可在任意尺寸的输入上提取特征。就像 MCP 神经元无法解决异或问题一样，单层卷积提取特征语义的能力有限。通过设计多层卷积，网络可以将浅层的边角，纹理等人类可以理解的、规则化的特征向高维映射，逐步提升为适合任务的、可以计算的复杂特征。

### 2.2.2.2 池化层

多年前，摄影领域专业器材单镜头反光式取景照相机(单反)的分辨率鲜有达到 4K 水平(一般指  $4096 \times 2160$  像素分辨率)，现如今手机内置相机的分辨率已经达到 4K 画质。高分辨率影像在生活中随处可见，但仅仅是一张  $512 \times 512$  的图像经过卷积层将通道拓展到 256 后，元素总量高达 6000 万，以 32 位浮点型存放在显存中需要占用 256MB 显存，现代深度卷积神经网络一般包含几十层卷积，并且需要多个样本同时训练优化梯度曲线。若不对卷积后特征进行一定处理，将耗费大量计算资源和显存资源。

一般来说，图像非边缘部位像素值是近似的，经过卷积后的值也是近似的，这意味着在浅层特征中包含大量冗余信息，我们可以通过池化层保留局部特征的关键信息。池化层可视为只有一个卷积核的特殊卷积层，只是在窗口内计算的不再是系数加权和，而是根据池化方式不同采取不同计算方式。常见的池化层有最大值池化和均值池化。对于最大值池化，只保留窗口内最大值作为输出，即：

$$f_{output}(u, v) = \max_{-\frac{K}{2} \leq m, n \leq \frac{K}{2}} f_{input}(u + m, v + n) \quad (2-21)$$

对于均值池化，输出为窗口内所有特征的平均值，即：

$$f_{output}(u, v) = \frac{1}{K \times K} \sum_{m=-\frac{K}{2}}^{\frac{K}{2}} \sum_{n=-\frac{K}{2}}^{\frac{K}{2}} f_{input}(u + m, v + n) \quad (2-22)$$

同时，为了减少池化后特征含有的冗余信息，一般将池化步长(可视为卷积步长)设置为核的大小，保证每个核覆盖元素不重复，每个元素仅参与一次池化计算。

如图2-7所示，一个输入为  $4 \times 4$  的单通道特征经过核大小为 2，步长为 2 的最大值池化后仅为  $2 \times 2$  大小，体积降低了 4 倍并且保留了每个区域最大特征。卷积

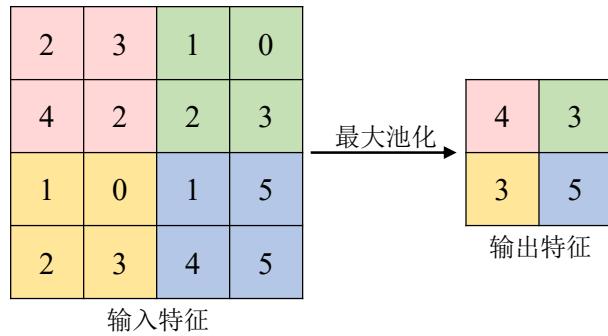


图 2-7 最大池化示意图

层输出可视为输入特征与模板系数的相关性，从线性系统的角度来看又可认为是输入特征对卷积核的响应，在浅层特征中，可认为最大池化选择了边角等与特征模板有较高相关性的区域，用于后续特征计算。池化层在降低特征大小的同时也会导致细节的丢失，核的大小需要谨慎设计，若上图2-7中采用的核大小为4，最终输出只有“5”一个元素，其余特征信息则被丢弃，可能导致网络无法正常收敛。

### 2.2.2.3 全连接层

全连接层多出现在分类任务和自然语言处理中，一般用于输出最后的分类结果。全连接是指输入的每个特征元素都通过系数与每个输出元素连接，一般认为全连接层在整个卷积网络中起到将高维特征映射到样本标记空间的作用。

图2-8是一个由全连接层构成的单隐层前馈网络，输入层和隐藏层，隐藏层和输出层之间采用全连接层计算。卷积特征经过重新排列后作为输入层，通过全连接层前向传递给隐藏层。对于第*i*层的第*j*个元素 $x_{ij}$ ，是由第*i-1*层所有元素加权计算得来，即：

$$x_{ij} = \sum_{k=1}^K w_{kj} x_{i-1k} \quad (2-23)$$

输入特征通过全连接层逐层映射到输出结果。

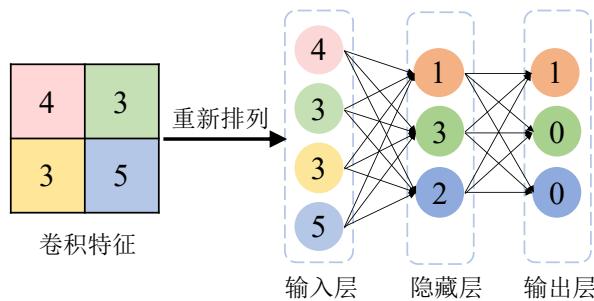


图 2-8 隐藏层示意图

多个全连接层叠加后仍可视为线性系统，但很多任务并非是线性可分的，因此我们需要在全连接层之间引入非线性变换，增加网络表达非线性函数的能力。这种非线性变化称为激活函数，一般为连续可微的非线性函数。常见的激活函数有 Sigmoid, Tanh, ReLU 和 LeakyReLU 等，定义如下：

$$f_{Sigmoid}(x) = \frac{1}{1 + e^{-x}} \quad (2-24)$$

$$f_{Tanh}(x) = \frac{1 - e^{-2x}}{1 + e^{-2x}} \quad (2-25)$$

$$f_{ReLU}(x) = \max(0, x) \quad (2-26)$$

$$f_{LeakyReLU}(x) = \begin{cases} x & x > 0 \\ \lambda x & x \leq 0 \end{cases} \quad (2-27)$$

### 2.2.3 反向传播算法

单层线性网络的训练可以直接利用误差的梯度更新参数，现代深度学习网络层数一般高达几十层，简单的参数更新方式已不再适用，需要更强大的学习算法。逆误差反向传播 (Back Propagation, BP) 算法<sup>[53]</sup> 是目前深度学习中应用最为广泛的神经网络学习算法，利用链式法则将误差逐层传递到输入，并根据选择的优化器规则对网络中参数进行更新，具体如下。

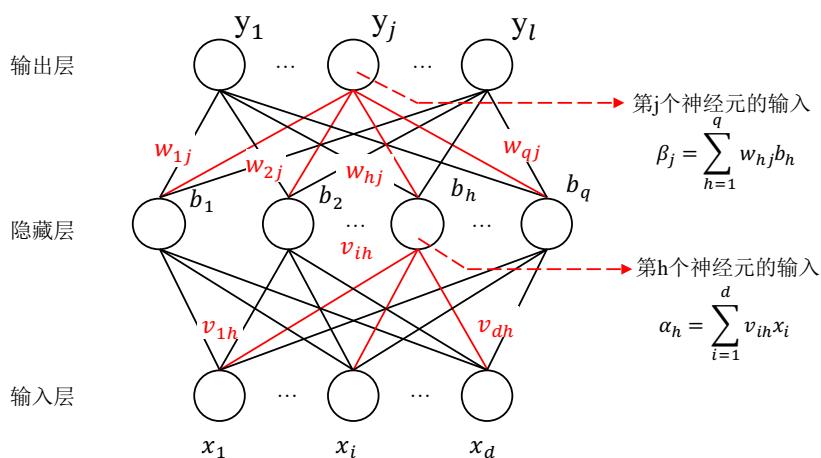


图 2-9 反向传播示意图

图2-9是一个单隐层前馈网络，输入层拥有  $d$  个神经元，隐藏层具有  $q$  个神经

元，输出层由  $l$  个神经元构成。对输入样本  $\mathbf{x} = (x_1, x_2, \dots, x_d)$ ，网络预测的输出为  $\hat{\mathbf{y}} = (\hat{y}_1, \hat{y}_2, \dots, \hat{y}_l)$ ，即：

$$\hat{y}_j = f(\beta_j - \theta_j) \quad (2-28)$$

其中， $f$  为输出层激活函数， $\theta_j$  为第  $j$  个输出神经元阈值， $\beta_j$  为第  $j$  个神经元输入，即：

$$\beta_j = \sum_{h=1}^q w_{hj} b_h \quad (2-29)$$

其中  $b_h$  为隐藏层第  $h$  个神经元输出， $w_{hj}$  为  $b_h$  到输出层第  $j$  个神经元输入的权重。同样，对于隐藏层中第  $h$  个神经元，阈值为  $\gamma_h$ ，输入为  $\alpha_h$ ，则有：

$$b_h = f(\alpha_h - \gamma_h) \quad (2-30)$$

$$\alpha_h = \sum_{i=1}^d v_{ih} x_i \quad (2-31)$$

其中  $v_{ih}$  为输入样本第  $i$  维特征值到隐藏层第  $h$  个神经元的权重。在整个网络训练中采用均方误差作为目标函数，则误差为：

$$E = \frac{1}{2} \sum_{j=1}^l (\hat{y}_j - y_j)^2 \quad (2-32)$$

$y_j$  为样本标注的第  $j$  维。为了使均方误差最小，可以沿着负梯度方向更新参数，给定学习率  $\eta$ ，则更新量为：

$$\Delta w_{hj} = -\eta \frac{\partial E}{\partial w_{hj}} \quad (2-33)$$

根据链式法则：

$$\frac{\partial E}{\partial w_{hj}} = \frac{\partial E}{\partial \hat{y}_j} \cdot \frac{\partial \hat{y}_j}{\partial \beta_j} \cdot \frac{\partial \beta_j}{\partial w_{hj}} \quad (2-34)$$

对于  $\frac{\partial \beta_j}{\partial w_{hj}}$ ，根据式 (2-29)，显然有：

$$\frac{\partial \beta_j}{\partial w_{hj}} = b_n \quad (2-35)$$

令激活函数  $f$  采用 Sigmoid 函数, 则其导数有:

$$\begin{aligned} f'(x) &= \frac{e^{-x}}{(1 + e^{-x})^2} \\ &= \frac{1 + e^{-x} - 1}{(1 + e^{-x})^2} \\ &= f(x)(1 - f(x)) \end{aligned} \quad (2-36)$$

根据式 (2-31) 和 (2-32) 有:

$$\begin{aligned} g_j &= -\frac{\partial E}{\partial \hat{y}_j} \cdot \frac{\partial \hat{y}_j}{\partial \beta_j} \\ &= -(\hat{y}_j - y_j) f'(\beta_j - \theta_j) \\ &= \hat{y}_j(1 - \hat{y}_j)(y_j - \hat{y}_j) \end{aligned} \quad (2-37)$$

综上, 可以得到:

$$\Delta w_{hj} = \eta g_j b_h \quad (2-38)$$

在输入层和隐藏层之间重复以上推导可以得到:

$$\Delta \theta_j = \eta g_j \quad (2-39)$$

$$\Delta v_{ih} = \eta e_h x_i \quad (2-40)$$

$$\Delta \gamma_h = -\eta e_h \quad (2-41)$$

其中:

$$\begin{aligned} e_h &= -\frac{\partial E}{\partial b_h} \cdot \frac{b_h}{\partial \alpha_h} \\ &= b_h(1 - b_h) \sum_{j=1}^l w_{hj} g_j \end{aligned} \quad (2-42)$$

综上我们得到了网络中所有参数的更新方法, 在训练时首先将输入在网络中前向传递并计算损失函数, 得到误差后按照公式逐层更新参数, 经过多轮迭代可最小化损失函数。上述讨论仅针对单个样本训练, 在实际应用中为了防止样本间梯度相互抵消导致计算资源浪费, 多采用累积多个样本梯度后更新一次参数的方式。

## 2.3 双目立体匹配

随着深度学习在图像处理中应用越来越广泛，传统双目立体匹配技术中各个步骤也逐渐被深度学习替代。得益于深度学习自我发掘特征表征的能力，基于深度学习的双目立体匹配技术在多个评价指标上均超越人工设计特征的传统技术。基于深度学习的双目立体匹配技术的网络设计也遵循着传统技术的基本步骤，即匹配代价计算，匹配代价聚合和视差预测，如图2-10所示。本节将从传统特征和深度学习两个方向介绍双目立体匹配技术。

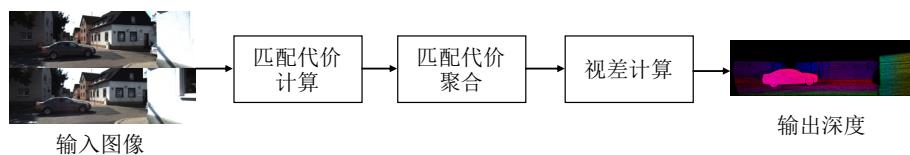


图 2-10 双目立体匹配流程图

### 2.3.1 基于传统特征的双目立体匹配

基于传统特征的双目立体匹配技术中应用较为广泛的是半全局匹配算法(Semi-Global Matching, SGM)<sup>[36]</sup>，该算法在计算局部特征匹配的同时引入全局信息使得算法在弱纹理和重复区域表现较好，同时利用多个一维方向近似二维代价聚合，有效降低了问题的复杂度。

双目匹配的第一步是匹配代价计算，一个简单想法是比较像素强度大小，这种方法易受到采集设备和光照变化的影响。Heiko 提出可以利用互信息熵进行匹配<sup>[36]</sup>，两块区域越匹配联合熵越低。具体来说，两幅图像的互信息熵  $M|I_1, I_2$  定义为：

$$M|I_1, I_2 = H_{I_1} + H_{I_2} - H_{I_1, I_2} \quad (2-43)$$

其中，图像的熵由像素强度的分布概率计算得到：

$$H_I = - \int_0^1 P_I(i) \log P_I(i) di \quad (2-44)$$

$$H_{I_1, I_2} = - \int_0^1 \int_0^1 P_{I_1, I_2}(i_1, i_2) \log P_{I_1, I_2}(i_1, i_2) di_1 di_2 \quad (2-45)$$

当两个区域匹配度增加，表明用一个区域预测另一个区域的未知量减少，两个区域提供的信息量降低，则互信息熵  $H_{I_1, I_2}$  降低， $M|I_1, I_2$  增加。上式无法直接用于离

散的图像计算, Kim 等人<sup>[54]</sup>提出可以通过泰勒展开将积分转化为求和形式, 即:

$$H_{I_1, I_2} = \sum_p h_{I_1, I_2}(I_{1p}, I_{2p}) \quad (2-46)$$

其中:

$$h_{I_1, I_2}(i, k) = -\frac{1}{n} \log(P_{I_1, I_2}(i, k) \otimes g(i, k)) \otimes g(i, k) \quad (2-47)$$

$$P_{I_1, I_2}(i, k) = \frac{1}{n} \sum_p T[(i, k) = (I_{1p}, I_{2p})] \quad (2-48)$$

$\otimes$  表示卷积,  $g(i, k)$  表示高斯核,  $T[\cdot]$  表示判真函数, 条件为真值为 1, 否则值为 0。最终, 互信息熵损失的定义为:

$$C_{MI}(\mathbf{p}, d) = -mi_{I_b, f_D(I_m)}(I_{bp}, I_{mq}) \quad (2-49)$$

$$mi_{I_1, I_2}(i, k) = h_{I_1}(i) + h_{I_2}(k) - H_{I_1, I_2}(i, k) \quad (2-50)$$

$$\mathbf{q} = e_{bm}(\mathbf{p}, d) = [p_x - d, p_y]^T \quad (2-51)$$

$I_{bp}, I_{mq}$  分别表示基本图像和匹配图像,  $\mathbf{q}$  和  $\mathbf{p}$  为坐标向量。 $f_D(\cdot)$  表示使用视差进行重建, 初始视差来自随机生成, 该算法需要多次迭代。

像素级别的匹配代价计算易受到噪声干扰, 导致错误匹配的损失低于正确匹配, 因此需要对局部视差的梯度变化添加惩罚项, 保证估计视差的平滑化, 可在视差图  $D$  上定义如下能量函数  $E(D)$ :

$$E(D) = \sum_p \left( C(\mathbf{p}, \mathbf{D}_p) + \sum_{q \in N_p} P_1 T[|D_p - D_q| = 1] + \sum_{q \in N_p} P_2 T[|D_p - D_q| > 1] \right) \quad (2-52)$$

最外层括号中第一项为像素级匹配代价函数; 第二项为对像素周围小视差的惩罚项, 较小的系数  $P_1$  可以调整算法更好的适应曲面; 第三项为对所有视差变化的惩罚项, 系数  $P_2$  控制算法在物体边缘处的性能。最小化式 (2-52) 可以找到最佳匹配的视差图  $D$ , 但形如式 (2-52) 的全局最优化问题需要在二维空间上进行搜索, 为了简化计算可以用多个一维方向上的搜索近似。因此式 (2-52) 可以转变为:

$$S(\mathbf{p}, d) = \sum_r L_r(\mathbf{p}, d) \quad (2-53)$$

$$\begin{aligned}
 L_r(\mathbf{p}, d) &= C(\mathbf{p}, d) + \min(L_r(\mathbf{p} - \mathbf{r}, d), \\
 &L_r(\mathbf{p} - \mathbf{r}, d - 1) + P_1 \\
 &L_r(\mathbf{p} - \mathbf{r}, d + 1) + P_1 \\
 &\min_i L_r(\mathbf{p} - \mathbf{r}, i) + P_2) - \min_k L_r(\mathbf{p} - \mathbf{r}, k)
 \end{aligned} \tag{2-54}$$

$\mathbf{r}$  为二维平面上的方向向量, 如水平方向, 垂直方向, 45 度方向等。利用动态规划最小化上式可求得视差图  $D$ 。

### 2.3.2 基于深度学习的双目立体匹配

早期, 深度学习仅用来代替匹配代价计算或代价聚合, 后来研究者们遵循光流研究的思路提出了端到端的双目立体匹配卷积神经网络。初期的立体匹配网络虽然能在输入图像一半分辨率的尺寸上预测视差, 但精度较低且网络泛化性能较差。随着 cost-volume 和 3D 卷积的引入, 双目立体匹配卷积神经网络的视差预测性能有了显著提高。

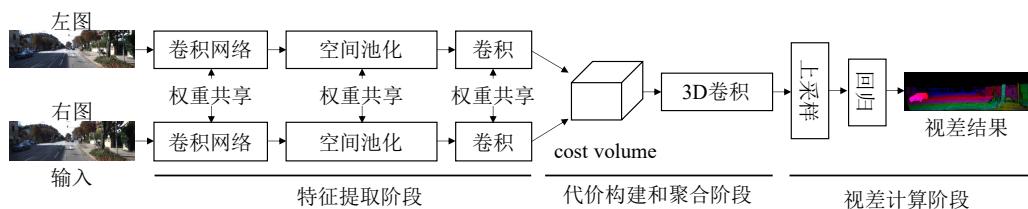


图 2-11 典型的双目立体匹配网络结构

一个典型的双目立体匹配网络结构如图2-11所示, 输入左右目图像首先通过多个卷积层和池化层提取双目特征, 接着一般通过平移和级联的方式构建 cost-volume, 并通过 3D 卷积聚合 cost-volume, 最后将聚合特征上采样到原始尺寸通过回归网络预测视差。在特征提取阶段双目图像首先被送入一个小型的卷积神经网络, 主要是提取图像的边角, 纹理等特征。接着特征被送入空间金字塔池化层, 目的是从多个尺度提取纹理特征, 构建物体(如汽车)和物体部件(如汽车的轮胎、窗户)之间的关系。上述处理一般采用权重共享网络对左右目图像分别处理, 采用权重共享的原因是希望在左右目图像中相同的物体可以具有近似的语义特征。在特征聚合阶段, 首先利用提取特征构建 cost-volume, 目前常用的做法是将特征复制多次, 分别在视差方向上平移不同距离然后级联形成新的特征, 接着利用 3D 卷积聚合潜在视差特征。在视差计算阶段, 目前多采用估计视差分布概率的方式实现亚像素级预测精度。

与早期深度模型相比, 具有极线约束的理论支撑, 并且遵循匹配代价计算, 聚

合和视差预测思路设计的网络，在普通场景的预测精度具有显著的精度。双目立体匹配的目的是提取场景三维信息，为后续高层视觉任务如机器人避障，路径规划等提供感知环境的基础。这些场景不可避免的会出现过曝场景，现有算法在过曝场景下都不可避免的会出现错误匹配，这些错误匹配严重影响后续任务的稳定性，因此我们需要研究在过曝场景下如何提高双目立体匹配的精度。

## 2.4 本章小结

本章从小孔成像模型出发，分析了世界坐标系与相机成像面坐标系的相对位置关系，接着通过对极约束建立了自由物体与双目图像视差间关系，为从双目图像获取物体深度信息提供了理论支撑。接着介绍了目前深度学习中常用的卷积层，池化层和反向传播原理等基础知识，为解决本课题的问题提供了有力工具。最后分别介绍了传统的和基于深度学习的双目立体匹配技术，为后续的研究建立了基本的模型。

### 第三章 过曝场景数据集构建

得益于在虚拟数据 Sceneflow Dataset<sup>[1]</sup> 上的预训练，现有算法在 KITTI Stereo 2012<sup>[4]</sup>/2015<sup>[5]</sup> 等自动驾驶场景表现较好。然而现有公开数据集缺乏过曝场景视差标注，限制了过曝场景下双目立体匹配技术进一步的研究。本章从分析现有数据集出发，研究如何构建与列车过曝场景类似的双目立体匹配数据集。

#### 3.1 公开数据集分析

双目立体匹配技术在不断发展，双目视差数据集也在不断更新。早期的 Middlebury Stereo Datasets<sup>[6-10]</sup> 仅包含 6 例标注样本，现如今的 Sceneflow Dataset 已包含上万例标注样本。常用的双目视差数据集基本信息如表3-1：

表 3-1 常用数据集基本信息

数据集	标注样本数	场景	数据采集方式	标注完整性
Middlebury Stereo Datasets (2001-2006)	38	室内	结构光	稠密
Middlebury Stereo Datasets (2014)	33	室内	多组结构光	稠密
KITTI Stereo 2012	388	交通场景	激光雷达	稀疏
KITTI Stereo 2015	400	交通场景	激光雷达 + 模型修复	稀疏
Sceneflow Datasets	35000	虚拟场景	Blender 渲染	稠密
Virtual KITTI <sup>[12]</sup>	21260	虚拟场景	Unity 渲染	稠密

MiddleBury Stereo Dataset 是双目立体匹配领域的经典数据集，最早被用于验证基于传统特征的双目立体匹配算法的正确性。该数据集最早由 Scharstein 等人于 2002 年在文献 [6] 中发表，共计 6 例样本，随后于 2005 年和 2006 年扩充后共计 38 例样本。MiddleBury Stereo Dataset 利用结构光编码像素，通过编码信息标注两幅图像间像素对应关系。结构光法需要一对相机和至少一部投影仪，投影仪将编码后的光线投影在场景上，根据图像中的编码信息寻找匹配像素，可计算两幅图像的对应关系。该方法需要根据投影编码前后的图像强度差异获取像素编码信息，因此一般仅适用于室内场景。在光滑表面，物体反射会引起编码信息位置紊乱导致无法获得正确匹配关系，因此该方法无法用于列车高反光部件的视差标定。

KITTI Vision Benchmark 是卡尔斯鲁厄理工学院和丰田工业大学针对双目视觉、光流预测、SLAM 和 3D 物体检测而构建的自动驾驶场景数据集。该数据集由一辆装配有双目相机，激光雷达等设备的采集车采集。如图3-1所示，双目相机

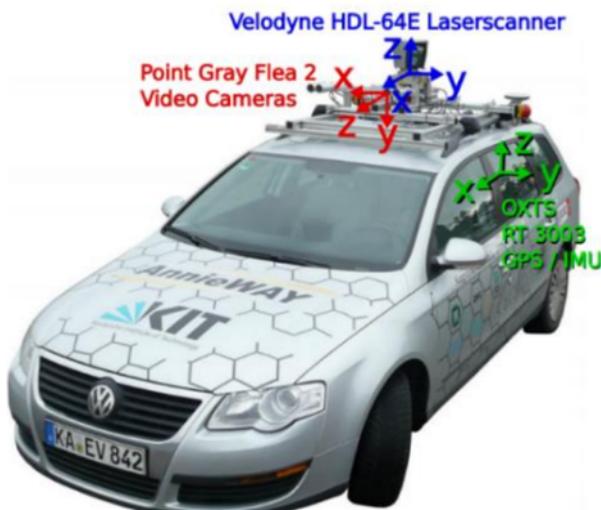


图 3-1 KITTI 数据集采集设备

通过同步设备可以同时记录当前场景的左右目图像；激光雷达同时逐行扫描，接受反射信号感知当前场景三维信息，但只能产生稀疏三维标注。根据文献 [4] 中说明，该数据集利用迭代最近点法对前后 5 帧三维数据进行配准，利用相机内外参数将三维数据投影在两幅图像中寻找对应关系，最后可获得约 50% 的标注率。基于激光雷达的标注方法虽然精度高，但无法接受高反光区域光线中的编码信息，难以用于图像过曝区域的视差标定。

受到深度学习在光流预测领域相关工作的启发，Mayer 等人认为视差也可以通过卷积神经网络进行预测，基于深度学习的卷积神经网络需要大量的数据进行训练，为了解决已有的视差数据集 (MiddleBury、KITTI) 数据量较少无法用于深度学习网络训练的问题，Mayer 等人利用三维软件 Blender 以渲染场景的方式制作了超过 30000 例双目视差标注数据 Sceneflow Dataset。根据文献 [11] 中相关实验，在合成数据 Sceneflow Dataset 上对网络进行预训练，可以提高算法在 KITTI Stereo 等真实场景数据集上的表现。虽然 Mayer 等人在渲染 Sceneflow Dataset 时设置了多种物体，但并未对环境光效做进一步控制，该数据集缺少过曝场景相关标注。

常见的双目视差数据集如图3-2所示，通过上述对已有数据集的分析，我们可以得知 KITTI Stereo Dataset 利用激光雷达采集真实场景三维数据，融合 11 帧数据后标注率约 50%，同时在物体边缘和遮挡区域存在大量未标注现象，而采用结构光的 MiddleBury Stereo Dataset 则可以获取较为稠密的视差标注，但标注过程繁琐，仅适用于室内少量样本标注。无论是结构光还是激光雷达都属于主动探测方法一种，需要工作在漫反射区域根据接收的反射光强和飞行时间计算距离。在真实场景中，过曝往往伴随着镜面反射，这意味着上述数据采集方法往往无法在过曝区

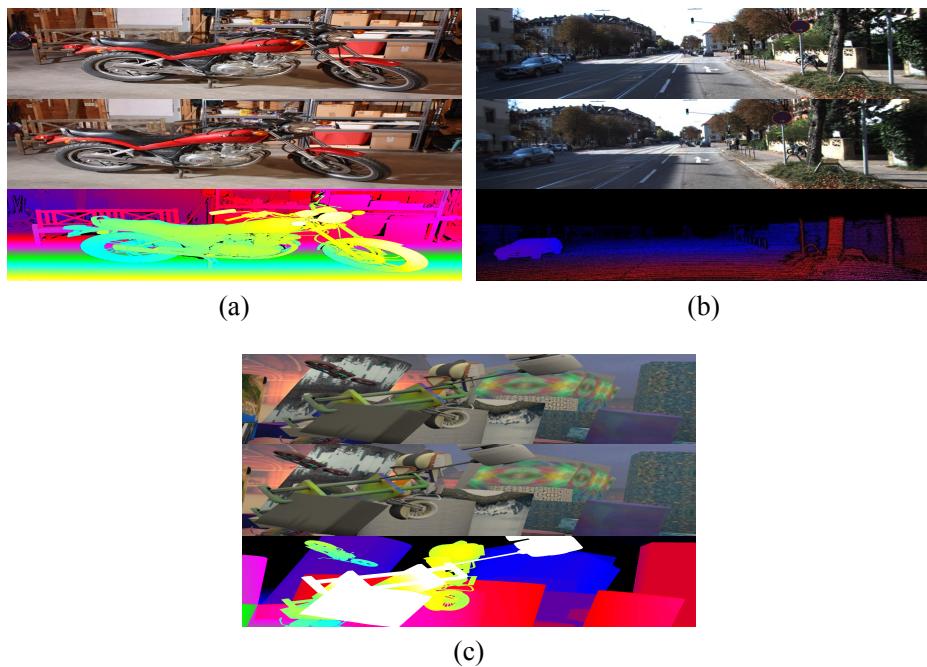


图 3-2 常见数据集样例。(a)Middlebury Stereo Dataset; (b)KITTI Stereo Dataset; (c)Sceneflow Dataset

域接收到正确反射光线。因此在真实场景(如列车高反光部件)中标定过曝区域视差存在困难,基于模型渲染的数据集制作方法可以很好的避免这一问题。但基于Blender渲染的Sceneflow Dataset旨在为深度学习方向的双目立体匹配提供通用的预训练数据,缺少针对过曝现象的渲染场景及其标注数据。综上,现有双目视差数据集缺少过曝场景及其标注信息,为了研究过曝场景下的双目立体匹配技术,实现列车高反光部件的测量,我们首先需要解决缺少相关场景数据集的问题。

### 3.2 数据采集方法

Sceneflow Dataset 和 Virtual KITTI Dataset 通过实验向我们证明了两点:一是在合成数据训练的网络直接应用在真实场景中的效果是可以接受的;二是在合成数据上预训练的网络在真实场景数据上微调(Fine Tune)后表现更好<sup>[11,12]</sup>。考虑到真实场景难以获取过曝区域的视差标注,本课题选择合成数据的方式构建过曝场景数据集,一是为过曝场景下双目立体匹配技术的性能研究提供标注数据,进行定量评估;二是为列车过曝场景提供预训练权重,给予模型过曝区域匹配的先验知识。

合成数据法一般利用3D建模软件如Blender、Unity等渲染数据,首先建立数据集模型,设计捕捉视角和物体运动轨迹,改变渲染引擎内部流水线输出RGB图

像和深度信息，最后利用相机几何模型将深度信息转换为视差标注。为了构建类似列车高反光部件测量场景的过曝数据，可对场景中物体表面给予不同的反射系数，并在相机前添加强光源保证每例样本均有过曝现象。考虑到 Blender 属于开源软件，并且提供相关 Python 接口，方便修改渲染引擎内部流水线以及控制光源的开关和位置等，本课题选择利用 Blender 进行数据集渲染。

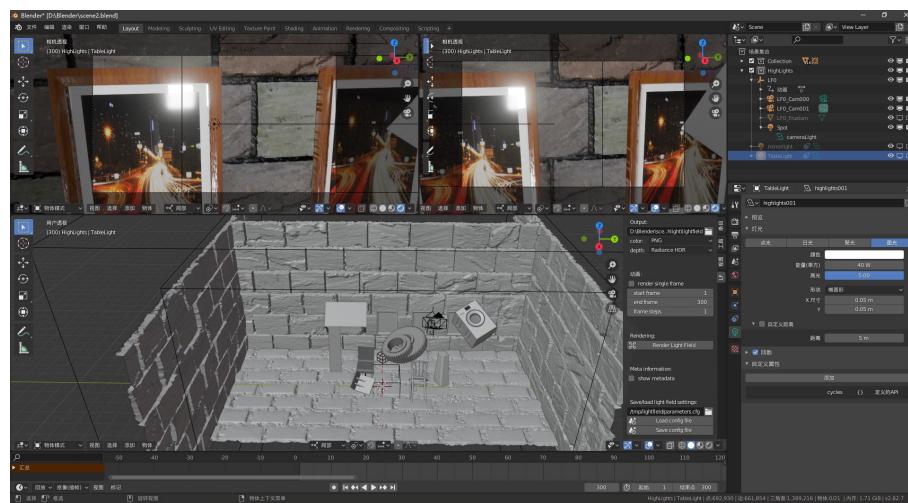


图 3-3 Blender 工作区界面

过曝场景数据集至少需要包含双目图像以及视差标注。为了产生双目图像，需要在 Blender 中设置两台相机，为了方便后续移动可将两台相机绑定在同一物体上；为了产生过曝图像，需要在 Blender 中添加额外的强光源，同时将强光源同相机绑定，保证每例样本中都包含过曝现象。进行相关设置后本课题使用的 Blender 工作界面如图3-3所示。右边侧栏为属性栏，可对项目中相关物体的属性进行设置，如相机镜头焦距，底片大小，环境光效等。左边为编辑和预览区，上半部分分别为左右相机的实时预览画面。预览画面中白色光斑在左右视图中处于不同位置，这是由于左右相机的观察视角不同导致。这种不一致是符合真实世界物理规律的，这说明采用上述设置产生的数据可以较好模拟真实世界过曝场景。值得注意的是 Blender 只能输出深度信息，为获取视差标注需要对数据做进一步处理。

考虑图3-4所示双目相机系统，左右目相机成像面分别为  $I_l$  和  $I_r$ ，水平对齐且位于同一平面，左右目相机的光心分别为  $O_l$  和  $O_r$  且光轴平行，光心到相机成像平面的距离为相机焦距  $F_l$  和  $F_r$ ，设世界坐标系下某点的  $P$  的齐次坐标为  ${}^W P = [{}^W X, {}^W Y, {}^W Z, 1]^T$ ，点  $P$  在左右相机的成像点的齐次坐标分别为  $p_l = [u_l, v_l, 1]^T$

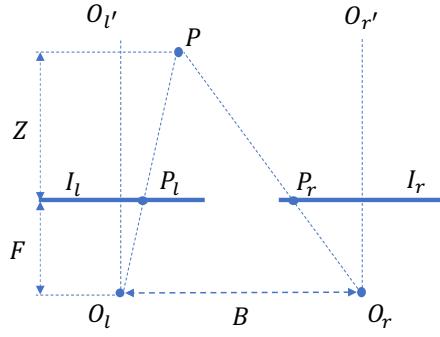


图 3-4 双目相机简化模型

和  $p_r = [u_r, v_r, 1]^T$ 。根据式 (2-8)，对于左目相机成像系统有：

$${}^C Z_l \begin{bmatrix} u_l \\ v_l \\ 1 \end{bmatrix} = \begin{bmatrix} f_{ul} & 0 & u_{0l} & 0 \\ 0 & f_{vl} & v_{0l} & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} R_l & t_l \\ 0^T & 1 \end{bmatrix} \begin{bmatrix} {}^W X \\ {}^W Y \\ {}^W Z \\ 1 \end{bmatrix} \quad (3-1)$$

以左目相机光心建立世界坐标系原点，则  $R_l$  退化成单位矩阵， $t_l$  为零矩阵，故有：

$$Z_l \begin{bmatrix} u_l \\ v_l \\ 1 \end{bmatrix} = \begin{bmatrix} f_{ul} & 0 & u_{0l} & 0 \\ 0 & f_{vl} & v_{0l} & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} X_l \\ Y_l \\ Z_l \\ 1 \end{bmatrix} \quad (3-2)$$

不难得到：

$$Z_l \times u_l = f_{ul} \times X_l + u_{0l} \times Z_l \quad (3-3)$$

同样对于右目相机有：

$$Z_r \times u_r = f_{ur} \times X_r + u_{0r} \times Z_r \quad (3-4)$$

由于左右目相机成像平面水平对齐且位于同一平面，根据坐标系间的相对位置关系有：

$$X_l = X_r + \text{baseline} \quad (3-5)$$

$$Z_l = Z_r = Z \quad (3-6)$$

将式 (3-5) 和 (3-6) 代入式 (3-3) 和 (3-4) 中整理可得:

$$u_l - u_r = (f_{ul} - f_{ur}) \times X_r + (u_{0r} - u_{0l}) \times Z + \frac{f_{ul} \times baseline}{Z} \quad (3-7)$$

根据定义  $u_l - u_r$  即为视差。对于 Blender 中的双目系统，可通过设置左右目相机使得  $f_{ul} = f_{ur}$ ,  $u_{0r} = u_{0l}$ , 对于真实世界中的双目系统可通过采用相同型号的相机实现。式 (3-7) 可简化为:

$$disp = \frac{f \times baseline}{Z} \quad (3-8)$$

通过上式可以将 Blender 输出深度值转化为视差值。列车部件非过曝区域的数据标注采用结构光法实现，利用两个周期不同的周期函数对投影光强进行唯一编码，通过左右目相机图像中的唯一编码确定对应关系，具体方法可参照文献 [7]。

### 3.3 数据采集场景

本课题在 Blender 虚拟场景中渲染了两类场景共计 900 例样本，每个样本包含两张过曝图像和关闭强光源后的未过曝图像共 4 张，左右目图像对应的视差和深度标注共 4 张。为了使渲染数据更加贴近列车高反光部件的过曝图像，对物体的材质设置较高的金属光泽和反射系数。同时采集了列车场景带有视差标注数据约 400 例，无标注数据约 4000 例。

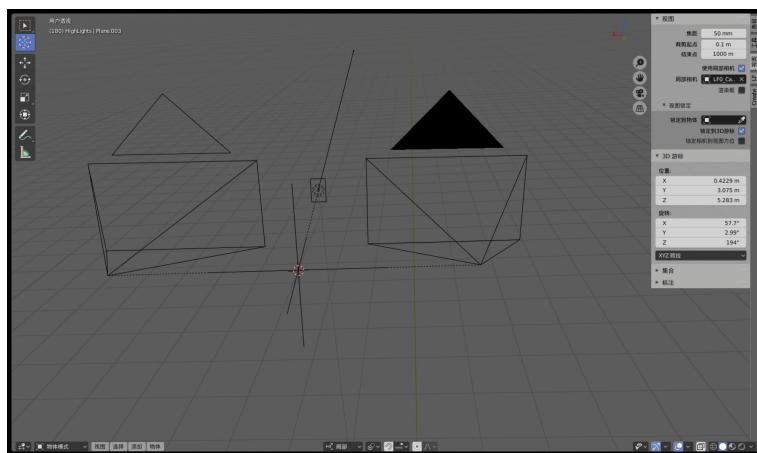


图 3-5 Blender 中的双目系统

根据上一节讨论，可将双目系统中相机参数设为相等以简化后续视差处理，本课题遵循文献 [11] 中相机设置，即相机焦距为 35mm，分辨率为  $1920 \times 1080$ ，baseline 为 4cm。搭建在 Blender 中的双目系统如图3-5所示，两部相机模型的位置同中间的十字架绑定，仅在水平位置

存在位移，大小为 baseline。同时与双目系统绑定的还有长方形光源，在图中以中间带有虚线圆圈的长方形表示，圆圈中引出的射线表示灯光方向。经过实验灯光亮度设置为 40W，高光系数设置为 10 可产生较好的过曝现象。该灯光布局同列车部件采集设备布局相似，有助于渲染更贴近实际场景的视差数据。



图 3-6 渲染场景建模

渲染场景由背景和悬浮物体组成，如图3-6所示。背景由多块带有凸起的墙面组成，目的是防止数据中出现无限远的物体。墙面的表面材质，反光系数，凸起均由真实世界物体采样而成，数据来自于公开素材库。悬浮物体中的相册和工件由建模工具生成，其中相册表面添加了一层玻璃材质，工件表面模拟了物理世界中的光滑金属面。其余模型则来自开源模型库，并在其表面设置了不同的反射系数用于模拟物理世界中常见的光滑表面。在 Blender 中，物体的连续移动通过插入关

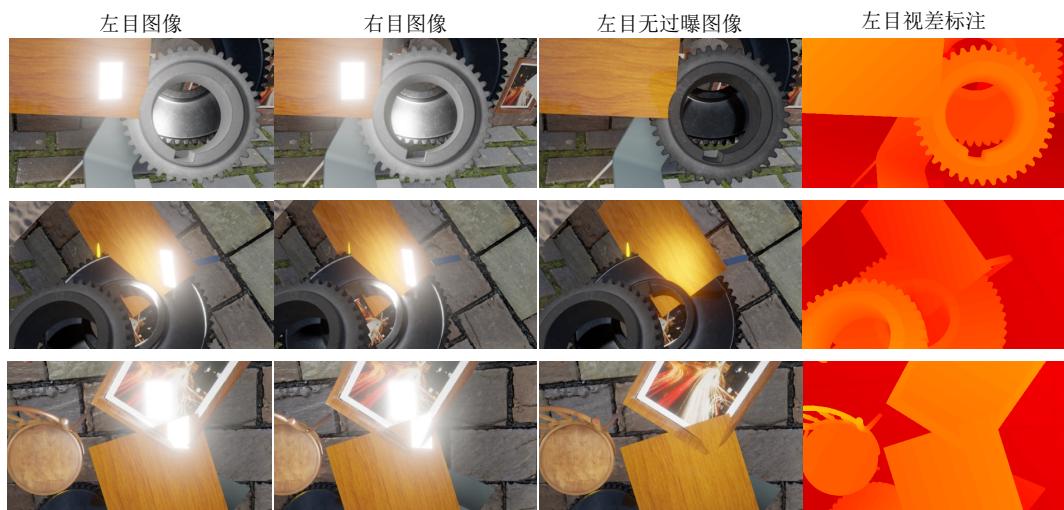


图 3-7 过曝场景数据样本

键帧，并在关键帧间对物体的几何坐标以及尺度因子进行调整实现。基于关键帧变换，本课题实现了两种随机渲染模式。模式一同文献 [11] 保持一致，仅有双目系统在场景中随机移动，强光源与相机的相对位置保持不变。为了提高数据的多

样性和随机性，在模式二中添加了物体和光源的随机移动用于模拟行车环境、工厂质检等复杂光源环境。模式二部分数据如图3-7所示，前两列为过曝场景图像用作网络输入，最后一列为伪彩色的视差图用作 GroundTruth 训练网络和测试性能，倒数第二列为关闭强光源后渲染左目图像，可用作图像修复的 GroundTruth 或者其他目的，右目无过曝图像和视差标注图以及左右目图像的深度图均未列出。

### 3.4 数据处理和视差验证

图像的每个像素都存在唯一深度值与其对应，通过深度值转化的视差值在遮挡和边缘区域无法找到对应像素，属于无效标注，在上述数据用于训练网络前需要剔除无效视差。根据文献 [11]，可通过左右视差的一致性检验清洗数据。具体来说，令  $D_l(u_l, v_l)$  表示在左目图像对应视差图  $D_l$  中坐标为  $(u_l, v_l)$  的视差值，则在右图中与该点对应的坐标为  $(u_l - D_l(u_l, v_l), v_l)$ ，即有：

$$D_l(u_l, v_l) = D_r(u_l - D_l(u_l, v_l), v_l) \quad (3-9)$$

式 (3-9) 说明对于左图中任意一点  $d(u_l, v_l)$ ，可以比较该点视差值和右图中对应位置  $(u_l - D_l(u_l, v_l), v_l)$  的视差值的差是否超过一定阈值，判断该点是否存在对应点，即该点是否处于无法匹配区域(左图左边缘因为视角移动消失在右图的区域、视角移动导致被遮挡的区域)。

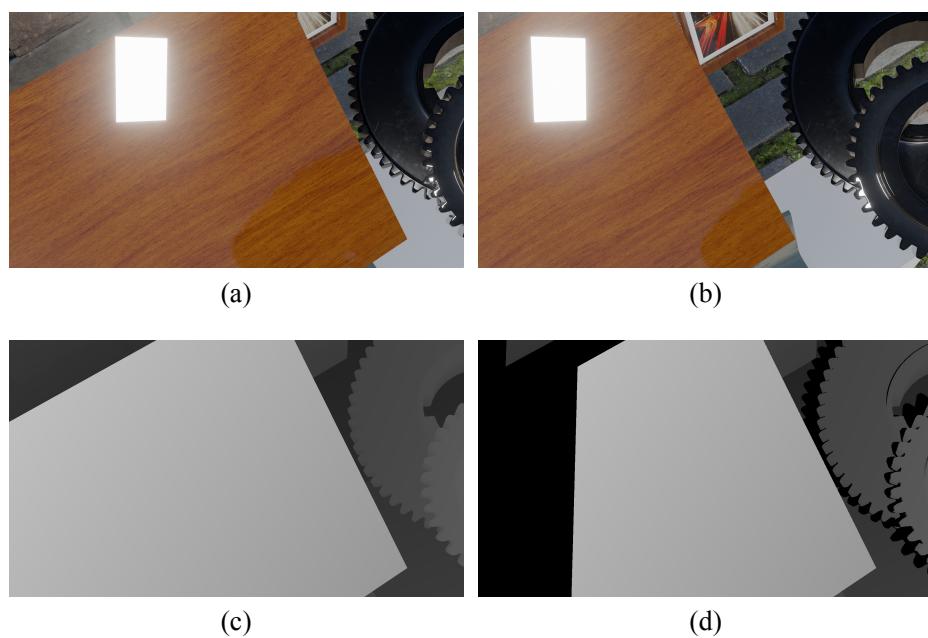


图 3-8 左右一致性检验。(a) 左目图像；(b) 右目图像；(c) 左目图像原始视差标记；(d) 左目图像有效视差标记

如图3-8所示，由于左右相机视角发生变化，左图中左半部分桌子的左边缘消失在右图中，同时左图右边缘的齿轮在右图中占据了更多像素。在原始输出视差中上述无法匹配区域均包含标注，这种标注的意义是当图像尺寸无限大和该物体不会被遮挡时对应像素的位置差，从双目立体匹配角度，这些区域的三维信息无法通过匹配方式感知。消除上述无效标注后左图视差图如3-8(d)所示，该标注更加符合视差物理意义。

完成视差有效性检查后，仍需要对视差的正确性做进一步检查。由式（3-9）不难得出，在不考虑物体反射的情况下应该有：

$$I_l(u_l, v_l) = I_r(u_l - D_l(u_l, v_l), v_l) \quad (3-10)$$

然而根据反射模型我们知道当视角发生改变时，物体反射的强度是不一样的。即使左右相机对同样光线产生同样响应，由于视角不同，同一物体在左右相机中产生的像素值也可能存在差异，即上式（3-10）一般不成立。虽然物体RGB值在左右相机记录中可能不一致，但物体位置利用式（3-10）重建后不会发生改变，因此可以通过比对边缘、纹理等位置验证视差正确性。

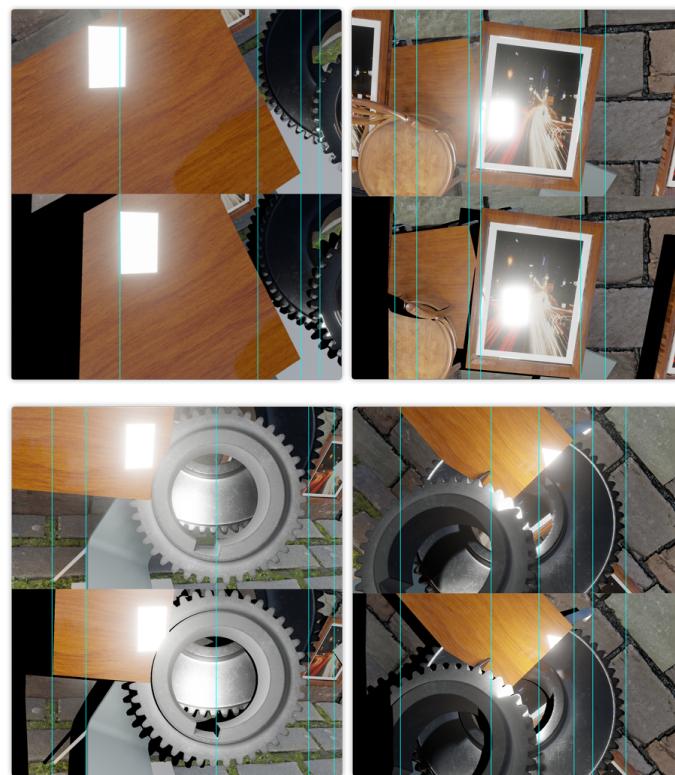


图 3-9 重建对比图

图3-9中共有四例重建对比样本，其中每例上部分为原始左图，下部分为利用左目视差和右图重建的左图。沿着图中垂直标线不难发现相框的边角、齿轮的边

缘在两幅图像中均已对齐，证明通过深度转化的视差是正确的。同时在图中也发现过曝产生的光斑无法对齐，这是因为左右相机存在视角变化，反射强光源产生的过曝光斑本身在左右图像中相对桌面的位置会发生改变。这进一步说明：一是通过 Blender 合成的数据是符合物理规律的，可以有效的模拟过曝场景；二是过曝场景下双目立体匹配的困难性，双目立体匹配是基于特征的匹配，过曝光斑形成强特征对匹配过程产生严重干扰。

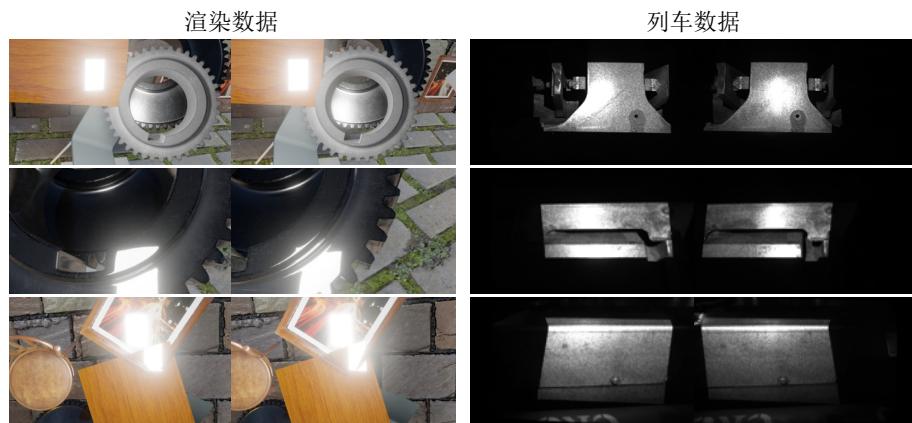


图 3-10 数据对比

图3-10为渲染数据和列车部件采集数据对比图，在进行数据渲染时，对物体材料设置了较高的金属性和反射系数，过曝光源采用了与列车场景采集设备相似布局，渲染数据较好的模拟了列车数据弱纹理，过曝显著的特点。渲染数据的光斑边界更明显，产生错误匹配的概率更高，适合用于验证视差精度以及作为预训练数据为模型提供先验知识。

### 3.5 本章小结

本章针对公开数据集缺乏过曝场景的问题，比较了各种数据采集方式的优缺点，选择了 Blender 渲染数据的方式构建过曝场景数据集。接着介绍了使用 Blender 渲染数据的主要工作流程，并详细推导了如何将 Blender 输出深度值转化为视差标注。同时指明了 Blender 中双目系统，过曝光源等相关参数均以模拟列车过曝场景为目的进行设置，并进一步说明如何在 Blender 中渲染过曝场景。最后利用视差对图像进行重建，验证了过曝现象的存在以及视差标注的正确性，表明该数据集可用于后续过曝场景下双目立体匹配的相关研究。

## 第四章 基于特征一致性的双目匹配特征提取模型

上一章详细描述了如何构建过曝场景双目视差数据集，其中在验证视差正确性时，原始左图中的光斑位置与重建左图中的光斑位置是不同的，两幅“左图”中丢失的图像信息也不完全一致。这表明左图和右图记录的场景信息存在冗余，对两者进行综合可以恢复出过曝导致丢失的部分信息。

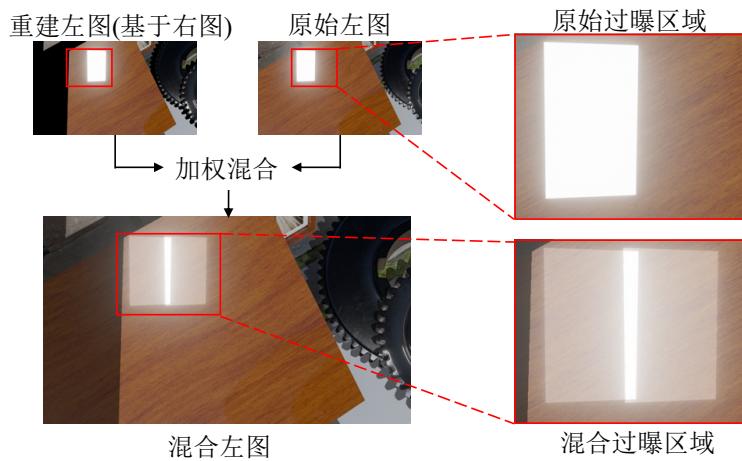


图 4-1 左右目信息混合对比图

将过曝左目图像和利用右图重建出的左目图像以平均加权的方式混合后，如图4-1所示。对比原始过曝区域和混合过曝区域可以发现，融合了右图信息的过曝区域基本可以恢复物体表面信息。图像修复常用来修复图像丢失的信息，其基本任务是通过图像丢失区域的邻域信息和图像全局信息重绘丢失区域。注意到双目图像包含过曝区域的冗余信息，如果在计算特征时引入冗余信息，利用图像修复相关技术修复过曝区域丢失的颜色、纹理和几何信息，可以恢复曝区域原本的语言特征，进一步提高过曝场景下视差估计的精度。本章从图像修复出发，研究如何利用图像修复的相关技术解决过曝区域特征丢失的问题。

### 4.1 多特征模型

为了修复过曝区域丢失特征，需要在特征计算过程中引入冗余信息，我们需要一种可以方便的将双目图像特征计算过程交织在一起的网络结构代替原本的特征提取结构。注意到图像修复属于生成任务的一种，其特点是将输入高分辨率图像转换到另一个特征空间高分辨率输出，根据任务不同保持某些属性不变。例如，对于修复任务，需要保持受损区域以外信息不变；对于图像风格转换，需要保持

转换前后物体形状不变。卷积层池化层在提取特征时会导致上述纹理、形状等低维特征消失，基于高维特征生成图像存在失真现象。因此研究者多采用带有跳跃连接的编解码网络解决上述问题。在双目立体匹配的网络设计中，特征提取模块目的是针对输入的双目图像选择合适特征表征，因此可以考虑直接将编码网络视为双目立体匹配特征提取模块。与单目图像不同，双目图像修复需要满足极线约束等成像物理规律，保证物体三维信息不发生改变。因此为解决过曝区域特征丢失的问题，首先需要考虑如何设计一种网络，一是能在输入输出之间传递低维信息，保证修复前后未过曝区域纹理不发生改变；二是在左目右目图像之间传递冗余信息，并保证修复后两幅图像构成的三维信息不发生变化。

#### 4.1.1 基于跳跃连接的 U型网络

在卷积神经网络发展初期研究者们设计的网络多针对分类问题，对于一例测试样本只需要输出一个分类标签。在许多视觉任务中，尤其是医疗影像任务中，不仅需要模型输出类别信息还需要输出位置信息，这意味着对输入的每个像素网络都需要预测一个分类标签。Ciresan<sup>[55]</sup> 等人将图片分为多个区域并利用滑动窗口的方法逐个输出位置信息初步解决了该问题。该方法有两个明显缺点，一是对于单例样本需要分成多个图像块计算，图像块间存在重叠导致大量重复计算；二是图像块大小和计算精度间难以达到平衡，需要针对问题具体设计。为了解决卷积网络难以传递低维特征的问题，Ronneberger<sup>[56]</sup> 等人提出了基于跳跃连接的 UNet。

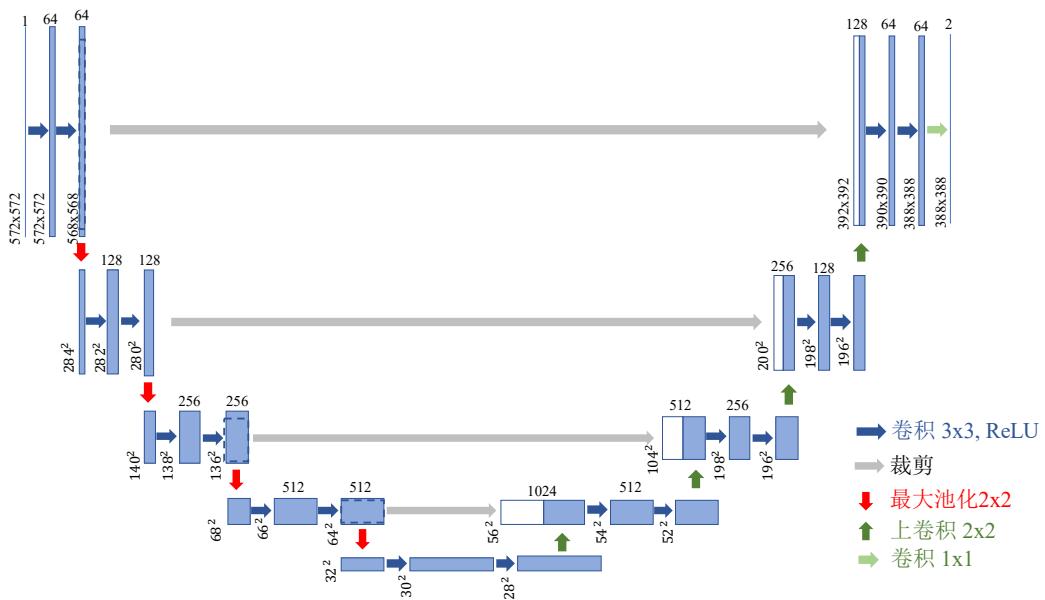


图 4-2 UNet 结构

UNet 的网络结构如图4-2所示，因其采用了对称的编解码结构整体类似一个 U

形，所以作者将其命名为 UNet。UNet 左侧部分称为收缩模型，用于捕捉上下文信息；右侧部分为对称拓展模型，用于综合多个特征。左侧和右侧对应特征层之间通过复制和级联方式形成新的特征。左侧收缩模型采用典型的卷积网络结构，由多个卷积块组成，每个卷积块包含两层  $3 \times 3$  卷积层以及 ReLU 激活函数。每个卷积块输出通过  $2 \times 2$  的最大池化层进行降采样，并倍增通道数后输入到下一层卷积块。对于右侧拓展模型中每一步，首先对输入特征上采样一倍并通过  $2 \times 2$  卷积将通道数减半，将其和收缩模型中对应层特征在通道维度上级联，形成新特征。接着将新特征通过两个带有 ReLU 激活函数的  $3 \times 3$  卷积层形成输出。最后通过  $1 \times 1$  卷积层将每个像素的 64 维特征映射为所需的分类向量。为了保证结果的准确性，在连接特征时仅保留没有 padding 像素参与计算的特征，所以分类结果的尺寸略小于输入图像。

该模型的关键点在于将收缩模型中高分辨率特征同拓展模型中上采样特征结合在一起，通过后续卷积层实现更精确输出。因为在上采样过程中引入了收缩模型中低维特征，保证了网络可以将纹理等信息传递到网络高层特征。图像 RGB 信息具有空间相似性，其分类结果也具有空间相似性，由底层传递到高层的语义信息有助于网络理解局部像素间、局部像素和整体图像间的关系，提高结果的平滑度和准确性。

设计 UNet 的目的是为了解决医疗影像任务中全像素分类问题，对称网络结构和对应层特征连接的方式使该网络实现了低维信息传递。生成任务中经常需要生成图像和输入图像在某些低维特征上保持一致，因此许多生成模型的设计都沿用 UNet 的思路。如伯克利人工智能实验室<sup>[57]</sup> 在利用生成对抗网络解决图像转换任务时，生成器采用了类 UNet 的网络结构，区别在于在 U 型的底部添加了多层不改变尺度的中间层。英伟达公司<sup>[58]</sup> 在解决不规则图像修复问题时也采用了相似的 UNet 型网络结构，区别在于使用其提出的“部分卷积层”替代了网络中的普通卷积层。

考虑到在进行过曝场景下的图像修复时需要在单目浅层特征和深层特征之间、左目特征和右目特征之间相互传递信息，而带有跳跃连接的 U 型网络结构可以很好的在多个特征间传递信息，对称的编解码结构可以方便的拓展至多特征网络并保证语义连贯性，因此本课题基于该思想设计了过曝场景下多特征提取模块。

#### 4.1.2 由单目到双目的多特征网络

目前，双目立体匹配神经网络中的特征提取阶段多采用卷积加空间金字塔池化的模型设计，并以权重共享的方式实现左右目图像特征提取。采用空间金字塔

模型的原因是仅从单个像素强度难以确定特征上下文关系，需要丰富纹理信息帮助网络估计像素对应关系<sup>[42]</sup>。采用权重共享的方式是为了保证左右目特征提取网络在训练过程中保持特征语义一致。

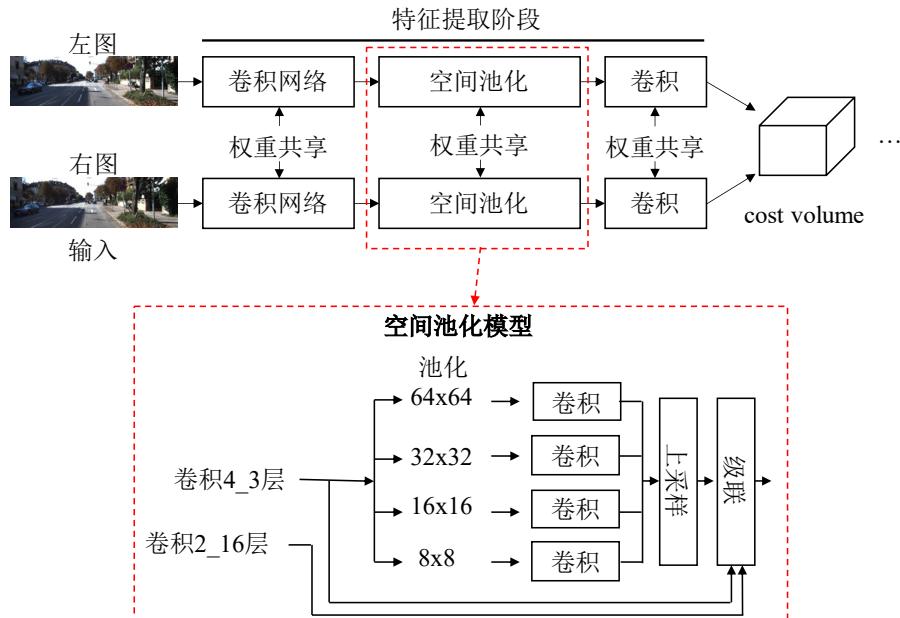


图 4-3 特征提取及空间金字塔池化

特征提取的一般结构如图4-3所示，其输出特征用于进一步构建 cost volume。对于单张输入图像，首先经过一个包含 4 层卷积层的小型卷积网络，卷积核大小为  $3 \times 3$ ，并分别在第二层和第四层设置卷积步长为 2 使图像尺寸降低两倍，其余层步长均为 1。接着将第四层卷积输出特征分别通过  $64 \times 64$ ,  $32 \times 32$ ,  $16 \times 16$ ,  $8 \times 8$  的平均池化层获取不同尺度特征，将这些特征均上采样至第二层卷积层输出尺寸，并同第二层卷积特征在通道维度上级联，形成包含丰富纹理信息的分层特征，最后通过卷积对多层特征聚合形成输出。对于输入左右目图像，采用具有相同系数的卷积层和池化层进行上述处理，保证在左右目输入图像中提取的特征可匹配。

空间金字塔池化的核心思想是在多个尺度上提取场景特征，但使用经过多层卷积输出的单个特征构建的分层特征存在以下问题：1) 在卷积过程中许多边、角和纹理等低层特征已经丢失；2) 忽略了不同层特征间递进关系，如车窗和轮胎一般而言会出现汽车所在区域而非行人像素附近。UNet 在编解码的过程中形成了多尺度特征，因此可以考虑基于 UNet 编解码的不同尺度特征构建包含递进关系的分层特征。在左右目输入图像间采用权重共享方式固然保证了提取特征的语义一致性，但在过曝场景中若对过曝光斑产生相同特征则会获得错误匹配关系，UNet 编

解码的对称结构设计允许我们方便的混合多个特征，因此可以考虑混合解码阶段特征实现过曝区域丢失特征修复。

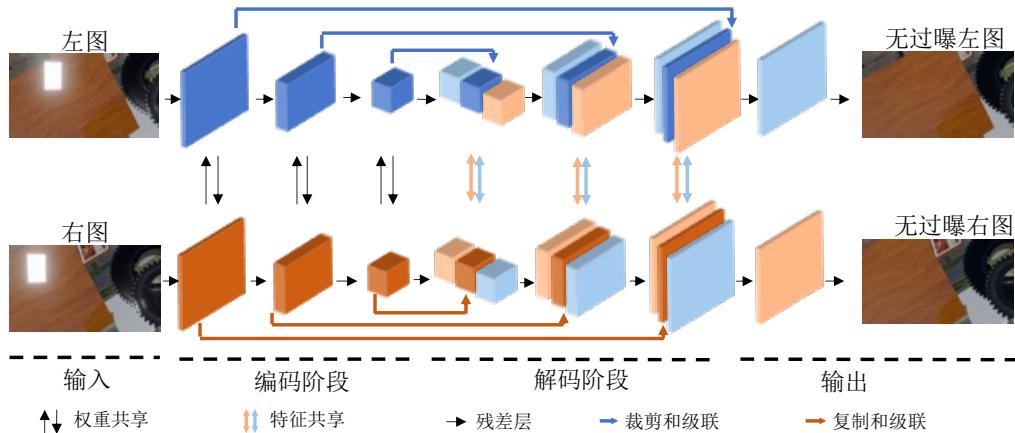


图 4-4 过曝场景下的特征提取模块

本课题基于 UNet 设计的过曝场景下双目特征提取模块如图4-4所示，整个特征提取模型可分为三个阶段：特征编码阶段，特征解码阶段以及图像解码阶段。编码阶段和解码阶段采用类似 UNet 中的对称设计，并在 U 型的底部添加了多个重复且不改变特征尺寸的卷积层，目的是进一步扩大模型的感知野。为了实现不同尺度间特征信息的递进传递，使用残差卷积块替代了 UNet 中单层卷积层。残差卷积块是 Kaiming<sup>[59]</sup> 等人提出的带有快捷连接的模型，通过添加一条从输入到输出的路径，解决已有模型难以模拟恒等变化的问题，其常见结构如下图4-5所示。在解码阶段，每层的输入不仅有来自上一层解码特征和其对应层编码层特征，而且额外引入了来自对应图像的解码特征，目的是引入对应图像冗余信息(对应图像是指左目图像对应的右目图像，或右目图像对应的左目图像)。为了帮助模型学习融合冗余特征，模块中添加了图像解码器，输出修复后无过曝的图像。

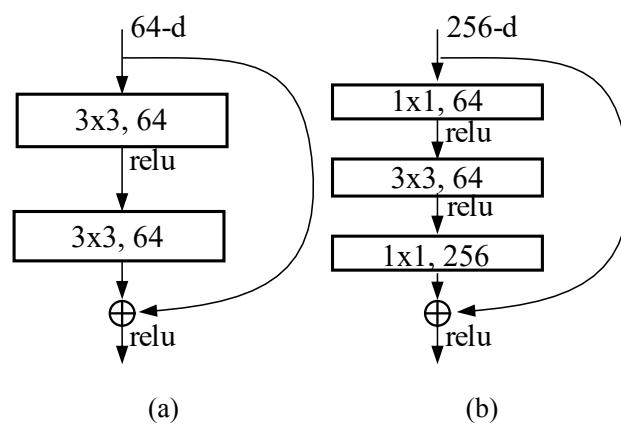


图 4-5 残差卷积网络结构。(a) 基本结构；(b) 带有通道变换的结构

该结构与现有特征提取模块最大的区别在于：1) 整个模块中只有编码阶段采用了权重共享方式，而在解码阶段采用了特征共享方式；2) 采用了编解码的对称设计代替空间金字塔池化。编码阶段权重共享是指对左右目输入图像采用相同的网络处理，共享梯度并绑定更新值，目的是保持左右目图像特征的语义一致性。解码阶段特征共享是指左(右)目图像特征在解码过程中共享来自右(左)目图像上一层解码特征，目的是通过引入右(左)目冗余信息修复过曝导致丢失的语义特征，减小过曝光斑的特征匹配度。具体来说，若编解码模型共有  $n$  层，则左目图像第  $i$  层解码层的输入包含来自左目图像  $n-i$  层编码特征， $i-1$  层解码特征以及来自右目图像  $i-1$  层解码特征。相比于直接从单层特征中池化分层特征，编解码对称设计允许特征自底向上，自顶向下的充分流动，保证了特征的空间递进关系。从左到右，从右到左，自底向上，自顶向下的多特征流动共同构成了该模块特征的基本流动方式。

## 4.2 多特征融合

上一节对本课题设计的特征提取模块的结构做了详细说明，其中核心是通过在解码阶段共享特征，将冗余信息引入到双目特征的计算路径中。卷积神经网络特点是逐层映射特征，但无法改变特征的空间位置。左右目图像中虽然包含冗余信息，但特征空间位置不同，在上节中，解码阶段对来自不同深度和路径的多个特征在通道维度上级联难以取得好的效果，针对双目图像的空间特点需要一种新的多特征融合方式。

### 4.2.1 Attention 机制

基于通道维度级联的特征融合方式面临的问题是，网络感知野大小受限于网络深度和卷积核大小，仅靠卷积层难以获取中远距离对应关系。注意力机制最早出现在自然语言处理中<sup>[60]</sup>，用于获取文本各个分词间关系。该机制利用矩阵乘法可以较好的获取中长距离特征关系，弥补卷积网络受限于感知野的不足，在图像处理中得到广泛应用。

注意力机制的一般结构如图4-6所示，对于输入特征  $F_1 \in \mathbb{R}^{B \times C \times H \times W}$  和  $F_2 \in \mathbb{R}^{B \times C \times H \times W}$ ，首先通过  $1 \times 1$  卷积核  $W_\theta^1$ ,  $W_\psi^2$  和  $W_\gamma^3$  将其变换到某高维空间内，即：

$$\theta_1 = W_\theta^1(F_1), \quad \psi_2 = W_\psi^2(F_2), \quad \gamma_3 = W_\gamma^3(F_3) \quad (4-1)$$

然后对重新排列后的  $\theta_1 \in \mathbb{R}^{(BH) \times W \times C}$  和  $\psi_2 \in \mathbb{R}^{(BH) \times C \times W}$  进行块矩阵乘法，并通过

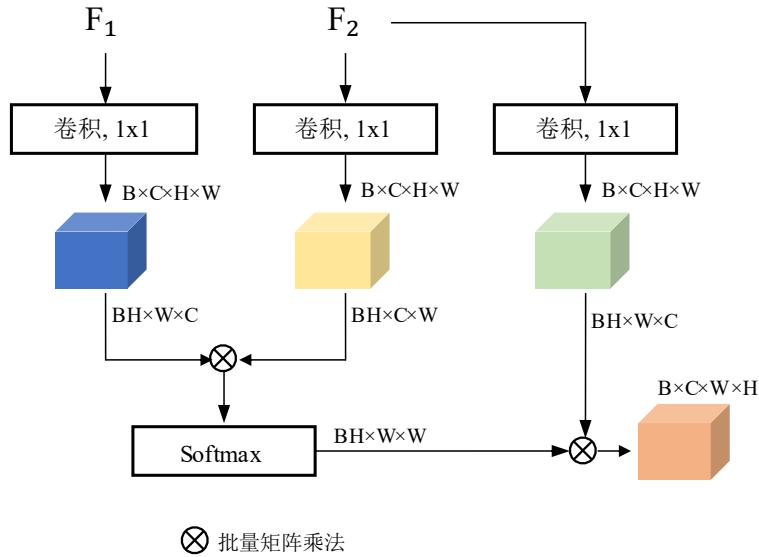


图 4-6 注意力机制

Softmax 归一化获得注意力图  $M_{2 \rightarrow 1} \in \mathbb{R}^{(BH) \times W \times W}$ ,  $M_{2 \rightarrow 1}$  代表了两个输入特征间的协方差矩阵。为了获得全局相关性, 将  $M_{2 \rightarrow 1}$  同  $\gamma_2$  作块矩阵乘法后同输入特征  $F_1$  在通道维度上级联, 最后通过  $1 \times 1$  卷积核  $W_O$  聚合特征并降低通道数, 获得最终输出特征  $O_1$  如下所示。

$$O_1 = W_O(cat(F_1, M_{2 \rightarrow 1} \times \gamma_2)) \quad (4-2)$$

#### 4.2.2 基于 Parallax-Attention 的多特征融合

受注意力机制启发, 本课题设计了如下视差注意模块用于实现 4.1 节中多特征融合, 其结构如图4-7所示。

以左目图像为例在某个解码层中有三个输入, 分别是来自对应编码层编码特征  $F_{encode}^l \in \mathbb{R}^{B \times C \times H \times W}$ , 来自上一层解码层输出  $F_{decode}^l \in \mathbb{R}^{B \times C \times H \times W}$  以及来自上一层右目图像解码层输出  $F_{decode}^r \in \mathbb{R}^{B \times C \times H \times W}$ 。将  $F_{decode}^l$  和  $F_{decode}^r$  通过卷积核大小为  $1 \times 1$  权重共享的连续卷积层  $\varphi$  汇聚多个通道中特征信息, 尺度变换为  $B \times H \times W$ , 即有:

$$f_l = \varphi(F_{decode}^l), f_r = \varphi(F_{decode}^r) \quad (4-3)$$

令特征在该层最大视差为  $d$ , 则构建  $cost_l \in B \times d \times H \times W$ , 对于  $cost_l$  的第二个维度的第  $i$  个元素, 有:

$$cost_l[B, i, H, i : W] = f_l[B, H, i : W] \times f_r[B, H, 0 : W - i] \quad (4-4)$$

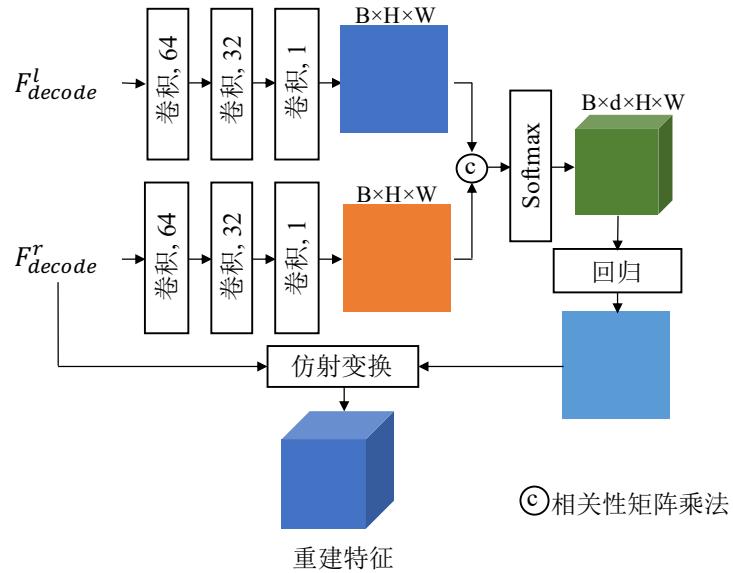


图 4-7 视差注意力机制

我们将该操作称为相关性矩阵乘法，其中  $[\dots, a : b, \dots]$  表示该维度上从位置  $a$  到位置  $b$  的所有元素。此时第二维度上聚集了该像素在所有潜在视差位置的匹配度，通过 Softmax 将其转换为概率分布函数，获得特征在视差分布上的注意力图  $M_{l \rightarrow r}$ ：

$$M_{l \rightarrow r} = \text{Softmax}(cost_l, dim = 1) \quad (4-5)$$

其中  $\text{Softmax}(cost_l, dim = 1)$  表示在  $cost_l$  第一维度上进行 Softmax 操作，Softmax 操作是指对于一组数  $[x_1, x_2, \dots, x_m]$ ，第  $i$  个数输出为：

$$\text{Softmax}([x_1, x_2, \dots, x_m], x_i) = \frac{e^{-x_i}}{\sum_{j=1,2,\dots,m} e^{-x_j}} \quad (4-6)$$

经过 Softmax 后，所有数之和为 1。注意到经过 Softmax 后可以得到每个像素视差注意力图，即其在可能视差上的概率分布函数  $\sigma(c_l)$ ，因此可以通过文献 [41] 中提出的 soft argmin 估计特征视差：

$$\hat{d} = \sum_{d=0}^{D_{\max}} d \times \sigma(c_l) \quad (4-7)$$

得到特征视差后结合右目特征  $F_{decode}^r$  重构左目特征，重构后的特征经过视差搬移后位于左目对应特征附近，因此可以直接通过在通道维度上级联特征实现融合。整体流程如算法4-1所示。

就设计思路而言，采用  $F_{decode}^l$  和  $F_{decode}^r$  计算视差注意力图的原因是两者在上

**算法 4-1** 基于 Parallax-Attention 的特征融合

**Data:** 编码特征  $F_{encode}^l \in \mathbb{R}^{B \times C \times H \times W}$ ,  $F_{encode}^r \in \mathbb{R}^{B \times C \times H \times W}$ , 解码特征  $F_{decode}^l \in \mathbb{R}^{B \times C \times H \times W}$ ,  $F_{decode}^r \in \mathbb{R}^{B \times C \times H \times W}$ , 最大视差  $d$ ,

**Result:** 融合特征  $F_{fuse}^l$ ,  $F_{fuse}^r$

- 1 将特征  $F_{decode}^l$ ,  $F_{decode}^r$  分别通过卷积核大小为  $1 \times 1$ , 输出通道为  $64$ ,  $32$ ,  $1$  的卷积层, 将通道数变换为  $1$ , 以  $f_l \in \mathbb{R}^{B \times H \times W}$  和  $f_r \in \mathbb{R}^{B \times H \times W}$  表示; 令  $i = 0$ , 并对  $cost_l \in \mathbb{R}^{B \times d \times H \times W}$  和  $cost_r \in \mathbb{R}^{B \times d \times H \times W}$  初始化;
- 2 **while**  $i < d$  **do**
- 3    $cost_l(:, i, :, i : W) = f_l(:, :, i : W) \odot f_r(:, :, 0 : w - i)$ ;
- 4    $cost_r(:, i, :, 0 : w - i) = f_l(:, :, i : W) \odot f_r(:, :, 0 : w - i)$ ;
- 5   *i*=*i*+1;
- 6 **end**
- 7 计算视差概率分布函数;
- 8    $prob_l = Softmax(cost_l, dim = 1)$ ;
- 9    $prob_r = Softmax(cost_r, dim = 1)$ ;
- 10 初始话  $init \in \mathbb{R}^{B \times d \times H \times W}$ , 值由在第二维度上重复的  $[1, 2, \dots, d]$  填充;
- 11 预测视差;
- 12    $\hat{d}_l = prob_l \otimes d_{init}$ ;
- 13    $\hat{d}_r = prob_r \otimes d_{init}$ ;
- 14 重构特征;
- 15    $f_{recon}^l = Warp(F_{decode}^r, \hat{d}_l)$ ;
- 16    $f_{recon}^r = Warp(F_{decode}^l, \hat{d}_r)$ ;
- 17 在通道维度上对特征进行融合;
- 18    $f_{fuse}^l = cat([F_{encode}^l F_{decode}^l f_{recon}^l], dim = 1)$ ;
- 19    $f_{fuse}^r = cat([F_{encode}^r F_{decode}^r f_{recon}^r], dim = 1)$ ;
- 20 输出融合特征  $f_{fuse}^l$ ,  $f_{fuse}^r$ ;

一层解码层中进行过一次特征融合, 两者特征相较于编码特征而言更为接近, 有助于网络获取冗余信息的位置关系。在计算相关性前通过卷积层将通道数变换到单通道, 一是为了汇聚各层特征, 二是为了降低计算量。设定最大视差  $d$  是因为考虑到左右目图像对应特征一般位于某个局部邻域内, 过曝光斑在两幅图像的位置差异不会跨越整张图像, 仅需考虑中短距离特征, 同时也可以进一步降低计算量。相较于注意力机制中直接计算整个向量行相关性  $W$  次,  $cost_l$  仅需要计算  $d$  次,  $d$  一般为  $(0.25 - 0.5) \times W$ 。

### 4.3 双目特征一致性修复

在上一节中, 本课题沿着注意力机制思路, 设计了基于视差注意力机制的小型特征视差预测网络, 实现了冗余特征的搬移和融合。来自右目冗余特征和左目过曝特征处于同一空间位置, 考虑到卷积层每层输出由输入各维度特征加权计算

得到，因此为了促进网络在左目过曝区域尽可能保留来自右目的冗余特征，修复右目丢失特征，本课题引入图像修复网络，即图4-4中的图像解码模块。

### 4.3.1 一致性预测

基于预测视差  $\hat{d}$  和右目特征  $F_{decode}^l$  可以得到重建左目特征  $F_{recon}^l$ ，在非过曝区域  $F_{recon}^l$  中特征与  $F_{decode}^l$  基本一致。为了将网络关注点集中在过曝区域，本课题设计了如下图4-8所示的双目特征一致性预测网络，考虑到仅仅通过差分也可到达相同目的，因此采用了简单的4层设计。

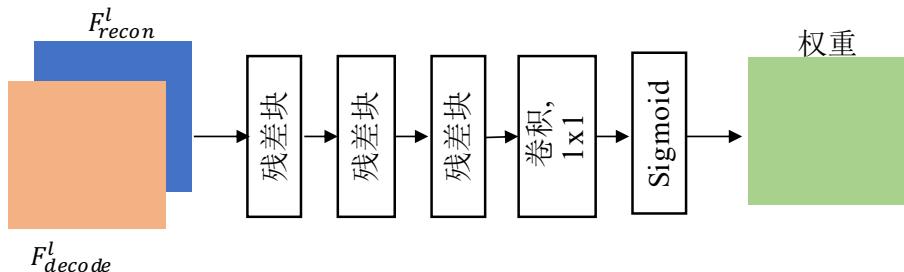


图 4-8 一致性预测网络

首先将输入  $F_{decode}^l$  和  $F_{recon}^l$  在通道维度上级联，接着送入三层残差卷积块，残差卷积块采用不改变尺度的基本构型。最后通过卷积核大小为  $1 \times 1$  的卷积层将通道数变为 1，最后通过 Sigmoid 函数将数值归一化到  $[0, 1]$  之间获得一致性预测权重  $w$ 。最后利用一致性权重对特征进行融合：

$$F_{fuse}^l = w \times F_{decode}^l + (1 - w) \times F_{recon}^l \quad (4-8)$$

考虑到图像解码阶段位于整个特征提取模块的尾部，特征空间与图像 RGB 空间较为接近，因此由融合特征  $F_{fuse}^l$  到重建图像  $I_{recon}^l$  的网络采用了类似一致性预测的网络结构，将输入替换为  $F_{fuse}^l$  并对通道数做出相应的调整。

### 4.3.2 损失函数

在构建图像修复的损失函数时，不仅要考虑过曝区域的修复精度问题，还需要考虑修复前后的语义一致性以及修复后过曝区域和其周围区域的连贯性。给定过曝场景下图像  $I_{oe}$ ，和过曝区域掩码图  $M$ 。过曝区域掩码图由标注数据中过曝图像  $I_{oe}$  和无过曝图像  $I_{gt}$  对比得到，其中 0 代表过曝区域。令最后网络输出的修复图像为  $I_{pred}$ ，可以构建像素级损失函数如下

$$\mathcal{L}_{oe} = \frac{1}{N_{I_{gt}}} \|(1 - M) \odot (I_{pred} - I_{gt})\|_1 \quad (4-9)$$

$$\mathcal{L}_{normal} = \frac{1}{N_{I_{gt}}} \|M \odot (I_{pred} - I_{gt})\|_1 \quad (4-10)$$

其中  $N_{I_{gt}}$  表示  $I_{gt}$  中的元素个数。该损失函数保证了生成修复图像中各个像素值的精度。为了保证图像修复前后语义特征一致，目前深度学习常采用的方法是将图像映射到某个预训练好的神经网络中，在其不同的特征层计算损失，因此我们定义感知损失如下：

$$\mathcal{L}_{perceptual} = \sum_{p=0}^{P-1} \frac{\|\Psi_p^{I_{pred}} - \Psi_p^{I_{gt}}\|}{N_{\Psi_p^{I_{gt}}}} + \frac{\|\Psi_p^{I_{comp}} - \Psi_p^{I_{gt}}\|}{N_{\Psi_p^{I_{gt}}}} \quad (4-11)$$

其中  $I_{comp}$  由  $I_{pred}$  和  $I_{gt}$  组合得到：

$$I_{comp} = M \odot I_{gt} + (1 - M) \odot I_{pred} \quad (4-12)$$

即  $I_{comp}$  由  $I_{pred}$  非过曝区域直接替换为  $I_{gt}$  得到。 $\Psi_p^{I_*}$  表示输入  $I_*$  在网络中第  $p$  层中激活特征图。鉴于在 ImageNet 上预训练的 VGG16 模型在多个跨领域任务中都具有较好的鲁棒性，本课题选择其中 pool1, pool2 和 pool3 用于生成上述损失函数。为了保证修复后图像连贯性，损失函数中添加了一项梯度惩罚 (Total Variation, TV)<sup>[61]</sup> 如下：

$$\mathcal{L}_{tv} = \sum_{(i,j) \in R, (i,j+1) \in R} \frac{\|I_{comp}^{i,j+1} - I_{comp}^{i,j}\|_1}{N_{I_{comp}}} + \sum_{(i,j) \in R, (i+1,j) \in R} \frac{\|I_{comp}^{i+1,j} - I_{comp}^{i,j}\|_1}{N_{I_{comp}}} \quad (4-13)$$

最终图像修复的损失函数如下：

$$\mathcal{L}_{repair} = \mathcal{L}_{normal} + 6\mathcal{L}_{oe} + 0.05\mathcal{L}_{perceptual} + 0.1\mathcal{L}_{tv} \quad (4-14)$$

各个损失函数的权重由多次实验后确定。

#### 4.4 实验结果与分析

本节针对本章提出的特征提取模块进行实验，实验采用 PSMNet 作为基线模型进行对比实验，实验将 PSMNet 原本的特征提取模块替换本章提出的多特征模型，并在损失函数中添加上节图像修复损失函数，其余部分保持一致。实验在第三章中构建的双目过曝场景数据集下进行，本课题虽然涉及到图像修复，但目的是优化双目立体匹配中特征提取过程，因此实验中主要针对过曝场景下视差估计精度进行评估，实验采用的平台配置如下表4-1所示。

对视差估计精度评估之前首先明确视差评价指标，本课题采用双目视差估计中常见的评价指标 (1) 误点率；(2) 平均像素误差<sup>[11]</sup> 对模型的视差预测精度进行评

表 4-1 实验平台配置表

类型	参数
操作系统	Ubuntu 18.04.1 LTS
CPU	Intel(R) Core(TM) i7-7700 CPU @ 3.60GHz
GPU	NVIDIA GeForce GTX 1080 Ti
代码语言	Python 3.6
深度学习平台	Pytorch1.7

估。

给定预测视差  $d_{pred}$  和标注视差  $d_{gt}$ , 则两种评价指标如下:

### (1) 误点率

按照 KITTI Stereo 2012 的说明, 误点率是指预测视差同标注视差的误差超过某一阈值且误差百分比超过 5%:

$$ErrorPointRate = \frac{Num\left((|d_{pred} - d_{gt}| > threshold) \cap \left(\frac{|d_{pred} - d_{gt}|}{d_{gt}} > 5\%\right)\right)}{Num(d_{gt} > 0)} \quad (4-15)$$

其中  $Num(d_*)$  表示视差图  $d_*$  中满足条件的元素个数。实验中采用的阈值同文献 [11] 保持一致即 2px, 3px 和 5px。该评价指标主要针对单个像素视差估计的绝对误差和相对误差两个方面进行评估。

### (2) 平均像素误差

平均像素误差是指预测视差同标注视差的总像素误差平均到每个像素的误差大小:

$$AverPixelErr = \frac{\sum |d_{pred} - d_{gt}|}{Num(d_{gt} > 0)} \quad (4-16)$$

该指标反应了模型整体的视差预测精度。

在第三章中构建过曝场景数据集时采用了两种模式, 模式一中相机与光源保持相对位置不变, 视差范围较大, 称为场景 1(Scene1); 模式二在视角移动的同时光源与相机相对位置也在改变, 视差范围较小, 称为场景 2(Scene1)。每个场景各有 450 例样本, 随机将各个场景分 300 例训练样本和 150 例测试样本, 图像分辨率均为  $1920 \times 1080$ , 在训练网络时将图像缩放至  $512 \times 256$ , 目的是为了获取亚像素级的视差标注精度。本章提出的模型和 PSMNet 模型在实验时加载了预训练模型权重, PSMNet 加载的预训练权重来自其官方网页中公开的 Sceneflow Dataset 预训练权重, 本课题改进后的模型加载的权重来自按照 PSMNet 论文 [42] 中的预训练方式在 Sceneflow Dataset 重新训练后保存的权重。

表 4-2 Scene1 和 Scene2 客观指标对比

method	>2px		>3px		>5px		mean	
	all	oe	all	oe	all	oe	all	oe
PSM&S2	<b>1.94%</b>	3.09%	<b>1.75%</b>	2.83%	<b>1.34%</b>	2.15%	<b>0.520</b>	0.801
Ours&S2	3.60%	<b>2.88%</b>	3.05%	<b>2.62%</b>	2.16%	<b>1.98%</b>	0.728	<b>0.632</b>
PSM&S1	25.42%	31.51%	24.54%	31.27%	22.75%	30.48%	10.827	15.042
Ours&S1	20.62%	22.49%	19.82%	22.08%	18.24%	20.73%	6.926	8.743
PSM&S1*	16.63%	19.85%	15.78%	19.60%	14.79%	19.29%	5.559	9.676
Ours&S1*	<b>14.32%</b>	<b>13.66%</b>	<b>13.61%</b>	<b>13.52%</b>	<b>12.89%</b>	<b>13.08%</b>	<b>5.233</b>	<b>5.469</b>

表4-2为本课题提出模型(Ours 表示)和 PSMNet 的客观指标对比，在两个场景进行实验，首先在 Scene2 上对两个模型进行训练，并分别在 Scene2 和 Scene1 上进行测试，以后缀 S2 和 S1 表示。同时也在 Scene1 中进行了训练并测试，以 S1\* 后缀表示。>2px, >3px, >5px 分别表示阈值为 2 个像素，3 个像素和 5 个像素的误点率，mean 表示平均像素误差，all 表示在整张图像中统计的客观指标，oe 表示仅在过曝区域统计的客观指标，过曝区域由数据中的过曝掩码指定，每个场景中的最优指标在表中以加粗表示。

由表4-2可知，在 Scene2 场景上进行训练时，虽然 PSMNet 整体上性能略高于本课题方法，PSMNet 平均像素误差仅为 0.520 像素，本课题方法为 0.728 像素。但在过曝区域本课题预测精度远优于 PSMNet，且当场景发生变化时，即在 Scene1 上进行跨场景实验时，本课题方法要远超 PSMNet，平均像素误差为 6.926 像素，PSMNet 平均像素误差为 10.827 像素，Ours 与在 Scene1 上重新训练后的测试结果仅相差 32.35%，而 PSMNet 同重新训练后的结果相比差距达到了 94.77%。这说明本课题改进后模型的泛化性能高于基准模型 PSMNet。在 Scene1 场景训练的测试指标中，本课题改进的模型各个指标均优于 PSMNet，进一步说明基于图像修复的特征提取模块有助于提高双目立体匹配模型在过曝场景下的视差估计精度。

为探究上述多特征模型、多特征融合方法以及图像修复对过曝场景下视差预测性能的影响，针对本课题提出的多特征提取模块开展消融实验，实验结果如表4-3所示。表4-3中，PSM 表示基线模型测试结果；PSM w/u 表示在基线模型基础上添加多特征模型测试结果；PSM w/u+f 表示在 PSM w/u 基础上添加融合模块测试结果；PSM w/u+r 表示在 PSM w/u 基础上添加图像修复模块测试结果。PSM w/u+f+r 即为本章提出模型。由测试结果可知，上述各模块不同程度的提高了过曝区域的视差预测精度。PSM w/u 结果表明在解码阶段引入多特征可以帮助网络获得冗余信息；PSM w/u+f 结果表明提出的特征融合模块可以充分利用特征位置关系，将图像的冗余信息搬到丢失位置；PSM w/u+r 结果进一步说明了图像修复模

表 4-3 特征提取模块消融实验

method	>2px		>3px		>5px		mean	
	all	oe	all	oe	all	oe	all	oe
PSM	25.42%	31.51%	24.54%	31.27%	22.75%	30.48%	10.827	15.042
PSM w/u	31.67%	31.53%	30.78%	30.99%	29.43%	30.48%	14.888	14.691
SPM w/u+f	30.15%	27.60%	29.31%	27.31%	28.14%	26.93%	15.021	13.698
PSM w/u+r	27.40%	26.19%	26.59%	25.97%	25.20%	25.52%	11.766	12.423
PSM w/u+r+f (Ours)	<b>20.62%</b>	<b>22.49%</b>	<b>19.82%</b>	<b>22.08%</b>	<b>18.24%</b>	<b>20.73%</b>	<b>6.926</b>	<b>8.743</b>

块的目标函数可以促进网络利用冗余信息，修复特征。

图4-9(a)为本课题改进模型同 PSMNet 和消融模型在 Scene1 数据集下的测试结果对比图，前两行为输入过曝场景下双目图像左图和右图，第三行为标注视差，然后分别为 PSMNet 模型、本课题改进模型和消融模型的预测结果。为了方便观察视差变化，对视差图进行了伪彩色处理，颜色越接近红色代表视差越小，物体距离相机越远；颜色越接近蓝色代表视差越大，物体距离相机越近。具体颜色与视差映射关系可以参考图4-9(b)，下方标注为归一化后视差大小。由图4-9(a)中可知，PSMNet 在光斑附近产生了严重错误匹配，预测视差异常。本课题的各个模块和本章提出的模型可以在一定程度上减轻过曝带来左右目不一致的影响，输出较为正常视差，但在正常区域均有精度下降的现象，因此后文将研究如何提高整体视差预测精度。在两个模型预测结果的部分边缘处存在视差异常，原因是物体边缘处缺少匹配像素。

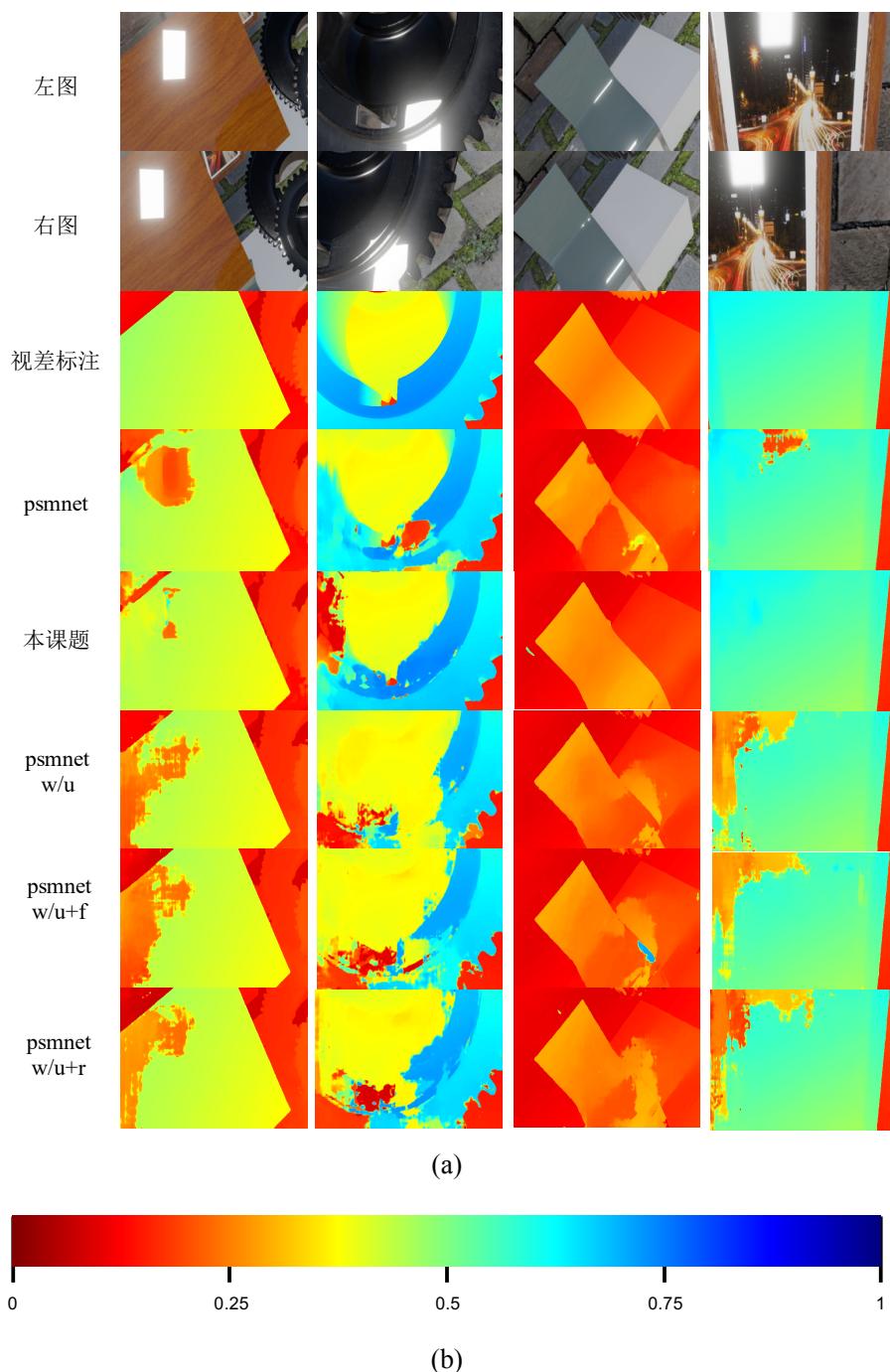


图 4-9 主观结果。(a) 中预测结果; (b) 伪彩色视差对应图

图4-10为本章提出的模型在 Scene2 数据集下的测试结果，前两列为过曝场景下双目图像左图和右图，第三列为图像修复模块输出的修复后图像，第四列为无过曝光光源渲染图像，作为图像修复 GroundTruth，最后一列为左图预测视差。由图可知，仅使用四层卷积的图像解码模块，仍能完成基本的图像修复工作。观察到第一行样本中光斑重叠区域未能修复，光斑的非重叠区域基本信息已经恢复，且



图 4-10 图像修复和视差预测结果

视差预测并未出现异常，这证明本课题设计的多特征融合模块能够使左右目冗余信息从右到左，从左到右的流动，修复丢失的语义特征，基于视差注意力机制的融合方式也能较好的对冗余信息进行融合。

## 4.5 本章小结

本章首先分析了在过曝场景下利用左右目图像中冗余信息恢复丢失信息的可能性，指出可以通过图像修复的方法引导网络融合冗余信息。接着分析了 UNet 网络特点以及现有特征提取网络采用权重共享带来的缺点，并详细介绍了本课题基于 UNet 和特征共享思路设计的多特征提取模块。然后针对左右目多特征融合问题，提出了基于视差注意力机制的特征融合方式，通过注意力机制获取特征视差概率分布函数，实现了基于特征重构的特征融合。接着详细介绍了后续图像修复的网络结构和损失函数。最后通过对比实验证明了该章中提出的方法可以有效提升网络在过曝区域的视差预测精度，并且可以在一定程度上提高网络泛化性能。

## 第五章 过曝场景下的双目立体匹配

在上一章中，主要研究了如何利用左右目图像冗余信息提高过曝区域视差估计精度。在本章中，主要针对提高过曝场景整体视差预测精度展开研究。首先对模型的整体进行详细介绍，分析结构上存在的问题并提出一种基于特征相关性的 cost volume 构建方案。接着从多任务角度构建联合损失函数进一步优化过曝场景下视差估计。最后对模型进行消融实验和对比实验，并测试模型在“高铁列车双目视差估计项目”应用场景下的表现。

### 5.1 基于一致性匹配特征的视差网络

第四章节中提出了基于多特征提取模块的双目视差估计模型，完整的网络结构如下图5-1所示。网络共分为三个阶段：特征提取阶段，代价聚合阶段和视差计算阶段。特征提取阶段即第四章中提出的多特征提取网络，本节着重介绍代价聚合阶段和视差计算阶段。

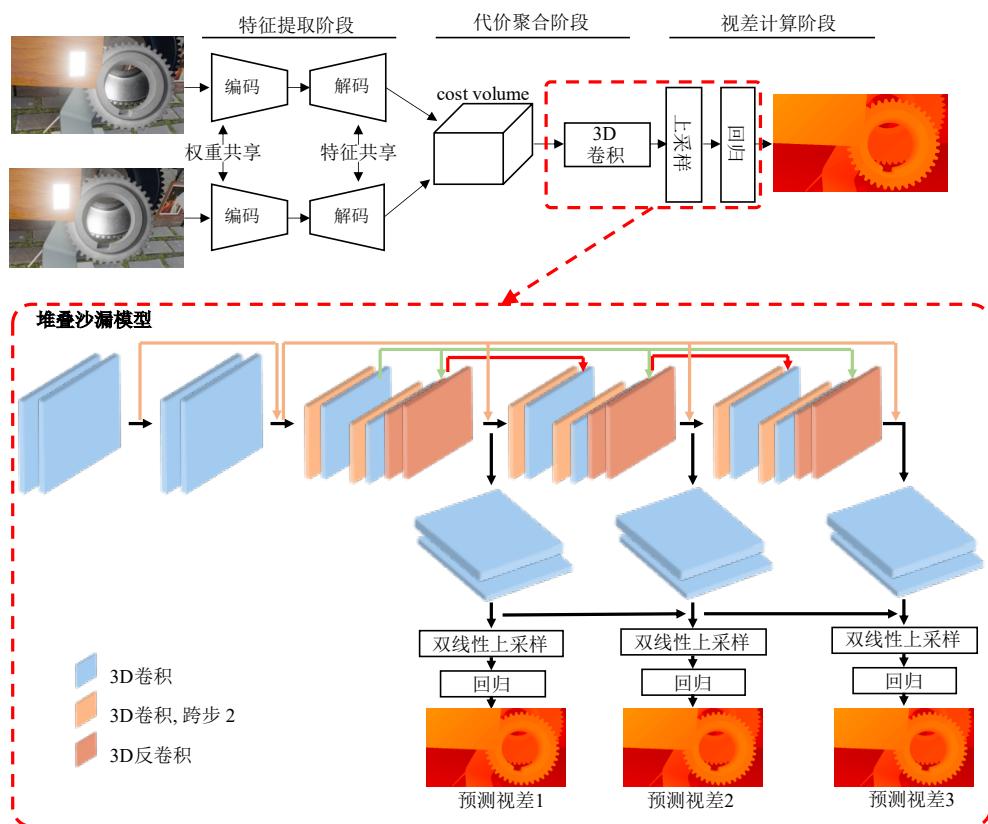


图 5-1 代价聚合和视差计算模块

代价聚合阶段首先需要构建 cost volume，cost volume 是以一定形式组织的五

维张量，用于描述双目图像特征（四维张量）在所有可能视差位置匹配关系的特征空间。PSMNet 选择将左右特征图在不同的视差层级上级联，并让网络学习对视差匹配度的评估。**cost volume** 具体构建方法如下，对于左右特征  $F_l \in \mathbb{R}^{B \times C \times H \times W}$ ,  $F_r \in \mathbb{R}^{B \times C \times H \times W}$ , 和最大视差  $d_{max}$ 。构建尺度为  $B \times 2C \times d_{max} \times H \times W$  的五维张量  $Volume_{cost}$ 。对于  $Volume_{cost}$  第三维的第  $i$  个元素，构建方法如下：

$$Volume_{cost}[:, 0 : C, i, :, i : W] = F_l[:, :, :, i : W] \quad (5-1)$$

$$Volume_{cost}[:, C : 2C, i, :, i : W] = F_r[:, :, :, 0 : W - i] \quad (5-2)$$

即  $Volume_{cost}$  在第二维度通道维度分为两部分，前  $C$  部分存放左目特征，后  $C$  部分存放右目特征，同时第三维度视差维度每增加一层，存放的左目特征在第四维度宽度维度上左移一个单位，右目特征右移一个单位，共移动  $d_{max}$  次。

构建 **cost volume** 后需要聚合其中的分层特征，并尽可能的学习上下文信息，因此采用了堆叠沙漏 (Stacked Hourglass) 形式的网络结构。该结构由重复的自上而下，自底而上的处理模块以及多层监督组成，具体如图5-1所示。堆叠沙漏模型最早由 Johnson<sup>[61]</sup> 等人为推理人体与人体关键部位空间关系而提出的网络结构，单个沙漏模型类似于前文提到的 UNet 结构，区别在于，一是 UNet 利用 2D 卷积处理图像特征，沙漏结构利用 3D 卷积处理包含额外视差维度的图像特征；二是沙漏结构编码特征在连接到解码特征时利用  $1 \times 1$  卷积层进行通道变换，连接的方式采取对应元素相加的形式，目的是允许特征在多级沙漏之间连接，促进网络学习更多的上下文信息。堆叠沙漏的多级结构允许在模型中间进行特征监督，添加中间监督的原因是考虑到在尺度较小的高维特征中物体特征更为明显，通过单独训练每个沙漏模型可以促进网络尽可能的理解场景对应关系。同时，后一级沙漏结构可以基于前一级沙漏结构的预测结果进行推理，形成从粗糙到细致的视差预测过程。

在传统双目立体匹配中，视差计算需要在最终的 **cost volume** 视差维度上进行 **argmin** 操作，即找到最小值对应的索引。在深度学习中这样操作存在两点问题：一是该操作不连续并且无法产生亚像素级精度；二是该操作不可微，反传过程复杂，难以训练。Kendall 等人<sup>[41]</sup> 提出了一种基于深度学习特征的视差计算方法 **soft argmin**。首先将进过聚合后的 **cost volume** 取反转换为预测损失  $c_d$ ，然后在视差维度进行 Softmax 对其进行归一化操作  $\sigma(\cdot)$ ，接着利用归一化后的概率分布函数对每个视差值加权求和：

$$soft\ argmin := \sum_{d=0}^{D_{max}} d \times \sigma(-c_d) \quad (5-3)$$

上式(5-3)是完全可微的，因此可用于网络训练和连续视差回归预测。

上述结构通过简单堆叠的方式构建 cost volume，在堆叠沙漏模型中聚合 cost volume 特征，寻找在视差维度上的对应关系，在视差计算阶段将聚合后特征转换为视差概率分布函数预测视差。PSMNet 基于上述结构在 KITTI 自动驾驶场景取得了较好成绩，本课题基于上述结构和双目修复特征提取模块在过曝场景上也表现出了远超基线模型的性能。

值得注意的是上述 cost volume 构建方式并没有对像素的对应关系进行表征，仅仅将左右目特征堆放在一起，需要模型在学习的过程选择合适的相似性度量方式，这在 PSMNet 上是合理的，因为其特征提取模块采用了权重共享形式，而左右目图像因为视角变化存在一定差异，因此对于同一物体，特征提取模块输出特征并不一致，无法直接计算匹配度，需要网络综合不同层特征和邻域特征综合考虑。但采用了第四章提出的基于特征共享的特征提取模块后，左右目特征在解码的过程中进行了多次融合，匹配程度有了一定提高，因此可以考虑利用特征相关性优化 cost volume 构建过程，进一步提高网络在过曝场景视差预测精度。

## 5.2 基于特征相关性的 cost-volume 构建方法

本节首先说明双目图像特征相关性，并分析左右目图像的相关性特点，接着研究如何利用特征相关性优化 cost volume 构建过程。

### 5.2.1 双目特征相关性

相关性是指两个变量的相关程度，而左右目图像可视为同一场景  $x$ ，在不同视角  $\theta_l$ 、 $\theta_r$  和仿射方程  $f_l(\cdot)$ 、 $f_r(\cdot)$  下的输出  $f_l(x, \theta_l)$ 、 $f_r(x, \theta_r)$ 。双目图像一般采用相同相机采集，因此有  $f_l = f_r$ ，双目图像经过校正后仅存在水平方向上较小位移，即  $\theta_l$  和  $\theta_r$  较为接近。 $f_l(x, \theta_l)$ 、 $f_r(x, \theta_r)$  具有相同的函数映射和相似的自变量，因此可以认为两者具有很强的相关性。我们可以通过图5-2对双目图像中的相关性建立直观的认识。

在图5-2中，我们首先关注处理框。存在物体“5”，在左右相机分别成像为左图和右图，左右相机水平位置的差异最终导致物体中心“5”在图像中的视差为1，具体表现为左图中“5”出现在第三个像素，右图中“5”出现在第二个像素，因为相机在成像过程中存在噪声和量化误差所以左右目图像的像素无法一一对应。

对图5-2执行如下操作，每次删去左图的左边  $d$  位和右图的右边  $d$  位并将对应位元素相乘后放在左图对应位置， $d \in [0, 3]$  得到结果框内输出，接着对每次操作的结果求和并对参与计算的元素个数求均值放在最左列。观察结果不难发现  $d = 1$

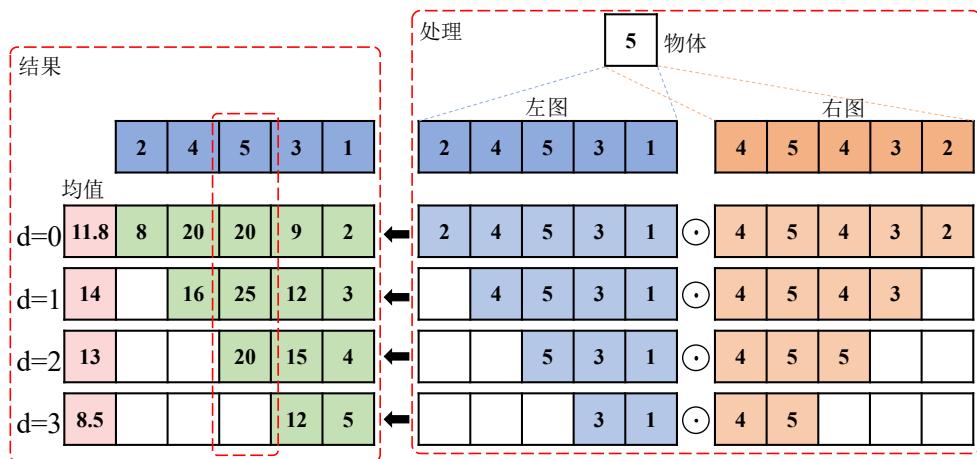


图 5-2 双目图像的相关性

的计算值最大，这是因为同一物体在左右目图像产生的图像虽然像素值不完全一致，但分布规律是一致的，每次计算时删除一个元素，相当于以左图为模块在右图上滑动寻找匹配块，整个过程类似于离散卷积后半段。这种非局部的相关性最早用来设计非局部均值滤波器，其核心思想是利用搜索框在整张图像上滑动寻找与当前点相似的区域，相似性由相似性度量函数评估，最终由相似区域确定当前点的加权系数。图像处理中的知名学者何恺明在文献 [62] 中对 Non-local 特性给出了详细的说明。

何恺明定义深度神经网络中非局部操作通用公式如下：

$$y_i = \frac{1}{\mathcal{C}(x)} \sum_{\forall j} f(x_i, x_j) g(x_j) \quad (5-4)$$

其中  $i$  为要计算其响应的输出位置索引， $j$  是所有可能位置的索引。 $x$  是输入信号（图像，序列，视频，通常是它们的特征）， $y$  是和  $x$  具有相同大小的输出信号。二元函数  $f$  用于计算位置  $i$  和所有位置  $j$  之间的某种关系，一元函数  $g$  表示输入信号在位置  $j$  的某种表示， $\mathcal{C}(x)$  用于对整个响应归一化。式 (5-4) 被称为非局部操作是因为：卷积操作往往只考虑局部输入加权和，递归操作一般只考虑当前帧和前一帧，相比之下非局部操作中输入信号所有元素均参与了计算过程。

对于函数  $f$  和  $g$ ，有多种选择。为了简化过程，对于  $g$  一般采用线性嵌入形式，即：

$$g(x_j) = W_g x_j \quad (5-5)$$

其中  $W_g$  是待学习的权重矩阵，一般通过尺寸为 1 的卷积核实现。二元函数  $f$  一般采用以下几种形式。

### (1) 高斯形式

沿着非局部均值和双边滤波器的思路，首先可以采用高斯形式的相似性度量函数：

$$f(x_i, x_j) = e^{x_i^T x_j} \quad (5-6)$$

其中  $x_i^T x_j$  为点积相似度。归一化因子采用求和形式：

$$\mathcal{C} = \sum_{\forall j} f(x_i, x_j) \quad (5-7)$$

### (2) 嵌入高斯形式

嵌入高斯形式是指在嵌入空间计算相似度，是高斯形式的一种扩展形式，可以表示为：

$$f(x_i, x_j) = e^{\theta(x_i)^T \varphi(x_j)} \quad (5-8)$$

在深度学习中，一般采用  $\theta(x_i) = W_\theta x_i$ ,  $\varphi(x_j) = W_\varphi x_j$  两种嵌入形式。归一化因子采用同式 (5-7) 相同的形式。Google 在自然语言处理中提出的自我注意力机制<sup>[60]</sup>可以视为嵌入高斯形式非局部操作的一种特殊形式。具体来说，对于某个给定的  $i$ ,  $\frac{1}{\mathcal{C}(x)} f(x_i, x_j)$  可视为在某个维度的位置  $j$  上进行 softmax 操作，因此我们有：

$$y = \text{Softmax}(x^T W_\theta^T W_\varphi x) g(x) \quad (5-9)$$

该式即为文献 [60] 中自我注意力机制。

### (3) 点积形式

相似性度量函数  $f$  同样可以被描述为点积的形式：

$$f(x_i, x_j) = \theta(x_i)^T \varphi(x_j) \quad (5-10)$$

为了简化梯度计算，该形式下的归一化因子可以选择元素总数，即：

$$\mathcal{C}(x) = N \quad (5-11)$$

其中  $N$  是输入中的元素总数，点积形式和嵌入高斯形式最大区别在于是否使用 Softmax 操作将其转换为概率分布。

### (4) 级联形式

级联形式一般用于视觉推理，基本形式如下：

$$f(x_i, x_j) = \text{ReLU}(w_f^T [\theta(x_i), \varphi(x_j)]) \quad (5-12)$$

其中  $[.]$  表示对输入在某一维度级联,  $w_f$  为权重矢量, 用于将级联后的特征映射到标量空间, 归一化因子可以采用元素个数。

以上是何恺明提出的四种相似性度量函数, 根据其实验结果以上四种形式的非局部操作对网络精度影响差异不大, 对预测精度起决定性作用的是网络是否引入非局部操作。在本节开始对左右目图像的操作实际上可以视为利用点积进行相似性度量的一种特殊形式, 特殊性在于点积计算整张图像间的相似性, 而对于视差, 相似性存在于同一行像素不同位置间, 因此我们可以将点积相似性度量推广为同一行像素间不同位置的 Hadamard 乘积形式。基于 Hadamard 相似性度量, 我们可以得到一种基于特征行相关性的 cost volume 构建方法。

### 5.2.2 基于特征相关性的 cost-volume 构建

构建 cost volume 的目的是以保留立体视觉几何学知识的方式约束模型<sup>[41]</sup>, 现有网络模型基本采用级联左右目图像语义特征的方式, 缺乏对特征相关性的利用。在上一节中对深度学习特征相关性做了详细说明, 基于上述 Hadamard 积相似性度量方法, 可以利用特征相关性按如下方式构建 cost volume。

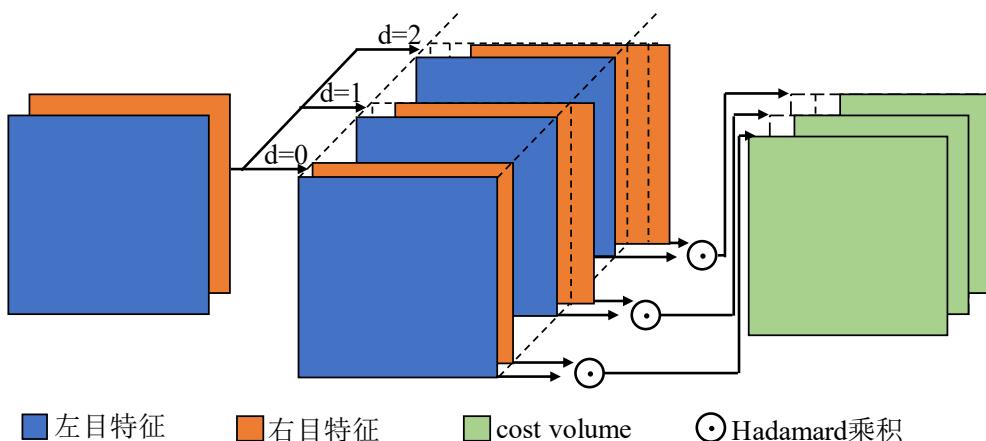


图 5-3 基于特征相关性的 cost volume 构建

对于特征提取网络输出的左右目特征, 取在通道维度上某一层特征  $f_l$  和  $f_r$ , 构建该层 cost volume 的方法如图5-3所示。设在该层特征尺度  $H \times W$  下视差最大值为  $d_{max}$ , 为在所有可能的  $d \in [0, d_{max} - 1]$  上构建 cost volume, 对于每一个  $d$ , 将  $f_r$  右移  $d$  位得到  $f_r[:, 0 : W - d]$ ,  $f_l$  左移  $d$  位得到  $f_l[:, d : W]$ 。然后将对应  $d$  的左右特征中移位后仍保留的元素逐位置相乘, 即计算 Hadamard 积, 并将结果存放在 cost volume 第三维, 在第三维的位置由  $f_l[:, d : W]$  在  $f_l$  中位置决定。对于单层特征可以按照上述过程构成三维特征, 在块维度和通道维度上重复以上过程最后可以得到五维 cost volume 特征。五个维度分别为块维度, 通道维度, 视差维度, 高度维度,

宽度维度。完整算法流程如算法5-1所示。

### 算法 5-1 cost volume 计算流程

```

Data: 特征提取模块输出特征  $F_l \in \mathbb{R}^{B \times C \times H \times W}$ ,  $F_r \in \mathbb{R}^{B \times C \times H \times W}$ ,  $H \times W$  尺寸  

    下对应的最大视差  $d_{max}$   

Result:  $Volume_{cost} \in \mathbb{R}^{B \times C \times d_{max} \times H \times W}$   

1 初始化  $Volume_{cost} = 0 \in \mathbb{R}^{B \times C \times d_{max} \times H \times W}$ ;  

2 for  $b = 0$ ;  $b < B$ ;  $b = b + 1$  do  

3   for  $c = 0$ ;  $c < C$ ;  $c = c + 1$  do  

4      $f_l = F_l[b, c, :, :] \in \mathbb{R}^{H \times W}$ ;  

5      $f_r = F_r[b, c, :, :] \in \mathbb{R}^{H \times W}$ ;  

6     for  $d = 0$ ;  $d < d_{max}$ ;  $d = d + 1$  do  

7        $Volume_{cost}[b, c, d, :, :] = f_l \odot f_r$ ;  

8       将  $f_l$  左移一个单位;  

9       将  $f_r$  右移一个单位;  

10      end  

11    end  

12  end  

13 返回  $Volume_{cost}$ ;
```

注意到在上述算法5-1中，输入特征为  $B \times C \times H \times W$  的四维特征，输出为  $B \times C \times d_{max} \times H \times W$  的五维特征  $Volume_{cost}$ 。即额外添加了视差维度，该维度下每个索引代表一个可能视差大小，每个元素则代表了左右目图像在该视差下匹配程度。具体来说，可以将  $Volume_{cost}$  中每个值视为：对于参与计算的每例样本(第一维度块维度  $B$ )，每个通道的特征(第二维度特征通道维度  $C$ )，在某个特定的视差(第三维度视差维度  $d_{max}$ )下，特征的每一行(第四维度高度维度  $H$ )里，每个左目特征元素(第五维度宽度维度  $W$ )与其对应的右目特征(位置由视差维度索引决定)的匹配程度。

虽然  $Volume_{cost}$  在视差维度上每个元素仅包含在特定视差下相似程度，但整个视差维度对输入左目特征中每个元素都建立了其与右目元素在  $[0, d_{max} - 1]$  所有可能视差范围内的相似性度量，这意味着网络在卷积时仅需要访问视差维度对应索引即可获取图像的匹配信息，无需通过多层卷积学习。相比较于级联 cost volume 构建方式，可以有效减少构建左右目特征匹配度所需要的网络深度。

例如，在使用网络进行推理时，若输入图像为  $1920 \times 1080$ ，最大视差为  $\frac{1}{4}$  图像宽度即 480，若采用级联方式构建 cost volume，并且在后续 3D 卷积中，卷积核大小采用  $3 \times 3 \times 3$ ，当卷积层深度达到 6 层时模型感知野为  $3^6 = 729$ ，仅能刚好覆盖视差范围 480。即对于最大视差处特征需要在第六层 3D 卷积中才能聚合，进一步评估相似性则需要更多卷积层。而 3D 卷积参数量远高于 2D 卷积，对于单个

模块 3D 卷积层数一般不超过 6 层，因此采用本节提出方法构建相似度 cost volume 可使网络聚焦于学习从相似度提取对应关系，建立概率分布模型并降低模型所需深度。

本节提出的 cost volume 构建方法同第四章中 Parallax-Attention(PA) 操作具有一定的相似性，但两者本质并不相同。首先，两种方法目的不同，本节方法旨在提出一种 cost volume 构建方法，PA 旨在实现特征融合；其次两者语义不同，本节方法的输出为特征在各个维度上的匹配程度，用于建立匹配关系，而 PA 直接建立对应关系，目的是用于视差计算；最后从计算过程来看，本节方法属于点积相似性度量的特殊推广，PA 则采用了嵌入高斯相似性度量作为中间过程计算视差。

### 5.3 多任务模型优化及目标函数

上一节中介绍基于特征相关性的 cost volume 构建方法时，需要用到第四章模型提取的双目特征；在双目特征修复模型中，对特征进行融合时，需要用特征的视差信息，在整个网络中存在两个任务，一是视差估计，二是图像修复，两个任务共享前端模型，如图5-4所示。根据文献[63]中对多任务深度学习模型的定义，该模型属于参数硬共享机制，可在一定程度上减轻模型过拟合风险，并且可以通过联合优化提升整体性能。

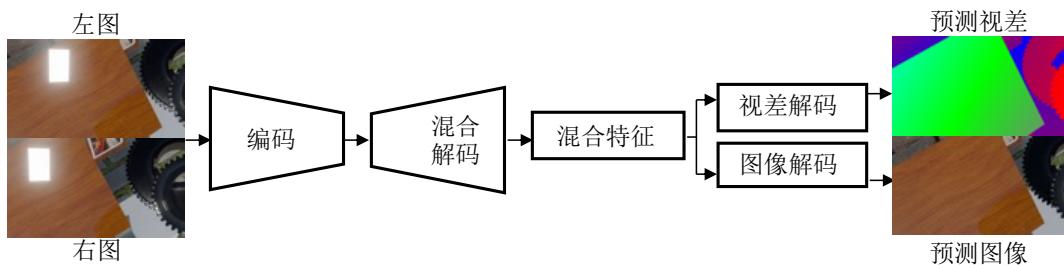


图 5-4 多任务模型

由于双目任务的特殊性，无论是双目图像超分辨，双目图像风格迁移还是双目图像修复都需要满足双目的对极约束，既保证视差前后不发生改变。视差信息作为双目图像场景的骨架从模型的输入流向输出。如果将视差解码模块中较为准确的视差信息引入到图像修复解码模块，可提升冗余信息融合的准确性，进一步提高图像修复精度；同时更为准确的冗余信息会通过反向传播影响特征提取模型的特征融合，为后续视差估计提供更加符合左右目一致性的特征，使得视差估计更为准确。因此下文将介绍如何建立视差和图像修复间关系实现联合优化。

### 5.3.1 多任务联合优化

#### 5.3.1.1 基于图像重构的损失函数优化

在本课题的多任务模型中主要有预测视差和修复图像两个输出，通过视差构建多任务优化的一个简单且直接的方法是利用视差和修复右图重建修复左图并与原始修复左图对比形成损失函数。直接采用第三章中的重构方法固然可以检验修复图像是否一致，但基于索引的剪切操作不连续不可微，无法用于反向传播，因此本课题采用 Jaderberg 提出的空间转换 (Spatial Transformer)<sup>[64]</sup> 实现图像重建。

为了将修复后尺寸为  $H \times W \times C$  的右图重建为左图，需要一组表示采样位置的采样点集  $G(x_i^s, y_i^s)$ ，对于采样点集中每一点都应用采样核采样得到输出  $V$ :

$$V_i^c = \sum_n^H \sum_m^W U_{nm}^c k(x_i^s - m; \Phi_x) k(y_i^s - n; \Phi_y) \quad (5-13)$$

$$\forall i \in [1, \dots, HW] \quad \forall c \in [1, \dots, C]$$

其中  $\Phi_x$  和  $\Phi_y$  表示通用采样内核  $k$  的参数， $k$  定义了采样方式。 $U_{nm}^c$  为输入在通道  $c$  位置  $(n, m)$  的值， $V_i^c$  表示第  $i$  个采样点输出，位置在输出  $c$  通道的  $(x_i^t, y_i^t)$  处。以双线性采样为例，则式 (5-13) 可以修改为：

$$V_i^c = \sum_n^H \sum_m^W U_{nm}^c \max(0, 1 - |x_i^s - m|) \max(0, 1 - |y_i^s - n|) \quad (5-14)$$

根据文献 [64]，式 (5-14) 相对于  $U$  和  $G$  的偏导数如下：

$$\frac{\partial V_i^c}{\partial U_{nm}^c} = \sum_n^H \sum_m^W \max(0, 1 - |x_i^s - m|) \max(0, 1 - |y_i^s - n|) \quad (5-15)$$

$$\frac{\partial V_i^c}{\partial x_i^s} = \sum_n^H \sum_m^W U_{nm}^c \max(0, 1 - |y_i^s - n|) \begin{cases} 0 & if \quad |m - x_i^s| \geq 1 \\ 1 & if \quad m \geq x_i^s \\ -1 & if \quad m < x_i^s \end{cases} \quad (5-16)$$

偏导  $\frac{\partial V_i^c}{\partial y_i^s}$  与式 (5-16) 类似。该采样方式对输入  $U$  和采样点  $G$  均有偏导，若将修复后右图  $I_{predr}$  视为输入  $U$ ，预测视差  $d_l$  视为采样点  $G$ ，则可以将误差梯度通过反向传播分别传递到图像修复模型和视差估计模型，实现多任务优化。根据定义，采样点  $(x_i^s, y_i^s)$  为待采样像素坐标的归一化值，为了将视差转换为采样点需要将视差减去自身横坐标索引，并对图像宽度归一化。

利用上述采样方式，可以构建如下图像重构损失函数：

$$\mathcal{L}_{unsup} = \|warp(I_{predr}, d_l) - I_{predl}\|_2 \quad (5-17)$$

其中  $warp$  表示使用 Spatial Transformer 重构左图， $I_{predl}$  和  $I_{predr}$  表示图像修复网络输出， $d_l$  表示视差预测网络输出。 $\mathcal{L}_{recon}$  以图像重建方式联合了修复图像和预测视差，在损失函数层面建立了图像修复任务和视差估计任务的关系，促进两个任务协同学习。该损失函数通过图像实现对视差的监督学习，因此在一般场景中也可用于视差估计的无监督学习。

### 5.3.1.2 基于特征重构的网络结构优化

视差估计模块和图像修复模块为两个独立模块，两者在前向传递，反向传播和推理时互不干扰，不存在先后顺序之分。注意到进行图像解码时已经可以通过视差解码模块获得较为准确的视差估计  $d_{pred}$ 。因此我们可以通过下采样  $d_{pred}$  至合适尺寸用于图像修复模块中特征融合，为冗余信息提供更为准确的位置信息。同时意味着在实际应用中，本课题提出的模型在推理视差时无需通过图像修复模块，可以节省部分计算资源。

使用  $d_{pred}$  融合图像修复特征  $f_l$  和  $f_r$  相当于将视差的计算路径引入图像修复中， $\varphi_{repair}$  表示图像解码模块，则修复后图像  $I_{repair}$  可以表示为：

$$I_{repair} = \varphi_{repair}(warp(d_{pred}, f_r), f_l) \quad (5-18)$$

若将式 (5-18) 代入第四章图像修复损失函数 (4-14) 中，可以发现图像修复损失函数将对视差产生梯度，即网络在学习修复图像的同时也会提高视差估计精度。

### 5.3.2 目标函数

在第四章，针对图像修复模块采用了多个损失函数如下

$$\mathcal{L}_{repair} = \mathcal{L}_{normal} + 6\mathcal{L}_{oe} + 0.05\mathcal{L}_{perceptual} + 0.1\mathcal{L}_{tv} \quad (5-19)$$

对于视差估计的损失函数同文献 [42] 保持一致，网络在堆叠沙漏结构下分别产生预测视差  $d_{pred1}$ ， $d_{pred2}$  和  $d_{pred3}$ ，视差标注为  $d_{gt}$ ，视差估计损失函数为：

$$\begin{aligned} \mathcal{L}_{disp} = & 0.5 \times \mathcal{L}_{pred}(d_{pred1}, d_{gt}) \\ & + 0.7 \times \mathcal{L}_{pred}(d_{pred2}, d_{gt}) \\ & + \mathcal{L}_{pred}(d_{pred3}, d_{gt}) \end{aligned} \quad (5-20)$$

其中  $\mathcal{L}_{pred}$  为单层估计结果损失函数:

$$\mathcal{L}_{pred}(\hat{d}, d) = \frac{1}{N} \sum_{i=1}^N smooth_{L1}(d_i - \hat{d}_i) \quad (5-21)$$

$N$  表示有效标注的视差个数,  $smooth_{L1}$  表示平滑的  $L_1$  范数, 具体如下:

$$smooth_{L1}(x) = \begin{cases} 0.5x^2 & if |x| < 1 \\ |x| - 0.5 & otherwise \end{cases} \quad (5-22)$$

同  $L_2$  范数相比,  $smooth_{L1}$  鲁棒性更好且对离群值敏感度更低。综合以上各个损失函数, 网络最终的目标函数为:

$$\mathcal{L} = \mathcal{L}_{repair} + \mathcal{L}_{disp} + 0.1\mathcal{L}_{unsup} \quad (5-23)$$

无监督损失函数的权重仅为 0.1, 因为即使修复后图像完全正确, 其像素值不可能完全一致, 添加无监督损失函数的目的更多的是为了在缺少标注数据的位置优化梯度。

## 5.4 实验结果与分析

本章针对过曝场景下视差预测精度问题, 研究了基于特征相关性的 cost volume 构建方法和多任务联合优化。为了验证本课题提出的方法是否有效, 首先在过曝场景 Scene1 进行消融实验, 并与现有技术进行对比实验。接着测试了模型在列车过曝场景中的表现, 针对无标注的列车场景, 进一步开展了无监督相关实验。

为防止过拟合, 过曝场景下实验采用在 Scene2 上训练, Scene1 上跨场景测试的形式进行, 在过曝场景 Scene1 上消融实验的客观指标如表5-1所示, PSMNet 表示本课题基线模型; PSMNet w/mix 表示本课题在第四章中提出的模型; PSMNet w/corr 表示将本章提出的 cost volume 构建方法应用于 PSMNet 中得到的模型; Ours w/o optim 表示 PSMNet w/mix 和 PSMNet w/corr 结合得到的模型; Ours 表示在 Ours w/o optim 基础上添加多任务联合优化得到的模型, 即本课题提出的完整模型。PSMNet 采用文献 [42] 官方代码, 所有模型均基于 Sceneflow Dataset 预训练权重进一步训练得到。

根据表5-1不难发现, 采用多特征提取模块, PSMNet w/mix 在过曝场景上预测平均像素误差由 10.827px 降低到 6.926px, 性能得到了显著提升。PSMNet w/mix 在过曝区域的预测性能基本与未过曝区域一致, 这得益于本文提出的特征提取模块中对左右特征进行了充分融合, 利用冗余信息填充了过曝区域丢失特征, 使得

表 5-1 Scene1 过曝场景消融实验

method	>2px		>3px		>5px		mean	
	all	oe	all	oe	all	oe	all	oe
PSMNet	25.42%	31.51%	24.54%	31.27%	22.75%	30.48%	10.827	15.042
PSMNet w/mix	20.62%	22.49%	19.82%	22.08%	18.24%	20.73%	6.926	8.743
PSMNet w/corr	24.90%	23.78%	23.73%	22.83%	21.78%	21.35%	7.944	9.375
Ours w/o optim	20.60%	<b>19.43%</b>	19.59%	<b>18.62%</b>	17.87%	<b>17.11%</b>	7.239	<b>5.139</b>
Ours	<b>19.22%</b>	20.01%	<b>18.24%</b>	19.60%	<b>16.35%</b>	18.00%	<b>5.830</b>	6.658

网络可以正确计算对应关系。采用基于相关性的 cost volume 后，PSMNet w/corr 在整个场景上的表现得到了进一步的提升，但在过曝区域的预测结果仍不理想。该现象说明了基于相关性的 cost volume 可以减轻网络学习压力，使其聚焦于学习计算对应关系任务本身，但是无法解决过曝区域无法匹配的问题。综合以上两点得到模型 Ours w/o optim，其无论是在普通场景还是过曝场景，视差预测性能都有较大的提升，这一方面说明第四章中提出的模型可充分利用冗余信息构建高度一致的左右目特征，另一方面说明本课题改进的 Hadamard 积相似性度量方式可以较好反映特征匹配关系。对比 Ours w/o optim 和 Ours 两个模型的客观指标可知，本课题从多任务模型思路出发，在网络结构和损失函数上的改进可以帮助图像修复任务和视差估计任务更好的工作，实现模型整体预测精度的提高。

部分模型的主观结果如图5-5所示，每列为单例样本的各模型测试结果，前两行为输入左右目图像，第三行为标注视差，第四行到第七行分别为 Ours、PSMNet、PSMNet w/ mix 和 PSMNet w/ corr 模型视差预测结果。就主观结果而言，Ours 模型在过曝区域基本实现正常预测视差，效果最好，PSMNet 模型在过曝区域基本失效，添加了第四章提出的多特征提取模块后效果有所改善。

本课题对比了多个模型在过曝场景下的性能，客观指标如表5-2所示。其中 DispNet 为文献 [11] 中提出的方法，网络结构采用了简单的全卷积形式，因此泛化性能较差。DispNetC 为在 DispNet 基础上添加 cost volume 后的模型，GANet 为 Feihu 等人<sup>[2]</sup> 等人针对 PSMNet 的改进模型，采用了一种多方向聚合的类卷积层代替网络中原本的 3D 卷积，对过曝场景有一定的适用性。AANet 为 Xu 等人<sup>[3]</sup> 对 GANet 的改进。由表可知，现有模型在过曝场景下均不如本课题提出的模型。

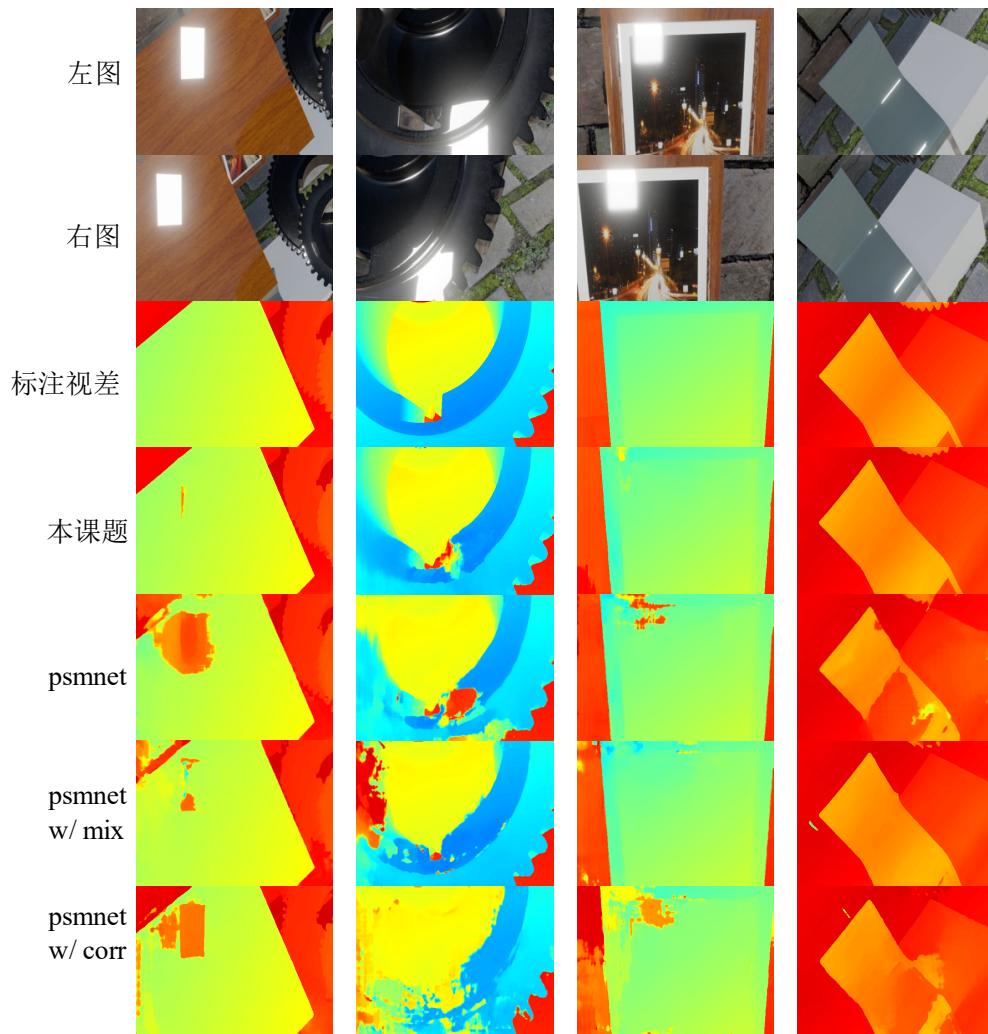


图 5-5 Scene1 场景主观结果

表 5-2 多模型过曝场景下性能对比

method	>2px		3px		5px		mean	
	all	oe	all	oe	all	oe	all	oe
DispNet	45.89%	47.96%	44.35%	47.12%	40.86%	45.10%	20.169	23.407
DispNetCorr	46.95%	48.09%	45.54%	47.24%	42.34%	45.12%	20.926	23.459
GANet	26.31%	25.44%	25.58%	25.16%	24.26%	24.46%	10.710	11.396
AANet	26.92%	28.64%	25.95%	28.27%	23.49%	27.29%	9.523	14.163
PSMNet	25.42%	31.51%	24.54%	31.27%	22.75%	30.48%	10.827	15.042
Ours	<b>19.22%</b>	<b>20.01%</b>	<b>18.24%</b>	<b>19.60%</b>	<b>16.35%</b>	<b>18.00%</b>	<b>5.830</b>	<b>6.658</b>

为了验证模型在列车过曝场景数据集 Texture 上的表现，利用课题构建的过曝数据对模型预训练，在 texture 上 finetune 后测试的客观指标见表5-3中 Texture 标

表 5-3 真实场景客观指标

method	>2px	>3px	>5px	mean
PSMNet&KITTI	1.97%	1.27%	0.63%	0.523
Ours&KITTI	1.74%	1.15%	0.64%	0.507
PSMNet&Texture	1.91%	1.90%	1.80%	1.001
Ours&Texture	2.28%	2.27%	2.18%	1.009

注，KITTI 标注为在 KITTI Stereo 2015 数据集上测试结果。KITTI 测试结果表明本课题提出的模型在普通场景的视差估计精度优于基线模型。Texture 数据缺少过曝区域视差标注，所以过曝区域的视差估计精度仅能通过主观结果评价。Texture 场景的主观测试结果如图5-6所示，考虑到列车底部部件多为平滑连续的金属面，视差分布应该连续、平滑，在平板区域不存在视差突变现象。观察本课题结果和 PSMNet 结果可知，在过曝区域本课题模型的预测视差几乎不存在突变现象，与过曝区域周围的视差值衔接平滑且连续，相较于 PSMNet 中大量的异常视差点，能够更好的反应部件的三维信息。

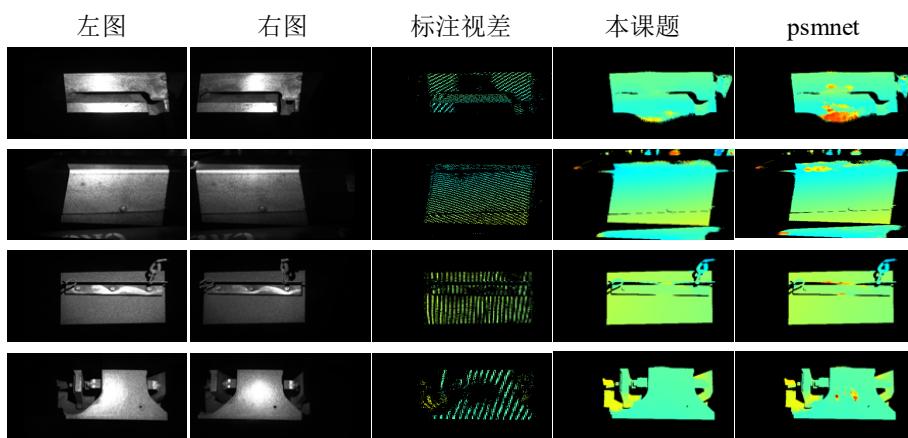


图 5-6 Texture 场景主观结果

在列车部件的光学测量场景中，想获取整个节车厢的训练数据是难以实现的，比较可行的方法是以少量样本对模型 finetune 后采用无监督学习实现大面积的应用。无监督学习的难点在于左右目图像的像素值无法一一对应，难以构建匹配关系监督网络学习。如图5-7所示，得益于本课题多特征模型以及图像修复模块，利用修复后图像进行无监督学习的视差预测结果，在物体边缘，复杂光源区域，相机畸变区域等左右目一致性被破坏区域的视差预测结果要优于 PSMNet。因此本课题提出的模型在列车部件测量实际应用中，实现了更好的视差预测结果。

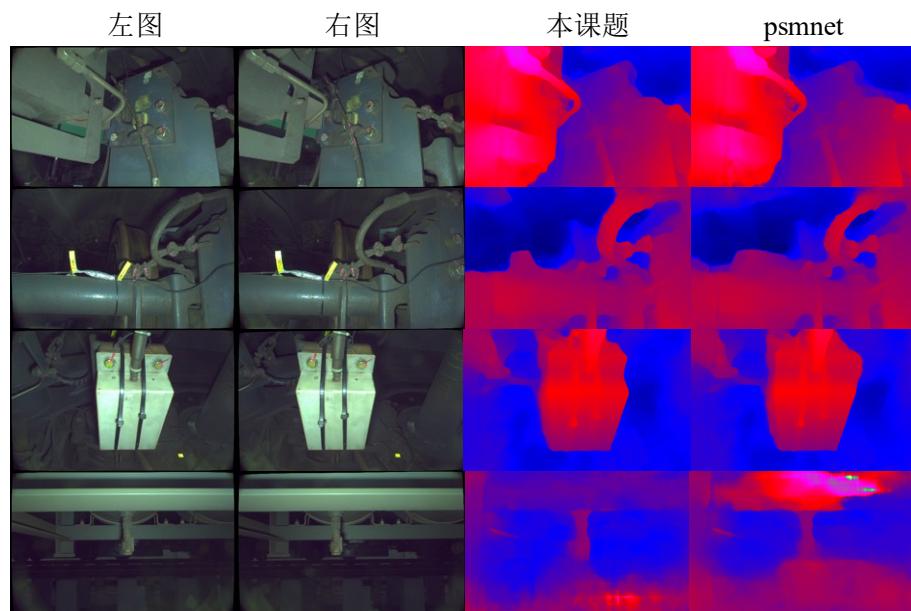


图 5-7 无监督主观结果

## 5.5 本章小结

本章首先介绍了本课题采用的双目立体匹配网络主要结构，对各个模块原理做了详细说明。接着分析了现有模型中 cost volume 不合理性，并说明了深度学习中特征相关性，基于特征相关性提出了一种 cost volume 构建方法。然后从多任务模型的角度分析了本课题提出的模型结构，并从目标函数和网络结构两个方向进行优化。最后通过消融实验和对比实验证明了本课题提出的模型在过曝场景下的有效性。在列车高反光部件的光学测量场景中，实现了显著优于现有方法的视差预测结果。无监督学习的实验进一步证明了本课题提出的模型在工业测量中具有更高的实用价值。

## 第六章 全文总结与展望

### 6.1 全文总结

双目立体匹配作为计算机视觉中一项基础且重要的任务，广泛应用于路径规划、光学测量和即时定位与地图构建等领域。但当场景中存在高反光物体引起的过曝现象时，面临着误匹配，视差精度低的问题。本文围绕“高铁列车双目视差估计项目”需求，研究了过曝场景下双目立体匹配技术，重点研究了过曝场景数据集构建、双目过曝特征修复和精度联合优化，论文的主要工作和研究成果如下：

1、本文分析了现有数据集的场景特点和采集方式，阐明了如何利用 Blender 渲染双目视差数据集，如何设置相关参数模拟列车底部高反光部件特性，建立了深度视差转换模型，构建了过曝场景下双目视差数据集，填补了相关领域的空白。图像重构实验表明，由深度值转换得来的视差值可精确重建图像，视差标注正确，可用于精度评估和真实场景模型预训练。

2、本文利用双目冗余性解决过曝场景下特征不一致的问题。本文提出的多特征提取模块，以编码阶段权重共享，解码阶段特征共享的形式，同时提取双目图像语义特征，解决了基于权重共享的特征提取模块需要分别计算双目图像特征，导致语义特征来源单一，无法修复过曝区域丢失特征的问题；本文提出的特征融合模块，基于视差注意力机制，从来自模型不同深度和路径的多语义特征中提取冗余信息，基于冗余信息和特征视差重构双目特征，实现多特征融合，解决了传统卷积层难以利用特征位置关系的问题；本文提出的图像修复模块，利用预测视差重构无过曝图像，并以目标函数的形式促进了冗余信息在左右目特征间的流动。消融实验表明，将上述模块分别添加到基线模型 PSMNet 中，同 PSMNet 相比，将过曝区域视差预测的平均像素误差由 15.04 像素分别降低至 14.69 像素、13.70 像素和 12.42 像素，证明上述模块对改善过曝区域的视差预测结果均有积极作用。将上述模块均添加到 PSMNet 中，则同 PSMNet 相比，可将整体预测的平均像素误差由 10.82 像素降低至 6.92 像素，预测精度提升 36%。

3、本文提出了基于上述多特征提取的双目立体匹配网络，从 cost volume 和多任务两个方向改进网络。本文提出的基于特征相关性的 cost volume 构建方法，利用双目立体匹配具有行特征相关性的特点，将点积相关性度量推广至 Hadamard 积形式，在双目匹配领域中解决了卷积神经网络学习中长距离对应关系困难的问题。为了进一步提高模型在过曝场景的整体预测精度，从网络结构和目标函数两方面对模型联合优化。实验表明，在过曝场景下，同主流方法相比，本文提出的模型可

将过曝场景下视差估计的平均像素误差由  $10.17(\pm 0.65)$  像素降低至 5.83 像素，精度提升约 40%。在“高铁列车双目视差估计项目”的生产环境中存在大量高反光部件，双目相机拍摄的图像中具有明显的过曝现象，本文提出的模型在上述过曝场景的视差预测结果显著优于现有方法，为后续故障检测、定位与分析提供了准确的三维信息。

## 6.2 后续工作展望

本文提出的改进方法虽然提高了双目立体匹配模型在过曝场景下的预测精度，但仍存在以下问题尚未解决：

1、本文利用 Blender 构建的过曝场景数据集包含 900 例样本，数量上已超过常用的真实场景数据集，但相比于虚拟数据集而言远远不够，同时虚拟数据和真实数据的分布并不完全一致。后续研究可针对构建更符合真实场景的过曝数据集展开。

2、本文的图像修复解码模块采用了较为简单的模型结构，虽然起到了引导模型学习融合冗余特征的作用，但图像修复的结果并不理想。后续研究可针对提高修复图像的准确性展开。

3、本文的视差解码模块中包含多个 3D 卷积模块，3D 卷积参数量大而且计算耗时，网络的推理速度难以满足自动机器人导航和自动驾驶场景的实时性要求。后续研究可针对降低模型复杂度展开。

## 致 谢

在攻读硕士期间，许多人对我的学术研究和日常生活给予了莫大的帮助，在此我向你们表达我由衷的感谢。

首先感谢我的导师刘光辉教授，三年来刘老师从我的兴趣出发，尽可能的为我提供科研平台和研究机会，很大程度的锻炼和提升了我的科研能力。在论文的选题时，刘老师结合我的科研经历和我交流并分析了双目立体匹配技术面临的难题。在实验过程中，刘老师帮助我制定了详细的实验计划和实施方案。在论文写作过程中，从章节逻辑到遣词造句，刘老师给我提供了仔细的指导，反复的帮助我修改论文。刘老师这种严于律己，对待科研认真负责的态度，对我的学习和生活产生了深刻的影响。

其次，我要感谢孟凡满老师。感谢孟老师在研一期间带领我阅读最新的会议论文，给予我科研上的指导。同时也感谢孟老师在我撰写毕业论文时，提出了宝贵的修改意见。

接着，我要感谢我的母校电子科技大学。是电子科技大学给我了这样浓厚的学术环境，在这里的七年，我不仅学到了知识，更认识到了许多志同道合的伙伴。感谢教研室的各位师兄师姐对我科研生活上的帮助，感谢陈谧、李林洲和张泰东三位同窗在这三年里给予我的帮助。感谢我的室友谌昕宇、海宇和李世超，三年来我们朝夕相处，共同进步。

最后，感谢我的父母，对我的生活和学习给予了持久的付出，在我遇到困难时，给予我鼓励和希望。

## 参考文献

- [1] 中华人民共和国发展和改革委员会. 智能汽车创新发展战略 [EB/OL]. [https://www.ndrc.gov.cn/xxgk/zcfb/tz/202002/t20200224\\_1221077.html](https://www.ndrc.gov.cn/xxgk/zcfb/tz/202002/t20200224_1221077.html), Feb 24, 2020
- [2] F. Zhang, V. Prisacariu, R. Yang, et al. Ga-net: Guided aggregation net for end-to-end stereo matching[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, 2019, 185-194
- [3] H. Xu, J. Zhang. Aanet: Adaptive aggregation network for efficient stereo matching[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2020, 1959-1968
- [4] A. Geiger, P. Lenz, R. Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Providence, 2012, 3354-3361
- [5] M. Menze, C. Heipke, A. Geiger. Joint 3d estimation of vehicles and scene flow[J]. ISPRS annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences, 2015, 2: 427-428
- [6] D. Scharstein, R. Szeliski. A taxonomy and evaluation of dense two-frame stereo correspondence algorithms[J]. International Journal of Computer Vision, 2002, 47(1-3): 7-42
- [7] D. Scharstein, R. Szeliski. High-accuracy stereo depth maps using structured light[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Madison, 2003, 214-231
- [8] D. Scharstein, C. Pal. Learning conditional random fields for stereo[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Minneapolis, 2007, 1-8
- [9] H. Hirschmüller, D. Scharstein. Evaluation of cost functions for stereo matching[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Minneapolis, 2007, 1-8
- [10] D. Scharstein, H. Hirschmüller, Y. Kitajima, et al. High-resolution stereo datasets with subpixel-accurate ground truth[C]. German Conference on Pattern Recognition, Münster, 2014, 31-42
- [11] N. Mayer, E. Ilg, P. Hausser, et al. A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, 2016, 4040-4048
- [12] A. Gaidon, Q. Wang, Y. Cabon, et al. Virtual worlds as proxy for multi-object tracking analysis[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, 2016, 4340-4349

- [13] S. Mattoccia, S. Giardino, A. Gambini. Accurate and efficient cost aggregation strategy for stereo correspondence based on approximated joint bilateral filtering[C]. Proceedings of the Asian Conference on Computer Vision, Xi'an, 2009, 371-380
- [14] A. Arranz, Á. Sánchez, M. Alvar. Multiresolution energy minimisation framework for stereo matching[J]. IET Computer Vision, 2012, 6(5): 425-434
- [15] L. Xu, O. Au, W. Sun, et al. Stereo matching by adaptive weighting selection based cost aggregation[C]. Proceedings of the IEEE International Symposium on Circuits and Systems, Beijing, 2013, 1420-1423
- [16] J. Salmen, M. Schlippling, J. Edelbrunner, et al. Real-time stereo vision: making more out of dynamic programming[C]. Proceedings of the International Conference on Computer Analysis of Images and Patterns, Salerno, 2009, 1096-1103
- [17] J. M. Pérez, P. Sánchez. Real-time stereo matching using memory-efficient belief propagation for high-definition 3d telepresence systems[J]. Pattern Recognition Letters, 2011, 32(16): 2250-2253
- [18] Y. Wang, C. Tung, P. Chung. Efficient disparity estimation using hierarchical bilateral disparity structure based graph cut algorithm with a foreground boundary refinement mechanism[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2012, 23(5): 784-801
- [19] D. Min, J. Lu, M. N. Do. A revisit to cost aggregation in stereo matching: How far can we reduce its computational redundancy?[C]. Proceedings of the International Conference on Computer Vision, Barcelona, 2011, 1567-1574
- [20] C. C. Pham, J. W. Jeon. Domain transformation-based efficient cost aggregation for local stereo matching[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2012, 23(7): 1119-1130
- [21] B. J. Tippetts, D. Lee, J. K. Archibald, et al. Dense disparity real-time stereo vision algorithm for resource-limited systems[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2011, 21(10): 1547-1555
- [22] S. H. Lee, S. Sharma. Real-time disparity estimation algorithm for stereo camera systems[J]. IEEE Transactions on Consumer Electronics, 2011, 57(3): 1018-1026
- [23] R. K. Gupta, S. Cho. Window-based approach for fast stereo correspondence[J]. IET Computer Vision, 2013, 7(2): 123-134
- [24] K. Sharma, K. Jeong, S. Kim. Vision based autonomous vehicle navigation with self-organizing map feature matching technique[C]. Proceedings of the International Conference on Control, Automation and Systems, Singapore, 2011, 946-949

- [25] J. Liu, X. Sang, C. Jia, et al. Efficient stereo matching algorithm with edge-detecting[C]. Proceedings of the Optoelectronic Imaging and Multimedia Technology III, Beijing, 2014, 927335-927346
- [26] Q. Yang, P. Ji, D. Li, et al. Fast stereo matching using adaptive guided filtering[J]. Image and Vision Computing, 2014, 32(3): 202-211
- [27] H. Hirschmüller, P. R. Innocent, J. Garibaldi. Real-time correlation-based stereo vision with reduced border errors[J]. International Journal of Computer Vision, 2002, 47(1): 229-246
- [28] J. Lu, G. Lafruit, F. Catthoor. Anisotropic local high-confidence voting for accurate stereo correspondence[C]. Proceedings of the Image Processing: Algorithms and Systems VI, California, 2008, 536-547
- [29] K. Chen, C. Su. Reducing computation complexity for disparity matching[C]. Proceedings of the IEEE International Symposium on Circuits and Systems, Beijing, 2013, 2916-2919
- [30] J. Fang, A. L. Varbanescu, J. Shen, et al. Accelerating cost aggregation for real-time stereo matching[C]. Proceedings of the International Conference on Parallel and Distributed Systems, Singapore, 2012, 472-481
- [31] K. Yoon, I. S. Kweon. Adaptive support-weight approach for correspondence search[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2006, 28(4): 650-656
- [32] C. Cigla, A. A. Alatan. Information permeability for stereo matching[J]. Signal Processing: Image Communication, 2013, 28(9): 1072-1088
- [33] K. Zhang, J. Lu, Q. Yang, et al. Real-time and accurate stereo: A scalable approach with bitwise fast voting on cuda[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2011, 21(7): 867-878
- [34] Z. Lee, J. Juang, T. Q. Nguyen. Local disparity estimation with three-moded cross census and advanced support weight[J]. IEEE Transactions on Multimedia, 2013, 15(8): 1855-1864
- [35] H. Wang, M. Wu, Y. Zhang, et al. Effective stereo matching using reliable points based graph cut[C]. Proceedings of the Visual Communications and Image Processing, Kuching, 2013, 1-6
- [36] H. Hirschmüller. Stereo processing by semiglobal matching and mutual information[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2007, 30(2): 328-341
- [37] R. Haeusler, R. Nair, D. Kondermann. Ensemble learning for confidence measures in stereo vision[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Portland, 2013, 305-312

- [38] J. Zbontar, Y. LeCun. Stereo matching by training a convolutional neural network to compare image patches.[J]. *J. Mach. Learn. Res.*, 2016, 17(1): 2287-2318
- [39] J. Pang, W. Sun, J. S. Ren, et al. Cascade residual learning: A two-stage convolutional neural network for stereo matching[C]. Proceedings of the IEEE International Conference on Computer Vision Workshops, Venice, 2017, 887-895
- [40] Z. Liang, Y. Feng, Y. Guo, et al. Learning for disparity estimation through feature constancy[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, 2018, 2811-2820
- [41] A. Kendall, H. Martirosyan, S. Dasgupta, et al. End-to-end learning of geometry and context for deep stereo regression[C]. Proceedings of the IEEE International Conference on Computer Vision, Venice, 2017, 66-75
- [42] J. Chang, Y. Chen. Pyramid stereo matching network[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, 2018, 5410-5418
- [43] D. A. Forsyth, J. Ponce. Computer vision: a modern approach[M]. Prentice Hall Professional Technical Reference, 2002, 168-169
- [44] H. C. Longuet-Higgins. A computer algorithm for reconstructing a scene from two projections[J]. *Nature*, 1981, 293(5828): 133-135
- [45] W. S. McCulloch, W. Pitts. A logical calculus of the ideas immanent in nervous activity[J]. *The Bulletin of Mathematical Biophysics*, 1943, 5(4): 115-133
- [46] G. E. Hinton, S. Osindero, Y. Teh. A fast learning algorithm for deep belief nets[J]. *Neural Computation*, 2006, 18(7): 1527-1554
- [47] I. Goodfellow, Y. Bengio, A. Courville, et al. Deep learning[M]. MIT press Cambridge, 2016, 32-33
- [48] F. Rosenblatt. The perceptron: a probabilistic model for information storage and organization in the brain.[J]. *Psychological Review*, 1958, 65(6): 386-387
- [49] A. Krizhevsky, I. Sutskever, G. E. Hinton. Imagenet classification with deep convolutional neural networks[J]. *Advances in Neural Information Processing Systems*, 2012, 25: 1097-1105
- [50] K. He, X. Zhang, S. Ren, et al. Deep residual learning for image recognition[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, 2016, 770-778
- [51] G. Huang, Z. Liu, L. Van Der Maaten, et al. Densely connected convolutional networks[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Hawaii, 2017, 4700-4708

- [52] Y. LeCun, L. Bottou, Y. Bengio, et al. Gradient-based learning applied to document recognition[J]. Proceedings of the IEEE, 1998, 86(11): 2278-2324
- [53] D. E. Rumelhart, G. E. Hinton, R. J. Williams. Learning representations by back-propagating errors[J]. Nature, 1986, 323(6088): 533-536
- [54] J. Kim, V. Kolmogorov, R. Zabih. Visual correspondence using energy minimization and mutual information[C]. Proceedings of the IEEE International Conference on Computer Vision, Nice, 2003, 1033-1040
- [55] D. Ciresan, A. Giusti, L. Gambardella, et al. Deep neural networks segment neuronal membranes in electron microscopy images[J]. Advances in Neural Information Processing Systems, 2012, 25: 2843-2851
- [56] O. Ronneberger, P. Fischer, T. Brox. U-net: Convolutional networks for biomedical image segmentation[C]. Proceedings of the International Conference on Medical Image Computing and Computer Assisted Intervention, Munich, 2015, 234-241
- [57] P. Isola, J. Zhu, T. Zhou, et al. Image-to-image translation with conditional adversarial networks[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, 2017, 1125-1134
- [58] G. Liu, F. A. Reda, K. J. Shih, et al. Image inpainting for irregular holes using partial convolutions[C]. Proceedings of the European Conference on Computer Vision, Munich, 2018, 85-100
- [59] K. He, X. Zhang, S. Ren, et al. Deep residual learning for image recognition[J]. CoRR, 2015, abs/1512.03385
- [60] A. Vaswani, N. Shazeer, N. Parmar, et al. Attention is all you need[J]. arXiv preprint arXiv:1706.03762, 2017
- [61] J. Johnson, A. Alahi, F. Li. Perceptual losses for real-time style transfer and super-resolution[C]. Proceedings of the European Conference on Computer Vision, Amsterdam, 2016, 694-711
- [62] X. Wang, R. Girshick, A. Gupta, et al. Non-local neural networks[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, 2018, 7794-7803
- [63] S. Ruder. An overview of multi-task learning in deep neural networks[J]. CoRR, 2017, abs/1706.05098
- [64] M. Jaderberg, K. Simonyan, A. Zisserman, et al. Spatial transformer networks[J]. Advances in Neural Information Processing Systems, 2015, 23: 2017-2025

## 攻读硕士学位期间取得的成果

### 申请的国家发明专利：

- [1] 刘光辉, 朱志鹏, 孙铁成, 李茹, 徐增荣. 一种基于多尺度网络的稀疏深度稠密化方法 [P]. 中国, CN109685842A, 2019.04.26
- [2] 刘光辉, 孙铁成, 朱志鹏, 李茹, 徐增荣, 廖岳鹏, 朱树元. 一种摄像头和激光雷达融合的端到端目标检测方法 [P]. 中国, CN111027401A, 2020.04.17

### 参与的科研项目：

- [1] 成都主导科技有限责任公司, 受电弓三维数据展示项目, 2017.12-2018.12
- [2] 成都铁安科技有限责任公司, 高铁列车双目视差估计项目, 2019.11-2020.11

### 获奖情况：

- [1] 2018-2019 学年度 研究生一等奖学金
- [2] 2019-2020 学年度 研究生三等奖学金