智能系统学报
CAAI TRANSACTIONS ON INTELLIGENT SYSTEMS

车载双目视觉动态级联修正实时立体匹配网络

何国豪, 翟涌, 龚建伟, 王羽纯, 张曦

引用本文:
何国豪,翟涌,龚建伟,王羽纯,张曦. 车载双目视觉动态级联修正实时立体匹配网络[J]. 智能系统学报, 2022, 17(6): 1145–1153.
HE Guohao,ZHAI Yong,GONG Jianwei,WANG Yuchun,ZHANG Xi. Real–time stereo matching network for vehicle binocular vision based on dynamic cascade correction[J]. *CAAI Transactions on Intelligent Systems*, 2022, 17(6): 1145–1153.

在线阅读 View online: https://dx.doi.org/10.11992/tis.202111013

DOI: 10.11992/tis.202111013
网络出版地址: https://kns.cnki.net/kcms/detail/23.1538.TP.20220922.1259.002.html

车载双目视觉动态级联修正实时立体匹配网络

何国豪[1], 翟涌[1], 龚建伟[1,2], 王羽纯[1], 张曦[2]
(1. 北京理工大学 机械与车辆学院, 北京 100081; 2. 北京理工大学 重庆创新中心, 重庆 401120)

摘 要: 针对目前基于双目视觉的高精度立体匹配网络消耗计算资源多、运算时间长、无法用于智能驾驶系统实时导航的问题, 本文提出了一种能够满足车载实时性和准确性要求的动态融合双目立体匹配深度学习网络。该网络加入了基于全局深度卷积的注意力模块完成特征提取, 减少了网络层数与参数数量; 通过动态代价级联融合、多尺度融合以及动态视差结果修复优化 3D 卷积计算, 加速了常用的 3D 特征融合过程。将训练好的模型部署在车载硬件例如 NVIDIA Jetson TX2 上, 并在公开的 KITTI 立体匹配数据集上进行测试。实验显示, 该方法可以达到与目前公开在排行榜中最好方法相当的运行精度, 3 像素点误差小于 6.58%, 并且运行速度小于 0.1 s/f, 能够达到车载实时使用性能要求。

关键词: 双目视觉; 深度学习; 立体匹配; 视差估计; 动态计算; 特征融合; 车载视觉
中图分类号: TP29 **文献标志码**: A **文章编号**: 1673−4785(2022)06−1145−09

中文引用格式: 何国豪, 翟涌, 龚建伟, 等. 车载双目视觉动态级联修正实时立体匹配网络 [J]. 智能系统学报, 2022, 17(6): 1145–1153.
英文引用格式: HE Guohao, ZHAI Yong, GONG Jianwei, et al. Real-time stereo matching network for vehicle binocular vision based on dynamic cascade correction[J]. CAAI transactions on intelligent systems, 2022, 17(6): 1145–1153.

Real-time stereo matching network for vehicle binocular vision based on dynamic cascade correction

HE Guohao[1], ZHAI Yong[1], GONG Jianwei[1,2], WANG Yuchun[1], ZHANG Xi[2]
(1. School of Mechanical Engineering, Beijing Institute of Technology, Beijing 100081, China; 2. Chongqing Innovation Center, Beijing Institute of Technology, Chongqing 401120, China)

Abstract: Given the shortcoming of high-precision stereo matching networks based on binocular vision, such as high computing resource consumption, long operating time, and inability to be used in real-time navigation by intelligent driving systems, this study proposes a dynamic fusion stereo matching deep learning network that can meet real-time and accuracy requirements in vehicles. The network includes a global deep convolution-based attention module to complete feature extraction while reducing the number of network layers and parameters and optimizing 3D convolution calculations through dynamic cost cascade fusion, multi-scale fusion, and dynamic disparity change to accelerate the commonly used 3D feature fusion process. The trained model is tested on KITTI Stereo 2015 dataset using onboard hardware such as the NVIDIA Jetson TX2. Experiments show that the method can achieve the same operating accuracy as the state-of-the-art method currently in the leaderboard, 3 pixels error is less than 6.58%, and the operating duration is less than 0.1 seconds per frame, meeting real-time performance requirements.

Keywords: binocular vision; deep learning; stereo matching; disparity estimation; dynamic computation; feature fusion; on-board vision

随着智能驾驶技术的快速发展, 车用视觉传感器及其数据处理方法也得到了广泛的关注。与单目相机相比, 在没有其他传感器辅助的情况下, 双目相机也可以获取环境的深度信息。所以双目视觉相关技术也得到快速发展。

从双目立体图像获取环境深度信息的关键步骤是视差估计, 也称为立体匹配。双目立体匹配需要解决的关键问题是如何对左右两目图像进行特征提取并进行相似性匹配。正确的相似特征匹

收稿日期: 2021−11−06. **网络出版日期**: 2022−09−23.
基金项目: 国家自然科学基金项目(U19A2083, 61703041).
通信作者: 龚建伟.E-mail:gongjianwei@bit.edu.cn.

配才能带来正确的视差与深度计算结果。传统的立体匹配算法，例如图割 (graph cuts, GC)[1]、区块匹配 (block matching, BM)、半全局匹配 (semi-global matching, SGM)、半全局区块匹配 (semi-global block matching, SGBM)[2] 遵循匹配代价计算、匹配代价聚合、视差计算和视差优化四步进行相似特征的搜索与匹配。虽然不需要大量的运算资源，但大多需要较长的运行时间，且无法得到理想的结果[3]，匹配空洞较多。

随着卷积神经网络 (convolutional neural networks, CNNs) 的发展，部分学者发现使用卷积核计算匹配代价比人造代价函数计算的匹配代价更准确[4]，于是开始出现使用 CNN 计算图像的相似性，以及计算每像素点视差大小的方法，如匹配代价卷积 (MC-CNN)[5]，半全局匹配网络 (SGM-Net)[6]。但这些方法仅仅是用 CNN 网络替换了传统方法的代价计算与聚合步骤，仍需进行图像后处理才能得到视差输出结果[7]，所以匹配空洞仍然存在[8]。

为了获取稠密的立体匹配结果，减少匹配空洞，近来端到端的学习网络开始应用到立体匹配方法中。GC-Net[9] 是第一个端到端的立体匹配深度学习网络，该网络使用了一个 3D 的编码-解码器完成立体匹配任务。PSM-Net[10] 应用了一个紧凑沙漏状的全 3D 卷积网络来得到最终的视差图。GA-Net[3] 对 GC-Net 进行了改进，GA-Net 在 3D 卷积阶段集成了传统 SGBM 匹配算法结果的特征图。以上的方法都有较深的 3D 卷积核或者大量 3D 特征融合计算，要求配以高性能硬件才能达到想要的效率，否则需要较长的运算时间。为了减轻计算负担，一些网络开始加入了替代 3D 卷积网络的结构，例如 AA-Net[11] 和 AART-Stereo[12]，但又出现了特征提取效果差、特征代价融合不完备、场景通用性不强等问题。

针对上述学习网络层数深、计算量大、特征融合效果不佳、场景通用性不强等问题，本文针对车载实时需求，提出了一种端到端的级联多尺度融合与动态融合的双目立体匹配网络。

首先采用了一个高效的三阶段降采样特征提取模块，该模块使用了基于全局深度卷积的可学习注意力模块和可学习空间金字塔池化模块，不但可以在较浅的网络中提供一个理想的多尺度融合特征，还能对每个分辨率尺度的特征图提供可学习的融合权重，从网络深度与卷积参数数量上缓解了运算负担，加速了运算过程。之后，借鉴了 AART-Stereo 对于 3D 卷积网络的替代方案，采用多阶段分辨率输出结果进行融合，使用低分辨率结果优化高分辨率的输出，并且还计算了每一阶段的动态视差代价以完成 3 像素点误差修正，用于最后的视差回归。最后，在公开 KITTI 立体匹配数据集[13] 上完成了对本文方法的验证。

1　相关技术

1.1　视差估计

视差指的是两张图像上相关联的两个像素点在像素坐标系下的位置差异。如果对图像对进行预处理，那么位置差异就指在水平方向上的像素坐标差值。如果在图像对的左目图像(u,v)处找到一个像素点，与它相对应的点在右图中的$(u-d,v)$处找到，就称视差值是 d，随即可以由 fb/d 得到深度值 Z，其中 f 是焦距，b 是双目相机基线长度。视差估计就是对图像对的左图上所有像素点进行视差计算，得到一张稠密视差图的过程，该过程也称为立体匹配。

1.2　高效 CNN 模块

分组卷积[14] 是减少每秒浮点运算次数（floating-point operations per second, FLOPS）并增加卷积网络效率的最常用的方法。对于一组输入输出，若将输入卷积通道分为 G 组，参加运算的卷积核总参数量将会减少到原先参数量的 $1/G$。在分组卷积中，若输入通道数 C 等于输出通道数 N 也等于分组数 G，则可将这样的分组卷积操作称为深度可分离卷积[15](depth-wise separable convolution, DS-Conv)。深度可分离卷积是一个精度与速度共存的卷积方法，虽然参与运算的卷积核参数减少了，但是通常会获得比常规卷积更优秀的结果。

最近的研究发现，卷积神经网络每一层的卷积核有内在的关联性，意味着可以将该层内的所有卷积核用一个到两个卷积核配以一定的权重大小来描述[16]。蓝图分离卷积 (blueprint separable convolution, BS-Conv) 就采用这样的思路。在本文提出的特征提取模块中就运用蓝图分离卷积完成特征提取，显著减少了运算参数量与运行时间。

1.3　注意力模块

注意力模块已经广泛运用在了卷积神经网络中，特别是拥有大量层数的网络结构。网络结构越深，提取到的特征相比原始输入就越抽象，越难找到原始输入中感兴趣的区域，这也是过深的网络结构反而会出现网络性能下降的原因。注意力模块通常是一个加入先验知识或者从网络较浅层提取出的一个权重分布，将该模块应用到网络的深处，可以强化网络在深层处的有效特征表示，同时抑制无效特征和噪声。这样能够强化网

络的鲁棒性, 使网络有更强的场景通用性以及更高的运算准确率[17]。

1.4　全局特征融合

为了拓展学习网络的感知域, 获得多尺度的特征图, 在特征提取框架中做代价与特征融合是至关重要的。池化操作是最常用的降采样与特征融合操作, 不同的池化核大小和池化核步长可以获得不同尺度的特征图。空间金字塔池化模块(spatial pyramid pooling, SPP) 对输入特征图采用不同大小的池化核, 进行池化操作之后, 又将获得的所有特征图上采样到相同大小, 完成多尺度特征融合以得到多尺度特征图。在通道端, 全局平均池化 (global average pooling, GAP) 可以将每一个输入通道的特征进行融合, 获取一个基于通道的权重图。

但是, 不管是空间金字塔池化模块还是全局平均池化模块都是不可学习的, 所以无法对不同尺度或不同通道特征的重点性进行区分。本文针对这种情况, 给予池化模块可学习的参数。在上文提到的深度可分离卷积操作中, 若卷积核的大小等于输入特征图的大小, 则称为全局深度卷积(global depth-wise convolution ,GDC)。此类卷积可以看作全局加权池化, 其输出大小与全局平均池化相同; 不同的是, 全局深度卷积会赋予每个输入通道一个可学习的卷积核, 对不同通道特征的重要程度进行区分[18]。

2　立体匹配网络

根据双目立体匹配需要解决的问题, 如准确地进行特征选择与匹配等, 以及本文提出的应用场景, 结合上文介绍的当下较先进的模块, 本文提出了一个具有浅层高效的多尺度特征提取模块、轻量化 3D 卷积网络以及多尺度动态融合模块的立体匹配网络, 网络结构如图 1 所示, 图中®代表动态视差计算操作, ⊕代表像素逐点相加。

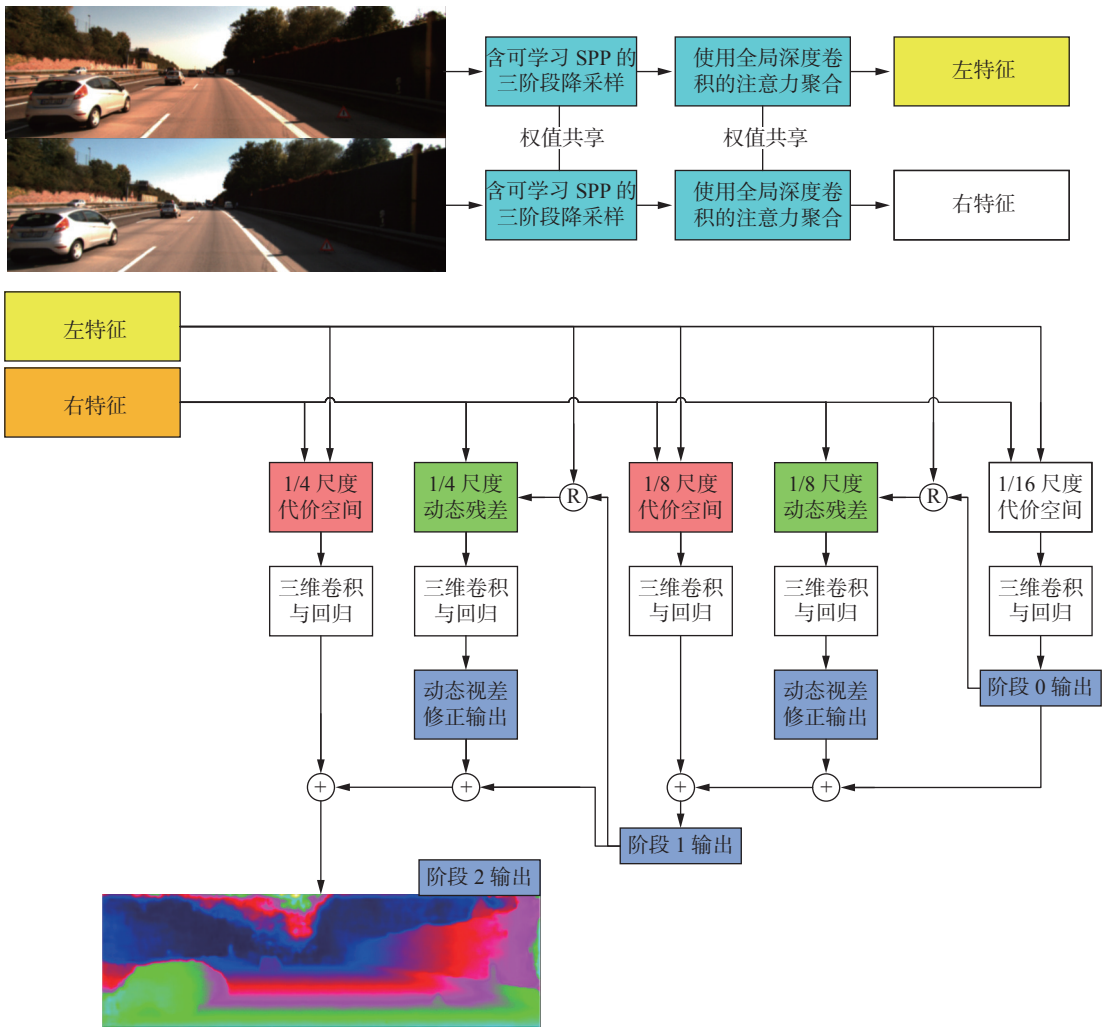


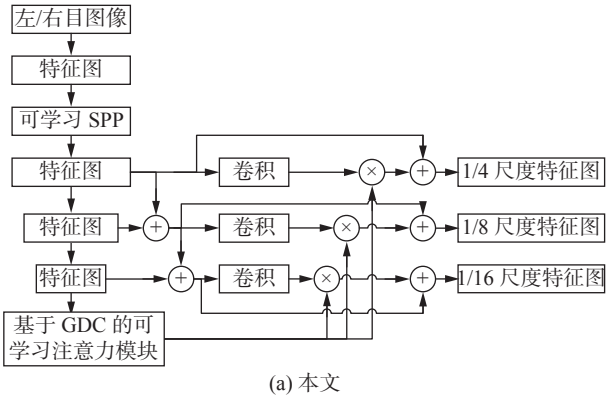
图 1　网络结构

Fig. 1　Sturcture of the network

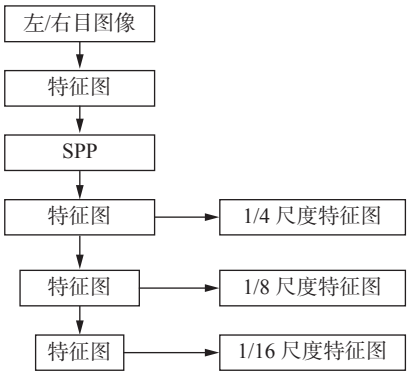
2.1　特征提取模块

目前着眼于实时性的立体匹配网络，通常因为采用过分简化的特征提取模块，导致无法提供有效的高质量特征图。因为本文想利用不同分辨率尺度的特征图来完成视差回归，所以一个良好的特征提取框架是十分必要的。若仅仅进行简单的降采样以满足后续视差回归所需的输入大小，不但无法节省运算时间，反而会造成精度丢失。

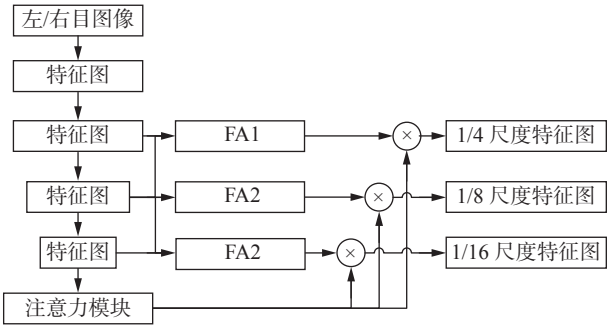
基于这样的想法，本文将语义分割领域以及光流领域常用的空间金字塔池化模块加入到了特征提取网络中。不同的是，本文为该模块输出的不同尺度特征图都赋予了可学习的参数，不但拓展了网络的感知域，还提供了参与融合的多尺度特征图所对应的权重，增强了场景通用性。为了减小运算负荷以及规避梯度消失问题，对本文提出的可学习空间金字塔池化模块进行了测试，最终选定仅仅在第一阶段降采样时采用该模块。本文提出的网络一共有三阶段降采样，获取了对应到原图大小的 1/4、1/8 和 1/16 的特征图。该可学习空间金字塔池化模块均为 2D 池化操作，模块中仅仅加入了少量的卷积核来引入可学习参数。整个特征提取框架中的卷积操作都采用蓝图可分离卷积进行，所以运行耗时短。本文的特征提取模块如图 2(a) 所示，并与几款当下较先进的立体匹配网络的特征提取模块做了比较。



(a) 本文



(b) PSM-Net



(c) AART-Stereo

图 2　特征提取模块

Fig. 2　Feature extraction module

为了可以充分发挥空间金字塔池化模块的性能，本文将经过该模块之后的每一层特征图都逐点相加到了下层的特征图上。在后续测试中可以发现，图 2(c) 的 AART-Stereo 网络由于过分简化了特征提取步骤，尽管有特征聚合模块（图中 FA），却仍然会导致较差的结果。

2.2　GDC 注意力模块

本文受到全局深度卷积和注意力模块功能的启发[19]，提出了一个基于全局深度卷积的通道端可学习注意力模块，如图 3 所示。

图 3　注意力模块

Fig. 3　Attention module

图 3 中，AVG 代表全局池化。第三阶段降采样的特征图经过第一次卷积后对每通道进行全局池化，将每个通道的特征图处理到1×1大小后，再采用1×1大小的卷积核进行全局深度卷积。由上文所述的全局深度卷积的特性，即输入通道数等于输出通道数等于卷积分组数，可以给该注意力模块上每个维度一个可学习的权重参数，用来体现不同通道特征的重要性。注意力模块随后被分别施加在 3 个阶段的降采样当中，用来体现每一阶段降采样结果的重要程度。

经上述处理之后，得到一$C\times1\times1$大小的向量，简化去除第二维即等价变换成$C\times1$大小的一维向量。因为本文想要将该向量应用到每一阶段的降采样当中，所以C需要满足$C=C_1+C_2+C_3$，其中C_i是第 i 阶段降采样后的通道个数，每一阶段的降采样结果只融合注意力模块内分割出的对应部分。

三阶段降采样特征提取结果将采用式 (1) 进行重新整合：

$$F_{\text{new}}=C(F_{\text{pre}})W_{\text{A}}+F_{\text{pre}} \tag{1}$$

式中：F_{pre} 代表融合注意力模块之前的特征图；F_{new} 是融合注意力模块之后的特征图；$C(*)$代表卷积

操作; W_A 为由注意力模块分割获得的权重。由于W_A的维度与F_{pre}的通道数相同, 式(1)中的乘法操作是由F_{pre}每个通道的张量乘以W_A内对应索引上的权重完成的, 给该通道上的所有值赋予一个权重。

2.3　动态代价与特征融合的 3D 卷积模块

本文提出了一个级联特征融合与动态修正代价融合的方法, 对特征提取阶段得到的不同尺度特征图进行级联融合, 级联方向从最小分辨率的特征图到最大分辨率的特征图。在底层, 由于没有更小尺度的特征提取结果, 匹配代价张量通过插值法进行回归构造, 完成视差结果的上采样。流程如图 4(a) 所示。

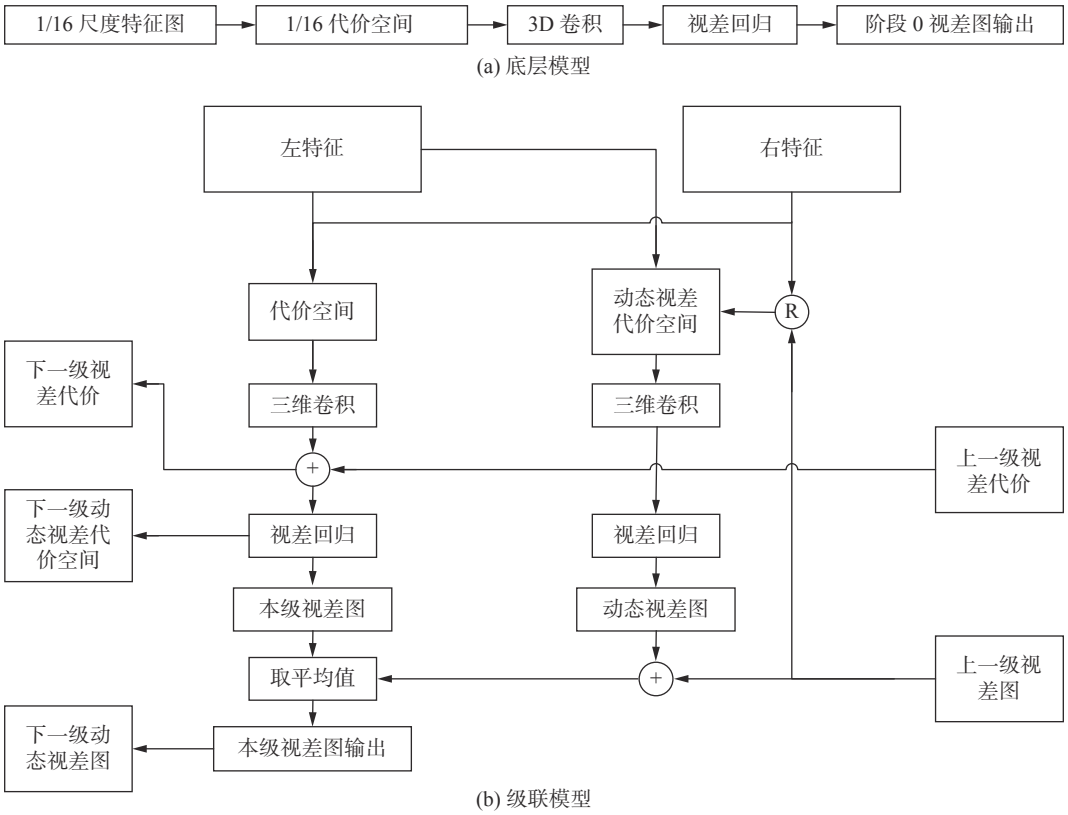


图 4　新型的 3D 卷积网络级联模型

Fig. 4　Novel 3D-CNN with cascade fusion module

非底层则使用级联模型, 首先采用与底层相同的方法得到完整的视差结果。与底层结构不同的是: 1) 上层将采用下层获得的视差结果进行均值融合, 如图 4(b) 所示; 2) 由于在视差估计任务中, 通常采用 3 像素点误差(3-pix err), 即用估计的视差值大小与真值大小相差 3 个像素坐标以上的像素点所占的比例来评判算法的性能, 所以本文将左右图特征以及下层得到的视差特征融合成一个代价张量, 并采用一维向量 [−2, −1, 0, 1, 2] 来处理该代价张量, 获取动态视差代价张量, 再进行插值回归, 如图 4(b) 所示。这样的操作将得到一个 5 通道的动态视差补偿结果, 每个通道张量的宽、高与原始图像大小相同。张量上每一点的值代表着该点动态视差补偿大小取 [−2, −1, 0, 1, 2] 其中某个值的概率。

本文将动态视差补偿结果与下层的视差输出结果进行叠加融合, 可以修正得到一个新的视差输出, 来达到修正视差输出的目的。最后, 将动态处理后得到的输出以及之前均值融合并插值回归得到的该层原始输出进行逐点相加再求均值, 就得到了本层的最终输出。

2.4　视差回归

当采用端到端的卷积神经网络进行立体匹配任务时, 一定会在得到最终视差结果之前得到一个降采样的代价张量, 通常这个代价张量如式 (2) 所示:

$$\boldsymbol{S} = BCDWH \tag{2}$$

式中: $\boldsymbol{S}$ 指的代价张量的大小; B 为批大小; C 为通道数; D 为视差等级, 即人为规定的视差上限值; W 与 H 分别代表张量的宽与高。

式(2)中第三维 D 表示视差维度, 共 192 维, 对 D 维的所有代价张量做 softmax 操作后, D_i 对应的$\boldsymbol{W}_i \times \boldsymbol{H}_i$张量上某点的数值大小将对应原图中该像素点视差大小为 i 的概率值。由此可以得到

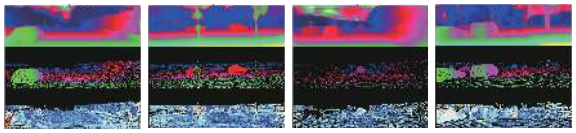
完整视差图：

$$D_{\text{res}} = \sum_{i=0}^{D}(i\text{softmax}(\textbf{Cost}_i)) \tag{3}$$

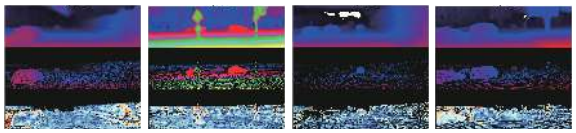
式中$\textbf{Cost}_i$代表每一个视差值的代价张量。动态视差补偿图的回归方式与式 (3) 回归方法同理。

3　实验与结果分析

本文在 KITTI 2015 Stereo[13] 数据集上完成了对上文所提出的网络中的模块的验证与测试，网络总体效果如图 5 所示，图中 (a) 与 (b) 两组图中的上、中、下 3 个部分分别代表网络输出结果图，真值图以及误点图。本文网络输出结果的平均 3 像素点误差率为 3.92%，AART-Stereo 网络输出结果的平均 3 像素点误差为 7.85%。



(a) 本文网络



(b) AART-Stereo

图 5　网络输出结果样例

Fig. 5　Output example of a network

3.1　数据集

KITTI 是真实城市环境的数据集，通过一辆家庭乘用轿车在城市道路下行驶采集数据并处理后得到。与电脑合成的数据集（如 SceneFlow Stereo 数据集[20]）相比，KITTI 数据集包含了更多的光照变化、明暗对比、过曝光等在现实世界中使用视觉传感器常出现的场景[21]。鉴于本文的目标是提出一个能够在智能驾驶领域使用的实时立体匹配网络，所以 KITTI 数据集是十分合适的一个选择。

KITTI 2015 Stereo 数据集包含了 200 对大小为 376×1 240 的用于训练的 RGB 双目图像对。相对应的，每组图像对都有一个稀疏的视差真值，该真值通过激光雷达获取并经过后处理得到。数据集中还有另外 200 对 RGB 双目图像对用作测试，这些图像对没有对应的真值，测试结果可以上传到 KITTI 官方进行评估。

3.2　实验设置

数据集与运行环境：训练集、验证集以及测试集的比例为 4∶1∶4。没有特殊说明时，网络运行在配置为 RTX 2 060 GPU，8 GB 内存，AMD 锐龙 4800H 的电脑上。

图像：输入为经过预处理的双目相机左右目图像，即纵向像素坐标已经对齐。为了满足网络运行要求，减轻运算负担，在图像送入网络前对其进行

裁剪：由于本文的网络采用三阶段降采样来输出 3 个不同分辨率尺度的结果，且底层的大小是原始输入的 1/16，所以经裁剪的图像长宽大小必须是 16 的整数倍。最终本文选择使用 256×512 大小的图像用作网络输入。

训练基本参数：本文在数据集上进行 300 个周期的训练，每个周期中采用 1 的批大小进行训练。

优化器：对于回归问题来说，Adam 优化器是最常用的优化器之一，它对训练过程中的学习率可以自适应地调整并且可以有效防止过拟合。本文根据 Adam 优化器作者的建议设置 Adam 优化器的参数$\beta_1 = 0.9, \beta_2 = 0.999$, weight decay = 0.000 1[22]。同时，由于发现有的作者会在使用 Adam 优化器的同时采用手动定义每个循环学习率的方式而不是使用自适应调整，本文会在后文对这两种方式进行比较，并加入常用的 SGD 优化器进行对比。

损失函数：由于本文的方法通过三阶段不同分辨率尺度进行插值回归并输出，所以需要将三阶段的输出损失相加来定义最终的损失大小。本文在默认 Adam 优化器下对均方误差损失 (MSE)、L_1 损失、平滑 L_1 损失进行了比较，总损失函数为

$$L = \alpha L_1(\text{St}_1) + \beta L_2(\text{St}_2) + \gamma L_3(\text{St}_3) \tag{4}$$

式中：St_i（i=1,2,3）代表第 i 尺度的输出；L_i是第 i 尺度所采用的损失函数。在计算总损失函数大小时，每一种损失函数不与其他损失函数共用，即在一种计算方法中，L_i的类别相同。为了强调每一阶段输出的重要程度，本文将每阶段的输出值做了加权处理，其中α、β、γ取值分别是 0.33、0.66、1。整网运算结果如表 1 所示。

表 1　损失函数测试

Table 1　Loss function test　　%

方法	精度
L_1 损失	91.33
MSE损失	90.58
平滑L_1损失	**94.67**
Huber损失	93.79

需要明确的是，尽管 Huber 损失与平滑 L_1 损失[23] 通常被认定为相同的损失函数（δ 通常取值为 1），但是 Pytorch 对这两种损失函数有不同的实现：

Huber 损失:

$$L(x)=\begin{cases}0.5x^2, & |x|\leqslant\delta=1\\ 0.5+|x|, & |x|>\delta=1\end{cases} \tag{5}$$

平滑 L_1 损失:

$$L(x)=\begin{cases}0.5x^2, & |x|\leqslant 1\\ |x|-0.5, & |x|>1\end{cases} \tag{6}$$

显然, Huber 损失与平滑 L_1 损失是 MSE 与 L_1 损失的折中方案, 其收敛速度快于 L_1 损失, 慢于 MSE。根据研究, Huber 损失与平滑 L_1 损失可以显著地缓解在回归问题中网络过拟合的情况, 并且收敛速度也能够接受[24]。

学习率: Adam 优化器作者推荐初始学习率为 0.001。如上文所述, Adam 优化器可以在训练过程中对学习率进行自适应的调整, 但部分作者仍采用自定义学习率的方式完成优化。本部分将对这两种情况进行对比, 并且加入最常用的 SGD 优化器进行自定义学习率的比较。自定义学习率: 在前 100 个周期中, 采用 0.001 的学习率; 在 101~200 个周期中, 采用 0.000 5 的学习率; 在 201~300 个周期中, 采用 0.000 05 的学习率。结果如表 2 所示。

表 2　学习率测试

Table 2　Learning rate test　　%

方法	精度
默认的Adam	**94.67**
手动调整学习率的Adam	91.58
手动调整学习率的SGD	92.44

可以看出, 对 Adam 优化器手动设置学习率很难使最终结果达到理想效果, 效果不如简单的 SGD 算法, 而 Adam 自适应调整的特性可以让模型更快地收敛到更精准的位置。

3.3　网络结构缺省性能测试与分析

如前文所述, 本文在第一阶段降采样中使用了可学习空间金字塔池化模块, 并且在后文的 3D 特征融合中也加入了复杂的动态融合机制。这些模块会引入相应数量的计算参数, 可能会影响网络的整体表现, 如运行时间、结果精度等。本文在同一环境下对比了未加入本文模块的网络, 以及加入了本文模块的网络, 评估了网络运算结果的 3 像素点误差。结果如表 3 和图 5 所示。

表 3 中 Ours-1 代表使用一次可学习金字塔池化, Ours-3 代表使用 3 阶段可学习金字塔池化。从表 3 可以注意到, 在第一阶段降采样加入了可学习空间金字塔池化模块之后, 网络整体运算精度得到大幅度提升。保留该模块, 再加入本文设计的轻量化 3D 特征融合与动态修正模块, 精度又出现了大幅度提升。鉴于可学习空间金字塔池化模块的良好性能, 尝试对 3 个阶段的降采样都采用该模块进行多尺度特征提取与融合。最终发现, 由于参数引进过多, 层数过深, 并且除蓝图可分离卷积以外又缺乏处理较深网络的其他结构, 导致精度出现了大幅衰减。

表 3　不同网络的精度与参数量测试

Table 3　Accuracy and parameter quantity of different networks

方法	3像素点误差/%	参数量/10^4
AART-Stereo	10.65	2.31
AART-Stereo+SPP	6.91	40
Ours-1	**3.45**	46
Ours-3	8.99	130

3.4　时间对比

在 3.3 节中, 对网络的整体运算精度进行了测验, 发现本文提出的网络有大幅度的精度提升。但是由于引入了额外的运算参数, 运行时间可能出现增加。本节比较了本文提出的网络与一些相似的端到端立体匹配网络的运算时长, 主要对比了在特征提取、视差回归以及网络整体所消耗的时间。所有的测试都以相同的训练环境与参数设定运行在 Nvidia RTX 2 060 GPU 上, 结果如表 4 所示。

表 4　耗时测试

Table 4　Time consumption test

方法	每步耗时/s		总耗时/s	3像素点误差/%
	特征提取	3D代价张量的计算与3D特征融合		
PSM-Net	0.03	0.70	0.73	1.86
AART-Stereo	0.01	0.02	0.03	6.91
AA-Net	0.01	0.05	0.06	3.74
Ours	0.01	0.05	0.06	3.45

可以发现, 每种方法耗时最多的步骤都是 3D 代价张量的计算与 3D 特征融合, 因为该操作涉及 3D 卷积网络的计算。PSM-Net 使用了一个全 3D 卷积网络的紧凑沙漏结构完成代价张量的计算与特征融合, 所以运算时间较长, 但特征融合完备, 运算精度高。本文提出的网络采用动态视差补偿机制以及多分辨率尺度级联互补机制, 可以在不损失大量精度的情况下替代部分 3D 卷积网络, 大幅度减少了运算参数量, 加速了运算过程。

3.5　KITTI 数据集结果分析

本文在 KITTI Stereo 2015 数据集上对该数据集提供的用于测试的 200 组 RGB 图像对进行了测试验证，结果如表 5 所示。

表 5　不同网络在 KITTI 数据集上测试

Table 5　KITTI stereo 2015 benchmark results

方法	D1-bg/%	D1-fg/%	D1-all/%	参考量M	运行时间/ms
AA-Net	1.39	5.39	2.55	4.00	1 500
DeepPruner-Fast	2.32	3.91	2.59	7.47	1 180
DispNetC	4.32	4.41	4.34	38.14	127
GC-Net	2.21	6.16	2.87	—	—
GANet-deep	1.485	3.46	1.81	—	—
GA-Net-15	1.55	3.76	1.95	—	—
AART-Stereo	6.37	13.95	7.54	0.023	75
本文方法	5.09	14.02	**6.58**	0.46	**100**

测试结果显示，通过改进特征提取方案，增加动态视差修复与多尺度结果融合后，本文提出的网络总 3 像素点误差为 6.58%，这个精度与当下较先进的网络相当，但需求的运算量更小。表中提到的 AA-Net 与 DeepPruner-Fast[25] 也都可以满足运行的实时性要求（<0.1 s/f），但根据 Chang 等的研究[12]，这些网络的实时性都需要高性能计算硬件的保证，无法运行在如 NVIDIA Jetson TX2 等车载运算设备上，难以满足车载需求。本文也将常见的网络部署到 TX2 设备上进行验证，运行结果如表 5 时间一栏所示。表 5 中未给出运行时长的方法意味着对应网络将无法在 TX2 计算环境中运行。由表 5 可知，本文提出的网络由于拥有较少的参数，可以部署在 NVIDIA Jetson TX2 上进行实时地运行。

4　结束语

本文提出了一个动态级联修正的实时立体匹配网络。该网络融合了三阶段降采样特征提取框架、基于全局深度卷积的注意力模块以及 3D 动态融合与修正模块，可以在短时间内提供高质量多尺度特征图，并区分各尺度重要性，同时完成实时的特征融合与输出修正以输出最终的立体匹配结果。与其他模型相比，本文的模型有更少的参数量以及较高的准确率，可以在车用计算平台上完成实时运算并满足车载使用精度要求。在未来，我们将探索人类先验知识在本文提出的网络中的应用，在降低网络计算复杂度的同时还可以强化网络的效果，为车载深度估计提供新的技术手段。

参考文献：

[1] MARTULL S, PERIS M, FUKUI K. Realistic CG stereo image dataset with ground truth disparity maps[J]. Technical report of ieice prmu, 2012(430): 117–118.

[2] HIRSCHMULLER H. Stereo processing by semiglobal matching and mutual information[J]. IEEE transactions on pattern analysis and machine intelligence, 2008, 30(2): 328–341.

[3] ZHANG Feihu, PRISACARIU V, YANG Ruigang, 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). Long Beach: IEEE, 2019 : 185−194.

[4] 张娣. 基于双目视觉的道路场景语义分割技术研究[D]. 南京: 南京理工大学, 2020.
ZHANG Di. Research on semantic segmentation of road scene based on binocular vision[D]. Nanjing: Nanjing University of Science and Technology, 2020.

[5] ŽBONTAR J, LECUN Y. Stereo matching by training a convolutional neural network to compare image patches[J]. Journal of machine learning research, 2016, 17: 1–32.

[6] SEKI A, POLLEFEYS M. SGM-nets: semi-global matching with neural networks[C]//2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu: IEEE, 2017 : 6640−6649.

[7] 王笛. 基于双目立体匹配的三维视觉方法研究 [D]. 西安: 西安理工大学, 2021.
WANG Di. Research on 3D vision method based on binocular stereo matching[D]. Xi'an: Xi'an University of Technology, 2021.

[8] 吴玉晗. 基于双目立体视觉的立体匹配算法研究 [D]. 成都: 电子科技大学, 2021.
WU Yuhan. Research on stereo matching algorithm based

on binocular stereo vision[D]. Chengdu: University of Electronic Science and Technology of China, 2021.

[9] 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. Venice: IEEE, 2017: 66−75.

[10] CHANG Jiaren, CHEN Yongsheng. Pyramid stereo matching network[C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018: 5410−5418.

[11] XU Haofei, ZHANG Juyong. Aanet: adaptive aggregation network for efficient stereo matching[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle: IEEE, 2020: 1959−1968.

[12] CHANG Jiaren, CHANG Peichun, CHEN Yongsheng. Attention-aware feature aggregation for real-time stereo matching on edge devices[C]//Asian Conference on Computer Vision. Cham: Springer, 2021: 365−380.

[13] MENZE M, GEIGER A. Object scene flow for autonomous vehicles[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Boston: IEEE, 2015: 3061−3070.

[14] KRIZHEVSKY A, SUTSKEVER I, HINTON G E. Imagenet classification with deep convolutional neural networks[J]. Communications of the ACM, 2017, 60(6): 84–90.

[15] SANDLER M, HOWARD A, ZHU M, et al. Mobilenetv2: inverted residuals and linear bottlenecks[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018: 4510−4520.

[16] HAASE D, AMTHOR M. Rethinking depthwise separable convolutions: how intra-kernel correlations lead to improved MobileNets[C]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle: IEEE, 2020 : 14588−14597.

[17] 郑秋梅, 温阳, 王风华. 基于注意力机制和可分离卷积的双目立体匹配算法 [J]. 微电子学与计算机, 2021, 38(5): 42–47.
ZHENG Qiumei, WEN Yang, WANG Fenghua. Stereo matching based on attention mechanism and separable convolution[J]. Microelectronics & computer, 2021, 38(5): 42–47.

[18] CHEN Sheng, LIU Yang, GAO Xiang, et al. MobileFaceNets: efficient CNNs for accurate real-time face verification on mobile devices[C]//Chinese Conference on Biometric Recognition. Cham: Springer, 2018: 428−438.

[19] 吴俊劼, 陈震, 张聪炫, 等. 基于特征级联卷积网络的双目立体匹配 [J]. 电子学报, 2021, 49(4): 690–695.
WU Junjie, CHEN Zhen, ZHANG Congxuan, et al. Binocular stereo matching based on feature cascade convolutional network[J]. Acta electronica sinica, 2021, 49(4): 690–695.

[20] MAYER N, ILG E, HAUSSER 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). Las Vegas: IEEE, 2016: 4040−4048.

[21] 龚伟, 秦岭, 任高峰, 等. 基于多维特征融合的双目立体匹配算法研究 [J]. 激光与光电子学进展, 2020, 57(16): 299–306.
GONG Wei, QIN Ling, REN Gaofeng, et al. Binocular stereo matching algorithm based on multidimensional feature fusion[J]. Laser & optoelectronics progress, 2020, 57(16): 299–306.

[22] KINGMA D, BA J. Adam: a method for stochastic optimization[EB/OL]. (2017−01−30)[2021−11−06]. https://arxiv. org/abs/1412.6980.

[23] REN Shaoqing, HE Kaiming, GIRSHICK R, et al. Faster R-CNN: towards real-time object detection with region proposal networks[J]. IEEE transactions on pattern analysis and machine intelligence, 2017, 39(6): 1137–1149.

[24] NATEKIN A, KNOLL A. Gradient boosting machines, a tutorial[J]. Frontiers in neurorobotics, 2013, 7: 21.

[25] DUGGAL S, WANG Shenlong, MA W C, et al. DeepPruner: learning efficient stereo matching *via* differentiable PatchMatch[C]//2019 IEEE/CVF International Conference on Computer Vision (ICCV). Seoul: IEEE, 2019 : 4383−4392.

作者简介:

何国豪, 硕士研究生, 主要研究方向为智能驾驶、智能系统视觉感知。

翟涌, 副教授, 主要研究方向为车辆电子控制。获得授权发明专利 5 项, 发表学术论文 10 篇。

龚建伟, 教授, 汽车研究所长, 主要研究方向为地面无人平台相关技术。主持国家级或省部级项目 10 余项, 授权发明专利 30 项。发表学术论文 13 篇, 参编专著和教材 5 部。