分类号：TH789　　学校代码：10363
密　级：公开　　学　　号：2210120133

安徽工程大学
Anhui Polytechnic University

硕士学位论文

题目　基于无监督学习的单目视觉深度估计算法研究

论文作者　李屹
校内导师　时培成（教授）
校外导师　倪绍勇（高级工程师）

专业学位类别　机械
研究方向（领域）　自动驾驶环境感知

论文提交日期：2024 年 5 月 30 日

分类号：TH789　　学校代码：10363

密 级：公开　　学　号：2210120133

基于无监督学习的单目视觉深度估计算法研究

RESEARCH ON MONOCULAR VISUAL DEPTH ESTIMATION ALGORITHM BASED ON UNSUPERVISED LEARNING

学生姓名：李屹

校内导师：时培成（教授）

校外导师　倪绍勇（高级工程师）

专　业：机械

研究方向：自动驾驶环境感知

论文答辩日期：2024 年 5 月 26 日

安徽工程大学学位论文原创性声明

本人郑重声明：我恪守学术道德，崇尚严谨学风。所呈交的学位论文，是本人在导师的指导下，独立进行研究工作所取得的成果。除文中已明确注明和引用的内容外，本论文不包含任何其他个人或集体已经发表或撰写过的作品及成果的内容。论文为本人亲自撰写，我对所写的内容负责，并完全意识到本声明的法律后果由本人承担。

学位论文作者签名： 

日期：2024 年 5 月 30 日

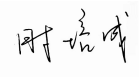
安徽工程大学学位论文版权使用授权书

学位论文作者完全了解学校有关保留、使用学位论文的规定，同意学校保留并向国家有关部门或机构送交论文的复印件和电子版，允许论文被查阅或借阅。本人授权安徽工程大学可以将本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和汇编本学位论文。

保密 □，在　　年解密后适用本版权书。

本学位论文属于

不保密 □。

学位论文作者签名：  　　指导教师签名： 

日期：　2024 年 5 月　30 日　　日期：　2024 年　5 月　30 日

基于无监督学习的单目视觉深度估计算法研究

摘 要

获取场景中的精确深度，是计算机视觉中三维环境感知中的核心课题。考虑到深度传感器获取场景中的深度成本较高、传统单目深度估计算法获取场景中的深度精度较低、深度学习单目深度估计算法中监督学习算法需要大量人工标注数据集等缺陷，近年来研究人员提出了基于无监督学习的单目视觉深度估计算法，一方面模型部署成本比专业深度传感器要低，另一方面与其它单目深度估计算法相比不需要人工标注数据集更方便模型训练。但是无监督单目深度估计算法也存在一些问题：从横向角度分析（与同类算法相比），当前基于无监督学习的单目深度估计算法存在深度图伪影较多、细节缺失严重；从纵向角度相比（与激光雷达传感器和深度相机等专业深度传感器相比），在动态场景下当前基于无监督学习的单目深度估计算法存在被估计的运动对象深度边缘缺失、网络模型收敛速度慢等问题。

本文针对上述无监督单目深度估计算法存在的问题，主要做出了以下贡献：

（1）横向改进方案。针对当前无监督单目视觉深度算法在估计过程中出现的边缘缺失，伪影过多、模型的泛化能力和鲁棒性较差等问题，提出一种基于增强密集连接和全局特征交互的无监督单目深度估计模型。模型由两个网络组成，深度估计网络和位姿估计网络。深度估计网络基于 Unet++网络结构，引入本文提出的增强密集连接模块和全局特征交互模块，其中增强密集连接模块加入了CBAM注意力机制，利用从不同尺度上捕捉图像的深度语义信息，以加强网络对小目标深度的捕捉；全局特征交互模块，基于自注意力机制设计，是为增强模型对不同分辨率和多纹理图像的深度估计的鲁棒性。本文方法在 KITTI 数据集和 Cityscapes 数据集上与现有的无监督单目模型相比，深度估计精度和收敛效果有较显著的提升，在误差指标和预测准确率上优于现有的算法，

均方根误差和相对均方误差相较于当前最先进的算法减少了 7%和 4%。

（2）纵向改进方案。为了减少无监督单目深度估计模型与专业深度传感器的差距，特别是智能驾驶车辆在动态场景下，单目相机深度估计过程中估计运动对象的边缘深度难以准确估计的问题，提出一种针对动态场景下的基于边缘增强的无监督单目深度估计模型。该模型由两个网络组成，深度预测网络和运动估计网络，两者均采用编码器-解码器结构。深度预测网络是基于 ResNet18 的 U-Net 结构。运动估计网络是基于 Flow-Net 的 U-Net 结构，在运动估计网络解码程序中提出了一种边缘强化解码器，其在解码过程中插入卷积注意力模块 CBAM 来加强残差平移场对运动对象的边缘特征的提取，同时引入了条状卷积模块来提高残差平移场对离散运动目标的捕捉。为了加快模型收敛，结合拉普拉斯算子，本文提出了一种新的边缘正则化。在 KITTI 数据集和 Cityscapes 数据集上与现有的动态无监督单目模型相比，深度估计精度和收敛速度均有较显著的提升，均方根误差相较于先进的 DepthMotion 算法降低了 4.8%，训练收敛速度提高了 36%。

（3）为了进一步验证横向改进方案和纵向改进方案的有效性、真实性以及分析改进方案的不足之处，利用智能驾驶测试样车对街道中多种天气交通场景进行数据采集来制作数据集，将自制数据集来进行训练和测试横向改进方案和纵向改进方案。实验结果表明横向改进方案和纵向改进方案无论在哪一种天气的交通场景下，性能都优于改进之前的算法。但是，也暴露出了一些问题在复杂天气（夜晚、阴天、雨天等恶劣天气）交通场景，现阶段无监督单目深度估计算法整体估计精度是要低于良好天气交通场景下的无监督单目深度估计算法，这也是研究人员未来需要努力改进的方向。

关键词：深度估计；计算机视觉；无监督学习；编解码器；注意力机制；智能驾驶

RESEARCH ON MONOCULAR VISUAL DEPTH ESTIMATION ALGORITHM BASED ON UNSUPERVISED LEARNING

ABSTRACT

Accurate depth acquisition in a scene is a core issue in three-dimensional environmental perception within computer vision. Considering the high cost of depth sensors for scene depth acquisition, the limited precision of traditional monocular depth estimation algorithms, and the necessity for extensive manually annotated datasets in supervised monocular depth estimation algorithms within deep learning, researchers have recently introduced unsupervised monocular depth estimation algorithms. On one hand, these algorithms offer lower deployment costs compared to specialized depth sensors, and on the other hand, they do not require manually annotated datasets, making model training more convenient compared to other monocular depth estimation algorithms. However, unsupervised monocular depth estimation algorithms also present some challenges: From a horizontal perspective (compared to similar algorithms), the current unsupervised learning-based monocular depth estimation algorithms often suffer from significant depth map artifacts and severe detail loss. From a vertical perspective (compared to specialized depth sensors like LiDAR sensors and depth cameras), in dynamic scenes, these algorithms frequently encounter issues such as missing depth edges of estimated moving objects and slow convergence speeds of the network models.

The article addresses the issues present in unsupervised monocular depth estimation and primarily makes the following contributions:

(1) This article proposes an unsupervised monocular depth estimation model based on enhanced dense connections and global feature interaction to

address issues such as edge deficiency, excessive artifacts, poor model generalization, and robustness in the current unsupervised monocular depth estimation process. The model consists of two networks: a depth estimation network and a pose estimation network. The depth estimation network is based on the Unet++ network architecture and incorporates the proposed enhanced dense connection module and global feature interaction module. The enhanced dense connection module includes a CBAM attention mechanism to capture depth semantic information from different scales, enhancing the network's ability to capture the depth of small objects. The global feature interaction module, designed based on self-attention mechanism, aims to enhance the model's robustness in depth estimation for images with different resolutions and multiple textures. Compared with existing unsupervised monocular models on the KITTI dataset and Cityscapes dataset, the proposed method shows significant improvements in depth estimation accuracy and convergence performance. It outperforms existing algorithms in error metrics and prediction accuracy, reducing the root mean square error and relative mean square error by 7% and 4% respectively compared to the current state-of-the-art algorithms.

(2) To address the challenge of accurately estimating the depth of objects' edges in dynamic scenes during the process of monocular camera depth estimation for autonomous driving vehicles, a novel unsupervised monocular depth estimation model based on edge enhancement is proposed. This model consists of two networks: a depth prediction network and a motion estimation network, both utilizing encoder-decoder structures. The depth prediction network is based on a U-Net architecture derived from ResNet18. The motion estimation network adopts a U-Net structure inspired by Flow-Net, incorporating an edge-enhanced decoder in its decoding process. This decoder incorporates a Convolutional Block Attention Module (CBAM) to enhance the extraction of edge features of moving objects by the residual translation field. Additionally, a stripe convolution module is introduced to improve the capturing of the residual translation field for discrete moving targets. To

expedite model convergence, a novel edge regularization technique leveraging the Laplacian operator is proposed. When evaluated against existing dynamic unsupervised monocular models on the KITTI dataset and Cityscapes dataset, the proposed model demonstrates significant improvements in both depth estimation accuracy and convergence speed. Specifically, compared to the state-of-the-art DepthMotion algorithm, the root mean square error is reduced by 4.8%, and the training convergence speed is increased by 36%.

（3）To further validate the effectiveness and authenticity of the horizontal improvement strategy and the vertical improvement strategy, and to analyze the shortcomings of these improvement schemes, we utilized an intelligent driving test vehicle to collect data across various weather and traffic scenarios on the streets to create a dataset. We used this self-made dataset to train and test both the horizontal and vertical improvement strategies. Experimental results indicate that regardless of the weather conditions, the performance of the new algorithms outperforms the algorithms prior to improvement. However, it also revealed some issues: in complex weather conditions such as nighttime, overcast, and rainy weather, the overall estimation accuracy of the current unsupervised monocular depth estimation algorithm is lower than that in favorable weather conditions. This highlights an area that researchers need to focus on for improvement in the future.

KEY WORDS: depth estimation, computer vision, unsupervised learning, encoder-decoder, attention mechanism, intelligent driving

目 录

摘 要 I
Abstract III
目 录 VI
第 1 章 绪 论 1
1.1 课题背景及研究意义 1
1.2 单目深度估计算法研究现状 4
1.2.1 传统的单目深度估计算法研究现状 5
1.2.2 基于有监督学习的单目深度估计算法研究现状 6
1.2.3 基于无监督学习的单目深度估计算法研究现状 7
1.3 无监督单目深度估计面临的挑战 9
1.4 论文的研究内容及结构安排 10
第 2 章 无监督单目深度估计相关工作及原理 13
2.1 相机模型和深度 13
2.1.1 相机成像模型和坐标系转换 13
2.1.2 相机测距与深度 15
2.1.3 相机的视差与深度的关系 17
2.1.4 常见的深度图的表达方式 17
2.2 图像重构原理 18
2.2.1 图像扭曲 18
2.2.2 双线性插值 20
2.2.3 图像重构 21
2.3 基于无监督学习的单目深度估计的网络结构设计 23
2.3.1 编解码结构 23
2.3.2 注意力机制 29
2.4 常用数据集与评价指标 32
2.4.1 常用数据集 32
2.4.2 常用评价指标 33
2.5 本章小结 34
第 3 章 基于增强密集连接和全局特征交互的无监督单目深度估计 35
3.1 引言 35
3.2 本章无监督单目深度估计网络结构及其原理 35
3.3 本章方法 36

3.3.1 深度预测阶段 38
3.3.1.1 增强密集连接模块 39
3.3.1.2 全局特征交互模块 41
3.3.2 位姿估计阶段 42
3.3.3 损失计算阶段 42
3.4 实验结果分析 43
3.4.1 实施细节 43
3.4.2 实验结果分析 43
3.4.3 消融实验 45
3.5 本章小结 47
第 4 章 针对动态场景下的基于边缘增强的无监督单目深度估计 49
4.1 引言 49
4.2 网络结构及其原理 50
4.2.1 深度预测阶段 50
4.2.2 运动估计阶段 51
4.2.3 图象重构阶段 52
4.3 运动网络边缘强化解码器 53
4.3.1 边缘强化解码器总体架构 53
4.3.2 CBAM 卷积注意力机制 54
4.3.3 条状卷积 56
4.4 损失函数 58
4.5 实验 60
4.5.1 实施细节 60
4.5.2 实验结果分析 61
4.5.3 消融实验 62
4.6 本章小结 64
第五章 自制数据集算法验证 66
5.1 引言 66
5.2 数据采集平台构建 66
5.2.1 传感器选择 66
5.2.2 相机的标定 67
5.2.3 计算平台 69
5.3 数据集制作及标注 70
5.4 算法测试和结果分析 70
5.4.1 参数设置 71

5.4.2 定量分析 71
5.4.3 定性分析 72
5.5 本章小结 73
第 6 章 结论与展望 74
6.1 研究成果总结 74
6.2 展望 75
参考文献 76

第 1 章 绪 论

1.1 课题背景及研究意义

人类的每一次生产力的发展都推动了社会的巨大进步，第一次工业革命是蒸汽动力的发展；第二次工业革命是电力和内燃机动力的发展；第三次工业革命是信息技术的发展；第四次工业革命是人工智能的发展。当前人们正处在第四次工业革命人工智能时代的浪潮中，生产力几何倍增长，特别是随着计算机视觉技术（Computer Vision，CV）在机器人、医药、航天、交通等行业的广泛应用。计算机视觉技术是一种通过相应的算法和相关的传感器来模拟人类视觉的过程，为后续人工智能系统的判断和执行提供重要的决策依据，计算机视觉技术的发展对人工智能有着举足轻重的意义。

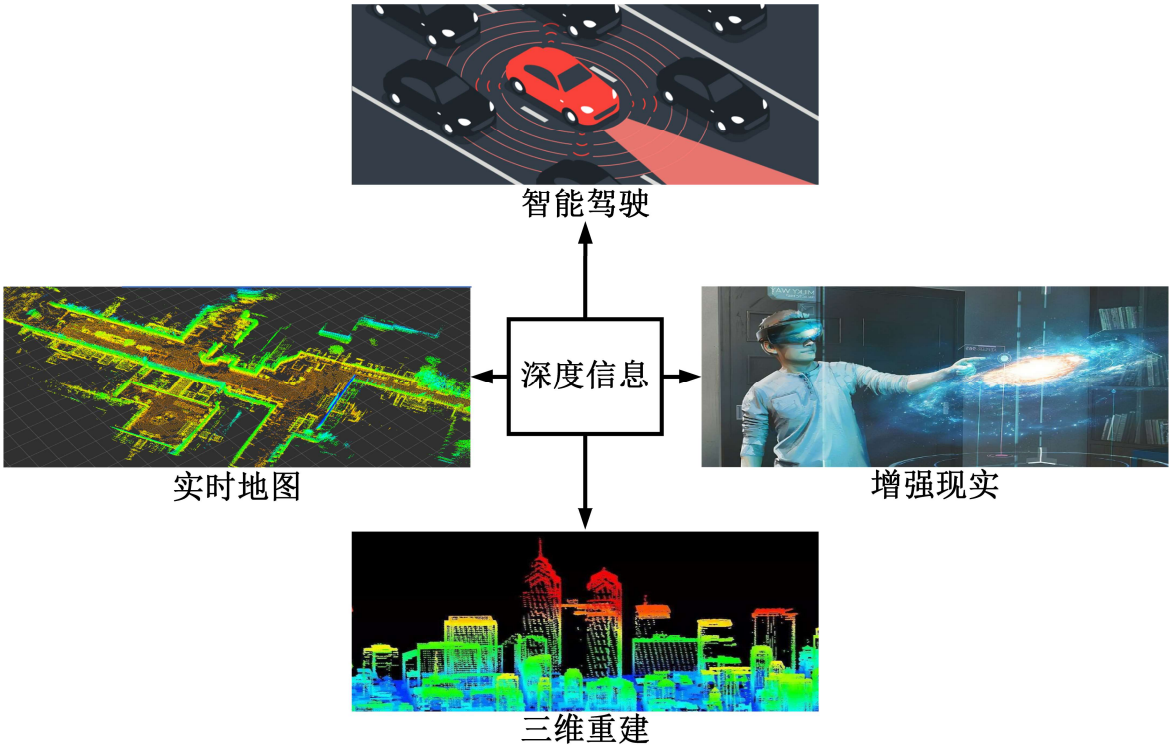


图 1-1 深度信息的应用场景

当前计算机视觉技术在目标识别、运动追踪等方面的应用都越来越成熟，比如公共场所的车牌识别、口罩识别、工业生产线上的缺陷检测、车辆的自适应巡航等。而计算机视觉在感知三维环境方面仍然没有得到大规模普及应用，精确的三维

环境信息可以辅助智能驾驶车辆主动识别前方的障碍物、让机器人实时地图构建、在增强现实中让用户沉浸体验虚拟世界和现实世界的融合、以及三维重建等其它应用。而衡量三维环境的一个重要参数即深度，深度即传感器采集的图像中每一个像素点距离传感器的距离，只有获得场景中的深度信息才能感知高精确的三维环境。如图 1-1 所示深度信息在智能驾驶、实时地图、增强现实和三维重建方面的应用。

目前获取场景中的深度的方式主要是使用一些深度传感器和双目立体视觉（两个相机）。其中深度传感器主要包括两类：深度相机和激光雷达。对于深度相机，实验室用的较多的索尼 DepthSense 系列以及 Occipital 的 Structural Sensor 系列，深度相机如图 1-2 所示。深度相机的工作原理大多数使用结构光法，即通过深度相机

图 1-2 深度相机图

图 1-3 Velodyne 激光雷达、SICK 激光雷达

的内置光源（红外光或激光）发射结构光（发射的结构光通常是网格和条纹图案）到被测对象上，当光线和被测对象相互作用时图案发生形变，通过分析这些形变来推断出被测对象各点到相机的距离，从而生成深度图像。由于深度相机使用了高频率的结构光，所以其实时性比较好；同时深度相机的软件开发工具包（SDK）较易编辑方便二次开发。深度相机的缺点也很明显，深度相机由于使用了昂贵的激光和红外发射光源其成本通常比普通的相机更昂贵甚至到达了激光雷达的成本；同时深度相机采集的深度图像需要复杂的数据处理方法来提取有用的信息，这将消耗更多的计算资源，对芯片算力要求较高。对于激光雷达，智能车辆中使用的较多的 Velodyne 的 32 线激光雷达 HDL-32E 和 SICK 的 16 线激光雷达 TM 系列，如图 1-3 所示 Velodyne 和 SICK 激光雷达。激光雷达的工作原理比较简单，通过发射激光脉冲，让发射的激光脉冲作用于被测对象表面，通过返回时间来计算物体到传感器之间的距离。激光雷达能够提供非常高的测量精度同时可以跟踪多个目标。但是激光雷达的制造成本较高且难以维护；同时激光雷达生成的点云数据通常比较稀疏需要进行复杂的数据处理和算法分析，占用芯片较多的计算资源。双目立体视觉，使用左右两个单目相机来模拟人的两只眼睛，产生深度。左右相机同时捕获同一场景的图像时，由于它们的位置不同，因此在两个图像中同一位置会有所偏移，这种位置

偏移即视差。结合相机的基线长度和焦距等参数，使用三角测量法计算出每个像素点对应物体表面到相机的距离（深度），双目相机如图 1-4 所示。双目立体视觉深度估计优点是深度估计原理简单芯片处理速度快，故双目立体视觉深度估计具有较高的实时性，可以在较短时间内完成深度估计；同时双目立体视觉可以方便调整摄像头的参数和位置，使得双目立体视觉深度估计比较灵活满足不同场景。但是双目立体视觉深度估计视差计算中匹配过程中会受到遮挡、光照变化等因素或背景复杂等因素影响，导致匹配错误；同时双目立体视觉要求两个相机进行精准的同步，这增加了硬件配置和系统部署的复杂性。

图 1-4 双目相机

随着机器学习和深度学习技术在计算机视觉领域的应用，研究人员为了摒弃深度相机和激光雷达成本较高难以维护的缺点、双目立体视觉中两个相机难以同步的缺点，通过深度学习技术将单目深度估计作为研究重点，并将模型初步部署到了产品中，比如国外的 Tesla 电车的 Tesla AI 中纯视觉方案中的单目深度估计模型 RegNet，如图 1-5 所示 RegNet 模型的效果图；在国内独角兽公司 Momenta 的智驾方

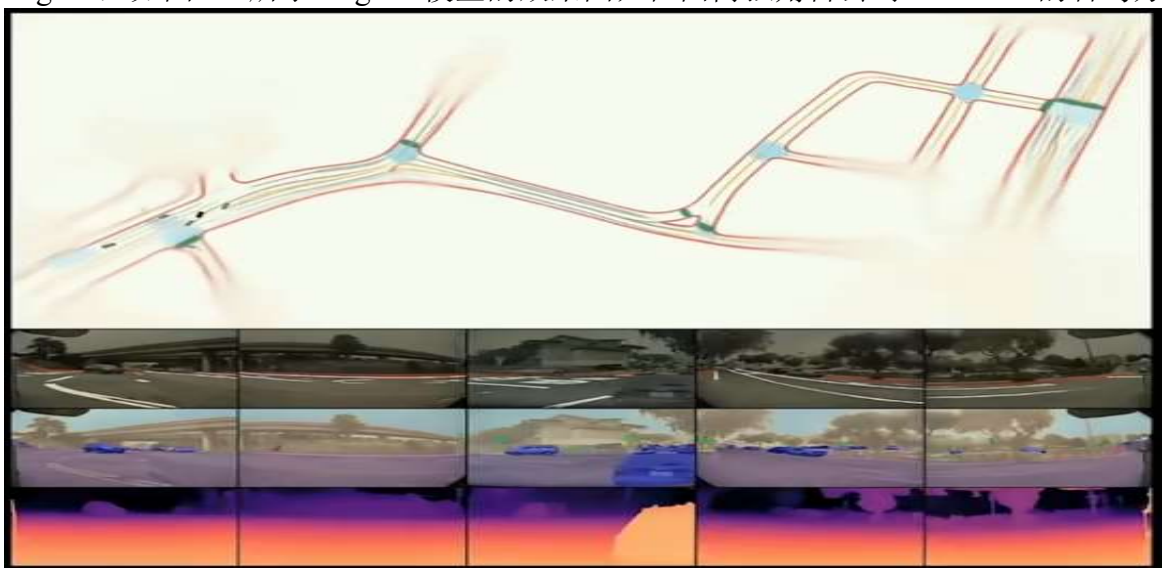


图 1-5 Tesla RegNet 效果图

案中也有自己的单目深度估计方案，专注重卡领域的图森科技和智加科技等也采用了纯视觉方案的单目深度估计，甚至在日常的手机照相中为了体现出拍摄对象的深度信息也采用了单目深度估计模型如 OPPO 手机的 MonoIndoor 模型[1]，如图 1-6 所

示 MonoIndoor 网络结构。上述应用在一定程度上标志着单目深度估计技术的进步但是大部分情况下还是辅助人类工作并不能完全替代人类工作，研究人员仍然还有很长一段路要走。应用深度学习技术通过单目相机进行深度估计面临着数据集获取困难、训练模型成本高等问题。早期人们选择使用有监督学习模型进行单目深度估计，模型在训练过程中需要使用大量人工标注好深度真值（ground truth depth）的数据集，以深度真值作为监督信号，导致数据集比较稀少，以及人工标注过程中成本高昂等缺点，同时由于使用的数据集有限模型的泛化能力和精度也较差。为了克服监有督学习的缺点，科研人员利用图像重构和极几何约束等知识，以重构的图像作为主要监督信号来监督网络的训练，提出了无监督学习的单目深度估计模型。基于无监督学习的单目深度估计模型，不需要数据集提供的真实深度值作为监督信号，节省了大量人工标注的成本而且使用训练的数据集也更易采集，泛化性能和精度也更高。

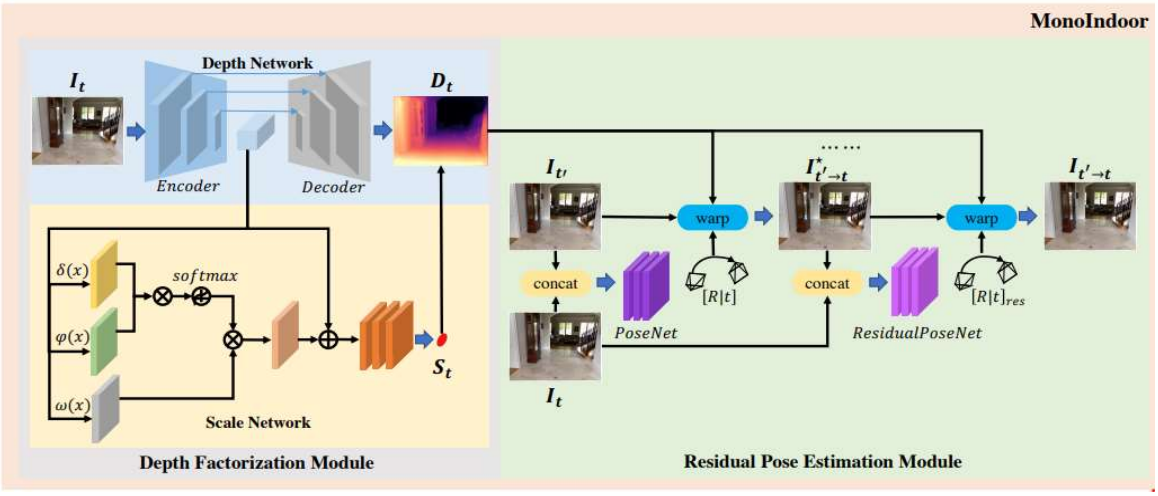


图 1-6 MonoIndoor 网络结构

本文依托课题《特定场景无人通用化全线控底盘开发与产业化应用》和课题《基于多模态融合的智能车辆多任务感知技术研究》，针对智能车辆在驾驶过程中使用纯视觉方案的单目深度估计算法进行重点研究，无监督单目深度估计算法是本文的主要研究内容。

1.2 单目深度估计算法研究现状

单目深度估计是用简单的二维 RGB 图像去推断整个三维环境，相当于从二维空间到三维空间的一个映射，在数学上这种问题大多数是欠定问题，因此在计算机视

觉中单目深度也是一项非常具有挑战的难题。根据数字图像处理技术、机器学习技术和深度学习技术人们将单目深度估计算法大致划分为三类：传统的单目深度估计算法、基于有监督学习的单目深度估计算法和基于无监督学习的单目深度估计算法。

1.2.1 传统的单目深度估计算法研究现状

传统的单目深度估计算法主要使用了两种思路：将深度估计问题转化为随机场中的随机变量思想和恢复三维结构思想。

将深度估计问题转化为随机场中的随机变量思想的主要为基于马尔科夫随机场（Markov Random Field，MRF）的深度估计方法。最先提出来使用马尔科夫随机场的是 Chaudhuri[2]，将深度图以及场景的原始图像被建模为具有平滑先验的马尔可夫随机场，并通过使用模拟退火最小化合适的能量函数来获得。由于 Chaudhuri[2]提出的使用马尔科夫随机场进行深度估计是基于局部一致性假设的，这就导致了局部相邻像素之间的深度比较相似，使得局部预测的深度值偏差比较大。随后 Saxena[3]对 Chaudhuri[2]进行改进，将马尔科夫随机场结合了多尺度局部和全局图像特征，并对各个点的深度以及不同点的深度之间的特征关系进行了建模，提高了局部预测的深度值精度。虽然 Saxena[3]对 Chaudhuri[2]的局部深度提高是有效的但是效果仍然不是非常明显，Sun[4]对 Saxena[3]提出的方法进一步改变。Sun[4]首先将被预测平面切割成很多小平面补丁，然后使用马尔可夫随机场来推断每一组“平面参数”，捕获补丁的三维空间位置，通过马尔科夫随机场对图像深度线索以及图像不同部分之间的关系进行建模，最终能捕获比 Saxena[3]更详细的局部深度。

利用恢复三维结构思想来进行深度估计主要的方法是运动恢复结构（Structure From Motion，SFM)和阴影中恢复形状(Shape From Shading，SFS)。SFM 旨在通过连续帧图像恢复出场景中每一像素点的深度以及相机的运动参数。早期经典的 SFM 算法是 Soatto[5]提出：首先使用 SIFT 算法提取关键特征点，然后通过特征描述符来匹配图像中的对应特征点；根据匹配的特征点使用八点法算出相机的旋转和平移，利用旋转平移等参数和三角测量求出三维点深度；最后通过计算重投影误差来优化相机的旋转参数和平移参数来提高三维环境中的深度。由于 Soatto[5]提出的算法对于无序图像不能处理多个实例中的同一结构场景，当包含不同实例的图像对进行视觉相似性进行匹配时，成对的几何关系以这样的方法推断出的对应关系是错误的。

Robert[6]对 Soatto[5]的算法提出了改进，提出了联合推断姿势和匹配概率期望最大化算法，使用匹配概率期望最大化替代最小二乘法，减少在匹配过程中将不同实列的图像错误匹配的问题。SFS 算法是另一种利用恢复三维结构思想来进行深度估计算法，SFS 旨在根据一张物体表面的灰度图像的亮度来计算该表面的三维形状。最早的根据灰度图像的亮度来估计三维环境的深度算法是 Horn[7]提出的：首先进行光学方程建模对物体表面的反射和散射过程进行分析；根据建立好的光学方程进行图像的像素值分析，建立光度约束，得到法线和光照之间的联系；基于建立的光度约束，利用 bundle adjustment[8]优化方法来估计图像中每一个点对应的表面法向量；最后利用得到的表面法线使用表面积分来求解每一个像素的深度。

上述经典方法在一定程度上突破了深度估计只能用深度传感器和多相机的限制，但是也存在很多不足。使用马尔科夫随机场进行单目深度估计，不仅仅要面对局部一致性假设不适用的问题，而且当场景中出现遮挡和透视变化的时候，马尔科夫随机场会让随机变量的分布变得不均匀导致在深度估计过程中深度值精度比较差。同时由于马尔科夫随机场的模型较大导致模型包含的参数较多如邻域大小、平滑权重参数等，这些参数需要进行大量的人工验证才能得到最优效果。SFM 算法由于模型使用对极几何作为几何分析基础。而对极几何的假设前提一般多是静态的，所以在一些动态场景下单目深度估计的精度比较差。SFS 算法由于本身深度估计理论依赖于灰度图像中较好光照条件和阴影信息，当光照条件较差和阴影不均匀时，会导致模型的深度估计性能严重下降。

1.2.2 基于有监督学习的单目深度估计算法研究现状

在深度学习中，有监督学习是一种使用比较多的深度学习方法，其基本思想是将从带有详细标签的数据中学习输入和输出之间的映射关系。对于有监督学习过程，通常输入的数据和对应输出的标签一起组成了训练的数据集，且将标签的真值作为唯一的监督信号，最终的目的是让模型根据输入的数据输出正确的标签。在深度学习中常见的有监督学习模型包括深度神经网络（DNN）、循环神经网络（RNN）、卷积神经网络（CNN）。在有监督单目深度估计中，使用最多的是基于卷积神经网络（CNN）的单目深度估计。Eigen[9]开辟了基于有监督学习的单目深度估计算法先河，首先使用包含 5 个卷积层和池化层的粗尺度网络，然后再用两个全连接层上采样输出精细的深度图。由于 Eigen[9]等人首次使用的全尺度卷积神经网络输

出深度图像很多局部细节出现了缺少，Eigen[10]受到多尺度卷积神经网络的启发对原先的模型进行了改进提出了基于多尺度卷积架构的深度估计模型，将粗网络的特征与单层卷积和池化的特征连接起来同时将连接后的特征进行与原图像特征拼接，以达到精细深度图的目的。为了解决有监督学习模型在训练回归网络时出现的模型预测的深度图分辨率较低的问题，Fu[11]等人在卷积神经网络训练过程中增加了离散化策略通过离散深度来抑制住在较大深度值出现误差较大以及降低相邻像素点之间深度的影响，使得模型能够完成高分辨率深度图的预测任务。

有监督学习方式摆脱了传统方法需要大量几何推导的烦琐，同时相对于传统算法在精度有一定的提高，但是也暴露出一些问题：由于训练的数据集需要人工进行大量标注导致训练成本过高，又因为有监督模学习网络模型都是以回归问题作为理论基础使得网络模型存在过拟合现象，预测出的深度图像出现局部性和稀疏性。

1.2.3 基于无监督学习的单目深度估计算法研究现状

在深度学习中，无监督学习是另一种使用比较多深度学习方法，其目标是从未标记的数据中学习出数据的内在结构或特征表示，而无需使用标签信息。相比较于有监督学习，无监督学习不需要人工标记大量数据，因此在许多情况下更具实用性。无监督学习方法在深度估计算法中也得到了大量应用，目前主流的无监督单目深度估计方法主要可以分为三类：1）基于光流和几何约束的深度估计；2）基于生成对抗网络的深度估计；3）基于多尺度和自监督学习（自监学习方法是无监督学习的一个子类）的深度估计。

基于光流和几何约束的深度估计方法主要有 Mahjourian[12]和 Dosovitskiy[13]等，Mahjourian[12]通过增强三维空间的几何约束来达到场景的深度和相机自我运动的一致性，其直接通过模型预测出的两张相邻帧的深度图（深度图可以转换为三维点云图）对应点的点云深度信息，并利用相邻帧之间的点云信息和相机自运动信息重构两张新的目标图像(本文没有特殊说明的图像均为 RGB 图像)，再使用几何一致性作为监督信号让模型进行学习。Dosovitskiy[13]首先对两个输入图像建立两个独立但相同的处理光流，再通过一系列的卷积操作分别提取两个图像的特征表示，随后在更高的卷积层上将这些特征表示组合到一起进行后续的下采样生成深度图像。基于生成对抗网络的深度估计方法主要有，Aleotti[14]和 Pilzer[15]等，其中 Pilzer[15]是提出生成对抗网络方法来进行单目深度估计的第一个研究者。Pilzer[15]利用两个子生成网

络来构建两个视图之间的对应场（即两个视图相对应的视差图），然后再利用两个子网络之间的对抗学习来重构视差图（视差图和深度图可以相互转换），最终形成一个闭环监督结构来进行相互约束和监督。Aleotti[14]利用生成器网络和鉴别器网络分别生成的真实图像和虚假图像来对抗学习重构视差图，完成深度估计。基于多尺度和自监督方法的单目深度估计与前面提到的两种方法相比是目前精度最高的一类方法，主要包括 Garg[16]、Goard[17]、和 Zhou[18]以及近年来提出的 Sun[19]、Guizilini[20]、Chen[21]，其中 Zhou[18]提出的 SfmLearner 和 Goard[17]提出的 Monodepth2 方法最经典。Zhou[18]的思路是将整个网络模型设置为两个子网络即深度估计网络和位姿估计网络，在训练的过程中将多张连续帧图像作为输入信号（输入信号包含目标图像和源图像），目标图像是用来通过深度预测网络得到相应的深度图，目标图像和多张源图像作为位姿估计网络的输入得到相应的位姿信息（位姿信息包括连续帧图像间的旋转和平移），利用得到的深度图和位姿信息反向扭曲重构目标图像，最后将原目标图像与重构的目标图像进行光度损失和平滑损失计算，构成整个网络的自监督信号。Goard[17]的整体网络架构与 Zhou[18]相似，不同的是弥补了 Zhou[18]的网络对一些运动物体和遮挡物体估计出的深度图边缘缺失等问题。Goard[17]利用一种固定像素自动掩膜来捕捉移动目标的深度，应用最小重投影损失函数来避免遮挡现象，同时在深度图像的解码阶段将不同的尺度的深度图都上采样到输入图像的分辨率大小，为避免估计出来的深度图纹理和边缘丢失将上采样到输入图像分辨率的大小的深度图也进行损失计算。最新的一些方法仍旧使用了 Monodepth2 的主体思想，其中 Sun[19]等人提出了 SC-DepthV3 网络模型，与 Monodepth2 方法不同的是多了预训练深度估计网络，使用预训练深度估计网络来生成一张伪深度图，使用伪深度图作为辅助监督信号来监督深度图的生成。Guizilini[20]提出了 ZeroDepth 网络模型，与 Monodpth2 方法 Unet 结构直接编解码不同的是使用了嵌入级几何先验，通过单帧信息来解耦解码器和编码器生成深度图，同时使用了交叉注意力机制来加强对弱纹理区域的提取。为了解决深度图像的边缘过于粗糙，Chen[21]对 Monodepth2 网络的损失函数进行了改进，提出了一种基于补丁的三元组（三元组：中心像素作为锚元组；与锚共享相同语义类的像素作为阳性元组；而与锚具有不同语义类的像素作为阴性元组）损失，加强了单目深度图像的边缘精度。

无监督单目深度估计算法，从本质上解决了有监督单目深度估计需要人工大量标注好的数据集的问题，让单目深度估计适用于更多的场景而不是仅仅局限于标注

好的数据集提供的单一场景。同时，对于一些复杂的遮挡场景和动态场景的深度估计也提供了明确的思路，总体上无监督单目深度估计算法是优于先前的有监督单目深度估计算法的。

1.3 无监督单目深度估计面临的挑战

虽然上述无监督单目深度估计方法取得了重大突破，但是也存在明显不足，面临很多挑战。本节从横向和纵向两个角度分析：

横向角度分析。从无监督单目深度估计算法自身角度分析吗，存在以下 3 点缺陷：1）利用光流和几何约束进行无监督单目深度估计是基于对被估计对象以低速或者静止的运动状态，这样才能保证相邻像素之间的空间连续性，由于被估计对象很多是高速运动的状态特别是一些智能驾驶的场景，导致被估计的车辆深度不够准确特别是对小目标深度的提取。2）利用生成对抗网络进行无监督单目深度估计，相比较 CNN（卷积神经网络）的训练过程 GAN（生成对抗网络）是一种动态训练，动态调整生成器的参数，导致训练时间较长，从而限制了模型的迭代和调试，模型在面对噪声（如估计一些多纹理对象的深度）时产生不稳定的深度估计结果。3）在基于多尺度和自监督学习（自监督学习也是无监督学习的一个子类）方式的无监督单目深度估计过程中，RGB 图像不明显的纹理区域会出现训练梯度消失现象，导致估计出来的图像边缘等细节出现缺失的现象；使用多尺度方法预测的度深度图，在计算损失过程容易忽略深度图的全局与局部的联系导致预测出来的深度图伪影缺较多，缺乏局部细节。

纵向角度分析。以专业深度传感器为基准分析，经过大量测试，发现大多数传统无监督单目深度估计方法中存在以下 3 点明显的缺陷：1）由于运动对象自身的轨迹是不断变化的，导致运动对象的纹理等特征难以被神经网络有效捕捉，所预测的深度图会出现很多伪影和深度边缘模糊等现象，这与激光雷达可以实时读取运动对象的深度轮廓有很大差距；2）在解码器解码过程中大多使用比较常见的方形卷积核，导致在对预测对象上采样过程中边缘附近处的帧没有被充分计算，编码器对浅层特征的细节信息并没有充分利用和部分离散运动对象没有被捕捉，最终预测出的深度图边缘产生缺失等现象，与鱼眼深度相机可以进行 220° 的深度估计并清晰捕捉被预测对象的边缘相比有一定的差距；3）由于网络整体过于复杂使得网络收敛较慢，与双目或者多相机深度估计的相应时间相比较慢且鲁棒性较差。

故无论是从无监督算法自身分析还是与专业传感器为基准分析无监督单目深度估计算法仍存在一些缺陷，这正是本文研究无监督单目深度估计算法的意义。

1.4 论文的研究内容及结构安排

基于无监督单目深度估计算法自身的缺陷和与专业深度传感器为基准分析产生的差距，本文从以下两个角度进行改进：

横向改进。从无监督单目深度估计算法网络结构本身改进。受到 Goard[17]等人的启发，提出一种基于增强密集连接和全局特征交互的无监督单目深度估计模型。本文提出了两点创新，以解决无监督单目深度估计算法自身需要改进的缺陷：1）增强密集连接结构中将每一个增强密集连接模块加入了 CBAM 注意力机制，该结构可让网络更好地利用不同层采样的尺度特征，捕捉不同尺度上图像的深度语义信息，增加网络的感受野，加强对小目标深度的捕捉。增强密集连接结构同时有助于缓解学习过程中梯度消失问题，由于每一层都连接到先前的所有特征层，让梯度可以更容易地通过跨越多个特征层传播，从而帮助网络训练程中克服梯度消失的现象，加强对弱纹理区域的深度估计减少深度图出现的深度边缘模糊等现象。2）全局特征交互结构是本文提出的另一个改进方法，该方法基于 Alaaeldin[22]提出的自注意力机制，本文引入了“查询-键值对”方式来计算沿着特征通道的注意力，这样可使注意力的权重更均匀，以保证模型预测出的不同的分辨率和多纹理图像的深度具有更强的鲁棒性。

纵向改进。从与专业深度传感器的差距进行改进。受到 Zhou[18]等人的启发，本文提出针对动态场景下的基于边缘增强的无监督单目深度估计模型。本文提出了三点创新来弥补无监督单目深度估计与专业深度传感器的差距：1）在运动估计网络中添加卷积注意力模块 CBAM[23]，CBAM 注意力机制可以帮助网络自动调整对运动对象的关注度，减少对运动背景或其他干扰因素的敏感性，从而使被估计对象的深度图边缘更加清晰。2）在运动估计网络解码阶段引入了一种新的卷积方法即通过将传统的方形卷积核替换为条状卷积核，使得运动网络更加有效地提取边界和全局信息以及捕捉到离散的运动对象。3）在损失函数里结合高斯拉普拉斯算子提出边缘正则化，以加快网络收敛。

论文共分为 6 章，各章安排如下：

第一章 绪论。首先，分析了人工智能时代背景下计算机视觉领域在各行各业的

主要应用及其重要意义。其次，分析了计算机视觉领域在三维环境感知方面的不足尤其是对三维环境中深度感知的不足。然后，分析了深度估计在智能驾驶和其它工业领域的重要意义，以及当前获取深度的主要方法。最后，介绍了当前无监督单目深度估计算法的一些优缺点和挑战，针对这些缺点和挑战提出了基于无监督学习的单目深度估计算法的一些主要改进方式。

第二章 无监督单目深度估计相关工作和原理。首先，介绍了相机模型的建立以及相机获取场景深度的一些基本方法。其次，介绍了相机的视差和深度的转换关系以及目前一些主要深度图表达方法。再次，介绍了图像的扭曲原理作为后续无监督单目深度估计监督信号的来源。然后，介绍了无监督单目深度估计网络结构的设计，尤其是网络的编解码结构和注意力的结构的设计。最后，介绍了无监督单目深度估计中常用的数据集和评价指标。

第三章 基于增强密集连接和全局特征交互的无监督单目深度估计。针对无监督单目深度估计算法自身出现的边缘缺失、伪影过多以及小目标对象深度估计不准确问题和鲁棒性差的问题，本章提出了 2 点改进。1）通过无监督单目深度估计将网络的编解码器进行改进，提出了增强密集连接编解码器并添加了注意力机制加强对小目标和深度边缘的捕捉；2）在解码器中添加全局特征交互模块增强模型对不同分辨率和多纹理图像的深度估计的鲁棒性。最后与现阶段先进的算法相比该算法精度有一定的提高。

第四章 针对动态场景下的基于边缘增强的无监督单目深度估计。针对与激光雷达和深度相机等传感器相比，无监督单目深度估计在动态场景下移动对象出现的边缘缺失鲁棒性较差等问题，本章提出了 3 点改进。1）在运动网络插入注意力机制，帮助网络自动调整对运动对象的关注度，减少对运动背景或其他干扰因素的敏感性 2）在运动网络解码阶段添加条状卷积模块来加强对运动对象的边缘特征的提取。3）结合拉普拉斯算子提出新的边缘正则化，来提高模型的鲁棒性和收敛速度。最后与现阶段先进的算法相比该算法在动态场景下的深度估计性能有一定的提高。

第五章 为了进一步验证第三章和第四章提出的算法有效性，本章利用智能驾驶测试样车对公路及街道多种天气交通场景进行数据采集来制作数据集。将自制数据集进行训练和测试第三章和第四章提出算法，根据其表现的优缺点来指导未来的工作。

第六章 总结与展望。总结了本文对无监督单目深度估计算法的一些贡献，分析

了本文的创新性和局限性，结合当前研究现状提出了进一步改进方向。

第 2 章 无监督单目深度估计相关工作及原理

2.1 相机模型和深度

2.1.1 相机成像模型和坐标系转换

相机作为一种通过光学系统和感光元器件将外部场景的光线转换成图像的设备，在社会各行各业应用特别广泛，特别是计算机视觉中。相机最基本（不考虑鱼眼相机等其它特殊相机）的成像模型都是以小孔成像（针孔模型）为基础。如图 2-1 所示，中学物理蜡烛小孔成像实验：将点燃的蜡烛和投影屏放在带有小孔的隔板中间，可以看到蜡烛的像投影到投影屏上。

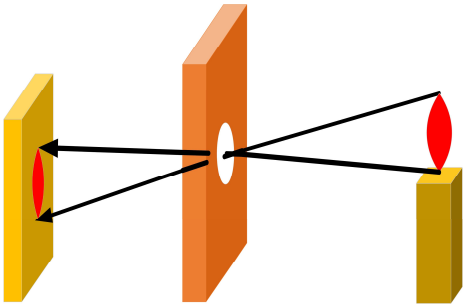


图 2-1 蜡烛小孔成像实验

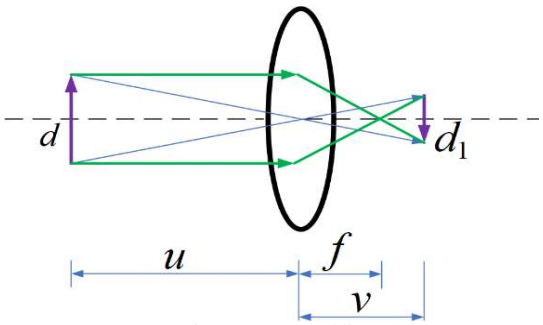


图 2-2 相机光学几何模型

相机的成像原理与小孔成像的模型类似，唯一不同的是，隔板的小孔变成了相机的透镜（镜头），以及投影屏幕变成了胶片或者数字传感器。相机的物距，焦距，和相距存在以下关几何关系如图2-2所示。其中u为物距，f为焦距，v为相距，d为源像的高度，d_1为像的高度。根据相似几何关系物距u、焦距f、相距v三者满足推导关系式：

$$\begin{cases}\frac{u}{d}=\frac{v}{d_1}\\ \frac{f}{d}=\frac{v-f}{d_1}\end{cases}\Rightarrow\frac{u}{v}=\frac{f}{v-f}=\frac{d}{d_1}\Rightarrow uv-uf=vf\Rightarrow\frac{uv}{u+v}=f\Rightarrow\frac{1}{f}=\frac{u}{uv}+\frac{v}{uv}=\frac{1}{u}+\frac{1}{v}\quad(2\text{-}1)$$

为了进一步建模分析出三维环境中的每一个点与像素点之间的关系，本文在相机模型中建立了四个坐标系：世界坐标系、相机坐标系、图像坐标系和像素坐标系。世界坐标系（World Coordinate System）：用于定位三维环境中物体的具体位置的坐标系，通常以三维环境中某个固定的参考点为原点，以固定的坐标轴方向表示物体的位置和朝向。相机坐标系（Camera Coordinate System）：用来描述相机的方向和位置的坐标，通常以相机的光学中心为相机坐标系的原点，以相机的光轴方向为相坐标系的 Z 轴，以相机的水平和垂直方向为相机坐标系的 X 轴和 Z 轴。图像坐标

系（Image Coordinate System）：用来描述相机捕获的目标映射到二维平面上的坐标系，图像坐标系的原点通常是以图像的左上角为原点，X 轴方向为水平向右，Y 轴方向为垂直向下。像素坐标系（Pixel Coordinate System）：像素坐标系是图像坐标系的一种特例，用来表示图像上每个像素的坐标，使用整数坐标来表示，通常以图像的左上角为像素坐标系的原点，X 轴方向为水平向右，Y 轴方向为垂直向下。为了更详细的理解这几个坐标系，参考图 2-3，其中$O_1-x-y-z$为图像坐标系、$O_2-X_c-Y_c-Z_c$为相机坐标系、O_3-u-v为像素坐标系，由于世界坐标系中的原点可以是任意位置，故没有表示。图 2-3 中，对于空间中的一点P，$P(X_w,Y_w,Z_w)$为世界坐标系中的对应点，$P(u,v)$为像素坐标系中的对应点。为了方便科学计算，研究人员一般将世界坐标中的点转换为像素坐标系中的点。空间中的世界坐标系中的点到像素坐标系中的点依次存在以下 3 步转换关系：

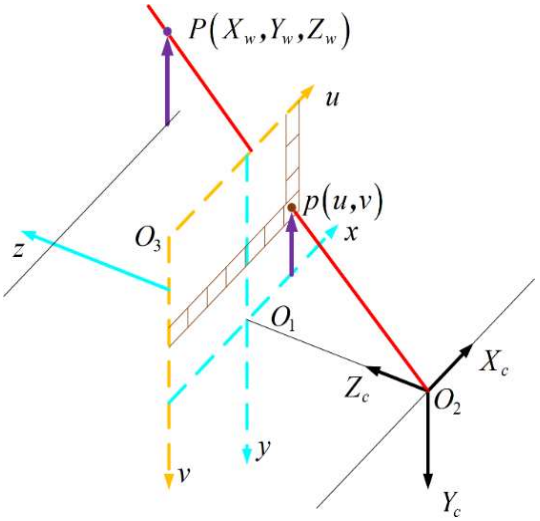


图 2-3 相机模型坐标系之间的转换

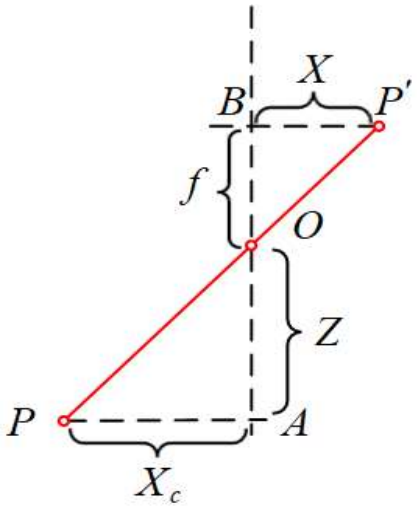


图 2-4 图像坐标系中的相似关系

（1）从世界坐标系到相机坐标系，设相机坐标系中的对应点为$P(X_c,Y_c,Z_c)$，根据仿射变换的知识有：

$$\begin{bmatrix}X_c\\Y_c\\Z_c\end{bmatrix}=R\begin{bmatrix}X_w\\Y_w\\Z_w\end{bmatrix}+T\Rightarrow\begin{bmatrix}X_c\\y_c\\Z_c\\1\end{bmatrix}=\begin{bmatrix}R&T\\0_3^T&1\end{bmatrix}\begin{bmatrix}X_w\\Y_w\\Z_w\\1\end{bmatrix}\qquad(2-2)$$

其中，$\boldsymbol{R}$表示相机的旋转变换，$\boldsymbol{T}$表示相机的平移变换。

（2）从相机坐标系到图像坐标系，设图像坐标系中的对应点为$P(X,Y,Z)$，如图 2-4 所示由相似三角形几何关系有，其中f表示焦距：

$$\frac{Z}{f}=-\frac{X_c}{X}=-\frac{Y_c}{Y}\qquad(2-3)$$

化简并整理后有：

$$\begin{cases} X = f\dfrac{X_c}{Z_c} \\ Y = f\dfrac{Y_c}{Z_c} \end{cases} \tag{2-4}$$

根据式 2-4 本文可以观察得到在图像投影的过程中，深度信息Z_c丢失，这也是人们进行深度估计算法研究的原因。

（3）从图像坐标系到像素坐标系，由于像素坐标系和图像坐标系相交与二维平面，从图像坐标系到像素坐标系只存在缩放和平移的关系，故图像坐标系中的点和像素坐标系中的点存在以下关系：

$$\begin{cases} u = \alpha X + C_x \\ v = \beta Y + C_y \end{cases} \tag{2-5}$$

其中α、β为缩放因子，C_x、C_y为平移因子；设：$\alpha f = f_x$，$\beta f = f_y$联立代入式 2-5 有：

$$\begin{cases} u = f_x\dfrac{X_c}{Z_c} + C_x \\ v = f_y\dfrac{Y_c}{Z_c} + C_y \end{cases} \tag{2-6}$$

将上式转化为其次坐标形式有：

$$Z_c\begin{pmatrix} u \\ v \\ 1 \end{pmatrix} = \begin{pmatrix} f_x & 0 & C_x \\ 0 & f_y & C_y \\ 0 & 0 & 1 \end{pmatrix}\begin{pmatrix} X_c \\ Y_c \\ Z_c \end{pmatrix} \tag{2-7}$$

联立 2-7 式、2-4 式、2-2 式，最终可以得到世界坐标系与像素坐标系的关系：

$$Z_c\begin{pmatrix} u \\ v \\ 1 \end{pmatrix} = \begin{pmatrix} \dfrac{f_x}{Z_c} & 0 & u & 0 \\ 0 & \dfrac{f_y}{Z_c} & v & 0 \\ 0 & 0 & 1 & 0 \end{pmatrix}\begin{pmatrix} \boldsymbol{R} & \boldsymbol{T} \\ 0^T & 1 \end{pmatrix}\begin{pmatrix} X_w \\ Y_w \\ Z_w \\ 1 \end{pmatrix} \tag{2-8}$$

在式2-8的右边，左起第一个矩阵称为相机的内参$\boldsymbol{K}$（大都数相机的内存参出厂已经标定好由相机本身决定）；左起的第二个矩阵中参数$\boldsymbol{R}$和$\boldsymbol{T}$称为相机的外参，由相机的自身的运动状态决定，其中$\boldsymbol{R}$表示相机的旋转矩阵，$\boldsymbol{T}$表示相机的平移矩阵。

2.1.2 相机测距与深度

相机作为人类生产生活中最为常见的光学传感器，相机不仅仅可以用来记录不同场景中的事物和事件，相机也可以用于科学推理比如场景中进行估计距离，又称为相机测距。相机测距即用相机采集的图像信息来计算目标物体与相机之间的距离

的方法。通常相机的测距方法有三角测量和利用视差来实现，视差测距多数利用双目或者多目相机来实现的，本文仅仅介绍基于三角测距的深度估计原理：

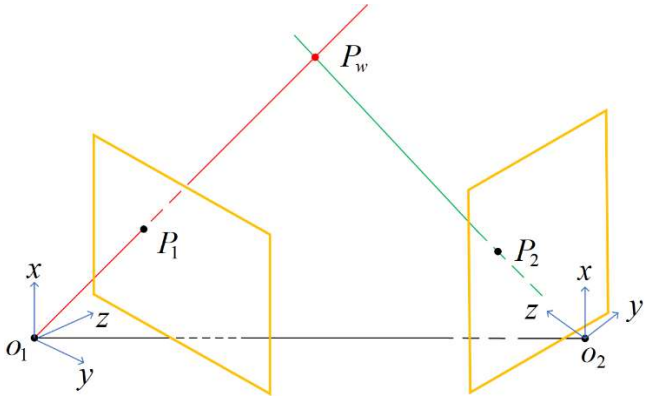


图 2-5 三角测量原理

三角测量法最初是由高斯针对不同季节观察到的星星的角度提出的，来估计星星与地球之间的距离。在基于无监督学习的单目深度估计过程中，使用连续帧图像来估计像素深度就是基于三角测量原理。如图 2-5 所示，点P_w是三维空间中的点，在两个相机坐标系$o_1-x-y-z$、$o_2-x-y-z$（在无监督单目深度估计中由于仅仅使用一个单目相机故坐标系两个位置内参相同，只是移动的位置不同）中的投影点分别是点P_1和P_2。由式 2-8 可以得到：

$$Z_1P_1 = K_1P_w \tag{2-9}$$

$$Z_2P_2 = K_2(RP_w + t) \tag{2-10}$$

其中，Z_1和Z_2表示空间点到相机光心的距离即深度（也是相机坐标系的z轴坐标），K表示相机的内参，R和t表示相机的外参。为了方便计算深度，定义$x_1 = K_1^{-1}P_1$、$x_2 = K_2^{-1}P_2$（由于使用的是同一个相机，故两个相机的内参相同即$K_1 = K_2$），代入 1 式有：

$$Z_2x_2 = Z_1Rx_1 + t \tag{2-11}$$

对于式 2-11，其中相机外参R和t可以通过本质矩阵分解得到，唯一不知道的是深度Z_2和深度Z_1，首先求Z_1，将上式两边同时左乘x_2^{-1}有：

$$Z_2x_2^{-1}x_2 = 0 = Z_1x_2^{-1}Rx_1 + x_2^{-1}t \tag{2-12}$$

等式右边可以看成是一个方程，只有Z_1是未知的，可以求出Z_1同理可以求出Z_2，这就是整个相机的三角测量原理。

在无监督单目深度估计过程中，研究人员的主要工作是通过神经网络预测出连续帧图像之间的外参R和T，而不是经过复杂的三角测距中本质矩阵分解来完成，这也是无监督单目深度估计算法与传统深度深度估计算法的本质区别。

2.1.3 相机的视差与深度的关系

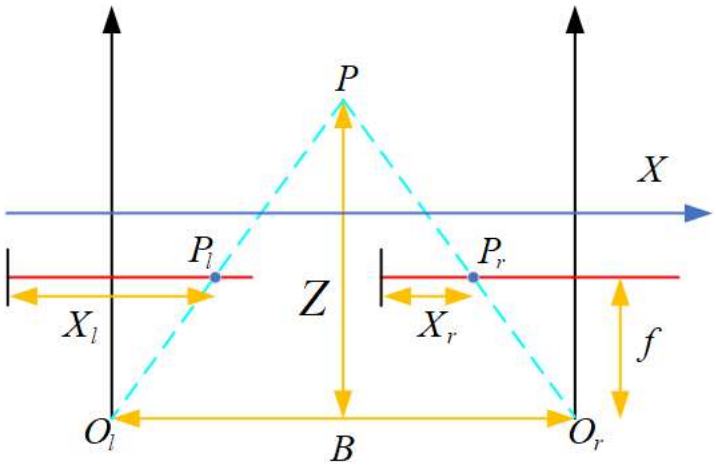


图 2-6 视差于深度之间的几何关系

在神经网络进行深度估计的过程中，为了方便相机进行推理深度和重构图片，深度一般不会直接获取而是先获取视差（disparity），让视差再转化为深度。视差即两个（多个）不同的相机或者是一个相机在运动过程中从不同角度获得空间中三维像素点位置的差异，所有的像素点的差异就一起构成了视差图。视差与深度的关系如图 2-6 所示，设空间中一点P在，相机O_l和O_r上的投影点为P_l和P_r，两条红线代表成像平面，以成像左边界为X方向的原点，B 表示相机O_l和O_r之间的距离又称为基线距离，投影点P_l和P_r在X方向的坐标分别是X_l和X_r，Z表示P点的深度，为方便推导将两相机O_l和O_r的焦距都设为f。由几何关系得到，视差为$X_l - X_r$，令$s = X_l - X_r$；P_l和P_r之间的距离为$B - s$，令$b = B - s$。根据相似几何关系有：

$$\frac{b}{B} = \frac{Z - f}{Z} \tag{2-13}$$

$$\frac{B - s}{B} = \frac{Z - f}{Z} \tag{2-14}$$

由式 2-14 可以得到：$Z = f\frac{B}{s}$

由于基线距离B和焦距f都是常数，所以可以得到深度与视差存在反比的关系，这也是为什么在日常生活中人们观察到的景物的距离越远（深度值越大）视差越小的原因，视差和深度的转换是后文图像重构的关键。

2.1.4 常见的深度图的表达方式

有了视差人们就可以得到想要的深度图，常见的深度图可以有三种表达方式：（1）灰度图像，灰度图像由于每一个像素只需一个灰度值（0~255）即可表示，不

需要复杂的颜色映射编码，对不同平台的设备平台兼容性较好，如图 2-7 所示。

图 2-7 用灰度图像表示深度图

（2）伪彩色图像，为了更加直观的表现出不同的深度图的深度范围，将不同距离的深度映射成不同的颜色，例如红色表示较近的距离，蓝色表示较远的距离，一般在室内场景应用较多，如图 2-8 所示。

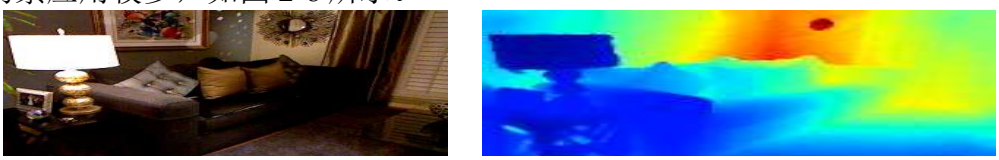


图 2-8 用伪彩色图像表示深度图

（3）彩色图像，彩色图像表示根据 Python 中的 matplotlib 库中包含的多种颜色来表示深度图的，这种表示方法最大的优点是深度图细节较为具体，本文深度图表示方法就是基于这种方式来表达深度图的，如图 2-9 所示。

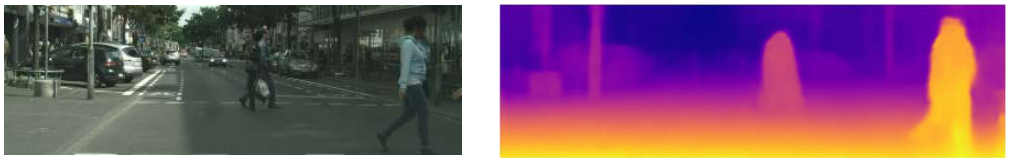


图 2-9 用彩色图像表示深度图

2.2 图像重构原理

图像重构是指根据已有的图像数据，使用一些科学方法去生成或重建原始图像的过程。在无监督单目深度估计过程中，采用图像重构损失（输入的源图像和重构目标图像之间的差异）作为其中的监督信号，来监督整个网络的训练。在无监督单目深度估计模型中图像的重构以图像扭曲（warping）方法和线性插值方法为基础，下面介绍如何使用这两个方法对图像重构。

2.2.1 图像扭曲

图像扭曲（Image Warping）又称为图像变形，图像扭曲的主要任务是通过一系列的空间变换以及形状变换来调整图像中每一个像素的位置来满足不同计算机视觉任务中不同映射的要求。图像扭曲过程如图 2-10 所示，已知一张源图像$I(x,y)$和坐标变换关系$T(x,y)$，如何计算得到一张变换后的图像$I'(x',y')$这就是一个最简单的

图像扭曲的过程，如何得到扭曲后的图像上的像素点就是最基本的扭曲问题。

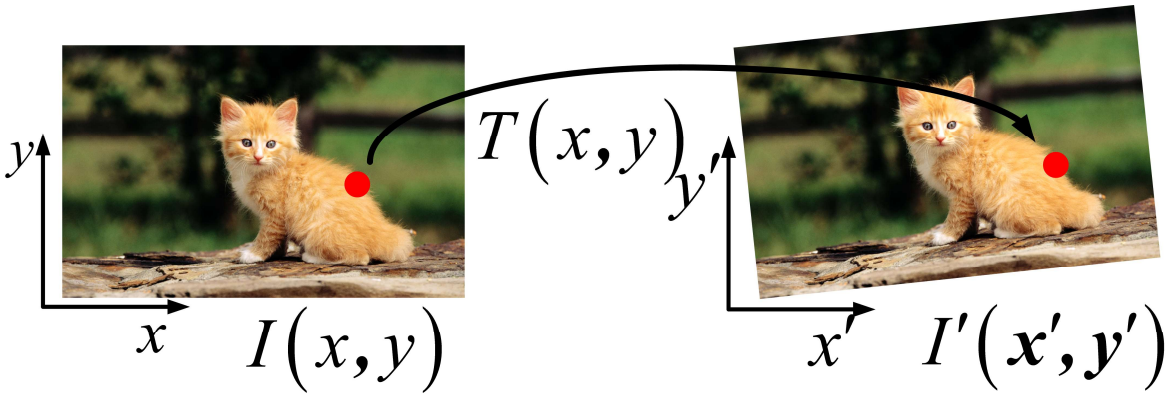


图 2-10 图像扭曲过程

常见的图像扭曲有仿射变换、投影变换、弯曲变换。研究人员在深度估计模型中使用到的就是投影变换。投影变换的本质即在仿射变换[24]的基础上加上透视变换，本质上是将图像从一个平面投影到另一个新的平面，投影变换的公式可以表示为：

$$\begin{pmatrix} x' \\ y' \\ w' \end{pmatrix} = \begin{pmatrix} a & b & c \\ d & e & f \\ g & h & i \end{pmatrix} \begin{pmatrix} x \\ y \\ w \end{pmatrix} \tag{2-15}$$

其中，(x,y,w)是源图像的像素坐标，(x',y',w')表示变换后的图像像素坐标，右边等式左起第一个矩阵表示投影矩阵。

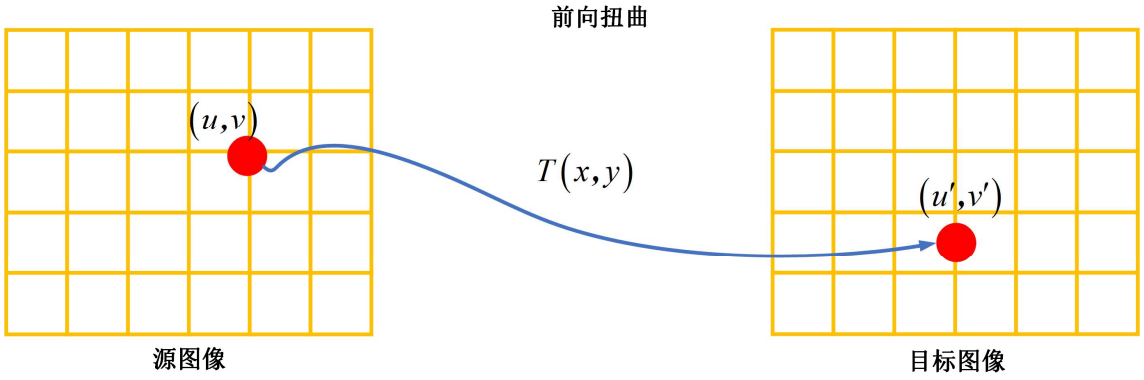


图 2-11 图像前向扭曲过程

图像扭曲方法分为前向扭曲和反向扭曲：

（1）前向扭曲。前向扭曲将源图像$I(x,y)$中的每一个像素值，根据坐标变换关系$T(x,y)$映射到目标图像$I'(x',y')$中相应的位置$(x',y') = I'(x,y)$上。如图 2-11 所示图像前向扭曲过程，源图像的像素点(u,v)被映射$T(x,y)$映射到目标图像的像素点(u',v')。但是根据坐标变换关系$T(x,y)$会存在 3 个问题：1）源图像像素映射到的位置不一定是整数值（由于重构图像的 RGB 值需要为整数）。2）源图像上的不同的像

素经过坐标变换关系$T(x,y)$可能被映射到目标图像的同一个像素点位置。3）源图像的像素映射到目标图像像素时出现没有对应的像素值。

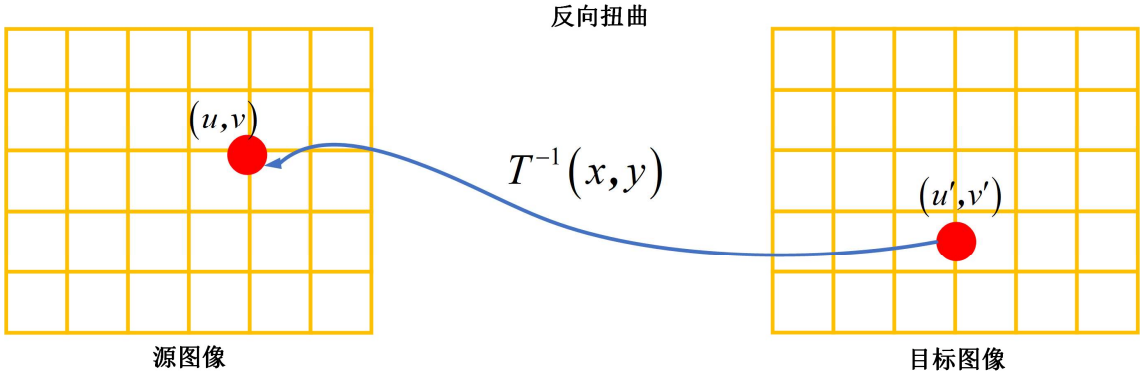


图 2-12 图像反向扭曲过程

（2）反向扭曲。为了解决前向扭曲过程中出现的目标像素没有对应的像素值情况，人们提出了反向扭曲。反向扭曲是指根据目标图像$I'(x',y')$中像素点位置(x',y')通过逆映射关系$T^{-1}(x',y')$找到源图像中像素点$(x,y)=T^{-1}(x',y')$的位置，并通过这种方法恢复出整个源图像。如图 2-12 所示，就是反向扭曲的过程，目标图像的像素点(u',v')被逆映射$T^{-1}(x,y)$映射为源图像的像素点(u,v)。由于反向扭曲时目标图像不存在空洞，每一个目标像素都可以找到对应的源像素，故一般情况下研究人员使用反向扭曲较多，只有当反向扭曲的映射关系$T^{-1}(x',y')$无法获得时才考虑使用前向扭曲。和前向扭曲一样反向扭曲也存在映射之后的源像素值不是整数的情况。

2.2.2 双线性插值

由于在进行图像扭曲的过程中会出现被扭曲之后图像的像素值不是整数的情况，为了解决这个问题本文选择使用线性插值方法，让非整数像素值逼近为整数像素值。常见的线性插值方法有：最近邻法、双三次插值、三线性插值和双线性插值。在无监督单目深度估计中，研究人员使用最多的是双线性插值方法。

当图像进行扭曲时，图像中一些像素的对应点并不存在或者不是整数的情况下，可以使用双线性插值方法。比如，当源图像的大小为$a\times b$，目标图像的大小为$m\times n$，且存在约束$m>a$，$n>b$时，目标图像上某一个像素点的坐标(i,j)对应的源图像上像素点的对应坐标为$\left(i\frac{a}{m},j\frac{b}{n}\right)$。如果存在$m,n$不是$a,b$的对应整数倍时，此时两个坐标$\left(i\frac{a}{m},j\frac{b}{n}\right)$也不是整数。故存在目标图像的像素点坐标$(i,j)$无法确定情况（灰度值和 RGB 值必须为整数），这时可以通过源图像上的坐标$\left(i\frac{a}{m},j\frac{b}{n}\right)$最近的 4 个

相邻的像素坐标来插值确定。

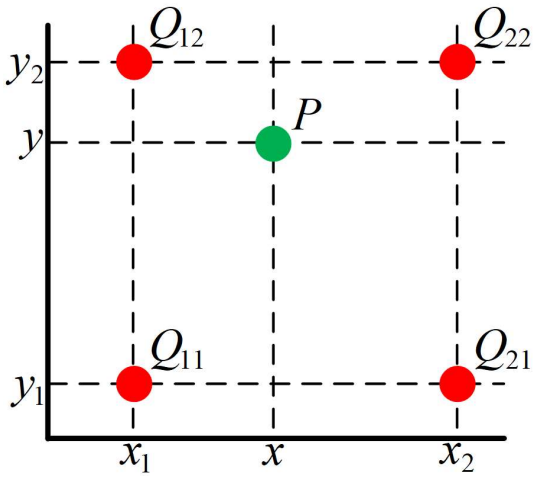


图 2-13 图像双线性插值

如图 2-13 所示，假设已知点$P=(x,y)$，已知函数$f(x)$在点Q_{11}、Q_{12}、Q_{21}、Q_{22}处的四个点值分别为(x_1,y_1)、(x_1,y_2)、(x_2,y_1)、(x_2,y_2)。首先在x方向进行插值、得到：

$$f(x,y_1)=f(R_1)\approx\frac{x_2-x}{x_2-x_1}f(Q_{11})+\frac{x-x_1}{x_2-x_1}f(Q_{21}) \quad (2-16)$$

$$f(x,y_2)=f(R_2)\approx\frac{x_2-x}{x_2-x_1}f(Q_{12})+\frac{x-x_1}{x_2-x_1}f(Q_{22}) \quad (2-17)$$

然后在y方向进行线性插值，得到：

$$f(P)\approx\frac{y_2-y}{y_2-y_1}f(R_1)+\frac{y-y_1}{y_2-y_1}f(R_2) \quad (2-18)$$

综合起来就是双线性插值的结果：

$$f(x,y)\approx\frac{f(Q_{11})}{(x_2-x_1)(y_2-y_1)}(x_2-x)(y_2-y)+\frac{f(Q_{21})}{(x_2-x_1)(y_2-y_1)}(x-x_1)(y_2-y)+\frac{f(Q_{12})}{(x_2-x_1)(y_2-y_1)}(x_2-x)(y-y_1)+\frac{f(Q_{22})}{(x_2-x_1)(y_2-y_1)}(x-x_1)(y_2-y) \quad (2-19)$$

又因为图像双线性插值只会用相邻的 4 个点，故上述公式分母都为 1，最终公式为：

$$f(x,y)\approx(x_2-x)(y_2-y)f(Q_{11})+(x-x_1)(y_2-y)f(Q_{21})+(x_2-x)(y-y_1)f(Q_{12})+(x-x_1)(y_2-y)f(Q_{22}) \quad (2-20)$$

以上就是整个双线性插值过程，是图像扭曲重构的必要步骤。

2.2.3 图像重构

根据 2.1 节的针孔相机模型，即相机模型是从三维空间到二维空间的一个映射，可以记相机模型为$T:R^3\longrightarrow R^2$。相机模型T可以把世界坐标系的三维点$P(X,Y,Z)$映射到像素坐标系中的二维点$p(u,v)$，即存在映射关系：

$$T(P)=\left(f_x\frac{X}{Z}+c_x,f_y\frac{Y}{Z}+c_y\right) \tag{2-21}$$

其中，(f_x,f_y,c_x,c_y)是相机内参。根据 2.1.2 三角测量知识有可以把Z转化为深度，故上式存在一个逆映射关系：如果已知一个像素点p的深度$D(p)$，也可以把它投影回 3 维空间，也就是逆映射关系$T^{-1}:R^2\times R\rightarrow R^3$：

$$T^{-1}\left(p,D(p)\right)=D(p)\left(\frac{u-c_x}{f_x},\frac{v-c_y}{f_y},1\right)^T \tag{2-22}$$

假设已知三维空间下两个拍摄位置之间的位姿变换(在无监督单目深度估计中通常由神经网络预测出位姿变换)，考虑两张图像，源图像和目标图像，则存在以下 3 个推论：1）若已知目标图像上每个像素的深度值，以及相机的内参，则根据式 2-22 可以得出，目标图像的像素坐标可以转换到目标图像的相机坐标系。2）若已知两个拍摄位置之间的位姿，根据式 2-12 可以在 1）的前提下将目标图像的相机坐标系下的点继续转换到源图像的相机坐标系下的点。3）在 2）的前提下，根据式 2-7 可以继续找到源图像对应的像素坐标。

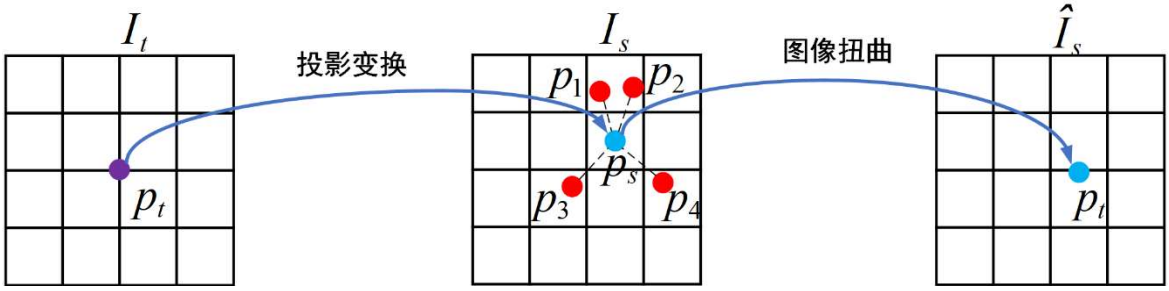


图 2-14 图像重构过程

总而言之，如果有了目标图像的深度和位姿，可以根据目标图像找到对应源图像的像素值来重构源图像。又，推论 1）~3）相当于 3 个映射关系，且出现的 3 个映射都是可逆映射，根据可逆映射的传递性：如果有了深度和位姿，可以在源图像上找到对应的像素值来重构目标图像。将上述重构关系总结为式 2-23：

$$p_s\sim K\hat{T}_{t\rightarrow s}\hat{D}_t(p_t)K^{-1}p_t \tag{2-23}$$

其中I_t表示目标图像(相机坐标系下)，I_s表示源图像（相机坐标系下），$\hat{I}_s$表示源图像（像素坐标系下），p_t表示目标图像的点，p_s表示对应源图像上的点，$\hat{p}_s$表示源图像从相机坐标系下转换到像素坐标系下的点。具体的过程如图 2-14 所示，其中p_1、p_2、p_3、p_4表示对p_s进行双线性插值的点。

这就是无监督单目深度估计过程中图像的重构过程，也是后续神经网络产生监督信号的重要来源。

2.3 基于无监督学习的单目深度估计的网络结构设计

基于无监督学习的单目深度估计网络模型的任务是通过输入单目图像预测相应的深度图像，网络通过对输入图像的重构来产生监督信号完成模型的训练，同时对输入的单目图像完成图像的推理过程。由于无监单目深度估计网络模型在训练过程中需要通过自身图像重构产生监督信号，这就对网络模型的结构提出了要求。不仅仅要求神经网络在重构的过程中产生准确的相机内外参数，同时也要求产生清晰深度语义信息。较为常见的单目深度估计模型中使用的网络结构有编解码结构（Encoder-Decoder）、卷积神经网络结构（Convolution Neural Network）、全卷积网络结构（Fully Convolutional）、自动解码器结构（Auto-Decoder）等。由于在处理语义分割、语义分类以及目标识别等相似任务中基于编解码器框架的性能最为优秀，结合基于无监督学习的单目深度估计模型要求，选择编解码结构为后续章节网络框架。编解码器结构通过对多层次特征、多尺度信息的提取和重建来保证输出高质量的深度图像，同时在编解码网络结构中可以方便添加注意力机制提高模型的泛化性能和鲁棒性。后续章节提出的创新点大部分都是基于编解码器的改进以及注意力机制的使用来提高深度估计网络的整体性能，下面依次展开介绍基于无监督学习的单目深度估计模型中编解码网络结构和注意力机制的原理和设计。

2.3.1 编解码结构

在基于编解码结构的神经网络中，有两部分构成分别是：编码器（Encoder）和解码器（Decoder）两部分构成。编码器的作用是负责将输入数据编码成低维表示，解码器的作用是使用上采样等方法将编码后的特征提高维度表述为原始数据。下面对无监督单目深度估计过程中使用的编码器和解码器的特征提取方式详细介绍。

1）编码器结构

在深度估计中使用较多的有卷积神经网络编码器、VGG[25]网络编码器和ResNet[26]网络编码器等。卷积神经网络是使用最早的一种编码器结构，通过卷积层的卷积核（滤波器）对输入图像进行卷积操作，捕获数据中的边缘纹理特征，并通过池化层对特征进行降维操作减少数据的冗余性来提高网络的泛化能力。VGG 编码器是另一种在深度估计中使用比较多的编码器。VGG 编码器与卷积神经网络编码器

类似，不同的是 VGG 使用的卷积核较小，仅使用了3×3的卷积核和2×2的最大池化层，同时其包含了 16 或 19 个卷积层，相较于卷积神经网络 VGG 网络深度较深。由于 VGG 编码器网络较深所以其捕获的深度语义信息更丰富，在深度估计过程中表现出来的泛化性能较佳。但是卷积神经网络网络编码器和 VGG 编码器随着网络的加深容易出现梯度消失和参数较多收敛缓慢的问题，在基于无监督学习的单目深度估计过程中本文选择使用 ResNet 编码器结构，ResNet 编码器最大的特点是使用了残差连接结构，这种跨层连接方式使得网络梯度能够更加平顺地传播，收敛速度更快。

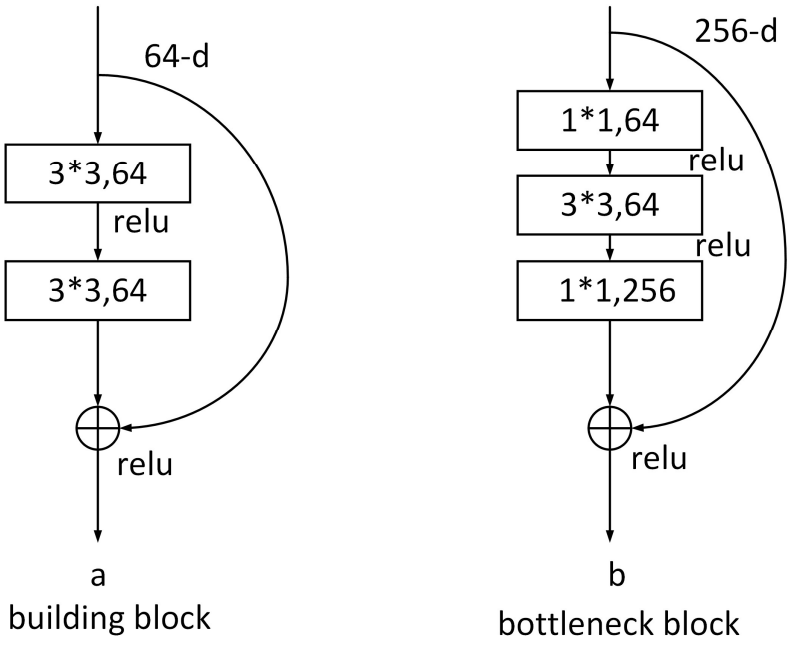


图 2-15 ResNet 残差单元

ResNet 编码器存在两种类型的残差单元如图所示 2-15，其中图 2-15 中，类型 a 是“building block”对应是浅层 ResNet 网络如：ResNet18 和 ResNet34；类型 b 是“bottleneck block”对应的是深层 ResNet 网络如：ResNet50、ResNet101、ResNet152。对于类型 a 残差单元分析：将输入的特征通过两个 3×3 卷积，且通道数保持 64 不变，右边通过从输入到输出的 shortcut（跳跃连接）结构，将主线经过两次卷积的输出特征和支路的 shortcut 特征相加（注意此时两个分支的维度相同），然后再进行输出，很好的保证了特征在传递过程中避免由于激活函数造成特征丢失。对于类型 b 残差单元分析：与左边 a 残差单元不同的是在输入和输出的时候添加了 1×1 的卷积层，同时通道数也会增加。由于输入的矩阵深度是 256-d，通过第一层 1×1 的卷积层时，输入的长和宽不变但是通道数减少到了 64（第一层起到降维的作用），而

在第三层的 1×1 的卷积时通道数从 64 升到 256（第三层起到升维的作用），这样做的好处是对于深层次网络可以大幅度降低运算量来提高运算效率。

表 2-1 是本文在训练过程中使用的 ResNet18、ResNet34、ResNet50 编码器，从表中可以看出 ResNet18 是一个相对较浅的编码器结构，ResNet18 每一层含有两个残差单元而，且 ResNet18 的残差单元类型都是 building block 形式的；而 ResNet34/50 网络结构较深每一层至少含有三个以上的残差单元，ResNet34 残差单元类型是 building block 形式的而 ResNet50 残差单元类型是 bottleneck block 形式的。

表 2-1 不同层数 ResNet 的参数设置

Layer name	Output size	18-layer	34-layer	50-layer
Conv1_x	112*112	7*7, 64, stride2		
Conv2_x	56*56	3*3 最大池化, stride2		
		$\begin{bmatrix}3*3 & 64\\3*3 & 64\end{bmatrix}*2$	$\begin{bmatrix}3*3 & 64\\3*3 & 64\end{bmatrix}*3$	$\begin{bmatrix}1*1 & 64\\3*3 & 64\\1*1 & 256\end{bmatrix}*3$
Conv3_x	28*28	$\begin{bmatrix}3*3 & 128\\3*3 & 128\end{bmatrix}*2$	$\begin{bmatrix}3*3 & 128\\3*3 & 128\end{bmatrix}*4$	$\begin{bmatrix}1*1 & 128\\3*3 & 128\\1*1 & 512\end{bmatrix}*4$
Conv4_x	14*14	$\begin{bmatrix}3*3 & 256\\3*3 & 256\end{bmatrix}*2$	$\begin{bmatrix}3*3 & 256\\3*3 & 256\end{bmatrix}*6$	$\begin{bmatrix}1*1 & 256\\3*3 & 256\\1*1 & 1024\end{bmatrix}*6$
Conv5_x	7*7	$\begin{bmatrix}3*3 & 512\\3*3 & 512\end{bmatrix}*2$	$\begin{bmatrix}3*3 & 512\\3*3 & 512\end{bmatrix}*3$	$\begin{bmatrix}1*1 & 512\\3*3 & 512\\1*1 & 2048\end{bmatrix}*3$
	1*1	平均池化, 1000-d 全连接层, softmax		
浮点数运算量（FLOPs）		$1.8*10^9$	$3.6*10^9$	$3.8*10^9$

2）解码器结构

在编解码结构中，解码器对编码器编码过程中提取的深度语义特征进行升维操作来恢复出原始图像的分辨率和维度，输出最终的深度。在深度学习领域使用较多的有：上采样结构、反卷积结构、自注意力机制结构、递归神经网络结构等。目前在单目深度估计中，为了提高预测出的深度图的整体精度，研究采人员提出了上述方法中改进的解码器结构，如多尺度上采样解码器、密集连接上采样解码器结构、细化特征反卷积解码器结构。本节结合相关无监督单目深度估计网络模型对无监督单目深度估计中常见的解码器结构进行介绍和分析。

（1） 分辨率多尺度上采样解码器结构

MonodepthNet[27]是无监督单目深度估计方法中使用最早和最经典的方法，如图 2-16 所示其网络结构图。首先，通过输入一对左右立体像对I^l和I^r作为源帧，左源帧I^l通过 MonodepthNet 输出左右立体图像对的深度图d^l和d^r，根据已知的源帧I^l、

I^r和估计的深度d^l、d^r进行扭曲（warp）得到重构后的重建帧$\hat{I}^l$、$\hat{I}^r$，将得到的

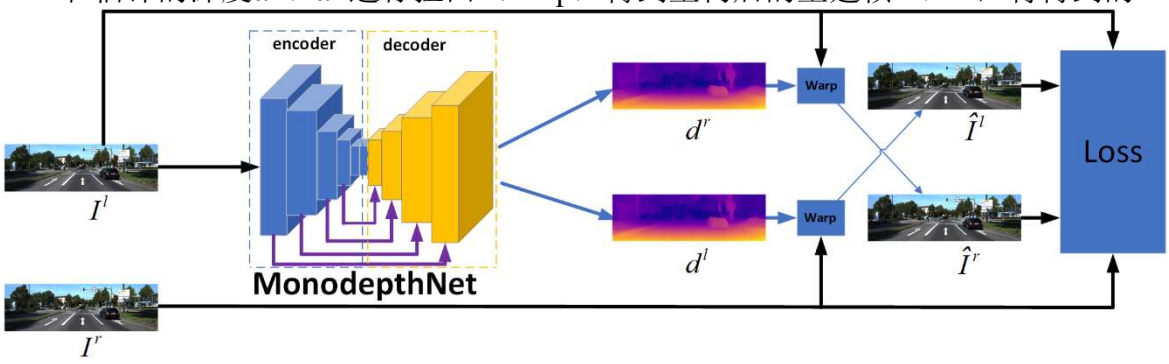


图 2-16 MonodepthNet

重建帧$\hat{I}^l$、$\hat{I}^r$和源帧I^l、I^r构成损失完成整个训练。注意到 MondepthNet 使用了具有跳跃连接功能的网络编解码器结构—全卷积神经网络(FCN)编解码器结构[28]，FCN 的具体结构如图 2-17 所示（仅以 FCN-8s 为例进行分析），可以将其分为两个阶段，第一阶段：编码器进行编码，使用经典的 VGG16 的卷边部分作为 backbone，同时将 VGG16 的最后三个全连接层也进一步加深变为三个卷积层。第二阶段：解码器进行解码，对编码后的特征进行上采样，首先使用双线性插值对转置卷积进行初始化，对编码器的第三个池化结果、第四个池化结果以及最后一层卷积层输出分别进裁剪以及上采样操，最后将三个分支的输出相加进行 8 倍的上采样还原到输出图像的分辨率大小。

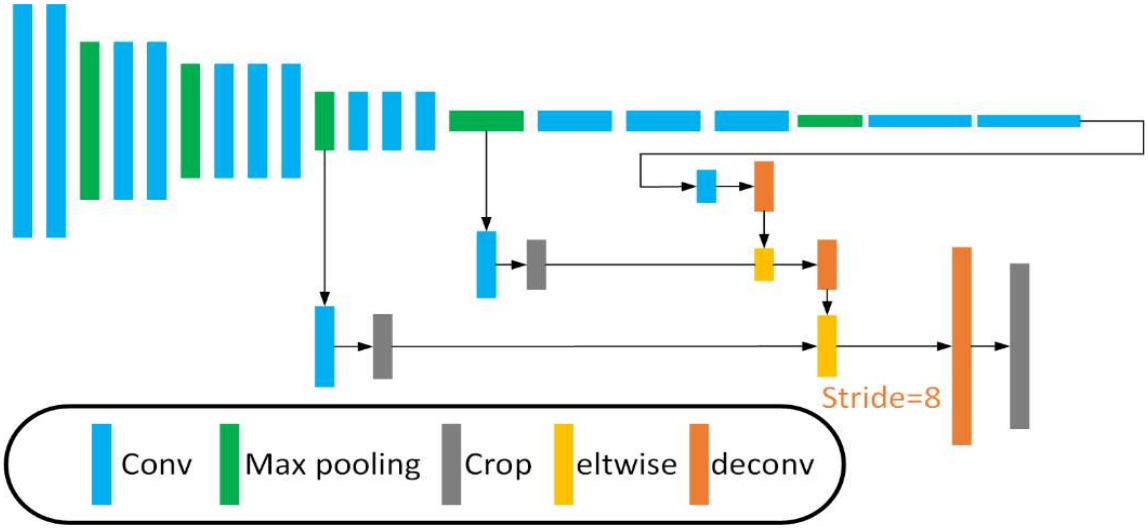


图 2-17 FCN 编解码器结构

相较于 CNN 编解码器，基于 FCN 的编解码器结构克服了在浅层的卷积层感知域较小的问题、对于一些较深的卷积层感知域较大可以学习到更多的抽象特征。但是使用 FCN 编解码器作为无监督单目深度估计的编解码器会存在未考虑到像素与像

素之间的位置关系即缺乏空间一致性。故 Godard[17]针对 MonodepthNet 算法进行改进提出了 Monodepth2 算法，如图 2-18 所示，整个无监督单目深度估计网络由两个

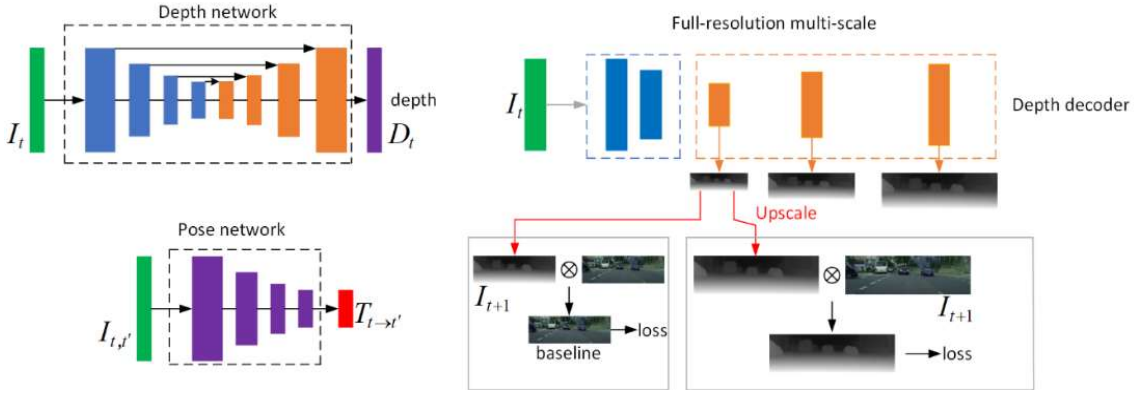


图 2-18 MonodepthNet 编解码结构

网络构成，深度网络（Depth network）和位姿网络（Pose network）。Godard[17]主要对 MonodepthNet 网络进行了两点改进：1）使用了位姿信号$T_{t\to t'}$来重构目标图像。2）在编解码结构中使用了全分辨率多尺度上采样解码器结构。Monodepth2 的编解码器结构使用了基于 Unet[29]结构的编解码结构，使用 Unet 编解码器结构的优点是可以方便的将不同尺度分辨率的深度图从中间层上采样到输入图像分辨率，同时在深度估计的总损失中添加了上采样过程中每个尺度的单个损失。Unet 编解码器相较于 FCN 编解码器比较相似，不同是 Unet 的解码器和 FCN 解码器的跳跃连接部分是不一样的，FCN 直接将特征进行叠加而 Unet 结构使用的是拼接操作，这种特殊结构

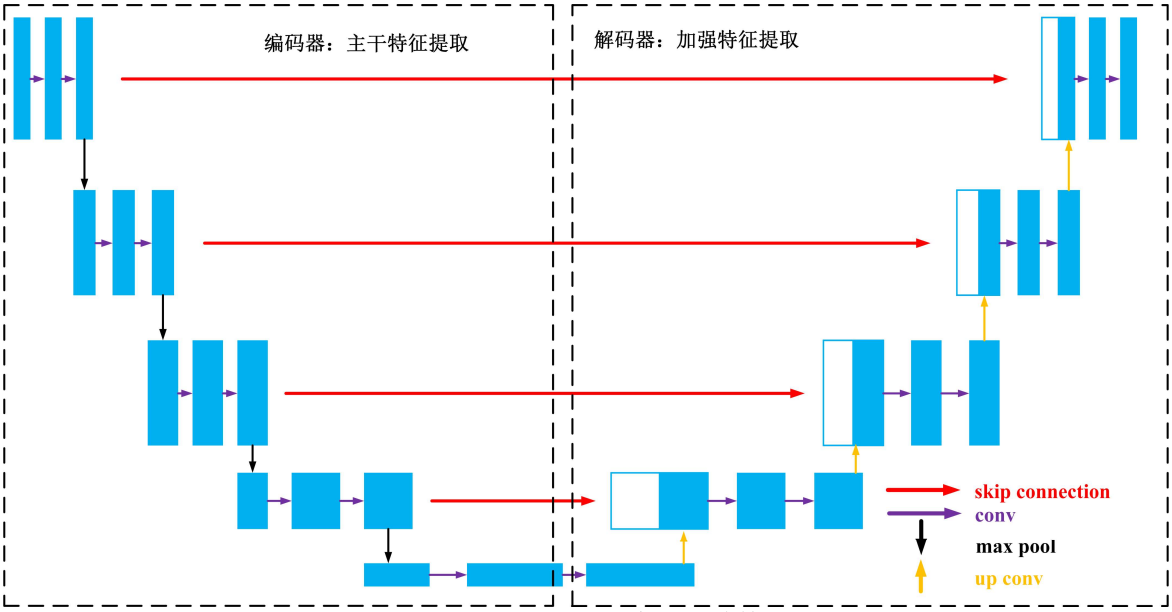


图 2-19 Unet 编解码结构

使得网络能够很好地保留图像的空间信息，有助于提高语义的准确性特别是边缘部分，Unet 编解码器的基本框架如图 2-19 所示。可以把 Unet 编解码结构分为两个部

分，第一个部分：第一部分的结构是基于 VGG16 的编码器结构，主要是进行卷积和最大池化的堆叠，进行主干特征的提取，利用提取出的五个初步有效特征层在第二部分将这五个有效特征层进行特征融合。第二部分：加强特征提取的解码器结构，利用在第一部分获得的五个初步有效特征层进行上采样和融合，最终获得一个与输入图像分辨率相同的有效特征层完成相应的深度估计任务。在后续第三章算法中 Unet 编解码器结构是一种很好的编解码器基准方便进行其它改进。

（2） 密集连接上采样解码器结构

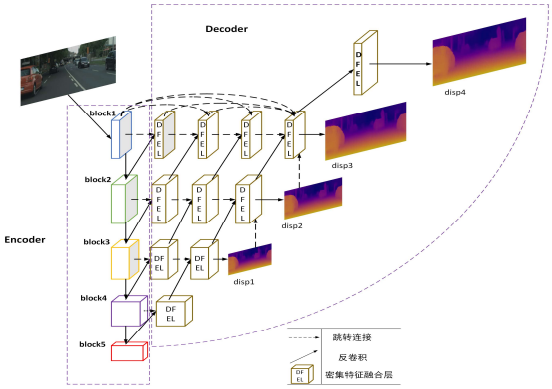


图 2-20 密集连接编解码结构

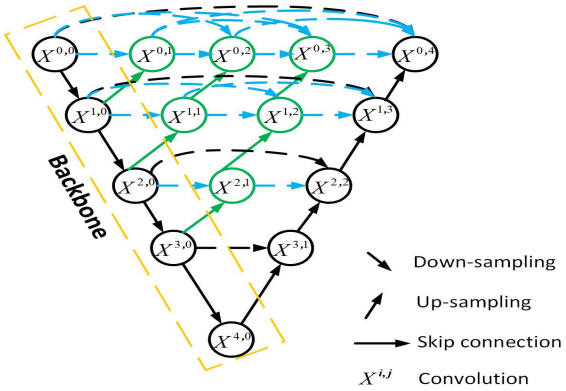


图 2-21Unet++密集连接编解码器

随着 Unet 编解码器结构在目标识别、语义分割、实例分割等多种计算机视觉任务的应用，人们根据 Unet 结构进一步提出了具有密集连接上采样解码器的 Unet++[30]结构。在无监督单目深度估计中，研究人员也将基于 Unet++的编解码结构应用到无监督单目深度估计算法中。陈莹[31]等人提出了基于 Unet++的密集连接编解码结构的无监督单目深度估计结构，如图 2-20 所示，他们在 Unet++的编解码网络基础上引入了密集特征融合层 DFEL，不仅仅提高了各级编解码器之间的信息互通和不同层次特征的重复利用率，同时在上采样过程中也加强了编码器提取抽象特征的能力。Unet++编解码器结构如图 2-21 所示：在解码器部分与 Unet 不同的是没有直接使用跳跃连接的方式进行上采样，而是进行稠密连接的方式，每一个特征的输出至少经历两个卷积层的上采样和下采样的操作，例如图 2-21 所示卷积层$X^{0,1}$输出是通过卷积层$X^{0,0}$进行一次下采样卷积和$X^{1,0}$进行一次上采样卷积拼接起来的过程。多次经过上述操作使得网络在解码过程中更容易捕获不同层次的特征以及通过叠加特征的方式让网络的感受野对不同大小的物体更加敏感。但 Unet++不足的地方是参数变多，网络鲁棒性变差，这也是后续第三章本文解决的问题之一。

（3）细化特征反卷积解码器

为了提高无监督单目深度估计的实时性以及泛化性能，Dosovitskiy 等[32]提出了

基于光流的无监督单目深度估计网络，整个网络结构如图 2-21 所示。编码器编码过程：通过输入两张相同的 RGB 图像到两个相同的处理流中，经过三个卷积层（独立且相同）来提取两张图像的特征；在第三个卷积层之后，通过关联层匹配操作（Conv_redir）将两张特征图进行组合到一起，最后再进行 6 次下采样。编码器编码

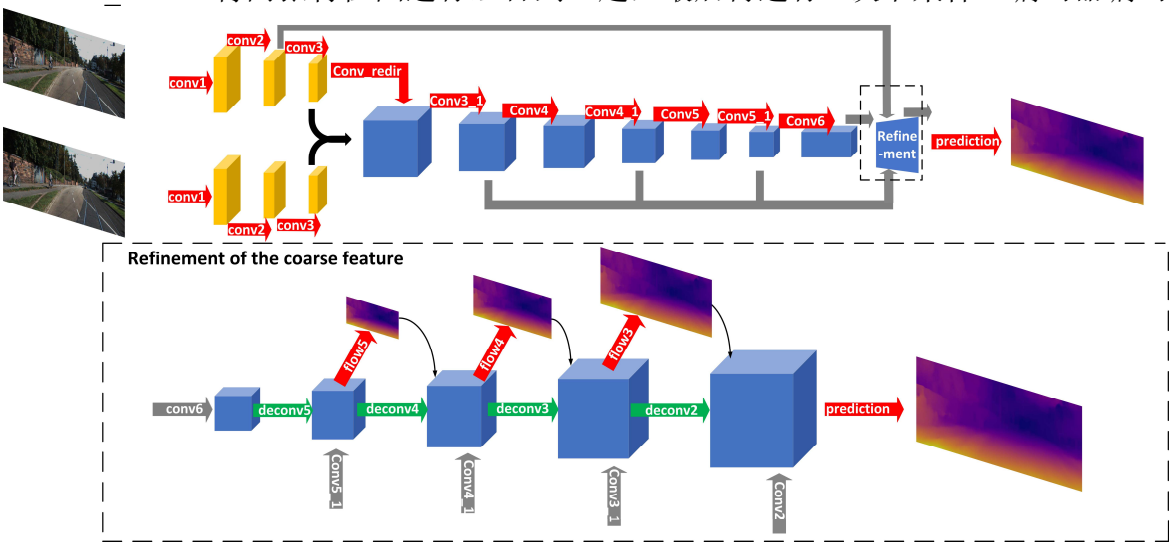


图 2-21 基于光流的无监督单目深度估计网络

过程，提出了细化特征反卷积解码器结构，该结构存在以下特点：首先主干部分用了 4 次上采样来进行图像尺度的恢复，同时将上一层输出的粗糙“光流”进行融合，这样既保证了编码过程中比较粗糙的特征图中的高级信息同时也保留了低层特征图精细的局部信息。该细化特征反卷积解码器由于解码过程可以高效抓取低层次和高层次的特征信息，在处理一些移动物体的深度估计中不失是一种好的思路，在第 4 章中针对动态场景下的基于边缘增强的无监督单目深度估计就是借鉴了这种解码器的思路。

2.3.2 注意力机制

神经网络是在基于模仿人类的神经元和突触等神经系统而来，不同的是在人类的视觉系统中，人类神经元与每一块小的且相邻区域的神经元相连接，而卷积神经网络中每个神经元只能与输入数据的整个区域相连，这就导致了人类的视觉系统对于视野中不同的对象分配的权重不一样，故对不同目标的注意程度不一样更容易捕捉关键物体，而卷积神经网络中并没有上述作用，研究人员将这种可以捕捉关键事物和关键特征的机制称为注意力机制（Attention Mechanism）。

为了模仿人类视觉系统中的注意力机制，在卷积神经网络中人们也提出了相应

的注意机制。注意力机制的核心思想在网络中引入可学习权重，这些权重的主要作用是根据网络任务的要求，可以通过自动调整权重来提取不同任务所需要的关键特征。在无监督单目深度估计过程中主要使用下三大类注意力机制：空间注意力机制、通道注意力机制和自注意力机制。

1）空间注意力机制

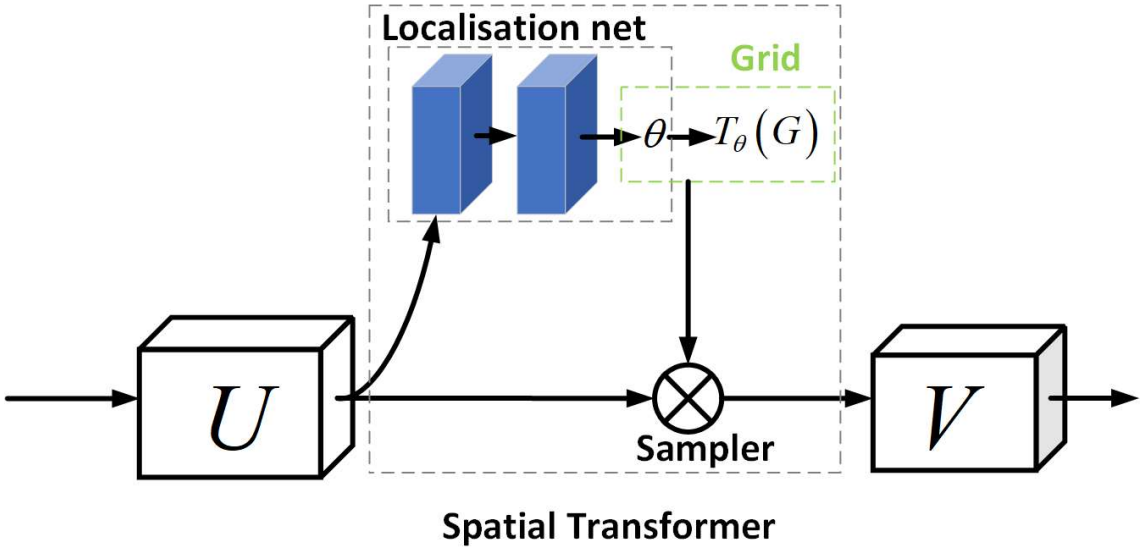


图 2-22 STN 空间注意力机制

空间注意力机制可以看作是对空间区域进行自适应选择的机制。常见的有 GENet[33]、RAM[34]等。第一个提出空间注意力机制的是 Google DeepMind[35]提出的 STN 网络如图 2-22 所示，这也是人们使用最多和最经典的网络结构之一。根据注意力计算过程可以将 STN 空间注意力机制网络计算过程分为 3 个步骤：第一步，输入图像U经过 Localisation net 的多重卷积之后得到相应的变换矩阵θ。第二步，Grid 映射处理阶段，将根据变换矩阵θ来计算映射之后的像素值位置$T_\theta(G)$；第三步，根据像素值位置$T_\theta(G)$通过双线性插值采样（Sampler）得到目标图像V。在 STN 的空间注意力机制结构中，利用网络通过调整图像本身来调整学习特征图像空间变换（且不改变网络的空间不变性特征），使得网络实现更高的精度。在第三章第四章的网络结构中也借鉴了这种思想方法。

2）通道注意力机制

在神经网络的训练过程，输入 RGB 图像除了要知道图像尺寸空间中的长和宽（即图 2-23 中的 H、W）还要知道输入图像（后文中没有特指的图像都是 RGB 图像）的通道数（即图 2-23 中的 C）。对于一些大型模型来说通道的数量多少以及通

道参数量的大小将直接影响模型对网络不同感受野的权重分配以及对于不同无效特

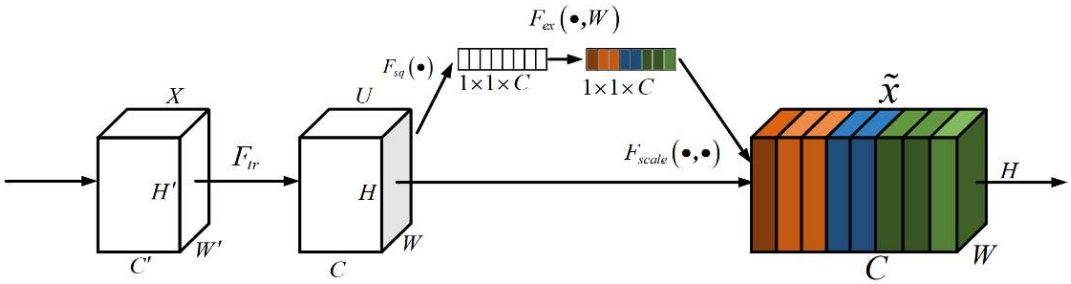


图 2-23 SENet 通道注意力机制

征的抑制。如图 2-23 所示网络结构，是 2017 年 ImageNet 使用的冠军通道注意力机制网络 SENet[36]。可以将 SENet 运算过程分为三个阶段：通道压缩阶段（F_{sq}）、自适应调整阶段(F_{ex})和权重重新加权阶段(F_{scale})。通道压缩阶段：将经过两次卷积之后得到的特征U通过 Squeeze 压缩操作，使其原有的空间维度变为1×1即将原来的全局空间信息压缩到通道描述符中，并对通道描符对其进行一次全局平均池化操作；自适应调整阶段：对压缩后的每一个特征应用通道自选门机制使其自动的加强有用特征权重抑制无用特征权重；权重重新加权阶段：将自适应阶段获得的新的权重通过逐通道乘法运算加权到新的通道权重上。由于 SENet 结构可以有效提到对重要特征的提取以及结构简单即插即用，在第 3 章和第 4 章算法研究中有一定的借鉴意义。

3）自注意力机制

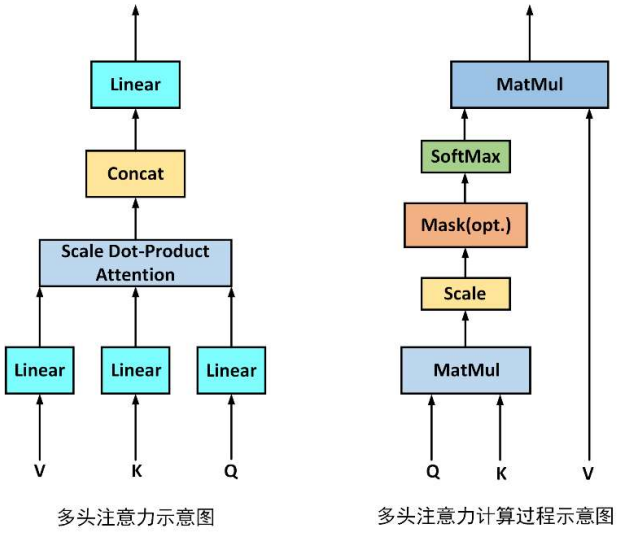


图 2-24 多头注意力机制

随着机器翻译模型 Transformer 的兴起，人们根据 Transformer 的结构提出了多头自注意力模型[37]，模型结构图 2-24 所示。多头注意结构最巧妙的应用是将“查询

-键值对”配对的方式引入自注意计算模型中。其中Q、K、V分别表示为查询（Query）、键（Key）、值（Value）。“查询-键值对”多头注意力计算过程可以分为两个步骤：第一步将输入特征线性投影到三个不同的空间中分别得到Q、K、V的权重值，且这三个空间的维度都相同；第二步根据线性投影得到的Q、K、V的值进行SoftMax 操作得到注意力的值，最后得到的值重新分配到原网络的卷积层。使用“查询-键值对”自注意力机制可以有效增加卷积神经网络的感受野以及不同场景中深度语义的理解，在第三章算法本文使用的特征交互模块就借鉴了这种结构加强了对于深度语义理解的。

在深度学习领域还有很多其它的注意力机制，在基于无监督单目深度估计模型中都有很强的借鉴意义，表 2-2 是近年来使用比较多的一些注意力机制，其中 Cls 表示分类、Icap 表示图像字幕、Det 表示图像字幕、FGCls 表示细粒度分类、SSeg 表示语义分割、ISeg 表示实例分割、Action 表示动作识别。

表 2-2 近年来使用较多的注意力机制类别

Category	Method	Publication	Tasks
RNN-based methods	RAM [34]	NIPS2014	CLs
	Hard and soft attention [38]	ICML2015	ICap
Predict the relevant region explictly	STN [35]	NIPS2015	Cls, FGCls
	DCN [39]	ICCV2017	Det, SSeg
Predict the relevant region explictly	GENet [33]	NIPS2018	Cls, Det
	PSANet [40]	ECCV2018	SSeg
Self-attention based methods	Non-Local[41]	CVPR2018	Action, Det, ISeg
	SASA[42]	NeurIPS2019	Cls, Det

2.4 常用数据集与评价指标

2.4.1 常用数据集

本节主要介绍一些用于无监督单目深度估计的公开数据集。

1) KITTI 数据集

KITTI 数据集是德国卡尔斯鲁厄理工学院和丰田美国技术研究院联合标定和测

量的，数据采集图像主要来自德国卡尔斯鲁厄地区，采集方式主要通过采集车的 2 个灰度摄像机、一个 64 线激光雷达、4 个光学镜头和一个 GPS 导航系统。是目前国际上主流使用的自动驾驶数据集之一。该数据集中包含市区、社区、乡村和高速公路等多种自动驾驶场景。每张图像中最多可以包含 15 辆汽车和 30 个行人，每张图像都有相应的激光雷达点云标签方便后续读取真实的深度值。本文使用 KITTI 数据集 22600 个图象对进行训练和 697 个图象对进行实验效果评估。

2) Cityscapes 数据集

Cityscapes 数据集是由德国斯图加特大学的计算机视觉小组创建，数据集图像主要来自德国斯图加特地区，采集方式与 KITTI 数据集类也使用了相机和激光雷达传感器。与 KITTI 数据集不同的是，Cityscapes 数据集主要针对城市街道不同的动态场景。Cityscapes 数据集包含了 30 个语义类别以及不同的天气条件，其中包含了 20000 张弱注释图像和 5000 张精细注释图像，500 张验证图像，1525 张测试图像。在动态场景下的深度估计是一种非常优秀的评估和训练数据集。

2.4.2 常用评价指标

深度估计模型性能有 5 个标准的评价指标，这些指标包括：均方根误差（RMSE）、相对均方误差（Sq Rel）、对数均方根误差（RMSE log）、平均相对误差（Abs Rel）、阈值精度（δ_i， i=1,2,3）。各指标计算公式如下：

$$\delta_i = max\left(\frac{y_{pred}}{y_{gt}}, \frac{y_{gt}}{y_{pred}}\right) < thr \tag{2-24}$$

$$Abs\ Rel = \frac{1}{n}\sum_{p}^{n}\frac{\left|y_{pred} - y_{gt}\right|}{y_{pred}} \tag{2-25}$$

$$Abs\ Rel = \frac{1}{n}\sum_{p}^{n}\frac{\left|y_{pred} - y_{gt}\right|^2}{y_{pred}} \tag{2-26}$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{p}^{n}\left(y_{pred} - y_{gt}\right)^2} \tag{2-27}$$

$$RMSE\ log = \sqrt{\frac{1}{n}\left|\log\left(y_{pred}\right) - \log\left(y_{gt}\right)\right|^2} \tag{2-28}$$

式中，y_{pred}为被预测图像像素的深度值；y_{gt}为被预测图像像素对应真实的深

度值；n为真实深度图中可获取的像素总数。除了阈值精度，其它四个误差值都是越小模型性能越佳

2.5 本章小结

本章介绍了无监督单目深度估计一些基本相关工作及原理，本章为后续提出的基于增强密集连接和全局特征交互的无监督单目深度估计模型和针对动态场景下的基于边缘增强的无监督单目深度估计模型起到抛砖引玉和理论支撑的作用。首先，建立起相机的几何模型，在相机几何模型的基础上介绍了单目相机测距的基本原理以及视差和深度的转换。其次，由于无监督单目深度估计信号来自于图像的重构本文又介绍了图像的重构和扭曲原理。再次，从构造无监督单目深度估计网络的角度，分析了一些经典的网络编解码结构以及注意力机制网络结构，这些结构为后续章节算法起借鉴改进意义。最后，介绍了无监督单目深度估计中使用的数据集和评价指标，这也是后文中需要使用的数据集和评价指标。

第 3 章 基于增强密集连接和全局特征交互的无监督单目深度估计

3.1 引言

受到 Goard[17]等人的启发，本章提出一种基于增强密集连接和全局特征交互的无监督单目深度估计模型。增强密集连接结构中将每一个增强密集连接模块加入了 CBAM 注意力机制，该结构可让网络更好地利用不同层采样的尺度特征，捕捉不同尺度上图像的深度语义信息，增加网络的感受野，加强对小目标深度的捕捉。增强密集连接结构同时有助于缓解学习过程中梯度消失问题，由于每一层都连接到先前的所有特征层，让梯度可以更容易地通过跨越多个特征层传播，从而帮助网络训练程中克服梯度消失的现象，加强对弱纹理区域的深度估计。全局特征交互结构是本章提出的另一个改进方法，该方法基于 Alaaeldin[22]提出的自注意力机制，本文引入了“查询-键值对”方式来计算沿着特征通道的注意力，这样可使注意力的权重更均匀，以保证模型预测出的不同的分辨率和多纹理图像的深度具有更强的鲁棒性。

3.2 本章无监督单目深度估计网络结构及其原理

本章所用的整体架构与 SFMlearner[18]框架类似，如图 3-1 所示，整个网络包括三个子模块，分别是 DepthNet（深度估计网络）、PoseNet（位姿估计网络）和损失函数（Loss）部分。在训练过程中，输入三张连续帧图像，将这三张连续帧图像分别记为I_t、I_{t+1}、I_{t-1}，其中I_t是目标图像帧，而相邻的两帧I_{t+1}和I_{t-1}称为源图像帧。整个网络的训练过程主要分为四个步骤。第一步是让目标帧I_t输入 DepthNet 预测出其相应的深度图 d。第二步在 PoseNet 依次输入两对帧即目标帧I_t、源图像帧I_{t+1}和目标帧I_t、源图像帧I_{t-1}，可以让 PoseNet 学习目标帧I_t、源帧I_{t+1}（I_{t-1}）之间的位姿信息即目标帧相对于源帧的旋转R（由于目标帧I_t和源帧I_{t+1}（I_{t-1}）之间是在极小的时间间隔内变化，一般忽略旋转影响）和平移T，将目标帧I_t和源帧I_{t+1}（I_{t-1}）之间预测出的平移记为$T_{t\to t+1}$（$T_{t\to t-1}$）。第三步进行目标图像的 project（重投影）以及目标图像的重构，将重构后的图像记作I_{t0}。重构的过程主要通过 warping（扭曲）来完成的。具体的 warping 过程为：如图 1，取深度图 d 上一点p'（目标帧I_t上对应的点为p），然后将深度图上的点p'通过 project（重投影）在两张源帧I_{t+1}、I_{t-1}上分别记为p^{t+1}、p^{t-1}；根据式（3-1）就可以估计出点p重构后的

点：

$$p_s \sim K T_{t\to s} d(p') K^{-1} pt \qquad (3-1)$$

其中，$T_{t\to s}$表示目标帧到源帧的平移（第二步已获得），K表示相机的内参（大部分相机内参出厂时就已知），$d(p')$表示p'相应的深度值（第一步已获得），p_s表示重构后的图像I_{t0}对应目标帧点p的点。最后将多个重构点p_s进行双线性插值，重构目标图像，这就是整个 warping 的过程。第四步进行损失计算。将重构帧I_{t0}和目标帧I_t进行比较，比较后的差异构成训练损失的主要分量，基于平滑损失函数和光度损失函数惩罚差异（在 3.3.3 节详细分析损失函数），网络会预测出正确的深度值。以上就是整个网络的训练过程。

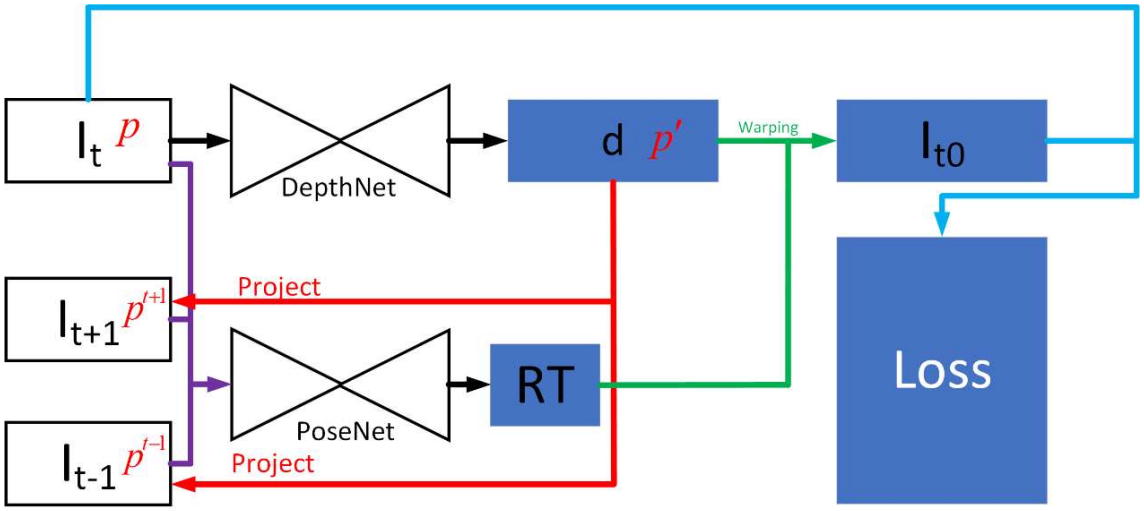


图 3-1 无监督单目深度估计网络结构及其原理

3.3 本章方法

本章提出新的网络模型如图 3-2 所示，基于第 1 节提出的整体网络结构及原理将本章设计的网络划为三个阶段：深度估计阶段、位姿估计阶段、损失计算阶段。深度估计阶段：深度估计阶段的主要任务是估计出目标帧的深度，是之后目标图像重建的基础，主要由 DepthNet（深度估计网络）构成。对于目前单目无监督深度估计，本章对传统的 DepthNet 进行了改进，设计了一种基于增强密集连接模块的密集连接编码和解码结构，加强了对弱纹理区域、小目标深度的提取，同时引入了全局特征交互模块用来抵抗输入不同尺度深度图序列中的噪声，让模型预测不同的分辨率和多纹理图像的深度具有更强的鲁棒性。位姿估计阶段：位姿阶段主要任务是估计出连续帧图像之间的平移和旋转信息，主要由 PoseNet（位姿估计网络）构成。

PoseNet 是基于 ResNet18[26]网络模型，可预测出 6 自由度的平移和旋转位姿。损失计算阶段：用来监督整个网络的训练。损失计算阶段主要由两组损失函数构成，光度损失和平滑损失，这两组损失作为监督信号监督整个网络学习过程。下面对各阶段组成部分详细介绍，以及各组分的贡献。

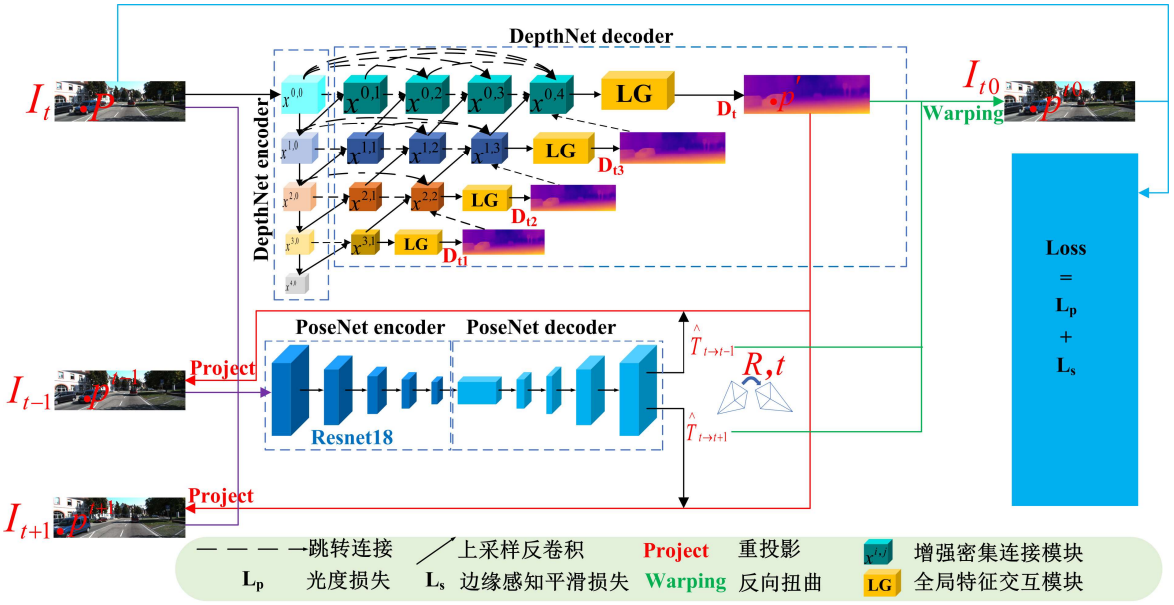


图 3-2 基于增强密集连接和全局特征交互的无监督单目深度估计

本章提出新的网络模型如图 3-2 所示，基于第 1 节提出的整体网络结构及原理将本章设计的网络划为三个阶段：深度估计阶段、位姿估计阶段、损失计算阶段。深度估计阶段：深度估计阶段的主要任务是估计出目标帧的深度，是之后目标图像重建的基础，主要由 DepthNet（深度估计网络）构成。对于目前单目无监督深度估计，本章对传统的 DepthNet 进行了改进，设计了一种基于增强密集连接模块的密集连接编码和解码结构，加强了对弱纹理区域、小目标深度的提取，同时引入了全局特征交互模块用来抵抗输入不同尺度深度图序列中的噪声，让模型预测不同的分辨率和多纹理图像的深度具有更强的鲁棒性。位姿估计阶段：位姿阶段主要任务是估计出连续帧图像之间的平移和旋转信息，主要由 PoseNet（位姿估计网络）构成。PoseNet 是基于 ResNet18 网络模型，可预测出 6 自由度的平移和旋转位姿。损失计算阶段：用来监督整个网络的训练。损失计算阶段主要由两组损失函数构成，光度损失和平滑损失，这两组损失作为监督信号监督整个网络学习过程。下面对各阶段组成部分详细介绍，以及各组分的贡献。

3.3.1 深度预测阶段

深度估计阶段：如图3-2，深度估计阶段主要是将目标帧I_t输入DepthNet之后提取出目标帧的特征并估计出其相应的深度，估计出的深度图D_t是之后的目标图像重建的基础，而整个DepthNet的网络拓扑结构也将直接决定了整个深度估计网络的性能。如图3-3所示，列举出了4个DepthNet的网络拓扑结构，其中Unet拓扑[29]和Unet++拓扑[30]是已经被应用过的拓扑结构，而基于增强密集连接和全局特征交互拓扑结构是本章提出来的新的拓扑结构，增强密集连接拓扑仅用来做消融对比实验（见3.4.3节）。Unet拓扑结构的不足是：特征仅仅在同一层编、解码器之间使用跳转连接，导致编码器对于不尺度同特征的使用是不充分的，容易让预测出的深度图纹理缺失和边缘模糊。而Unet++拓扑结构改变了特征仅仅只是在同一层编、解码器之间的跳跃问题，实现了特征在多层编码和解码之间的融合，网络性能得到了进一步提高，但是密集连接本身也导致了网络复杂程度变高，使得对不同的分辨率和多纹理图像的深度估计的鲁棒性变差。本章基于Unet拓扑和Unet++拓扑结构进一步提出新的基于增强密集连接和全局特征交互拓扑结构。如图3-3所示，新的拓扑结构弥补了由Unet及Unet++结构带来的特征利用不充分、不同尺度的目标的特征表达能力弱、鲁棒性差的问题。新的拓扑结构应用了增强密集连接模块和全局特征交互模块，增强密集连接模块加强了对弱纹理区域、小目标深度的提取；全局特征交互模块用来抵抗输入不同尺度深度图序列中的噪声，加强了网络的鲁棒性能和泛化性能。

本章的DepthNet的密集连接原理可以做如下定量表述，将每一个增强密集连接模块定义为$x^{i,j}$，将每一个增强密集连接模块的输出定义为$x_0^{i,j}$。其中上标i表示编码器下采样的方向，而上标j表示沿着跳过的路径对增强密集连接模块的卷积层进行连接，每个增强密集连接模块的输出为$x_0^{i,j}$，$x_0^{i,j}$可以表示为：

$$x_0^{i,j}=\begin{cases}H\left(x^{i-1,j}\right), & j=0\\ H\left(\left[\left[x^{i,k}\right]_{k=0}^{j-1},u\left(x^{i+1,j-1}\right)\right]\right), & j>0\end{cases} \tag{3-2}$$

式（3-2）中，$H(x)$表示卷积运算，$u(x)$表示表示上采样层，[⬚]表示concatenate层（特征拼接层）。参照图3-2和式（3-2），对于节点$j=0$的增强密集连接模块的输入仅仅来自编码器的前一层；对于节点$j=1$的增强密集连接模块的输入需要接收两个输入，分别来自两个相邻的$j=0$的增强密集连接模块而且分别是来自这一层的卷积以及下一层的上采样然后进行concatenate（拼接）操作；更一般的情

况，对于$j > 1$的增强密集连接模块，将接收$j + 1$个输入，其中j个输入是相同跳跃路径中的先前j个增强密集连接模块的输入，并且最后一个输入是来自较低跳跃路径的上采样输出。

下面将详细阐述增强密集连接模块和全局特征交互模块构建方法，以及各组分的贡献。

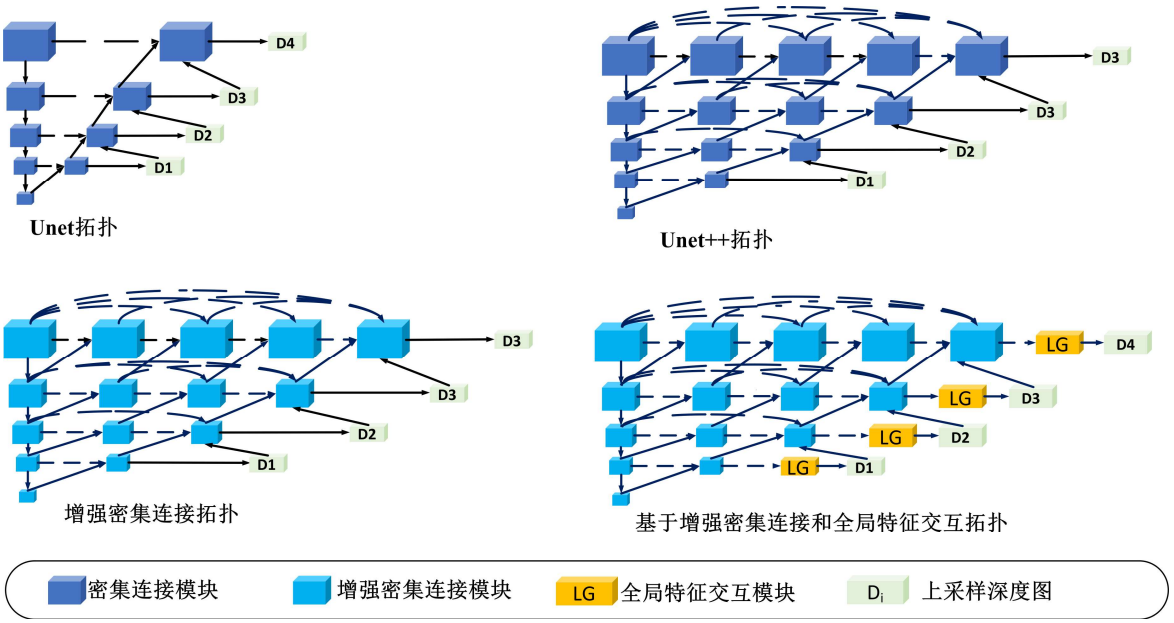


图 3-3 DepthNet 网络拓扑结构和本章网络拓扑结构

3.3.1.1 增强密集连接模块

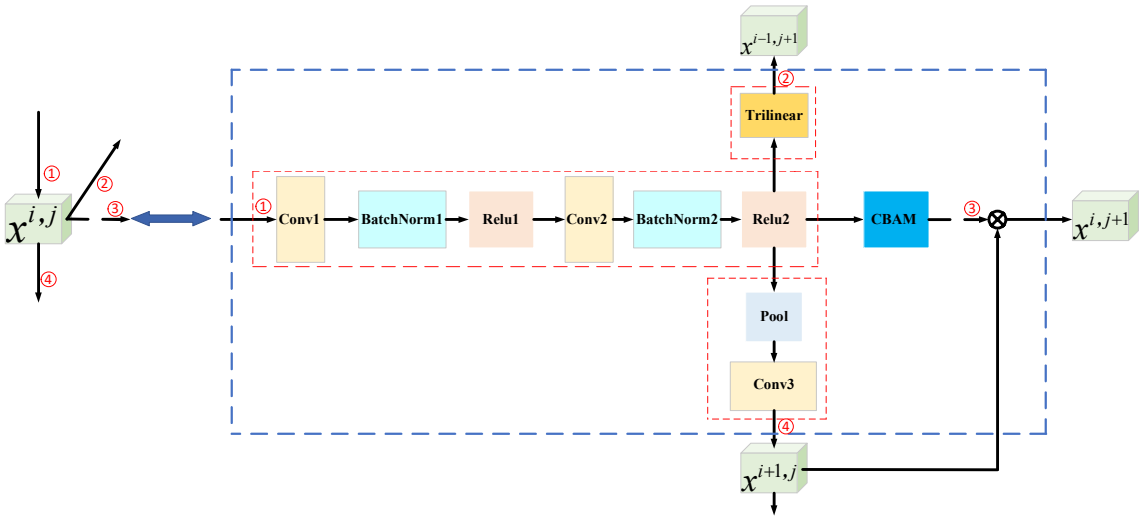


图 3-4 增强密集连接模块

基于 Unet 和 Unet++的密集连接模块基础上，本章提出了增强密集连接模块，

增强密集连接模块的结构如图 3-4 所示。将整个模块分为四个部分，即图 3-4 的①②③④。其中①表示输入部分，分别经过两次卷积、批归一化、Relu 激活函数操作；②表示对输入的特征使用三线性插值进行上采样，上采样后的输出作为为模块$x^{i-1,j+1}$的输入；④表示对输入的特征进行下采样操作，经过一次最大池化和卷积操作再进行下采样，下采样后的输出作为$x^{i+1,j}$的输入；③表示对①输出进行 CBAM[23]卷积注意力操作，同时③和模块$x^{i+1,j}$上采样后的结果进行一次拼接，拼接后的结果作为模块$x^{i,j+1}$的输入。

增强密集连接模块相对于传统密集连接模块做出了两点改进，将上采样的双线性插值改为三线性插值，同时在③的输出之后添加了 CBAM 卷积注意力机制模块。三线性插值相较于双线性上采样的精度更高，使得采样之后的深度特征更加平滑以及深度边缘更加清晰。CBAM 注意力机制的主要目标是增强卷积神经网络对于不同特征通道和空间位置的感知能力，从而提升网络的表现。CBAM 总体结构，如图 3-5：输入一个中间特征图，然后沿着两个独立的维度（通道维度和空间维度）依次推断注意力图，然后将注意力图乘以输入特征图以对特征进行自适应修饰。同时 CBAM 注意力机制可以明显地分割出小物体对象，也能够更易明确小物体对象和背景之间的相对位置关系，使得小物体深度出现的伪影更少，边缘信息更加明确。

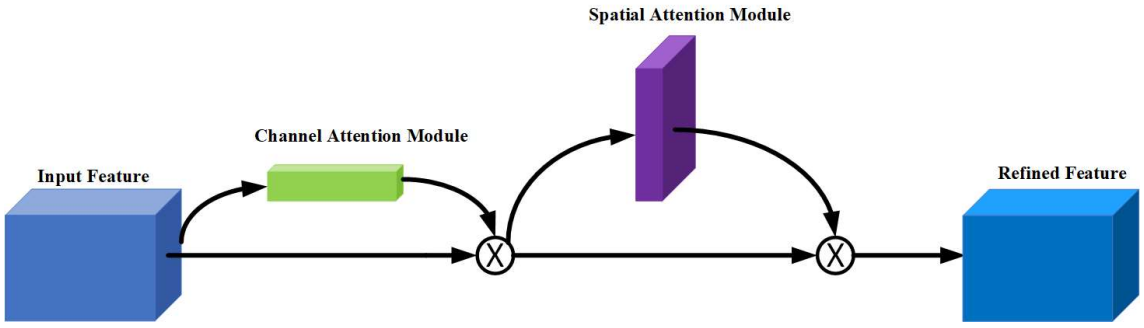


图 3-5 CBAM 注意力机制结构

用公式表述整个 CBAM 卷积注意力机制过程：

$$F' = M_c(F) \otimes F \tag{3-3}$$

$$F'' = M_s(F) \otimes F' \tag{3-4}$$

其中，⊗表示逐元素乘法；$F' \in \mathbb{R}^{c\times1\times1}$表示经过通道注意力$M_c$之后得到的特征，$F'' \in \mathbb{R}^{c\times H\times W}$是中间特征$F$输入空间注意力$M_s$之后得到的最终特征，$c$表示输入特征图的通道数，$H$、$W$表示特征图的长和宽。

3.3.1.2 全局特征交互模块

全局特征交互模块由 Alaaeldin[22]提出的自注意力机制改进而来，引入了“查询-键值对”注意力机制（Query-Key-Value，QKV）[41]如图 3-6 所示，由于此模块一开始就让 Query（所需要得到的查询信息）和 Key(键)拼接然后再学习其相关性（权重），让学习到的 Value（值）的权重与 Key（键）权重产生更有效的匹配。由于 Value 和 Key 的强匹配特性，全局特征交互模块抵抗输入不同尺度深度图序列中的噪声（多纹理图像）能力增强，同时提高了模型的鲁棒性和泛化能力。

全局特征交互模块具体结构如图 3-6 所示，设经过密集连接之后的特征为X_1，大小为$H \times W \times C$，经过 Layer Norm（层归一化），并将其线性投影（Linear）到三个不同的向量空间，即查询向量Q，键向量K，值向量V。根据这 3 个向量，网络可以估计出对它们应的三个权重矩阵W_q，W_k，W_v，其中Q、K、V用权重矩阵可以表示为：

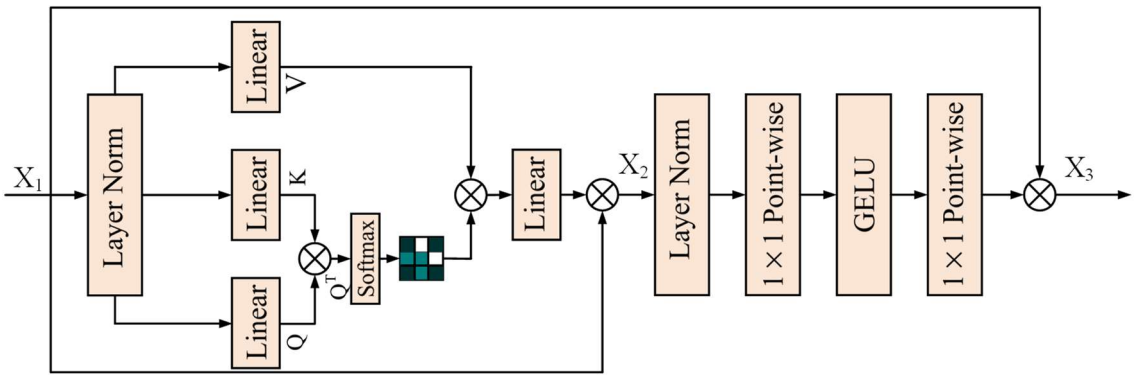


图 3-6 全局特征交互模块结构

$$Q = W_q X_1;\ K = W_k X_1;\ V = W_v X_1$$

将Q、K、V用向量表示为：

$$Q = \{q_1, q_2, q_3 \dots q_N\};\ K = \{k_1, k_2, k_3 \dots k_N\};\ V = \{v_1, v_2, v_3 \dots v_N\}$$

通过“查询-键值对”注意力机制的方式来计算输出X_2：

$$X_2 = \sum_{n=1}^{N} attention\big((K,V), q_n\big) + X_1 = \sum_{j=1}^{N} softmax\left(\frac{Q^T K}{\sqrt{D}}\right) v_j + X_1 \qquad (3-5)$$

其中$(,)$表示向量之间的拼接操作。为了使非线性特征更加明显可以对输出的特征X_2进行增强，最终的输出X_3为：

$$X_3 = X_1 + P\Big(G\Big(P\big(LN(X_2)\big)\Big)\Big) \qquad (3-6)$$

其中P表示进行大小为1×1的卷积降维操作获得网络提取更加高级的特征，G表示$GELU$激活函数，LN表示层归一化。

3.3.2 位姿估计阶段

在位姿估计阶段，在 PoseNet（位姿估计网络）输入目标帧与源帧，相邻的三帧作为 PoseNet 的输入，输出为 6 自由度的旋转和平移向量。由于 PoseNet 是基于 ResNet18 为编码器使用的网络权重内存较少，所以整个网络非常简洁容易实现轻量化。PoseNet 由 7 个步长为 2 的卷积层组成，解码器由 4 个步长为 1 的反卷积层组成，最后，将全局平均池化应用于所有空间位置的聚合预测，输出具有六自由度的相对姿态。

3.3.3 损失计算阶段

损失函数作为单目无监督深度学习的重要组成部分，是提供监督信号的唯一来源。本章损失函数主要包含两组函数：平滑损失函数和光度损失函数。

平滑损失函数构成了深度正则化分量，目的是规范低梯度区域的深度定义如下：

$$L_s(D_t,I_t)=|\partial_x D_t(x,y)|e^{-|\partial_x I_t(x,y)|}+\left|\partial_y D_t(x,y)\right|e^{-|\partial_y I_t(x,y)|} \qquad (3-7)$$

其中，$D_t(x,y)$表示估计出的深度图在点(x,y)处的深度，$I_t(x,y)$表示目标帧在点(x,y)处的像素值，$\partial_i D_t(x,y)$和$\partial_i I_t(x,y)$是梯度值，分别表示深度图D_t和对应目帧I_t在点(x,y)处x或y方向的梯度，其中$i\in\{x,y\}$。

光度损失函数构成了一致性正则化分量，目的是保证重构后的目标帧I_{t0}和源目标帧I_t在每一点处的相似性，定义如下：

$$L_p(I_t,I_{t0})=\frac{\alpha}{2}\left(1-SSIM(I_t,I_{t0})\right)+(1-\alpha)|I_t-I_{t0}| \qquad (3-8)$$

其中，$SSIM(x,y)$表示图像结构相似性损失函数，I_t和I_{t0}分别表示源目标帧和重构后的目标帧，α表示超参数。

考虑到遮挡和运动物体的影响，沿用 Monodepth2[6]中的最小光度误差方法，来自动掩蔽遮挡和运动物体的影响，光度损失重新表示为：

$$L_p'=min\left\{L_p(I_t,I_{t0}),L_p(I_{t'},I_{t0})\right\} \qquad (3-9)$$

其中，$t'\in(t+1,t-1)$。

综上，整个训练损失表示为：

$$L_{total}=\lambda L_s+\mu L_p' \qquad (3-10)$$

其中，λ和μ表示超参数。

3.4 实验结果分析

3.4.1 实施细节

本章使用深度学习库 Pytorch1.7.1 在 NVIDIA GeForce RTX3070 上训练了该模型，操作系统为 Ubuntu18.04。以 Adam 作为优化器对整个网络进行 10^6 次迭代训练，优化器的学习率设置为 10^{-4}。训练的批量大小为 16，训练的 RGB 图象分辨率大小为 1024×320。损失损失函数中几个超参数的设置：3-8 式中$\alpha = 0.85$；3-10 式中λ和μ分别设为 0.01 和 0.001。

3.4.2 实验结果分析

为了验证本章所提出的方法的有效性和先进性，在 KITTI 数据集上将本章方法与近几年其它方法进行对比，定量分析结果如表 3-1 所示，其中分辨率栏表示训练时使用的 RGB 图像分辨率大小，其它 7 栏表示前文介绍的评估指标。定性结果如图 7 所示，从表 3-1 可以看出：

表 3-1 与其它模型在 KITTI 数据集上的测试结果对比

方法	分辨率	误差指标（越小越好）				预测准确率（越大越好）		
		Abs Rel	Sq Rel	RMSE	RMSE log	δ<1.25	δ<1.25²	δ<1.25³
Zhou[18]	128×416	0.208	1.768	6.856	0.283	0.678	0.885	0.957
GeoNet[44]	128×416	0.155	1.296	5.857	0.233	0.793	0.931	0.973
DDVO[43]	128×416	0.151	1.257	5.583	0.228	0.810	0.936	0.974
Casser[45]	128×416	0.141	1.026	5.291	0.215	0.816	0.945	0.979
Ranjian[46]	256×832	0.148	1.149	5.464	0.226	0.815	0.935	0.973
Gordon[47]	128×416	0.128	0.959	5.230	0.212	0.845	0.947	0.976
Yang[48]	384×512	0.131	1.254	6.117	0.220	0.826	0.931	0.973
EPC++[49]	1024×320	0.141	1.029	5.350	0.216	0.816	0.941	0.976
Goard[17]	1024×320	0.115	0.882	4.701	0.190	0.879	0.961	0.982
Luo[50]	112×384	0.130	1.086	4.876	0.205	0.878	0.946	0.970
Sun[19]	640×192	0.108	0.744	4.709	0.188	0.864	0.960	**0.984**
Guizilini[20]	640×192	0.102	0.728	4.378	0.196	**0.892**	0.961	0.977
Chen[21]	640×192	0.105	0.745	4.537	**0.184**	0.883	0.963	0.983
Ours	1024×320	**0.095**	**0.716**	**4.178**	0.195	0.881	**0.972**	0.982

1）不同分辨率的情况下，与表现最优的方法相比，本章提出的模型在平均相对误差（Abs Rel）、相对均方误差（Sq Rel）、均方根误差（RMSE）相对于前者至少减少了 7%，4%，5%。

2）相同的分辨率情况下，与表现最优的方法相比，本章提出的模型在平均相对误差（Abs Rel）、相对均方误差（Sq Rel）、均方根误差（RMSE）相对于前者分别减少了 17%，19%，11%。

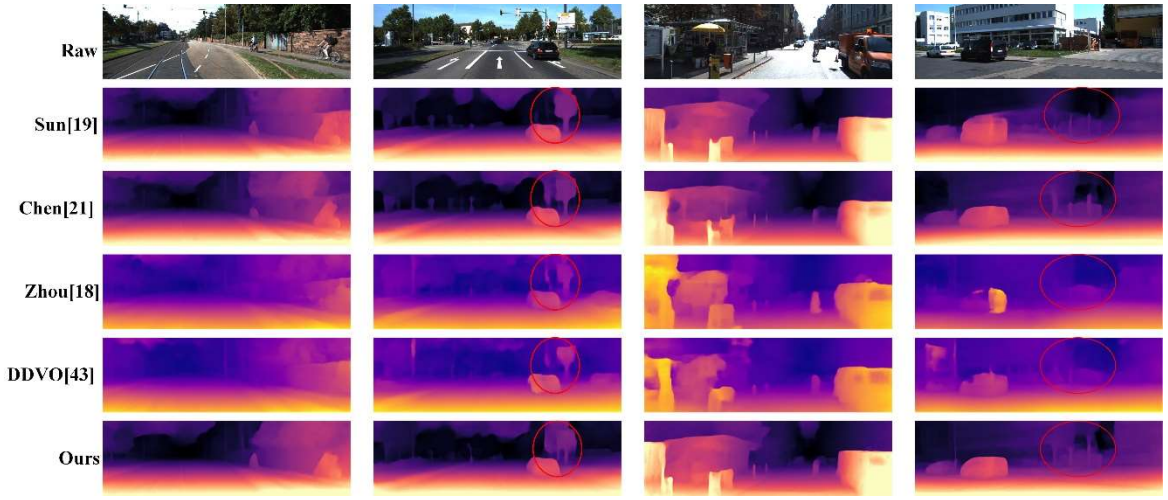


图 3-7 KITTI 数据集上的定性分析结果

为了更直观地定性表明本模型与其他模型之间的效果对比，抽取了表 1 中的部分模型进行复现并与本章的方法进行对比，结果如图 3-7 所示。其中 Raw 行为测试图像，Sun[19]、Chen[21]、Zhou[18]、DDVO[43]为与表 1 对应的模型预测出的深度图，其中用红色圆圈所圈出来的部分是重点观察的部分。

图 3-7 中，左起第 2 张图片，包含有红绿灯杆和路牌杆，经对比：本文提出的模型可以清晰的捕捉和区分红绿灯和路牌的杆，而其它模型容易出现伪影或者边缘不清楚的情况。左起第 4 张图片，包含有围栏和附近的汽车，经对比：本文提出的模型可以清晰捕捉围栏旁边的三根细杆，且预测出的汽车边缘较深，而其它对比模型会出现围栏旁边的三根细杆预测不全以及预测出的汽车边缘出现伪影。故通过定性与定量测试结果表明，在不同的分辨率模型以及相同分辨率模型情况下，本章提出的方法都是优于先前的方法。

为了突出本章提出的算法对小目标区域和弱纹理区域的捕捉能力，使用热力图映射的方式将小目标区域和弱纹理区域的定量误差定性展示出来。对被预测图像添加红色掩膜（便于突出热力值低的区域所以添加掩膜），并使用热力图将预测后的深度图的小目标区域和弱纹理区域深度的平均相对误差（Abs Rel）映射到被掩膜处理过的图像上（热力图蓝色映射代表相对均方误差最低）。选取表 1 中预测效果较好的两个模型 Sun[19]、Chen[21]与本章模型进行对比，如图 3-8 所示：

图 3-8 中，左起第二张图片，红绿灯灯牌区域，经热力图分析：相对于 Sun[19]和 Chen[21]，本章提出的模型热力图蓝色映射包含了 Sun[19]和 Chen[21]的蓝色映射，本文提出的模型在红绿灯区域平均相对误差（Abs Rel）最低。左起第四张图，道路上的汽车和围栏区域，经过热力图分析：相对于 Sun[19]和 Chen[21]，本章提出的模型热力图在汽车和栏杆区域蓝色映射最多，而 Sun[19]和 Chen[21]在汽车和围栏区域蓝色映射比较离散或较模糊，本文提出的模型在汽车和围栏区域平均相对误差（Abs Rel）最低。故通过热力图测试结果表明，本章提出的方法在小目标区域和弱纹理区域的精度都是优于先前的方法。

图 3-8 KITTI 数据集上小目标区域误差可视化分析

3.4.3 消融实验

为了分析本章提出的增强密集连接模块以及全局特征交互模块对深度预测能力的提升效果，以分辨率为 1024×2048 的 Cityscapes 数据集进行消融实验，使用不同的 Cityscapes 数据集的目的是对模型的泛化能力进行验证，其中表 3-2 表示了其定量分析，图 3-9 表示了其定性实验。

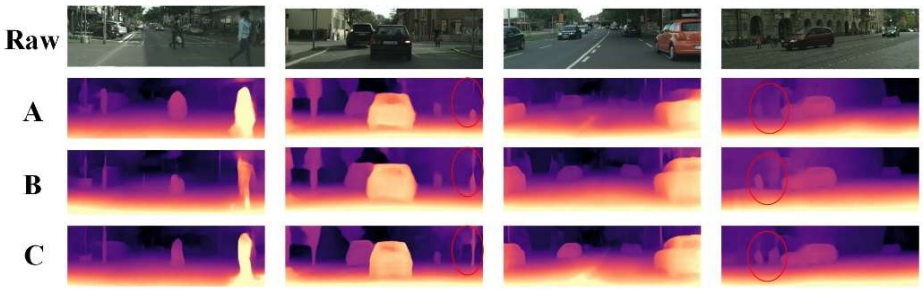
表 3-2 在 Cityscapes 数据集上消融实验测试结果对比

方法	增强密集连接模块	全局特征交互模块	误差指标（越小越好）				预测准确率（越大越好）		
			Abs Rel	Sq Rel	RMSE	RMSE log	$\delta<1.25$	$\delta<1.25^2$	$\delta<1.25^3$
Baseline			0.101	0.791	4.628	0.184	0.872	0.965	0.980
Ours	✓		0.096	0.771	4.331	0.185	0.876	0.967	0.981
		✓	0.097	0.782	4.493	**0.182**	0.879	0.969	0.979
	✓	✓	**0.095**	**0.716**	**4.178**	0.195	**0.881**	**0.972**	**0.982**

在 baseline（图 3 中的 Unet++拓扑结构）基础上，将 DepthNet 网络结构换为如图 3 所示的增强密集连接拓扑结构（Unet++拓扑架构+增强密集连接模块），进行消融实验，从表 3-2 结果中可以看出，网络在对弱纹理区域、小目标深度的提取精度上有了一定的提高。相较于 baseline，平均相对误差（Abs Rel），相对均方误差（Sq Rel）和对数均方根误差（RMSE）分别减少了 4%、3%和 6%。

在 baseline 上添加全局特征交互模块（Unet++拓扑架构+全局特征交互模块），即在 Unet++中加入全局特征交互模块，相较于 baseline，平均相对误差（Abs Rel），相对均方误差（Sq Rel）和对数均方根误差（RMSE）分别减少了 4%、1%和 3%，可见全局特征交互模块使用的“查询-键值对”注意力方式使得模型的鲁棒性和泛化能力进一步提高。

若在 baseline 的基础上同时使用增强密集连接模块和全局特征交互模块（Unet++拓扑架构+增强密集连接模块+全局特征交互模块）后，相较于 baseline，平均相对误差（Abs Rel），相对均方误差（Sq Rel）和对数均方根误差（RMSE）分别减少了 6%、9%和 10%，可见网络捕捉小物体以及边缘轮廓信息和整合的能力达到了最大化。



A：Unet++拓扑 B：Unet++拓扑 +增强密集连接模块 C：Unet++拓扑 +增强密集连接模块+全局特征交互模块

图 3-9 Cityscapes 数据集上的消融实验效果

消融实验的定性描述如图 3-9 所示：

1）左起第 2 张图片在汽车附近有电线杆，在 baseline 的方法中没有显示出电线杆的边缘信息；使用了增强密集连接模块后可显示出电线杆的边缘信息，但是也存在部分缺失；在既使用增强密集连接模块又使用全局特征交互模块后，可以清楚的看到电线杆边缘，清晰明显没有边缘缺失。

2）左起第 4 张图片在汽车前方有自行车和行人，在 baseline 方法中可以模糊的捕捉到行人，但伪影较多，没有捕捉到自行车深度；使用了增强密集连接模块后可显示出行人，且伪影比 baseline 的方法少，但是没有捕捉到自行车的深度；在既使

用增强密集连接模块又使用全局特征交互模块后，可以清晰的看到行人和自行车的边缘。

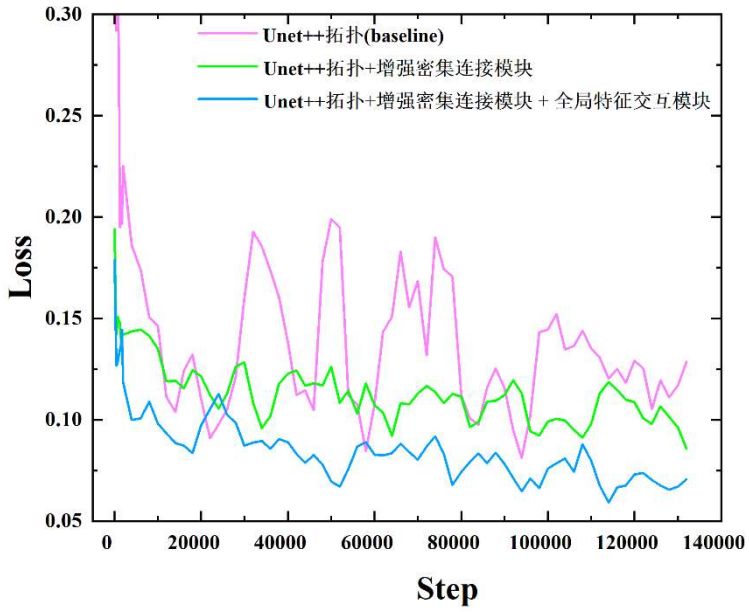


图 3-10 使用增强密集连接模块和全局特征交互模块网络损失变化情况

消融实验训练过程中损失函数变化如图 3-10 所示。根据图 3-10，仅使用 baseline 方法时容易出现梯度陡增现象，网络学习不稳定鲁棒性差；在 baseline 基础上增加增强密集连接模块后网络学习变得平顺但是依然出现收敛不均匀现象；使用增强密集连接和全局特征交互模块后，收敛梯度变化更均匀，网络学习更平顺鲁棒性更佳。

综上，由消融实验可表明，本章提出的增强密集连接模块和全局特征交互模块对整个网络的识别精度都是有效的，将两个结合在一起取得了最好的结果。

3.5 本章小结

为解决当前深度图中出现的边缘缺失，伪影过多、模型的泛化能力和鲁棒性较差等问题，本章提出了一种基于增强密集连接和全局特征交互的单目深度估计方法。该方法的核心基于 Unet++网络结构进行了两点改进包含：

1）提出了一种增强密集连接模块，该模块中引入了 CBAM 注意力机制，该结构让网络更好地利用不同层采样的尺度特征，从不同尺度上捕捉图像的深度语义信息，加强了网络对小目标深度的捕捉。

2）提出了基于自注意力机制的全局特征交互模块，增强了模型对不同的分辨

率和多纹理图像深度估计的鲁棒性。

在 KITTI 数据集和 Cityscapes 数据集上进行的测试结果表明：

1.本模型相较于其它模型有更细致的深度边缘和更强捕获小目标能力，均方根误差相较于目前最先进算法减少了 7%。

2.提出的各个加入组分对整个网络深度估计性能均有所提高，特别是增强密集连接模块和全局特征交互模块让模型均方根误差减少了 4%，捕捉到的多尺度分辨率对象也更加清晰。

3.根据训练过程中损失函数的变化，表明本章提出的增强密集连接模块和全局特征交互模块有更好的收敛效果，网络收敛更平稳，鲁棒性更佳。

第 4 章 针对动态场景下的基于边缘增强的无监督单目深度估计

4.1 引言

估计动态场景中的深度一直以来都是有挑战性的课题，这也是与专业深度传感器之间精度产生较大差距的来源。最近也有一些工作提出了从单目视频序列中学习场景的深度、相机的运动和对象运动的若干方法。Yin[44]等人提出了 GeoNet，通过神经网络联合预测了深度、自我运动和光流等信息，他们的方法分为了两个阶段，第一个阶段估计深度和相机运动，第二阶段根据对象相对于场景的运动导致的残余光流来掩蔽移动对象。Luo[50]等人使用整体运动解析器来联合优化深度、相机运动和光流估计，其立体图象对的输入可以消除深度估计过程中产生的歧义。Casser[45]等人提出了在预先训练的分割模型的帮助下估计场景中对象的运动，显著改进了移动对象的深度估计。Godard[17]等人提出了处理遮挡的新的方法将那些可能移动的对象进行人为的分割，估计效果有一定的提高但是工作量非常大。Li[51]等人对 Godard[17]的方法进行改进提出了 DepthMotion 算法，使其不需要对移动对象进行分割，而是用一种运动平移场去定位那些可能移动的对象，但是不足的地方是由于平移场有些时候会产生歧义有些运动对象会被屏蔽。

虽然以上方法看起来非常有前景，但它们的深度估计性能尚未能与一些专业测量深度信息的传感器相提并论。为了尽可能减小与专业传感器测量结果的差距，经过大量测试观察，发现大多数传统无监督单目深度估计方法中存在三个明显的缺陷：1）由于运动对象自身的轨迹是不断变化的，导致运动对象的纹理等特征难以被神经网络有效捕捉，所预测的深度图会产生很多伪影和深度边缘模糊等现象；2)在解码器解码过程中大多使用比较常见的方形卷积核，导致在对预测对象上采样过程中边缘附近处的帧没有被充分计算，编码器对浅层特征的细节信息并没有充分利用和部分离散运动对象没有被捕捉，最终预测对象深度图的边缘产生缺失等现象；3)由于网络整体过于复杂使得网络收敛较慢。在单目相机深度估计的过程中，这些因素严重影响了深度估计的精度。因此，本文引入了几个创新，以解决由对象的自身运动和网络卷积核结构所造成的深度估计精度差的问题：1)在运动估计网络中添加卷积注意力模块 CBAM[23]，CBAM 注意力机制可以帮助网络自动调整对运动对象的关注度，减少对运动背景或其他干扰因素的敏感性，从而使被估计对象的深度图边缘更加清晰。2)在运动估计网络解码阶段引入了一种新的卷积方法即通过将传统

的方形卷核积替换为条状卷积核，使得运动网络更加有效地提取边界和全局信息以及捕捉到离散的运动对象。3)在损失函数里结合高斯拉普拉斯算子提出边缘正则化，以加快网络收敛。

4.2 网络结构及其原理

本章设计的网络总体框架如图 4-1 所示，整个动态场景下的基于边缘增强的无监督单目深度估计网络分为三个阶段：深度预测阶段、运动估计阶段和图象重构阶段。第一阶段深度估计阶段，主要目的是粗略估计深度图；第二阶段运动估计阶段，估计出相机的内参以及提取视频帧中出现的运动信息；第三阶段是对图象的重构。下面对各个阶段工作进行详细介绍。

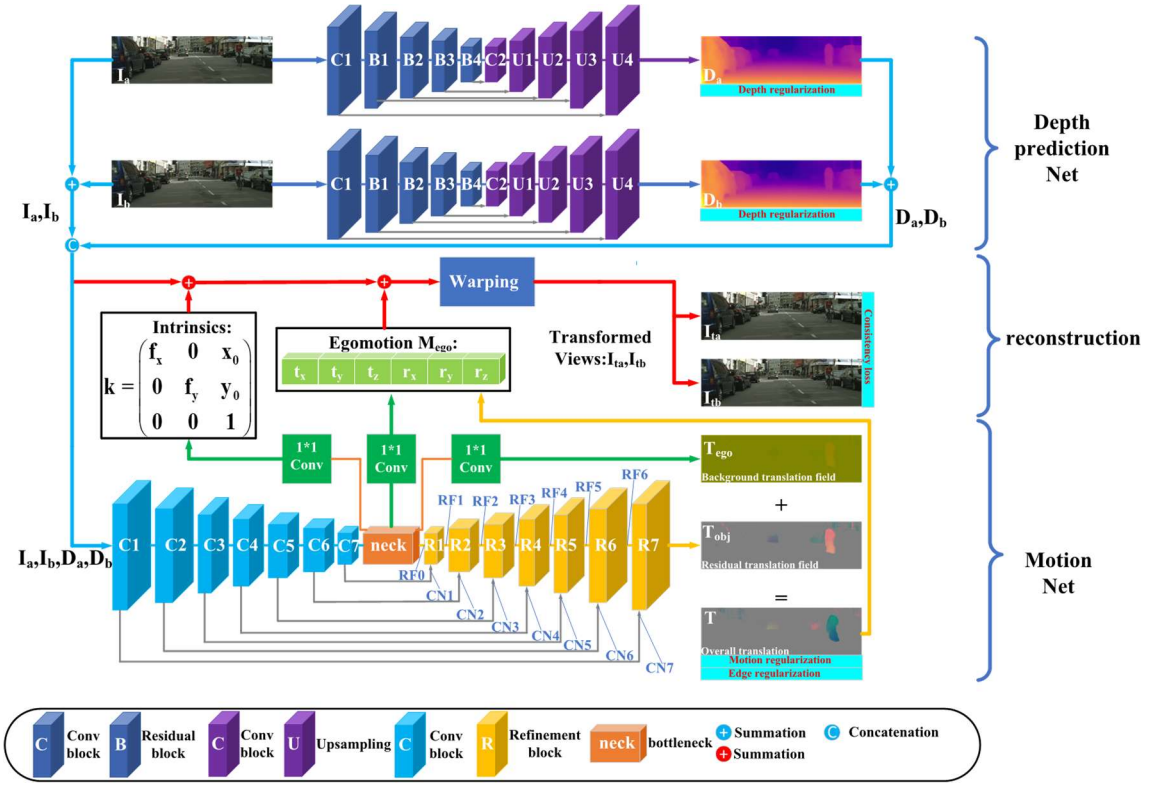


图 4-1 深度估计网络结构图

4.2.1 深度预测阶段

深度预测阶主要是深度预测网络对连续帧图象深度的学习，深度预测网络以两张连续帧 RGB 图象I_a、I_b作为输入，输出相应图象的深度图D_a、D_b，如图 4-1 所示 Depth prediction Net。深度预测网络是基于 ResNet18[26]的 U-Net[29]结构，采用

ResNet18 作为编码器，通过 ResNet18 深度叠加的卷积块对图象特征进行高度压缩获得初步特征空间，其大小为原特征空间的 1/16，其中蕴含了丰富的深度信息；深度预测网络的解码阶段与 U-Net 解码阶段相似是加强特征提取阶段，利用先前编码主干部分获取到的初步有效特征层进行上采样，并且进行特征融合。最终获得一个融合所有深度特征的有效特征层，对最终有效特征层的深度信息进行归类获得初步预测的深度图D_a、D_b。

4.2.2 运动估计阶段

运动估计阶段主要是运动网络对连续帧图象中出现的动态信息的学习，运动网络除了要以一对连续帧 RGB 原图象作为输入外，还要 concate（拼接）D_a、D_b的深度信息通道，即除了 RGB 三个通道之外还有一个包含深度信息的通道，总共要输入四个通道的图象信息，如图 4-1 所示 Motion Net。运动网络输出为相机内参K、相机的自我运动M_{ego}、背景平移场T_{ego}和残差平移场T_{obj}。运动网络是基于 Flow-Net[32]的 U-Net 结构。在编码阶段，输入四个通道的图象信息后经过七次卷积下采样提取运动对象自我运动的有效特征，然后再经过 bottle neck 模块处理掉高频运动冗余噪声，最后通过引出三个1×1卷积预测出相机的内参K、相机自我运动M_{ego}和背景平移场T_{ego}。其中，相机的自我运动M_{ego}是相机相对于背景的运动，由相机自身相对于背景的旋转及平移构成，包含 SE3 变换[14]的平移向量$t=[t_x,t_y,t_z]$和旋转矩阵$R=[r_x,r_y,r_z]$；背景平移场T_{ego}是由相机相对于背景的运动向量构成。为了消除预测出来的运动参数之间的歧义（运动歧义证明[47]），还需要知道相机与运动对象之间的相对运动关系。可以将相机相对于背景的运动T_{ego}关联上运动对象相对于背景之间的运动来得到相机与运动对象之间的运动关系。为此，在解码阶段对预测出来的T_{ego}进一步学习，对编码阶段的七个包含粗略运动信息的有效特征层进行特征细化及上采样，并且进行特征融合，最终融合了所有的特征的有效特征层就是需要得到的运动对象相对于背景之间的运动，将运动对象相对于背景之间的运动称为残差平移场T_{obj}，将相机相对与运动对象之间的运动关系称为总的运动场T。其中，每一步的特征细化都是通过细化模块（Refinement block）来执行的。背景平移场T_{ego}、残差平移场T_{obj}和总的运动场T存在以下数学关系：

$$T_{ego}(u,v)+T_{obj}(u,v)=T(u,v) \tag{4-1}$$

其中，(u, v)为像素坐标中的点。如图 4-2 是本文在 KITTI 数据集和 Cityscapes 数据集上测试得到的深度图和残差平移场。

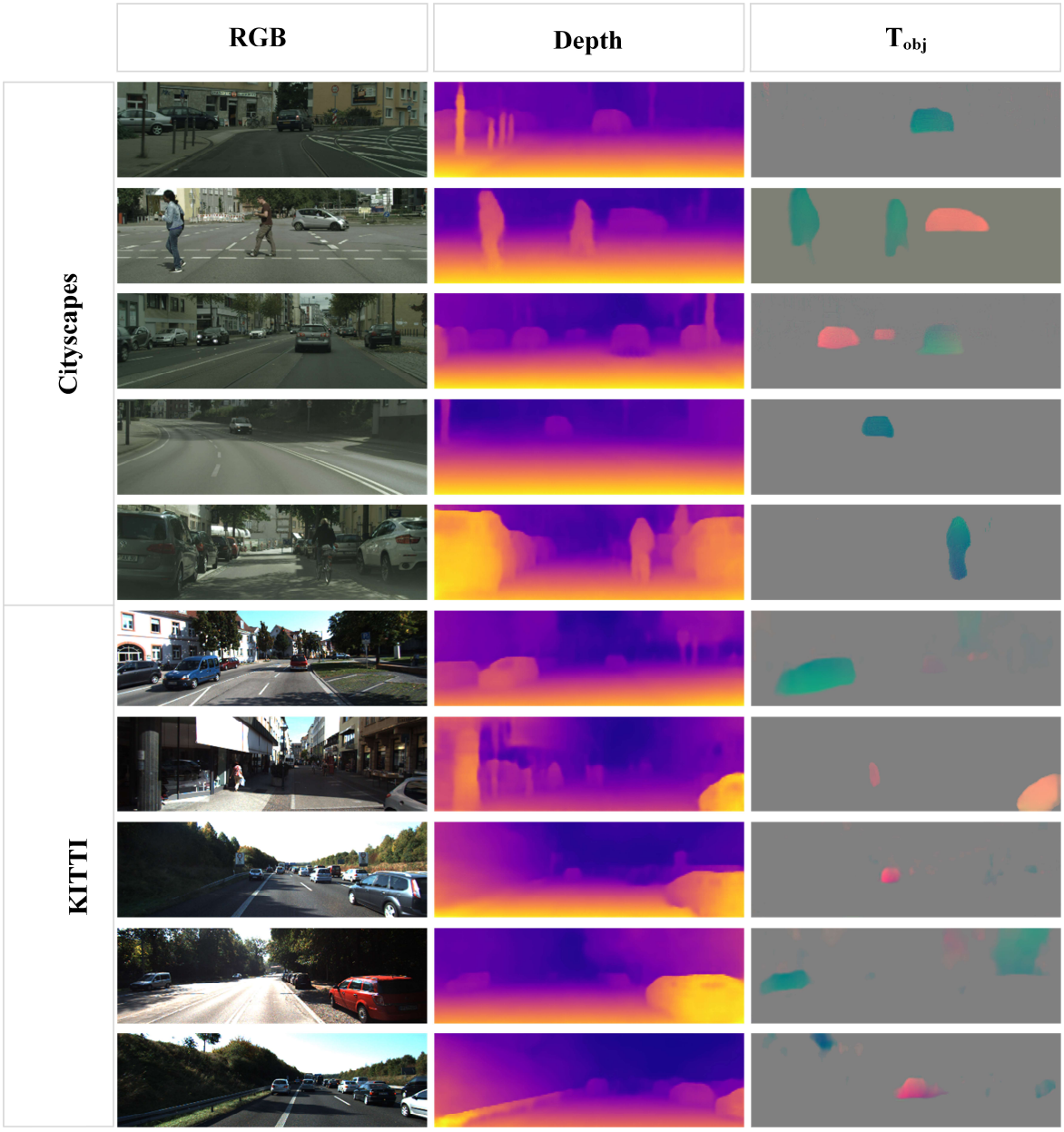


图 4-2 在 KITTI 数据集和 Cityscapes 数据集上测试得到的深度图和残差平移场

4.2.3 图象重构阶段

连续帧图象重构原理与 Zhou[18]方法类似，本章方法主干是使用深度图和相机矩阵将两个相邻帧联系在一起：

$$z'p' = KRK^{-1}zp + Kt \qquad (4-2)$$

其中K表示相机内参矩阵：

$$K=\begin{pmatrix} f_x & 0 & x_0 \\ 0 & f_y & y_0 \\ 0 & 0 & 1 \end{pmatrix} \tag{4-3}$$

f_x、f_y、x_0、y_0表示相机内部的具体内参；t和R分别是 SE3 变换的平移向量和旋转矩阵；P和P'是在由旋转矩阵R和平移向量t表示的变换之前和之后的齐次像素坐标；z和z'分别代表变换前后像素的对应的深度。在深度预测网络中可以得到参数z，在运动网络中可以得到R、t和K，式（1）表示将一个视频帧扭曲（warping）到另一个相邻的帧上，扭曲之后的帧将构成新的视频图象。如图 4-1 所示，输入的一对连续帧I_a和I_b通过扭曲分别重构成了图象I_{tb}和I_{ta}。最后将重构图象的帧与图象的实际帧进行比较，重构后的图象I_{ta}和I_{tb}分别会与I_b和I_a进行比较，比较后的差异构成训练损失的主要分量，通过惩罚差异，网络将学会正确预测z、K、R和t。

4.3 运动网络边缘强化解码器

本节将详细介绍边缘强化解码器其网络内部各节点的结构，阐述网络内部各模块及其组分的具体构建方法，并通过输出特征图分析各组分的贡献。

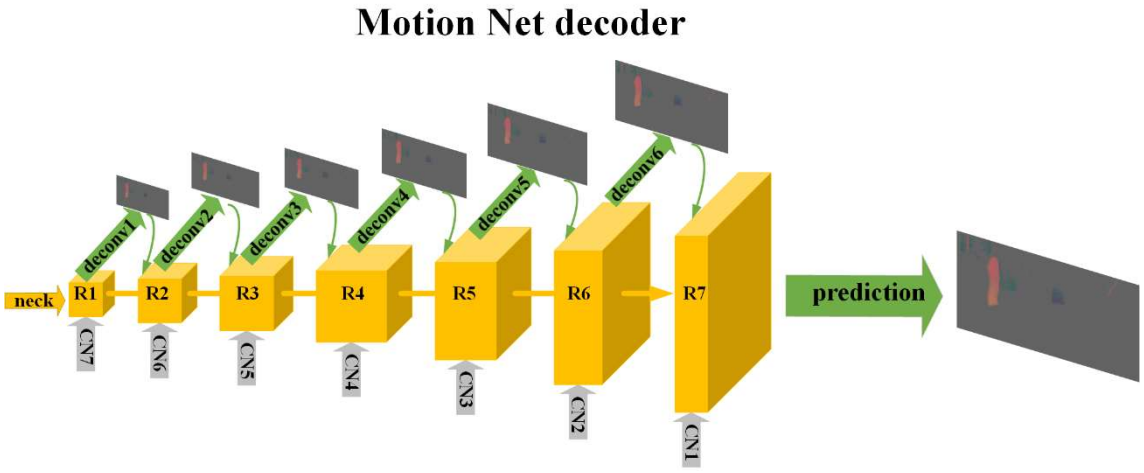


图 4-3 运动网络解码器结构

4.3.1 边缘强化解码器总体架构

解码器的具体结构如图 4-3 所示，解码器结构主要由 7 个 Refinement（细化模块）级联而成。由于卷积神经网络大多数通过卷积层和池化层串联来提取图象的高

级特征，即在空间上收缩特征图。池化对于网络训练在计算上可行是必要的，并且更重要的是允许在输入图象的大区域上聚合信息。然而，池化会导致分辨率降低，因此为了提供密集针对逐像素的预测，需要一种方法来细化粗略的池化表示。因此在上采样过程中，将 Refinement 部分与编码阶段每一层卷积的输出特征融合在一起。这样，既保留了从较粗的特征图传递的高级信息，又保留了较低层特征图中提供的精细局部信息。

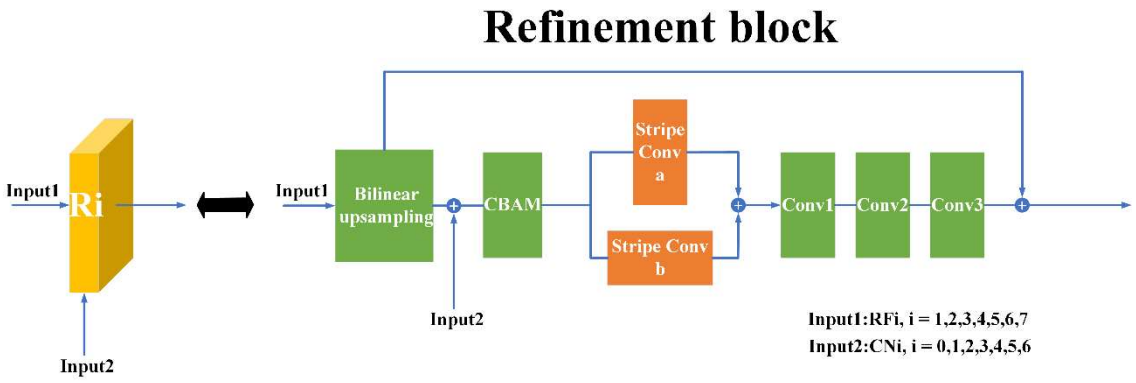


图 4-4 边缘强化细化模块

对于每一个 Refinement 模块的结构如图 4-4 所示，分为四步：1)为了尽可能保持输入特征图的在不引入大量失真的情况下改变图像分辨率，在第一步需要对输入的特征图进行双线性插值操作。2)第二步经过 CBAM 卷积注意力机制[23]，通过结合空间注意力和通道注意力使神经网络关注特征图中的不同和位置信息，来加强对特征图中运动信息的提取。3)第三步经过条状卷积模块，提取离散的运动对象以及加强对边缘语义信息的提取。4)第四步，将获取的边缘信息输入三层卷积并且对第一步插值后的特征进行融合输出细化的特征图。

4.3.2 CBAM 卷积注意力机制

本章使用结合通道注意力和空间注意力的 CBAM（Convolutional Block Attention Module）卷积注意力机制来作为边缘强化解码器的重要组成部分。CBAM 注意力机制的主要目标是增强卷积神经网络对于不同特征通道和空间位置的感知能力，从而提升网络的表现。CBAM 总体结构，如图 4-6：输入一个中间特征图，然后沿着两个独立的维度（通道维度和空间维度）依次推断注意力图，然后将注意力图乘以输入特征图以对特征进行自适应修饰。

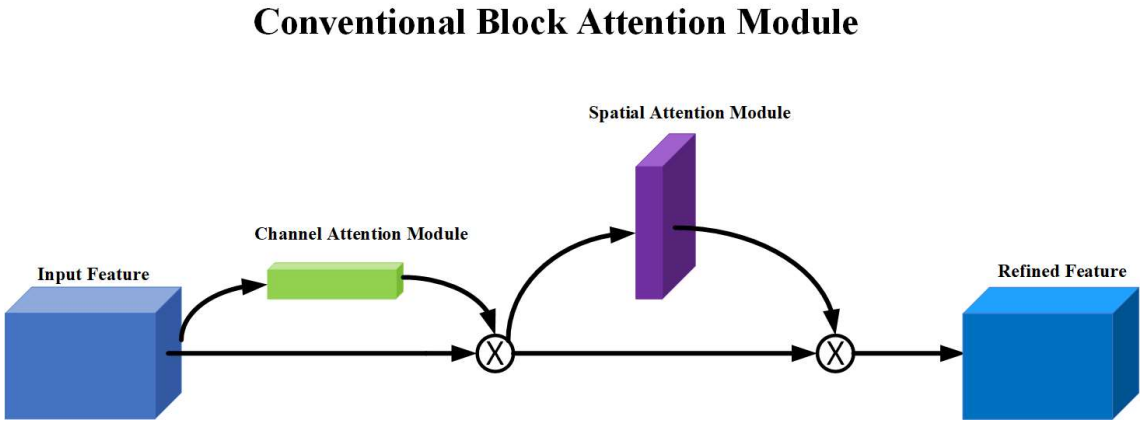


图 4-6 卷积注意力机制总体结构

对于通道注意力，如图 4-7：第一步输入大小为$H \times W \times C$的特征图，经过两个并行的 MaxPool（全局最大池化）层和 AvgPool（全局平均池化）层，将特征图从大小为$H \times W \times C$变为$1 \times 1 \times C$；第二步经过 Share MLP（多层共享神经网络）模块，将通道数压缩为原来的$1/r$（r为减少率），再扩张至原通道数，经过 ReLU 激活函数得到两个激活后的特征；最后，将激活后的特征进行元素逐行相加，再通过 sigmoid 激活函数非线性处理，得到通道注意力的输出结果$M_c \in \mathbb{R}^{C\times1\times1}$。定义为：

$$M_c(F) = \sigma\left(MLP(AvgPool(F)) + MLP(MaxPool(F))\right) = \sigma\left(W_1\left(W_0(F_{avg}^c)\right) + W_1(W_0(F_{max}^c))\right) \quad (4-4)$$

其中，σ表示 sigmoid 激活函数；$W_0 \in \mathbb{R}^{C\times C/r}$表示 MLP 的权重，$W_1 \in \mathbb{R}^{C\times C/r}$表示 ReLU 激活函数的权重；$F_{avg}^C$和$F_{max}^C$表全局平均池化特征和全局最大池化特征。

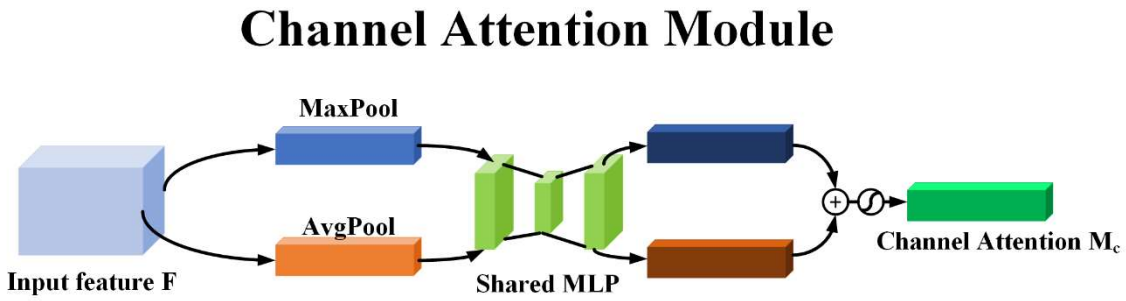


图 4-7 通道注意力机制结构

对于空间注意力，设通道注意力输出的特征图为F'，如图 4-8：第一步，输入特征图F'，依次经过 MaxPool（全局最大池化层）和 AvgPool(全局平均池化层)，得到两个大小为$H \times W \times 1$的特征图；第二步，将得到的两个特征图进行 concat（通道拼接）操作，再经过一个7×7卷积操作，降维成 1 个通道；最后，将降维的特征通

过 sigmoid 激活函数，将输出结果乘以原图变回$H \times W \times C$的大小，最终得到的空间注意力的输出结果为$M_s \in \mathbb{R}^{C\times H\times W}$。空间注意力公式：

$$M_s(F) = \sigma\left(f^{7\times7}([AvgPool(F); MaxPool(F)])\right) = \sigma\left(f^{7\times7}\left([F^s_{avg}; F^s_{max}]\right)\right) \quad (4-5)$$

其中，σ表示 sigmoid 激活函数；f表示尺寸大小为7×7的卷积运算。

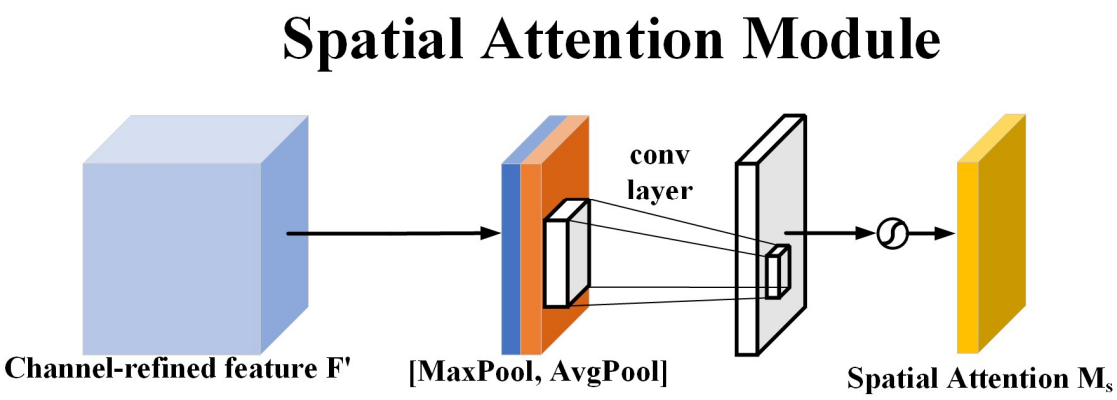


图 4-8 空间注意力机制结构

综上，用公式表述整个 CBAM 卷积注意力机制过程：

$$F' = M_c(F) \otimes F \quad (4-6)$$

$$F'' = M_s(F) \otimes F' \quad (4-7)$$

其中，$\otimes$表示逐元素乘法；$F' \in \mathbb{R}^{C\times1\times1}$表示经过通道注意力之后得到的特征，$F'' \in \mathbb{R}^{C\times H\times W}$是中间特征$F$输入 CBAM 注意力之后得到的最终特征。

图 4-5 边缘强化模块各组分可视化结果，（a）输入 RGB 图象、（b）初步提取特征、（c）经过 CBAM 后的特征、（d）经过 CBAM 和条状卷积后的特征

从图 4-5（c）和（b）可以看出，经过 CBAM 注意力模块后，解码器可以明显地分割出运动对象，也能够更易明确运动对象和背景之间的相对位置关系，出现的伪影也更少，边缘信息更加明确。

4.3.3 条状卷积

传统的卷积核大多使用正方形卷积核，无论是在分类任务还是语义分割任务以

及深度估计任务中，这种卷积核只对每一个像素的局部特征进行聚合而并没有完全利用边界和全局上下文特征，同时这也限制了神经网络在捕捉不同对象相对背景运动时存在的各向异性。

图 4-9 方形卷积对运动对象的捕获

如图 4-9 所示，对于此场景中离散出现的两个骑自行车的行人，使用方形卷积就容易让卷积核采样附近与行人无关的特征，加权合并之后容易让预测出的平移场中出现的运动对象边缘出现模糊的不清晰的现象。如图 4-10 所示，与图 4-9 是同样的场景，不一样的是这次采样使用的是条状卷积，条状卷积可以偏向于特定的方向来部署条形卷积核，在相同卷积核尺寸下，条状卷积可以捕捉远距离相对位置关系实现更大的感受野；沿其他空间维度保持窄的内核形状，在对图片采样提取特征过程中有助于联系上下文信息并且有效防止在不相干的区域出现采样干涉的情况。所以条状卷积可以作为传统方形卷积一个非常好的补充，帮助运动网络提取更细致和更离散的运动对象的信息。因此，本章为了增强对离散运动对象的捕捉和对运动对象边缘信息的提取在运动网络解码过程中采取了条状卷积。

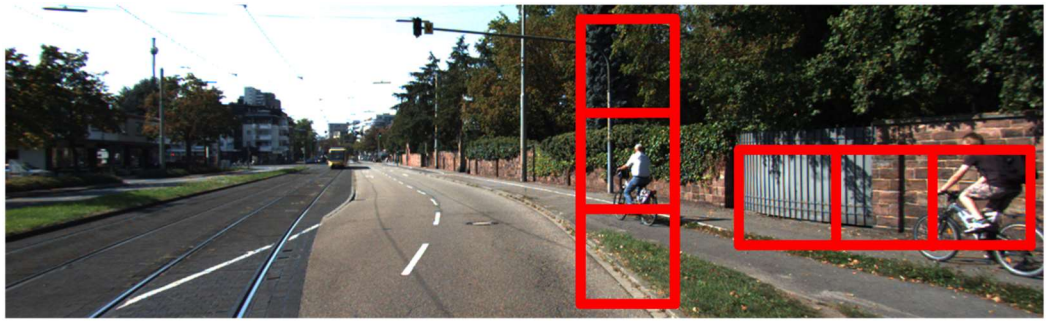


图 4-10 条状卷积对运动对象的捕获

条状卷积具体的细节如图 4-11 所示，其中 a 为 11×3 卷积层，b 为 3×11 卷积层，令卷积 a 在水平方向，卷积 b 在垂直方向分别大范围聚合边界附近的像素，然后进行元素级的 add 操作(像素级的叠加)来融合两条状卷积提取的特征，为保证融合特征的多样性，条状卷积后不接激活函数。由于条状卷积 a 和 b 是沿着正交方向对全局

上下文信息采样，这就为增强运动对象的边缘信息和判断离散运动对象的捕获提供重要线索，因此通过该条状卷积可以有效的增强对运动对象的边缘信息和运动信息的获取。

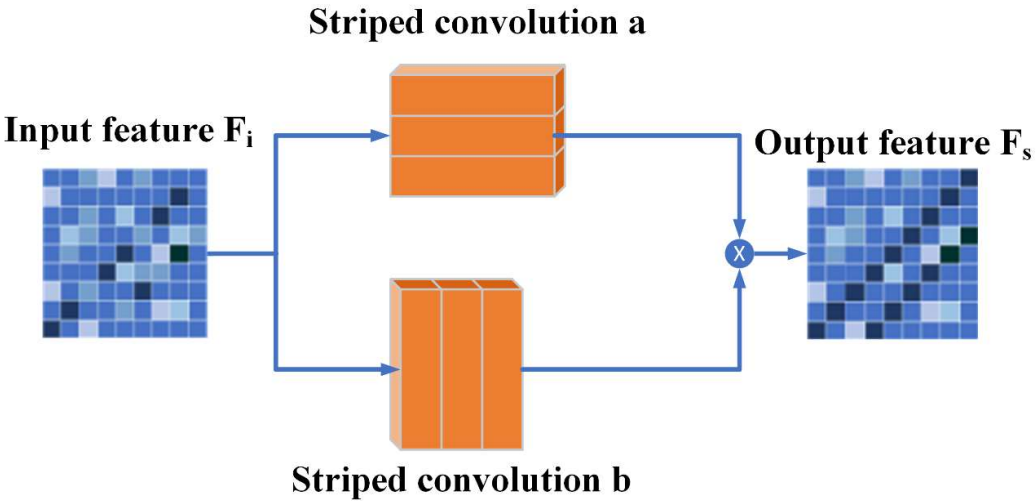


图 4-11 条状卷积细节

从图 4-5（d）和（c）可以看出，经过条状卷积后，能够明确区分离散对象，如从图象右端倒数第二辆车在图（d）中更加清晰。同时，车辆轮廓也要比图（c）更加清晰，边缘细节信息更加丰富。

4.4 损失函数

损失函数作为单目相机无监督深度学习的重要的组成部分，是提供训练监督信号的唯一来源。本章损失函数主要包含了四个正则化分量：1、深度正则化$L_{reg,dep}$。2、运动正则化$L_{reg,mot}$。3、一致性正则化L_{cyc}。4、边缘强化正则化L_{edge}。

深度正则化的目的是规范低梯度区域的深度，定义如下：

$$L_{reg,dep} = \alpha_{dep} \iint \left(|\partial_u D(u,v)|e^{-|\partial_u I(u,v)|} + |\partial_v D(u,v)|e^{-|\partial_v I(u,v)|}\right)dudv \qquad (4-8)$$

其中，α_{dep}是一个超参数，$D(u,v)$表示预测出来的深度图，$I(u,v)$表示输入的 RGB 视频帧，$\partial D(u,v)$表示图象深度的梯度，$\partial I(u,v)$表示图象像素的梯度。

运动正则化，根据残差平移场的两个性质：（1）稀疏性，因为通常一帧中的大多数像素都属于背景或静态对象。（2）三维空间中，整个刚性移动物体在平移过程中形状往往是恒定的。故运动场$T(u,v)$上的运动正则化$L_{reg,mot}$包括组平滑损失L_{g1}和稀疏性损失$L_{1/2}$组成，定义为：

$$L_{g1}[T(u,v)] = \sum_{i\in\{x,y,z\}} \iint \sqrt{\left(\partial_u T_i(u,v)\right)^2 + \left(\left(\partial_v T_i(u,v)\right)^2\right)}\, dudv \qquad (4-9)$$

其中，$\partial T(u,v)$表示运动场运动向量的梯度。

稀疏性损失$L_{1/2}$定义为：

$$L_{1/2}[T(u,v)] = 2\sum_{i\in\{x,y,z\}} \langle|T_i|\rangle \iint \sqrt{1+|T_i(u,v)|/\langle|T_i|\rangle}\, dudv \qquad (4-10)$$

其中，$\langle|T_i|\rangle$是$|T_i(u,v)|$的空间平均值，这种正则化是自归一化。最终的运动正则化损失：

$$L_{reg,mot} = \alpha_{mot} L_{g1}\left[T_{obj}(u,v)\right] + \beta_{mot} L_{1/2}\left[T_{obj}(u,v)\right] \qquad (4-11)$$

一致性正则化，一致性正则化包含两个分量运动周期一致损失L_{cyc}和遮挡和光度一致性损失L_{rgb}，且L_{cyc}鼓励任何一对帧之间的向前和向后运动彼此相反：

$$L_{cyc} = \alpha_{cyc}\frac{\|RR_{inv}-1\|}{\|RR-1\|^2+\|RR_{inv}-1\|^2} + \beta_{cyc}\iint \frac{\left\|R_{inv}T(u,v)+T_{inv}(u_{warp},v_{warp})\right\|}{\|T(u,v)\|^2+\left\|T_{inv}(u_{warp},v_{warp})\right\|^2} dudv \qquad (4-12)$$

其中“inv”下标表示输入帧按顺序反转的情况下获得的相同的量，“warp”下标表示 warp 操作；α_{cyc}和β_{cyc}表示超参数。L_{rgb} 鼓励光度一致性包括 RGB 空间中的L_1损失和 SSIM 损失：

$$L_{rgb} = \alpha_{rgb}\iint \left|I(u,v)-I_{warp}(u,v)\right|_{D(u,v)>D_{warp}(u,v)} dudv + \beta_{rgb}\frac{1-SSIM\left(I,I_{warp}\right)}{2} \qquad (4-13)$$

其中，I和I_{warp}是原始图象和重构图象。

前三个正则化大部分化沿用了以前的工作，本章提出了一种边缘正则化。边缘正则化是基于拉普拉斯算子改进而来，其具有旋转不变性的特性，且可以通过图象梯度来寻找边缘的强度和方向。对于所预测的平移场，边缘正则化可以很好的捕捉到运动对象的边缘强度，加快了神经网络的收敛速度，公式表述如下：

$$L_{edge}[T(u,v)] = \sum_{i\in\{x,y,z\}} \iint \sqrt{(\partial^2 T_i(u,v)/\partial^2 u^2)^2 + (\partial^2 T_i(u,v)/\partial^2 v^2)^2}\, dudv \qquad (4-14)$$

其中，$\partial^2 T(u,v)$表示求运动场中运动向量的梯度的变化。

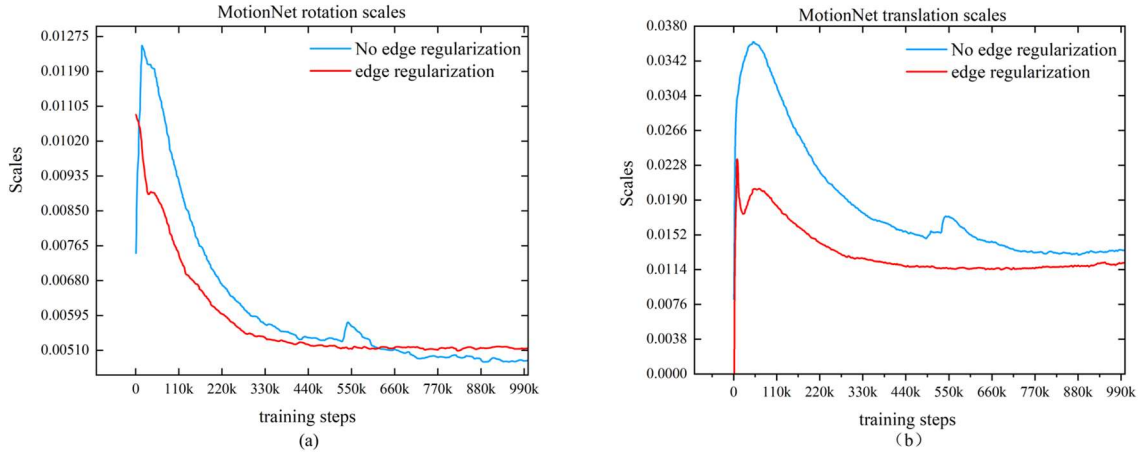


图 4-12 训练过程运动网络尺度因子变化

当施加深度正则化、运动正则化、一致性正则化和在前三个正则化基础上继续施加边缘正则化后，随着训练 steps（轮数）的增多旋转和平移尺度因子的变化，如图 4-12（a）和图 4-12（b）所示。无论对于旋转尺度因子还是平移尺度因子，添加边缘正则化后旋转尺度因子和平移尺度因子在 440k 左右 steps 后收敛，而没有添加边缘正则化旋转尺度因子和平移尺度因子在 770k 左右 steps 后收敛，添加边缘正则化后收敛速度提高了 42%。同时，没有添加边缘正则化的运动网络，容易在 550k 左右 steps 出现梯度陡增现象，网络学习不平稳，而添加边缘正则化后梯度变化更加均匀，网络学习更加平稳，网络收敛效果更佳。

综上所述整个网络的损失函数为：

$$L_{total} = L_{reg,mot}\left(L_{g1}(group\ smoothnessloss) + L_{1/2}(sparsity\ loss)\right) + L_{reg,depth} + L_{cyc}(rotation + translation) + L_{rgb}\left(rgb_{consistency} + ssim\right) + L_{edge} \quad (4\text{-}15)$$

4.5 实验

4.5.1 实施细节

本章使用 Tensorflow 框架实现了所提整个网络结构，在 NVIDIA GeForce RTX3070 上训练了该模型。以 Adam 作为优化器对深度预测网络和运动估计网络 10^6 次迭代训练，优化器的学习率设置为 10^{-4}。训练的批量大小为 16，训练的 RGB 图象分辨率大小为 1248×128。损失损失函数中几个超参数的设置：4-8 式中，α_{dep}为 10^{-3}；4-11 式中，α_{mot}和β_{mot}分别设置为 1.0 和 0.2；4-12 式中，α_{cyc}和β_{cyc}分别设为 10^{-3} 和 5×10^{-2}，4-13 式中α_{rgb}和β_{rgb}分别设置为 1.0 和 0.8。

4.5.2 实验结果分析

为证明对本章提出的改进方法的有效性和先进性进行验证，在 KITTI 数据集和 Cityscapes 数据集上将本章方法的实验结果与其他无监督单目深度估计方法进行对比，定量结果分析如表 4-1 和所示，定性结果分析如图 4-13 所示。

表 4-1 在 KITTI 数据集上的测试结果对比

方法	分辨率	运动模型	误差指标（越小越好）				预测准确率（越大越好）		
			Abs Rel	Sq Rel	RMSE	RMSE log	$\delta<1.25$	$\delta<1.25^2$	$\delta<1.25^3$
Zhou[18]	128×416		0.208	1.768	6.856	0.283	0.678	0.885	0.957
GeoNet[44]	128×416	✓	0.155	1.296	5.857	0.233	0.793	0.931	0.973
DDVO[43]	128×416	✓	0.151	1.257	5.583	0.228	0.810	0.936	0.974
Casser[45]	128×416	✓	0.141	1.026	5.291	0.215	0.816	0.945	0.979
Ranjian[46]	256×832		0.148	1.149	5.464	0.226	0.815	0.935	0.973
Gordon[47]	128×416	✓	0.128	0.959	5.230	0.212	0.845	0.947	0.976
Yang[52]	384×512		0.131	1.254	6.117	0.220	0.826	0.931	0.973
EPC++[49]	256×832		0.141	1.029	5.350	0.216	0.816	0.941	0.976
Li[51]	128×416	✓	0.130	0.950	5.138	0.209	0.843	0.948	0.978
Luo[50]	112×384	✓	0.130	1.086	4.876	**0.205**	0.878	0.946	0.970
Ours	128×416	✓	**0.119**	**0.916**	**4.778**	0.227	**0.881**	**0.952**	**0.982**

其中，运动模型一栏，“✓”表示使用运动模型；粗体标注为在使用运动模型的情况下各项评估指标的最优结果。从表 4-1 可以看出：1、在都使用运动模型的情况下，分辨率不同的情况下，与最优的方法对比，本章提出的模型在平均相对误差、相对均方误差、均方根误差相对于前者至少提高了 8%，15%，2%。2、在相同的分辨率情况下、与没有使用运动模型的情况下，与最优的方法对比，本章提出的模型在平均相对误差、相对均方误差、均方根误差相对于前者分别提高了 4%，4%，3%。3、在相同的分辨率情况下、且都使用运动模型的情况下，与最好的方法对比，本章提出的模型在平均相对误差、相对均方误差、均方根误差上相对于前者分别提高了 7%，4%，7%。

为了更直观地定性表明本模型与其他模型效果对比，抽取了表 4-1 中的部分模型进行复现与本章的方法进行对比，如图 4-13 所示。其中 Raw 为测试图象，Zhou[18]、GeoNet[44]、DDVO[43]、Yang[52]、Li[51]为与表 4-1 对应的模型预测出的深度图，其中用红色圆圈所框出来的部分是需要重点观察的部分。在第二张测试图中(从左到右)，注意观察房屋前面的大树部分，本文预测出的模型可以清晰的捕捉到大树

的树干部分并可以与旁边的广告牌杆加以区分，而其他模型区分的比较模糊。在第四张测试图中，汽车附近的三角路牌，本文预测出来的深度图可以捕捉到三角路牌的边缘轮廓，而其它测试模型并不能区分出三角路牌轮廓或比较模糊。故通过定性与定量测试结果表明，在不同的分辨率以及与无论是否应用动态模型，本章提出的方法都是由于先前的方法，能够实现深度估计质量的整体提升。

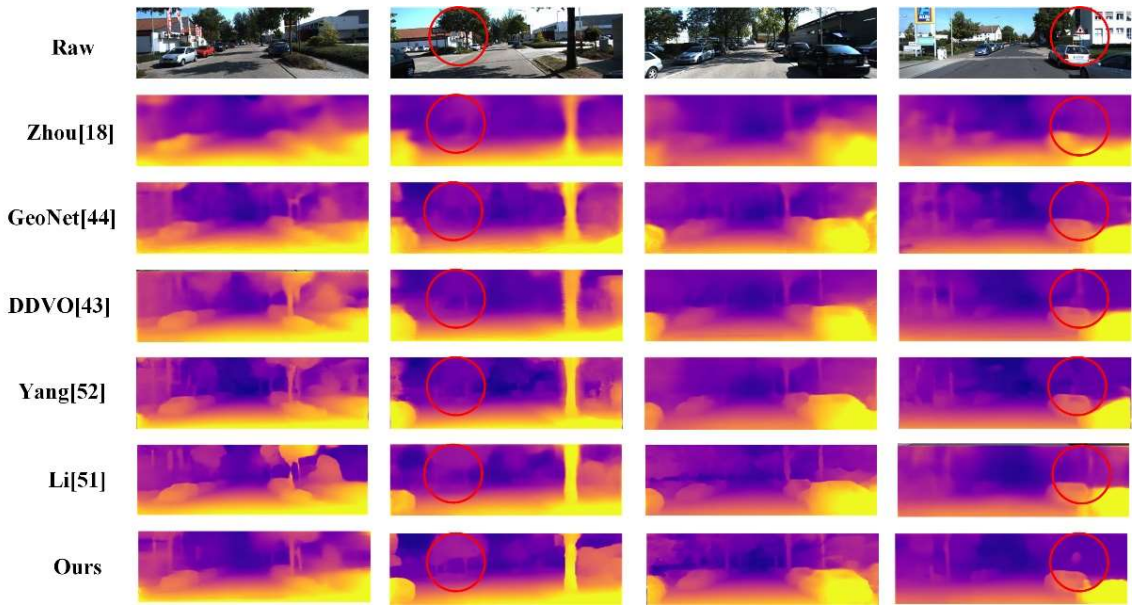


图 4-13 与表 1 结果对应的部分方法在 KITTI 数据集上的定性结果分析

4.5.3 消融实验

为了验证本章提出的边缘强化解码器对提升深度预测能力的有效性，以 128×416 分辨率为例，通过消融实验进行对比，分析了边缘强化解码器主要构成组分 CBAM 注意力模块和条状卷积模块对预测结果的影响。

在 baseline（基准）上添加条状卷积后，由于网络充分利用边界和全局上下文特征，对离散运动对象的捕捉也得到了加强，从表 4-2 中可以看出，网络在精度上有了一定的提高。其中，平均相对误差，相对均方误差和对数均方根误差分别提高了 3%、6%和 2%。在 baseline 上添加 CBAM 注意力模块后，由于 CBAM 注意力模块增强了卷积神经网络对于不同特征通道和空间位置的感知能力，同时加强了运动网络对于运动对象的分割和明确了运动对象与背景之间的空间关系，从表 4-2 中可

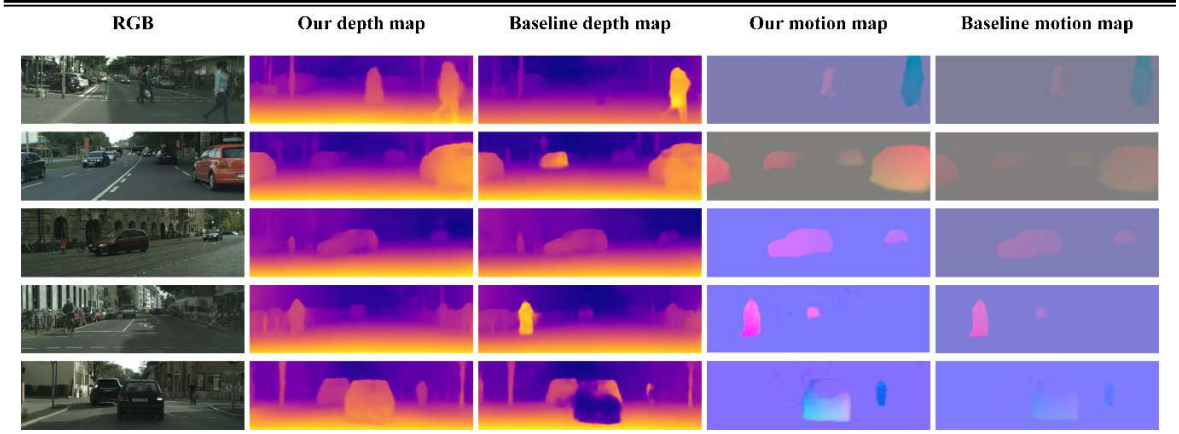


图 4-14 使用边缘强化解码器与没有使用边缘强化解码器在 Cityscapes 数据集上的定性结果分析

以看出，网络在精度上也有小幅的提高。其中，平均相对误差，相对均方误差和对数均方根误差分别提高了 5%、1%和 2%。若在 baseline 的基础上同时使用条状卷积和 CBAM 注意力模块后，网络边缘信息的提取和整合的能力达到了最大化，可以定性的描述，如图 4-14：1、在第一张测试图中可以看出 baseline 的深度图中没有清晰捕捉到中间行走的行人，使用条状卷积和 CBAM 注意力模块后的深度图可以捕捉到中间的行人；2、在第三张、第四张和第五张测试图可以看到 baseline 的描述车辆的深度图中出现了空洞的现象，使用条状卷积和 CBAM 注意力模块后的深度图比较平滑；3、在这五张所有测试图中，baseline 预测出的运动场中运动对象边缘都比较模糊，同时部分图出现了边缘缺失的现象如第五张测试图，使用条状卷积和 CBAM 注意力模块后的运动场中运动对象轮廓更加清晰也没有存在边缘缺失的现象。同时根据表 4-2 进行定性分析：在 baseline 上添加条状卷积和 CBAM 注意力模块后，平均相对误差，相对均方误差，均方根误差，对数均方根误差分别提高了，8%，6%，2%，4%。

表 4-2 在 Cityscapes 数据集上消融实验测试结果对比

方法	条状卷积	CBAM 注意力	误差指标（越小越好）				预测准确率（越大越好）		
			Abs Rel	Sq Rel	RMSE	RMSE log	$\delta<1.25$	$\delta<1.25^2$	$\delta<1.25^3$
Baseline			0.119	1.295	6.984	0.190	0.846	0.952	0.982
	✓		0.115	**1.214**	7.012	0.185	0.846	0.962	0.981
Ours		✓	0.116	1.317	6.893	0.186	0.849	0.969	**0.989**
	✓	✓	**0.110**	1.223	**6.831**	**0.183**	**0.852**	**0.971**	**0.989**

最后本章分别使用了五种不一样的大小的条状卷积进行条状卷积大小的消融实验，由表 4-3，根据误差指标和预测准确率可以看出，本章所选的 11×3 的条状卷积

可以最大限度发挥出条状卷积充分发挥其充分利用边界信息的作用。

由消融实验可表明，本章涉及边缘强化解码器的每一组分对于网络的精度的提升都是有效的，且将两个组分结合效果最好。同时对于条状卷积的大小选择 11×3 效果为最佳。

表 4-3 条状卷积消融实验结果

方法	条状卷积大小	误差指标（越小越好）				预测准确率（越大越好）		
		Abs Rel	Sq Rel	RMSE	RMSE log	$\delta<1.25$	$\delta<1.25^2$	$\delta<1.25^3$
Baseline	无	0.119	1.295	6.984	0.190	0.846	0.952	0.982
Ours	9×1	0.120	1.31	6.952	0.21	0.832	0.955	0.979
	9×3	0.125	1.33	6.960	0.198	0.838	0.961	0.981
	11×1	0.119	0.125	6.880	0.192	0.846	0.963	0.980
	11×7	0.118	**0.125**	**6.753**	0.189	0.849	0.972	0.975
	11×3	**0.110**	1.232	6.852	**0.187**	**0.852**	**0.978**	**0.981**

4.6 本章小结

本章提出了一种面对动态场景下的单目深度估计方法改进。该方法的特点是解决当前，在动态场景下深度估计边缘模糊和运动对象难以捕捉的问题。该方法的核心包括两点：

1、提出了一种边缘强化解码器，通过 CBAM 卷积注意力模块首先捕捉到并分割出运动对象，然后再经过条状卷积让其充分利用边界和上下文信息加强对边界的提取。

2、提出了一种新的损失函数，边缘损失函数，利用拉普拉斯算子的旋转不变性，使运动网络的收敛速度更快，防止运动网络出现梯度爆炸的现象。

3、在 KITTI 数据集和拥有更多动态场景下的 Cityscapes 数据集上训练，并在测试图象中表现出了本模型相较于其他运动模型和静态模型有更细致的边缘，均方根误差相较于先进的 Li[51]提出的算法降低了 4.8%。在 Cityscapes 数据集上的消融实验表明，提出的各个加入的组分均对整个网络深度估计性能均有所提高，条状卷积和 CBAM 注意力机制分别让模型均方根误差降低了 3%和 5%，消融实验总的运动场表现出了运动网络对运动对象的捕捉也更强。同时，根据训练过程中运动网络尺度因子的变化表现出提出的边缘损失函数有更快的收敛效果，收敛速度提高了 36%。总之，与之前的方法相比，本章提出的方法性能优越，有效增强了网络对边缘区域的

预测能力，输出的深度图边缘更细致、捕捉到的运动对象也更加清晰，网络收敛速度更快。

第五章 自制数据集算法验证

5.1 引言

为了进一步验证第三章和第四章提出的算法有效性，本章依托智能驾驶测试样车来采集街道中不同交通场景的数据集，将采集后的数据集进行进一步训练和测试来分析本文提出的算法的优缺点，根据优缺点来指导未来工作。

5.2 数据采集平台构建

智能驾驶数据采集平台是一个专门用于收集、存储、标注智能驾驶场景下车辆对周围环境的感知的数据。数据收集包括主要是依赖于相机、雷达、组合导航等多种传感器对交通环境中行人、车辆、路障等信息的实时记录。数据存储主要是通过车载 GPU 实时读取采集的数据，再通过 CPU 处理器实时分配内存将读取的数据存储到磁盘中。数据标注主要是人工对提取目标的真实深度值、目标种类、目标的距离进行标记和校核。

图 5-1 智能驾驶测试样车

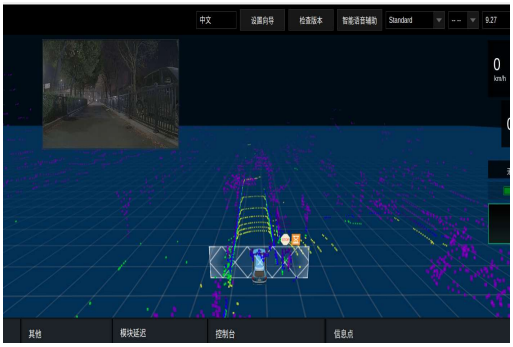


图 5-2 Apollo 系统终端显示

本章使用基于百度 Apollo7.0 智能驾驶系统的智能驾驶测试样车为平台，对数据采集平台进行构建。测试样车如图 5-1 所示，Apollo 系统终端显示如图 5-2。测试样车的收集系统主要包括：单目相机、激光雷达、组合导航；测试样车的存储系统主要为内置工控机中的 GPU 和 CPU；测试样车的标注系统主要为终端显示，对于系统错误标注进行人工实时筛选。本章利用智能驾驶测试样车对智能驾驶场景的数据进行大量采集，为第三章和第四章提出的算法的验证提供重要现实依据。

5.2.1 传感器选择

基于无监督学的单目深度估计算法只依赖于单个相机来完成整个三维环境的感

知任务，故传感器仅用到单目相机，单目相机实物如图 5-3 所示。相较于室内静态场景，智能驾驶场景大多是一种动态、且预测对象繁多复杂的场景。这就要求单目相机的性能要求远远大于其它工业相机的要求，不仅仅要求相机的分辨率较高可以清晰捕捉各种目标，还要求单目相机的帧率不能过低否则影响采集的数据实时性要求。

图 5-3 BFLY-PGE-23S6M 单目相机

图 5-4 单目相机安装位置

基于上述要求，本文在构建智能驾驶数据采集平台的时候选用了美国 FLIR 高性能计算机视觉面阵单目相机 BFLY-PGE-23S6M，采用高动态车规级图像传感器（夏普 Sharp Rj32S3AA0DT），协同主流串行传输芯片进行图像的采集，其主要参数如表 5-1 所示。相机安装在测试样车顶部的可标记导轨中间位置并用螺栓固定，相机离底盘高度为 60cm，如图 5-4 所示。

表 5-1 相机主要参数

参数	数值
靶面尺寸	1/1.9 inch CMOS
像元尺寸	$3\mu m$
帧率	35FPS
分辨率	2448×2048
有效焦距	10mm
电源	10~20V

5.2.2 相机的标定

相机标定的主要目的是确定相机的内外参数，确保相机捕获的图像可以真实反映相机成像各个目标像素点之间的位置关系。本节所用的相机标定方法采用了经典

的张正友标定法[53]，标定原理是将世界坐标系固定于棋盘上，利用已知的棋盘上角点的像素坐标来反向推导出相机的内外参数和畸变参数。在第二章 2.1 节相机模型和深度的内容介绍时，已经详细介绍过世界坐标系和像素坐标系的对应关系，本节不再详细介绍标定原理。

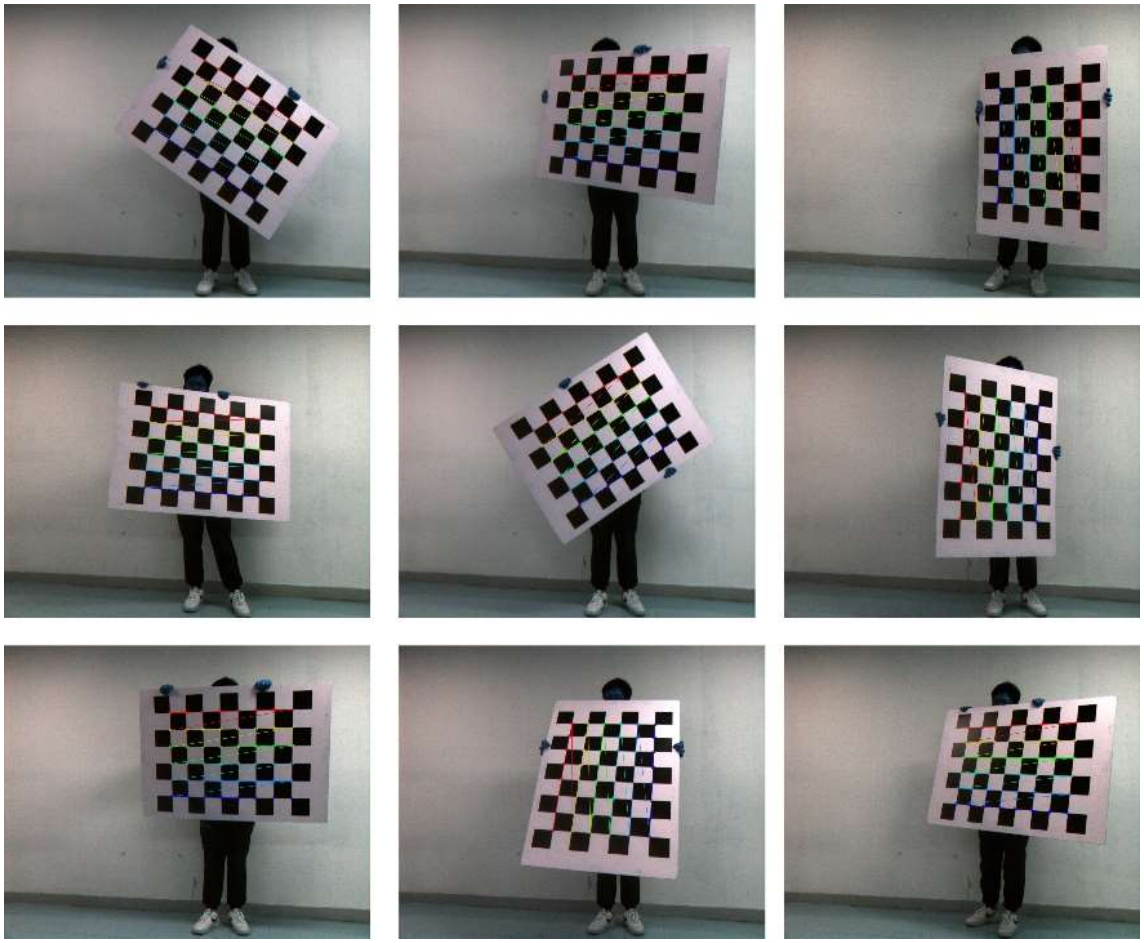


图 5-5 角点检测结果

本节基于 Opencv 视觉工具包，使用 Python 对 FLIR 单目相机进行标定，具体标定过程如下：

（1）制作棋盘格标定板。使用仿射率较好的 PVC 板材质将 9 行 7 列的黑白相间的棋盘打印在其上，其中棋盘长为 1.189mm、宽为 841mm。

（2）采集棋盘标定图像。根据不同距离、不同角度、不同方向对棋盘进行拍摄。本节采用了 10m 至 5m 的远近距离，向纸内核纸外旋转 45 度角，横向核纵向两个方向进行标定图像的采集。

（3）使用标定好的算法，对采集好的图像进行标定。经过 cv2.findChessboardCorners()函数确认标定的角点在相应区域内，打印内存标定结果。角点检测结果如图

5-5，表 5-2 表示相应的内参标定结果。其中，f_x/d_x、f_y/d_y分别表示x轴方向和y轴方向的归一化焦距；u_0、v_0表示图像中心；w_1、w_2为径向畸变，p_1、p_2为切向畸变。

表 5-2 相机内参标定结果

内部参数	标定结果
焦距	$f_x/d_x = 1211.3$，$f_y/d_y = 1041.4$
图像中心	$u_0 = 720.8$，$v_0 = 530.4$
径向畸变	$w_1 = -0.456$，$w_2 = 0.314$
切向畸变	$p_1 = -0.0021$，$p_2 = -0.0136$

5.2.3 计算平台

测试样车计算平台使用的嵌入式开发套件为英伟达 Jeston AGX Orin CLB（如图 5-6），搭载 12 核 Arm Cortex-A78AE 的 CPU 处理器，GPU 为搭载 1792 个英伟达 CUDA 核心及 56 个 Tensor Core 核心的英伟达 Ampere 架构，视频编码最大速度为 2×4K60Hz，视频解码最大速度为 1×8K30Hz，功耗为$40w$，更多详细参数参照表 5-3。相较于同类型的 Jeston AGX Xavier，Jeston AGX Orin CLB 处理器算力是其八倍以上，Jeston AGX Orin CLB 开发套件完全支持 AI 边缘计算和图像采集的要求。其在工控机的安装位置如图 5-7。

图 5-6 Jeston AGX Orin CLB 开发套件

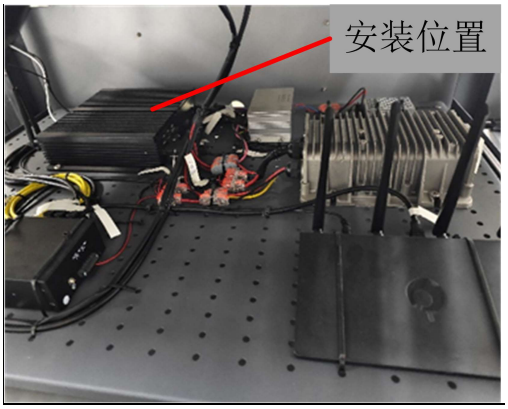


图 5-7 开发套件安装位置

表 5-3 英伟达 Jeston AGX Orin CLB 开发套件参数

参数	数值
GPU	搭载 1792 个 CUDA 核心的 VoltaGPU
CPU	12 核 ARM 的 64 位 Cortex-A78AE CPU
内存	64GB 256 位 LPDDR4x
网络	RJ45 连接器，支持 10Gbe
USB	2×USB-3.2 1×USB-2.0 2×USB-Type-c
尺寸	104×104×75mm

5.3 数据集制作及标注

本章数据集构建主要目的在于验证算法在实际交通场景下对车辆、行人等对象的深度检测效果。通过测试样车对不同街道、不同天气情况、以及遮挡等多种天气的交通场景进行捕捉，来验证算法的可行性。

本章数据集是通过测试样车对城区街道进行采集而来，包括晴天、夜晚、雨天、阴天等多种天气。在采集过程中，通过设置单目相机的抓拍帧率，5 帧取一张图片，最后一共采集到了 2000 张各种场景的图片，部分采集图像如图 5-8 所示。

数据集的标注。由于本文提出的算法是基于无监督学习的算法，故不需要标记真实深度值，这也是无监督深度估计方法相对于有监督单目深度估计的巨大优势，真正的做到了数据集随采随用。

图 5-8 部分采集图像展示

5.4 算法测试和结果分析

本节将介绍，基于自制数据集进行相关的算法验证实验，分别将实验分为 4 组对应 4 个实验算法：使用第三章 baseline 算法进行实验（没有添加增强密集连接模块和特征交互模块的网络），将其命名为实验算法Ⅲ；使用第三章在 baseline 基础上改进的算法（在 baseline 基础上添加增强密集连接模块和特征交互模块的网络）进行实验，将其命名为实验算法Ⅲ+；使用第四章 baseline 算法（没有使用边缘强化解码器的网络，同时没有添加边缘正则化）进行实验，将其命名为实验算法Ⅳ；第四章在 baseline 基础上改进的算法（在 baseline 基础上使用边缘强化解码器和使用边缘正则化）进行实验，将其命名为实验算法Ⅳ+。下面分别对这 4 组实验，进行参数设置、以及定性和定量分析。

5.4.1 参数设置

对于实验算法III和实验算法III+的参数设置。使用深度学习库 Pytorch1.7.1 在 NVIDIA GeForce RTX3070 上训练了该模型，操作系统为 Ubuntu18.04。以 Adam 作为优化器对整个网络进行 106 次迭代训练，优化器的学习率设置为 10-4。训练的批量大小为 16，训练的 RGB 图象分辨率大小为 1024×320。损失损失函数中几个超参数的设置：3-8 式中$\alpha = 0.85$；3-10 式中λ和μ分别设为 0.01 和 0.001。

对于实验算法Ⅳ和实验算法Ⅳ+的参数设置。使用 Tensorflow 框架实现了所提整个网络结构，在 NVIDIA GeForce RTX3070 上训练了该模型。以 Adam 作为优化器对深度预测网络和运动估计网络 10^6 次迭代训练，优化器的学习率设置为 10^{-4}。训练的批量大小为 16，训练的 RGB 图象分辨率大小为 1024×320。损失损失函数中几个超参数的设置：4-8 式中，α_{dep}为 10^{-3}；4-11 式中，α_{mot}和β_{mot}分别设置为 1.0 和 0.2；4-12 式中，α_{cyc}和β_{cyc}分别设为 10^{-3} 和 5×10^{-2}，4-13 式中α_{rgb}和β_{rgb}分别设置为 1.0 和 0.8。

5.4.2 定量分析

如表 5-4 所示，使用方法[17]对这 4 组实验算法进行评估。本节从两个角度分析：实验算法改进效果和自制数据集本身分析。

从实验算法改进效果分析。将实验算法III与验方法III+对比分析，网络在对弱纹理区域、小目标深度的提取精度上有了一定的提高，均相对误差（Abs Rel），相对均方误差（Sq Rel）和对数均方根误差（RMSE）分别减少了 7%、3%和 4%。将实验算法Ⅳ与验方法Ⅳ+对比分析，网络对运动目标的边缘深度准确率一定的提高，平均相对误差（Abs Rel），相对均方误差（Sq Rel）和对数均方根误差（RMSE）分别减少了 6%、3%和 5%。将这 4 个实验算法III、III+、Ⅳ、Ⅳ+放在一起分析可以得出改进实验算法III+和Ⅳ+是最优、而实验算法III+的效果是优于实验算法Ⅳ+，且均相对误差（Abs Rel），相对均方误差（Sq Rel）和对数均方根误差（RMSE）分别减少了 4%、3%和 6%。

从自制数据集自身分析。将表 5-4 与表 3-1 及表 4-1 进行对比，可以看出无论是使用哪个实验算法，使用自制数据集训练出来的模型精度要比，使用 KITTI 数据集和 Cityscapes 数据集的整体精度要低 10%，一方面与自制数据集的采集较少有关，

另一方面由于自制数据集中有大量的阴雨天气和夜晚等多样天气，使得模型的预测准确率降低。

表 5-4 不同实验算法误差和精度测试结果

算法	分辨率	误差指标（越小越好）				预测准确率（越大越好）		
		Abs Rel	Sq Rel	RMSE	RMSE log	δ<1.25	δ<1.25^2	δ<1.25^3
III+	1024×320	**0.126**	**1.089**	**5.456**	**0.223**	0.843	**0.941**	0.982
IV+	1024×320	0.135	1.171	5.778	0.247	**0.847**	0.892	**0.988**
III	1024×320	0.139	1.211	5.866	0.281	0.811	0.887	0.979
IV	1024×320	0.143	1.312	5.911	0.334	0.796	0.891	0.971

5.4.3 定性分析

为了更直观地看出基于自制数据集 4 个实验算法的效果，本节将 4 个模型对公路采集的部分交通场景的深度进行推理，如图 5-9 所示五张图片。五张图片中包含了三种交通场景：晴天、雨天和夜晚，其中左起前三张图片是晴天交通场景、左起第四张图片是雨天交通场景、左起第五张图片是夜晚交通场景。下面从两个角度进行分析：从实验算法改进效果和自制数据集自身分析。

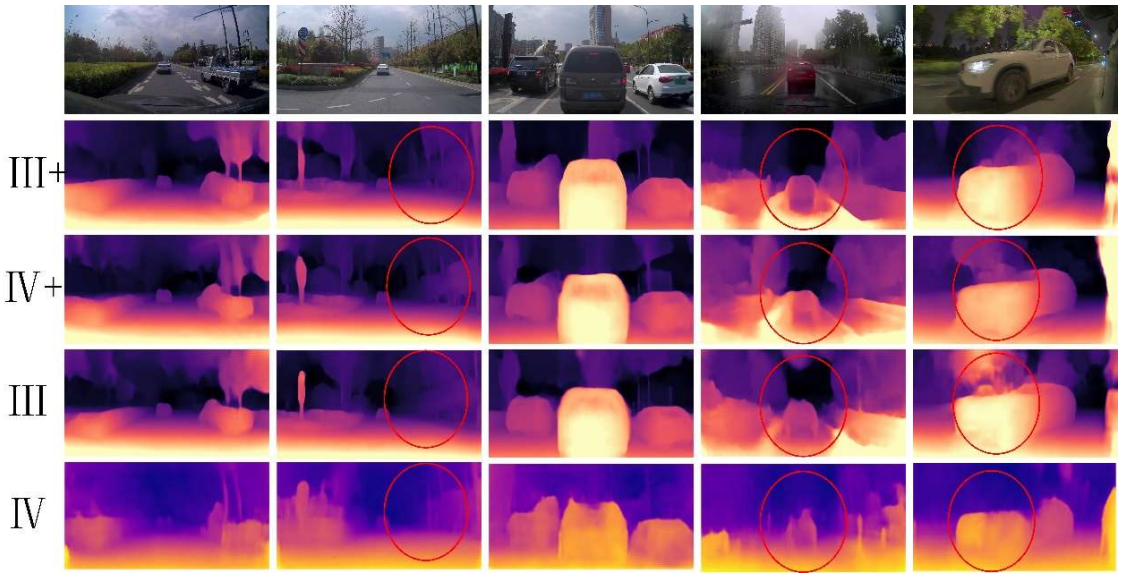


图 5-9 不同实验算法可视化分析

从实验算法改进效果分析。对于实验算法III和实验算法III+：观察左起第二张图片对应的深度图可以看出，对于马路右侧的树干边缘实验算法III+可以较清楚分辨出树干的边缘信息的，实验算法III可以预测出的树干细节缺失；观察左起第四张图片的雨天道路场景，实验算法III+可以实验算法III可以清晰预测出车轮；观察左起第五张图片的夜晚道路场景，实验算法III可预测出的车灯区域出现伪影而实验算

法III+没有出现。对于实验算法Ⅳ和实验算法Ⅳ+：观察左起第二张图片对应的深度图可以看出，实验算法Ⅳ+对于马路右侧的树干的边缘和远处小车的边缘深度较实验算法Ⅳ更加清晰；观察左起第四张图片的雨天道路场景的汽车和第五张夜晚道路场景的汽车边缘，实验算法Ⅳ不能完全预测整个汽车的深度轮廓而实验算法Ⅳ+可以清楚辨别出。将这 4 个实验算法III、III+、Ⅳ、Ⅳ+预测的深度图放在一起分析可知，无论是哪一种场景实验算法III+的预测效果最好、实验算法Ⅳ+次之。

从自制数据集自身分析。4 个实验算法在晴天交通场景的深度预测效果要远远优于雨天和夜晚交通场景的深度预测。从中可以分析出，一方面无监督单目深度估计模型在复杂天气的预测效果还是比较差，另一方面需要采集更多的复杂天气的数据对模型进行训练，让模型对复杂天气的鲁棒性更强。

5.5 本章小结

首先，介绍了基于智能技术测试样车数据采集平台的构建；然后将构建好的数据采集平台对多种交通场景进行数据集采集工作；最后将采集好的数据对本文提出的算法进行训练测试，证明了本文提出的算法模型无论在什么样的天气状况下都是优于改进之前的算法模型。但是，也暴露出了一些问题在复杂天气（夜晚、阴天、雨天等恶劣天气）交通场景，现阶段无监督单目深度估计算法整体估计精度是要低于良好天气交通场景下的无监督单目深度估计算法，这也是研究人员未来需要努力改进的方向。

第 6 章 结论与展望

6.1 研究成果总结

随着人工智能时代的到来，计算机视觉三维感知技术的应用也越来越广泛，特别是智能驾驶领域中的障碍物检测、机器人领域中的地图实时生成、虚拟现实领域中沉浸式人机交互等。这些领域的发展都离不开传感器要获取一个高精度的深度值，在兼顾模型经济性和普适性的条件下，人们提出了基于无监督学习的单目深度估计算法。经过充分的调研，分析出当下基于无监督单目深度估计算法的不足。首先是横向比较（同类算法）：度估计过程中出现的边缘缺失，伪影过多、小目标模糊、模型的泛化能力和鲁棒性较差等问题。纵向比较（与激光雷达和深度相机传感器相比）：运动对象边缘缺失、模型收敛速度慢、边缘缺失等问题。从两个角度分析，本文提出了两种解决方案。同时，为了进一步验证提出方案的合理性和真实性以及后续改进方向，本文利用智能驾驶测试样车对公路、街道多种天气交通场景进行数据采集制作数据集，将采集好的数据对本文提出的算法进行训练测试，验证了算法改进的有效性以及后续要改进的方向。具体工作如下：

（1）由于无监督单目深度估计算法本身的编解码器结构，特征仅仅在同一层编、解码器之间使用跳转连接，导致编码器对于不同尺度同特征的使用是不充分的，容易让预测出的深度图纹理缺失和边缘模糊。本文对编解码器结构进行改进，将原先的编解码器结构改为带有注意力机制的增强密集连接模块的编解码结构，减少了伪影过多和小目标模糊的现象。同时在解码之后加入全局特征交互模块利用自注意力机制来加强整个模型的鲁棒性能，使网络学习更平顺。

（2）与专业传感器相比在城市复杂的移动行人与车辆场景下深度估计精度不准确且鲁棒性较差。在编码器结构中，本文引入了基于注意力机制的条状卷积，传统的卷积核大多使用正方形卷积核，这种卷积核只对每一个像素的局部特征进行聚合而并没有完全利用边界和全局上下文特征，同时这也限制了神经网络在捕捉不同对象相对背景运动时存在的各向异性。使用状卷积可以偏向于特定的方向来部署条形卷积核，在相同卷积核尺寸下，条状卷积可以捕捉远距离相对位置关系实现更大的感受野，继而将移动物体的边缘提取更具清晰。同时引入新的基于拉普拉斯算子的边缘正则化，让网络收敛速度变快且学习更稳定。

（3）在使用自制训练集进行训练测试下，本文提出的算法模型无论在什么样

的天气状况下都是优于改进之前的算法模型。但是，也暴露出了一些问题在复杂天气（夜晚、阴天、雨天等恶劣天气）交通场景，现阶段无监督单目深度估计算法整体估计精度是要低于良好天气交通场景下的无监督单目深度估计算法，这也是研究人员未来需要努力改进的方向。

6.2 展望

本文从横向和纵向两个分别对基于无监督学习的单目深度估计模型进行了改进，并且与现有算法相比精度有定的提高，但是仍存在一些不足，需要在后续的研究中继续改进，后续可以从以下几个角度进行研究：

（1）增加语义动态分割网络。目前大部分无监督单目深度估计损失函数都是基于投影关系和极几何约束的，而这些假设都是在低速静态场景下提出的。对于一些高速动态的移动对象很多几何关系都不成立。可以借助神经网络首先将车、人等动态对象分割出来，然后再根据动态先验将动态场景下的对象映射到相应的静态场景然后再进行深度估计。

（2）引入生成对抗网络（GAN）判别器来预测复杂天气及夜晚的深度。目前大部分深度估计网络，预测出的场景比较单一，大部分都是晴朗的天气，对于一些雨雪雾天气或者是夜晚的深度的预测比较空缺。可以改进本文已有的无监督单目深度估计网络编码器，通过引入 GAN 判别器作为编码器用于训练复杂天气以及夜晚图像的深度提取，通过将复杂天气及夜间特征编码器和已有的白天特征解码器组合起来预测复杂天气和夜间的深度。

（3）提高算法的实时性。本文着重提高了深度估计的精度而忽略了算法的相应时间。如在智能驾驶过程中，要求实时感知不同路况不同交通场景的三维深度信息，否则存在安全隐患。有提高无监督单目深度估计的实时性，才能被应用到更广阔的领域中。在后续研究中可以让网络变得更轻量化减少参数冗余来加快网络运算速度，另一方面模型在推理深度图的过程中可以采用权重更少的解码结构减少模型推理使用的时间。

参考文献

[1] Ji P, Li R, Bhanu B, et al. Monoindoor: Towards good practice of self-supervised monocular depth estimation for indoor environments[C]//Proceedings of the IEEE/CVF international conference on computer vision. 2021: 12787-12796.

[2] Rajagopalan A N, Chaudhuri S, Mudenagudi U. Depth estimation and image restoration using defocused stereo pairs[J]. IEEE transactions on pattern analysis and machine intelligence, 2004, 26(11): 1521-1525.

[3] Saxena A, Chung S, Ng A. Learning depth from single monocular images[J]. Advances in neural information processing systems, 2005, 18.

[4] Saxena A, Sun M, Ng A Y. Make3d: Learning 3d scene structure from a single still image[J]. IEEE transactions on pattern analysis and machine intelligence, 2008, 31(5): 824-840.

[5] Ambiguities L. Optimal Structure From Motion: Local Ambiguities and Global Estimates[J].

[6] Roberts R, Sinha S N, Szeliski R, et al. Structure from motion for scenes with large duplicate structures[C]//CVPR 2011. IEEE, 2011: 3137-3144.

[7] Horn B K P, Brooks M J. The variational approach to shape from shading[J]. Computer Vision, Graphics, and Image Processing, 1986, 33(2): 174-208.

[8] Triggs B, McLauchlan P F, Hartley R I, et al. Bundle adjustment—a modern synthesis[C]//Vision Algorithms: Theory and Practice: International Workshop on Vision Algorithms Corfu, Greece, September 21–22, 1999 Proceedings. Springer Berlin Heidelberg, 2000: 298-372.

[9] Eigen D, Puhrsch C, Fergus R. Depth map prediction from a single image using a multi-scale deep network[J]. Advances in neural information processing systems, 2014, 27.

[10] Eigen D, Fergus R. Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture[C]//Proceedings of the IEEE international conference on computer vision. 2015: 2650-2658.

[11] Fu H, Gong M, Wang C, et al. Deep ordinal regression network for monocular depth estimation[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2018: 2002-2011.

[12] Mahjourian R, Wicke M, Angelova A. Unsupervised learning of depth and ego-motion from monocular video using 3d geometric constraints[C]//Proceedings of the IEEE conference on

computer vision and pattern recognition. 2018: 5667-5675.

[13] Dosovitskiy A, Fischer P, Ilg E, et al. Flownet: Learning optical flow with convolutional networks[C]//Proceedings of the IEEE international conference on computer vision. 2015: 2758-2766.

[14] Aleotti F, Tosi F, Poggi M, et al. Generative adversarial networks for unsupervised monocular depth prediction[C]//Proceedings of the European conference on computer vision (ECCV) workshops. 2018: 0-0.

[15] Pilzer A, Xu D, Puscas M, et al. Unsupervised adversarial depth estimation using cycled generative networks[C]//2018 international conference on 3D vision (3DV). IEEE, 2018: 587-595.

[16] Garg R, Bg V K, Carneiro G, et al. Unsupervised cnn for single view depth estimation: Geometry to the rescue[C]//European conference on computer vision. Springer, Cham, 2016: 740-756.

[17] Godard C, Mac Aodha O, Firman M, et al. Digging into self-supervised monocular depth estimation[C]//Proceedings of the IEEE/CVF international conference on computer vision. 2019: 3828-3838.

[18] Zhou T, Brown M, Snavely N, et al. Unsupervised learning of depth and ego-motion from video[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2017: 1851-1858.

[19] Sun L, Bian J W, Zhan H, et al. Sc-depthv3: Robust self-supervised monocular depth estimation for dynamic scenes[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2023.

[20] Guizilini V, Vasiljevic I, Chen D, et al. Towards zero-shot scale-aware monocular depth estimation[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2023: 9233-9243.

[21] Chen X, Zhang R, Jiang J, et al. Self-supervised monocular depth estimation: Solving the edge-fattening problem[C]//Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. 2023: 5776-5786.

[22] Ali A, Touvron H, Caron M, et al. Xcit: Cross-covariance image transformers[J]. Advances in neural information processing systems, 2021, 34: 20014-20027.

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

[24] 廖干洲,曾霞.基于亮度均衡化和仿射变换的交通标志图像修复研究[J].机电工程技

术,2024,53(02):212-216.

[25] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition[J]. arXiv preprint arXiv:1409.1556, 2014.VGG

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

[27] Godard C, Mac Aodha O, Brostow G J. Unsupervised Monocular Depth Estimation with Left-Right Consistency[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017: 270-279

[28] Long J, Shelhamer E, Darrell T. Fully convolutional networks for semantic segmentation[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2015: 3431-3440.

[29] Ronneberger O, Fischer P, Brox T. U-net: Convolutional networks for biomedical image segmentation[C]//Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18. Springer International Publishing, 2015: 234-241.

[30] Zhou Z, Rahman Siddiquee M M, Tajbakhsh N, et al. Unet++: A nested u-net architecture for medical image segmentation[C]//Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support: 4th International Workshop, DLMIA 2018, and 8th International Workshop, ML-CDS 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 20, 2018, Proceedings 4. Springer International Publishing, 2018: 3-11.

[31] 陈莹,王一良.基于密集特征融合的无监督单目深度估计[J].电子与信息学报,2021,43(10):2976-2984.

[32] Dosovitskiy A, Fischer P, Ilg E, et al. Flownet: Learning optical flow with convolutional networks[C]//Proceedings of the IEEE international conference on computer vision. 2015: 2758-2766.

[33] Lin M, Chen H, Sun X, et al. Neural architecture design for gpu-efficient networks[J]. arXiv preprint arXiv:2006.14090, 2020.

[34] Mnih V, Heess N, Graves A. Recurrent models of visual attention[J]. Advances in neural information processing systems, 2014, 27.

[35] Jaderberg M, Simonyan K, Zisserman A. Spatial transformer networks[J]. Advances in neural

information processing systems, 2015, 28.

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

[37] Dosovitskiy A, Beyer L, Kolesnikov A, et al. An image is worth 16x16 words: Transformers for image recognition at scale[J]. arXiv preprint arXiv:2010.11929, 2020.

[38] Shen T, Zhou T, Long G, et al. Reinforced self-attention network: a hybrid of hard and soft attention for sequence modeling[J]. arXiv preprint arXiv:1801.10296, 2018.

[39] Xiong C, Zhong V, Socher R. Dcn+: Mixed objective and deep residual coattention for question answering[J]. arXiv preprint arXiv:1711.00106, 2017.

[40] Zhao H, Zhang Y, Liu S, et al. Psanet: Point-wise spatial attention network for scene parsing[C]//Proceedings of the European conference on computer vision (ECCV). 2018: 267-283.

[41] Wang X, Girshick R, Gupta A, et al. Non-local neural networks[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2018: 7794-7803.

[42] Chen C, Chen Z, Zhang J, et al. Sasa: Semantics-augmented set abstraction for point-based 3d object detection[C]//Proceedings of the AAAI Conference on Artificial Intelligence. 2022, 36(1): 221-229.

[43] Wang C, Buenaposada J M, Zhu R, et al. Learning depth from monocular videos using direct methods[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2018: 2022-2030.

[44] Yin Z, Shi J. Geonet: Unsupervised learning of dense depth, optical flow and camera pose[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2018: 1983-1992.

[45] Casser V, Pirk S, Mahjourian R, et al. Unsupervised monocular depth and ego-motion learning with structure and semantics[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. 2019: 0-0.

[46] Ranjan A, Jampani V, Balles L, et al. Competitive collaboration: Joint unsupervised learning of depth, camera motion, optical flow and motion segmentation[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2019: 12240-12249.

[47] Gordon A, Li H, Jonschkowski R, et al. Depth from videos in the wild: Unsupervised monocular depth learning from unknown cameras[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2019: 8977-8986.

[48] Yang Z, Wang P, Wang Y, et al. Every pixel counts: Unsupervised geometry learning with holistic 3d motion understanding[C]//Proceedings of the European conference on computer vision (ECCV) workshops. 2018: 0-0.

[49] Luo C, Yang Z, Wang P, et al. Every pixel counts++: Joint learning of geometry and motion with 3d holistic understanding[J]. IEEE transactions on pattern analysis and machine intelligence, 2019, 42(10): 2624-2641.

[50] Luo X, Huang J B, Szeliski R, et al. Consistent video depth estimation[J]. ACM Transactions on Graphics (ToG), 2020, 39(4): 71: 1-71: 13.

[51] Li H, Gordon A, Zhao H, et al. Unsupervised monocular depth learning in dynamic scenes[C]//Conference on Robot Learning. PMLR, 2021: 1908-1917.

[52] Yang Z, Wang P, Wang Y, et al. Every pixel counts: Unsupervised geometry learning with holistic 3d motion understanding[C]//Proceedings of the European conference on computer vision (ECCV) workshops. 2018: 0-0.

[53] 林绿开,钮倩倩,李毅.基于棋盘标定板的优化相机参数标定方法[J].计算机技术与发展,2023,33(12):101-105.

攻读硕士学位期间承担的科研任务与主要成果

（一）参与的科研项目

[1] 基于多模态融合的智能车辆多任务感知技术研究，安徽省自然科学基金资助项目，课题编号: 2208085MF173.

[2] 特定场景无人车通用化全线控底盘开发与产业化应用，安徽省重点研究与开发计划项目，课题编号: 202104a05020003.

（二）发表的学术论文

[1] 基于增强密集连接和全局特征交互的无监督单目深度估计[J]，智能科学与技术学报（终审）

[2] Unsupervised Monocular Depth Estimation with Edge Enhancement for Dynamic Scenes[J]，IEEE Sensors Journal（Accept）

（三）申请及已获得的发明专利

[1] 一种具有减震功能的电动汽车驱动桥控制器[P].CN115163741A，2022-07-28.（发明专利已授权）

[2] 一种基于新能源汽车安全防护的驱动电机[P].CN114583903A，2022-06-03.（发明专利实质审查）

（四）获奖情况

[1] 第十届“挑战杯·华安证券”安徽省大学生课外学术科技作品竞赛二等奖

致 谢

首先，衷心感谢导师时老师对本人的精心指导。时老师对于前沿科技有着长远的目光，对自动驾驶的发展有着精准的判断，矢志不渝的带领、鼓励我从事智能感知的研究；时老师具有严谨的科研态度和一丝不苟的工作精神，他的言传身教将使我终生受益。

感谢我的同门江彤、张建国、毛飞、张庆。感谢你们为我带来的欢乐、灵感、慰藉、动力、安全感，也愿我们顶峰相见。同时，感谢室友吴朗、徐腾、纪旭阳在生活上的照顾与帮助，也愿我们友谊长存。

感谢周梦如师妹、徐柳柳师妹、潘艺馨师弟、吴文超师弟、杨礼师弟、董心龙师弟、李帅师弟。感谢你们对我工作、实验的协助，与你们的每一次学术交流都驱动着我向更广阔的领域探索，也愿你们学业有成、效力祖国。

感谢父母对我的养育之恩以及最亲近人的支持。感谢我的父母对我人生观、价值观、世界观的引导，感谢我最亲近的人对我无穷的鼓励。

最后，再一次感谢所有关注、支持、帮助我的人，并衷心感谢为审阅本论文付出辛勤劳动的各位专家和学者。

尚德致学　唯实惟新