

中图分类号：TP391.4

论文编号：10006SY1706124

北京航空航天大學
硕 士 学 位 论 文

实时且高精度双目立体匹配算法
研究与实现

作者姓名 张友敏

学科专业 计算机科学与技术

指导教师 百晓教授

培养院系 计算机学院

Research and Implementations of Real-time and High Accuracy Binocular Stereo Matching Algorithms

A Dissertation Submitted for the Degree of Master

Candidate: Zhang Youmin

Supervisor: Prof. Bai Xiao

School of Computer Science and Engineering
Beihang University, Beijing, China

中图分类号：TP391.4

论文编号：10006SY1706124

硕士 学位 论文

实时且高精度双目立体匹配算法研究与实现

作者姓名	张友敏	申请学位级别	工学硕士
指导教师姓名	白晓	职 称	教授
学科专业	计算机科学与技术	研究方向	模式识别
学习时间自	2017 年 9 月 1 日	起 至	2020 年 7 月 30 日止
论文提交日期	2020 年 8 月 3 日	论文答辩日期	2020 年 7 月 30 日
学位授予单位	北京航空航天大学	学位授予日期	年 月 日

关于学位论文的独创性声明

本人郑重声明：所呈交的论文是本人在指导教师指导下独立进行研究工作所取得的成果，论文中有关资料和数据是实事求是的。尽我所知，除文中已经加以标注和致谢外，本论文不包含其他人已经发表或撰写的研究成果，也不包含本人或他人为获得北京航空航天大学或其它教育机构的学位或学历证书而使用过的材料。与我一同工作的同志对研究所做的任何贡献均已在论文中做出了明确的说明。

若有不实之处，本人愿意承担相关法律责任。

学位论文作者签名： 张友波

日期： 2020年 8月 1日

学位论文使用授权书

本人完全同意北京航空航天大学有权使用本学位论文（包括但不限于其印刷版和电子版），使用方式包括但不限于：保留学位论文，按规定向国家有关部门（机构）送交学位论文，以学术交流为目的赠送和交换学位论文，允许学位论文被查阅、借阅和复印，将学位论文的全部或部分内容编入有关数据库进行检索，采用影印、缩印或其他复制手段保存学位论文。

保密学位论文在解密后的使用授权同上。

学位论文作者签名： 张友波

日期： 2020年 8月 1日

指导教师签名： 白晓

日期： 2020年 8月 1日

摘要

近年来，虚拟现实和增强现实技术得到大众的广泛关注，深度估计因此也成为计算机视觉领域一个备受瞩目的研究课题，在室内场景构建，自动驾驶以及人体和面部跟踪等研究和产品领域也被大量应用。双目立体视觉是一个经典的计算机视觉、深度估计问题，它旨在利用像素匹配方法确定双目视觉系统获取的左右目图像中各像素间的对应关系，以便确定拍摄物体的深度信息。随着双目立体匹配研究的深入，实时化且高性能已成为该课题研究的方向和热点。

本文对双目立体匹配方法和相关技术进行了系统研究，专注于用几何先验构建更加结构性的双目匹配过程，减少现有双目匹配方法中的模型参数冗余，从而实现实时且高精度的双目立体匹配方案。主要工作及创新点如下：

(1) 提出一种自适应的单峰匹配代价滤波方案。对于每个像素点，以真实视差值为中心， L_1 距离为相似度度量，根据真实视差构造一个单模态（单峰）匹配代价分布；并且对交叉熵损失函数改进，用于约束网络估计得到的与真实匹配代价分布之间的一致性。为防止网络在一些极具挑战的区域（比如遮挡）过拟合，本文提出置信度评估网络，能够根据网络匹配不确定度自适应调整真实单峰分布的平缓程度。实验结果表明，该方法能够加强网络对匹配代价计算的学习，即鲁棒的图像特征和相似度估计函数，实现更加高效匹配过程，并且在新的场景中表现出更优的泛化性能。在保证与目前的先进水平算法，如PSMNet，相同精度条件下，单峰匹配代价滤波方案能大量减少匹配代价聚合过程所需的3D卷积数量。

(2) 设计一个轻量级的高精度立体匹配框架。本文提出一种对视差搜索空间进行快速剪枝方案，通过提出一种高效且轻量的视差参选值推荐模块，结合自适应的单峰匹配代价滤波方案实现10ms内锁定视差参选值，并利用3D代价聚合网络对各参选值进行评估和筛选，进而得到高精度的视差匹配结果；实验结果表明，在保持高精度的同时，我们的立体匹配网络能够在KITTI数据集分辨率下提供超过20FPS的视差推断效率，在现有实时双目视觉算法中有着十分明显的性能优势。

关键词：深度估计，双目立体视觉，滤波，实时，置信度评估

Abstract

Recent years, with virtual reality and augmented reality technologies received extensive attention from the public, depth estimation has therefore become a high-profile research topic in the field of computer vision. Lots of applications have been developed in the field of indoor scene construction, automatic driving, and human and facial tracking. As a classic computer vision and depth estimation problem, binocular stereo matching aims to determine the depth information by using the pixel matching algorithm to find the correspondence between the pixels in the left and right images acquired by the binocular vision system. With the fast development of stereo matching research, real-time and high performance has become the hot topics of this subject.

Through systematically studying the binocular stereo matching methods and related technologies, this paper targets at constructing a more structural binocular matching process using geometric priors. By further reducing the model parameter redundancy in the existing stereo matching methods, our stereo matching method achieves both real-time and high accuracy. The main contributions and innovations are as follows:

(1) An adaptive unimodal cost volume filtering scheme is proposed. For each pixel, with the true disparity value as the center and the L_1 distance as the similarity measure, a unimodal matching cost distribution is constructed; and the cross-entropy loss is further improved to constrain consistency between the obtained and true matching cost distribution. In order to prevent the network from overfitting in some extremely challenging areas (such as occlusion), we propose a confidence estimation network that can adaptively adjust the smoothness of the true unimodal distribution according to the network matching uncertainty. Experimental results show that our method can enhance the learning of matching cost computation, that is, more robust image features and similarity measure functions. By constructing a more efficient matching process, our algorithm also shows better generalization performance in new scenarios. Furthermore, under the condition of reaching the same accuracy as state-of-the-art, e.g., PSMNet, our unimodal cost volume filtering scheme can greatly reduce the number of 3D convolutions required in the matching cost aggregation process.

(2) A high accuracy but lightweight stereo matching architecture is designed by proposing

a fast pruning scheme for disparity searching space. The core we achieved it is an efficient and lightweight disparity proposal network. With our adaptive unimodal cost volume filtering scheme further embedded, the pruned disparity searching space is determined within 10ms. The 3D cost aggregation network is thereafter used to evaluate and filter each disparity candidate to obtain high-accuracy matching results. Experimental results show that, while maintaining high accuracy, our stereo matching network can provide more than 20FPS inference rate under the resolution of the KITTI data, which shows an obvious advantage in the existing real-time stereo matching methods.

Key words: Depth Estimation, Binocular Stereo Matching, Filtering, Real Time, Confidence Measure

目 录

第一章 绪论	1
1.1 论文选题背景	1
1.2 国内外相关研究现状	1
1.3 课题研究目标及内容	4
1.4 本文组织结构	5
第二章 相关理论和技术	6
2.1 双目立体视觉成像基本原理	6
2.1.1 相机模型	7
2.1.2 对极几何	9
2.1.3 相机和场景 3D 结构重建.....	10
2.2 基于传统算法的立体匹配方法	11
2.2 基于深度学习的立体匹配方法	14
2.3 本章小结	17
第三章 自适应的单峰匹配代价滤波	19
3.1 引言	19
3.2 基于 3DCNN 的立体匹配算法	19
3.3 自适应的单峰匹配代价滤波模块	21
3.3.1 单峰匹配代价分布生成	22
3.3.2 置信度估计网络	22
3.3.3 立体聚焦损失	23
3.3.4 全部损失函数	23
3.4 实验验证	24
3.4.1 数据库及评价指标和实现细节	24
3.4.2 实验结果及讨论	25
3.5 本章小结	31
第四章 基于视差推荐网络的实时双目立体匹配	32
4.1 引言	32

4.2 实时网络算法	33
4.2.1 特征提取网络	33
4.2.2 视差推荐网络	34
4.2.3 匹配代价聚合	36
4.2.4 视差图优化	37
4.2.5 全部损失函数	37
4.3 实验验证	38
4.3.1 数据库及实验协议和实现细节	38
4.3.2 实验结果及讨论	39
4.4 本章小结	42
总结与展望	44
参考文献	45
攻读硕士学位期间取得的学术成果	53
致 谢	54

图目

图 1 视差维度上的匹配代价分布样例.....	3
图 2 双目相机拍摄图像与视差图.....	6
图 3 双目立体视觉模型.....	7
图 4 针孔相机几何模型.....	8
图 5 世界坐标与相机坐标间的欧几里得转换.....	9
图 6 外极线约束示意图.....	9
图 7 三角形法求解 3D 空间信息.....	11
图 8 AcfNet 网络结构框架.....	19
图 9 PSMNet 网络结构图.....	20
图 10 AcfNet 不同超参的消融实验结果.....	25
图 11 AcfNet 在 Scene Flow 测试集上方差 σ 的分布直方图.....	26
图 12 AcfNet 方差调节有效性示意图.....	27
图 13 AcfNet 在 Scene Flow 测试集上定性评估结果.....	28
图 14 AcfNet 在 KITTI 2012 上可视化结果	29
图 15 AcfNet 在 KITTI 2015 上可视化结果	30
图 16 DPN-Stereo 实时网络结构框架	33
图 17 视差推荐网络所预测的视差搜索空间可视化.....	40
图 18 DPN-Stereo 在 Scene Flow 测试集上定性评估结果	42
图 19 DPN-Stereo 在 KITTI 2015 数据集上的定性评估结果.....	42

表目

表 1 AcfNet 网络模块有效性分析	27
表 2 自适应单峰匹配代价滤波有效性分析	28
表 3 匹配代价滤波对比分析	29
表 4 AcfNet 在各公开数据集上的结果	30
表 5 DPN-Stereo 网络模块有效性分析	39
表 6 DPN-Stereo 网络各模块运行时间统计	40
表 7 定性评估双目立体匹配方法运行时间和精度	41
表 8 DPN-Stereo 在 KITTI 2015 上的定性评估	41

第一章 绪论

1.1 背景和意义

从图像中进行3D几何重建是一个经典的计算机视觉问题，已有超过30年的研究历史。而直到最近这些技术才足够成熟，可以从实验室中严格控制的环境应用到室外场景，为工业提供精确、可量产的深度度量方案。获得密集且精确的深度图是处理一些更高阶段任务的关键所在，比如3D重建^{[1][2][3]}、建图、定位^{[4][5][6]}和自动驾驶^{[7][8][9]}等。本文主要讨论双目立体匹配^{[10][11]}，是一项被动深度估计方案。当然，也有许多技术设备用于从设备中主动估计深度，比如结构光投影、ToF度量和激光雷达等。借用这些深度感知设备一般情况下能够在场景中测量得到非常精确的深度信息，但环境因素的干扰也是这些设备所需共同面对的问题。以激光雷达为例，为实现对环境场景全面且精确的扫描，一般需要多个雷达设备共同参与，这要求技术人员对各雷达的摆放和校正也有着十分严格的要求，失对齐有时在所难免。而且，发射的雷达信号可能由于镜面反射或者多路效应而无法再次被雷达设备接收。值得一提的是，雷达扫描只能提供稀疏的深度点云信息，对点云密集度有要求的情况下只能投入更多的雷达，因此成本投入大也成为一个问题。

从图像中直接推断深度信息成为克服以上难题的潜在解决方案。经过几十年的技术发展，已发展出各式各样的解决方案。双目深度估计^{[12][13]}技术作为计算机视觉的一个重要分支，是替换昂贵的主动深度估计设备的一个十分可靠的解决方案。传统立体匹配算法一般基于人工设计的特征和相似度度量函数。随着机器学习、深度学习技术的繁荣和成功，双目立体匹配受益匪浅，性能也得到了极大提升。然而，巨大的内存消耗和计算量需求仍然是这些方法的一大诟病，也阻碍了双目立体视觉方法在现实场景中的推广和应用。本文从双目立体匹配的实时性和精度之间的权衡展开研究，在保持精度在现有方法的先进水平前提下，将立体匹配的时间消耗降到可实用范围，推进双目视觉在基于深度图像的研究领域发挥更为重要的作用，顺应计算机视觉的发展潮流。

1.2 国内外相关研究现状

随着计算机视觉技术的越来越广泛的应用，立体匹配技术受到越来越多的关注。在最近的几十年中，立体匹配技术发展迅速。在2002年，Scharstein和Szeliski^[14]的研究工作提出了一种适用于双目深度估计算法的框架，该框架规定了整个深度估计过程的四个步

骤：匹配代价计算，匹配代价聚合，视差图计算，视差图优化。这次调研工作第一次引入Middlebury^[14]数据集和相关的匹配结果评价指标。而后，一个更大的数据集KITTI^{[15][16]}公布了大量的双目街景图像，并且提供对应的激光点云深度信息。这些数据集中包含大量的复杂场景，对双目立体匹配方法提出了新的挑战。对应的，相继有大量的研究工作提出了对各个匹配模块的改进算法，大幅提升了双目立体匹配在各方面的性能。

匹配代价是立体匹配的基础，设计一种能够抵抗噪声干扰并且对照明变化不敏感的匹配代价可以提高立体匹配的准确性。具体地，计算匹配代价是计算参考图像上的每个像素点 $I_R(\mathbf{p})$ 与目标图像上的对应点 $I_T(\mathbf{pd})$ 在所有视差可能性范围内的代价值。传统算法通常直接使用图片的像素强度作为特征，使用差的绝对值（AD）、差的平方和（SD）、归一化互相关（NCC）和其他指标作为匹配代价。基于图像梯度或者二分类模式的局部描述子，比如CENSUS^[17]和BRIEF^{[18][19]}后来也相继被应用。据Tombari等人^[20]提供的技术调研可知，在假设相邻像素点更可能来自相同物体表面和视差的假设下，图像内容语义层面的认知可以更快速地合并对具有相似外观的相邻像素的视差预测。因此，局部匹配成本的优化也可以在全局框架内进行，通常将结合局部数据项和成对平滑项的能量函数最小化，该方法可以使用图形切割^[21]或置信度传播^[22]来实现全局优化，也可以扩展到倾斜的曲面^[23]。Hirschmuller^[24]提出半全局匹配（SGM）也是对全局优化的一种流行而有效的近似方法，因为其设计的动态规划算法可以在多个方向上优化能量函数。

真实的深度信息不仅可以用于评估算法的鲁棒性，也提供了将机器学习算法引入双目匹配框架的机缘。Zhang和Seitz^[25]交替优化视差估计和Markov随机场正则化参数。Scharstein和Pal^[26]学习条件随机场（CRF）参数，而Li和Huttenlocher^[27]使用结构化支持向量机训练非参数CRF模型。机器学习也可以用来估计传统立体匹配算法的可信度，例如Haeusler等人的随机森林方法^[28]。据Park和Yoon^[28]显示，这种置信度可以改善SGM的视差估计结果。

近年来，利用卷积神经网络（CNN）进行立体匹配的方法越来越多。Zbontar和LeCun^[29]展示了一个经过 9×9 的图像块训练的匹配代价计算网络，紧接着使用无需学习的成本聚合和正则化方法，也可以产生可比肩当时最好方法性能的结果。后续，Luo等人^[30]设计了一个孪生网络计算局部匹配代价，并且用多分类标记训练网络，在保证精度的条件下，得到了一个更快的匹配网络架构。Tulyakov等人^[31]则设计离散的拉普拉斯分布软化多分类标签，进一步提升了双目匹配精度。但是这些基于分类的方法将视差回归和网络训练分离，他们一般需要传统的正则化项和后处理才能得到光滑的视差图。而且

自然场景中的立体匹配任务是复杂且多变的，每个像素点使用一致的多分类标签是不合理也无法应对所有情况，甚至容易导致网络过拟合。

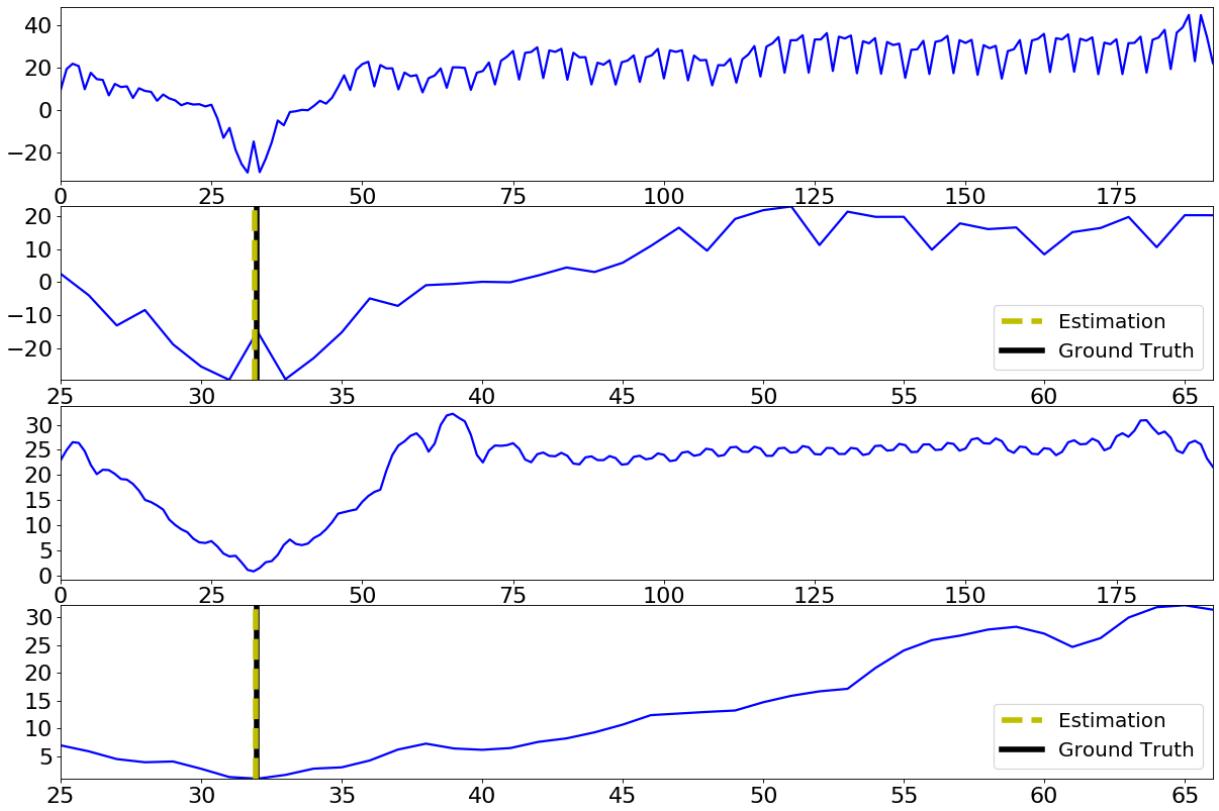


图1 视差维度上的匹配代价分布样例

不过，也有一部分工作提出直接端到端的网络训练并且直接回归视差图。Mayer等人^[32]创建了一个大型合成数据集用于训练一个视差估计（以及光流）的网络，从而改善了现有技术。作为网络的一种变型，它沿着视差维度计算一维相关性（1-D correlation）作为对立体匹配代价的乘积近似。Kendall^[33]等人则提出一种基于3D卷积端到端学习视差的方法。该方法利用图像的几何特性构造3D匹配代价体，大幅提升网络语义理解和视差估计精度。并且，依赖soft argmin（公式(3.1)）视差回归算法，该方法能够端到端地实现亚像素精度的视差学习，无需额外的后处理和正则化。这套匹配代价构建模式被不断沿用和改善，比如PSMNet^[34]、GANet^[35]，并且成为现有方法中精度最高的网络框架。然而，视差回归算法soft argmin在回归过程中受所有视差模态影响，所得到的匹配代价经常呈现多峰，回归得到的视差值也不会对应匹配代价最小的视差值。如图1所示，第一行为PSMNet只采用soft argmin算法用于视差回归训练得到的匹配代价分布，第三行为本文提出方法AcfNet（Adaptive unimodal cost volume filtering Network，自适应的单峰匹配代价滤波网络）训练收敛后的匹配代价分布，其中第二、四行分别对应第一、三行缩放

视差范围[25, 66]内的匹配代价分布。和本文方法相比，PSMNet回归的匹配代价分布中最小匹配代价对应的视差值与真实的视差偏离较远，这也就意味着模型在训练过程中学习到的相似度估计是不可靠的。

总体而言，这些基于神经网络的方法能够提取鲁棒的特征用于克服复杂的现实世界场景，并且通过把传统立体匹配方法实现为网络层，立体匹配模型能够实现端到端的训练。然而，目前的高精度立体匹配算法高度依赖于3D卷积实现匹配代价聚合过程。为有效聚合各个视差以及不同像素点间的匹配信息，网络结构上会堆积大量的3D卷积。相应的，模型推断深度的速度也会大幅减小，因此实时性问题成为深度估计算法落地现实场景的一大阻碍。对应的解决方案普遍采用由粗到精的网络架构^{[32][36][37][38]}。

具体而言，其网络架构与特征金字塔网络^[40]相似，通过一系列的下采样操作，得到多尺度的特征图；在最低分辨率特征图上，网络进行全视差搜索范围内的匹配并得到粗糙的视差图，对于更高分辨率特征图，则以上一尺度预测的视差图为中心，只进行小范围内的视差搜索。通过在低分辨率匹配过程实现视差参选值筛选，减少了高维特征和匹配过程的计算量，大幅提升网络匹配算法。但缺点也十分明显，低分辨率特征易丢失细节特征信息，且由于高分辨的匹配高度依赖于低分辨率的匹配结果，在低分辨率便丢失的细小且与背景视差相差较大的物体将一直无法被网络检测并完成深度估计。

最近，Duggal等人^[39]则认为立体匹配的视差搜索空间过于密集，有许多的视差参选值并不需要参与训练与评估；而且根据Tombari等人^[20]的技术调研结果，相邻像素点可能来自于同一视差假设，所以许多像素点的匹配结果可以依靠信息传播的方式获得。基于以上观察，Duggal等人^[39]将传统块匹配（Patch Matching）^[76]方法嵌入到神经网络中对全视差匹配样本空间进行剪枝，并利用3D聚合网络对剪枝后的样本空间完成代价聚合和精细的视差估计。然而缺陷是仅依靠上下文信息有限的两个并行分支网络预测剪枝后空间的上下界是病态的，比如一些遮挡或者无条纹区域的上下界预测是不准确的，那么后续的优化过程也将是在错误的视差搜索空间进行。因此，现有的实时方案在性能上仍然和目前精度最好的一些方法存在很大差距。

1.3 课题研究目标及内容

本文主要对双目立体视觉领域的实时且高精度双目立体匹配问题展开研究。首先对现有双目立体匹配算法进行调研，了解传统方法和现有基于深度学习的算法框架，并着重研究匹配代价计算过程，将几何先验用于网络监督，提升匹配性能。并且，实现双目

立体匹配实时应用的轻量化网络设计，解决网络参数量过剩、视差搜索空间过大问题，实现高精度（即平均像素误差为1，这与目前领域内大型网络架构所具有的深度推测准确率持恒）且实时（即在普通级显卡，如GTX 1080Ti上实现20帧/秒以上的深度推测速度）的立体匹配算法。

综上所述，本文主要有两部分的研究内容：

1、重点研究匹配代价计算过程。结合传统双目立体匹配算法中的几何架构，本文将针对匹配代价在理想情况下呈现单模态（单峰）的几何性质，从网络设计、网络约束等层面进行改进与创新，实现网络能够根据上下文信息自适应调整学习策略，且收敛后每个像素的匹配代价分布更具备几何意义。

2、为进一步促进立体匹配网络对实时应用的拓展，本文将分析现有双目立体匹配算法的步骤及特点，特别是针对实时应用的网络设计方法。逐像素的视差搜索空间是导致计算量暴涨和时间消耗剧增的主要原因。本文将研究一种高效且轻量的视差参选值推荐网络，通过小量计算成本即可快速搜索视差搜索空间，从而大幅提升双目立体匹配效率，实现性能可靠的实时立体匹配结果。

1.4 本文组织结构

本文共分为四章，具体内容如下：

第一章是绪论部分，介绍选题背景、国内外研究现状以及论文的主要研究目标和内容。

第二章是国内外研究现状综述，针对主流的双目立体匹配方法进行介绍，并对近年来出现的基于深度学习方法进行比较，分析其优缺点及面对实时应用的网络设计。

第三章是自适应的单峰匹配代价滤波，首先对基于深度学习的双目立体匹配卷积神经网络进行概述，并在此基础上提出一些改进方案和具体的实现细节，最后给出实验结果及讨论。

第四章是基于视差推荐网络的实时双目立体匹配，首先对现有实时立体匹配方法进行概述，在此基础上提出视差推荐网络，以大幅剪枝视差搜索空间降低网络的运行时耗，最后给出实验结果及讨论。

文章最后，我们对本课题的总体研究过程进行了概括和总结，并对该课题的后续研究内容进行展望。

第二章 相关理论和技术

双目立体视觉是计算机视觉的一个重要分支。经过几十年的研究，双目测距这一课题已经有了成熟的理论体系，以下将详细介绍双目立体匹配算法的基础理论知识，并详细介绍传统立体匹配方法和现有基于深度学习的立体匹配模型。

2.1 双目立体视觉成像基本原理

双目立体视觉讨论的是两个透视图像之间的几何关系。这些图像可以由双目设备获得，也可以通过相机在场景中的相对运动获得，这两种方法是几何等价的。以图2所展示的KITTI 2015数据集中经过矫正的双目相机拍摄街景图像为例，图2 (a) 为相机左视图（参考图像），图2 (c) 为相机右视图（目标图像）。双目立体视觉所要解决的问题就是建立两个视图之间的匹配关系，从而恢复场景的3D结构信息。观察图2 (a) 和2 (c) 可以发现，右视图中的像素相对左视图产生了向左的偏移，也就是右视图中的像素与左视图中的像素存在对应关系，只是存在位置差异，而这个偏移量就是真实视差图像素值的大小。公式(2.1)描述了左视图，右视图与视差图之间的关系：对于左视图上的点 $I_L(\mathbf{p})$ ，其坐标为 (x, y) ，在视差图上对应点的视差值为 $D(x, y)$ ；而横坐标 x 与视差值 $D(x, y)$ 之差，即为点 $I_L(\mathbf{p})$ 在右视图上所对应的位置 $I_T(\mathbf{p}d)$ ，坐标为 (x', y') 。

$$(x', y') = (x - D(x, y), y) \quad (2.1)$$

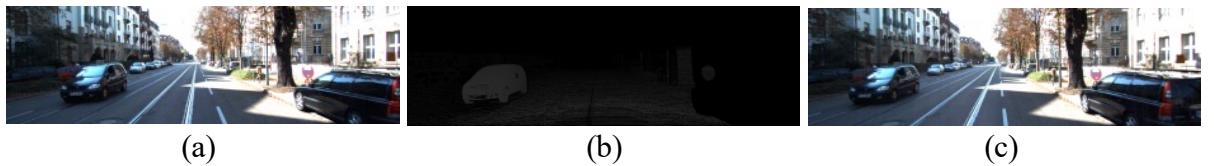


图 2 双目相机拍摄图像与视差图

具体地，如图3是一个双目立体视觉基础模型，左右两个摄像头在同一平面且与光轴平行， O_R ， O_T 分别是左右相机的光心，记 $O_R O_T$ 之间的基线距离为 b ， f 为相机焦距， X 为现实场景中物体的任意一点，其在两个相机的光屏成像投影为 x 和 x' ， Z 为点 X 到基线的垂直距离即为场景深度，其中 $X_R - X_T$ 成为视差，记为 d 。由 Xxx' 与 $XO_R O_T$ 三角相似可得：

$$\frac{b}{Z} = \frac{(b + X_T) - X_R}{Z - f} \rightarrow Z = \frac{b \cdot f}{X_R - X_T} = \frac{b \cdot f}{d} \quad (2.2)$$

Z 为物体到相机的距离，根据公式(2.2)可以从视差 d 推算得到物体到相机的距离。

为建立以上匹配关系，需要回答三个问题^[81]：

- i. 对应几何：给定第一个（左）视图中的一点，第二个（右）视图中的对应点存在什么样的位置限制关系；
- ii. 相机几何：给定一系列的图像对应点，如何求得两个视图下的相机矩阵；
- iii. 场景几何：给定各视图中的像素匹配关系和相机矩阵，如何恢复场景的3D信息。

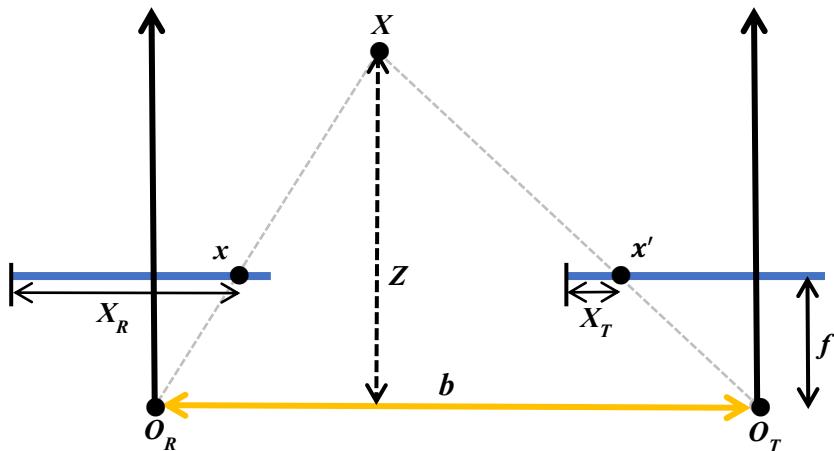


图 3 双目立体视觉模型

给定两个视图之间的对极几何关系，可以直接回答第一个问题：给定一个视图中的任意一点，其在另一个视图中的对应点位置由对极几何关系被限定在另一视图的对极线上。对极几何只取决于相机，比如相机之间的相对位置和他们各自的相机内部参数，而完全脱离于场景结构信息。对于第二、三个问题，可由图像点匹配关系同时解决。无需依赖于其他任何信息，相机和场景结构重建都可以直接从图像匹配关系求解。接下来，我们将对对极几何和点匹配关系对相机和场景3D结构的重建进行描述，不过我们将先简单介绍相机模型以为后续描述作铺垫。

2.1.1 相机模型

相机实现了3D世界到2D图像的映射，作为一般相机投影模型的特例，本文主要讨论相机中心投影模型，结合投影几何工具进行解析，因此相机平面和投影中心等相机参数均可用矩阵表示进行计算。特别的，我们将主要分析基本的针孔相机模型，比如CCD相机就是这一类，当然也可泛化到其他的相机种类。

假设投影中心为欧几里得坐标系系统的坐标原点，并且相机平面在\$Z = f\$处。在针孔摄像机模型下，如图4所示，以相机中心为相机坐标系原点，空间中的一点\$X = (x_w, y_w, z_w)\$通过中心投影在相机平面上点\$x = (x_c, y_c, z_c)\$处。通过简单的三角相似性，我

们可以用公式(2.3)描述世界坐标到图像平面点坐标的关系:

$$\begin{bmatrix} x_c \\ y_c \\ z_c \\ 1 \end{bmatrix} = \begin{bmatrix} f \times x_w / z_w \\ f \times y_w / z_w \\ f \\ 1 \end{bmatrix} \quad (2.3)$$

这是一个从三维空间到二维空间的映射。投影中心亦称为相机中心，通过相机中心并且垂直于相机平面的线称为相机主线，而主线与相机平面的交点称为主点。并且，穿过相机中心平行于相机平面的平面称为相机主平面。如果世界坐标和图像坐标用齐次向量表示，那么中心投影则可在齐次坐标下通过简单的线性映射得到。特别的，公式(2.3)则可表示为矩阵乘法：

$$\begin{bmatrix} x_c \\ y_c \\ z_c \\ 1 \end{bmatrix} = \begin{bmatrix} f / z_w & 0 & 0 & 0 \\ 0 & f / z_w & 0 & 0 \\ 0 & 0 & f / z_w & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \cdot \begin{bmatrix} x_w \\ y_w \\ z_w \\ 1 \end{bmatrix} = \begin{bmatrix} K & 0 \\ 0^T & 1 \end{bmatrix} \cdot \begin{bmatrix} x_w \\ y_w \\ z_w \\ 1 \end{bmatrix} \quad (2.4)$$

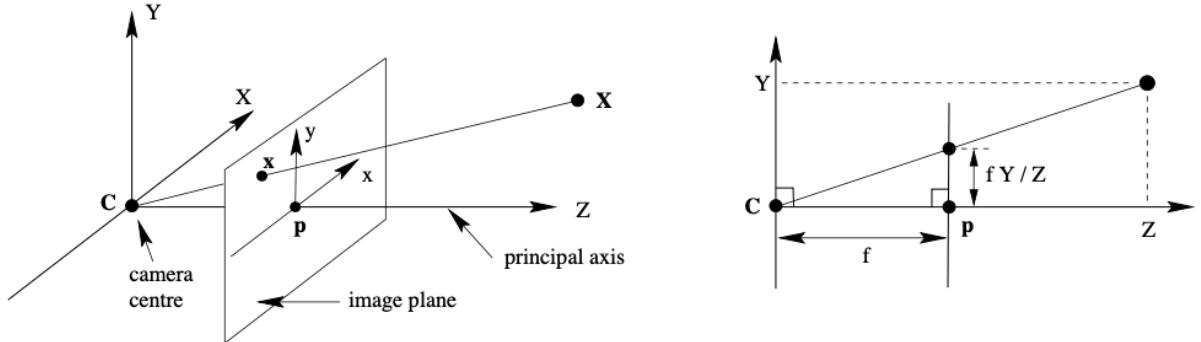


图 4 针孔相机几何模型

一般情况下，空间中的点可以表示为不同的欧几里得坐标帧，也就是世界坐标帧。而相机坐标和世界坐标通过平移和旋转矩阵联系，如图5所示。那么根据旋转矩阵R与平移矩阵t，我们可以得到：

$$\begin{bmatrix} x_w \\ y_w \\ z_w \\ 1 \end{bmatrix} = \begin{bmatrix} R & t \\ 0^T & 1 \end{bmatrix} \begin{bmatrix} x'_w \\ y'_w \\ z'_w \\ 1 \end{bmatrix} = \begin{bmatrix} r_{11} & r_{12} & r_{13} & t_1 \\ r_{21} & r_{22} & r_{23} & t_2 \\ r_{31} & r_{32} & r_{33} & t_3 \\ 0 & 0 & 0 & 1 \end{bmatrix} \cdot \begin{bmatrix} x'_w \\ y'_w \\ z'_w \\ 1 \end{bmatrix} \quad (2.5)$$

其中 $[x_w, y_w, z_w]$ 为空间点在相机坐标系下坐标， $[x'_w, y'_w, z'_w]$ 为空间点在世界坐标系下坐标。

结合公式(2.4)和公式(2.5)，我们可以得到空间中一点到图像平面的转换关系：

$$\begin{bmatrix} x_c \\ y_c \\ z_c \\ 1 \end{bmatrix} = \begin{bmatrix} K & 0 \\ 0^T & 1 \end{bmatrix} \cdot \begin{bmatrix} R & t \\ 0^T & 1 \end{bmatrix} \cdot \begin{bmatrix} x_w \\ y_w \\ z_w \\ 1 \end{bmatrix} \quad (2.6)$$

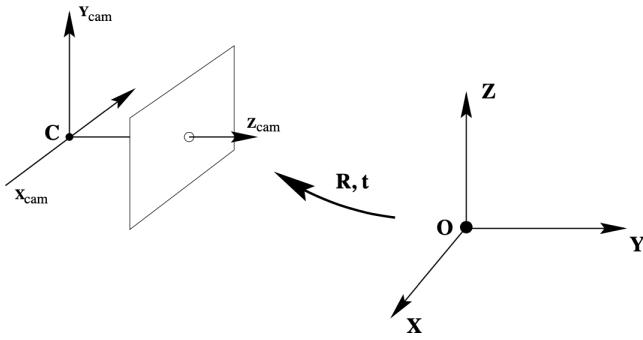


图 5 世界坐标与相机坐标间的欧几里得转换

2.1.2 对极几何

对极几何是两个视图之间的内部投影几何关系，它只与相机内部参数和相对位姿有关，而与场景结构信息独立。基本矩阵F则集中了这几个内在几何关系的精华。它是一个秩为2的 3×3 矩阵。如图6所示，如果3D空间中的一点X在第一个视图中映射为点x，在第二个视图中映射为x'，则满足关系 $x'^T F x = 0$ 。

那么图像中的点x和x'在空间上有什么样的关系呢。如图6 (a) 所示，图像中的点x和x'、空间点X和相机中心共平面 π 。以x和x'为起点的射线相交于空间点X。当然，各射线也是共平面的，这个性质是后续左右视图对应点搜索的关键。

平面 π 是由基线和以点x为起点的射线所共同决定的。从以上描述可知以点x'为起点的射线也在平面 π 内，因此对应点x'一定坐落在第二个视图与平面 π 相交的直线 l' 上。直线 l' 称为点x的对极线。从这里可以看出，得益于该对极线约束，双目匹配方法只需在对极线 l' 上计算匹配点，而无需在整个图像范围内匹配。

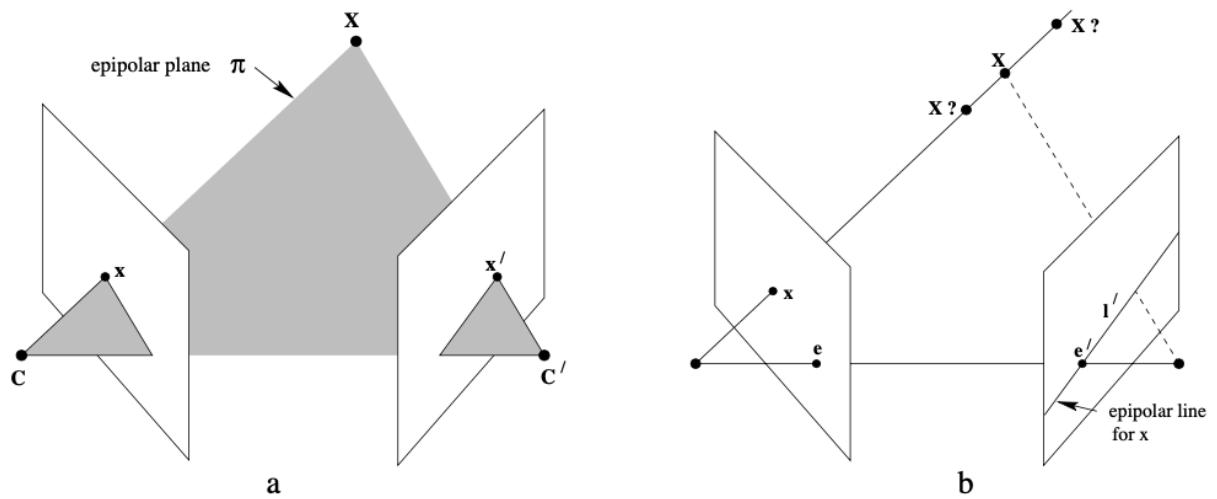


图 6 外极线约束示意图

综上，对极几何主要存在以下几何元素：

- 对极点为相机中心连线（即基线）与图像平面的交点，因此，对极点是在一视图中另一视图相机中心的像，也是基线方向的消失点。
- 对极平面为包含基线的平面。
- 对极线为极平面与图像平面的交线。所有的对极线都相交于对极点。一个对极面与左右图像平面交叉于对极线，也就是对极线之间存在对应关系。

本质矩阵 F 是对极几何的代数表示，如图6 (b) 表示，它描述的是一视图中一点 x 与另一视图对极线 l' 之间的映射关系。求解本质矩阵 F 可以分解为两步。第一步，将点 x 映射到另一视图中坐落在对极线 l' 的点 x' ；第二步，极线 l' 通过点 x' 与对极点 e' 的连线得到。

2.1.3 相机和场景 3D 结构重建

给定一系列的对应点 $x_i \leftrightarrow x'_i$ ，并且假设这些对应点来自于3D空间中的未知点集 X_i ，可直接恢复场景的空间结构和相机空间信息。重建任务关键在于重建相机矩阵 P 、 P' 和3D点 X_i 并满足如下关系：

$$x_i = PX_i, \quad x'_i = P'X_i \quad \text{for all } i. \quad (2.7)$$

给定足够多的点，是可以独一无二的求解出相机矩阵的，这也是未校正方法的魅力所在，不过恢复的重建场景仍存在一定得投影模糊，可通过相机和场景附加信息消除。

根据两视图进行重构的方法包含以下步骤：

- 从点匹配关系中计算基本矩阵 F 。
- 从基本矩阵 F 中计算相机矩阵。
- 对于每个点对 $x_i \leftrightarrow x'_i$ ，计算其在3D空间中对应的点 X_i 。

基本矩阵 F 计算：给定一系列的匹配点对 $x_i \leftrightarrow x'_i$ ，对任意点对均需满足 $x'_i F x_i = 0$ 。因为 x_i 和 x'_i 均已知，对于基本矩阵 F 中的所有参数求解方程都是线性的。实际上，每个点对都能生成与基本矩阵参数的一个线性方程，一般给定8个点对就能得到基本矩阵的解。若多于8个点对关系，还可用最小二乘法查找。

相机矩阵求解：从基本矩阵 F 计算相机矩阵 P 、 P' 则可以直接由上述基本矩阵 F 的分解过程得到。

3D空间点求解：给定相机矩阵 P 、 P' ，并且两视图中的点 x_i 和 x'_i 满足 $x'_i F x_i = 0$ 关系，3D空间信息可根据三角形法求解。根据对极几何关系，3D空间点 X_i 存在于分别以点 x_i 和 x'_i 为起点的射线平面内。特别的， x'_i 坐落在对极线 $F x_i$ 上，由此可推测得到两视图中的以点 x_i 和 x'_i 为起点的射线反向投影将经过相机中心。因此，在已知相机中心与匹配点对

点 x_i 和 x'_i 共平面的条件下，3D空间点 X_i 必存在与射线交叉点处，也就是说，3D空间点 X_i 通过两相机中心投影交图像平面于点 x_i 和 x'_i 处。具体可参加图7。

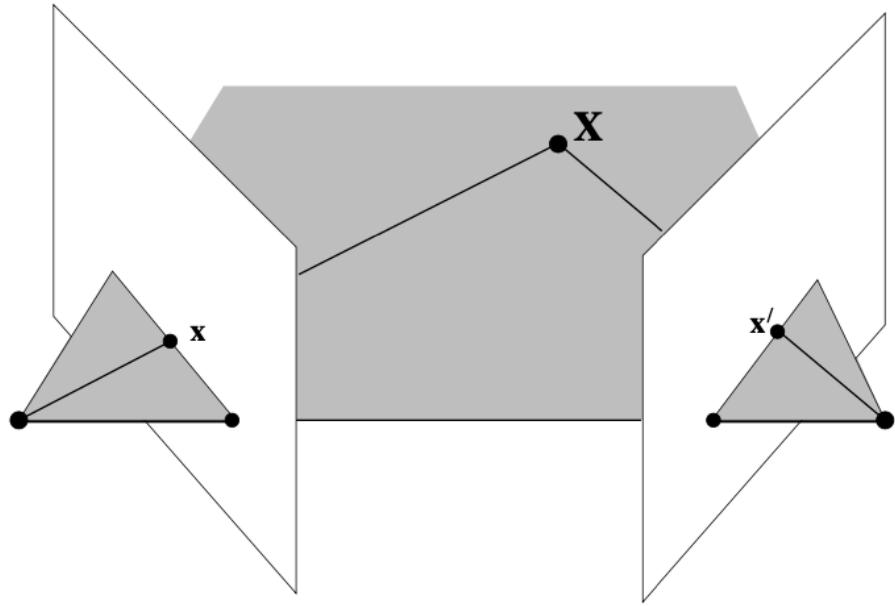


图 7 三角形法求解 3D 空间信息

2.2 基于传统算法的立体匹配方法

经过几十年的研究，双目立体匹配这一课题已经有了成熟的理论体系，KITTI^{[15][16]}, Middlebury^[14]等标准立体数据集也相继被构建出来。2002年，Scharstein和Szeliski^[14]提出了一种普遍适用于双目深度估计的算法框架，划分了整个深度估计过程的四个步骤：匹配代价计算，匹配代价聚合，视差计算和优化，视差图细化。

(1) 匹配代价计算

计算匹配代价，即计算参考图像上的每个像素点 $I_R(\mathbf{p})$ 与目标图像上的对应点 $I_T(\mathbf{pd})$ 在所有视差可能性范围内的代价值，因此计算得到的代价值可以存储在一个 $h \times w \times d_{max}$ 的三维数组 $C(x, y, d)$ 中，其中 h, w, d_{max} 分别为图像的长、宽和最大的视差值，通常称这个三维数组为视差空间图（Disparity Space Image, DSI）或者匹配代价体（Cost Volume）。匹配代价是立体匹配的基础，设计一种能够抵抗噪声干扰并且对照明变化不敏感的匹配代价可以提高立体匹配的准确性。最常用的基于像素级别的匹配代价计算方法包含差的绝对值（AD）、差的平方和（SD）、归一化互相关（NCC）等。在视频处理任务中，这些匹配标准称为均方误差（MSE）和平均绝对差（MAD）量度，也经常用术语“位移帧差异”描述^[41]。不过，相对鲁邦的度量方法，比如截断二次方、加噪高斯

分布等方法^[42]后来也被相继提出。因为这些方法限制了聚合过程不匹配像素点的影响，所以也更加实用。其他传统匹配代价包括归一化互相关^[17]，其原理类似于平方差和（SSD），以及二进制匹配代价和基于边缘的二元特征等方法。但是，二进制匹配成本在密集的立体匹配方法中并不常用。也有一些代价计算方法对摄像机增益或偏差的差异不敏感，例如基于梯度的度量和非参数度量（比如排序或者CENSUS转换^[17]）。当然，也可以通过执行用于偏置增益或直方图均衡的预处理步骤来校正不同的相机特性。其他匹配标准包括相位和滤波器组响应。最后，Birchfield和Tomasi^[44]提出了对图像采样不敏感的匹配代价。他们不只是比较整数视差值的像素（这可能会错过有效匹配），而是将参考图像中的每个像素与其他图像的线性插值函数进行比较。

(2) 匹配代价聚合

一般局部或基于窗口的方法通过对DSI: $C(x, y, d)$ 中所支持区域求和或求平均值来汇总匹配代价。这个支持区域可以是固定视差的二维平面（比如一个平行面），也可以是 $x - y - d$ 空间中的三维平面（比如倾斜表面）。而窗口的选择可以使用正方形窗口或高斯卷积，通过在不同像素位置锚定的多个窗口（即可移动窗口或者自适应大小的窗口^[45]），可以实现二维维度的信息聚合。对于三维情况下的支持函数设计，包含视差差异限制，视差梯度限制和连续性限制。在一个固定区域的匹配代价聚合，一般可以通过2D或者3D卷积实现：

$$C(x, y, d) = w(x, y, d) * C_0(x, y, d) \quad (2.8)$$

当然，如果是一个长方形窗口，用盒卷积会更加有效。另一种聚合方式是通过迭代扩散的方式完成，也就是每次迭代操作是通过对相邻像素点的匹配代价进行加权求和用于更新当前点的匹配代价。

(3) 视差计算与优化

基于局部的方法：在这一类方法中，主要强调的是匹配代价计算和匹配代价聚合过程，而计算最终的视差值是相对简单的：对于每个像素点，只是简单的取最小的匹配代价所对应的视差值，也就是对局部使用“胜者为王”策略（Winner Takes All,WTA）。这种方法的局限性在于只强调匹配结果的单一性，但是实际情况可能存在多个匹配点。

基于全局优化的方法：和基于局部的方法不同，基于全局的方法将工作的重点放在视差计算阶段，而经常会直接跳过聚合阶段。许多全局化方法是在能量最小化框架^[46]中定义问题。目标是找到一个视差 d 使得全局能量最小化：

$$E(d) = E_{data}(d) + \lambda E_{smooth}(d) \quad (2.9)$$

其中数据项 $E_{data}(d)$ 定义了视差 d 的情况下输入图像对的匹配程度。一般是在视差空间下定义的：

$$E_{data}(d) = \sum_{(x,y)} C(x, y, d(x, y)) \quad (2.10)$$

其中 C 为（初始的或者聚合后的）匹配代价视差空间（DSI）。

而 $E_{smooth}(d)$ 是根据算法设计定义的光滑假设。为了让计算损耗限制在一定范围内，光滑项一般限定在度量相邻像素点间的差异：

$$E_{smooth}(d) = \sum_{(x,y)} \rho(d(x, y) - d(x + 1, y)) + \rho(d(x, y) - d(x, y + 1)) \quad (2.11)$$

其中 ρ 为视差差异的单调增函数。在基于正则化的视觉任务中， ρ 为二次函数，能够实现视差图全局光滑，但是会导致物体边缘的结果较差。而基于边缘保留的能量函数就可以克服这个问题，Geman 等人^[47]给出贝叶斯插值能量函数，并且基于马尔科夫随机场（MRFs）提出一种非连续性保留的能量函数，这种能量函数可以线性时间内处理。当然， $E_{smooth}(d)$ 还可以根据图像像素值差异进行定义，比如：

$$\rho_d(d(x, y) - d(x + 1, y)) \cdot \rho_I(I(x, y) - I(x + 1, y)) \quad (2.12)$$

其中 ρ_I 为像素值差异的单调减函数，能够在图像像素值梯度大的地方减少光滑代价。这个想法的本意是鼓励视差不连续的地方和图像边缘对齐，以此取得一个比较好的全局优化结果。

一旦全局能量函数被定义，大量的算法即可用于寻找能量的最小值。传统方法中，结合正则化和马尔科夫随机场的方法包含模拟退火^[48]和平均场退火^[49]。相对更加有效的方法，比如最大流和图切割法^[21]被提出用于求解某一类特殊的全局优化问题。

动态规划：相对特殊的一类全局优化方法是基于动态规划的算法。公式(2.9)中的 2D 优化问题为 NP 难问题，动态规划可以在多项式时间内找到独立扫描线的全局最小值。动态规划第一次被引入立体视觉任务是基于边缘的方法，而后才集中于密集的扫描线优化问题^[50]。这些方法通过计算两条相应扫描线之间所有成对匹配代价矩阵的最小代价路径来实现全局优化。相应的，动态规划应用到立体匹配任务所面临的挑战是对遮挡像素点分配合适的匹配代价以及强调扫描线间的一致性。同时单调性和有序性也是使用动态规划的两个必备前提条件，此约束要求在两个视图之间的扫描线上像素的相对顺序必须保持相同，而在包含狭窄前景对象的场景中可能就不能被满足。

协作算法：最后，受人类立体视觉计算模型启发的协作算法是最早提出的视差计算方法之一。这样的算法反复执行局部计算，但是使用非线性运算，从而导致总体行为类似于全局优化算法。实际上，对于其中的某些算法，可以显式建立全局函数用于最小化^[42]。

(4) 视差图细化

大多数立体匹配算法在某些离散空间中计算一组视差估计值，一般情况下为整数视差。对于诸如机器人导航或人员跟踪的应用，这些可能就足够了。但是，对于基于图像的渲染，这样的量化结果图会导致非常糟糕的视图合成结果（场景似乎由许多薄的剪切层组成）。为了纠正这种情况，许多算法在初始的离散匹配阶段之后采用亚像素优化以获得亚像素的匹配视差图。

亚像素视差估计可以由许多方式计算得到，其中包含迭代的降梯度算法，以及在一个离散的视差水平拟合一个匹配代价曲线^[51]。这种方法简单易实现，计算量小，并且增加了立体匹配的分辨率。但是该方法有效的前提是保证待匹配的像素点强度变换连续且光滑，而且这些像素点所在的区域要在同一平面上。

除了亚像素计算，还有许多后处理方法用于细化视差图。遮挡区域的检测可以采用左右一致性检查；中值滤波可以用于清除不匹配点；由于遮挡造成的视差图空洞可以通过平面拟合填充，当然也可以通过相邻视差值插值。

2.2 基于深度学习的立体匹配方法

传统的立体匹配算法多围绕匹配成本计算和视差优化进行研究，受人工设计的局限性，对于病态区域（如遮挡，纹理少的区域等）往往表现出不尽人意的效果。基于深度学习的立体匹配方法则精度上远超人工设计的相应算法，这也表明了深度学习方法对上下文信息提取和语义信息分析的卓越能力。

(1) 神经网络用于匹配代价计算

这一类方法主要是集中于设计巧妙且有效的网络结构用于计算两个图像块之间的匹配代价计算。网络的输出代表图像块中心像素点的匹配代价，而其他匹配过程仍然采用手工设计的方法，比如基于交叉的匹配代价聚合，半全局匹配^[24]，左右一致性检查，亚像素增强^[51]以及双边滤波^[52]。为了利用神经网络计算匹配代价，Zbontar和LeCun^[29]堆积了几个卷积层用于提取左右 9×9 图像块间的匹配代价。从图像块提取的特征向量要么计算点积，要么直接衔接在一起，然后用几层全连接层计算相似性。这种方法计算量很

大，因为需要多个前向过程（次数等于视差参选值的数量）在所有的视差参选值上计算匹配代价。基于这个框架，Luo等人^[30]设计了一个孪生网络提取图像块特征，并利用乘积操作计算匹配代价，最终在多分类标记的监督下训练网络，在保证精度的条件下，得到了一个更快的匹配网络架构。Park和Lee等人^[53]为了提升网络感受野，提出一个金字塔池化方案，相比Zbontar和LeCun^[29]取得了更加准确的匹配结果。Shaked和Wolf等人^[54]则提出利用多层加权残差短接的方式拼接一个高速网络结构并用于增加匹配代价计算网络深度，有效提升了网络特征提取及表示能力。这些方法展示了提升网络匹配代价计算能力是提升立体匹配精度的根本，但其他步骤仍需依赖手工方法是这一类方法的最大缺点和不足。

(2) 神经网络用于视差图细化

这一类方法设计卷积神经网络解决视差图的后处理和细化。Gidaris和Komodakis等人^[55]采用^[30]的方法获得初始视差估计结果，然后附加三个神经网络用于视差细化。特别地，第一个子网络用于检测初始视差图中错误的区域，第二个网络用于将错误区域的视差预测结果替换掉，而第三个网络则用于对替换后的视差结果进一步优化和光滑化，最终产生更加精确的匹配结果。Pang等人^[56]则模型化视差细化过程为残差学习过程，通过一个网络学习初始视差图与真实视差图之间的差量。StereoNet^[37]则首先生成一个低分辨率的视差图，接着用边缘已知的上采样子网络进行优化。对于初始的视差图，利用上采样操作对视差图上采样两倍，并把原始彩色图像下采样到对应尺度，根据视差图和原始图像信息，优化子网络会输出一个残差视差图并作为增量与上采样的初始视差图相加，叠加后的结果即为优化结果。左右递归对比网络（LRCR）^[57]将左右一致性检查与视差生成集成到一个统一的网络中。具体的，一个递归神经网络用于学习左右一致性并作为一种软集中引导机制让模型能够选择性的对某些区域进行优化。Batsos和Mordohai等人^[58]也用一个递归神经网络做视差细化。和左右递归对比网络^[57]不同的是，他们根据多尺度残差信息估计优化视差图，这也使得他们的方法能够矫正不同类型的错误，比如不同区域可能需要不同程度的光滑化。Ye等人^[59]的优化方案主要受^[55]启发，在离群点检测前，先将初始最优的或者局部最优的视差图囊括在一起，然后用不同的基学习器对光滑区域和细节部分进行优化。和^[7]的方法差不多，Liang等人的视差优化方案侧重于对初始视差估计和视差优化阶段分享特征并且在特征空间计算重建误差，从而得到一个更加压缩并且可解释的网络结构。

(3) 端到端的立体匹配网络

这一类方法直接端到端的回归视差图，没有后处理和正则化操作，主要依赖于特征提取网络的潜能以及卷积神经网络的拟合能力。在网络的顶端，一个度量估计视差与真实视差差异的回归损失函数被用于监督整个立体匹配网络的学习，估计的误差会通过反向传播用于更新网络参数。

基于这个设计准则，Mayer等人^[32]提出了一个编解码网络结构并且端到端回归视差，匹配代价计算是沿着视差维度计算左右特征图之间的一维相关性（1-D correlation），相当于一种乘积近似，并且创建了一个大型合成数据集用于训练视差估计（以及光流）网络，从而改善了现有技术。但是只用一维相关近似匹配代价会丢失大量的信息，因此Gwc-Net^[61]提出组相关（group-wise correlation）的匹配代价计算函数。它的核心原理在于将左右特征图在通道维度进行分组，然后计算左右各组之间的相关性，相比Mayer等人^[32]对每个通道都进行correlation操作能够保留更多的原始图像信息。

而Kendall^[33]等人则提出一种基于3D卷积端到端学习视差的方法。和Mayer等人^[32]不同的是，该方法对网络提取的左右图像特征采用衔接的方式构造三维视差空间图DSI: $C(x, y, d)$ ，也称为代价体，并且利用堆叠的3D卷积结构提取特征并实现上下文信息融合，最后依赖soft argmin（公式(3.1)）视差回归算法，从而实现端到端的亚像素精度的视差学习，无需额外的后处理和正则化。这套匹配代价构建模式被不断沿用和改善，比如PSMNet^[34]、GANet^[35]，并且成为现有方法中精度最高的网络框架。具体的，PSMNet^[34]主要强化了特征提取和代价聚合过程。特征提取阶段，空间池化金字塔结构被用于提取多尺度上下文信息；代价聚合网络则引入多个堆叠的沙漏3D卷积神经网络结构，并且对每个结构都添加中间监督信号，实现了更加高效和全面的上下文信息提取实现。不过GCNet^[33]和PSMNet^[34]为了实现视差空间内的代价有效聚合，引入了大量的3D卷积层，比如PSMNet中有25层3D卷积，而这庞大的计算量，对计算资源的消耗是十分严重的，而且对实时应用十分不友好。而GANet^[35]则主要是将传统经典方法SGM^[24]也嵌入到网络中，并且实现可微分和梯度可反向传播的运算。虽然Sgm-nets^[62]也对SGM算法进行了网络可学习化，但也只是停留在某些超参上的可学习，所以在代价聚合方面，其效果和GANet还是相距甚远。

不过，这些方法所共存的一个问题是视差回归完全依赖于soft argmin算法，而该算法在回归过程中受所有视差模态影响，所得到的匹配代价经常呈现多峰，回归得到的视差值也不会对应匹配代价最小的视差值，如图1所示。相应的，图像特征提取和代价计算函数也无法得到有效的学习。虽然也有些方法尝试把立体匹配作为多分类任务训练和学

习，比如Tulyakov等人^[31]设计离散的拉普拉斯分布软化多分类标签，但是这些基于分类的方法将视差回归和网络训练分离，他们一般需要传统的正则化项和后处理才能得到光滑的视差图。而且自然场景中的立体匹配任务是复杂且多变的，而固定每个像素点的多分类标签保持一致是不合理也无法应对所有情况，甚至容易导致网络过拟合。

(4) 多任务结合的立体匹配算法

这类方法通过引入多任务学习从而使得立体匹配算法能够受利于其他任务中有相互促进的语义信息。EdgeStereo^[63]从边缘检测任务中提取中级特征信息用于恢复视差图估计中缺失的细节信息；SegStereo^[64]则将语义特征信息嵌入到神经网络中，并且利用语义分割损失项对视差图学习施加约束项。置信度估计^{[68][69][70]}则致力于检测匹配过程中失配的像素点并根据这些信息有针对性的提出解决方案，比如^[70]则通过置信度分数对每个像素点的匹配代价分布进行调节，最终保留置信度高的像素点对应的匹配代价而抑制置信度低的像素点，最终提升视差估计精度。

(5) 实时立体匹配网络

现有基于深度学习的立体匹配网络一般对计算资源要求较大，且十分消耗显存，对实时应用不太友好，比如目前的state-of-the-art网络PSMNet在Nvidia Jetson TX2 GPU上的帧率甚至低于0.3帧每秒。而其中也有不少工作在实时立体匹配算法设计上起着积极的推动作用。其中DispNetC^[32]，Tonioni等人^[36]，StereoNet^[37]，AnyNet^[38]等工作均采用由粗到精的网络结构，通过大尺度的下采样操作，有效减少网络计算量和推断时间，但细节丢失却是这一类方法的致命问题，因此这些工作在视差图后处理和正则化方面都提出了各自的解决方案。而DeepPruner^[39]则是另辟蹊径，从视差搜索空间剪枝进而减少不必要的计算量，有效提升了网络视差匹配效率。不过缺陷是仅依靠上下文信息有限的两个并行分支网络预测剪枝后空间的上下界是病态的，比如一些遮挡或者无条纹区域的上下界预测是不准确的，那么后续的优化过程也将是在错误的视差搜索空间进行。与DeepPruner^[39]不同，我们提出一个轻量级的视差参选值推荐网络，在保证真实视差在剪枝后的视差搜索空间的前提下，得到压缩的视差搜索空间，并利用3D代价聚合网络对视差参选样本进行评估和最终的视差预测，实现了更加高效和高精度的立体匹配过程。

2.3 本章小结

本章主要介绍了双目立体视觉成像原理，传统和基于深度学习的立体匹配算法。

对于双目立体视觉成像原理，主要从相机模型、对极几何、相机和场景3D结构重建

三个部分进行介绍。

对于立体匹配算法设计，2002年Scharstein和Szeliski^[14]提出的深度估计框架，即匹配代价计算、匹配代价聚合、视差计算和优化、视差图细化，不仅准确地总结了先前工作的核心算法流程，后续提出的许多启发性工作和改进工作基本也是围绕这个框架和流程进行。对于传统立体匹配算法，本文依照该算法框架进行展开介绍和详细对比分析。总体而言，传统算法主要受限于手工设计的函数表示能力有限，无法得到一种普适的算法能应付各种各样的场景结构和变化因素。

对于基于深度学习的立体匹配算法，则主要受益于深度卷积神经网络强大的学习能力和特征抽象能力，而随之带来的问题是计算资源的需要增大，且十分消耗显存。推进实时应用的工作主要是采用由粗到精的网络结构和视差搜索空间剪枝策略，但性能与速度上仍未达到一个理想的均衡状态。

综上所述，针对现有端到端的立体匹配算法，我们可采用自适应的单峰匹配代价滤波方法，并且设计视差参选值推荐网络，大幅提升网络视差推断效率，且保持网络鲁棒的匹配效果。

第三章 自适应的单峰匹配代价滤波

3.1 引言

理想情况下，每个像素点的匹配代价分布是以真实视差为中心的单峰分布。为了明确约束网络学习这种代价分布以学习更加鲁棒的图像特征和代价计算函数，我们提出根据真实视差图为每个像素生成以真实视差为中心的单峰代价分布，并且用于对网络预测的匹配代价体（Cost Volume）直接施加监督。为了揭示每个像素点的匹配不确定度，我们设计了一个置信度估计网络去估计每个像素点的置信度并用于调节其对应的真实单峰分布。图8详细展示了我们的整体网络结构框架。由于PSMNet^[34]是目前的state-of-the-art立体匹配模型，我们采用它作为我们的基础网络。对于输入的左右图像对，PSMNet输出经过堆叠的3个沙漏3D卷积神经网络聚合后的匹配代价体；对于每个匹配代价体，我们分别用一个置信度评估网络（Confidence Estimation Network）估计置信度图并用于调节真实的匹配代价体（Ground Truth Cost Volume），以生成像素级别的单峰分布（Unimodal Distribution）作为网络训练标记；并且我们提出立体聚焦损失（Stereo Focal Loss）约束估计的和真实的匹配代价体之间的一致性。最终，通过Soft Argmin函数根据估计的匹配代价体生成亚像素的视差图，并且计算 L_1 损失用于监督估计的和真实的视差图。算法伪代码如下所示：

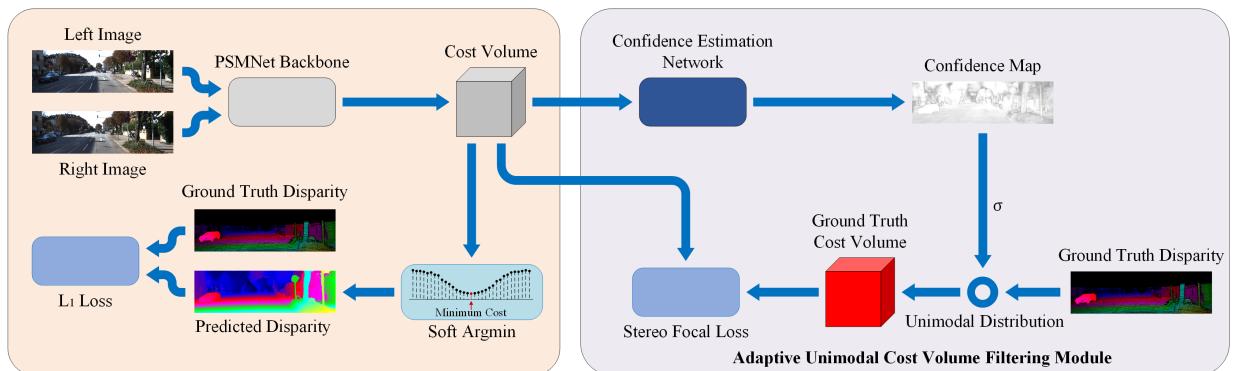


图 8 AcfNet 网络结构框架

3.2 基于 3DCNN 的立体匹配算法

在矫正的图像对中，对于左图中的每个像素点 $p(x, y)$ ，双目立体匹配的目标是找到右图中的对应点，也就是 $p'(x + d, y), d \in \mathbb{R}^+$ ，对于亚像素的立体匹配，视差值 d 为浮点数。为了方便计算和内存访问，视差一般离散为一些列可能的视差参选值，也就是 $\{0, 1, \dots, D - 1\}$ ，因此可以构建一个 $H \times W \times D$ 的匹配代价体（Cost Volume），其中 H, W, D

分别为图像高度、宽度和最大视差值。为了恢复亚像素的视差匹配结果，每个视差对应的匹配代价都会用于加权插值获得最终的视差回归。整个网络实现过程如图9所示。

对于采用的PSMNet模型，其网络结构如图9所示，主要由4个部分组成，分别是特征提取，匹配代价体计算，基于3D卷积神经网络（CNN）的匹配代价聚合和视差回归。

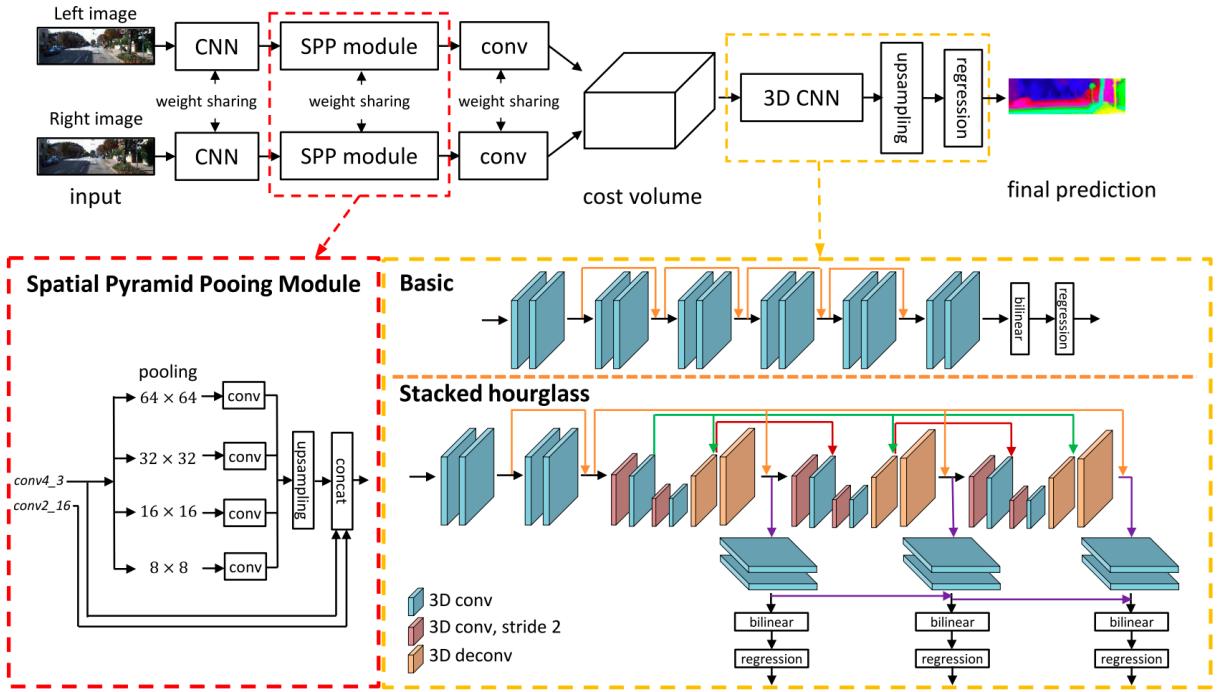


图 9 PSMNet 网络结构图

仅从像素强度确定上下文关系是十分困难的，而包含物体级别的语义信息将对匹配十分有利，特别是对于病态区域的视差估计。为了学习和提取物体之间的相对关系，PSMNet提出了空间金字塔模块（Spatial Pyramid Pooling Module）用于图像特征提取，通过4个并行的固定大小的平均池化模块，最终的图像特征表示包含多尺度的上下文信息。匹配代价体的构成采用的是GCNet^[33]的衔接方式，为左右特征匹配保留了最原始的图像信息。为了在视差维度和空间维度聚集特征信息，PSMNet提出了一种沙漏型（也就是编解码结构）的3D CNN架构，包含重复的从上到下、自底向上的处理过程，还对网络三个阶段输出的匹配代价体都进行监督学习。整体而言，PSMNet实现了非常优越的立体匹配性能，也是选择该网络框架作为我们基础网络的原因。

一般而言，匹配代价体（Cost Volume）为每个像素点都构建了 D 个匹配代价 $\{c_0, c_1, \dots, c_{D-1}\}$ ，即匹配代价分布。为了从该分布中估计亚像素的视差值，GCNet^[33]提出使用soft argmin函数进行回归：

$$\text{soft argmin} : \hat{d} = \sum_{d=0}^{D-1} d \times \hat{P}(d) \quad (3.1)$$

其中，

$$\hat{P}(d) = \text{softmax}(-c_d) = \frac{\exp(-c_d)}{\sum_{d'=0}^{D-1} \exp(-c_{d'})} \quad (3.2)$$

匹配代价最小的视差值对最终的插值结果贡献越大。网络训练阶段，对于像素点 p ，其真实视差值为 d_p ，一般采用smooth L_1 损失约束：

$$\mathcal{L}_{\text{regression}} = \frac{1}{|\mathcal{P}|} \sum_{p \in \mathcal{P}} \text{smooth}_{L_1}(d_p - \hat{d}_p) \quad (3.3)$$

其中

$$\text{smooth}_{L_1} = \begin{cases} 0.5x^2, & \text{if } |x| < 1 \\ |x| - 0.5, & \text{otherwise} \end{cases} \quad (3.4)$$

由于整个过程都是可导的，所以网络可直接用真实视差图进行监督训练。但是从公式(3.1)的soft argmin回归过程可以看出，匹配代价体只是作为视差插值过程的加权权重，由于只需得到真实视差值即可，这导致匹配代价体可以以任意状态参与视差插值过程（如图1所示），而对匹配代价体的数学分布无任何要求。这一事实和每个像素的匹配代价分布应该呈现单峰分布矛盾，而直接原因则是缺乏对匹配代价分布的直接监督约束，这也启发我们提出自适应的单峰匹配代价滤波方案。

3.3 自适应的单峰匹配代价滤波模块

如图8所示，我们提出的AcfNet网络结构在PSMNet的基础上，仅嵌入一个自适应的单峰匹配代价滤波模块即可完成对单峰匹配代价分布的学习。对于PSMNet输出的3个经过聚合后的匹配代价体，由单峰匹配代价分布生成、置信度估计网络、立体聚焦损失三个部分实现自适应的滤波效果。算法伪代码如下：

算法1：自适应的单峰匹配代价滤波

输入数据：PSMNet网络估计的匹配代价 c ，自信度评估网络confidence_model，真实视差图 d_{gt} ，单峰匹配代价方差 σ ，方差调整常数 $s = 1.0$ ， $\epsilon = 1.0$ ，最大视差值 D ，立体聚焦损失 \mathcal{L}_{SF}

输出数据：滤波后的匹配代价 c'

```

1: FUNCTION filter( $c$ , confidence_model,  $d_{gt}$ ,  $\sigma$ ,  $s$ ,  $\epsilon$ ):
2:    $f \leftarrow \text{confidence\_model}(c)$  // 估计每个像素的匹配自信度

```

```

3:    $\sigma \leftarrow s \times (1 - f) + \epsilon$  // 根据匹配自信度计算单峰匹配代价方差
4:   FOR  $d \leftarrow 0$  TO  $D - 1$  DO:
5:      $c_d^{gt} = \frac{|d - d^{gt}|}{\sigma}$  // 根据 $\sigma$ 和 $d_{gt}$ 为每个像素点生成真实匹配代价 $c^{gt}$ 
6:   END FOR
7:    $c' \leftarrow \mathcal{L}_{SF}(c, c^{gt})$  // 约束 $c$ 与 $c^{gt}$ 保持一致, 得到滤波后的匹配代价 $c'$ 
8:   RETURN  $c'$ 
9: END FUNCTION

```

3.3.1 单峰匹配代价分布生成

匹配代价体反应的是待匹配像素对的相似性, 真实的匹配对之间的匹配代价应该是最小的, 而其他参数视差值的匹配代价应该随着和真实视差的距离而增加。这个性质要求每个像素的匹配代价分布都应该以真实视差为中心。给定真实视差 d^{gt} , 单峰分布定义为:

$$P(d) = softmax\left(-\frac{|d - d^{gt}|}{\sigma}\right) = \frac{\exp(-c_d^{gt})}{\sum_{d'=0}^{D-1} \exp(-c_{d'}^{gt})} \quad (3.5)$$

其中 $c_d^{gt} = \frac{|d - d^{gt}|}{\sigma}$, $\sigma > 0$ 为方差, 可以控制真实视差周围的峰型尖锐程度。

一般情况下, 每个像素点的上下文信息是不一样的。因此, 让每个像素点保持一致的真实匹配代价分布 $P(d)$ 是不合理的。比如对于位于桌子角上的一个像素点更偏向于十分锋利的单峰, 而对于纹理少的区域, 则更偏向于相对平缓的分布。为建立更加合理的匹配代价分布, 我们设计了一个置信度评估网络去自适应的调节每个像素点 p 的单峰分布方差 σ_p 。

3.3.2 置信度估计网络

在传统的置信度评估方法中, 大量的研究工作^{[68][69]}集中于研究聚合后的匹配代价分布曲线, 进而有效的检测出预测的视差图中的离群点并用于提升视差图预测准确率。其中Park和Yoon^[70]提出基于置信度引导的匹配代价滤波方法。这些方法一般都是直接将置信度评估作为先验信息或者附加特征信息优化匹配代价和视差图, 但是我们直接根据网络预测的置信度分数用于调节真实匹配代价分布的平缓度, 从而让每个像素能够根据上下文信息自适应的调节单峰分布平缓程度。具体的, 我们设计了一个置信度评估子网,

包含一个 3×3 的卷积(Convolution)层，一个批归一化(BN)层，一个激活层(ReLU)，和一个 3×3 卷积层，最后接一个激活层(Sigmoid)输出估计的置信度图。对于输入的经过聚合后的匹配代价体，网络直接输出置信度图 $f \in [0, 1]^{H \times W}$ 。对于像素点 p ，如果其预测的置信度 f_p 很大，则意味着网络可以十分自信地找到其独一无二的匹配点；相反，如果其预测的置信度值很小，则意味着存在匹配模糊。因此，真实匹配代价分布的方差可以由估计的置信度值进行动态调节：

$$\sigma_p = s(1 - f_p) + \epsilon \quad (3.6)$$

其中 $s \geq 0$ 为常量，反应了方差 σ_p 对置信度值 f_p 改变的敏感程度，而 $\epsilon > 0$ 定义的是 σ 的下界，并且可以有效的防止除0的数学问题。相应的 $\sigma_p \in [\epsilon, s + \epsilon]$ 。在我们的实验中，两种类型的像素很可能具有较大的方差值 σ ：纹理少和遮挡像素。对于纹理少的区域，可能存在多个匹配的像素点；而对于遮挡像素，则找不到正确的匹配点。由于每个像素点的 σ_p 可以动态调节，真实匹配代价体可以根据公式(3.5)和(3.6)做出相应的改变。

3.3.3 立体聚焦损失

对于像素点 p ，我们已经得到了估计的匹配代价分布 $\hat{P}_p(d)$ 和真实的匹配代价分布 $P_p(d)$ ，用交叉熵损失计算分布误差是最直接的方式。但是，根据Zbontar和LeCun等人^[29]的描述，对于每个像素点，都面临着严重的视差样本不均衡问题，也就是每个像素点只有一个真实的视差值(正样本)和上百个不匹配的视差值(负样本)。因此，受focal loss^[71]解决一阶段目标检测中样本不均衡的启发，我们提出了立体聚焦损失(Stereo Focal Loss)用于聚焦正样本的预测，以防网络训练被负样本所主导，

$$\mathcal{L}_{SF} = \frac{1}{|\mathcal{P}|} \sum_{p \in \mathcal{P}} \left(\sum_{d=0}^{D-1} (1 - P_p(d))^{-\alpha} (-P_p(d) \cdot \log \hat{P}_p(d)) \right) \quad (3.7)$$

其中 $\alpha \geq 0$ 是聚焦参数，当 $\alpha = 0$ 时，该损失函数则直接退化为交叉熵损失，当 $\alpha > 0$ 时，立体聚焦损失(Stereo Focal Loss)会根据 $P_p(d)$ 的大小分配更多的权重到正视差样本上。

3.3.4 全部损失函数

总的来说，我们最终的损失函数一共包含三个部分：

$$\mathcal{L} = \mathcal{L}_{SF} + \lambda_{regression}\mathcal{L}_{regression} + \lambda_{confidence}\mathcal{L}_{Confidence} \quad (3.8)$$

其中 $\lambda_{regression}$, $\lambda_{confidence}$ 为两个权衡参数。 \mathcal{L}_{SF} 监督匹配代价体的学习, $\mathcal{L}_{regression}$ 监督视差回归, 而 $\mathcal{L}_{Confidence}$ 则作为一个正则化器鼓励更多的像素点拥有大的置信度值:

$$\mathcal{L}_{Confidence} = \frac{1}{|\mathcal{P}|} \sum_{p \in \mathcal{P}} -\log f_p \quad (3.9)$$

3.4 实验验证

3.4.1 数据库及评价指标和实现细节

(1) 数据库

为了定性和定量评估我们提出的方法 AcfNet, 我们将在三个具有挑战性的公开数据集 (Scene Flow^[32], KITTI 2012^[15], KITTI 2015^[16]) 上进行评测。Scene Flow^[32]是一个合成数据集, 包含 35454 张训练图片对和 4370 张测试图片对, 而且提供密集的真实视差标注信息, 非常适合训练和测试网络模型。KITTI 2012^[15], KITTI 2015^[16]是两个真实街景数据集, 所提供的视差标注信息都是通过雷达扫描得到的, 因此都是稀疏的。前者包含 194 张训练图片对和 195 张测试图片对; 而后者包含 200 张训练图片对和 200 张测试图片对。两个 KITTI 数据集对于训练神经网络来说数据量都太小了, 因此非常具有挑战性。因此, 我们参考 GC-Net^[33]主要在 Scene Flow^[32]上设计消融实验并且对网络设计分析。

(2) 评价指标

实验中, 我们采用两个标准的评估指标: (1) 3-pixel-error (3PE), 指预测的视差和真实的视差大于 3 像素的像素点占总数的百分比; (2) end-point-error (EPE), 指预测的视差和真实的视差的平均差异。EPE 更加注重亚像素的误差, 而 3PE 则重点刻画离群点所占的百分比。并且, 为了进一步评估 AcfNet 在处理遮挡区域上的性能, 我们将 Scene Flow^[1] 测试集根据左右一致性检查分为遮挡 (occluded regions, OCC) 和非遮挡区域 (not occluded regions, NOC)。首先, 我们标注左真实视差图 D^L 中像素点坐标为 \mathbf{p} , 那么根据以下公式:

$$\text{NOC} \quad \text{if } |d - D^R(\mathbf{p} - \mathbf{d})| \leq 1 \quad \text{for } d = D^L(\mathbf{p}), \quad (3.10)$$

$$\text{OOC} \quad \text{otherwise.} \quad (3.11)$$

其中 D^R 为右真实的视差图, $\mathbf{p} - \mathbf{d}$ 为右图中对应的位置 \mathbf{p} 往左平移 \mathbf{d} 个像素值。根据

我们的统计结果，遮挡像素占整个测试集的 16%。

(3) 实现细节

我们的AcfNet采用PyTorch实现，所有的模型都是采用RMSProp的标准设置端到端训练。对于所有数据集中的图像，都会采用颜色归一化进行数据处理。训练的时候，我们随机截取 $H = 256, W = 512$ 图像块，并且最大的视差值 D 设置为192。对于网络训练，我们随机初始化网络参数并且在Scene Flow上以一个恒定的学习率0.001训练10个周期（Epoch），并且直接用训练好的模型进行测试。对于KITTI数据集，我们用Scene Flow上预训练的模型进行微调600个周期（Epoch）。初始的微调学习率设置为0.001，并且在100和300个周期的时候衰减 $\frac{1}{3}$ 。当提交到KITTI公开榜单的时候，为了获得更好的预训练模型，我们会在Scene Flow上延长训练到20个周期（Epoch）。训练的批数据大小为3，总共3块NVIDIA GTX 1080Ti GPUs，所以每张显卡上放1个批数据。

3.4.2 实验结果及讨论

(1) 消融实验结果分析

由于Scene Flow有足够的数据量用于网络端到端训练且不用担心过拟合问题，我们的所有实验都是在Scene Flow数据集上进行。并且，在所有的实验中，Stereo Focal Loss均采用 $\alpha = 5.0$ 进行正负视差样本均衡。考虑到大多数的视差预测误差均为亚像素误差，即误差均小于1个像素点，而3PE评估的3像素误差已经无法准确的揭示网络性能，我们只利用EPE误差来研究网络在不同参数设置下的性能差异。

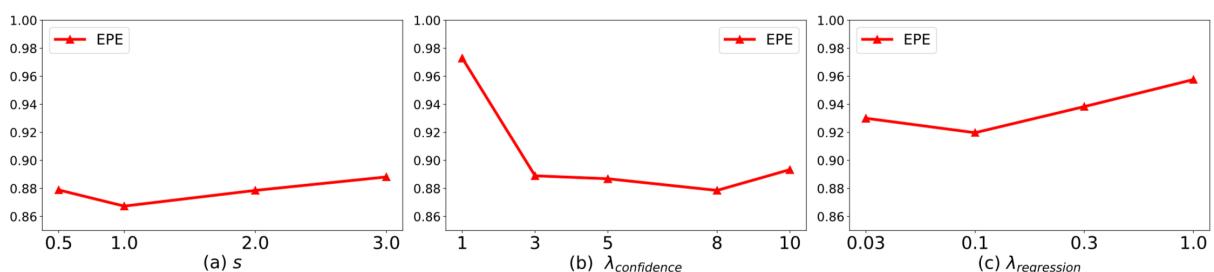


图 10 AcfNet 不同超参的消融实验结果

单峰分布方差 σ 分析

方差 σ 的大小反映了单峰分布的锋锐程度，对我们提出的方法AcfNet起着至关重要的作用。在我们的方法中，方差主要由 ϵ 和 s 限制，即 $\sigma \in [\epsilon, s + \epsilon]$ 。

首先，我们研究方差 σ 被固定的情况，也就是所有像素点的方差均为一个相同的值 $(s = 0, \sigma = \epsilon)$ 。通过网格搜索，我们发现 $\sigma = 1.2$ 时网络预测结果最好。这也暗示了对

于大多数的像素点，它们更偏向于用 $\sigma = 1.2$ 建立单峰分布。因此，我们设定 σ 的下限 ϵ 为1.0来探究自适应的方差学习。

接着，我们研究方差敏感度调节参数 s ，它控制着方差 σ 的上限。图10 (a) 展示了调节参数 s 得到的结果，其中 $s = 1$ 效果最好，而且当 s 从0.5变化到3.0的过程中，性能表现相当稳定。并且，当网络收敛后，我们在图11中展示了当 $s = 1$ ，也就是 $\sigma \in [1.0, 2.0]$ 时Scene Flow测试集中所有像素点的方差分布直方图。可以看出，大多数的像素点偏向于小的方差，而还有些像素点需求更大的方差来平缓单峰分布。

损失均衡权重

$\lambda_{confidence}$ 调节的是置信度网络的损失与其他损失的均衡，也隐式控制着方差的学习。

从图10 (b) 可以看出，当变化 $\lambda_{confidence}$ 时，让每个像素点的匹配置信度过大或过小都会导致更差的结果，而当 $\lambda_{confidence} = 8.0$ 时取得了最好的性能。

$\lambda_{regression}$ 均衡的是现有网络中广泛使用的视差回归损失，而过大的 $\lambda_{regression}$ 会消去文中提出的其他两项损失。图10 (c) 展示了性能变化曲线，可以看到，适当的均衡该回归损失与其他两项损失的影响能够大幅提升网络匹配性能。

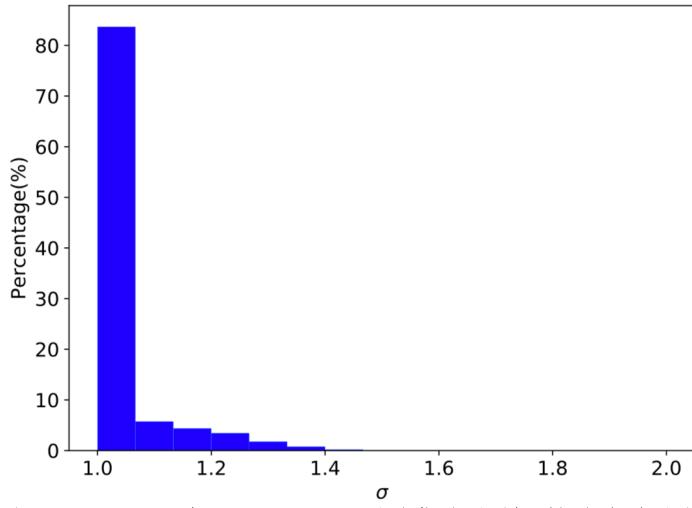


图 11 AcfNet 在 Scene Flow 测试集上方差 σ 的分布直方图

(2) 方差分析

方差估计是我们实现自适应匹配代价滤波的重要设计，它能够根据网络学习的难度自适应调整单峰分布的平缓度。为了定量评估它的性能，我们采用Ilg等人^[66]在论文中采用的sparsification plots技术。它能够揭示我们预测的置信度评估结果和真实的误差大小的吻合性。如图12所示，我们绘出AcfNet在Scene Flow测试集上的sparsification plots。该

图展示的是置信度相对较小的那部分像素点不断被移除后其余像素点的EPE误差；而Oracal曲线对应的是误差相对较大的那部分像素点不断被移除后其余像素点的EPE误差；同时我们还给出了随机移除像素点后的EPE误差曲线，即Random曲线。结果显示，我们的置信度评估曲线和Oracal曲线十分接近，仅移除6.9%的像素点，误差就下降了一半，并且性能上远胜于随机排除像素点的情况。这充分证明了我们的置信度评估在检测和解释离群点方面有着十分优越的性能。而且，通过图13，我们给出了几个可视化例子。可以看出，难学的区域主要在遮挡区域（1a, 1c, 2a），无条纹区域（1b, 3a）和细小物体（3a）。在这些难学的区域，我们的网络都给了非常低的置信度，这也证明了网络能够将这些区域的分布变缓以减小他们的影响，从而有效防止网络在这些区域过拟合。

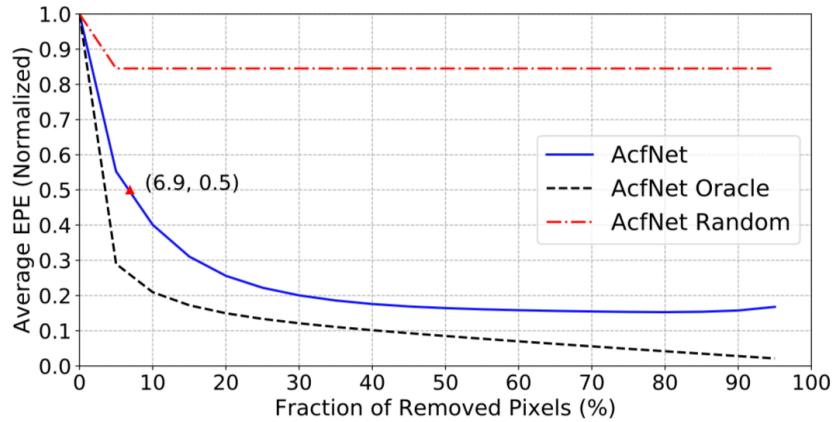


图 12 AcfNet 方差调节有效性示意图

表 1 AcfNet 网络模块有效性分析

Method	Scene Flow					
	EPE[px]			3PE[%]		
	ALL	OCC	NOC	ALL	OCC	NOC
PSMNet	1.101	3.507	0.637	4.56	17.64	2.12
PSMNet + CE	0.965	2.962	0.556	4.60	16.63	2.30
PSMNet + SF	0.917	2.874	0.524	4.47	16.19	2.26
PSMNet + CENet + SF	0.867	2.736	0.495	4.31	15.77	2.13
AcfNet	0.867	2.736	0.495	4.31	15.77	2.13

(3) 网络模块有效性分析

以PSMNet^[34]为基础网络，通过不断加入我们设计的模块以验证他们的有效性。实验结果如表1所示。相对基础网络PSMNet，我们首先验证对匹配代价体施加单峰分布约束的有效性，并且通过交叉熵损失（CE）约束分布学习。可以看出，单峰约束在各个指标上均有显著的性能提升，这也证明了单峰匹配代价滤波的优越性；接着我们采用Stereo

Focal Loss (SF) 解决交叉熵损失 (CE) 中面临的正负视差样本问题，这带来了各个指标上的进一步提升；最后我们加入置信度评估网络 (CENet)，同样在各个指标上都大幅提升准确率。值得一提的是，在评价指标 ALL EPE 上，我们的方法相对 PSMNet 从 1.101 降到 0.867，也就是将近 20% 的性能提升，这充分体现了自适应单峰匹配代价滤波的优越性和高性能。同时，我们还在图 13 中给出了几个可视化结果，从左到右依次为：左图，右图，真实的视差图，预测的视差图，误差图，置信度图；在误差图中，暖色调意味着误差越大；在置信度图中，颜色越暗表示越不确信。我们的方法预测的结果和真实的视差图基本一致，即使在结构复杂的区域预测的结果也依然很好。

(4) 自适应单峰匹配代价滤波有效性分析

AcfNet 直接在 PSMNet 基础上添加单峰匹配代价滤波约束。表 2 展示了两个版本的 AcfNet 与 PSMNet 的性能比较，其中每个像素点统一方差的 AcfNet(uniform) 相对 PSMNet 有大幅性能提升，并且自适应版本的 AcfNet(adaptive) 进一步提升了匹配准确率。这充分

表 2 自适应单峰匹配代价滤波有效性分析

Method	Scene Flow					
	EPE[px]			3PE[%]		
	ALL	OCC	NOC	ALL	OCC	NOC
PSMNet	1.101	3.507	0.637	4.56	17.64	2.12
AcfNet(uniform)	0.917	2.874	0.524	4.47	16.19	2.26
AcfNet(adaptive)	0.867	2.736	0.495	4.31	15.77	2.13

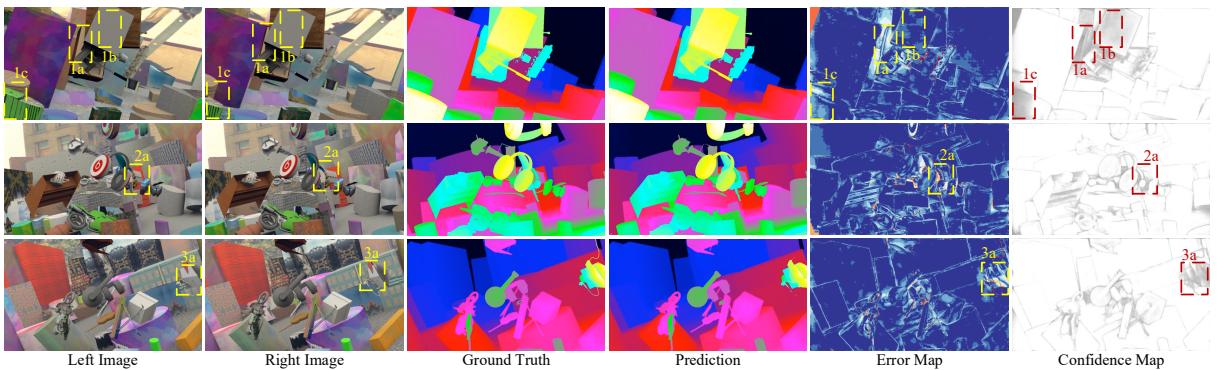


图 13 AcfNet 在 Scene Flow 测试集上定性评估结果

展现了单峰监督的有效性和自适应方差调整的优越性。相比 AcfNet(uniform)，AcfNet(adaptive) 在 OCC (也就是遮挡) 区域的提升更加明显，这与方差分析中得出我们的置信度评估网络能够有效检测并且防止网络在这些区域过拟合的结论一致。

(5) 匹配代价滤波对比分析

为了进一步证明我们滤波方法的优越性，我们设计实验与我们工作最相关的Poggi等人^[65]的工作进行对比。尽管有许多经典的基于匹配代价的滤波方法^{[70][72]}，但是这些方法已经无法和现有基于深度学习的方法对比。Poggi等人^[65]的匹配代价增强策略是先生成一个以真实视差为中心的高斯分布，然后作为权重，加权到未经过聚合的匹配代价上，从而加强以真实视差为中心的单峰匹配代价分布。他们的方法和我们的方法主要有两个不同点：（1）他们的单峰分布是作为权重影响匹配代价分布，但是我们直接用做网络监督项，可以直接引导网络将匹配代价滤波成一个单峰。（2）他们的真实视差信息在训练和测试阶段都需要，而我们仅需在训练过程使用。如表3所示，表中所有方法都是先在 Scene Flow 数据集上随机初始化训练，然后直接测试他们在 KITTI 2012 和 KITTI 2015 的泛化性能。所有的方法都是以 PSMNet 为基准网络，并且所有可用的视差信息都被使用。Poggi 等人^[65]在 KITTI 2015 上的泛化性能取自他们的论文实验汇报结果。作为滤波性能比较，我们的方法远胜 Poggi 等人^[65]的结果。而且，从泛化性能上看，我们相对 PSMNet 在 KITTI 2012 上有 11.64% 的提升，在 KITTI 2015 上有 10.74% 的提升。这也就是说明明确的单峰约束能够让网络学习到更好的相似度度量和特征提取方式，从而在不同数据集上表现出优越的泛化性能。

表 3 匹配代价滤波对比分析

Method	EPE[px]		3PE[%]	
	Scene Flow	KITTI 2012	KITTI 2012	KITTI 2015
PSMNet	1.101	29.18	-	30.19
Poggi et.al. ^[65]	0.991	-	-	23.13
AcfNet	0.867	17.54	17.54	19.45

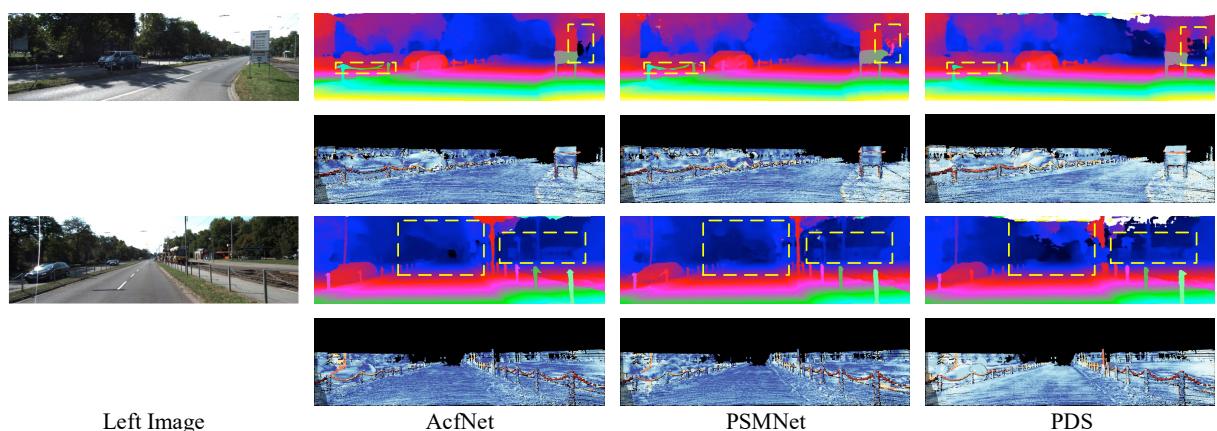


图 14 AcfNet 在 KITTI 2012 上可视化结果

(6) 与双目立体匹配的state-of-the-art方法对比分析

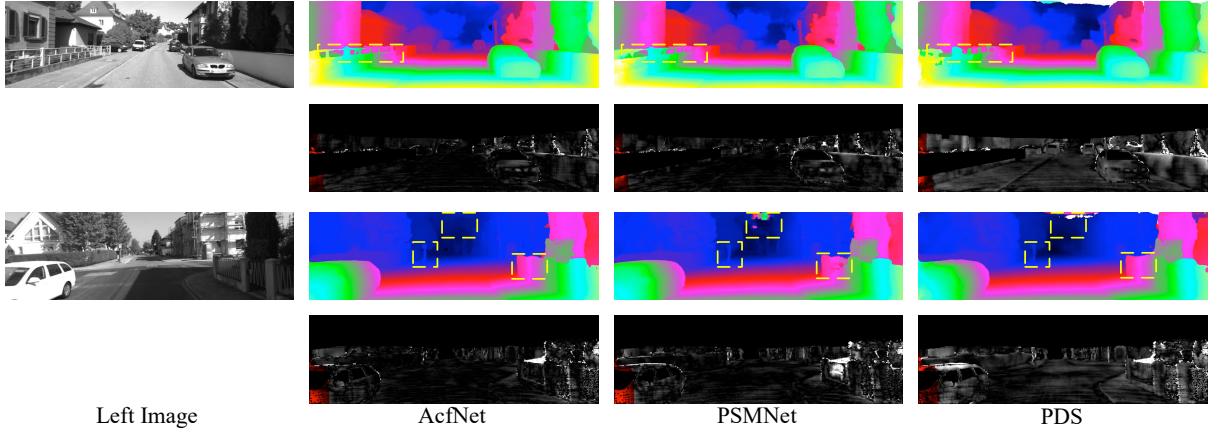


图 15 AcfNet 在 KITTI 2015 上可视化结果

表 4 AcfNet 在各公开数据集上的结果

Method	EPE	Scene		KITTI 2012				KITTI 2015			
		Flow		2px		3px		4px		5px	
		Noc	All	Noc	All	Noc	All	Noc	All	D1-all	D1-all
MC-CNN ^[29]	3.79	3.90	5.45	2.43	3.63	1.90	2.85	1.64	2.39	3.88	3.33
GC-Net ^[33]	2.51	2.71	3.46	1.77	2.30	1.36	1.77	1.12	1.46	2.67	2.45
iResNet ^[60]	1.40	2.69	3.34	1.71	2.16	1.30	1.63	1.06	1.32	2.44	2.19
PSMNet ^[34]	1.09	2.44	3.01	1.49	1.89	1.12	1.42	0.90	1.15	2.32	2.14
EdgeStereo ^[63]	1.12	2.79	3.43	1.73	2.18	1.30	1.64	1.04	1.32	2.16	2.00
SegStereo ^[64]	1.45	2.66	3.19	1.68	2.03	1.25	1.52	1.00	1.21	2.25	2.08
PDS ^[31]	1.12	3.82	4.65	1.92	2.53	1.38	1.85	1.12	1.51	2.58	2.36
Gwc-Net ^[61]	0.77	2.16	2.71	1.32	1.70	0.99	1.27	0.80	1.03	2.21	1.92
HD3-Stereo ^[73]	1.08	2.00	2.56	1.40	1.80	1.12	1.43	0.94	1.19	2.02	1.87
GA-Net ^[35]	0.84	2.18	2.79	1.36	1.80	1.03	1.37	0.83	1.10	1.93	1.73
AcfNet	0.87	1.83	2.35	1.17	1.54	0.92	1.21	0.77	1.01	1.89	1.72

为了进一步评估我们的模型 AcfNet 的性能，我们在表 4 中提供了我们在 Scene Flow^[32], KITTI 2012^[15], KITTI 2015^[16]三个数据集上与目前 state-of-the-art 方法的对比，其中包含：基于分类的方法（MC-CNN, PDS, HD3-Stereo），增强匹配成本计算的方法（Gwc-Net），堆积优化子网络的方法（iResNet-i2），有着十分强大的成本聚合网络的方法（PSMNet, GA-Net）和增加额外信息的方法（EdgeStereo, SegStereo）。尽管他们都试图改进网络去得到更加鲁棒的立体匹配结果，但是我们的方法仍然在性能上领先他们一大截。其中图 14 和图 15 可视化了我们在 KITTI 2012 和 KITTI 2015 的几个例

子，并且标出了和 PDS, PSMNet 对比明显效果更好的地方。对于每个数据集都提供了 2 个可视化样例，每个样例中，第一行为视差图预测结果，第二行为误差图可视化结果，其中 KITTI 2012 可视化图中，白色表示预测不准确，KITTI 2015 中，暖色调表示预测不准确。可以看出，我们的方法在细小物体，图片和天空边缘表现的更好。

3.5 本章小结

本章主要介绍了自适应的单峰匹配代价滤波方法用于端到端的双目深度估计网络。该方法对目前基于深度学习的双目立体匹配方法在匹配代价学习方面的欠缺进行求解。提出的AcfNet网络对网络预测的匹配代价体施加以真实视差为中心的单峰匹配代价分布，并且每个像素点的单峰分布方差能够根据该像素点的上下文信息自适应调整。最后，实验表明，该方法在KITTI等公开数据集上的深度估计精度达到了领域先进水平，并且在交叉数据验证中表现出了优越的泛化性能。

第四章 基于视差推荐网络的实时双目立体匹配

4.1 引言

自Kendall^[33]等人提出基于3D卷积端到端学习视差的方法GCNet以来，大量工作不断发展和改善其网络架构，比如PSMNet^[34]、GANet^[35]，并且成为现有方法中精度最高的网络框架。然而，这套框架特别消耗计算资源，推断一张 540×960 的图像一般需要500毫秒左右，这对实时应用十分不友好。而核心原因是其庞大的视差搜索空间，即 $\{0, 1, \dots, D - 1\}$ ， D 为最大视差搜索值。对此，大量的加速方法被提出以实现实时性能，比如由粗到精的网络架构^{[32][36][37][38]}和Duggal等人^[39]提出的视差空间剪枝方法DeepPruner。前者采样多尺度特征空间和扭曲（warp）策略，而问题在于低分辨率特征易丢失细节特征信息，且由于高分辨的匹配高度依赖于低分辨率的匹配结果，在低分辨率便丢失的细小且与背景视差相差较大的物体将一直无法被网络检测并完成深度估计。对于后者，视差空间剪枝和视差预测、优化等过程均在 $\frac{1}{4}$ 尺度进行，有效降低了细节信息丢失的概率，不过缺陷是仅依靠上下文信息有限的两个并行分支网络预测剪枝后空间的上下界是病态的，比如一些遮挡或者无条纹区域的上下界预测是不准确的，那么后续的优化过程也将是在错误的视差搜索空间进行。

我们的目标是大幅提升目前基于深度学习的双目立体匹配方法的深度估计帧率，同时具备与大网络（如PSMNet^[34]、GANet^[35]）一致的高匹配精度，使实时双目深度估计成为可能。我们的模型构建主要基于两个关键的观察发现：1. 双目匹配的视差搜索空间很大，但是许多的视差参选值可以非常确信地丢弃而没必要对整个搜索空间的视差都进行评估；2. 由于自然场景的连续性，相邻的像素一般拥有相似的视差值。这表明一旦我们知道了一个像素的视差值，我们可以有效地传播该信息到它的相邻像素点。基于此，我们提出DPN-Stereo网络，一个实时双目匹配模型，其网络框架图如图16所示。特别的，对于输入的左右图像对，一个U型特征提取网络会输出 $\frac{1}{4}$ 分辨率的特征图。在种子视差推荐模块（Disparity Proposal Seed Module）中通过计算左右特征图的互相关性得到匹配代价并用于估计粗糙的视差图。基于该粗糙的视差图，2个并行的2D沙漏型网络将学习可变形卷积的偏移参数，并利用可变形卷积完成对该粗糙视差图的视差采样，得到可靠且压缩的搜索空间上下界。和以往采用全视差范围建立3D匹配代价体的思路不同，我们构

建一个压缩的3D匹配代价体，其深度维度（Depth Dimension）为8。紧接着，3D代价聚合网络实现对匹配代价体的代价聚合，根据聚合后得到的代价体，Soft Argmin函数将用于估计亚像素的视差图。为了保留边缘等细节信息，边缘已知的优化模块（Edgeaware Refinement）对预测的视差图进行优化，并且计算 L_1 损失用于监督估计的和真实的视差图。其中视差推荐网络伪代码如下：

算法2：视差推荐网络

输入数据：经特征提取网络得到的左右图特征 F_l, F_r ，种子视差推荐模块dseed_model，

视差采样模块dsample_model

输出数据：视差搜索样本集 D_{set}

```

1: FUNCTION disparity_proposal_network( $F_l, F_r, dseed\_model, dsample\_model$ ):
2:    $d_{seed} \leftarrow dseed\_model(F_l, F_r)$  // 估计粗糙的视差图作为种子视差
3:    $d_{min}, d_{max} \leftarrow dsample\_model(d_{seed})$  // 采样得到视差搜索空间上下界
4:    $D_{set} \leftarrow \text{uniform sample 8 disparities in interval } [d_{min}, d_{max}]$  // 均匀采样
5:    $D_{set} \leftarrow \{D_{set}, d_{seed}\}$  // 将种子视差也放入视差搜索空间
6:   RETURN  $D_{set}$ 
7: END FUNCTION

```

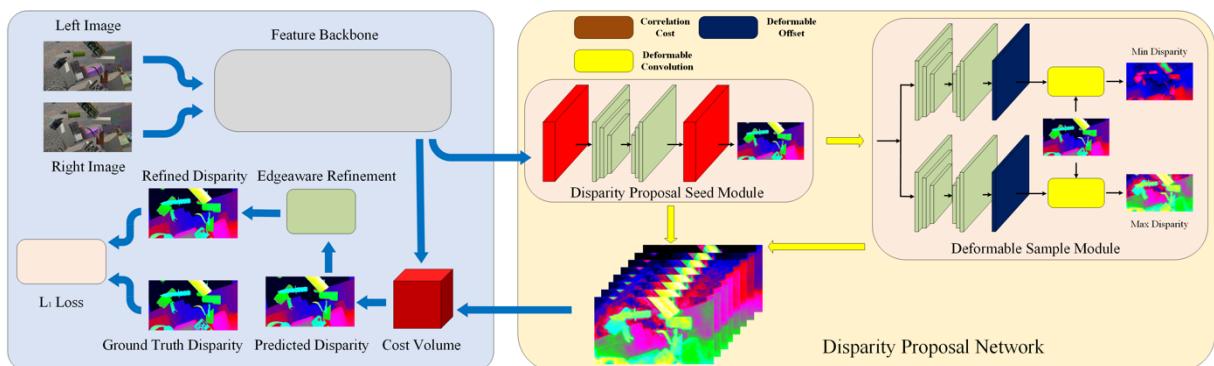


图 16 DPN-Stereo 实时网络结构框架

4.2 实时网络算法

4.2.1 特征提取网络

网络架构的第一步是从图像中提取有意义的特征表示，为后续实现精准的匹配提供帮助。在双目匹配过程中，条纹稀少的区域一般很难处理，而传统匹配方法为了求解该问题一般使用非常大的窗口进行匹配代价聚合。为了确保网络能够有足够大的上下文信

息，我们通过设计大的感受野进行图像特征提取。特别的，我们用一个参数共享的特征提取网络（也就是孪生网络）对两张输入图像进行特征提取。首先，我们用3个通道数为8的卷积层实现对原图像的2倍下采样；紧接着，我们使用3个残差块^[74]，即卷积核为 3×3 ，步幅为2，且紧跟归一化层和Relu激活层，完成对图像特征的大幅下采样，得到图像分辨率分别为4, 8, 16的图像特征。并且，对于下采样尺度为16的图像特征，我们使用DenseAspp^[75]模块实现更大的感受野提取。紧接着，我们利用U型网络结构对多尺度特征进行融合，对得到的下采样尺度为4的图像特征，一个附加的DenseAspp^[75]进一步完成上下文信息融合和处理。最后，一个 3×3 的卷积层输出通道数为16的图像特征，不再附加归一化层和Relu激活层。这种低分辨率的特征表示是十分重要的，主要原因有两个：1) 对于无条纹等具有挑战性的区域，所提取的特征具有足够大的感受野。2) 特征高度压缩，能够有效降低网络计算消耗。

4.2.2 视差推荐网络

现有的双目匹配方法一般在全视差搜索范围内生成匹配代价体（cost volume）^{[31][32][33][34][35][36][37][38]}。如此庞大的视差空间不仅增加了内存计算消耗，而且增加了计算负担。以PSMNet^[34]为例，3D匹配代价体构建和匹配代价聚合过程会费超过250ms的时间。这两个操作导致实时应用变成不可能。最近的工作DeepPruner^[39]尝试对视差空间进行大尺度间隔采样得到10个左右的视差参选值，接着采用PatchMatch^[76]和两个并行3D卷积网络完成对这些视差参选值的评估和可能的视差搜索空间上下界的预测。而矛盾就在于既希望快速锁定搜索空间又希望上下文信息有限的两个网络准确预测搜索空间上下界。一些比较难以处理的区域（例如遮挡和无条纹区域）即使是AcfNet, GANet^[35]这种参数量庞大的网络都是很难准确掌握其可能的视差范围，更何况一个未加入任何约束和几何先验的轻量级网络。所以，为解决上述问题并快速、准确锁定视差搜索空间，本文提出视差推荐网络（Disparity Proposal Network）。我们将搜索空间上下界的确定过程分解为两个步骤：第一步负责估计场景结构和深度信息，第二步负责根据得到的场景结构信息推测得到相对深度估计并利用采样策略分别从相对更近或更远的区域得到匹配空间上下界，该上下界确保了真实视差值在搜索范围内。接着，在上下界内均匀采用，得到少量推荐视差参选值用于构建稀疏的3D匹配代价体，让网络集中于高度可能的区域进行细匹配过程，从而大幅减少计算量。具体的，我们先用一个种子视差推荐模块（Disparity Proposal Seed Module）估计粗略的场景结构和深度信息，即种子视差图；根据该种子视

差图和场景中物体中具有相对远景的层次结构先验，设计可变形视差采样模块(Disparity Sample Module) 实现自适应变化采样距离的视差采样方案并得到视差搜索空间的上下界。最终，通过在预测的上下界范围内均匀采样，即得到最终的视差值参选集合并用于后续的筛选和优化过程。

种子视差推荐模块 (Disparity Proposal Seed Module)：人观察自然场景的过程一般为，先快速扫描全局，得到一个大概的轮廓和层次远近信息。启发于这种现象，我们将利用2D卷积网络实现预匹配过程，实现快速掌握场景上下文和深度信息。具体的，我们将利用左右特征图之间的一维相关性 (1-D correlation)^[32] 沿着视差维度 (即 $d = \{0, 1, \dots, \frac{D-1}{4}\}$) 计算匹配代价，并利用沙漏型结构的2D卷积神经网络结构完成匹配代价的聚合。对于聚合得到的匹配代价体，我们用一个置信度评估网络 (Confidence Estimation Network) 估计置信度图并用于调节真实的匹配代价体 (Ground Truth Cost Volume)，以生成像素级别的单峰分布 (Unimodal Distribution) 作为网络训练标记；并且我们将利用立体聚焦损失 (Stereo Focal Loss) (即公式(3.7)) 约束估计的和真实的匹配代价体之间的一致性，对应的损失项记为 $\mathcal{L}_{SF-CORR}$ 。同样的，我们将采用soft argmin函数 (公式 (3.1)) 从聚合的匹配代价体中回归视差图，该视差图即为我们设计视差推荐网络的基石，因此我们设定其为种子视差图，可能的视差参选值也将在这个种子视差图采样得到。虽然相似度函数固定为一维相关性 (1-D correlation)，在一定程度上牺牲了精度，却为网络节省了大量的计算资源消耗和时间花费，自适应单峰匹配代价滤波的监督作用强化了图像特征的学习，为种子视差图学习到大概的场景深度层次信息提供了保障。对于所采用的沙漏型结构的2D卷积神经网络，其具体网络参数如下：首先两个卷积层完成大幅度的匹配代价下采样，将单个点的匹配代价扩散到相邻像素点，其卷积核为 3×3 ，步幅为 2，且紧跟归一化层和Relu激活层。接着两个卷积核为 3×3 ，步幅为 1 的卷积层继续对下采样后的匹配代价进行聚合，最后两层反卷积层对匹配代价进行聚合并恢复到原有尺度。该网络结构的直观表示可参考图16。

可变形视差采样模块 (Deformable Sample Module)：给定种子视差图，接下来就是如何生成高度可靠且有效的视差参选值了。与Duggal等人^[39]提出的视差空间剪枝方法 DeepPruner采用相同的轻量级并行网络直接从构建的粗糙匹配代价体 (cost volume) 上预测搜索空间上下界方案不同，我们提出可变形采样模块 (Deformable Sample Module) 根据种子视差图采样得到上下界。核心思想在于：具有挑战性的区域一般分布在图像中

的某些局部区域，即匹配模糊仅限定在一定范围。若从整个图像的感受野来看，它的相对更近或更远的区域是一定存在且相对更好预测的，虽然这个更近和更远区域的深度相差很大，但至少确保了后续的3D匹配代价聚合和优化过程在存在真值的区间内筛选和优化。因此，为了实现自适应的采样方案，我们引入可变形卷积^{[78][79]}。和普通卷积操作固定采样点不同，可变形卷积可对采样位置进行学习，实现自适应的特征采样和聚合。因此，对可变形卷积的偏移参数（Offset）进行学习是视差采样的关键，为其学习过程提供左右特征图和种子视差图的上下文信息即可实现自适应且合理的可变形视差采样。具体的，我们将一维相关性（1-D correlation）匹配代价，左图特征和种子视差图在通道维度进行衔接并作为可变形采样模块的输入。为了实现信息的广泛传播与聚合，我们采用沙漏型结构的2D卷积神经网络，其网络结构的直观表示可参考图16的可变形视差采样模块（Deformable Sample Module）。每个沙漏型结构的2D卷积神经网络都将输出一个可变形卷积的偏移参数特征，结合可变形卷积，即可实现对种子视差图的自适应视差样本采样。对于得到的上下界视差图，为保证所采样视差值为相对更近或更远的深度值，我们利用排序损失^[77]约束：

$$\mathcal{L}_{Rank-k} = \frac{1}{|\mathcal{P}|} \sum_{p \in \mathcal{P}} \ln(1 + e^{k \times (d_p - \hat{d}_p)}) \quad (4.1)$$

其中 $k \in \{-1, 1\}$ ，当所学视差图为下界时 $k = -1$ ；上界时 $k = 1$ 。同时，为驱使学习的上下界视差图趋近于真实视差图，我们采用smooth L_1 损失（公式(3.3)）监督其学习，该损失项记为 $\mathcal{L}_{1-Prop(k)}$ ，其中 $k = \{-1, 1\}$ 。至此，对于得到的视差空间上下界，我们采用均匀采样的方法等间隔采样得到8个视差参选值样本。相比于原视差样本空间 $\{0, 1, \dots, \frac{D-1}{4}\}$ ，我们实现了6倍的空间压缩，为后续3D匹配代价和视差估计过程节省了大量的计算成本和资源消耗，是实时双目立体匹配得以实现的核心因素。

4.2.3 匹配代价聚合

基于视差推荐网络所预测的视差推荐样本空间，我们建立3D匹配代价体聚合网络进行空间聚合。结合先前双目匹配方法^{[33][34][35]}的实践经验，我们采用GCNet^[33]的衔接方式，对输入的左右图特征和4个视差参选值，输出对应的3D匹配代价体。相比于先前的工作^{[33][34][35]}，我们的匹配代价体压缩了6倍，使得聚合过程十分高效。并且，根据公式(3.1)，soft argmin函数将用于估计亚像素的视差图。同样，我们采用smooth L_1 损失（公式(3.3)）监督其学习，该损失项记为 \mathcal{L}_{1-Pred} 。

4.2.4 视差图优化

仅依赖粗糙的匹配，所得到的视差图一般缺乏细节信息。为了保持压缩紧凑的模型设计，我们采用边缘保留的优化网络（即图16中的Edgeaware Refinement网络）解决这个问题。值得一提的是，该优化网络的职责在于对粗糙的视差图进行膨胀或者消融操作，通过原始图像信息作为引导渲染高频率细节信息，所以用一个压缩紧凑的网络学习像素到像素之间的映射关系是十分合理的。特别的，我们的优化网络只需要学习一个残差视差图添加到粗糙的视差图上完成优化过程。

首先，我们将soft argmin函数预测的视差图通过双线性上采样方法得到左右图像分辨率的视差图。根据该视差图，我们采用扭曲（warp）方法将右图插值到左图，并与原始左图做差得到误差图。我们的优化网络即可将左右图，上采样的视差图和生成的左图及误差图作为网络输入。先通过一个 3×3 的卷积（Convolution）层得到32维度的表示，接着通过5个 3×3 的卷积（Convolution）层，采样膨胀率（dilation）分别为1、2、4、8、1；并且每个卷积层都紧跟归一化层和Relu激活层。最后，网络的输出是通过一个 3×3 的卷积（Convolution）层得到一个1维度的视差残差图并加到网络输入的上采样视差图，并且利用Relu激活层确保视差图中的所有值都为正。同样，我们采用smooth L_1 损失（公式(3.3)）监督其学习，该损失项记为 $\mathcal{L}_{1-Refine}$ 。我们的优化网络原理和双边滤波相似，本质上是通过原始图像的引导实现边缘保留的优化过程。

4.2.5 全部损失函数

总的来说，我们最终的损失函数一共包含六个部分：

$$\mathcal{L} = \mathcal{L}_{SF-CORR} + \lambda_{1-Pred} \mathcal{L}_{1-Pred} + \lambda_{Rank-k} \mathcal{L}_{Rank-k} + \lambda_{1-Prop(k)} \mathcal{L}_{1-Prop(k)} + \lambda_{1-Refine} \mathcal{L}_{1-Refine} + \lambda_{confidence} \mathcal{L}_{Confidence} \quad (4.2)$$

其中 λ_{1-Pred} 、 $\lambda_{1-Refine}$ 、 λ_{Rank-k} 、 $\lambda_{1-Prop(k)}$ 和 $\lambda_{confidence}$ 为损失权衡参数。 \mathcal{L}_{SF} 监督匹配代价体的学习； λ_{1-Prop} 监督视差推荐网络推选的各个视差图，使可变形卷积学习到最优的采样位置偏移； λ_{Rank} 确保学习的视差搜索空间上下界包含真实的视差值； λ_{1-Pred} 监督匹配代价聚合过程，只有学习到鲁棒的图像特征与相似度估计函数，才能从视差搜索空间中挑选中最匹配的视差图； $\lambda_{1-Refine}$ 监督优化网络实现边缘保留的视差图优化，在原始图像的引导下学习渲染高频率细节信息；根据公式(3.9)， $\mathcal{L}_{Confidence}$ 作为一个正则化器鼓励更多的像素点拥有大的置信度值。

4.3 实验验证

4.3.1 数据库及实验协议和实现细节

(1) 数据库

为了定性和定量评估我们提出的方法 DPN-Stereo，我们将在三个具有挑战性的公开数据集（Scene Flow^[32], KITTI 2012^[15], KITTI 2015^[16]）上进行评测。Scene Flow^[32]是一个合成数据集，包含 35454 张训练图片对和 4370 张测试图片对，而且提供密集的真实视差标注信息，非常适合训练和测试网络模型。KITTI 2012^[15], KITTI 2015^[16]是两个真实街景数据集，所提供的视差标注信息都是通过雷达扫描得到的，因此都是稀疏的。前者包含 194 张训练图片对和 195 张测试图片对；而后者包含 200 张训练图片对和 200 张测试图片对。两个 KITTI 数据集对于训练神经网络来说数据量规模太小，非常具有挑战性。因此，参考 GC-Net^[33]我们只在 Scene Flow^[32]上设计消融实验并且对网络设计分析。

(2) 评价指标

实验中，我们采用两个标准的评估指标：(1) 3-pixel-error (3PE)，指预测的视差和真实的视差大于 3 像素的像素点占总数的百分比；(2) end-point-error (EPE)，指预测的视差和真实的视差的平均差异。EPE 更加注重亚像素的误差，而 3PE 则重点刻画离群点所占的百分比。

(3) 实现细节

我们的DPN-Stereo采用PyTorch实现，所有的模型都是采用RMSProp的标准设置端到端训练。对于所有数据集中的图像，都会采用颜色归一化进行数据处理。训练的时候，我们随机截取 $H = 256, W = 512$ 图像块，并且最大的视差值 D 设置为 192。对于网络训练，我们随机初始化网络参数并且在 Scene Flow 上以一初始的学习率 0.001 训练 64 个周期 (Epoch)，每 20 个周期学习率除以 2，并且直接用训练好的模型进行测试。对于 KITTI 数据集，我们用 Scene Flow 上预训练的模型进行微调 600 个周期 (Epoch)。初始的微调学习率设置为 0.001，并且在 100 和 300 个周期的时候衰减 $\frac{1}{3}$ 。训练的批数据大小为 16，总共 4 块 NVIDIA GTX 1080Ti GPUs，所以每张显卡上放 4 个批数据。对于各损失权衡参数，我们设置 $\lambda_{1-Pred} = 0.5$ 、 $\lambda_{1-Refine} = 1.0$ 、 $\lambda_{Rank-k} = 10.0$ 、 $\lambda_{1-Prop} = 0.001$ 和 $\lambda_{confidence} = 8.0$ 。

4.3.2 实验结果及讨论

由于 Scene Flow 有足够的数据量用于网络端到端训练且不用担心过拟合问题，我们的所有实验都是在 Scene Flow 数据集上进行。

(1) 消融实验结果分析

为了理解 DPN-Stereo 各个模块的有效性，我们首先研究各个模块在整个网络中所起的效果。实验结果如表 5 所示。我们的基础网络架构是：对输入的左右图像进行特征提取，利用一维相关性（1-D correlation）计算匹配代价并回归得到视差图（即种子视差推荐模块（Disparity Proposal Seed Module））。通过对表 5 中数据，可以看出即使网络规模较小，自适应的单峰匹配代价滤波策略也表现出了极佳的双目匹配性能，在极高的运行速度下（54.56 帧每秒）仍能得到 EPE=2.052 的性能优势。相比之下，GCNet^[33]的运行帧率仅 1.1 帧每秒，而 EPE 仅 2.51。值得注意的是，边缘保留(Edgeaware Refinement)的视差图优化结构对网络提升性能异常显著。这也表明了，当网络规模较小时，其细节保留及全局上下文信息提取能力较差，而通过边缘保留的优化结构，能够辅助网络实现细节补偿和物体级光滑性约束，从而得到更加优越的性能。视差推荐网络作为我们模型的核心部分，同样有着举足轻重的地位：其中种子视差推荐模块为后续视差搜索空间预测和 3D 匹配代价聚合过程提供了良好的场景结构和深度信息，可变形采样模块（Deformable Sample Module）实现了精确的搜索空间上下界定位。为了消除由于 3D 匹配代价聚合网络所引进的参数量对模型性能的影响，我们将种子视差推荐模块所得到的视差图重复 8 次输入 3D 匹配代价聚合网络，也就是表 5 中的（*）。可以看出，3D 代价聚合网络对有效性较差的视差推荐值空间优化效果微弱。相对之下，可变形采样模块则实现了 EPE 从 1.426 到 1.012 的优化。

表 5 DPN-Stereo 网络模块有效性分析

Network Component				Inference Speed	Scene Flow
Disparity Proposal Network	3D Cost	Edgeaware		FPS	EPE
Deformable Proposal Seed Module	Deformable Sample Module	Aggregation	Refinement	54.56	2.052
✓	*	✓	✓	31.13	1.426
✓	*	✓	✓	24.52	1.329
✓	✓	✓	✓	22.61	1.012

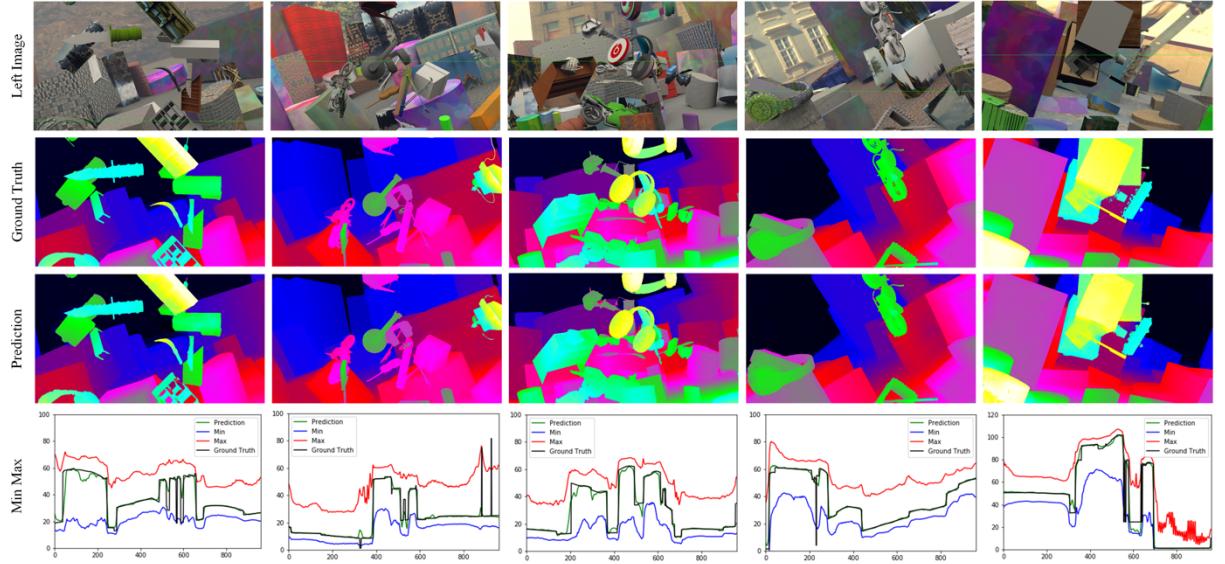


图 17 视差推荐网络所预测的视差搜索空间可视化

(2) 视差搜索空间上下界可视化

视差推荐网络的目标是为了对视差搜索空间进行剪枝，将不可能的匹配排除，从而确保计算资源都集中在一部分有意义的视差参选值上。为了理解视差推荐网络的有效性，我们给出了图 17 所示可视化结果。对于每一张原始左图，我们随机在图中描一条横线，并给出其对应的真实视差、估计视差以及视差推荐网络所估计的视差搜索空间。从图 17 中可以看出，无论是对于纹理丰富区域，亦或是遮挡、纹理稀疏等具有挑战性的区域，我们的视差推荐网络均能够给出覆盖真实视差的搜索空间。相比原先 $\{0, 1, \dots, D - 1\}$, $D = 192$ 的视差搜索空间，我们剪枝后的空间大幅缩减，有效的减少了匹配代价聚合过程所需承担的计算量和内存消耗。并且，经匹配代价聚合网络优化后得到的视差值，与真实视差十分接近，从图 18 中给出了几个可视化误差图也可看出，大部分区域的误差均在亚像素级别。

表 6 DPN-Stereo 网络各模块运行时间统计

Feature Backbone	Disparity Proposal Network		3D Cost	Edgeaware Refinement
	Disparity Proposal	Disparity		
	Seed Module	Sample Module		
	13.5ms	3.5ms	5.5ms	11.5ms
				13.5ms

(3) 运行时间和内存分析

表 7 定性评估双目立体匹配方法运行时间和精度

Method	GCNet ^[33]	PSMNet ^[34]	DeepPruner ^[39]	StereoNet ^[37]	DispNetC ^[32]	Ours
EPE	2.51	1.12	0.86	1.53	1.68	1.01
Runtime	900ms	500ms	292ms	52.2ms	60ms	47.5ms

在 KITTI 数据集分辨率下，也就是 384×1248 ，并且设置网络为测试模式。通过测试 50 次迭代，计算平均耗时，我们将网络各个模块的时间消耗给出在表 6 中，这对于对实时双目匹配感兴趣的读者来说能清晰的知道和理解网络时间消耗瓶颈在哪。我们的模型在 Nvidia 1080Ti 上能够以 22 帧每秒的速率运行，其中特征提取、边缘保留的优化和 3D 匹配代价聚合过程耗时均为网络整体耗时的四分之一左右。视差推荐网络耗时则相对更低，其中种子视差推荐模块和可变形视差采样模块耗时均在 5 毫秒左右，这和我们当初设计一个快速的视差空间锁定方案初衷一致。为了进一步描述网络的有效性，在同一测试环境下，我们将对比基于全视差空间匹配代价体的方法（如 PSMNet^[34]）和视差空间剪枝方案 DeepPruner^[39]。给定一对全分辨率的 KITTI 图像，PSMNet 在测试过程中需要消耗 4351MB 的内存，DeepPruner 消耗 1161MB 的内存，我们的方法与 DeepPruner 相似，消耗 1093MB 内存。而通过表 7 所示目前先进双目立体匹配方法在运行时间和精度与我们方法的对比，我们的方法在同一等级的时间损耗方法中，精度上取得了极大领先，而相比帧率在 1-2 帧每秒左右的方法，如 PSMNet，GCNet，我们的方法在性能和帧率上都有着绝对的领先优势。同时，我们还在图 18 中给出了几个可视化结果，从左到右依次为：左图，右图，真实的视差图，预测的视差图，误差图，置信度图；在误差图中，暖色调意味着误差越大；在置信度图中，颜色越暗表示越不确信。我们的方法预测的结果和真实的视差图基本一致，即使在结构复杂的区域预测的结果也依然很好。

表 8 DPN-Stereo 在 KITTI 2015 上的定性评估

Methods	Inference		Noc(%)			All(%)		
	Runtime	bg	fg	all	bg	fg	all	
GCNet ^[33]	900ms	2.02	3.12	2.45	2.21	6.16	2.87	
PSMNet ^[34]	500ms	1.71	4.31	2.14	1.86	4.62	2.32	
DeepPruner ^[39]	292ms	1.71	3.18	1.95	1.87	3.56	2.15	
StereoNet ^[37]	52.2ms				4.30	7.45	4.83	
DispNetC ^[32]	60ms	4.11	3.72	4.05	4.32	4.41	4.34	
AcfNet(Ours)	480ms	1.36	3.49	1.72	1.51	3.80	1.89	
DPN-Stereo(Ours)	47.5ms	1.46	3.35	1.77	1.63	3.73	1.98	

(4) 与双目立体匹配的state-of-the-art方法对比分析

为了进一步评估我们的模型 DPN-Stereo 的性能，我们在表 7 和表 8 中分别提供了我们在 Scene Flow^[32]和 KITTI 2015^[16]两个数据集上与目前 state-of-the-art 方法的对比。评估中的模型运行耗时均是我们在统一环境下测试得到，即在 KITTI 数据集分辨率下，也就是 384×1248 ，并且设置网络为测试模式，通过测试 50 次迭代，计算平均耗时。实验平台为 Nvidia 1080Ti。DPN-Stereo 取得了十分优越的性能，同时运行速度相当快。图 18 和图 19 分别给出了 DPN-Stereo 在 Scene Flow^[32]和 KITTI 2015^[16]两个数据集上的可视化结果。

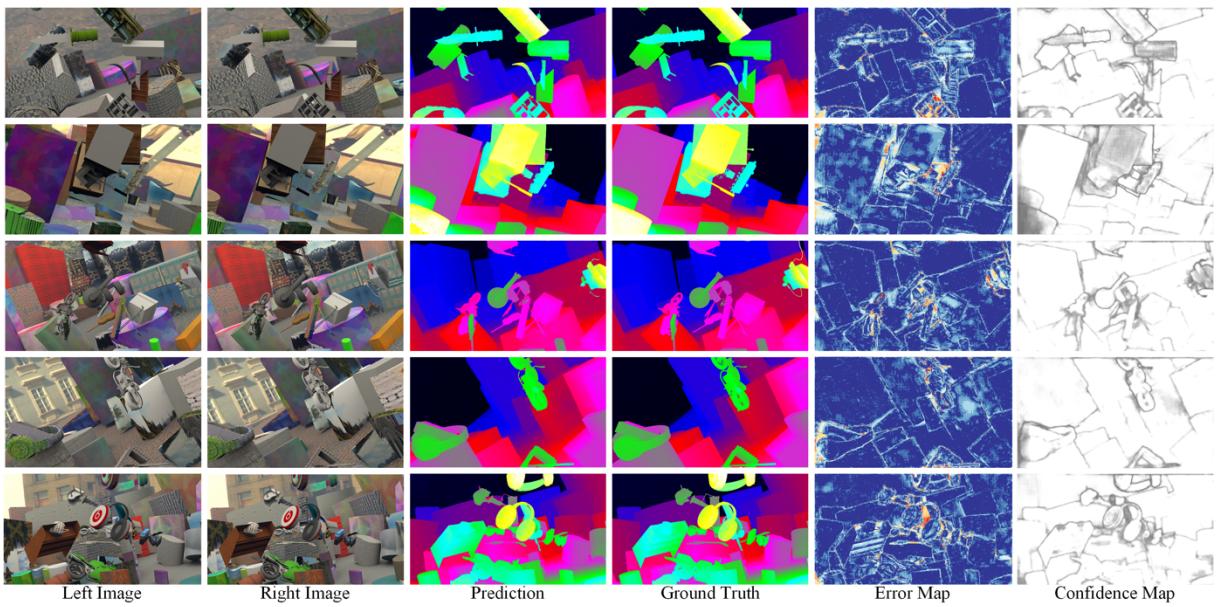


图 18 DPN-Stereo 在 Scene Flow 测试集上定性评估结果

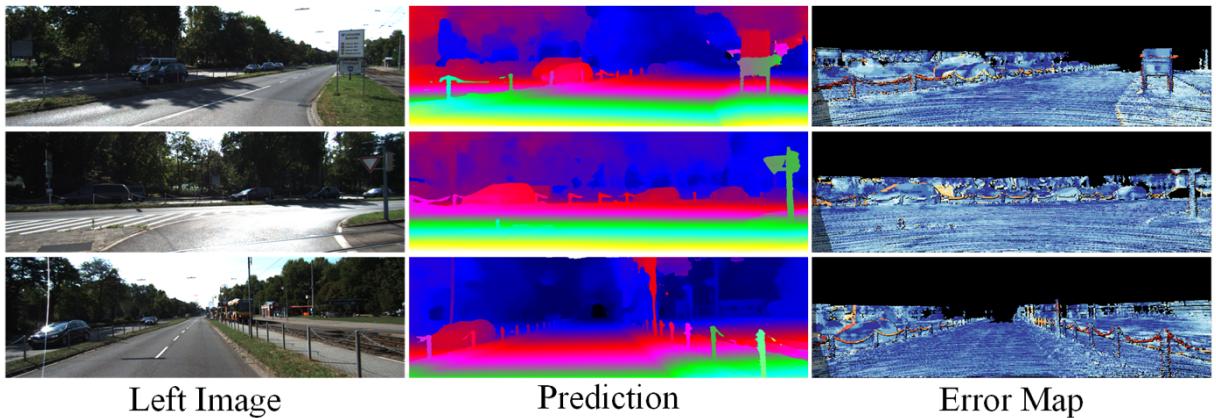


图 19 DPN-Stereo 在 KITTI 2015 数据集上的定性评估结果

4.4 本章小结

本章我们展示了对于全视差空间状态下的3D匹配代价体，如何在无需评估所有匹配分数的前提下实现快速的视差搜索空间剪枝。为了实现这个目标，我们设计了可端到端

训练的神经网络，而视差推荐网络是其核心结构。实验表明我们设计的实时立体匹配方法在速度上领先于现有处于先进水平的实时匹配方法，同时性能上也实现了超越。在未来，我们计划将该方法在光流、深度补全等任务上测试，为深度估计方法的实用性提供基础。

总结与展望

为了将双目立体匹配拓展到实际应用场景并保证实用性，本文主要讨论如何在保证高精度的同时，实现实时的立体匹配方案。自然场景中的物体结构和场景层次信息是有一定顺序的，正如我们可以用图像的二阶统计量对图像颜色信息去冗余^[80]，我们也可以结合几何先验亦或统计信息对基于深度学习的双目匹配过程去冗余。

本文首先介绍了研究背景及意义，引出了现有双目匹配方法在应用方面的技术瓶颈与研究高精度实时匹配方法的重要性。结合相关的研究现状分析，本文指出了现有基于视差回归的深度匹配方法在匹配代价计算过程所存在的缺陷，并提出自适应的单峰匹配代价滤波方案。针对不同的场景与语义信息，本文设计了一个置信度估计网络以自适应地调整单峰代价滤波学习。从在立体匹配的主流数据集 Sceneflow, KITTI 2015, KITTI 2012上的实验结果表明，我们的方法性能成为新的state-of-the-art。并且从给出的可视化图可以看到误差图和置信图很好的对应起来，说明AcfNet可以给予具有丰富信息的像素较高的置信度，同时防止信息较少的像素过拟。虽然我们的方法对匹配代价计算过程带来了新的建模视角，但也引入了大量的参数，增加了网络训练的复杂度。而终其原因是 我们还未揭示真实场景中的匹配代价单峰分布，毕竟采用某种分布形式进行模拟还是无法反应真正的匹配过程，这也是我们以后将继续研究并解决的重点难题。

接着，我们进一步分析了现有实时立体匹配的经典方法，并提出视差推荐网络。其核心在于，在极短的运行时间内推荐覆盖真实视差的搜索空间上下界的前提下，大幅缩减视差搜索范围，从而有效减少匹配代价聚合过程所需承担的计算量和内存消耗。实验表明我们设计的实时立体匹配方法在速度上领先于现有处于先进水平的实时匹配方法好几倍，同时性能上也实现了超越。不过，我们的视差推荐网络还是可以进一步优化的，比如我们的方法得到的视差搜索上下界不够紧致，在纹理丰富区域依然间隔较大，而这也是需要进一步探索的问题。当然，这种视差搜索空间剪枝方案是值得探索和研究的，特别是对于光流匹配任务，匹配空间所带来的计算量消耗是相当严重的，而将该方案成功推广到光流等类似的任务将是一个极大的贡献。

总体而言，我们的高精度且实时的匹配方案取得了十分优越的成果，也亟需在应用中检验。在双目匹配方法未广泛应用之前，这仍然是一个十分值得探索和研究的课题。

参考文献

- [1] Hadjitheophanous S, Ttofis C, Georghiades A S, et al. Towards hardware stereoscopic 3D reconstruction: a real-time FPGA computation of the disparity map[C]//Proceedings of the Conference on Design, Automation and Test in Europe. European Design and Automation Association, 2010: 1743-1748.
- [2] Geiger A, Ziegler J, Stiller C. Stereoscan: Dense 3d reconstruction in real-time[C]//2011 IEEE Intelligent Vehicles Symposium (IV). IEEE, 2011: 963-968.
- [3] Bruno F, Bianco G, Muzzupappa M, et al. Experimentation of structured light and stereo vision for underwater 3D reconstruction[J]. ISPRS Journal of Photogrammetry and Remote Sensing, 2011, 66(4): 508-518.
- [4] Murray D, Little J J. Using real-time stereo vision for mobile robot navigation[J]. autonomous robots, 2000, 8(2): 161-171.
- [5] Häne C, Zach C, Lim J, et al. Stereo depth map fusion for robot navigation[C]//2011 IEEE/RSJ International Conference on Intelligent Robots and Systems. IEEE, 2011: 1618-1625.
- [6] Stelzer A, Hirschmüller H, Görner M. Stereo-vision-based navigation of a six-legged walking robot in unknown rough terrain[J]. The International Journal of Robotics Research, 2012, 31(4): 381-402.
- [7] Geiger A, Lenz P, Urtasun R. Are we ready for autonomous driving? the kitti vision benchmark suite[C]//2012 IEEE Conference on Computer Vision and Pattern Recognition. IEEE, 2012: 3354-3361.
- [8] Howard A. Real-time stereo visual odometry for autonomous ground vehicles[C]//2008 IEEE/RSJ International Conference on Intelligent Robots and Systems. IEEE, 2008: 3946-3952.
- [9] Konolige K, Agrawal M, Bolles R C, et al. Outdoor mapping and navigation using stereo vision[C]//Experimental Robotics. Springer, Berlin, Heidelberg, 2008: 179-190.
- [10] Mattoccia S. Stereo vision: Algorithms and applications[J]. University of Bologna, 2011, 22.
- [11] Szeliski R. Computer vision: algorithms and applications[M]. Springer Science & Business

Media, 2010.

- [12]Marr D, Poggio T. A computational theory of human stereo vision[J]. Proceedings of the Royal Society of London. Series B. Biological Sciences, 1979, 204(1156): 301-328.
- [13]Krizhevsky A, Sutskever I, Hinton G E. Imagenet classification with deep convolutional neural networks[C]//Advances in Neural Information Processing Systems. 2012: 1097-1105.
- [14]Scharstein D, Szeliski R. A taxonomy and evaluation of dense two-frame stereo correspondence algorithms[J]. International Journal of Computer Vision, 2002, 47(1-3): 7-42.
- [15]Geiger A, Lenz P, Urtasun R. Are we ready for autonomous driving? the kitti vision benchmark suite[C]//2012 IEEE Conference on Computer Vision and Pattern Recognition. IEEE, 2012: 3354-3361.
- [16]Menze M, Geiger A. Object scene flow for autonomous vehicles[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2015: 3061-3070.
- [17]Zabih R, Woodfill J. Non-parametric local transforms for computing visual correspondence[C]//European Conference on Computer Vision. Springer, Berlin, Heidelberg, 1994: 151-158.
- [18]Calonder M, Lepetit V, Strecha C, et al. Brief: Binary robust independent elementary features[C]//European Conference on Computer Vision. Springer, Berlin, Heidelberg, 2010: 778-792.
- [19]Heise P, Jensen B, Klose S, et al. Fast dense stereo correspondences by binary locality sensitive hashing[C]//2015 IEEE International Conference on Robotics and Automation (ICRA). IEEE, 2015: 105-110.
- [20]Tombari F, Mattoccia S, Di Stefano L, et al. Classification and evaluation of cost aggregation methods for stereo correspondence[C]//2008 IEEE Conference on Computer Vision and Pattern Recognition. IEEE, 2008: 1-8.
- [21]Zabih Z. Computing visual correspondence with occlusions using graph cuts[C]//Eighth IEEE International Conference on Computer Vision. 2001, 2: 508-515.
- [22]Klaus A, Sormann M, Karner K. Segment-based stereo matching using belief propagation and a self-adapting dissimilarity measure[C]//18th International Conference on Pattern

- Recognition (ICPR'06). IEEE, 2006, 3: 15-18.
- [23]Bleyer M, Rhemann C, Rother C. PatchMatch Stereo-Stereo Matching with Slanted Support Windows[C]//BMVC. 2011, 11: 1-11.
- [24]Hirschmuller H. Stereo processing by semiglobal matching and mutual information[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2007, 30(2): 328-341.
- [25]Zhang L, Seitz S M. Estimating optimal parameters for MRF stereo from a single image pair[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2007, 29(2): 331-342.
- [26]Scharstein D, Pal C. Learning conditional random fields for stereo[C]//2007 IEEE Conference on Computer Vision and Pattern Recognition. IEEE, 2007: 1-8.
- [27]Haeusler R, Nair R, Kondermann D. Ensemble learning for confidence measures in stereo vision[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2013: 305-312.
- [28]Park M G, Yoon K J. Leveraging stereo matching with learning-based confidence measures[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2015: 101-109.
- [29]Zbontar J, LeCun Y. Computing the stereo matching cost with a convolutional neural network[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2015: 1592-1599.
- [30]Luo W, Schwing A G, Urtasun R. Efficient deep learning for stereo matching[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016: 5695-5703.
- [31]Tulyakov S, Ivanov A, Fleuret F. Practical deep stereo (pds): Toward applications-friendly deep stereo matching[C]//Advances in Neural Information Processing Systems. 2018: 5871-5881.
- [32]Mayer N, Ilg E, Hausser P, et al. A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016: 4040-4048.
- [33]Kendall A, Martirosyan H, Dasgupta S, et al. End-to-end learning of geometry and context for deep stereo regression[C]//Proceedings of the IEEE International Conference on

- Computer Vision. 2017: 66-75.
- [34]Chang J R, Chen Y S. Pyramid stereo matching network[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018: 5410-5418.
- [35]Zhang F, Prisacariu V, Yang R, et al. GA-Net: Guided Aggregation Net for End-to-end Stereo Matching[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2019: 185-194.
- [36]Tonioni A, Tosi F, Poggi M, et al. Real-time self-adaptive deep stereo[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2019: 195-204.
- [37]Khamis S, Fanello S, Rhemann C, et al. Stereonet: Guided hierarchical refinement for real-time edge-aware depth prediction[C]//Proceedings of the European Conference on Computer Vision (ECCV). 2018: 573-590.
- [38]Wang Y, Lai Z, Huang G, et al. Anytime stereo image depth estimation on mobile devices[C]//2019 International Conference on Robotics and Automation (ICRA). IEEE, 2019: 5893-5900.
- [39]Duggal S, Wang S, Ma W C, et al. DeepPruner: Learning Efficient Stereo Matching via Differentiable PatchMatch[J]. arXiv preprint arXiv:1909.05845, 2019.
- [40]Lin T Y, Dollár P, Girshick R, et al. Feature pyramid networks for object detection[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017: 2117-2125.
- [41]Tekalp A M, Tekalp A M. Digital video processing[M]. Upper Saddle river, NJ: Prentice Hall PTR, 1995.
- [42]Scharstein D, Szeliski R. Stereo matching with nonlinear diffusion[J]. International Journal of Computer Vision, 1998, 28(2): 155-174.
- [43]Bolles R C. The JISCT stereo evaluation[C]//Proc. of Image Understanding Workshop. 1993: 263-274.
- [44]Birchfield S, Tomasi C. A pixel dissimilarity measure that is insensitive to image sampling[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1998, 20(4): 401-406.
- [45]Kang S B, Szeliski R, Chai J. Handling occlusions in dense multi-view stereo[C]//Proceedings of the 2001 IEEE Computer Society Conference on Computer

- Vision and Pattern Recognition. CVPR 2001. IEEE, 2001, 1: I-I.
- [46]Terzopoulos D. Regularization of inverse visual problems involving discontinuities[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1986 (4): 413-424.
- [47]Geman S, Geman D. Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1984 (6): 721-741.
- [48]Barnard S T. Stochastic stereo matching over scale[J]. International Journal of Computer Vision, 1989, 3(1): 17-32.
- [49]Geiger D, Girosi F. "Parallel and Deterministic Algorithms for MRFs: Surface Reconstruction and Integration[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1991, 12: 401-412.
- [50]Bobick A F, Intille S S. Large occlusion stereo[J]. International Journal of Computer Vision, 1999, 33(3): 181-200.
- [51]Kanade T, Okutomi M. A stereo matching algorithm with an adaptive window: Theory and experiment[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1994, 16(9): 920-932.
- [52]Elad M. On the origin of the bilateral filter and ways to improve it[J]. IEEE Transactions on Image Processing, 2002, 11(10): 1141-1151.
- [53]Park H, Lee K M. Look wider to match image patches with convolutional neural networks[J]. IEEE Signal Processing Letters, 2016, 24(12): 1788-1792.
- [54]Shaked A, Wolf L. Improved stereo matching with constant highway networks and reflective confidence learning[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017: 4641-4650.
- [55]Gidaris S, Komodakis N. Detect, replace, refine: Deep structured prediction for pixel wise labeling[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017: 5248-5257.
- [56]Pang J, Sun W, Ren J S J, et al. Cascade residual learning: A two-stage convolutional neural network for stereo matching[C]//Proceedings of the IEEE International Conference on Computer Vision. 2017: 887-895.
- [57]Jie Z, Wang P, Ling Y, et al. Left-right comparative recurrent model for stereo

- matching[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018: 3838-3846.
- [58]Batsos K, Mordohai P. Recresnet: A recurrent residual cnn architecture for disparity map enhancement[C]//2018 International Conference on 3D Vision (3DV). IEEE, 2018: 238-247.
- [59]Ye X, Li J, Wang H, et al. Efficient stereo matching leveraging deep local and context information[J]. IEEE Access, 2017, 5: 18745-18755.
- [60]Liang Z, Guo Y, Feng Y, et al. Stereo Matching Using Multi-level Cost Volume and Multi-scale Feature Constancy[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2019.
- [61]Guo X, Yang K, Yang W, et al. Group-wise Correlation Stereo Network[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2019: 3273-3282.
- [62]Seki A, Pollefeys M. Sgm-nets: Semi-global matching with neural networks[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017: 231-240.
- [63]Song X, Zhao X, Hu H, et al. Edgestereo: A context integrated residual pyramid network for stereo matching[C]//Asian Conference on Computer Vision. Springer, Cham, 2018: 20-35.
- [64]Yang G, Zhao H, Shi J, et al. Segstereo: Exploiting semantic information for disparity estimation[C]//Proceedings of the European Conference on Computer Vision (ECCV). 2018: 636-651.
- [65]Poggi M, Pallotti D, Tosi F, et al. Guided stereo matching[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2019: 979-988.
- [66]Ilg E, Cicek O, Galessio S, et al. Uncertainty estimates and multi-hypotheses networks for optical flow[C]//Proceedings of the European Conference on Computer Vision (ECCV). 2018: 652-667.
- [67]Kendall A, Gal Y. What uncertainties do we need in bayesian deep learning for computer vision?[C]//Advances in Neural Information Processing Systems. 2017: 5574-5584.
- [68]Kim S, Min D, Kim S, et al. Unified confidence estimation networks for robust stereo matching[J]. IEEE Transactions on Image Processing, 2018, 28(3): 1299-1313.

- [69]Fu Z, Fard M A. Learning Confidence Measures by Multi-modal Convolutional Neural Networks[C]//2018 IEEE Winter Conference on Applications of Computer Vision (WACV). IEEE, 2018: 1321-1330.
- [70]Park M G, Yoon K J. Learning and selecting confidence measures for robust stereo matching[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2018, 41(6): 1397-1411.
- [71]Lin T Y, Goyal P, Girshick R, et al. Focal loss for dense object detection[C]//Proceedings of the IEEE International Conference on Computer Vision. 2017: 2980-2988.
- [72]Hosni A, Rhemann C, Bleyer M, et al. Fast cost-volume filtering for visual correspondence and beyond[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2012, 35(2): 504-511.
- [73]Yin Z, Darrell T, Yu F. Hierarchical Discrete Distribution Decomposition for Match Density Estimation[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2019: 6044-6053.
- [74]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.
- [75]Yang M, Yu K, Zhang C, et al. Denseaspp for semantic segmentation in street scenes[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018: 3684-3692.
- [76]Besse F, Rother C, Fitzgibbon A, et al. Pmbp: Patchmatch belief propagation for correspondence field estimation[J]. International Journal of Computer Vision, 2014, 110(1): 2-13.
- [77]Chen W, Fu Z, Yang D, et al. Single-image depth perception in the wild[C]//Advances in neural information processing systems. 2016: 730-738.
- [78]Dai J, Qi H, Xiong Y, et al. Deformable convolutional networks[C]//Proceedings of the IEEE international conference on computer vision. 2017: 764-773.
- [79]Zhu X, Hu H, Lin S, et al. Deformable convnets v2: More deformable, better results[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2019: 9308-9316.
- [80]Van der Schaaf A, van Hateren J H. Modelling the power spectra of natural images:

- statistics and information[J]. Vision research, 1996, 36(17): 2759-2770.
- [81]Hartley R, Zisserman A. Multiple view geometry in computer vision[M]. Cambridge university press, 2003.

攻读硕士学位期间取得的学术成果

- [1] **Zhang Y**, Chen Y, Bai X, et al. Adaptive Unimodal Cost Volume Filtering for Deep Stereo Matching[C]//AAAI. 2020: 12926-12934.
- [2] Yu S , **Zhang Y** , Wang C , et al. HMFlow: Hybrid Matching Optical Flow Network for Small and Fast-Moving Objects[C]// International Conference on Pattern Recognition (ICPR). 2020.

致 谢

这篇论文记录了我在研究生期间的成长，科研、生活，受到了许多人的支持和帮助，十分感谢大家，陪伴我成长。

首先，我要特别感谢我的指导教师——百晓老师。学术、生活都给了许多的帮助和指导，充足的计算资源、有意思的科研课题，让我真正体会到了科研的热情洋溢和用知识解释生活的快感。老师的悉心指导和栽培让我不断认识和反省自己的不足，逐渐走出科研低谷，打磨自己的科研价值观和人生观。我犯过无数大大小小的错误，老师都是非常耐心的点出我的问题，给我机会改正。您的信赖和支持，是我研究生涯最大的幸福。

感谢各位评审老师对本论文做出的修改建议，辛苦各位老师！

同时，也十分感谢深动科技的各位伙伴们，谢谢你们！

我也要感谢实验室的各个小伙伴，让我的研究生生活苦中带甜。王栋、张雪妮、余岁寒金，还有王翔，我们是同一时间进入到这个实验室的小伙伴，面对科研中的各种难题让我们互相鼓励和一起坚强。感谢严程、周雷、孙鹏、许凡各位师兄师姐，带我们玩乐，也带我们搞科研。感谢陈益民、仲崇明、王欣、黄晓丽、关晗金、焦履安、张竹君，给科研的枯燥增添了无数的欢乐。重点感谢一下陈益民，我们共同的努力最终成就了我们的AAAI论文。感谢徐清华、宋嘉钰、刘伟，让我对NLP有了诸多了解。祝徐清华、宋嘉钰还有我，往后的博士生涯顺利；当然，希望刘伟明年能跟我们汇合，在欧洲一起读博，一起欢乐。

在日常生活中，要特别感谢我的室友。跟刘伟同学快10年了，我们俩算是知根知底又志趣相投，希望有机会在欧洲一起读博搞科研；刘华帅，是大学一起参加了四年ACM程序设计竞赛的战友，也是共同奋斗了三年的研究生室友，最佩服的莫过于你的睿智，希望以后还有机会合作，有你hold my back让我很安心；王星河，像极了一个游荡于民间的打油诗诗圣，放荡不羁、行云流水而已。室友四人，京城的大大小小各个馆子基本都曾留下我们的身影，不论忙碌，不论悲喜，上桌即侃侃而谈，烦恼皆烟消云散。人生若得几饭友、几同寐、几知己，足以！也感谢其他未在此一一道谢的挚友们、老师们，谢谢你们，让我研究生生活如此精彩、顺利！

这里，也要特别感谢我的父母以及我的姐姐。感谢你们对我无微不至的关怀和爱，有了你们的支持和信任，我更加勇往直前。还有得感谢我姨妈、姨夫、表妹、表哥、表嫂、表姐，今年遇上疫情，没能按时返校。但是在姨妈家住的这两个月，我感受到了无

微不至的关怀，姨夫给我普及了很多的人生哲理和社会经验，教我为人处世。谢谢你们！

当然，还应该感谢自己，感谢自己没有放弃自己，辛酸、艰苦、快乐、幸福，我们都一起度过，希望未来的自己继续加油，继往开来！