文章编号　1004-924X(2024)03-0445-11

空洞卷积和双边格网的立体匹配网络

张晶晶[1,2,3*]，杜兴卓[1,2,3]，支　帅[4,5]，丁国鹏[4,5*]

(1. 中国地质大学(武汉) 自动化学院，湖北 武汉 430074；
2. 复杂系统先进控制与智能自动化湖北省重点实验室，湖北 武汉 430074；
3. 地球探测智能化技术教育部工程研究中心，湖北 武汉 430074；
4. 中国科学院 微小卫星创新研究院，上海 201203；
5. 上海微小卫星工程中心，上海 201203)

摘要：为解决基于深度学习的立体匹配方法面临着网络规模大、网络结构复杂等问题，提出了一个网络规模较小、精度较高的网络结构。该网络在特征提取模块删减修改了复杂冗余的残差层并引入了空洞卷积金字塔池化模块来扩大视野范围，提取更多有用的上下文信息；在代价计算模块中使用了三维卷积层以成本聚合提升立体匹配的精度；最后，在代价聚合模块引用了双边格网模块以较低分辨率的成本量来获取精度较高的视差图。将该网络在 KITTI 2015 数据集和 Scene Flow 数据集等主流数据集上进行实验，结果显示，相较于其他主流优秀网络类如金字塔立体匹配网络(Pyramid Stereo Matching Network，PSM-Net)，网络规模参数量减少了约 38%，并取得了较高的实验精度，其中 Scene Flow 数据集的终点误差(End-point Error，EPE)为 0.86，是一个同时兼顾速度与精度的立体匹配网络。

关　键　词：计算机视觉；立体匹配；人工神经网络；视差

中图分类号：TP394.1；TH691.9　　**文献标识码**：A　　**doi**：10.37188/OPE.20243203.0445

Atrous convolution and Bilateral grid network

ZHANG Jingjing[1,2,3*], DU Xingzhuo[1,2,3], ZHI Shuai[4,5], DING Guopeng[4,5*]

(1. *School of Automation, China University of Geosciences (Wuhan), Wuhan* 430074, *China*;
2. *Hubei Provincial Key Laboratory of Advanced Control and Intelligent Automation for Complex Systems, Wuhan* 430074, *China*;
3. *Engineering Research Center of Earth Exploration Intelligent Technology, Ministry of Education, Wuhan* 430074, *China*;
4. *Innovation Academy for Microsatellites of Chinese Academy of Sciences, Shanghai* 201203, *China*;
5. *Shanghai Microsatellite Engineering Center, Shanghai* 201203, *China*)
* *Corresponding author, E-mail: dinggp@microsate.com*

Abstract: To address the challenges of large-scale and complex network structures in deep learning-based stereo matching, this work introduces a compact yet highly accurate network. The feature extraction module simplifies by removing complex, redundant residual layers and incorporating an Atrous Spatial Pyramid

收稿日期：2023-06-20；**修订日期**：2023-07-24.

基金项目：中国科学院国防科技创新实验室基金资助项目(No. CXJJ-19S012)；国家自然科学基金资助项目(No. 42001408)

Pooling (ASPP) module to broaden the field of view and enhance contextual information extraction. For cost calculation, three-dimensional (3D) convolutional layers refine stereo matching accuracy through cost aggregation. In addition, a bilateral grid module is integrated into the cost aggregation process, achieving precise disparity maps with reduced resolution demands. Tested on widely-used datasets like KITTI 2015 and Scene Flow, our network demonstrates a significant reduction in parameters by approximately 38% compared to leading networks like Pyramid Stereo Matching Network (PSM-Net), without compromising on experimental accuracy. Notably, it achieves an end-point error (EPE) of 0.86 on the Scene Flow dataset, outperforming many top-performing networks. Thus, our network effectively balances speed and accuracy in stereo matching.

Key words: computer vision; stereo matching; artificial neural network; parallax

1 引　言

立体匹配是三维重建的关键步骤。随着计算机视觉技术和图像处理技术的快速发展,基于双目视觉图像的立体匹配技术在三维建模、机器人视觉导航、增强现实以及汽车自动驾驶领域得到了广泛的应用[1-2],其思想是通过估计校正后的立体图像对中同一水平线上像素,找到空间像素的对应关系。

近年来,基于深度学习的方法在立体匹配领域展示出巨大的潜力[3]。相对于传统方法,三维卷积神经网络(Three-Dimensional Convolutional Neural Networks,3D CNN)可以大幅度地提升立体匹配的精度,也有众多优秀网络脱颖而出。目前,基于3D CNN的立体匹配方法面临的主要问题就是如何以尽可能小型的网络利用到更多的上下文信息。Geometry and Context Network (GC-Net)[4], 金字塔立体匹配网络(Pyramid Stereo Matching Network, PSM-Net)[5]以及Guided Aggregation Net(GA-Net)[6]则采用不同的模型实现了较高精度的立体匹配。Attention Concatenation Volume Network(ACV-Net)[7]采用一种新颖的注意力权重代价体构建方法,设计的立体匹配网络精度得到了大幅度提升。Cascaded Recurrent Network(CRE-Net)[2]设计一个层次网络以提取更细致的特征,同时提出自适应的群体关联层来减轻错误校正的影响。尽管这些网络都具有较高的精度,但是网络庞大、消耗大以及实时性差仍是不可忽略的问题。

为解决这些问题,本文提出了一种精度较高且较为轻量的立体匹配网络,称之为基于空洞卷积和双边格网的立体匹配网络(Atrous convolution and Bilateral grid Network, AB-Net)。首先,使用一个简化的残差模块以略微降低网络精度的代价大幅度缩减网络规模;其次,采用空洞卷积的池化金字塔模块(Atrous Spatial Pyramid Pooling,ASPP)[8]来进一步提升感受视野,目的是提取更多的上下文细节信息以提升立体匹配精度;最后,在引用堆叠沙漏的3D CNN模块的同时,在网络中引入双边网络模块[9]以整合各个维度的图像特征并建立其对应关系,从而在保证网络精度的同时进一步提高效率。AB-Net在KITTI 2012、KITTI 2015数据集[10]以及Scene Flow数据集[11]都实现了较高的精度。

2 相关工作

近年来,基于深度学习的立体匹配取得了飞速发展。Kendall等[4]提出了GC-Net,该网络是一个使用3D卷积层进行匹配代价计算端到端的网络,使用编码器-解码器的架构来合并多尺度的特征以实现代价聚合。为了更加有效地利用上下文信息,PSM-Net[5]使用空间金字塔池(SPP)模块来集成不同尺度的特征,并使用堆叠的沙漏结构3D卷积层进行成本聚合,有效地提高了立体匹配精度。Zhang等[6]提出的GA-Net使用了两个新的神经网络层,进一步提升了立体

匹配精度。Xu[7]提出了多级自适应补丁立体匹配，以提高匹配成本在不同差异下的显著性，进而提升立体匹配精度。为了更好地恢复深度细节，Li[8]提出 CRE-Net，该网络利用多次细化特征、叠加的级联结构以及自适应的群体关联层，以提高精细细节周围的渲染结果。同年，Wang[3]提出了一种不确定性估计方式，它从概率分布中学习相关结果，可以量化不确定性，加入到目前主流的立体匹配网络中以提升精度。

目前，精度已经趋于极限，学者们开始着手提升效率。为了追求实时性能，Stereo-Net[7]以低分辨率（例如 1/8 分辨率）使用 3D 卷积进行立体匹配计算，得到的网络能以 60 frame/s 的高帧率实时运行，但却降低了立体匹配的精确性。

3　基于空洞卷积和双边格网的立体匹配网络

本文提出的 AB-Net 包括用于有效合并全局上下文信息的 ASPP 的特征提取模块、用于代价聚合的堆叠沙漏模块以及双边格网模块。

3.1　网络体系架构

AB-Net 结构如图 1 所示。特征提取模块由 ASPP 模块和残差层组成，其作用是尽可能多地提取双目图像不同尺寸的特征；3D 卷积堆叠沙漏模块由多个 3D CNN 组成，其作用是聚合且正则化四维匹配代价卷的视差信息以及其余特征信息；双边格网模块的作用是对前序的数据进行切片操作，以低分辨率执行大部分计算，获得精度更高的视差图，最后进行上采样与视察回归计算即可完成双目立体匹配。

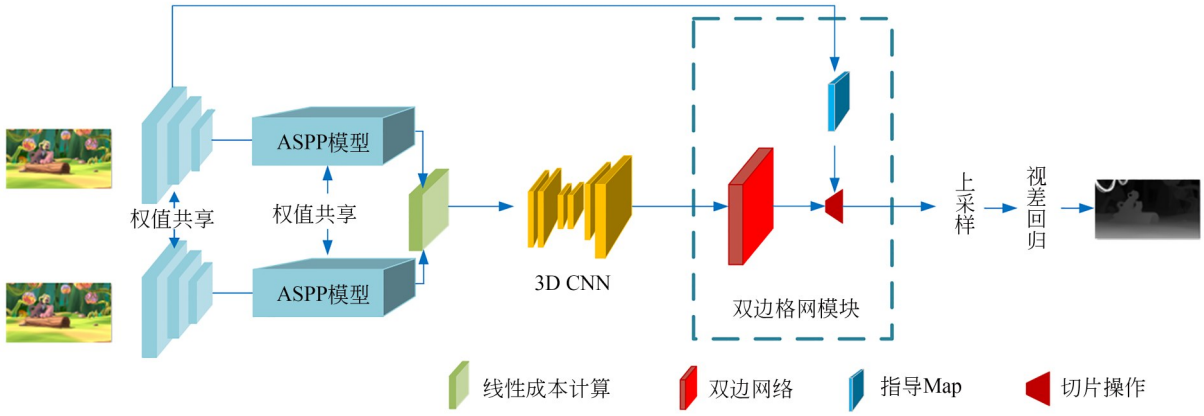


图 1　AB-Net 网络结构

Fig. 1　Architecture of AB-Net

3.2　特征提取模块

残差网络块可以在防止梯度消失的同时最大程度地提取图像特征信息，因此广泛应用在特征提取的任务中，但残差层的参数量巨大，非常消耗资源。与以往由较为复杂的残差层组成的特征提取模块不同，本文采用一个较为轻型的残差层来提取图像特征。最初的三层使用卷积核为 3×3，步长分别为 2，1，1 的三个卷积对输入图像进行下采样。然后使用步长为 1，2，2，1 的 4 个残差层，以 1/8 的分辨率快速生成图像的一维语义信息特征。PSM-Net 的参数量为 5 224 768 Byte，修改后参数量减少至 2 896 192 Byte，大幅缩减了网络的规模与体积。

单纯的依靠像素级别的特征来确定上下文之间的关系是不现实的，高效地利用物体周围的环境信息作为特征并加以提取则有助于一致性估计。由于 AB-Net 特征提取模块使用层数较少的残差层，感受视野的尺寸受限，后续必须使用感受视野更大的模块。本文采用 ASPP 结构以扩大网络的感受视野。

卷积层的感受视野受卷积核尺寸的影响，扩大其感受视野的方法主要有两种，分别是扩大卷积核的尺寸或者将多个小卷积核的卷积层级联。

它们都会扩大网络的规模，降低网络效率。与普通的卷积层相比，空洞卷积通过调整扩张率来扩大立体匹配网络的感受视野，其示意图如图 2 所示。

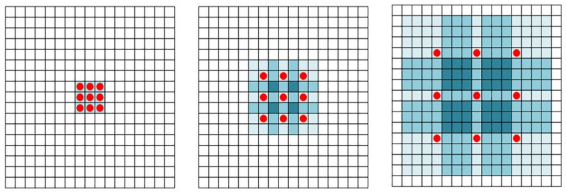


图 2　空洞卷积不同扩张率示意图

Fig. 2　Schematic diagram of atrous convolution with different expansion rates

ASPP 模块由空洞卷积组成，采用不同的扩张率(rate)：2，12，24，36 并采用上采样将由残差层输出的低维特征图像恢复到原始尺寸，再将 layer2，layer4 以及 branch1，branch2，branch3，branch4 进行级联操作。特征提取模块示意图如图 3 所示。

3.3　3D 卷积堆叠沙漏模块

与 PSM-Net 类似，为了聚合且正则化四维匹配代价卷的视差信息以及其余特征信息，本文采用 3D CNN 从多个维度提取特征信息。然后使用沙漏对称型架构，编码器的架构是 2 个步长为 2 的 3D CNN 卷积层，执行下采样操作；对称地，解码器的架构是 2 个步长为 2 的 3D CNN 反卷积层，执行上采样操作以恢复尺寸。编码器与解码器以跳跃方式连接。整体沙漏架构如图 4 所示。

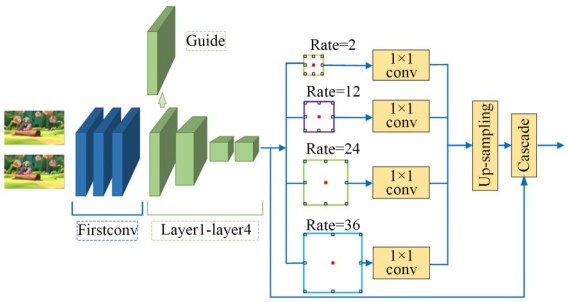


图 3　特征提取模块结构示意图

Fig. 3　Schematic diagram of feature extraction module structure

3.4　双边格网模块

为了搭建一种能够实现高精度立体匹配同时保持高效率的立体匹配网络，本文在 3D 卷积堆叠沙漏模块后使用一个基于双边网络的上采样模块，此模块通过双边网络处理的切片操作，以低分辨率执行大部分计算，同时还可以用高分辨率的成本量获得精度更高的视差图。该模块主要包括双边网络创建以及切片两个操作，将图像特征集合作为指导特征，对数据双边网络的低分辨率成本量进行切片操作，如图 5 所示。

对于双边网络的创建，首先输入一个低分辨率(本文采用的分辨率为 1/8)的四维成本量(包括宽度、高度、视差以及特征)，通过一个卷积核为 3 的三维卷积层即可转换为双边网络 B，包括宽度 x、高度 y、视差 d 以及指导特征 g 4 个维度，该双边网络表示为 $B(x,y,d,g)$。

对双边网络进行切片操作，目的是生成高分

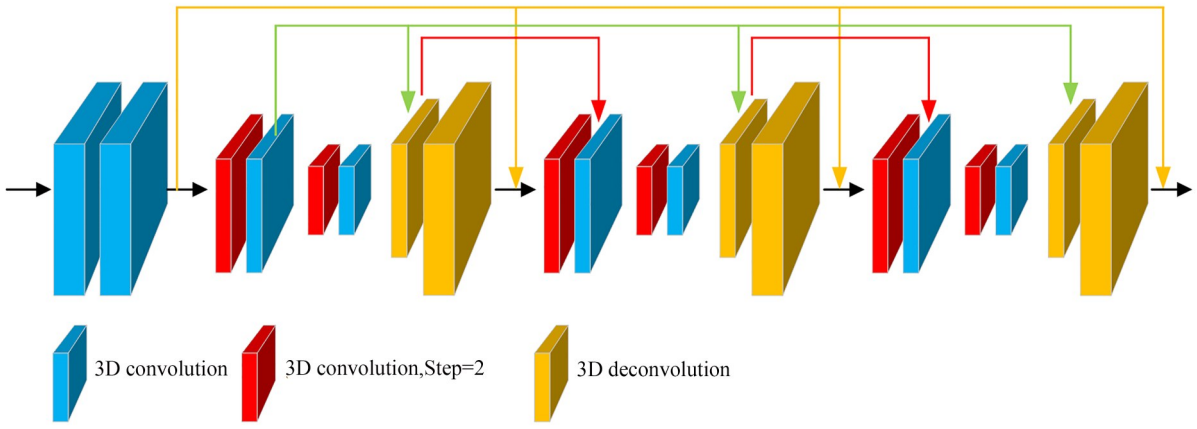


图 4　3D CNN 堆叠沙漏模块结构示意图

Fig. 4　Schematic diagram of stacked hourglass module of 3D CNN

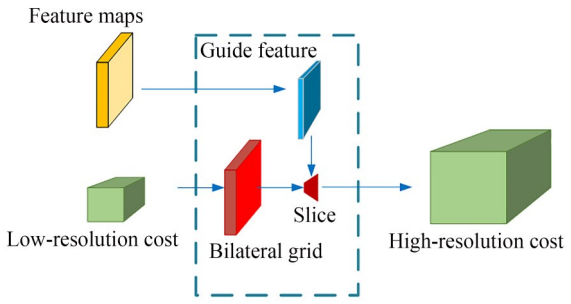


图 5　双边格网模块示意图

Fig. 5　Schematic diagram of Bilateral grid

辨率的成本量 C_{H}，操作过程如下：

$$C_{\mathrm{H}}(x,y,d)=B\left(sx,sy,sd,s_{\mathrm{G}}G(x,y)\right),\quad(1)$$

其中：G 是残差层输出的图像特征 maps 通过 1×1 的卷积层生成指导特征，$s\in(0,1)$ 是双边网络尺寸与高分辨率代价卷 C_{H} 的宽高比，$s_{\mathrm{G}}\in(0,1)$ 是双边网络的灰度值与指导特征 G 灰度值的比。

3.5　视差回归计算

这里使用 Soft Argmin[4] 方法，通过微分获得效果好的视差值图。经过网络处理后可获得每个图像在一定视差值范围内的匹配成本，成本越高表示越不匹配。然后，利用 Softmax 操作正则化可以算出各个图像属于一定区域内不同视差值的概率，通过加权求和可以得出各种像素点的平均视差值，即有：

$$D=\sum_{d=0}^{D_{\max}} d\times\mathrm{Softmax}\left(C_H(x,y,d)\right).\quad(2)$$

3.6　损失函数

本文采用的基础函数是 L_1 损失函数，该函数稳定性较强且对于数据异常的值不敏感，定义如下：

$$L\left(d,\hat{d}\right)=\frac{1}{N}\sum_{n=1}^{N}L_1\left(d_i-\hat{d}_i\right),\quad(3)$$

其中：N 是像素的数量，d_i 是真实的视差值，$\hat{d}_i$ 是预测的视差值。L_1 表达式如下：

$$L_1=\begin{cases}0.5x^2, & |x|<1\\ |x|-0.5, & |x|\geqslant 1\end{cases}.\quad(4)$$

4　实验与分析

为验证 AB-Met 网络的精度与性能，实验测试与分析在 Scene Flow 数据集和 KITTI 2015 数据集上进行。

4.1　数据集介绍

Scene Flow 立体匹配数据集是一个规模较大的人工数据集，由 35 434 对训练图像以及 4 370 对测试图像组成，其像素分辨率为 960×540，同时该数据集也为每对图像生成了一张高精度的稠密视差图作为真实值。在实验过程中，该数据集部分图像的部分像素视差值超过了本文所设定的最大视差，因此本文在计算误差与损失时将这部分忽略。

KITTI 数据集具体分为 KITTI 2012 以及 KITTI 2015，其是由德国卡尔斯鲁厄理工学院和丰田美国技术研究院联合创办，在真实的道路场景下利用双目相机以及激光雷达等设备获取的立体匹配数据集。KITTI 2012 提供 194 个训练和 195 个测试图像，KITTI 2015 提供 200 个训练和 200 个测试图像，每幅图像的像素分辨率为 1 240×376。该数据集还利用激光雷达为每对图像生成了一张高精度的稀疏视差图作为真实值。本文将训练集的图像数据随机划分为训练集（70%）与测试集（30%）。

Middlebury[12] 数据集是计算机视觉领域中广泛使用的一个数据集，包含多个场景下的图像序列和对应的视差图，共计 15 对训练图像与 15 对测试图像。

4.2　实验细节与性能指标

实验在 Ubuntu18.04 环境下，采用 Pytorch 深度学习框架，图形处理器为 NVIDIA GeForce 3090 完成 AB-Net 的训练与测试。在训练过程中采用了 Adam 优化器[10]，其延迟率参数设置分别为 $\beta_1=0.9$，$\beta_2=0.99$。所有训练数据的尺寸均设置为 512×256。对于 Scene Flow 数据集，本文以 0.001 的学习率训练 30 轮，再以 0.000 1 的学习率训练 10 轮；对于 KITTI 2015 数据集，由于其图像对较少，直接从 0 开始训练易导致网络过拟合，因此本文采用迁移学习的方法，将 Scene Flow 训练好的模型作为 KITTI2012 和 KITTI 2015 预训练的模型并对网络进行微调，先以 0.001 的学习率训练 200 轮，再以 0.000 1 的学习率训练 100 轮。

Scene Flow 数据集训练时，本文采用终点误差（End-point Error，EPE）作为评价指标。EPE

越大，匹配率越低。其表达式如下：

$$E_{\mathrm{EP}}=\frac{1}{N}\sum_{i\in N}\sqrt{\left(d_i-\hat{d}_i\right)^2}. \tag{5}$$

KITTI 2015数据集训练时，本文采用3像素误差(3px-Error)作为评价指标。3px-Error越大，匹配率越低。其表达式如下：

$$E_{3\mathrm{px}}=\frac{1}{N}\sum_{i\in N}\Phi\left(\left|d_i-\hat{d}_i\right|\right), \tag{6}$$

其中：

$$\Phi(a,b)=\begin{cases}1, a>b\\0, a\leqslant b\end{cases}, \tag{7}$$

式中：N是像素的数量，d_i是真实的视差值，$\hat{d}_i$是预测的视差值。

4.3 消融实验

AB-Net的基准网络为PSM-Net网络，首先对优化后的残差模块进行测试，然后对引入ASPP模块的网络进行测试，最后对引入双边格网的模块进行测试，并与PSM-Net进行对比。其中，Res-CV表示构建的成本量的分辨率，EPE为Scene Flow数据集的测试指标，结果如表1所示。

表1 消融实验结果

Tab. 1 Results of ablation experiment

网络名称	残差层简化	ASPP	双边格网	Res-CV	E_{pe}/pixel	t/ms
PSM-Net				1/4	1.09	2 310
AA-Net				1/4	0.87	1 147
	√			1/4	1.16	1 125
AB-Net	√	√		1/4	1.01	1 856
	√	√	√	1/8	**0.86**	**951**

由表1可知，仅进行残差层优化后的模型EPE提升为1.16，运行时间缩短约50%，仍能取得不错的精度；在引入ASPP模块后，随着更多的特征细节信息被提取，修改后网络的EPE进一步下降至1.01，但运行时间在增大；引入双边格网模块后，成本量的体积缩减为原尺寸的1/8，并对其加速处理，精度大幅度提升的同时网络的运行时间也大幅缩减。AB-Net的运行时间低于PSM-Net和AA-Net的运行时间。PSM-Net，AA-Net的EPE分别是1.09和0.87，AB-Net的EPE是0.86，误差下降了约21%和1%。

4.4 与其他算法对比结果

本文将AB-Net的测试结果与GC-Net[4]，PSM-Net[13]，CRL，AA-Net[14]和AED-Net[15]进行比较，首先在Scene Flow数据集上进行测试，结果如表2所示。可以看出，由于AB-Net对残差层进行了大量删减，网络规模相较于其他网络缩小很多，同时引入ASPP模块以及双边格网模块来保证网络具有较高的精度。与基准网络PSM-Net相比，AB-Net的网络规模参数量减少了约38%，立体匹配精度提升了约21%。

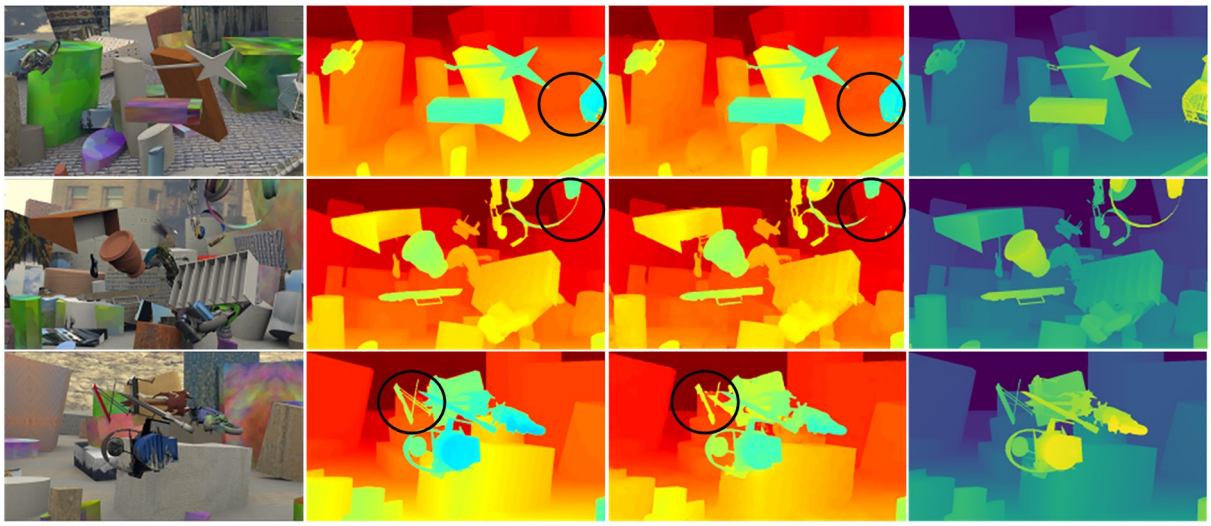
图6展示了3个测试实例，从图像中可以看出，AB-Net在非常复杂、重叠的场景下也能获取精准的视差图，并且在一些细节上的表现比PSM-Net更加优秀(见图中黑色圆圈部分)。

表2 SceneFlow测试集结果

Tab. 2 Result of different methods on SceneFlow dataset

网络名称	E_{pe}/pixel	Number of parameters/10^8
GC-Net[4]	2.51	3.50
PSM-Net[5]	1.09	5.20
CRL[16]	1.32	78.77
AA-Net[17]	0.87	—
AED-Net[18]	0.89	—
AB-Net	**0.86**	**3.20**

其次测试KITTI 2015数据集，将200对图像输入网络中得到预测的视差图，上传至KITTI官网以评估分析，并与其他网络进行比较，结果如表3所示。其中，D_1表示视差图中错误匹配点所占的比例，b_{g}表示背景区域，f_{g}表示前景区域，all表示整张视差图的全部区域。由表4可以看出，AB-Net在全部区域的匹配错误率均为最低，为2.26%；同时所有像素的前景区域、背景区域，以



(a) 图像对的左视图 (a) Left view of the image pair　(b) AB-Net预测的视差图 (b) Disparity map predicted by AB-Net　(c) PSM-Net预测的视差图 (c) Disparity map predicted by PSM-Net　(d) 真实视差图 (d) True disparity map

图 6　SceneFlow 数据集上不同算法的结果对比

Fig. 6　Result of different methods on SceneFlow dataset

表 3　KITTI 2015 双目立体匹配数据集测试结果

Tab. 3　Test result on KITTI 2015 binocular stereo matched dataset

网络名称	All Pixel/%			Non-Occluded Pixels/%			Number of parameters/10^8
	$D_{1\text{-bg}}$	$D_{1\text{-fg}}$	$D_{1\text{-all}}$	$D_{1\text{-bg}}$	$D_{1\text{-fg}}$	$D_{1\text{-all}}$	
GC-Net[4]	2.21	6.16	2.87	2.02	5.58	2.61	3.50
PSM-Net[5]	1.86	4.62	2.32	1.71	4.31	2.14	5.20
CRL[13]	1.99	5.39	2.55	4.93	2.32	1.89	78.77
AA-Net[14]	2.48	3.59	2.67	2.32	3.12	2.45	—
AB-Net	1.91	4.34	2.26	1.82	4.17	2.11	**3.20**

表 4　KITTI 2012 双目立体匹配数据集测试结果

Tab. 4　Test results on KITTI 2012 binocular stereo matched dataset

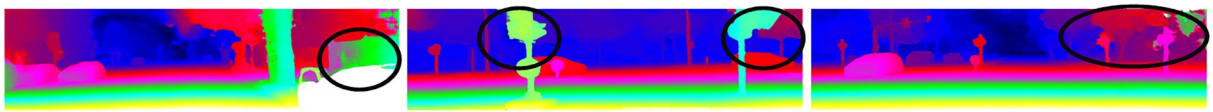
网　络	O_{Noc}/%	O_{All}/%	A_{Noc}/%	A_{All}/%	Number of parameters/10^8
GC-Net[4]	1.77	2.06	0.6	0.7	3.50
PSM-Net[12]	1.49	**1.89**	**0.5**	**0.6**	5.20
AA-Net[14]	2.54	3.18	0.6	0.7	—
AED-Net[15]	3.40	4.11	0.7	0.8	—
AB-Net	**1.44**	2.45	0.6	0.7	**3.20**

及非遮挡像素的前景区域、全部区域的误差也较低，分别是 1.91%，4.34%，1.82% 以及 2.11%；而非遮挡像素的背景区域误差则较高，为 4.17%。

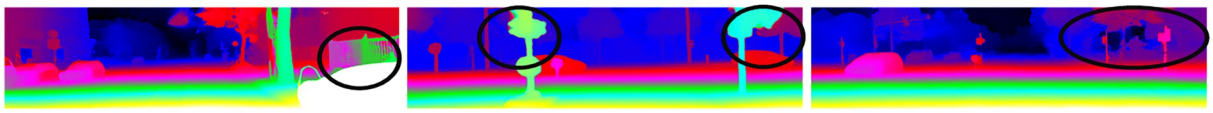
图 7 展示了 PSM-Net，GC-Net 以及 AB-Net 的预测视差图效果对比以及 AB-Net 的预测误差图。可以清楚地看到，AB-Net 的预测效果在细节方面相较于其他网络更胜一筹，能够更清晰地展示复杂背景、路灯以及栅栏的深度信息和清晰的轮廓（见图中黑色圆圈部分）。

最后测试 KITTI 2012 数据集，将 KITTI 2012 测试集的 195 对图像输入网络中得到预测的视差图，上传至 KITTI 官网以评估分析，并与其他网络进行比较，结果如表 4 所示。其中，O_{Noc}

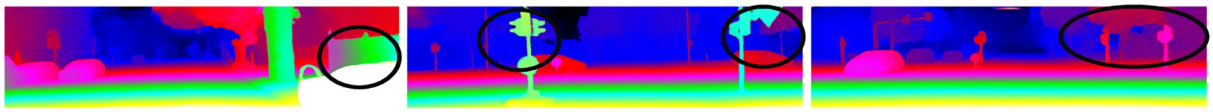
(a) 左视图
(a) Left image



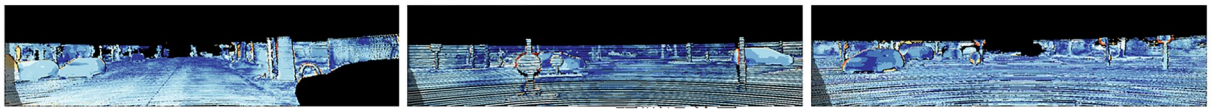
(b) PSM-Net预测视差图
(b) Result of PSM-Net



(c) GC-Net预测视差图
(c) Result of GC-Net



(d) AB-Net预测视差图
(d) Result of AB-Net



(e) AB-Net预测误差图
(e) Error image of AB-Net

图 7　不同方法在 SceneFlow 数据集上结果对比

Fig. 7　Result of different methods on SceneFlaw dataset

和 O_{All} 分别表示非遮挡区域和整个区域的视差图中误匹配点所占的比例；A_{Noc} 和 A_{All} 分别表示非遮挡区域和整个区域的视差图中匹配点的平均误差。

由表 4 可知，本文网络在特征提取模块对残差层进行删减，减少了网络的参数量，相较于其他网络而言，在网络规模上有了较大缩减，非遮挡区域的视差图中误匹配点所占的比例（O_{Noc}）为 1.44%，在几个网络中排名第一。AB-Net 在网络规模上有一定的优越性，与其他考虑运行效率、注重实时性的网络相比，本文网络的精度略高。实验表明，本文所提网络在保证精度的情况下能高效实现立体匹配。

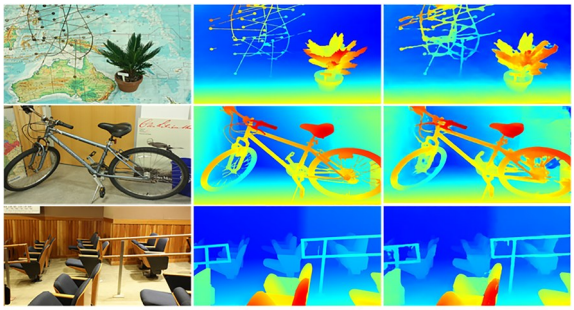
为进一步验证 AB-Net 网络的精度，本文在 Middlebury 2014 数据集[18]上进行评估。将该数据集中的 15 对训练集图像在上述网络模型中进行微调，测试结果如表 5 所示。其中，Bad2.0 指的是绝对误差大于 2 像素的点的百分比，该值越低表示网络对该数据集的预测能力越好。预测的视差图如图 8 所示。

表 5　Middlebury 2014 数据集测试分析

Tab. 5　Result on middlebury 2014 dataset analysis

网　络	Bad2.0/%
PSM-Net[5]	18.58
GA-Net[6]	17.43
AA-Net[13]	15.94
AB-Net	**7.56**

由表 5 可知，AB-Net 在该数据集上的 Bad2.0 误差为 7.56%，相较于 PSM-Net 的 18.58%、GA-Net 的 17.43% 均有较大的提升。通过量化分析说明 AG-Net 对于风格迥异的数据集拥有较强的预测能力。



(a) 左视图　(b) AB -Net　(c) PSM -Net
(a) Left image

图 8　Middlebury 2014 测试结果

Fig. 8　Test results on Middlebury 2014

图 8 展示了利用 PSM-Net 和 AB-Net 对未经训练的双目视觉图像的视差预测结果。可以清晰地看出，面对复杂地图背景的场景、杂物场景以及景深较大的教室场景，本文网络均能较好地生成高质量、边缘清晰、层次分明、深度信息一目了然的视差图，而 PSM-Net 的表现一般，场景物体边缘不清晰，图像边缘也出现了大量的匹配错误。

4.5　泛化实验

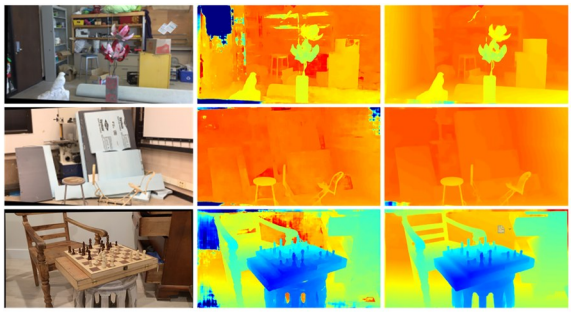
为验证 AB-Net 的泛化能力，本文在训练网络时，仅对 Scene Flow，KITTI2015 以及 KITTI2012 数据集训练后，就直接对 Middlebury 2014 数据集[18]进行预测评估。同时与 PSM-Net，GA-Net 和 AA-Net 进行泛化性测试对比，结果如表 6 所示。其中，Bad2.0 指的是绝对误差大于 2 像素的点的百分比，该值越低表示网络对该数据集的泛化性越好。预测的视差图如图 9 所示。

表 6　Middlebury 2014 数据集泛化能力数据分析

Tab. 6　Analysis of generalization ability data on Middlebury 2014 dataset

网　络	Bad2.0/%
PSM-Net[5]	24.8
GA-Net[6]	19.1
AA-Net[13]	18.7
AB-Net	**17.4**

由表 5 可知，AB-Net 在 Middlebury 2014 数据集上的 Bad2.0 误差为 17.4%，相较于 PSM-Net 的 24.8%、GA-Net 的 19.1% 和 AA-Net 的 18.7%，分别降低了 7.4%，1.7% 和 1.3%。通过量化分析说明 AG-Net 对于风格迥异、未经训练的数据集拥有较强的泛化能力。



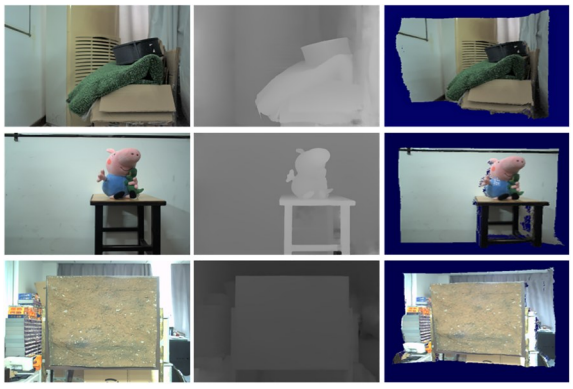
(a) 左视图　(b) AB-Net　(c) PSM-Net
(a) Left image

图 9　未经训练 Middlebary 2014 数据集预测结果对比

Fig. 9　Untrained results on Middlebury 2014 dataset

4.6　现实场景实验

本文使用双目相机对现实不同复杂度的场景进行拍摄，将获取到的双目图像输入训练好的网络中进行视差预测，效果如图 10 所示。



(a) 左视图　(b) AB-Net　(c) PSM-Net
(a) Left image

图 10　现实场景实验结果

Fig. 10　Experimental results of real scene

图 10 展示了利用 AB-Net 对现实场景进行视差预测并利用预测的视差图得到的三维重建结果。可以清晰地看出，面对杂物场景、复杂玩偶场景以及复杂地形场景，AB-Net 能较好地生成高质量、边缘清晰、层次分明的视差图，利用生成的视差图进行三维重建的结果也与现实场景无异。由此说明，AB-Net 拥有较强的泛化能力，面对复杂的、未经训练的现实场景也能取得较好的效果。

5 结　论

本文提出了一种AB-Net立体匹配网络。该网络通过精简冗余的残差层和引入ASPP模块，能够在保持较小网络规模的同时，扩大感受视野，提取更多细节信息，并获取足够的上下文信息。此外，本文还采用3D卷积层来提高立体匹配的准确性，并引入双边格网模块，在较低分辨率的成本量下获取更精确的视差图。

本文在KITTI 2012，KITTI 2015，Scene Flow及Middlebury 2014数据集上对AB-Net进行测试，结果显示与PSM-Net等立体匹配网络相比，AB-Net在参数量减少38%的情况下仍能保持较高的实验精度。对于KITTI 2015数据集，AB-Net在全部区域的匹配错误率为2.26%；对于KITTI 2012数据集，非遮挡区域的视差图中误匹配点所占比例为1.44%；而在Scene Flow数据集上，终点误差(EPE)为0.86。对于Middlebury 2014数据集，AB-Net也表现出较强的预测能力，Bad2.0误差为8.56%，优于对比网络。此外，使用AB-Net对现实场景数据进行预测，并获得了边缘清晰、深度信息明确且无遮挡的视差图，验证了AB-Net的高准确性和泛化性能。

参考文献：

[1] WEI H, MENG L J. An accurate stereo matching method based on color segments and edges[J]. *Pattern Recognition*, 2023, 133: 108996.

[2] 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*. IEEE, 2022: 16263-16272.

[3] WANG C, WANG X, ZHANG J W, *et al*. Uncertainty estimation for stereo matching based on evidential deep learning [J]. *Pattern Recognition*, 2022, 124: 108498.

[4] 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*. IEEE, 2017: 66-75.

[5] 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*. IEEE, 2018: 5410-5418.

[6] 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*. IEEE, 2019: 185-194.

[7] 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*. IEEE, 2022: 12981-12990.

[8] CHEN L C, PAPANDREOU G, KOKKINOS I, *et al*. DeepLab: semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs[J]. *IEEE Trans Pattern Anal Mach Intell*, 2018, 40(4): 834-848.

[9] 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*. IEEE, 2021: 12497-12506.

[10] MENZE M, GEIGER A. Object scene flow for autonomous vehicles[C]. 2015 *IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*). *June* 7-12, 2015. *Boston*, *MA*, *USA*. IEEE, 2015: 3061-3070.

[11] KINGMA D P, BA J. Adam: A method for stochastic optimization [J]. *arXiv preprint* arXiv: 1412.6980, 2014.

[12] 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, eds. Lecture Notes in Computer Science. Cham: Springer International Publishing, 2014: 31-42.

[13] KHAMIS S, FANELLO S, RHEMANN C, *et*

al. StereoNet: guided hierarchical refinement for real-time edge-aware depth prediction[J]. 2018.

[14] 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*. IEEE, 2017: 887-895.

[15] 杨戈，廖雨婷. 基于AEDNet的双目立体匹配算法[J]. 华中科技大学学报(自然科学版)，2022，50(3)：24-28.
YANG G, LIAO Y T. Algorithm of binocular stereo matching based on AEDNet[J]. *Journal of Huazhong University of Science and Technology* (*Natural Science Edition*), 2022, 50(3): 24-28. (in Chinese)

作者简介：

张晶晶(1987－)，男，湖北荆州人，副教授，2014年于中国科学院大学获得博士学位，主要从事计算成像、图像处理方面的研究。E-mail：jingjingzhang@cug.edu.cn

通讯作者：

丁国鹏(1986－)，男，江西兴国人，副研究员，2015年于中国科学院大学获得博士学位，主要从事智能载荷设计、图像处理方面的研究。E-mail：dinggp@microsate.com