激光与光电子学进展
先进成像

双目立体视觉研究进展与应用

杨晓立[1†]，徐玉华[1*†]，叶乐佳[1]，赵鑫[1]，王飞[2]，肖振中[1**]
[1]奥比中光科技集团股份有限公司，广东 深圳 518062；
[2]深圳奥芯微视科技有限公司，广东 深圳 518062

摘要 双目立体视觉模仿人类视觉系统对环境进行三维感知，通过对校正后的左右图像进行立体匹配获取两幅图像的视差，再根据三角测量原理计算出场景的深度，在近几十年中一直是计算机视觉领域的研究热点，取得了一系列的进展。传统的立体匹配方法采用手工设计的特征进行立体匹配，研究表明，这类方法对弱纹理或重复纹理区域以及遮挡区域表现不佳。近年来，基于深度学习的立体匹配方法取得了显著的进展，表现出了强大的应用潜力。本综述对这一不断发展的领域进行全面的调研，讨论不同方法之间的优点和局限性，介绍市面上的双目产品，并展望该领域的研究与应用前景。

关键词 立体视觉；立体匹配；深度学习
中图分类号 TP391 **文献标志码** A **DOI**：10.3788/LOP230457

Research Progress on Binocular Stereo Vision Applications

Yang Xiaoli[1†], Xu Yuhua[1*†], Ye Lejia[1], Zhao Xin[1], Wang Fei[2], Xiao Zhenzhong[1**]
[1]*Orbbec Inc, Shenzhen* 518062, *Guangdong, China*;
[2]*Shenzhen Oxin Technology Co., Ltd., Shenzhen* 518062, *Guangdong, China*

Abstract Binocular stereo vision simulates the human visual system to perceive the environment in three dimensions. The parallax between two images can be obtained by stereo-matching the corrected left and right images and calculating the scene depth based on the triangulation principle. This area has been identified as a research hotspot in computer vision and has made significant progress in the past few decades. Traditional stereo matching methods use hand-designed features that perform poorly in weak or repeated texture and occlusion areas. Recently, stereo matching methods based on deep learning have significantly progressed, showing strong application potential. In this review, we conducted a comprehensive survey on this developing field, discussed the advantages and limitations of the different methods, introduced the binocular products currently on the market, and assessed the research and application prospects in this field.

Key words binocular stereo; stereo matching; deep learning

1 引 言

双目立体视觉利用立体匹配算法对校正后的双目相机左右两幅图像进行密集匹配，建立两幅图像像素点之间的密集对应关系（用视差图表达），再根据相机参数恢复出场景的深度图像[1]。双目视觉的核心在于立体匹配算法，即给定左图像中的一个像素点(x,y)要确定它在右图中的对应点(x_R,y_R)，$d=x-x_R(y=y_R)$。深度Z与视差d的关系为$Z=BF/d$，其中，B为双目视觉系统的基线长度，F为等效焦距。双目视觉在三维建模[2]、汽车自动驾驶[3]、机器人自主导航[4]、无人机控制[5]、火星车定位[6]、月球车定位[7]、手机摄影（如背景虚化）[8]、机械手视觉引导[9]、摄影测量[10]、人流量分析[11]等方面具有广泛的应用。

立体匹配的难点在于如何降低传感器噪声、前后景遮挡、弱纹理或重复纹理区域、高反光区域、透明物体等因素的影响[12]。经过四十多年的研究，该问题仍然没有被完全解决，至今仍然是计算机视觉领域中研究的热点[12-13]。近年来，基于深度学习的立体匹配方法迅速发展，取得了显著的进展。2020年ECCV最佳

收稿日期：2023-01-05；**修回日期**：2023-02-12；**录用日期**：2023-02-22；**网络首发日期**：2023-03-02
基金项目：河套深港科技创新合作区深圳园区科研及创新创业项目（HZQB-KCZYB-2020098）
通信作者：*xyh_nudt@163.com；**xiaozhenzhong@orbbec.com
†共同第一作者

论文 RAFT[14]引领了基于多次迭代优化的网络结构[15-16]的发展，使双目深度估计在各大数据集上的精度达到新的高度。与此同时，一些基于3D卷积的网络结构[17-18]和基于Transformer的网络结构[19-20]也展现出了强大的应用潜力。在自动驾驶场景双目深度估计数据集KITTI2012[3]和KITTI2015[21]，以及Middlebury 2014数据集[22]上，与传统方法相比，基于深度学习的方法占据了绝对优势地位。

立体匹配精度的定量评价指标通常采用端点误差(E_{EPE})和误匹配率(R_{PBM})：

$$E_{\mathrm{EPE}}=\frac{1}{N}\sum_{i=1}^{N}\left|d-\hat{d}\right|, \tag{1}$$

$$R_{\mathrm{PBM}}=\frac{1}{N}\sum_{i=1}^{N}\left|d-\hat{d}\right|>\delta_d, \tag{2}$$

式中：N是有效像素的总数；d是预测视差；$\hat{d}$是真值视差标签；δ_d是误匹配的阈值。在评价时，通常还根据非遮挡区域误匹配率(nonocc)、所有像素点的误匹配率(all)、视差不连续区域误匹配率(disc)等进行评价。

然而，虽然基于深度学习的立体匹配方法取得了令人兴奋的进步，但仍然存在以下问题：

1)计算资源消耗问题：现有的精度较高的立体匹配网络需要在高端的GPU上才有可能达到实时效果。例如：GA-Net-15[23]处理一对分辨率为1242×375的图像需要1.8 s(TESLA P40 GPU)；PSMNet[24]需要0.41 s(Titan-Xp GPU)；AANet[12]需要62 ms(NVIDIA V100 GPU)；而实时网络BGNet和BGNet+[25]也分别需要25.4 ms和32.3 ms(GTX 2080Ti)。高端显卡从价格(比如英伟达的RTX 3090Ti GPU售价超过10000元)、功耗、体积等方面来说，都无法用于消费级应用。

2)泛化性能问题：现有大部分网络训练过程都是先在合成数据集上进行预训练，然后在某个特定数据集上微调(fine-tune)。训练出来的模型在这个数据集上性能很好，在别的数据集上性能急剧下降，这对于一个面向通用场景的深度传感器来说是不可接受的。

在接下来的第2节中，首先对传统立体匹配方法进行了简要的回顾；第3节对近8年基于深度学习的立体匹配方法进行了介绍；第4节介绍了立体匹配模型的加速方法；第5节对常用的立体视觉数据集进行了介绍；第6节介绍了双目视觉产品；最后是本文的总结和展望。

2 传统立体匹配方法

Scharstein等[1]在2002年把立体匹配分为4个步骤，即匹配代价计算、代价聚合(cost aggregation)、视差计算和视差优化，该框架一直沿用至今。

匹配代价计算的目的是计算左图中的一个像素点与右图中的某个像素点之间的匹配相似度。传统方法采用人工设计的代价计算函数，常用的匹配代价函数有基于单点灰度值比较的absolute intensity differences (AD)、考虑采样误差单点灰度值比较的BT[26]、结合单点亮度和梯度比较的代价函数、与图像亮度无关的Census变换、结合AD和Census变换的AD-Census[27]、归一化互相关(NCC)、互信息[28]等。Hirschmuller等[29]对各种匹配代价函数进行了系统的比较。

立体匹配方法可分为基于全局约束的方法和基于局部约束的方法。基于全局约束的方法利用整幅图像信息进行计算，能到达全局最优解。全局最优算法的本质是将对应点的匹配问题转化为寻找某一能量函数的全局最优问题，通常会跳过代价聚合步骤，直接计算视差值。这类算法的核心环节包括能量函数构造和能量函数的求解策略[30]。全局方法有图割法[31-32]、置信度传播[33-34]等。图割法的基本思想是将立体匹配问题转化为一种能量函数的形式，根据能量函数构造合适的图，并求其最小割(最大流)[35]。Sun等[33]把立体匹配问题考虑成在Markov网络框架下的全局匹配问题，用贝叶斯置信度传播算法在整个图像区域得到较好的结果。Scharstein等[1]将该方法与图割法进行了细致的比较。虽然全局方法具有精度高的优势，但计算速度很慢，难以用于实时的系统。基于局部约束算法利用兴趣点周围的局部信息进行计算，相应的计算复杂度较低。基于局部约束的立体匹配方法需要代价聚合的步骤，代价空间(cost volume)是代价聚合中的一个重要的概念。3D的代价空间C中的一点$C(x,y,d)$表示在像素(x,y)处、视差为d时的匹配代价。代价聚合通常在代价平面中进行。最基本的代价聚合方法是盒滤波(box filtering)[1]，等价于用一个矩形窗口进行均值滤波，其优点是计算效率高、计算速度与窗口的尺寸无关，缺点是不具备边缘保持特性。Yoon等[36]针对这个问题提出基于自适应窗口权重滤波的代价聚合方法，其本质是双边滤波[37]，同时考虑空间距离权重和颜色距离权重。双边滤波器具有良好的边缘保持特性，匹配时可以采用一个大的匹配窗口(比如35×35)，显著改善了视差图的边缘准确度。原始的双边滤波算法计算复杂度高，后续很多方法被提出，以提高自适应窗口权重滤波的效率。双边滤波器不是可分离滤波器，文献[38]采用一个近似的方法，分别在X和Y方向上用基于颜色和空间距离的权重进行1D滤波，以提高计算效率。He等[39]去掉了空间距离权重，只保留了颜色距离权重项，设计了可采用box filter高效实现的引导滤波器。Hosni等[40]采用引导滤波器进行代价聚合，是当年(2011年)性能最好的基于局部约束的立体匹配方法。自适应窗口权重滤波通常采用的是规则的矩形窗口，但窗口内像素点的权重是变化的。Zhang等[41]设计了一种基于十字叉的自适应形状窗口的代价滤波方法，虽然窗口形状不规则，但仍可以应用积分图进行高效计算。Hirschmuller[28]为了平衡精度和效率，提出半全局立体匹配方法，基于多方向的线扫描优化

进行代价聚合，该方法具有很强的实用性，如 Intel D435 双目深度传感器就用到了这种方法[42]。

基于窗口匹配方法通常都有一个隐含的假设，即该窗口内所有像素点都具有相同的视差值[43]，这个假设只有在相机的光轴垂直于目标平面时才成立。Bleyer 等[43]采用 3 个参数的视差平面对基于窗口的匹配过程进行更精确的建模，但同时显著扩大了参数优化的空间，为了提高效率，采用随机优化方法 PatchMatch[44]进行参数搜索。

Yang[45-46]提出一种基于最小生成树的代价聚合方法，每个改像素点都会受到其他所有像素点的支持（因此把这个方法称为非局部方法）。该方法具有边缘保持特性，且计算效率高。Li 等[47]利用多颗最小生成树为每个像素搜索 3 自由度的视差平面参数，在论文发表当年（2017 年）在 Middlebury 3.0 评测数据集[22]上取得最好的精度。

代价聚合完成以后，赢家通吃（WTA）[28]是最基本的视差计算方法。用 WTA 得到视差图后，通常还会采用一些视差优化方法进行处理，包括错误匹配剔除（如左右一致性检测、speckle filter 等）、亚像素视差值计算、遮挡区域空洞填充、中值滤波[28,40]，从而输出最终的视差图。

3 基于深度学习的立体匹配方法

3.1 混合方法

尽管卷积神经网络在目标检测、分类、识别等任务中已经流行了多年，但在 2015 年之前，立体匹配算法主要还是采用人工设计的代价函数。深度学习方法首次应用于立体匹配还是在 2015 年，Zbontar 等[48]设计了两种用于图像块（如 9×9 或 11×11 的图像块）匹配的网络结构，即快速结构 MC-CNN-Fast 和精确结构 MC-CNN-Acc。该方法只在代价计算的时候用到卷积网络，在代价聚合和视差后处理步骤中仍然采用传统方法，如十字臂滤波[41]、线扫描代价聚合[28]等。对于 MC-CNN-Fast，待匹配的图像块（如 9×9）输入网络后，通过一系列的卷积运算提取图像特征向量，再经过归一化操作把特征向量的模变为 1。最后，用相关运算对两个图像块的归一化特征计算匹配相似度。精确结构 MC-CNN-Acc 和 MC-CNN-Fast 的区别在于，卷积网络提取出特征以后没有用归一化和点积计算匹配度，而是把两个图像块的特征拼起来送入一个全连接网络，用网络去计算匹配度。这种结合深度学习和传统技术的方法通常被称为混合方法。

也有大量的工作将注意力放在更复杂的网络结构设计上，以获取更丰富的语义信息来提升立体匹配的精度。Zagoruyko 等[49]在他的工作中证明了更复杂的网络可以有效提升立体匹配的精度；Park 等[50]在文献[48]的基础上进行改进，采用金字塔池化来获得更大的感受野；Shaked[51]设计了一种新型的高速网络（highway network）来计算匹配代价。

除了特征提取和代价计算之外，也有工作将深度学习的方法融入立体匹配的其他模块之中。针对 SGM[28]中惩罚参数不易调节的问题，Seki 等[52]提出 SGM-Net，网络的输入为小图像块及其位置，输出为 3D 物体的惩罚参数。SGM-Net 采用路径代价与邻域代价这两种新型代价函数，使得网络在训练中可以有效使用现实场景中采集到的稀疏真值视差标签（比如激光雷达）。然而，由于无法直接获取 SGM 中的惩罚参数真值，只能通过三步法来生成弱标签，所以训练 SGM-Net 的过程较为复杂且耗时。

此外，还有不少非端到端立体匹配的方法[53-55]被提出，这些方法相较于传统方法在效果上有着不少的提升，但受限于缺乏图像全局信息、高额的计算负担，已经逐渐被端到端的方法所替代。

3.2 端到端立体匹配方法

基于卷积网络的图像块匹配（如 9×9 的图像块匹配）通常很耗时，且代价聚合和视差优化仍然采用传统方法，对遮挡区域和弱纹理区域的处理仍然存在很多问题。为了提高计算速度、改善遮挡区域和弱纹理区域的估计精度，采用端到端网络是一个更好的选择。端到端的立体匹配网络可以无缝包含立体匹配的全部步骤，直接从双目图像中估计出稠密深度图。自从最早的（2016 年）端到端视差估计网络 DispNet[56]被提出后，大量的工作都基于该结构展开并取得了有益的结果。代价空间可以分为 3D 代价空间与 4D 代价空间，通常分别由 2D 卷积网络和 3D 卷积网络进行代价聚合操作。DispNet[56]是基于 3D 的代价空间设计的，而基于 4D 卷积的 PSMNet[24]在论文发表当年取得了 KITTI 数据集第一名的成绩。2020 年，RAFT[14]获得了当年 ECCV 最佳论文的荣誉，在光流领域的多个数据集上均取得了第一名的成绩，是当时最先进的光流估计方法。鉴于光流估计和立体匹配两种任务的相似性，立体匹配领域开始有众多工作借鉴 RAFT[14]的方法和思想，使用 GRU 结构对视差进行多次迭代优化并在多个立体匹配数据集上取得了排名领先的成果。自从 2017 年 Transformer 模型[57]面世以来，其使用的自注意力（self-attention）结构取代了循环神经网络（recurrent neural network），并在自然语言处理领域取得突破性进展，就开始有不少工作尝试将 Transformer 模型应用于计算机视觉领域[58-60]，并取得了巨大的成功。同时在立体匹配和光流估计领域，也已经有工作将 Transformer 和自注意力结构应用于网络中，展现了令人惊叹的潜力。因此，本小节分别对基于 2D 卷积的网络、3D 卷积的网络、多次迭代优化的网络，以及 Transformer 的方法进行讨论和分析。

3.2.1 基于 2D 卷积的网络

DispNet[56]的结构和端到端光流估计网络 FlowNet[61]的结构（图 1）几乎是一样的，甚至比 FlowNet

更加简单，因为光流要考虑两个方向的位移估计，而视差估计只需要考虑水平方向的位置偏移。DispNet有两种结构：DispNetS和DispNetC。DispNetS把两张三通道的RGB图像拼成六通道作为网络的输入，经过具有9个卷积层的收缩网络后，再进入扩张网络。在扩张网络中，通过反卷积对特征图进行上采样，同时把收缩网络同分辨率的特征和上一层网络估计出来的视差图的上采样拼接（Concat）起来，再经过逐层上采样得到最终的视差图估计。DispNetC与DispNetS的主要区别在于，DispNetC会对左右图特征图进行相关运算，建立一个代价空间。随后将代价空间和左图特征拼接后，继续用卷积核滤波。由于DispNetC引入了立体匹配里面的代价空间的思想，其精度优于DispNetS。DispNetC在发表的时候，在KITTI2015[21]上的精度仅次于MC-CNN[48]，但计算速度比MC-CNN快1000倍。

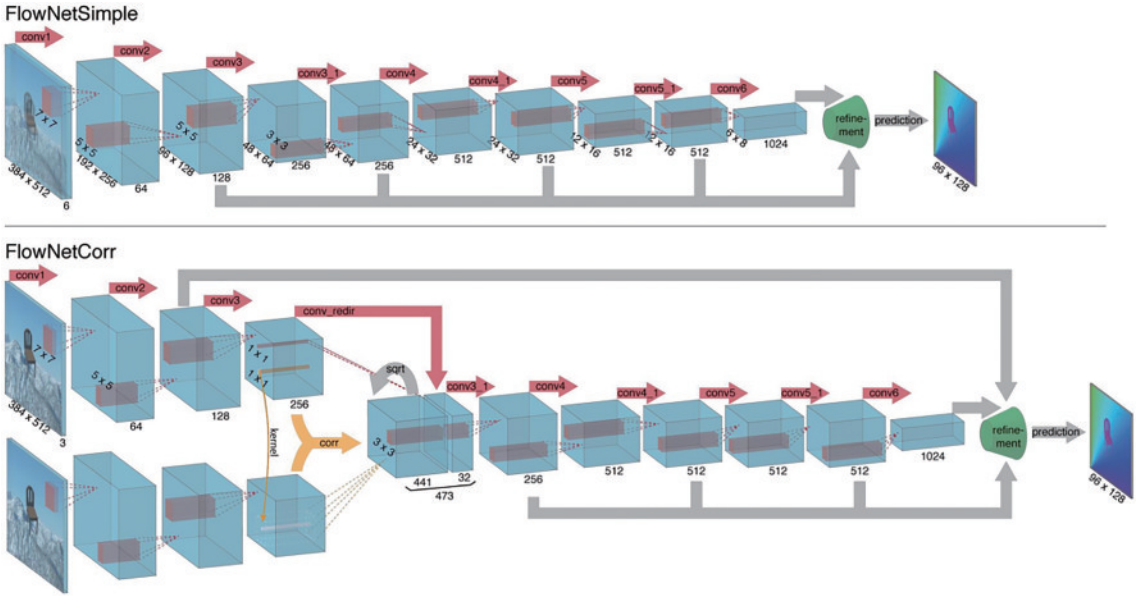

图1 FlowNetSimple与FlowNetCorr结构示意图[61]

Fig. 1 Schematic diagram of FlowNetSimple and FlowNetCorr structure[61]

为了进一步提高视差估计的精度，文献[62-63]设计了残差优化层（residual refinement layers）结构。残差优化层的输入通常包括当前估计的视差图、左图特征和重建误差（重建误差为左特征图与warping后的右特征图的差值）。此外，边缘信息[64]和图像分割[65]信息也被融入网络，以提高立体匹配的精度。AANet[13]将可变形卷积[66]引入立体匹配网络，设计两个有效且高效的成本聚合模块：自适应同尺度聚合模块（adaptive intra-scale aggregation）和自适应跨尺度聚合模块（adaptive cross-scale aggregation）来实现成本聚合，从而获得自适应形状的感受野，以完全替代现有SOTA的立体匹配模型中常用的3D卷积，在加快推理速度的同时保持较高的准确率。HITNet[67]把图像分成一系列的不重叠的图像块（tile），图像块的参数包含1个3自由度的几何参数和1个可学习的特征向量（可以解释为匹配的质量）。其视差估计过程主要包括特征提取、视差初始化和视差传播。该方法的特点在于：采用3自由度的视差平面表达每个tile的视差；视差初始化的时候，在全视差范围内计算视差值；在每个分辨率，初始匹配（看起来很糟糕，但包含细节信息）都用来恢复细的结构（在低分辨率可以重建大面积的弱纹理目标，但很难重建细小的结构）；没有完整的代价空间，所采用的窄代价空间只有3层，所以预测速度较快。Bi3D[68]将深度估计问题转变为分类问题，利用多个分类器对视差进行估计，判断场景内物体的深度值与给定深度距离的远近，从而提升网络效率，更契合应用需求。SMD-Nets[69]通过紧凑参数化的双峰混合密度对视差的前景和背景进行建模，以解决视差图边缘过于平滑的问题，提升了深度不连续处的锐利程度。同时，采用连续表达实现了有限的内存和计算量下的高分辨率输出。

基于2D卷积的立体匹配模型具有结构简洁和运行效率高的优点，在计算资源有限的情况下也能提供较高的精度。此外，由于2D卷积在各种算力平台（GPU、CPU、NPU等）上的成熟部署经验，该方法在工程化应用中具有天然优势。然而，由于三维代价空间提供的信息有限，该类网络结构在提高预测精度和鲁棒性方面仍需进一步研究。

3.2.2 基于3D卷积的网络

GCNet[70]的网络结构设计遵循着文献[1]中的立体匹配步骤，实现了端到端的立体匹配网络，如图2所示。首先使用2D卷积提取图像特征，并通过拼接左右图特征构建四维代价空间；然后采用3D卷积进行代价空间滤波，实现匹配代价聚合；最后采用soft argmax计算视差。虽然GCNet的精度被后续的很多网络结

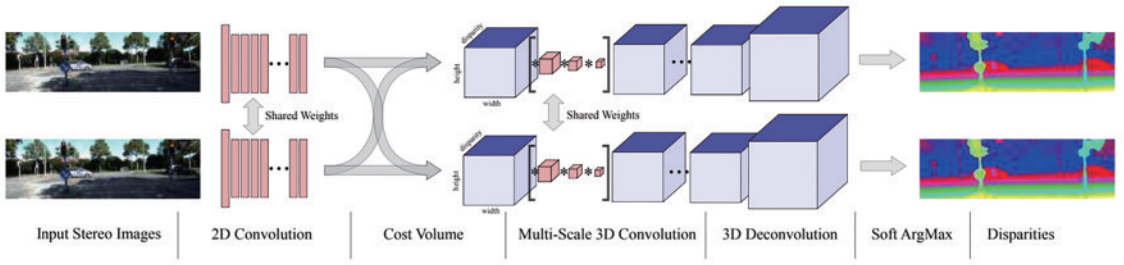

图 2 GCNet结构示意图[70]

Fig. 2 Schematic diagram of GCNet structure[70]

构超越，但该工作提出的3D卷积代价滤波和soft argmax被后面的工作广泛采用。

PSMNet[24]在刚发表的时候，占据KITTI排行榜第一名。特征提取用空间金字塔池化(SPP)模块，代价空间滤波采用堆叠沙漏(stacked hourglass)结构。GWCNet[71]的代价空间由两部分组成，其中，一部分和PSMNet类似(但通道数更少)，另外一部分把左右图特征按通道均匀地划分为若干组，按组进行相关来构建一个代价空间，减少网络参数的同时提高了网络的预测精度。StereoNet[72]是Google在2018年设计的一个实时立体匹配网络，处理速度可以达到60 frame/s。其加速策略是，在很低的分辨率(如1/8)计算代价空间计算出低分辨率的视差图，然后对视差图进行2倍上采样，将上采样的视差图与同分辨率的左图拼接后，送入视差优化网络。如此逐级优化，直至最高分辨率。GA-Net[23]将SGM[28]思想融入立体匹配网络，以提高网络的预测精度。DeepPruner[73]将PatchMatch[43]的思想引入网络中，基于预测的视差上下界建立一个窄的、用于加速的代价空间。Zhang等[74]为了改善立体匹配网络的泛化性能，提出域归一化(domain normalization)，其具体做法是先对每张特征图的每个通道进行归一化(减去该特征图通道的均值后除以标准差)，再在通道维度进行一次模归一化。Xu等[25]设计了一个高效的基于双边网格学习的代价空间上采样模块，可对现有的立体匹配网络(如GCNet[70]、PSMNet[24]、GANet[23]等)加速4～29倍，且保持相当的精度。基于这个具有边缘保持特性的代价空间上采样模块还设计了一个实时的网络BGNet，在当年(2021年)的实时网络中具有最高的精度(在PyTorch或TensorFlow平台)。Yao等[75]提出一种高效匹配高分辨率图像的方法，先在低分辨率上进行稠密匹配，估计出哪些区域存在细节损失，然后在高分辨率上对该区域进行稀疏匹配，从而提高网络效率、降低显存消耗，最终能够对5000×3500的图像进行立体匹配。为了解决立体匹配网络因数据集场景及视差分布差异而缺乏泛化性的问题，CFNet[18]通过融合级联代价空间的方式提升匹配算法的鲁棒性，并采用基于方差的不确定估计方法，通过估计级联中不同阶段的不确定度自适应调整下一阶段的视差搜索范围，以减少错误匹配。AcvNet[17]提出一种新的代价空间构建方法，生成注意力权重以抑制冗余信息并增强连接体中的匹配相关信息，引入多级自适应块匹配以提升在不同视差甚至无纹理区域时的匹配代价独特性。2022年发表时，AcvNet在KITTI、SceneFlow、ETH3D等立体匹配数据集上均取得排名前三的成绩。

基于3D卷积的网络结构设计一般遵循经典的立体匹配步骤，具有较好的可解释性。相比于三维代价空间，四维代价空间能够提供更多的细节信息，因此所预测出的视差图精度更高。然而，由于3D卷积的高昂计算成本，如何设计更高效率的网络架构成为了亟须解决的问题。

3.2.3 基于多次迭代优化的网络

RAFT[14]是用于光流估计的网络，但只需将网络结构进行小的修改之后，即可无缝应用到立体匹配任务中。RAFT的具体处理流程如下，其结构如图3所示。

1) 特征提取：使用由6个残差层组成的网络[76]对左右图进行特征提取，在特征提取的过程中不断降低特征图的分辨率，同时增加特征的通道数，从而获取分辨率为输入图片分辨率1/8的左右图的特征图。

2) 相似性计算：相似性为左右特征图的内积，基于相似性可得到一个四维张量的代价空间，它提供了关于大小像素位移的关键信息。需要注意的是，这里的相似性计算不应与FlowNetCorr中的相关操作混淆，RAFT由特征内积得到两个特征图的全相似性，最大视差为图像的宽度。之后再对代价空间进行池化，构建4层的代价空间金字塔，从而保证同时捕捉到较大和较小的像素位移。

3) 代价空间查找表：依据当前估计出的光流/视差，在构建好的代价空间金字塔上的对应区域进行取值。

4) 多次迭代更新：构建GRU卷积循环网络，使用上一次估计的光流/视差、相关信息张量和循环卷积的隐藏状态作为输入，输出光流或视差的更新值、更新后的隐藏状态以及上采样特征。

5) 光流/视差上采样：基于循环网络输出的上采样特征对低分辨率的光流/视差进行加权组合，可获得比双线性插值上采样更优秀的高分辨率视差图。

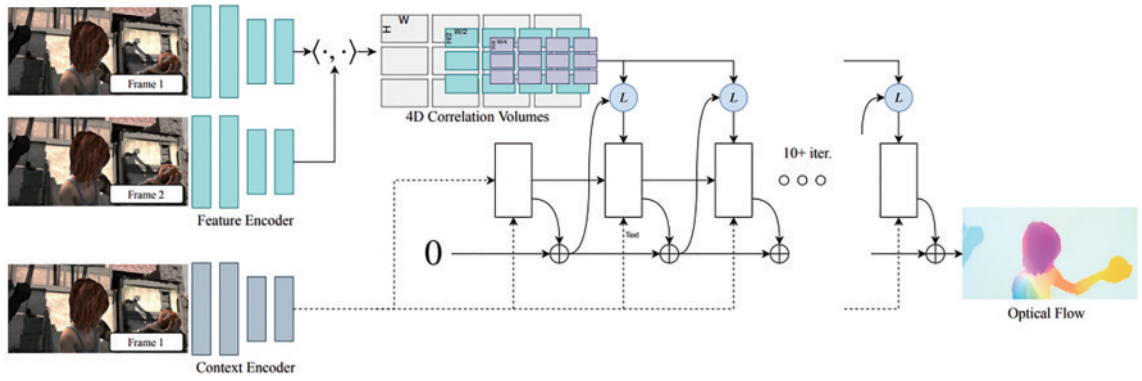

图 3　RAFT 结构示意图[14]

Fig. 3　Schematic diagram of RAFT structure[14]

RAFT-Stereo[15]是将光流算法 RAFT[14]应用于立体匹配的工作，并进行了以下优化：1）由于立体匹配任务只需要考虑水平方向视差，RAFT-Stereo 只在水平方向上进行相关操作；2）RAFT 的更新模块在固定的分辨率尺度上运行，而 RAFT-Stereo 使用多分辨率更新算子，能同时在不同分辨率的特征图上更新，多个 GRU 单元之间交错使用隐藏状态，但还是在最高分辨率的 GRU 上进行代价空间查找并输出最终视差；3）采用慢快 GRU 策略，由于在高分辨率特征上更新 GRU 模块的计算量远大于在低分辨率特征上更新的计算量，所以为了降低网络计算量，高频在低分辨率上更新、低频在高分辨率上更新。最终，RAFT-Stereo 在 ETH3D、Middlebury、KITTI 数据集上均取得了名列前茅的结果。

CREStereo[16]同样是基于 RAFT[14]的改进工作。针对真实场景中双目相机存在非理想矫正、相机模块不统一等问题，CREStereo 提出自适应分组局部相关模块（adaptive group correlation layer），通过局部特征注意力、2D-1D 局部交替搜索、可变性搜索窗口和分组相关等子模块提高了模型的鲁棒性。同时，CREStereo 还设计了递归更新模块和级联堆叠模块，使模型可以更好地恢复精细的深度细节。在模型推理时，使用堆叠级联的方案，即对输入图像进行下采样，构建一个图像金字塔，并将其输入同一个经过训练的特征提取网络以获取不同分辨率的特征信息。在论文发表时，CREStereo 在 Middlebury 和 ETH3D 数据集上均取得了第一的优异成绩。

RAFT 等基于迭代优化网络结构，通过设计巧妙的轻量 GRU 模块多次进行迭代更新，在光流估计和立体匹配领域取得了突破性的进展，具有强大的泛化能力与鲁棒性。然而受限于序列化的迭代更新结构，推理时间会随迭代次数线性增加，且难以进行速度优化。

3.2.4　基于 Transformer 的网络

Google 于 2017 年提出 Transformer 模型[57]之后，又于 2020 年提出了将 Transformer 模型应用在图像分类的模型 ViT[59]，其主要思路为将图像分成固定尺寸的图像块（patch），并通过线性变换将其转变为向量，再将其输入 Transformer 模型中分类。实际上 ViT 仅使用了 Transformer 模型中的编码器部分提取图像特征，而没有使用解码器部分。鉴于 ViT 取得的优异成果，之后就开始出现了将 Transformer 应用于立体匹配或光流估计的工作，比如 STTR[77]，从序列到序列（seq2seq）的角度去看待深度估计问题。STTR 首先使用卷积神经网络对双目图像进行特征提取，之后代价空间的建立使用 Transformer，包括自注意力机制和交叉注意力模块，分别对相同图像和两张图像的注意力信息进行提取，随着注意力层数的增加，注意力模块会更关注局部语义信息，比如在大范围无纹理区域中会更倾向于边缘等主要特征，这有助于 STTR 解决歧义。同时，STTR 认为仅仅依靠注意力模块是不够的，还使用相对位置编码的方法来提供位置信息，采用最优传输（optimal transport）理论和注意力模板（attention mask）的方法对立体匹配进行限制。STTR 在 KITTI 等数据集上取得了和基于卷积的方法相当的结果，甚至与为高分辨率图像设计的多分辨率网络相比也不逊色，证明了 Transformer 在立体匹配任务中的可行性。

在光流估计领域，基于 Transformer 框架的 FlowFormer[19]和 GMFlow[20]等模型，在 Sintel 数据集排名榜上已经超过 RAFT[14]，取得了主导地位。针对 RAFT 推理时间会随迭代次数线性增加而且难以进行速度优化的问题，GMFlow[20]验证了无需大量迭代，同样可以取得很好的光流估计效果，同时速度更快的假设。GMFlow 将光流重新定义为一个全局匹配问题，在使用卷积神经网络提取出图像特征后，通过 Swin Transformer[60]（包括位置编码模块、自注意力机制和交叉注意力模块）提取出更强力的图像特征，再进行全局匹配计算图像之间的相似性。FlowFormer[19]同样使用卷积神经网络提取图像特征，在计算出全局代价空间之后，使用 Transformer 模块的代价空间编码器对代价空间进行编码，再使用循环解码器通过迭代更新的方法对编码后的代价空间进行解码，并估计出光流。

在计算机视觉领域，基于Transformer的网络结构已经在检测、分割、光流估计等方面取得了显著的成果。随着对Transformer结构的不断深入研究，将会出现更多基于Transformer结构的立体匹配模型，展现出强大的应用潜力。

3.2.5 损失函数

有许多工作[24,70,72,15]使用L1误差[式(4)]或者Smooth-L1误差[式(5)]作为训练立体匹配网络的损失函数，直接回归视差，具有简单直观、对异常误差值不敏感的特点。在一些多任务学习工作中，还会额外使用辅助损失函数来引导立体匹配任务。

$$L_{\text{disparity}}(d)=\frac{1}{N}\sum_{i=1}^{N}L\left(d_i-\hat{d}_i\right), \tag{3}$$

式中：N是有效像素的总数；d是预测视差；$\hat{d}$是真值视差标签；L可以是L1误差：

$$L_1=|x|, \tag{4}$$

也可以是Smooth-L1误差：

$$L_{\text{Smooth-L1}}=\begin{cases}0.5x^2, & \text{if}|x|<1\\ |x|-0.5, & \text{otherwise}\end{cases}。 \tag{5}$$

比如Edgestereo[64]引入边缘损失函数[式(6)]，将边缘信息和边缘正则化整合到视差预测网络中，同时估计视差图与边缘信息，从而提高模型在细节处的预测效果。

$$L_{\text{ds}}(d)=\sum_{i=1}^{N}|\partial_x d|\exp(-|\partial_x\varepsilon|)+|\partial_y d|\exp(-|\partial_y\varepsilon|), \tag{6}$$

式中：∂d是视差图的梯度；$\partial\varepsilon$表示边缘概率图的梯度。

视差只是估计代价空间过程中的副产物，在学习中容易产生过拟合，可能会导致代价空间存在多峰分布，因此也有不少工作将关注点放在了如何优化代价空间上。AcfNet[78]对代价空间进行了单峰约束，将视差真值标签转换为单峰分布，并直接对代价空间进行监督。

$$L_{\text{SF}}(d)=\frac{1}{N}\sum_{i=1}^{N}\left\{\sum_{p=0}^{D-1}\left[1-P_i(p)\right]^{-\alpha}\cdot\left[-P_i(p)\cdot\ln\widehat{P}_i(p)\right]\right\}, \tag{7}$$

式中：$\hat{P}(p)=\text{Softmax}(-c_d)$；由真值标签构造的理想代价空间$P(p)=\text{Softmax}\left(-\frac{|\hat{d}-p|}{\sigma}\right)$；$c_d$是由网络预测出的代价空间；参数$\sigma$用于控制代价空间的尖锐程度，$\alpha>0$时为聚焦参数(focusing parameter)，$\alpha=0$时该损失函数为交叉熵损失。

3.3 自监督与弱监督立体匹配

训练立体匹配网络需要大量数据，但在实际使用中，获取大量拥有准确且稠密标签的双目数据是较为困难的，相对而言，获取无标签或拥有稀疏标签的数据则更为容易且可行。因此，一些基于无标签或稀疏标签数据的工作被提出，并取得了可喜的成果。

3.3.1 自监督立体匹配

自监督是数据能够提供监督信息的一种无监督学习方式，因此在没有标签数据的情况下，如何快速获取可靠的数据标签是自监督立体匹配的关键。在光流估计领域，文献[79-80]展示了自监督方法的可行性，主要思路为基于图像亮度一致性的假设，给定第1帧图像和估计出的光流，对第2帧图像进行重构，再使用重构图像的数损失函数替代标签作为监督信号。在立体匹配领域，Deep Stereo[81]通过选择相邻图片上的像素生成新视角下的图像。Deep3D[82]使用DispNet[56]的网络结构，将一组双目图像的左图作为输入对双目图像的右图进行重构，该工作为2D电影转成3D电影提供了新的实现方法。文献[83]使用双目图像作为输入预测视差图，再根据预测出的视差图和右图，使用扭曲(warp)操作重构出左图，再使用重构出的左图和输入左图作为损失函数训练模型。文献[84]通过引入可导的双线性插值来解决warp变换反向梯度传播的问题，使得整个算法可以端对端地训练学习；文献[78]提出了图像重构损失和光滑度损失两个自监督损失函数，进一步优化网络性能；文献[85]提出了图像重构损失和光滑度损失两个自监督损失函数，进一步优化网络性能，在KITTI数据集上，其精度甚至超过了一些监督学习的方法。

3.3.2 弱监督立体匹配

基于监督学习的立体匹配从带有标签的数据中学习，但是双目图像的标签不易获取，而且由于传感器对于透明物体、低反射物体等情况也存在一定困难，获取到的标签也存在噪声，因此也有不少工作致力于解决该问题。文献[86]首先使用传统立体匹配方法(AD-CENSUS[87]或SGM[28])制作出伪标签，再通过一些方法(比如文献[88])对伪标签进行置信度量，接着选择伪标签上的可靠标签作为稀疏标签对网络进行训练。文献[89]使用基于残差FlowNet[61]的结构估计出视差图，再使用监督学习的方法对从视差图和激光雷达获取到的稀疏深度图进行监督、自监督学习的方法对重构出的图像与原始图像进行监督。文献[90]提出了一种迭代的框架：首先随机初始化网络参数，生成初始深度图；再对深度图使用左右一致性校验获取高置信标签，接着使用高置信标签对网络进行训练；更新权重后再重复以上视差图预测和高置信度标签获取两个步骤。随着迭代次数的增加，视差图将会越来越准确，从而达到收敛状态。文献[91]首先使用传统方法估计双目图像的视差，再从中筛选出高置信度的稀疏点，并输入事先已经训练好的深度补全网络中，以获取稠密的视差图，再使用该稠密的视差图作为标签数据训练立体匹配网络。

4 立体匹配模型加速

随着新的立体匹配方法不断提高匹配模型的精度，模型的耗时也随之增加，高精度的匹配网络需要在高端GPU上才有可能实时运行，制约了立体匹配模型在消费级产品上的应用，因此也有不少工作开始关注如何对立体匹配模型进行加速。

4.1 轻量化模型结构

受限于计算代价过高的代价空间以及用来进行代价聚合时含参量较大的3D卷积，网络模型往往十分冗余，因此许多工作将目标放在如何设计更轻量化的模型结构上。一种思路是将3D卷积模型轻量化。StereoNet[72]是第一个实时的双目立体匹配网络，能够在英伟达Titan X上达到60 frame/s，StereoNet使用3D卷积进行立体匹配，但发现就算把代价空间设计得比较小巧，它仍然包含较多的特征信息，只会有较少的精度损失，网络在得到一个粗糙的低分辨率视差图后，通过一个边缘敏感的精修网络，得到更加细致的、保留边缘的视差图。AnyNet[92]同样基于视差图由粗糙到精细优化的思想，在获取到低分辨率初始视差后，使用更高尺度的左右图特征和扭曲操作计算视差的残差并不断纠正视差图。AnyNet能处理1242×375分辨率的双目图像，在Nivida Jetson TX2上达到10～35 frame/s，在实时性和精度上取得了平衡。BGNet[25]通过设计高效的基于双边网格学习的代价空间上采样模块，在实现实时网络的同时还保证了精度。

另一种思路是使用2D卷积网络，文献[93]指出仅仅使用2D卷积，同样可以取得速度与精度平衡的结果。MobileStereoNet[94]基于MobileNets[95-96]结构，设计了两种分别基于2D和3D代价空间的立体匹配网络，通过在模型之中大量使用可分离卷积，有效降低了模型的参数量和计算量。HITNet[67]采用更有效的视差优化方案，不建立完整的代价空间、不使用3D卷积，实现了模型的轻量化。文献[93]不使用卷积神经网络提取图像特征，而是直接利用传统算法快速得到初始的匹配代价，再通过2D卷积对其进行代价聚合，并基于U-Net2D卷积的编码-解码结构得到最终的视差图，其在1080Ti GPU上可以实现50 frame/s的速度。

4.2 模型压缩

另一种方法就是对立体匹配模型进行压缩。对于网络结构优化的研究于20世纪90年代就已经开始了[97-98]，但当时的应用对于网络压缩并无太多需求，因此没有引起太多关注。由Han等[99-100]于2016年发表的一系列经典论文，对网络模型压缩有着里程碑式的意义，引领了模型轻量化与加速方向的研究。文献[99]使用剪枝(pruning)、量化(quantization)、霍夫曼编码(Huffman coding)对网络进行压缩，在没有精度损失的前提下，使网络压缩了35～49倍。

4.2.1 模型剪枝

按照剪枝的粒度划分，剪枝可以分为层间剪枝、特征图剪枝、卷积核剪枝以及核内剪枝[101]。前3种为结构化剪枝，可以直接在硬件中加速。而核内剪枝属于非结构化剪枝，所得到的权重矩阵是稀疏的、不规则的，需要特别的硬件支持才能明显加速。

网络剪枝的核心问题就是如何有效裁剪模型并最小化精度损失。最直接的方法就是逐层判断权重的重要性，再将重要性较低的权重裁剪掉。这里的重要性判断可以基于权重的大小[102-103]、基于batch normalization(BN)层[104]的尺度、也可以基于激活层[105]的输出。而文献[106]则提出权重的重要性判断不一定基于权值，也有研究将最小化裁剪后的特征重建误差作为裁剪依据[107-108]。以上关于模型剪枝的研究关注如何裁剪模型参数，而需要裁剪多少模型参数也同样重要。按文献[109]的定义，模型权重的稀疏率可以分为预定义和自动两种方式。在预定义的方法中，会人为地设计某些规则，分析和设置网络中每一层的稀疏率。比如文献[95,110]将网络中卷积的宽度、深度和特征图的分辨率按特定比例一起调节。文献[111]使用强化学习自适应决定每一层的压缩比例。NetAdapt[112]设计了一种渐进式的剪枝方法，以硬件的耗时、功耗等作为返回指标对网络进行迭代裁剪。在对网络进行剪枝后，网络准确率会下降，这就需要使用微调(fine-tune)策略恢复其准确率。文献[113]提出了一种微调策略，在微调过程中，逐渐将稀疏率从初始的稀疏度值提升到目标稀疏率。

4.2.2 模型量化

对网络的量化一般包括训练后量化和训练量化[114]。训练后量化(posttraining quantization)是指在模型训练完成后再进行量化[115]，由于其不需要重新训练模型，易于工程化落地，因此在业界得到了广泛的应用[116]。对于训练后量化，核心问题就是如何减少由量化范围和量化位宽(bit-width)等因素引起的模型精度下降。对于量化范围的设定，一类方法是基于统计近似的方法[117]通过校准数据集来统计量化范围，另一类是基于优化的方法[115,118]。以上方法关注如何在预训练模型上找到最优的量化参数，而另一个思路则是使模型权重的分布更合适量化。文献[119]提出了功能不变的网络变换，可以减小网络的量化范围。文献[120]在训练网络时使用峭度正则化，使模型权重的分布对量化噪声有更高的容忍度。

相比于训练后量化，训练量化可以获得更高度精度[114]。对于训练量化，由于量化操作是个阶跃函数，它的导数不存在或者为0，这就给基于链式求导的网络训练带来了困难[121]。对这个问题，使用最广泛的方案是由Bengio等[122]提出的straight through estimator(STE)。该方法把量化操作的导数用1近似，其本质是对量化函数梯度的近似，会造成前向传播和反向近

似的梯度不匹配的问题。另一种思路是使用一个可微的函数逼近，比如在文献[123]中，训练时用带参数的tanh函数不断逼近不光滑的sign函数。

最近几年，业界对于低位宽量化的关注越来越多[121,124]，但这类量化会引起非常大的精度损失，所以业界的主要工作就是维持网络低位宽量化的准确率。文献[121]将模型权重和激活输出都量化为+1或−1，为了解决sign函数不可导的问题，在求导时使用hard tanh来近似量化操作。另外，为了进一步减少训练时BN的开销，给出了基于shift操作的shift-based batch normalization(SBN)。文献[125]提出DoReFa-Net，它基于STE[122]将量化位宽拓展到任意位，将量化对象扩展到权重、激活输出和梯度。针对低精度网络训练难的问题，INQ[126]对于网络权重的量化，基于渐进式神经网络量化的思想，引入参数分组、量化、重训练等3种操作，将预训练网络中的每一层参数分为两组，通过两组交替固定与训练来补偿量化给模型造成的精度损失。

4.3 NPU 计算平台

大幅减少权重的数量并不一定会提高现有加速器(如GPU)的性能和能效，这些加速器擅长处理规则和密集的神经网络，但缺乏对不规则稀疏模型的专门支持。新型神经网络加速器不仅可以有效地应对原始的密集神经网络，而且还可以有效应对严重修剪过的稀疏神经网络。以寒武纪的Cambricon-X[127]NPU为例，它是一个基于处理元素(PE)的体系结构，该体系结构由多个处理元素和一个缓冲区控制器(BC)组成，以利用神经网络模型的稀疏性和不规则性。具体而言，BC集成了一个有效的索引模块，用于从集中式神经元缓冲区中只选择所需的神经元，然后将此类神经元传输到连接的PE，减小带宽需求。在接收到此类神经元后，PE可以使用本地存储的压缩突触执行高效计算。此外，由于突触的不规则分布，多个PE以异步方式工作，以提高效率。与早期的神经网络加速器DianNao[128](不支持非结构化剪枝加速)相比，可以获得7.23倍的加速性能和6.43倍的能效提升。而与具有稀疏库(cuSPARSE)的GPU相比，该加速器平均可实现10.6倍的加速和29.43倍的能效提升。与带有稀疏库(即稀疏BLAS)的CPU相比，该加速器平均可实现144.41倍的加速。由此可见，支持非结构化剪枝的NPU在对稀疏网络加速方面具有显著优势。

5 立体视觉数据集

最初，研究人员为了评估立体匹配算法的性能，用高精度结构光三维扫描仪建立了评测数据集，如Middlebury[1,22,129]。近年来，为了训练基于深度学习的立体匹配网络，需要大规模的立体视觉数据集。但与目标识别、目标检测等任务不同，立体视觉数据集无法用人工的方式进行标注，一种方案是采用高精度的三维扫描设备采集立体视觉数据，但也存在效率低、对于透明物体、低反射率物体表面、远距离物体，扫描仪难以获得这些物体的表面深度[12,130]等问题。对于驾驶场景，通常采用激光扫描仪获得稀疏或半稠密的深度图[3,21,131]。在立体视觉研究领域，合成数据集(如SceneFlow[56]、IRS[132])是一个常用的方法。但基于合成数据集训练出来的模型，存在域适应的问题[74]。

表1列出了常用的双目立体视觉数据集。其中：Dataset name是数据集的名字；Type是数据的类型，分为真实数据(Real)和合成数据(Synthetic)两种；Year是制作数据的年份；Resolution是双目图像的分辨率；Scenes、Training scenes、Testing scenes分别是双目数据的总数、训练集数量以及测试集的数量；Label type

表1 立体视觉数据集

Table 1 Stereo vision dataset

Dataset name	Type	Year	Resolution	Scenes	Training scenes	Testing scenes	Label type
KITTI2012[3]	Real	2012	1240×376	389	194	195	Sparse
Middleburry[22]	Real	2014	2948×1988	30	15	15	Dense
KITTI2015[21]	Real	2015	1242×375	400	200	200	Sparse
Scene Flow[56]	Synthetic	2016	960×540	39049	34801	4248	Dense
ETH3D[133]	Real	2016	940×490	47	27	20	Dense
Falling Things[134]	Synthetic	2018	960×540	61500			Dense
IRS[132]	Synthetic	2019	960×540	100025	84946	15079	Dense
HR-VS[135]	Synthetic	2019	2464×2056	780			Dense
	Real	2019	2424×1918	33			Dense
DrivingStereo[131]	Real	2019	1762×800	182188	174437	7751	Sparse
ApolloScape[136]	Real	2019	3130×960	5165	4156	1009	Sparse
InStereo2K[12]	Real	2020	1080×860	2050	2000	50	Dense
Booster[130]	Real	2022	4112×3008	419	228	191	Dense
CRE[16]	Synthetic	2022	1920×1080	40000			Dense

为真值标签的类型，分为稀疏(Sparse)和稠密(Dense)两种。

6 双目视觉产品

目前市面上存在大量双目立体视觉产品，例如：D455[42]是由 Intel® RealSense™所出品的双目深度相机，被广泛应用在避障、物件测量、人脸识别等领域；Gemini 2 是基于奥比中光(ORBBEC)[137]全新的 MX6600 深度计算芯片所开发的，提供六轴惯性传感数据，支持深度和 RGB 帧同步以及多机同步，相机还集成了多分辨率下深度信息与彩色信息空间对齐功能，可实现高性能中远距离的室内/半室外环境感知。从实现的算法上来看：众多的产品[42,138-139]使用了传统立体匹配算法，比如 SGM 算法[28]；也有产品使用了基于深度学习的算法[140]。从实现的硬件算力芯片上来看，有的产品采用 ASIC 芯片计算深度图像[131-132]，有的产品[141]使用现场可编程逻辑门阵列(FPGA)，也有产品[142]使用专用的 VPU 芯片[143]，还有产品[138]使用移动端 GPU[144]。表 2 列出了市场上常见的双目视觉产品。其中：Company 为厂商的名称；Camera 是产品型号；Depth resolution and fps 是产品所输出深度图的分辨率和帧率；Depth accuracy 是深度图的精度；FOV 是相机深度图水平和竖直方向的视场角；Depth range 是深度相机的工作距离，单位是 m；Baseline 是深度相机的工作距离，单位是 mm。

表 2 双目视觉产品

Table 2 Stereo vision products

Company	Camera	Depth resolution and fps	Depth accuracy	FOV /(°×°)	Depth range /m	Baseline /mm
Intel RealSense[42]	D455	1280×720@30 848×480@90	＜2% up to 4 m	87×58	0.4-20	95
ORBBEC[137]	Dabai Pro	640×400@30 320×200@30	±6 mm@1 m	67.9×45.3	0.2-2.5	40
	Gemini 2	1280×800@30 640×400@60 320×200@60	≤2%(1280×800@2m & 81% ROI)	91×66	0.15-10	50
STEREOLABS[140]	ZED 2i	4416×1242@15 3840×1080@30 2560×720@60 1344×376@100	＜1% up to 3 m ＜5% up to 15 m	110×70	0.3-20	120
DUO[145]	DUO MLX	752×480@45 640×480@49 640×240@98 640×120@192 320×480@86 320×240@168 320×120@320		165° Wide angle lens		30
Rubedos[138]	VIPER	up to 1280×720 up to 40		70×43	1-15	200
Carnegie Robotics[146]	MultiSense KS21	up to 1920×1200 Up to 30	10 m:≤1.5% 20 m:≤3%	135×84	0.5-20	210
Blaxtair[139]	Omega	1024×512@11 512×256@18		100×70	0.5-10	100
Human+[141]	PSP010-800	1280×800@25	10 m:≤2% 30 m:≤6%	70×50	1-30	160
Percipio.XYZ[147]	PM801-E1	1280×960@1 640×480@1 320×240@1	3.45 mm@2500 mm	64×49	0.8-4.3	300
OAK[142]	OAK-D-Pro	1280×800@120		80×55	0.2-35	75

需要强调的是，基于 ASIC 芯片的双目深度相机，如 Intel RealSense D455 和奥比中光的 Gemini 2，这种深度相机在成本和功耗方面都具有显著优势。由于带有 ASIC 深度计算引擎，这些双目产品不需要使用应

用平台的算力就可以直接输出稠密的深度图像，便于大规模应用。

7 总结与展望

本综述对双目立体视觉的研究进展与应用进行了介绍，包括传统立体匹配方法和基于深度学习的立体匹配方法。其中，对于基于深度学习的立体匹配方法，介绍了混合方法、端到端立体匹配方法、自监督与弱监督立体匹配方法，同时也对如何加速立体匹配模型的方法进行了总结，还对常用的立体匹配数据和双目立体视觉产品进行了介绍。

目前市面上尽管存在大量双目立体视觉学术研究成果和产品（比如 Intel D435/D455、奥比中光的 Gemini 2等），但这些方法和产品仍存在着适应性和预测精度有待提高的问题。因此，对双目立体视觉技术的未来进行了以下展望：

1）基于深度学习的立体匹配方案有望在边缘端工程大规模应用。目前现有的双目立体视觉产品大多是基于传统方法而实现的，而基于深度学习的立体匹配方法在预测精度等指标上已经领先于传统方法。随着 NPU、移动端 GPU 计算平台的成熟、算力的不断提高，在边缘端的实现成为可能，这对基于深度学习的立体匹配方法的模型轻量化提出了更高的要求。同时，由于深度学习模型是在特定训练数据集上进行训练的，容易产生过拟合问题，因此如何提升模型的泛化能力还需要深入探索。

2）复杂光照环境与复杂物体材质问题仍有待解决。较差的光照环境（如雨、雪、雾环境）与复杂的物体材质（如透明物体）是现有立体匹配方法面临的一大难题，如何从硬件设计的角度或图像处理的方向提升图像质量仍需要研究。同时，如何改善网络的性能、提升网络对物体语义的理解能力，也是一种可能的解决方案。

3）高分辨率视差图估计。现有的大部分立体匹配方法和产品都只能获取到较低分辨率（小于 1920×1080）的视差图或深度图，而低分辨率的视差图感知较小物体或远距离物体的能力欠缺。同时，由于高分辨率视差图的获取往往会带来更多的时间消耗，因此如何高效获取高分辨率视差图成为了未来发展的挑战。

4）能够描述几何细节特征的视差图稀疏表达。现有的深度图或视差图大多都是用逐像素点的深度图或视差图进行表达的。比如，为了估计分辨率为 1280×800 的图像的深度，网络需要预测 102.4 万个变量。而实际上，场景的深度可以用更加紧凑的方式进行表达。比如，用 4 个点就可以表示很大的一面墙壁的深度。桌子、椅子也可以用一些稀疏的点对他们的深度进行完整的表达，虽然所需要的点会比表达平面更多，但远远小于百万这样的量级。场景中的几何细节，在有的应用中很重要，比如地板上 1 个 2 cm 高的积木，移动机器人应该能识别出来，从而在行进的过程中避开。如果能找到这样的有效、紧凑的表达，网络只需要预测更少的参数，有望提高网络的效率和泛化性能。

参 考 文 献

[1] Scharstein D, Szeliski R. A taxonomy and evaluation of dense two-frame stereo correspondence algorithms[J]. International Journal of Computer Vision, 2002, 47(1): 7-42.

[2] Se S, Jasiobedzki P. Stereo-vision based 3D modeling and localization for unmanned vehicles[J]. International Journal of Intelligent Control and Systems, 2008, 13(1): 47-58.

[3] 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, June 16-21, 2012, Providence, RI, USA. New York: IEEE Press, 2012: 3354-3361.

[4] Murray D, Little J J. Using real-time stereo vision for mobile robot navigation[J]. Autonomous Robots, 2000, 8(2): 161-171.

[5] Fan R, Jiao J H, Pan J, et al. Real-time dense stereo embedded in a UAV for road inspection[C]//2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), June 16-17, 2019, Long Beach, CA, USA. New York: IEEE Press, 2020: 535-543.

[6] Li R X, Squyres S W, Arvidson R E, et al. Initial results of rover localization and topographic mapping for the 2003 Mars exploration rover mission[J]. Photogrammetric Engineering & Remote Sensing, 2005, 71(10): 1129-1142.

[7] 王保丰, 周建亮, 唐歌实, 等. 嫦娥三号巡视器视觉定位方法[J]. 中国科学: 信息科学, 2014, 44(4): 452-460.
Wang B F, Zhou J L, Tang G S, et al. Research on visual localization method of lunar rover[J]. Scientia Sinica: Informationis, 2014, 44(4): 452-460.

[8] Luo C C, Li Y M, Lin K M, et al. Wavelet synthesis net for disparity estimation to synthesize DSLR calibre bokeh effect on smartphones[C]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 13-19, 2020, Seattle, WA, USA. New York: IEEE Press, 2020: 2404-2412.

[9] 朱代先. 基于双目视觉的工件定位与抓取研究[J]. 计算机测量与控制, 2011, 19(1): 92-94.
Zhu D X. Research on location and crawling of workpiece based on binocular vision[J]. Computer Measurement & Control, 2011, 19(1): 92-94.

[10] 龚健雅, 季顺平. 摄影测量与深度学习[J]. 测绘学报, 2018, 47(6): 693-704.
Gong J Y, Ji S P. Photogrammetry and deep learning[J]. Acta Geodaetica et Cartographica Sinica, 2018, 47(6): 693-704.

[11] 顾骋, 钱惟贤, 陈钱, 等. 基于双目立体视觉的快速人头检测方法[J]. 中国激光, 2014, 41(1): 0108001.
Gu C, Qian W X, Chen Q, et al. Rapid head detection

method based on binocular stereo vision[J]. Chinese Journal of Lasers, 2014, 41(1): 0108001.

[12] Bao W, Wang W, Xu Y H, et al. InStereo2K: a large real dataset for stereo matching in indoor scenes[J]. Science China Information Sciences, 2020, 63(11): 212101.

[13] Xu H F, Zhang J Y. AANet: adaptive aggregation network for efficient stereo matching[C]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 13-19, 2020, Seattle, WA, USA. New York: IEEE Press, 2020: 1956-1965.

[14] Teed Z, Deng J. RAFT: recurrent all-pairs field transforms for optical flow[M]//Vedaldi A, Bischof H, Brox T, et al. Computer vision-ECCV 2020. Lecture notes in computer science. Cham: Springer, 2020, 12347: 402-419.

[15] Lipson L, Teed Z, Deng J. RAFT-Stereo: multilevel recurrent field transforms for stereo matching[C]//2021 International Conference on 3D Vision (3DV), December 1-3, 2021, London, United Kingdom. New York: IEEE Press, 2022: 218-227.

[16] Li J K, Wang P S, Xiong P F, et al. Practical stereo matching via cascaded recurrent network with adaptive correlation[C]//2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 18-24, 2022, New Orleans, LA, USA. New York: IEEE Press, 2022: 16242-16251.

[17] Xu G W, Cheng J D, Guo P, et al. Attention concatenation volume for accurate and efficient stereo matching[C]//2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 18-24, 2022, New Orleans, LA, USA. New York: IEEE Press, 2022: 12971-12980.

[18] Shen Z L, Dai Y C, Rao Z B. CFNet: cascade and fused cost volume for robust stereo matching[C]//2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 20-25, 2021, Nashville, TN, USA. New York: IEEE Press, 2021: 13901-13910.

[19] Huang Z Y, Shi X Y, Zhang C, et al. FlowFormer: a transformer architecture for optical flow[EB/OL]. (2022-03-30)[2022-08-09]. https://arxiv.org/abs/2203.16194.

[20] Xu H F, Zhang J, Cai J F, et al. GMFlow: learning optical flow via global matching[C]//2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 18-24, 2022, New Orleans, LA, USA. New York: IEEE Press, 2022: 8111-8120.

[21] Menze M, Heipke C, Geiger A. Joint 3d estimation of vehicles and scene flow[J]. ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences, 2015, II-3/W5: 427-434.

[22] Scharstein D, Hirschmüller H, Kitajima Y, et al. High-resolution stereo datasets with subpixel-accurate ground truth[M]//Jiang X Y, Hornegger J, Koch R. Pattern recognition. Lecture notes in computer science. Cham: Springer, 2014, 8753: 31-42.

[23] Zhang F H, Prisacariu V, Yang R G, et al. GA-net: guided aggregation net for end-to-end stereo matching [C]//2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 15-20, 2019, Long Beach, CA, USA. New York: IEEE Press, 2020: 185-194.

[24] Chang J R, Chen Y S. Pyramid stereo matching network [C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, June 18-23, 2018, Salt Lake City, UT, USA. New York: IEEE Press, 2018: 5410-5418.

[25] Xu B, Xu Y H, Yang X L, et al. Bilateral grid learning for stereo matching networks[C]//2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 20-25, 2021, Nashville, TN, USA. New York: IEEE Press, 2021: 12492-12501.

[26] 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.

[27] Mei X, Sun X, Zhou M C, et al. On building an accurate stereo matching system on graphics hardware[C]//2011 IEEE International Conference on Computer Vision Workshops (ICCV Workshops), November 6-13, 2011, Barcelona, Spain. New York: IEEE Press, 2012: 467-474.

[28] Hirschmüller H. Stereo processing by semiglobal matching and mutual information[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2008, 30(2): 328-341.

[29] Hirschmuller H, Scharstein D. Evaluation of cost functions for stereo matching[C]//2007 IEEE Conference on Computer Vision and Pattern Recognition, June 17-22, 2007, Minneapolis, MN, USA. New York: IEEE Press, 2007.

[30] 白明, 庄严, 王伟. 双目立体匹配算法的研究与进展[J]. 控制与决策, 2008, 23(7): 721-729.
Bai M, Zhuang Y, Wang W. Progress in binocular stereo matching algorithms[J]. Control and Decision, 2008, 23(7): 721-729.

[31] Hong L, Chen G. Segment-based stereo matching using graph cuts[C]//Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2004. CVPR, June 27-July 2, 2004, Washington, DC, USA. New York: IEEE Press, 2004.

[32] 张令涛, 曲道奎, 徐方. 一种基于图割的改进立体匹配算法[J]. 机器人, 2010, 32(1): 104-108.
Zhang L T, Qu D K, Xu F. An improved stereo matching algorithm based on graph cuts[J]. Robot, 2010, 32(1): 104-108.

[33] Sun J, Zheng N N, Shum H Y. Stereo matching using belief propagation[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2003, 25(7): 787-800.

[34] Yang Q X, Wang L, Yang R G, et al. Stereo matching with color-weighted correlation, hierarchical belief propagation, and occlusion handling[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2009, 31(3): 492-504.

[35] Boykov Y, Veksler O, Zabih R. Fast approximate

energy minimization via graph cuts[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2001, 23(11): 1222-1239.

[36] Yoon K J, Kweon I S. Adaptive support-weight approach for correspondence search[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2006, 28(4): 650-656.

[37] Tomasi C, Manduchi R. Bilateral filtering for gray and color images[C]//Sixth International Conference on Computer Vision (IEEE Cat. No. 98CH36271), January 7, 1998, Bombay, India. New York: IEEE Press, 2002: 839-846.

[38] Wang L, Liao M, Gong M L, et al. High-quality real-time stereo using adaptive cost aggregation and dynamic programming[C]//Third International Symposium on 3D Data Processing, Visualization, and Transmission (3DPVT'06), June 14-16, 2006, Chapel Hill, NC, USA. New York: IEEE Press, 2007: 798-805.

[39] He K M, Sun J, Tang X O. Guided image filtering[M]// Daniilidis K, Maragos P, Paragios N. Computer vision-ECCV 2010. Lecture notes in computer science. Heidelberg: Springer, 2010, 6311: 1-14.

[40] 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.

[41] Zhang K, Lu J B, Lafruit G. Cross-based local stereo matching using orthogonal integral images[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2009, 19(7): 1073-1079.

[42] Keselman L, Woodfill J I, Grunnet-Jepsen A, et al. Intel (R) RealSense(TM) stereoscopic depth cameras[C]// 2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), July 21-26, 2017, Honolulu, HI, USA. New York: IEEE Press, 2017: 1267-1276.

[43] Bleyer M, Rhemann C, Rother C. PatchMatch stereo-stereo matching with slanted support windows[C]// Proceedings of the British Machine Vision Conference 2011, August 29-September 2, 2011, Dundee, UK. London: British Machine Vision Association, 2011: 1-11.

[44] Barnes C, Shechtman E, Finkelstein A, et al. PatchMatch: a randomized correspondence algorithm for structural image editing[J]. ACM Transactions on Graphics, 2009, 28(3): 1-11.

[45] Yang Q X. A non-local cost aggregation method for stereo matching[C]//2012 IEEE Conference on Computer Vision and Pattern Recognition, June 16-21, 2012, Providence, RI, USA. New York: IEEE Press, 2012: 1402-1409.

[46] Yang Q X. Stereo matching using tree filtering[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2015, 37(4): 834-846.

[47] Li L C, Yu X, Zhang S L, et al. 3D cost aggregation with multiple minimum spanning trees for stereo matching [J]. Applied Optics, 2017, 56(12): 3411-3420.

[48] Zbontar J, LeCun Y. Stereo matching by training a convolutional neural network to compare image patches [J]. Journal of Machine Learning Research, 2016, 17(1): 2287-2318.

[49] Zagoruyko S, Komodakis N. Learning to compare image patches via convolutional neural networks[C]//2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 7-12, 2015, Boston, MA, USA. New York: IEEE Press, 2015: 4353-4361.

[50] Park H, Lee K M. Look wider to match image patches with convolutional neural networks[J]. IEEE Signal Processing Letters, 2017, 24(12): 1788-1792.

[51] Shaked A, Wolf L. Improved stereo matching with constant highway networks and reflective confidence learning[C]//2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), July 21-26, 2017, Honolulu, HI, USA. New York: IEEE Press, 2017: 6901-6910.

[52] Seki A, Pollefeys M. SGM-nets: semi-global matching with neural networks[C]//2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), July 21-26, 2017, Honolulu, HI, USA. New York: IEEE Press, 2017: 6640-6649.

[53] Knöbelreiter P, Reinbacher C, Shekhovtsov A, et al. End-to-end training of hybrid CNN-CRF models for stereo[C]//2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), July 21-26, 2017, Honolulu, HI, USA. New York: IEEE Press, 2017: 1456-1465.

[54] Gidaris S, DetectKomodakis N., replace, refine: deep structured prediction for pixel wise labeling[C]//2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), July 21-26, 2017, Honolulu, HI, USA. New York: IEEE Press, 2017: 7187-7196.

[55] Güney F, Geiger A. Displets: Resolving stereo ambiguities using object knowledge[C]//2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 7-12, 2015, Boston, MA, USA. New York: IEEE Press, 2015: 4165-4175.

[56] Mayer N, Ilg E, Häusser P, et al. A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation[C]//2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 27-30, 2016, Las Vegas, NV, USA. New York: IEEE Press, 2016: 4040-4048.

[57] Vaswani A, Shazeer N, Parmar N, et al. Attention is all You need[C]//Proceedings of the 31st International Conference on Neural Information Processing Systems, December 4-9, 2017, Long Beach, California, USA. New York: ACM Press, 2017: 6000-6010.

[58] Wang X L, Girshick R, Gupta A, et al. Non-local neural networks[C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, June 18-23, 2018, Salt Lake City, UT, USA. New York: IEEE Press, 2018: 7794-7803.

[59] Dosovitskiy A, Beyer L, Kolesnikov A, et al. An image is worth 16x16 words: transformers for image recognition at scale[EB/OL]. (2020-10-22) [2022-11-12]. https://arxiv.org/abs/2010.11929.

[60] Liu Z, Lin Y T, Cao Y, et al. Swin transformer: hierarchical vision transformer using shifted windows[C]// 2021 IEEE/CVF International Conference on Computer Vision (ICCV), October 10-17, 2021, Montreal, QC, Canada. New York: IEEE Press, 2022: 9992-10002.

[61] Dosovitskiy A, Fischer P, Ilg E, et al. FlowNet: learning optical flow with convolutional networks[C]// 2015 IEEE International Conference on Computer Vision (ICCV), December 7-13, 2015, Santiago, Chile. New York: IEEE Press, 2016: 2758-2766.

[62] Liang Z F, Feng Y L, Guo Y L, et al. Learning for disparity estimation through feature constancy[C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, June 18-23, 2018, Salt Lake City, UT, USA. New York: IEEE Press, 2018: 2811-2820.

[63] Pang J H, Sun W X, Ren J S, et al. Cascade residual learning: a two-stage convolutional neural network for stereo matching[C]//2017 IEEE International Conference on Computer Vision Workshops (ICCVW), October 22-29, 2017, Venice, Italy. New York: IEEE Press, 2018: 878-886.

[64] Song X, Zhao X, Fang L J, et al. EdgeStereo: an effective multi-task learning network for stereo matching and edge detection[J]. International Journal of Computer Vision, 2020, 128(4): 910-930.

[65] Yang G R, Zhao H S, Shi J P, et al. SegStereo: exploiting semantic information for disparity estimation [M]//Ferrari V, Hebert M, Sminchisescu C, et al. Computer vision-ECCV 2018. Lecture notes in computer science. Cham: Springer, 2018, 11211: 660-676.

[66] Dai J F, Qi H Z, Xiong Y W, et al. Deformable convolutional networks[C]//2017 IEEE International Conference on Computer Vision (ICCV), October 22-29, 2017, Venice, Italy. New York: IEEE Press, 2017: 764-773.

[67] Tankovich V, Häne C, Zhang Y D, et al. HITNet: hierarchical iterative tile refinement network for real-time stereo matching[C]//2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 20-25, 2021, Nashville, TN, USA. New York: IEEE Press, 2021: 14357-14367.

[68] Badki A, Troccoli A, Kim K, et al. Bi3D: stereo depth estimation via binary classifications[C]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 13-19, 2020, Seattle, WA, USA. New York: IEEE Press, 2020: 1597-1605.

[69] Tosi F, Liao Y Y, Schmitt C, et al. SMD-Nets: stereo mixture density networks[C]//2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 20-25, 2021, Nashville, TN, USA. New York: IEEE Press, 2021: 8938-8948.

[70] Kendall A, Martirosyan H, Dasgupta S, et al. End-to-end learning of geometry and context for deep stereo regression[C]//2017 IEEE International Conference on Computer Vision (ICCV), October 22-29, 2017, Venice, Italy. New York: IEEE Press, 2017: 66-75.

[71] Guo X Y, Yang K, Yang W K, et al. Group-wise correlation stereo network[C]//2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 15-20, 2019, Long Beach, CA, USA. New York: IEEE Press, 2020: 3268-3277.

[72] Khamis S, Fanello S, Rhemann C, et al. StereoNet: guided hierarchical refinement for real-time edge-aware depth prediction[M]//Ferrari V, Hebert M, Sminchisescu C, et al. Computer vision-ECCV 2018. Lecture notes in computer science. Cham: Springer, 2018, 11219: 596-613.

[73] Duggal S, Wang S L, Ma W C, et al. DeepPruner: learning efficient stereo matching via differentiable PatchMatch[C]//2019 IEEE/CVF International Conference on Computer Vision (ICCV), October 27-November 2, 2019, Seoul, Korea (South). New York: IEEE Press, 2020: 4383-4392.

[74] Zhang F H, Qi X J, Yang R G, et al. Domain-invariant stereo matching networks[M]//Vedaldi A, Bischof H, Brox T, et al. Computer vision-ECCV 2020. Lecture notes in computer science. Cham: Springer, 2020, 12347: 420-439.

[75] Yao C T, Jia Y D, Di H J, et al. A decomposition model for stereo matching[C]//2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 20-25, 2021, Nashville, TN, USA. New York: IEEE Press, 2021: 6087-6096.

[76] He K M, Zhang X Y, Ren S Q, et al. Deep residual learning for image recognition[C]//2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 27-30, 2016, Las Vegas, NV, USA. New York: IEEE Press, 2016: 770-778.

[77] Li Z S, Liu X T, Drenkow N, et al. Revisiting stereo depth estimation from a sequence-to-sequence perspective with transformers[C]//2021 IEEE/CVF International Conference on Computer Vision (ICCV), October 10-17, 2021, Montreal, QC, Canada. New York: IEEE Press, 2022: 6177-6186.

[78] Zhang Y M, Chen Y M, Bai X, et al. Adaptive unimodal cost volume filtering for deep stereo matching [J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2020, 34(7): 12926-12934.

[79] Yu J J, Harley A W, Derpanis K G. Back to basics: unsupervised learning of optical flow via brightness constancy and motion smoothness[M]//Hua G, Jégou H. Computer vision-ECCV 2016 workshops. Lecture notes in computer science. Cham: Springer, 2016, 9915: 3-10.

[80] Ahmadi A, Patras I. Unsupervised convolutional neural networks for motion estimation[C]//2016 IEEE International Conference on Image Processing (ICIP), September 25-28, 2016, Phoenix, AZ, USA. New York: IEEE Press, 2016: 1629-1633.

[81] Flynn J, Neulander I, Philbin J, et al. Deep Stereo: learning to predict new views from the world's imagery [C]//2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 27-30, 2016, Las Vegas, NV, USA. New York: IEEE Press, 2016: 5515-5524.

[82] Xie J Y, Girshick R, Farhadi A. Deep3D: fully automatic 2D-to-3D video conversion with deep convolutional neural networks[M]//Leibe B, Matas J, Sebe N, et al. Computer vision-ECCV 2016. Lecture notes in computer science. Cham: Springer, 2016, 9908: 842-857.

[83] Garg R, B G V K, Carneiro G, et al. Unsupervised CNN for single view depth estimation: geometry to the rescue[M]//Leibe B, Matas J, Sebe N, et al. Computer vision-ECCV 2016. Lecture notes in computer science. Cham: Springer, 2016, 9912: 740-756.

[84] Ren Z, Yan J C, Ni B B, et al. Unsupervised deep learning for optical flow estimation[C]//Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, February 4-9, 2017, San Francisco, California, USA. New York: ACM Press, 2017: 1495-1501.

[85] Godard C, Mac Aodha O, Brostow G J. Unsupervised monocular depth estimation with left-right consistency [C]//2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), July 21-26, 2017, Honolulu, HI, USA. New York: IEEE Press, 2017: 6602-6611.

[86] Tonioni A, Poggi M, Mattoccia S, et al. Unsupervised adaptation for deep stereo[C]//2017 IEEE International Conference on Computer Vision (ICCV), October 22-29, 2017, Venice, Italy. New York: IEEE Press, 2017: 1614-1622.

[87] Zabih R, Woodfill J. Non-parametric local transforms for computing visual correspondence[M]//Eklundh J O. Computer vision-ECCV '94. Lecture notes in computer science. Heidelberg: Springer, 1994, 801: 151-158.

[88] Poggi M, Mattoccia S. Learning from scratch a confidence measure[C]//Proceedings of the British Machine Vision Conference 2016, September 19-22, 2016, York, UK. London: British Machine Vision Association, 2016: 1-13.

[89] Kuznietsov Y, Stückler J, Leibe B. Semi-supervised deep learning for monocular depth map prediction[C]// 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), July 21-26, 2017, Honolulu, HI, USA. New York: IEEE Press, 2017: 2215-2223.

[90] Zhou C, Zhang H, Shen X Y, et al. Unsupervised learning of stereo matching[C]//2017 IEEE International Conference on Computer Vision (ICCV), October 22-29, 2017, Venice, Italy. New York: IEEE Press, 2017: 1576-1584.

[91] Aleotti F, Tosi F, Zhang L, et al. Reversing the cycle: self-supervised deep stereo through enhanced monocular distillation[M]//Vedaldi A, Bischof H, Brox T, et al. Computer vision-ECCV 2020. Lecture notes in computer science. Cham: Springer, 2020, 12356: 614-632.

[92] Wang Y, Lai Z H, Huang G, et al. Anytime stereo image depth estimation on mobile devices[C]//2019 International Conference on Robotics and Automation (ICRA), May 20-24, 2019, Montreal, QC, Canada. New York: IEEE Press, 2019: 5893-5900.

[93] Yee K, Chakrabarti A. Fast deep stereo with 2D convolutional processing of cost signatures[C]//2020 IEEE Winter Conference on Applications of Computer Vision (WACV), March 1-5, 2020, Snowmass, CO, USA. New York: IEEE Press, 2020: 183-191.

[94] Shamsafar F, Woerz S, Rahim R, et al. MobileStereoNet: towards lightweight deep networks for stereo matching[C]//2022 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), January 3-8, 2022, Waikoloa, HI, USA. New York: IEEE Press, 2022: 677-686.

[95] Howard A G, Zhu M L, Chen B, et al. MobileNets: efficient convolutional neural networks for mobile vision applications[EB/OL]. (2017-04-17)[2022-10-05]. https://arxiv.org/abs/1704.04861.

[96] Sandler M, Howard A, Zhu M L, et al. MobileNetV2: inverted residuals and linear bottlenecks[C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, June 18-23, 2018, Salt Lake City, UT, USA. New York: IEEE Press, 2018: 4510-4520.

[97] LeCun Y, Denker J, Solla S. Optimal brain damage[M]. San Francisco: Morgan Kaufmann, 1989.

[98] Hassibi B, Stork D G, Wolff G, et al. Optimal brain surgeon: extensions and performance comparisons[C]// Proceedings of the 6th International Conference on Neural Information Processing Systems, November 29, 1993, Denver, Colorado. New York: ACM Press, 1993: 263-270.

[99] Han S, Mao H Z, Dally W J. Deep compression: compressing deep neural networks with pruning, trained quantization and Huffman coding[EB/OL]. (2015-10-01) [2022-11-05]. https://arxiv.org/abs/1510.00149.

[100] Han S, Liu X Y, Mao H Z, et al. EIE: efficient inference engine on compressed deep neural network[C]// 2016 ACM/IEEE 43rd Annual International Symposium on Computer Architecture (ISCA), June 18-22, 2016, Seoul, Republic of Korea. New York: IEEE Press, 2016: 243-254.

[101] Anwar S, Sung W. Coarse pruning of convolutional neural networks with random masks[M]. Amsterdam: Elsevier, 2017.

[102] Li H, Kadav A, Durdanovic I, et al. Pruning filters for efficient ConvNets[EB/OL]. (2016-08-31) [2022-10-08]. https://arxiv.org/abs/1608.08710.

[103] Liu B Y, Wang M, Foroosh H, et al. Sparse convolutional neural networks[C]//2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 7-12, 2015, Boston, MA. New York: IEEE Press, 2015: 806-814.

[104] Liu Z, Li J G, Shen Z Q, et al. Learning efficient convolutional networks through network slimming[C]// 2017 IEEE International Conference on Computer Vision (ICCV), October 22-29, 2017, Venice, Italy. New York: IEEE Press, 2017: 2755-2763.

[105] Hu H Y, Peng R, Tai Y W, et al. Network trimming: a data-driven neuron pruning approach towards efficient deep architectures[EB/OL]. (2016-07-12) [2022-10-09].

https://arxiv.org/abs/1607.03250.

[106] Ye J B, Lu X, Lin Z, et al. Rethinking the smaller-norm-less-informative assumption in channel pruning of convolution layers[EB/OL]. (2018-02-01) [2022-10-09]. https://arxiv.org/abs/1802.00124.

[107] Luo J H, Wu J X, Lin W Y. ThiNet: a filter level pruning method for deep neural network compression [C]//2017 IEEE International Conference on Computer Vision (ICCV), October 22-29, 2017, Venice, Italy. New York: IEEE Press, 2017: 5068-5076.

[108] Yu R C, Li A, Chen C F, et al. NISP: pruning networks using neuron importance score propagation[C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, June 18-23, 2018, Salt Lake City, UT, USA. New York: IEEE Press, 2018: 9194-9203.

[109] Liu Z, Sun M J, Zhou T H, et al. Rethinking the value of network pruning[EB/OL]. (2018-10-11)[2022-08-09]. https://arxiv.org/abs/1810.05270.

[110] Tan M X, Le Q V. EfficientNet: rethinking model scaling for convolutional neural networks[EB/OL]. (2019-05-28)[2022-10-09]. https://arxiv.org/abs/1905.11946.

[111] He Y H, Lin J, Liu Z J, et al. AMC: AutoML for model compression and acceleration on mobile devices[M]// Ferrari V, Hebert M, Sminchisescu C, et al. Computer Vision-ECCV 2018. Lecture notes in computer science. Cham: Springer, 2018, 11211: 815-832.

[112] Yang T J, Howard A, Chen B, et al. NetAdapt: platform-aware neural network adaptation for mobile applications[M]//Ferrari V, Hebert M, Sminchisescu C, et al. Computer vision-ECCV 2018. Cham: Springer, 2018, 11214: 289-304.

[113] Zhu M, Gupta S. To prune, or not to prune: exploring the efficacy of pruning for model compression[EB/OL]. (2017-10-05)[2022-10-09]. https://arxiv.org/abs/1710.01878.

[114] Jacob B, Kligys S, Chen B, et al. Quantization and training of neural networks for efficient integer-arithmetic-only inference[C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, June 18-23, 2018, Salt Lake City, UT, USA. New York: IEEE Press, 2018: 2704-2713.

[115] Ron B, Yury N, Elad H. Post training 4-bit quantization of convolution networks for rapid-deployment[C]// Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada. Canada: NIPS, 2019: 7948-7956.

[116] Migacz S. 8-bit inference with TensorRT[EB/OL]. (2017-05-08)[2022-10-09]. https://on-demand.gputechconf.com/gtc/2017/presentation/s7310-8-bit-inference-with-tensorrt.pdf

[117] Gong J, Shen H H, Zhang G M, et al. Highly efficient 8-bit low precision inference of convolutional neural networks with IntelCaffe[C]//Proceedings of the 1st on Reproducible Quality-Efficient Systems Tournament on Co-designing Pareto-efficient Deep Learning, April 24, 2018, Williamsburg, VA, USA. New York: ACM Press, 2018: 1-2

[118] Lin X F, Zhao C, Pan W. Towards accurate binary convolutional neural network[C]//Proceedings of the 31st International Conference on Neural Information Processing Systems, December 4-9, 2017, Long Beach, California, USA. New York: ACM, 2017: 344-352.

[119] Zhao R, Hu Y W, Dotzel J, et al. Improving neural network quantization without retraining using outlier channel splitting[EB/OL]. (2019-01-28) [2022-10-09]. https://arxiv.org/abs/1901.09504.

[120] Chmiel B, Banner R, Shomron G, et al. Robust quantization: one model to rule them all[C]//Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual. Canada: NIPS, 2020: 5308-5317.

[121] Courbariaux M, Hubara I, Soudry D, et al. Binarized neural networks: training deep neural networks with weights and activations constrained to +1 or-1[EB/OL]. (2016-02-09)[2022-10-09]. https://arxiv.org/abs/1602.02830.

[122] Bengio Y, Léonard N, Courville A. Estimating or propagating gradients through stochastic neurons for conditional computation[EB/OL]. (2013-08-15)[2022-10-08]. https://arxiv.org/abs/1308.3432.

[123] Cao Z J, Long M S, Wang J M, et al. HashNet: deep learning to hash by continuation[C]//2017 IEEE International Conference on Computer Vision (ICCV), October 22-29, 2017, Venice, Italy. New York: IEEE Press, 2017: 5609-5618.

[124] Hwang K, Sung W. Fixed-point feedforward deep neural network design using weights 1, 0, and −1[C]//2014 IEEE Workshop on Signal Processing Systems (SiPS), October 20-22, 2014, Belfast, UK. New York: IEEE Press, 2014.

[125] Zhou S C, Wu Y X, Ni Z K, et al. DoReFa-net: training low bitwidth convolutional neural networks with low bitwidth gradients[EB/OL]. (2016-06-20) [2022-10-08]. https://arxiv.org/abs/1606.06160.

[126] Zhou A J, Yao A B, Guo Y W, et al. Incremental network quantization: towards lossless CNNs with low-precision weights[EB/OL]. (2017-02-10) [2022-10-08]. https://arxiv.org/abs/1702.03044.

[127] Zhang S J, Du Z D, Zhang L, et al. Cambricon-X: an accelerator for sparse neural networks[C]//2016 49th Annual IEEE/ACM International Symposium on Microarchitecture (MICRO), October 15-19, 2016, Taipei, China. New York: IEEE Press, 2016.

[128] Chen T S, Du Z D, Sun N H, et al. DianNao: a small-footprint high-throughput accelerator for ubiquitous machine-learning[J]. ACM SIGARCH Computer Architecture News, 2014, 42(1): 269-284.

[129] Scharstein D, Szeliski R. High-accuracy stereo depth maps using structured light[C]//2003 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2003. Proceedings, June 18-20, 2003, Madison, WI, USA. New York: IEEE Press, 2003.

[130] Ramirez P Z, Tosi F, Poggi M, et al. Open challenges in deep stereo: the booster dataset[C]//2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition

(CVPR), June 18-24, 2022, New Orleans, LA, USA. New York: IEEE Press, 2022: 21136-21146.

[131] Yang G R, Song X, Huang C Q, et al. DrivingStereo: a large-scale dataset for stereo matching in autonomous driving scenarios[C]//2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 15-20, 2019, Long Beach, CA, USA. New York: IEEE Press, 2020: 899-908.

[132] Wang Q, Zheng S Z, Yan Q S, et al. IRS: a large naturalistic indoor robotics stereo dataset to train deep models for disparity and surface normal estimation[EB/OL]. (2019-12-20) [2022-10-09]. https://arxiv. org/abs/1912.09678.

[133] Schöps T, Schönberger J L, Galliani S, et al. A multi-view stereo benchmark with high-resolution images and multi-camera videos[C]//2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), July 21-26, 2017, Honolulu, HI, USA. New York: IEEE Press, 2017: 2538-2547.

[134] Tremblay J, To T, Birchfield S. Falling things: a synthetic dataset for 3D object detection and pose estimation[C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), June 18-22, 2018, Salt Lake City, UT, USA. New York: IEEE Press, 2018: 2119-21193.

[135] Yang G S, Manela J, Happold M, et al. Hierarchical deep stereo matching on high-resolution images[C]//2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 15-20, 2019, Long Beach, CA, USA. New York: IEEE Press, 2020: 5510-5519.

[136] Huang X Y, Cheng X J, Geng Q C, et al. The ApolloScape dataset for autonomous driving[C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), June 18-22, 2018, Salt Lake City, UT, USA. New York: IEEE Press, 2018: 1067-10676.

[137] ORBBEC[EB/OL]. [2022-10-08]. http://www.orbbec.com.cn/.

[138] RUBEDOS[EB/OL]. [2022-10-08]. https://www.rubedos.com/.

[139] Blaxtair[EB/OL]. [2022-10-08]. https://blaxtair.com/.

[140] Stereolabs[EB/OL]. [2022-10-08]. https://www.stereolabs.com/.

[141] Human+ [EB/OL]. [2022-10-08]. http://humanplustech.com/.

[142] OAKChina [EB/OL]. [2022-10-08]. https://www.oakchina.cn/.

[143] Intel[EB/OL]. [2022-10-08]. https://www.intel.com/content/www/us/en/products/details/processors/movidius-vpu.html.

[144] NVIDIA [EB/OL]. [2022-10-08]. https://www. nvidia.com/en-us/autonomous-machines/embedded-systems/.

[145] Duo3D[EB/OL]. [2022-10-08]. https://www.duo3d.com/.

[146] CARNEGIE ROBOTICS [EB/OL]. [2022-10-08]. https://www.carnegierobotics.com/.

[147] PERCIPIO. XYZ[EB/OL]. [2022-10-08]. https://www.percipio.xyz/.