计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2023,32(8):19−30 [doi: 10.15888/j.cnki.csa.009233]
©中国科学院软件研究所版权所有.
E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

以双目图像互为监督的相容解迭代优选算法①

翟记锋

(复旦大学 计算机科学技术学院, 上海 200438)
通信作者: 翟记锋, E-mail: 20210240027@fudan.edu.cn

摘　要: 在计算机视觉领域的双目立体匹配方向, 基于神经网络的深度学习算法需要场景数据集进行训练, 泛化能力差. 针对这两个问题, 根据神经网络能够模拟函数的特点, 提出一种无需在数据集上训练, 以双目图像互为监督的深度场景相容解迭代优选算法. 该算法使用场景位置猜测网络模拟关于当前双目图像的深度场景相容位置空间, 用与该网络匹配的互监督损失函数通过梯度下降法指导该网络在输入双目图像上迭代学习, 搜索深度场景相容位置空间中的可行解, 整个算法过程无需在数据集上训练. 与 CREStereo、PCW-Net、CFNet 等算法在 Middlebury 标准数据集图像上的对比实验表明, 该算法在非遮挡区域的平均误匹配率为 2.52%, 在所有区域的平均误匹配率为 7.26%, 比对比实验中的其他算法有更低的平均误匹配率.

关键词: 双目立体匹配; 立体匹配; 神经网络; 无监督学习; 自监督学习

引用格式: 翟记锋.以双目图像互为监督的相容解迭代优选算法.计算机系统应用,2023,32(8):19–30. http://www.c-s-a.org.cn/1003-3254/9233.html

Iterative Optimization Algorithm for Compatible Solution with Binocular Images as Mutual Supervision

ZHAI Ji-Feng

(School of Computer Science, Fudan University, Shanghai 200438, China)

Abstract: In the direction of binocular stereo matching in computer vision, deep learning algorithms based on neural networks require scene datasets for training and have poor generalization ability. In order to address these two problems, an iterative optimization algorithm of compatible solutions of deep scenes is proposed based on the ability of neural networks to simulate functions, and the algorithm requires no training on a dataset, with binocular images supervised by each other. The algorithm uses a scene location guessing network to simulate the compatible location space of a deep scene about the current binocular image, and a mutually supervised loss function matched with this network is used to guide the network to iteratively learn on the input binocular image by gradient descent. In addition, the feasible solution in the compatible location space of the deep scene is searched, and the whole process does not require training on the dataset. Comparison experiments with CREStereo, PCW-Net, CFNet, and other algorithms on Middlebury standard dataset images show that this algorithm has an average mismatching rate of 2.52% in non-occluded regions and 7.26% in all regions, which is lower than that of the other algorithms in the comparison experiments.

Key words: binocular stereo matching; stereo matching; neural networks; unsupervised learning; self-supervised learning

双目立体视觉一直是计算机视觉领域热门研究方向之一, 主要目标是根据双目立体相机拍摄的一对二维图像, 还原出三维场景的位置信息. 双目立体相机是模拟人眼的产物, 该相机由两个单目相机组成, 水平相隔一定距离, 同时各自成像, 形成一对图像. 双目立体匹配是双目立体视觉的关键一步, 目标是求解双目图

① 基金项目: 国家自然科学基金 (61771146)
收稿时间: 2022-12-24; 修改时间: 2023-02-09, 2023-04-07; 采用时间: 2023-04-13; csa 在线出版时间: 2023-06-30
CNKI 网络首发时间: 2023-07-03

© 中国科学院软件研究所 http://www.c-s-a.org.cn

像对应三维场景的深度信息, 用视差图表示. 有了视差图和这一对图像, 再联合双目立体相机参数就能够还原出真实的三维场景位置信息. 现实中, 像无人驾驶、移动机器人、虚拟现实、增强现实、星球探测、辅助视觉等[1–3] 都需要这种三维感知的能力, 因此双目立体匹配的应用十分广泛.

经过严格的数学推导, 求解双目立体匹配问题转化为求解一对双目图像之间像素点的匹配问题, 其深度可用该对匹配点坐标之差——视差表示. 双目立体匹配的难点在于双目图像中遮挡、弱纹理、高光、噪音、透射、重复图案等病态区域的存在, 导致求解过程中产生大量的多解. 为了解决这些难点, 大量的研究者提出了各种各样的双目立体匹配算法. 双目立体匹配算法大体上分为两种, 传统算法和深度学习算法. 传统算法是指无需通过数据集学习, 不使用神经网络, 直接计算出该对图像深度的算法. 深度学习算法指基于深层的神经网络, 一般是卷积神经网络, 在大量场景数据集下自主学习的算法.

传统算法一般由 4 个主要步骤[4]——代价计算、代价聚集、视差选择、视差优化组成. 传统算法的输入是双目立体相机拍摄的, 经过对极几何校正后的双目图像对. 双目图像对经过对极几何校正后, 左右图像的匹配点均在同一行, 故可极大减少像素匹配搜索空间. 代价计算是计算同行任意两个左右像素的差异, 一般为亮度差异, 用这种差异衡量两个像素匹配的程度, 如绝对亮度差 (absolute differences, AD[4])、平方亮度差 (squared differences, SD[5])、加和绝对亮度差 (sum of absolute differences, SAD[6])、加和平方亮度差 (sum of squared differences, SSD[7])、归一化相关系数 (normalized cross correlation, NCC[8])、级别变换 (rank transform, RT[9])、统计变换 (census transform, CT[10]) 等. 代价聚集是使用一定的方法, 一般是基于窗口的方法, 将代价计算的像素代价聚集起来, 利用邻居信息, 以减少匹配的不确定性. 常用的聚集方法有固定窗口 (fixed-size window, FW[11])、多窗口 (multiple windows, MW[12])、自适应窗口 (adaptive windows, AW[13])、自适应支撑窗口 (adaptive support weights, ASW[14]) 等. 视差选择是根据代价聚集的结果选择聚集代价最小的视差, 即用“赢者通吃”的方法. 视差优化是为了补救视差选择步骤部分选择错误的视差, 方法通常有左右一致性校验、填充和插值、亚像素拟合、中值滤波、双边滤波等.

双目立体匹配的深度学习算法通常分为有监督的学习方法和无监督的学习方法. 有监督的学习方法需要视差真值图. 视差真值图一般通过如激光、雷达等其他物理手段直接测距获得. 有监督的深度学习方法需要在大量的标注数据集上长时间的训练, 消耗的计算资源大, 且标注数据集的获取较困难, 虽然在和训练集场景相似的场景下表现较好, 但在其他场景下表现较差, 即泛化能力差. 无监督的学习方法不需要视差真值图, 而是利用左右图像蕴涵关系指导训练. 与有监督的学习方法相比, 无监督学习方法的最大优点是无需昂贵的标注数据集, 缺点是收敛效果差些.

传统算法已经被无数学者探索了几十年, 最近几年进展缓慢, 而深度学习算法发展迅速, 然而, 深度学习算法的发展是以算力提升为基础的, 并不改变需要数据集训练的现象且泛化能力差. 考虑到左右像是同一场景经两次确定性投影生成的, 由左右像逆推可以找到至少一个关于深度场景位置信息的相容解, 故而左右像天然可以互为导师数据 (由左像变换到右像, 由右像变换到左像). 基于此, 本文尝试从单个双目图像中学习其专有的匹配规则, 提出了一个以双目图像互为监督的深度场景相容解迭代优选算法. 该算法不同于传统算法和深度学习算法, 使用神经网络但不使用数据集训练, 用互监督损失函数通过梯度下降法指导网络收敛, 只在输入图像对上迭代学习, 优化相容解, 进而计算出双目图像的深度.

本文的主要贡献包括 3 个方面.

(1) 提出一个以双目图像互为监督的深度场景相容解迭代优选算法, 该算法不需要数据集训练, 只需在当前输入图像对上迭代学习, 故而摆脱了对数据集的依赖. 对比实验表明, 该算法有更低的误匹配率.

(2) 设计了场景位置猜测网络, 该网络避免了昂贵的 3D 卷积, 同时输出左右视差图, 进而能够使用视差优化技术, 提高匹配精度.

(3) 设计了互监督损失函数, 用于指导场景位置猜测网络的收敛, 该损失函数无需视差真值图.

1 相关工作

传统算法通常又可分为局部匹配算法和全局匹配算法[4]. 局部匹配算法依据的是图像局部区域的特征, 具有计算复杂度低、运算速度快的特点, 典型的方法

© 中国科学院软件研究所 http://www.c-s-a.org.cn

有 SAD[6]、SSD[7]、CT[10] 等. SAD 以窗口绝对亮度差之和计算代价量. SSD 是以窗口平方亮度差之和计算代价量. CT 是把窗口中心像素和邻居像素比较, 比较结果用二进制字符串表示, 以海明码距离计算代价量. 全局匹配算法依据的是图像全局的特征, 转化为求解能量函数最小化问题, 效果好, 但计算复杂度高, 典型方法有信任传播 (belief propagation, BP[15])、图割 (graph cut, GC[16]) 等. BP 方法把立体匹配问题建模为 3 个成对的马尔科夫随机场. GC 方法是用图割的方法转化为求解最小割问题. 还有个别方法不属于这两类, 如半全局匹配 (semi-global matching, SGM[17])、PatchMatch[18] 等. SGM 使用扫描线的方法沿直线方向聚集匹配代价. PatchMatch 把场景建模为由一个个三维平面拼接而成, 转化为求解三维标签问题. 总体上看, 传统算法的优点是无需数据集训练, 场景依赖性低, 占用的计算资源小, 缺点是求得的视差精度低.

在有监督的学习方法中, 除少数方法外, 如 MC-CNN[19], 其余都是端到端的学习方法, 即神经网络的输入是一对双目图像, 输出是目标视差图, 如 GC-Net[20]、PSMNet[21]、GA-Net[22]、HITNet[23]、CREStereo[24]、RAFT-Stereo[25]、EAI-Stereo[26]. MC-CNN 方法是把传统算法 4 个主要步骤中的代价计算由孪生网络替代, 其余步骤不变, 首先学习左右图像块的相似度, 用相似度计算代价量, 其效果仍然受限于传统的代价聚集方法. GC-Net 方法是把传统方式中的每个步骤都用神经网络替代, 是一个端到端的方法, 使用 3D 卷积对代价量进行视差回归. PSMNet 方法采用多尺度的神经网络架构, 首先用沙漏模型提取图像多尺度特征, 计算不同尺度的代价量, 再用 3D 卷积在不同尺度上进行视差回归. GA-Net 方法设计了两个特殊的网络层, 对代价量分别进行局部和全局的聚集, 不仅综合了全局和局部特征, 还避免了昂贵的 3D 卷积操作. HITNet 以多清晰度和倾斜的平面块建模, 避免显式构建代价量, 达到实时预测的程度. CREStereo 引入了级联的循环神经网络, 具有清晰的边缘且能够处理非理想校正的图片. RAFT-Stereo 是在光流网络 (RAFT[27]) 的基础上进行多级卷积门控循环单元 (gated recurrent unit, GRU) 改造, 能够迭代细化视差. 有监督的深度学习方法之间的不同之处在于网络架构和损失函数, 相同之处是都需要在大量的标注数据集上长时间的训练, 消耗的计算资源大, 且标注数据集的获取较困难, 虽然在和训练集场景相似的场景下表现较好, 但在其他场景下表现较差, 即泛化能力差.

无监督的学习方法相对于有监督的学习方法要少得多. Zhou 等人的方法[28] 使用无监督学习网络, 首先对网络参数初始化赋值, 通过左右一致性检测产生初始置信地图, 用迭代的方式更新置信地图, 指导网络训练. Wang 等人的方法[29] 利用视差注意力机制计算特征相似度, 避免了视差范围过大的限制. 与有监督的学习方法相比, 无监督学习方法的最大优点是无需昂贵的标注数据集, 缺点是在有监督学习的训练场景下收敛效果不及有监督学习.

无论是有监督还是无监督的深度学习方法, 都需要大量数据集训练, 训练数据集的规模不足和获取难度限制了其预测效果和应用范围.

2 算法框架

算法框架如图 1 所示, 主要由场景位置猜测网络 (scene location guessing network, SLGNet)、互监督损失函数、视差优化等 3 个部分组成, 其输入是经过对极几何矫正后的双目图像对, 输出是视差图.

因为双目图像对是由同一场景经两次不同角度确定性投影生成的左右像, 故左右像天然可以互为导师数据. 假设已知场景深度, 则左像能够变换到右像, 右像能够变换到左像. 利用左右像互为导师数据的关系, 可以验证深度场景位置信息的合理性. 神经网络 SLGNet, 就是模拟逆推的深度场景位置信息的相溶解空间, 网络输出是解空间中的一个解, 即猜测的一种深度场景位置, 用视差图表示. 有了猜测的深度场景位置信息, 可以得到深度场景按照输入图像对相同投影角度向左、向右的投影图像对, 即得到猜测投影图像对. 猜测投影图像对和输入图像对做比较, 可以验证这种猜测深度场景位置信息的相容性, 其差异用互监督损失函数表示, 当这种差异小到一定程度时, 即可认为这种猜测是其一相容解. 该算法用互监督损失函数指导 SLGNet 网络的学习, 迭代优化深度场景位置信息的相容解空间, 最终得到从左往右、从右往左都相容的优化解. 由于遮挡、弱纹理、高光、噪音、透射、重复图案等病态区域的存在, 导致相容解的多解性. 为了使猜测的视差图尽可能地接近真实解还需要视差优化技术, 综合左右视差图的信息, 最终得到以左视差图为参照的目标视差图.

© 中国科学院软件研究所 http://www.c-s-a.org.cn

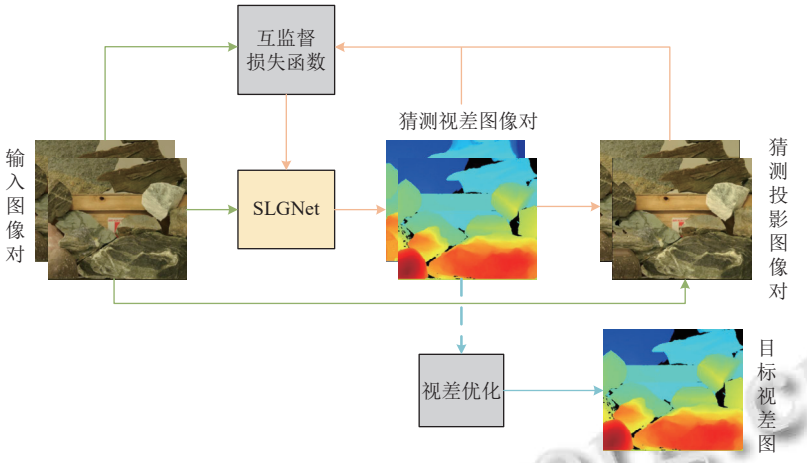


图 1 算法框架图

具体的算法如算法 1 所示. SLGNet 网络参数采用随机的方式初始化, 之后每一次迭代更新一次网络参数, 直到迭代达到一定次数为止, 再经过视差优化, 得到最终的视差图.

算法 1. 以双目图像互为监督的相容解迭代优选算法

输入: 双目图像I^{L}、I^{R}, 初始神经网络$SLGNet_0$;
输出: 视差图D^{L}.

1. /* 上标L、R分别表示左、右. */
2. /*神经网络参数采用随机的方式初始化*/;
3. FOR i IN $RANGE(0,\ n)$ /* n 为总迭代次数 */
4. $D_i^{\mathrm{L}}, D_i^{\mathrm{R}} \leftarrow SLGNet_i\left(I^{\mathrm{L}}, I^{\mathrm{R}}\right)$;
5. $P_i^{\mathrm{L}} \leftarrow project\left(D_i^{\mathrm{L}}, I^{\mathrm{R}}\right)$; /* $project$ 为投影函数 */
6. $P_i^{\mathrm{R}} \leftarrow project(D_i^{\mathrm{R}}, I^{\mathrm{L}})$;
7. $loss_i^{\mathrm{L}} \leftarrow supervise\left(I^{\mathrm{L}}, P_i^{\mathrm{L}}, D_i^{\mathrm{L}}\right)$; /* $supervise$ 为互监督损失函数 */
8. $loss_i^{\mathrm{R}} \leftarrow supervise\left(I^{\mathrm{R}}, P_i^{\mathrm{R}}, D_i^{\mathrm{R}}\right)$;
9. $loss \leftarrow loss_i^{\mathrm{L}} + loss_i^{\mathrm{R}}$; /* 总损失函数 */
10. $SLGNet_{i+1} \leftarrow update(SLGNet_i, loss)$;/* 梯度下降法更新网络参数 */
11. END FOR
12. $D^{\mathrm{L}} \leftarrow refine\left(D_n^{\mathrm{L}}, D_n^{\mathrm{R}}\right)$; /* $refine$ 为视差优化函数 */

2.1 SLGNet 网络架构

SLGNet 网络整体架构如图 2 所示, 包括特征提取模块、特征联合、视差预测模块等 3 个部分. 输入是经过对极几何矫正后的双目图像对, 经过特征提取网络模块分别提取左右图像特征, 然后一起送入一个卷积层, 进行特征联合, 再分别送入两个结构相同但非共享权重的视差预测模块, 分别输出左右视差图.

SLGNet 的每一个卷积层都是由卷积、归一化、非线性函数依次组成, 其中卷积核尺寸均是3×3, 归一化方法为实例归一化 (instance normalization[30]), 非线性函数为 ReLU[31]. SLGNet 网络无需批量数据集训练, 因此通常使用的批归一化 (batch normalization[32]) 并不适用 SLGNet, 而实例归一化计算归一化统计量时只需要考虑单个样本的单层中的单个通道, 更能考虑到单个像素的特征, 适合 SLGNet 架构.

2.1.1 特征提取模块

特征提取模块的主要功能是从图像中自动提取联合所需要的特征. 特征提取模块结构如图 3 所示, 图 3 中卷积指步长为 1 的卷积; 下卷积指步长为 2 的卷积, 会使宽度和高度各减小 1 倍, 通道增加 1 倍; 上卷积指步长为 2 的转置卷积, 会使宽度和高度各增加一倍, 通道减小一倍; 上采样指用双线性插值的方式增大空间尺寸, 用于弥补上卷积过程中填充的零值; 箭头表示数据流向. 首先, 双目图像通过 3 次下卷积逐层收缩网络, 再通过两个残差块扩大感觉野, 然后通过 3 次上卷积逐层恢复原尺寸, 在上卷积过程中, 同时进行上采样以增加信息向后传递, 最终得到特征向量组. 出于特征联合应该使用相同特征提取原则的考虑, 对于输入的左右图, SLGNet 使用同一个特征提取模块.

2.1.2 特征联合

特征联合是 SLGNet 中的关键一步, 让特征提取模块提取到的左右图像特征一起卷积, 形成视差预测所需要的综合特征. 该层只是简单的信息汇集, 故只有一个卷积层, 更深层的信息处理在视差预测模块. 只用一个卷积层的好处是既能够汇集信息, 又计算简单, 避免构造代价量、3D 卷积等计算复杂的操作.

© 中国科学院软件研究所 http://www.c-s-a.org.cn

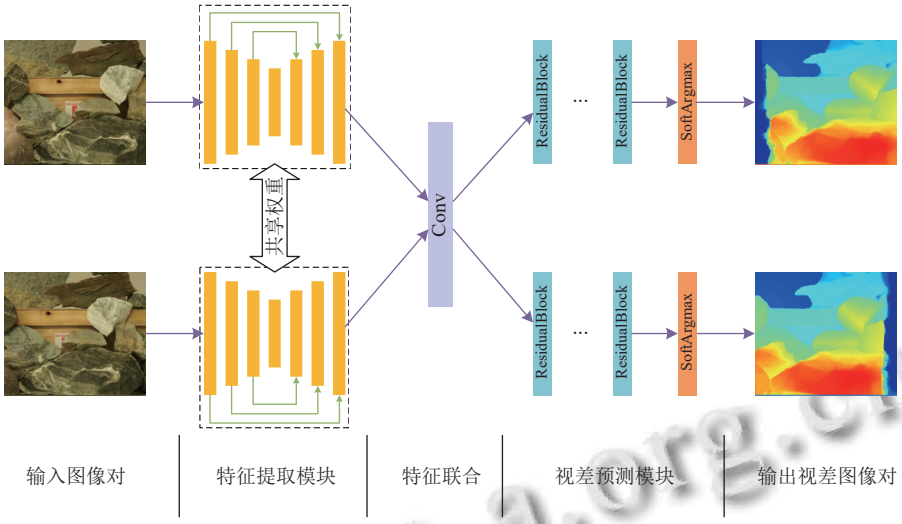


图 2 SLGNet 网络架构图

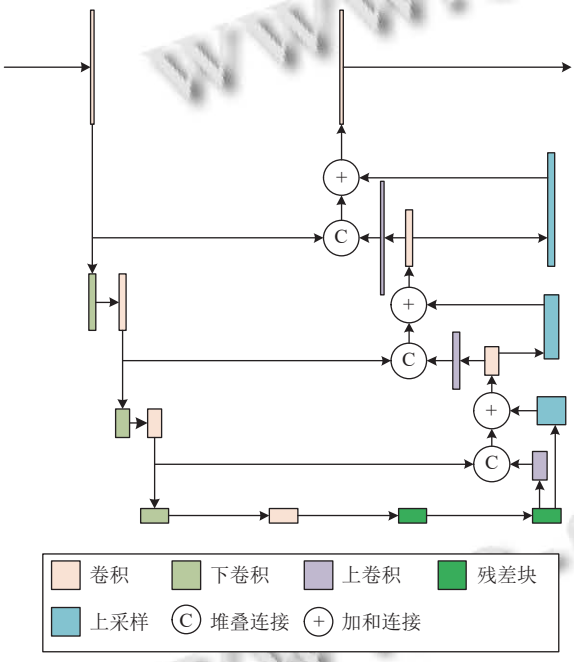


图 3 特征提取模块网络结构

2.1.3 视差预测模块

视差预测模块由若干个残差块 (ResidualBlock[33], 图 4) 后跟一个 *SoftArgmax* 组成, 主要功能是从联合特征中预测视差, 从而生成猜测视差图像对. 本文使用两个相同结构但非共享权重的视差预测模块分别预测左右视差. 两个视差预测模块之所以采用非共享权重的方式, 是因为左右图像对应视差的场景角度是不同的, 需要不同的预测方式. *SoftArgmax* 是 *Argmax* 的可微变体, 公式如下:

$$SoftArgmax(z)=\sum_i \frac{e^{z_i}}{\sum_j e^{z_j}}\cdot i \tag{1}$$

其中, z为一通道特征向量, z_i为特征向量分量, i, j为特征向量分量下标. *SoftArgmax* 以可微的方式计算特征向量的最大分量的近似位置, 进而用该位置表示预测视差.

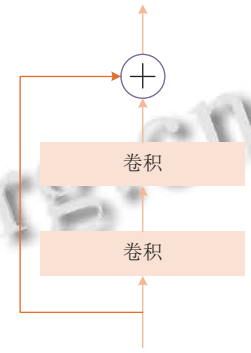


图 4 残差块结构

2.2 互监督损失函数

用d表示视差, 上标L、R分别表示左图和右图, x、y分别表示图像像素横坐标和纵坐标, 对于匹配像素点对, 根据视差定义, 则有关系式:

$$d^{\mathrm{L}}=d^{\mathrm{R}}=x^{\mathrm{L}}-x^{\mathrm{R}} \tag{2}$$

故已知右图像和左视差图可以计算出左投影图像, 已知左图像和右视差图可以计算出右投影图像.

互监督损失函数由投影差异函数、梯度投影差异函数和视差平滑函数组成.

用I表示图像亮度函数, $\|\cdot\|$表示L_1范式距离, 则以

© 中国科学院软件研究所 http://www.c-s-a.org.cn

左图为基准的亮度差异函数为:

$$B_1=\sum_x\sum_y\left\|I^{\mathrm{L}}(x,y)-I^{\mathrm{R}}(x-d^{\mathrm{L}},y)\right\|$$

同理, 以右图为基准的亮度差异函数为:

$$B_2=\sum_x\sum_y\left\|I^{\mathrm{L}}(x+d^{\mathrm{R}},y)-I^{\mathrm{R}}(x,y)\right\|$$

故, 总的投影差异函数为:

$$B=B_1+B_2 \tag{3}$$

用$\nabla_x I$表示图片x方向的亮度梯度函数, 则以左图为基准的x方向的亮度梯度差异函数为:

$$G_1=\sum_x\sum_y\left\|\nabla_x I^{\mathrm{L}}(x,y)-\nabla_x I^{\mathrm{R}}(x-d^{\mathrm{L}},y)\right\|$$

用$\nabla_y I$表示图片y方向的亮度梯度函数, 则以左图为基准的y方向的亮度梯度差异函数为:

$$G_2=\sum_x\sum_y\left\|\nabla_y I^{\mathrm{L}}(x,y)-\nabla_y I^{\mathrm{R}}(x-d^{\mathrm{L}},y)\right\|$$

同理, 以右图为基准的x、y方向的亮度梯度差异函数分别为:

$$G_3=\sum_x\sum_y\left\|\nabla_x I^{\mathrm{L}}(x+d^{\mathrm{R}},y)-\nabla_x I^{\mathrm{R}}(x,y)\right\|$$

$$G_4=\sum_x\sum_y\left\|\nabla_y I^{\mathrm{L}}(x+d^{\mathrm{R}},y)-\nabla_y I^{\mathrm{R}}(x,y)\right\|$$

故, 总的梯度投影差异函数为:

$$G=G_1+G_2+G_3+G_4 \tag{4}$$

亮度函数I及其梯度∇I都是离散的, 即对视差d不可导, 因此投影差异函数对d不可导. 为使投影差异函数对d可导, 投影变换时, 采用 Jaderberg 等人的空间变换的可导化处理方法[34].

对于弱纹理、遮挡、高光等病态区域, 投影差异函数和梯度投影差异函数均无能为力, 故还需要平滑约束函数. 用D表示视差函数, T表示指示函数, α为用户定义的阈值, 则左视差图的x方向平滑约束函数为:

$$S_1=\sum_x\sum_y\left|D^{\mathrm{L}}(x,y)-D^{\mathrm{L}}(x-1,y)\right|\cdot T\left[\left\|I^{\mathrm{L}}(x,y)-I^{\mathrm{L}}(x-1,y)\right\|<\alpha\right]$$

左视差图的y方向平滑约束函数为:

$$S_2=\sum_x\sum_y\left|D^{\mathrm{L}}(x,y)-D^{\mathrm{L}}(x,y-1)\right|\cdot T\left[\left\|I^{\mathrm{L}}(x,y)-I^{\mathrm{L}}(x,y-1)\right\|<\alpha\right]$$

同理, 右视差图的x、y方向的平滑约束函数分别为:

$$S_3=\sum_x\sum_y\left|D^{\mathrm{R}}(x,y)-D^{\mathrm{R}}(x-1,y)\right|\cdot T\left[\left\|I^{\mathrm{R}}(x,y)-I^{\mathrm{R}}(x-1,y)\right\|<\alpha\right]$$

$$S_4=\sum_x\sum_y\left|D^{\mathrm{R}}(x,y)-D^{\mathrm{R}}(x,y-1)\right|\cdot T\left[\left\|I^{\mathrm{R}}(x,y)-I^{\mathrm{R}}(x,y-1)\right\|<\alpha\right]$$

上述平滑约束函数的加和就是平滑约束损失函数, 即:

$$S=S_1+S_2+S_3+S_4 \tag{5}$$

故总的互监督损失函数为:

$$loss=B+G+S \tag{6}$$

2.3 视差优化

由于遮挡、弱纹理、高光等病态区域的存在, 匹配过程中难免存在误匹配、噪音的情况. 视差优化就是为了纠正这种误匹配和消除噪音. 本文算法仅使用 3 个视差优化步骤——左右一致性校验、填充、中值滤波.

2.3.1 左右一致性校验

已知左视差图和右视差图可以作左右一致性校验. 位置(x,y)的校验规则为:

(1) 若$d=D^{\mathrm{L}}(x,y)$, 满足$\left|d-D^{\mathrm{R}}(x-d,y)\right|\leqslant 1$, 则为正确匹配.

(2) 若$d\neq D^{\mathrm{L}}(x,y)$, 满足$\left|d-D^{\mathrm{R}}(x-d,y)\right|\leqslant 1$, 则为错误匹配.

(3) 其他则为遮挡.

根据上述校验规则, 每个像素位置只可标记为 3 种情况——正确匹配、遮挡、错误匹配的一种.

2.3.2 填充

根据上述校验标记结果, 对于正确匹配, 不做处理; 对于遮挡, 用背景的视差值进行填充, 即从当前位置向左寻找第 1 个标记为正确匹配的视差值填充, 若找不到, 则视为错误匹配, 用错误匹配的填充方法进行填充; 对于错误匹配, 按以下步骤计算出的视差进行填充.

(1) 从 16 个不同方向分别寻找前m个标记为正确匹配的视差值的位置, 得到 16 组位置集合.

(2) 对每一组位置集合, 累加集合中的位置与当前位置对应原图的绝对颜色差值.

(3) 取累加绝对颜色差值最小的位置集合为目标集合.

(4) 对目标集合中的m个位置及对应视差值进行线性拟合, 进而推断出当前位置的视差.

© 中国科学院软件研究所 http://www.c-s-a.org.cn

2.3.3 中值滤波

填充之后再进行 3 次中值滤波, 以消除残留的离群点和噪声.

3 实验

本文使用 Middlebury[35] 测评网站中的 Aloe、Cloth4、Cone、Rocks1、Teddy 等图像和 SceneFlow[36] 数据集作为实验对象. Middlebury 测评图像是在特定光照下拍摄的室内场景图像, 其真值视差图是用结构光的方式获得. SceneFlow 是一个大规模的合成数据集, 用于光流、视差估计、场景流估计. 实验代码是用 Python 编程语言编写, 版本号为 3.8.11, 使用 PyTorch 1.9.0 编程框架, CUDA 11.4, 其运行在硬件为 Intel(R) Xeon(R) Silver 4210R CPU @ 2.40 GHz、NVIDIA GeForce RTX 3090, 操作系统为 CentOS 7 的环境下. 对比实验代码同样运行在该环境下. 除特别说明外, 以下实验的参数均是: 视差预测模块的残差网络块数量为 15 个; 平滑约束函数自定义阈值α为 0.05; SLGNet 使用 Adam(0.9, 0.999) 优化器, 共迭代 3 000 次, 前 2 000 次迭代的学习率为 0.01, 后 1 000 次为 0.001; 视差优化步骤中的m为 6, 中值滤波大小为5×5.

参照立体视觉行业惯例, 评估指标使用误匹配率指标和 *EPE* (end-point error, 端点误差) 指标. 误匹配率是指预测视差图和真值视差图相比误差大于δ个像素所占的比率, 公式如下:

$$r_\delta = \frac{1}{N}\sum_x\sum_y T\left[\left|D_p(x,y)-D_t(x,y)\right|>\delta\right] \tag{7}$$

其中, $D_p(x,y)$表示预测的视差值, $D_t(x,y)$表示真实的视差值, N为像素个数, T是表示指示函数, δ为用户定义的阈值, 以下实验取$\delta=1$. 误匹配率越小表示匹配错误的像素越少, 效果越好.

EPE 指标公式如下:

$$EPE = \frac{1}{N}\sum_x\sum_y\left|D_p(x,y)-D_t(x,y)\right| \tag{8}$$

EPE 可以用来测量从立体匹配算法生成的视差图与基准视差图之间的差异. *EPE* 越低, 意味着立体匹配算法具有更高的精度, 从而得到更准确的深度估计.

指标又分为所有区域和非遮挡区域两种情况. 所有区域指整张图片区域, 而非遮挡区域指双目图像对中像素在左右图均可见的区域.

4 收敛过程分析

以 Teddy 图像为例, 分析该算法的收敛过程. 图 5 展示了 Teddy 的不同迭代次数的收敛情况, 图中视差图均未经过视差优化处理. 先以随机的方式初始化网络 SLGNet, 进而得到初始猜测视差左图, 再由初始猜测视差左图和输入右图经过投影变换得到初始猜测投影左图, 同理可得初始猜测视差右图和初始猜测投影右图. 由初始猜测视差图对、初始猜测投影图对和输入双目图对根据式 (6) 计算 SLGNet 网络的互监督损失函数, 再由梯度下降法更新 SLGNet 网络参数, 进而得到下一次迭代猜测视差图对、猜测投影图对, 如此循环迭代, 直至收敛. 从图 5 中可以看出, 初始时猜测视差图呈现出不连续的点, 迭代 10 次时视差已连成片, 迭代 100 次时视差图已有清晰的边界, 之后随着迭代次数的增加, 视差图越来越精细. 猜测投影图也和猜测视差图一样, 迭代次数越多, 和输入图的差异越小. 猜测视差左图最左边和猜测视差右图最右边没有匹配是因为这些区域对应的像素只在一张图片中出现, 即要么出现在左图, 要么出现在右图, 故没有匹配点.

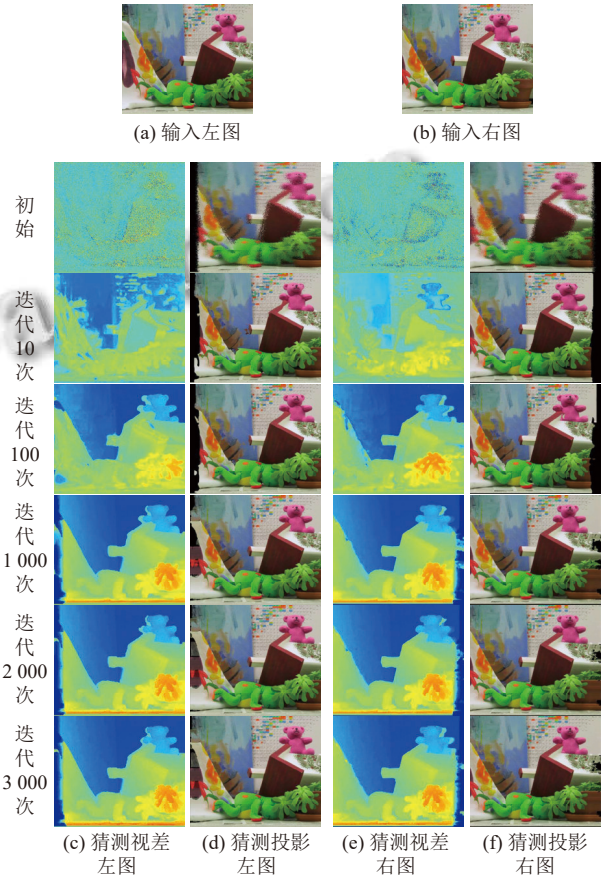


图 5 Teddy 收敛过程

© 中国科学院软件研究所 http://www.c-s-a.org.cn

5 对比实验

对比算法有 CREStereo、LBPS[37]、RAFT-Stereo、CDR-Fusion[38]、CFNet[39]、PSMNet、GA-Net、PCW-Net[40], 均为近 5 年发表的算法, 其中 CREStereo、CDR-Fusion、PCW-Net 为 2022 年发表, RAFT-Stereo、CFNet 为 2021 年发表. 对于深度学习算法, 其程序均是其作者公开的代码, 其中 CREStereo、RAFT-Stereo 使用其作者预训练好的 ETH3D 网络模型参数, PSMNet、GA-Net、CFNet、PCW-Net、LBPS 使用其作者预训练好的 kitti 2015 网络模型参数. CDR-Fusion 是无监督的方法, 使用其作者预训练好 SceneFlow 网络模型参数. 这里和深度学习算法作对比, 对比的是其泛化能力, 故使用的不是预训练好的和预测图像相同场景的模型参数. 所有算法均在同一台服务器下运行.

图 6 是 Middlebury 部分图像的对比效果. 算法得到的视差图和真值图均用相同方式着色, 故视差图与真值图相比, 越接近越好. 视差图下面是匹配错误分布图, 黑色的点表示匹配错误, 白色的点表示匹配正确, 从匹配错误分布图中可以清楚地看到, 匹配错误像素的位置. 从图 6 的匹配错误分布图可以看出, 匹配错误的像素集中在图像左边和纹理边缘处, 这些区域往往是遮挡区域. 遮挡区域的视差不能由匹配计算, 只能由周围像素推断填充, 而填充的结果并不可靠, 因此评价指标以非遮挡区域为主, 所有区域为辅. 从图 6 可以看出, 与 LBPS 算法相比, 本文算法的匹配错误区域要小, 与 PCW-Net 算法、CF-Net 算法相比, 本文算法的匹配错误的像素左则部分 (红色框框起区域) 没有明显差别, 纹理边缘处 (绿色框框起区域) 要更少.

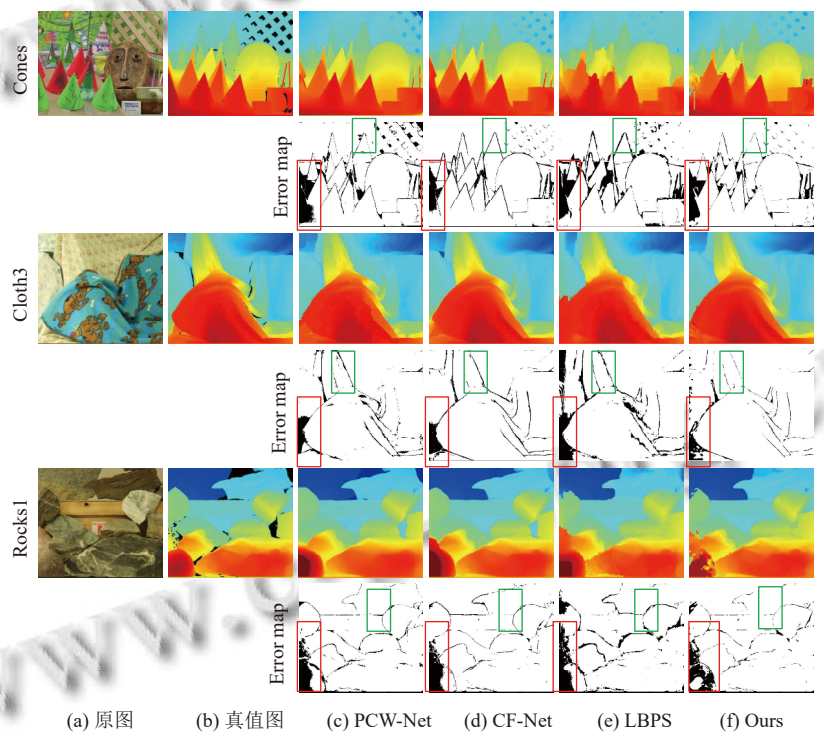


图 6 Middlebury 部分图像对比效果

表 1、表 2 是 Middlebury 对比实验的误匹配率数据. 表中黑体标注的数字表示该数字在各算法中排名第 1. 从表 1 中可以看出, 在非遮挡区域, 9 张图中, 本文算法有 6 张图的误匹配率排名第 1, CFNet 算法有 2 张, CREStereo 算法有 1 张, 本文算法平均误匹配率为 2.52%, 比表中其他算法都要低. 从表 2 中可以看出, 在所有区域, 9 张图中, 本文算法有 5 张图的误匹配率排名第 1, CFNet 算法有 2 张, CREStereo 算法和 LBPS 算法各有 1 张, 本文算法平均误匹配率为 7.26%, 比表中其他算法都要低.

表 3 是 Middlebury 对比实验的平均 *EPE* 指标, 在非遮挡区域, 本文算法与 CFNet 算法的平均 *EPE* 均为 0.30 像素, 并列第 1; 在所有区域, 本文算法的 *EPE* 为 0.54 像素, 次于 CFNet、RAFT-Stereo 算法, 优于

© 中国科学院软件研究所 http://www.c-s-a.org.cn

LBPS、PCW-Net、PSMNet、GA-Net、CREStereo 算法. 对于 *EPE* 指标, 误差大的像素比误差小的像素对指标的影响大, 而误差大的像素主要是由于遮挡区域的填充引起的. 在误匹配率对比中, 本文算法在非遮挡区域和所有区域均有优势, 在 *EPE* 对比中, 本文算法在非遮挡区域有优势, 在所有区域没有优势, 因此, 本文算法的优势主要在非遮挡区域.

表 4、表 5 是 30 张 SceneFlow 图像的平均指标. 从表 4 中可以看出, 在非遮挡区域, 本文算法的平均误匹配率为 3.01%, 小于其他算法, 效果最好; 在所有区域, 本文算法的平均误匹配率为 6.39%, 小于其他算法, 效果最好; 从表 5 中可以看出, 在非遮挡区域, 本文算法的平均 *EPE* 为 0.35 像素, 低于其他算法, 效果最好; 在所有区域, 本文算法的平均 EPE 为 0.76 像素, 只次于 PCW-Net 的 0.62 像素. 在 SceneFlow 图像的对比实验和在 Middlebury 图像的对比实验一致, 均表明本文算法有更低的误匹配率, 优势主要在非遮挡区域.

表 1 各算法在 Middlebury 的非遮挡区域误匹配率对比 (%)

算法	Aloe	Cloth1	Cloth2	Cloth3	Cloth4	Cones	Rocks1	Rocks2	Teddy	平均
LBPS[37]	6.53	0.93	6.46	3.94	2.57	7.41	3.67	2.9	12.94	5.26
PCW-Net[40]	6.69	0.80	4.70	3.21	3.26	6.74	2.64	3.06	7.65	4.31
PSMNet[21]	88.74	55.38	28.51	60.00	33.71	53.56	75.10	58.71	84.78	59.83
GA-Net[22]	13.01	1.58	10.65	10.94	8.34	17.61	5.72	7.10	15.27	10.02
CFNet[39]	6.15	0.55	2.73	2.60	2.47	4.55	**2.00**	**1.65**	5.04	3.08
CREStereo[24]	6.55	0.97	4.67	4.05	3.11	4.71	5.20	4.46	**4.46**	4.24
RAFT-Stereo[27]	7.08	0.64	2.93	2.28	2.33	4.85	4.34	2.50	4.99	3.55
Ours	**3.40**	**0.33**	**1.52**	**1.30**	**1.74**	**3.55**	2.77	1.78	6.28	**2.52**

表 2 各算法在 Middlebury 的所有区域误匹配率对比 (%)

算法	Aloe	Cloth1	Cloth2	Cloth3	Cloth4	Cones	Rocks1	Rocks2	Teddy	平均
LBPS[37]	11.58	7.52	14.09	10.33	**8.45**	13.71	10.05	11.09	16.49	11.48
PCW-Net[40]	8.88	4.18	12.45	5.84	13.25	12.38	7.06	7.89	11.13	9.23
PSMNet[21]	89.62	58.19	32.34	62.50	36.49	56.85	76.75	62.12	83.92	62.09
GA-Net[22]	20.24	8.96	19.08	14.75	20.05	23.94	11.57	14.43	18.89	16.88
CFNet[39]	10.22	3.86	8.76	5.66	11.82	**8.78**	6.77	**6.59**	7.15	7.73
CREStereo[24]	8.92	3.38	12.29	5.94	10.81	8.84	9.79	8.83	**5.98**	8.31
RAFT-Stereo[27]	9.37	4.42	10.47	**3.94**	11.97	9.73	9.15	8.19	7.59	8.31
Ours	**8.01**	**3.36**	**8.29**	4.00	8.69	**8.78**	**6.60**	7.34	10.30	**7.26**

表 3 各算法在 Middlebury 的平均 *EPE* 对比 (像素)

算法	非遮挡区域	所有区域
LBPS[37]	0.47	0.88
PCW-Net[40]	0.41	0.57
PSMNet[21]	1.38	1.67
GA-Net[22]	0.67	1.28
CFNet[39]	**0.30**	0.49
CREStereo[24]	0.71	0.82
RAFT-Stereo[27]	0.31	**0.48**
Ours	**0.30**	0.54

表 4 各算法在 SceneFlow 的平均误匹配率对比 (%)

算法	非遮挡区域	所有区域
LBPS[37]	5.42	11.03
CRD-Fusion[38]	6.49	11.98
CFNet[39]	8.53	12.90
PCW-Net[40]	3.51	7.14
Ours	**3.01**	**6.39**

表 5 各算法在 SceneFlow 的平均 *EPE* 对比 (像素)

算法	非遮挡区域	所有区域
LBPS[37]	0.64	1.26
CRD-Fusion[38]	0.87	1.20
CFNet[39]	0.71	1.06
PCW-Net[40]	0.40	**0.62**
Ours	**0.35**	0.76

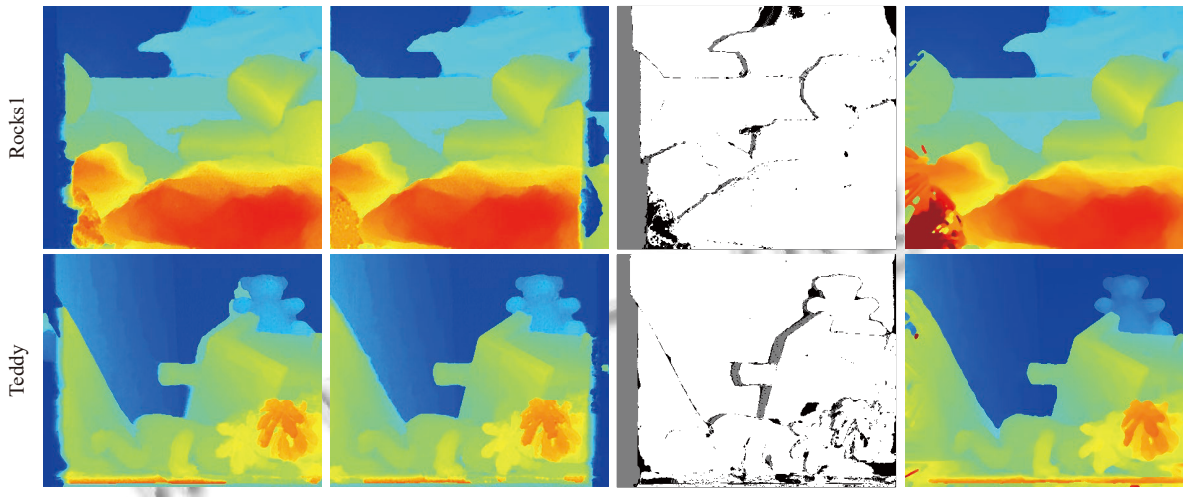
6 视差优化实验

为了证明视差优化的有效性, 本文以 Rocks1 和 Teddy 图像为例, 对比了视差优化前和视差优化后的效果, 见图 7. 图 7 中, 左右视差图是 SLGNet 生成的未经过视差优化的视差图; 校验图是视差优化步骤中左右一致性校验后形成的标记图, 其中白色为正确匹配区域, 灰色为遮挡区域, 黑色为误匹配区域; 目标视差图

© 中国科学院软件研究所 http://www.c-s-a.org.cn

是以左视差图为基础经过视差优化生成的图, 即算法最终生成的视差图. 图中, 目标视差图与左视差图相比, 填充了校验图中的遮挡区域和误匹配区域.

表 6 是视差优化前后的误匹配率的变化, 优化前后的最小值用黑色字体标出. 从表中可以看出, 视差优化前, 非遮挡区域平均误匹配率为 3.22%, 所有区域为 14.29%, 视差优化后, 非遮挡区域平均误匹配率为 2.52%, 所有区域为 7.26%, 均有所下降, 但所有区域的下降幅度更大. 无论是在所有区域, 还是在非遮挡区域, 所有图片的误匹配率均下降.



(a) 左视差图 (b) 右视差图 (c) 校验图 (d) 目标视差图

图 7 视差优化效果图

表 6 视差优化前后误匹配率对比 (%)

区域	优化前后	Aloe	Cloth1	Cloth2	Cloth3	Cloth4	Cones	Rocks1	Rocks2	Teddy	平均
非遮挡区域	优化前	4.50	0.48	2.48	1.83	1.96	4.76	3.45	2.25	7.24	3.22
	优化后	**3.40**	**0.33**	**1.52**	**1.30**	**1.74**	**3.55**	**2.77**	**1.78**	**6.28**	**2.52**
所有区域	优化前	15.97	10.28	15.99	12.00	15.87	15.41	13.10	13.09	16.88	14.29
	优化后	**8.01**	**3.36**	**8.29**	**4.00**	**8.69**	**8.78**	**6.60**	**7.34**	**10.30**	**7.26**

7 结束语

本文提出一种既不同于传统算法又不同于深度学习算法的双目立体匹配算法, 该算法使用神经网络, 但不需要训练数据集训练, 以双目图像互为监督学习, 仅在一对当前图像上反复学习, 用迭代寻优过程不断精确视差图, 在 Middlebury、SceneFlow 数据集图像上的对比实验表明该算法比 CREStereo、RAFT-Stereo、PCW-Net 等算法有更低的误匹配率. 该算法使用 SLGNet 神经网络架构, 该网络产生一对视差图, 故可进行视差优化, 从而可以充分利用视差优化技术增强视差预测效果. 针对 SLGNet 网络的收敛, 本文还设计了互监督损失函数, 该损失函数的特点是无需视差真值图, 其可以应用于非监督学习的双目立体匹配算法.

参考文献

1 Wei H, Xu C, Jin ZF. Binocular matching model based on hierarchical V1 and V2 receptive fields with color, orientation, and region feature information. IEEE Transactions on Biomedical Engineering, 2020, 67(11): 3141–3150. [doi: 10.1109/TBME.2020.2977350]

2 Geiger A, Lenz P, Urtasun R. Are we ready for autonomous driving? The KITTI vision benchmark suite. Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition. Providence: IEEE, 2012. 3354–3361.

3 Geiger A, Lenz P, Stiller C, *et al*. Vision meets robotics: The KITTI dataset. The International Journal of Robotics Research, 2013, 32(11): 1231–1237. [doi: 10.1177/0278364913491297]

4 Hamzah RA, Ibrahim H. Literature survey on stereo vision disparity map algorithms. Journal of Sensors, 2016, 2016: 8742920.

5 Yang QX, Yang RG, Davis J, *et al*. Spatial-depth super resolution for range images. Proceedings of the 2007 IEEE Conference on Computer Vision and Pattern Recognition.

© 中国科学院软件研究所 http://www.c-s-a.org.cn

Minneapolis: IEEE, 2007. 1–8.
6 Lee SH, Sharma S. Real-time disparity estimation algorithm for stereo camera systems. IEEE Transactions on Consumer Electronics, 2011, 57(3): 1018–1026. [doi: 10.1109/TCE.2011.6018850]
7 Fusiello A, Castellani U, Murino V. Relaxing symmetric multiple windows stereo using Markov random fields. Proceedings of the 3rd International Workshop on Energy Minimization Methods in Computer Vision and Pattern Recognition. Sophia Antipolis: Springer, 2001. 91–105.
8 Hirschmuller H, Scharstein D. Evaluation of cost functions for stereo matching. Proceedings of the 2007 IEEE Conference on Computer Vision and Pattern Recognition. Minneapolis: IEEE, 2007. 1–8.
9 Gu Z, Su XY, Liu YK, *et al*. Local stereo matching with adaptive support-weight, rank transform and disparity calibration. Pattern Recognition Letters, 2008, 29(9): 1230–1235. [doi: 10.1016/j.patrec.2008.01.032]
10 Ma L, Li JJ, Ma J, *et al*. A modified census transform based on the neighborhood information for stereo matching algorithm. Proceedings of the 7th International Conference on Image and Graphics. Qingdao: IEEE, 2013. 533–538.
11 Scharstein D, Szeliski R. A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. International Journal of Computer Vision, 2002, 47(1–3): 7–42.
12 Veksler O. Fast variable window for stereo correspondence using integral images. Proceedings of the 2003 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. Madison: IEEE, 2003. 556–561.
13 Zhao Y, Taubin G. Real-time stereo on GPGPU using progressive multi-resolution adaptive windows. Image and Vision Computing, 2011, 29(6): 420–432. [doi: 10.1016/j.imavis.2011.01.007]
14 Fang JB, Varbanescu AL, Shen J, *et al*. Accelerating cost aggregation for real-time stereo matching. Proceedings of the IEEE 18th International Conference on Parallel and Distributed Systems. Singapore: IEEE, 2012. 472–481.
15 Liang CK, Cheng CC, Lai YC, *et al*. Hardware-efficient belief propagation. IEEE Transactions on Circuits and Systems for Video Technology, 2011, 21(5): 525–537. [doi: 10.1109/TCSVT.2011.2125570]
16 Wang HQ, Wu M, Zhang YB, *et al*. Effective stereo matching using reliable points based graph cut. Proceedings of the 2013 Visual Communications and Image Processing (VCIP). Kuching: IEEE, 2013. 1–6.
17 Hirschmuller H. Stereo processing by semiglobal matching and mutual information. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2007, 30(2): 328–341.
18 Bleyer M, Rhemann C, Rother C. PatchMatch stereo-stereo matching with slanted support windows. Proceedings of the 2011 British Machine Vision Conference. Dundee: BMVC, 2011. 1–11.
19 Žbontar J, LeCun Y. Stereo matching by training a convolutional neural network to compare image patches. The Journal of Machine Learning Research, 2016, 17(1): 2287–2318.
20 Kendall A, Martirosyan H, Dasgupta S, *et al*. End-to-end learning of geometry and context for deep stereo regression. Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV). Venice: IEEE, 2017. 66–75.
21 Chang JR, Chen YS. Pyramid stereo matching network. Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018. 5410–5418.
22 Zhang FH, Prisacariu V, Yang RG, *et al*. GA-Net: Guided aggregation net for end-to-end stereo matching. Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Long Beach: IEEE, 2019. 185–194.
23 Tankovich V, Häne C, Zhang YD, *et al*. HITNet: Hierarchical iterative tile refinement network for real-time stereo matching. Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Nashville: IEEE, 2021. 14357–14367.
24 Li JK, Wang PS, Xiong PF, *et al*. Practical stereo matching via cascaded recurrent network with adaptive correlation. Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). New Orleans: IEEE, 2022. 16242–16251.
25 Lipson L, Teed Z, Deng J. RAFT-Stereo: Multilevel recurrent field transforms for stereo matching. Proceedings of the 2021 International Conference on 3D Vision (3DV). London: IEEE, 2021. 218–227.
26 Zhao HL, Zhou HZ, Zhang YJ, *et al*. EAI-stereo: Error aware iterative network for stereo matching. Proceedings of the 16th Asian Conference on Computer Vision. Macao: Springer, 2022. 3–19.
27 Teed Z, Deng J. RAFT: Recurrent all-pairs field transforms for optical flow. Proceedings of the 16th European Conference on Computer Vision. Glasgow: Springer, 2020. 402–419.

© 中国科学院软件研究所 http://www.c-s-a.org.cn

28 Zhou C, Zhang H, Shen XY, *et al*. Unsupervised learning of stereo matching. Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV). Venice: IEEE, 2017. 1576–1584.

29 Wang LG, Guo YL, Wang YQ, *et al*. Parallax attention for unsupervised stereo correspondence learning. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2022, 44(4): 2108–2125. [doi: 10.1109/TPAMI.2020.3026899]

30 Huang X, Belongie S. Arbitrary style transfer in real-time with adaptive instance normalization. Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV). Venice: IEEE, 2017. 1510–1519.

31 Krizhevsky A, Sutskever I, Hinton GE. ImageNet classification with deep convolutional neural networks. Communications of the ACM, 2017, 60(6): 84–90. [doi: 10.1145/3065386]

32 Ioffe S, Szegedy C. Batch normalization: Accelerating deep network training by reducing internal covariate shift. Proceedings of the 32nd International Conference on Machine Learning. Lille: JMLR.org, 2015. 448–456.

33 He KM, Zhang XY, Ren SQ, *et al*. Deep residual learning for image recognition. Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Las Vegas: IEEE, 2016. 770–778.

34 Jaderberg M, Simonyan K, Zisserman A. Spatial transformer networks. Proceedings of the 28th International Conference on Neural Information Processing Systems. Montreal: MIT Press, 2015. 2017–2025.

35 Scharstein D, Pal C. Learning conditional random fields for stereo. Proceedings of the 2007 IEEE Conference on Computer Vision and Pattern Recognition. Minneapolis: IEEE, 2007. 1–8.

36 Mayer N, Ilg E, Häusser P, *et al*. A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation. Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas: IEEE, 2016. 4040–4048.

37 Knobelreiter P, Sormann C, Shekhovtsov A, *et al*. Belief propagation reloaded: Learning BP-layers for labeling problems. Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle: IEEE, 2020. 7897–7906. [doi: 10.1109/CVPR42600.2020.00792]

38 Fan XL, Jeon S, Fidan B. Occlusion-aware self-supervised stereo matching with confidence guided raw disparity fusion. Proceedings of the 19th Conference on Robots and Vision (CRV). Toronto: IEEE, 2022. 132–139. [doi: 10.1109/CRV55824.2022.00025]

39 Shen ZL, Dai YC, Rao ZB. CFNet: Cascade and fused cost volume for robust stereo matching. Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville: IEEE, 2021. 13901–13910.

40 Shen ZL, Dai YC, Song XB, *et al*. PCW-Net: Pyramid combination and warping cost volume for stereo matching. Proceedings of the 17th European Conference on Computer Vision. Tel Aviv: Springer, 2022. 280–297. [doi: 10.1007/978-3-031-19824-3_17]

(校对责编: 牛欣悦)

© 中国科学院软件研究所 http://www.c-s-a.org.cn