文章编号：0253-374X(2024)10-1640-09

DOI：10.11908/j.issn.0253-374x.22483

基于两阶段代价矩阵和动态注意力的双目立体匹配网络

王志成，王泽灏

（同济大学 电子与信息工程学院，上海 201804）

摘要：目前大多数先进的双目立体匹配网络通过构建4D代价矩阵以保留图像的语义信息，增加了网络的计算量开销。为了解决上述问题，提出了两阶段的组合代价矩阵和多尺度动态注意力的EDNet++网络。首先从全局的、粗粒度的视差搜索范围上构建的基于相似度的代价矩阵作为引导，在局部的搜索范围上实现细粒度的组合代价矩阵，其次提出基于残差的动态注意力机制，其根据中间结果信息自适应地生成空间上的注意力分布，并且通过迁移实验证明了该方法的有效性，最后在各大公开数据集上的对比实验结果表明，相较于其他方法，EDNet++方法能够达到算法精度和实时性的良好平衡。

关键词：双目立体匹配；神经网络；注意力机制

中图分类号：TP391.4　　**文献标志码**：A

EDNet++: Improving Stereo Matching with Two-Stage Combined Cost Volume and Multiscale Dynamic Attention

WANG Zhicheng, *WANG Zehao*

(College of Electronics and Information Engineering, Tongji University, Shanghai 201804, China)

Abstract: Most state-of-the-art stereo matching networks construct 4D cost volume to preserve the semantic information of the image, which increases the computational cost of the network. To solve this problem, a network named EDNet++ with a two-stage combined cost volume and a multi-scale dynamic attention is proposed. First, a correlation cost volume is constructed based on global and coarse-grained disparity search range, which is used as a guide to construct a fine-grained combined cost volume on the local disparity search range. Then, the dynamic attention mechanism based on residuals can adaptively generate spatial attention distribution according to the intermediate result information, and the effectiveness of this method is proved by the transfer experiment. The comparison experiments on various public data sets show that EDNet++ can achieve a good balance between accuracy and real-time performance compared with other methods.

Keywords: stereo matching; neural network; attention mechanism

准确快速的深度估计对于机器人导航、三维重建、自动驾驶等应用具有重要意义。双目立体匹配作为一种深度估计的算法，被广泛运用在各个领域。双目立体匹配任务是在左右视图间进行相同像素的匹配，并且计算相同像素的距离偏差(俗称视差)。

传统方法则以人工设计的特征提取和匹配成本聚合算法为主，对于那些无纹理和重复纹理的区域匹配效果较差。卷积神经网络(CNNs)已经被广泛采用来克服传统立体匹配中的难点。文献[1-4]通过构建4D代价矩阵和三维卷积实现了最高的精度。虽然4D代价矩阵可以保留丰富的语义信息，但是它显著提高了计算量，并且很难达到实时的性能。文献[5]通过二维卷积进行左右图相似度的特征聚合。文献[6-8]采用了类似的方法，在速度和准确性之间达到良好的平衡。然而，基于相似度的方法在每个视差维度上都只存在一个单一的特征维度，这种表达能力始终是薄弱的，如何合理地将基于相似度和拼接的代价矩阵相融合是一个重点的难题。

1　相关工作

经典的立体匹配网络流程包括特征提取、构建代价体积、特征聚合和视差计算4个步骤[9]。近年来，CNN被引入到立体匹配任务[10]中，引起了人们的广泛关注。一些从真实场景中收集的数据集，如KITTI、

收稿日期：2022-11-21
基金项目：中国国防基础研究项目(JCKY 2020206B03)
第一作者：王志成，教授，工学博士，主要研究方向为模式识别、计算机视觉。E-mail：zhichengwang@tongji.edu.cn
通信作者：王泽灏，硕士生，主要研究方向为双目立体匹配。E-mail：2033063@tongji.edu.cn

论文拓展介绍

Middlebury，经常被用于微调和作为测试基准。

1.1　代价匹配计算

基于CNN的匹配代价计算方法对立体匹配精度有很大的贡献。有2种常用的匹配代价计算方法。一种是使用二维[11]或一维[5]卷积运算层（称为相关层）。文献[6-7，12-13]采用了这种特征向量之间的内积。Liang等[6]建立了初始视差估计的相关体积，它遵循通过特征一致性学习的视差细化模块。另一种流行的计算匹配成本的方法是通过连接来自对面立体图像的对应特征，形成一个4D代价矩阵。这个方法首先可以在文献[1]中找到。文献[14]在Scene Flow数据集上表现最好，其引入了DenseNet[15]的思想，进一步改进了PSMNet[2]。Zhang等[4]提出了带有2个引导聚合层和15个3D卷积的GANet，以实现最先进的性能。Gu等[16]将代价矩阵分解为多个阶段的级联形式，使用上一阶段的粗粒度视差图，可以获得更高空间分辨率和更窄视差范围的成本体积。除了级联成本，Shen等[17]提出了一种融合的代价矩阵表示方法来处理较大的域差。

1.2　立体匹配网络的残差网络

在立体匹配任务中，残差学习策略被广泛用于精化视差估计[6][18]。Gidaris等人[19]提出了像素级图像标记的深度网络架构，将任务分为3个部分：① 检测初始像素级标签，② 用新标签替换标签，③ 改进新标签。Pang等[8]提出了一种级联残差学习方案，采用两阶段CNN，其中第2阶段通过产生残差信号来细化估计。Jie等[20]提出了一种循环模型来增强基于左右一致性检查的视差估计，这意味着网络为2个视图生成视差图，并在每个循环步骤识别不匹配区域。Yu等[21]构建的代价矩阵依赖于像素级深度残差，通过使用由粗到细的策略来细化深度图并改变深度搜索范围。Stucker等[22]专门构建了一个基于U-Net[23]的网络，通过回归残差校正来增强重建。为了满足实时推理的需要，AnyNet[24]根据应用的需要，采用残差学习策略灵活地进行输出视差估计。通过聚合边缘信息进行残差学习，构建多任务网络进行边缘检测和立体匹配。Cheng等[25]首次将神经结构搜索（NAS）应用于立体匹配任务。该算法用于自动确定传统立体匹配流程中的特征网和匹配网。

2　方法

2.1　网络结构

提出的EDNet＋＋网络结构如图1所示。借鉴DispNetC[5]的网络框架作为特征提取网络（backbone），在特征提取网络上，利用权值共享编码器网络中conv3的最后左、右特征映射，形成一阶段的代价矩阵。关于两阶段的代价矩阵构建的更多信息见第3.2节。提出的两阶段组合代价矩阵构建的细节见图1。利用二维卷积对合并后的矩阵进行聚合并对视差进行回归。在解码器部分，网络遵循由粗到细的策略，逐步细化视差。为了生成注意力感知的残差特征，网络中使用了动态注意力模块。基于注意力的动态注意力在图1的右下角得到了很好的说明。与具有6个输出尺度的DisNetC[5]不同，本网络结构将视差预测减少到4个尺度，删除全分辨率的1/16和1/32的预测。

2.2　两阶段组合代价矩阵

提出了两阶段的组合代价矩阵，通过一阶段的相似度矩阵得到粗粒度视差结果，在二阶段，在局部视差搜索范围的基础上使用相似度矩阵和4D代价矩阵结合的方式构建代价矩阵，见图2。

在第1阶段，给定一对立体图像特征f_{L}和f_{R}，相似度代价矩阵计算如式（1）：

$$C_{\mathrm{c}}(d,x,y)=\frac{1}{N}\left\langle f_{\mathrm{L}}(x-d,y),f_{\mathrm{R}}(x,y)\right\rangle \tag{1}$$

式中：$\langle x_1,x_2\rangle$为2个特征向量$\boldsymbol{x}_1$和$\boldsymbol{x}_2$的内积；N为输入特征的通道数，代价矩阵的形状为$N\times D\times H\times W$；$d$为视差搜索范围$D$内的某一个视差值，空间大小为$H\times W$。$f_{\mathrm{L}}$和$f_{\mathrm{R}}$分别代表左视角图像特征和右视角图像特征。

假设在一阶段视差估计网络中已经通过相似性代价矩阵和聚合网络获得了粗粒度的基准视差图s_{-1}。接下来需要设置初始平面数n和初始平面间隔Δn来自定义二阶段的视差搜索范围。相对视差范围Δd_k定义为

$$\Delta d_k=k\Delta n,k\in\{-\frac{n}{2},-\frac{n}{2}+1,\cdots,0,\cdots,\frac{n}{2}-1\} \tag{2}$$

基准视差图s_{-1}的尺寸H和W与特征图的f_{L}和f_{R}相同，下一步则根据基准视差s_{-1}对特征图进行变形，建立第2阶段的组合代价矩阵。其中第2阶段的代价矩阵的计算方法为

$$C_{\mathrm{b}}(\Delta d_k,x,y)=\frac{1}{n}M\{f_{\mathrm{L}}(x-\Delta d_k+s_{-1}(x,y),y),f_{\mathrm{R}}(x,y)\} \tag{3}$$

式中：M代表组合代价计算；$s_{-1}(x,y)$为坐标点(x,y)在基准视差图s_{-1}对应的视差值。该算法将特征点转移到基准视差上，并使用它作为每个像素搜索的

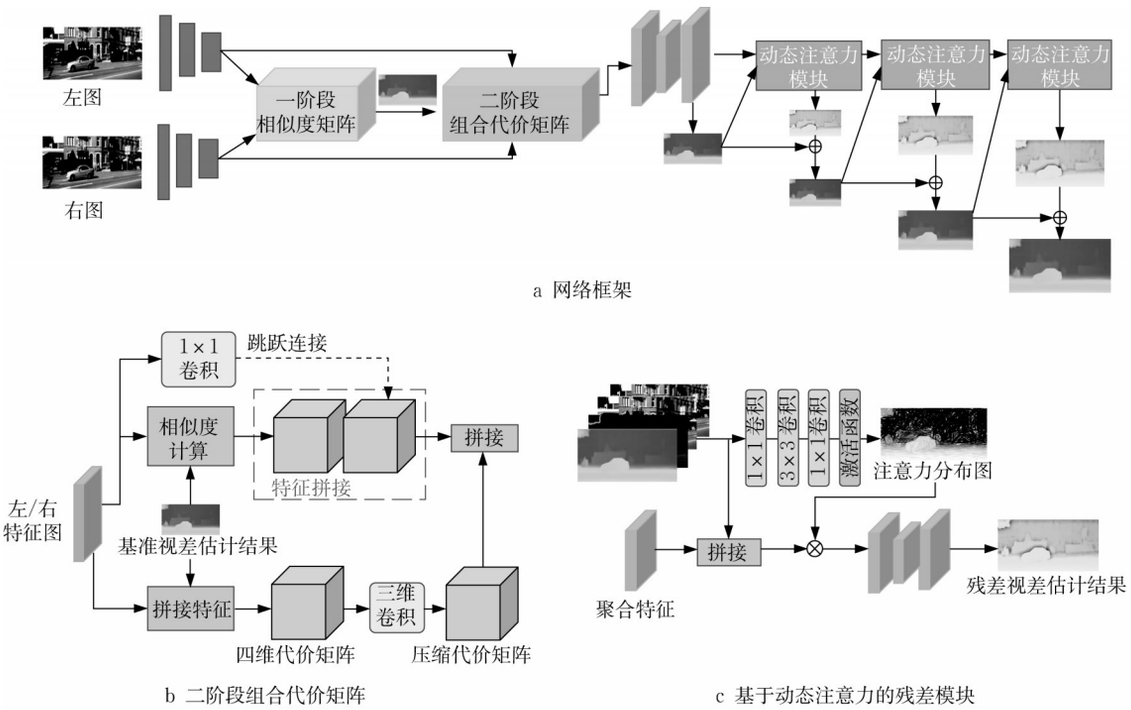


图1 EDNet++框架图总览

Fig. 1 Overview of proposed EDNet++

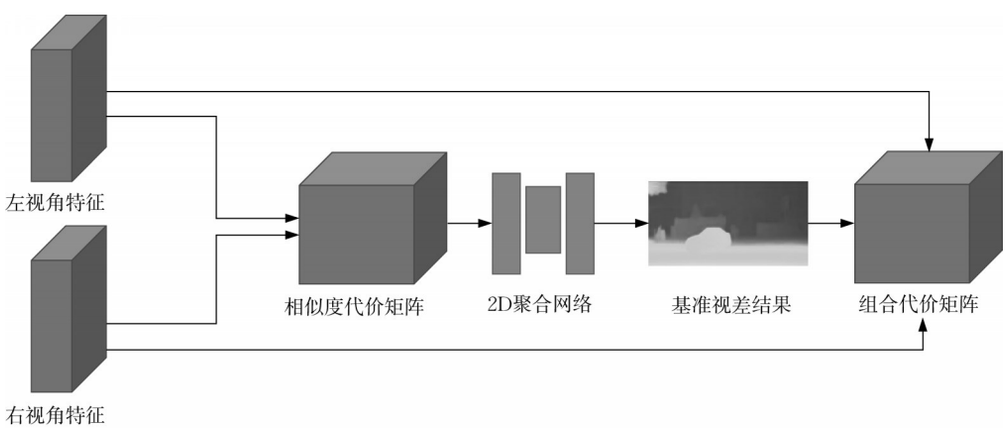


图2 两阶段代价矩阵构建流程

Fig. 2 Overview of two-stage cost volume construction

中心点,如图3所示。

在第2阶段,将基于拼接的4D代价矩阵信息融合进来,4D代价矩阵的构建方式定义为

$$C_{\mathrm{t}}(d,x,y)=T\{f_{\mathrm{L}}(x-d+s_{-1}(x,y),y),f_{\mathrm{R}}(x,y)\} \tag{4}$$

其中T为特征拼接操作,在得到形状为$N\times D\times C\times H\times W$的4D矩阵后,使用3个三维卷积进行聚合,将其压缩为$N\times D\times 1\times H\times W$的形状。最后将相似度矩阵和压缩4D代价矩阵级联形成组合代价矩阵。

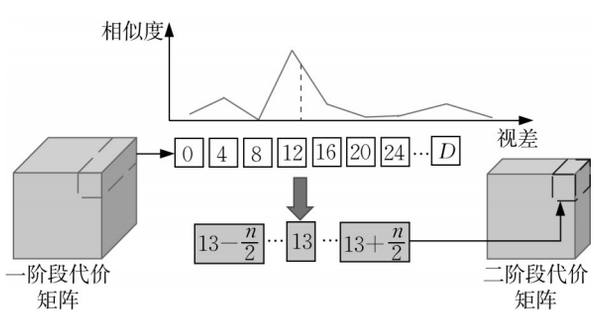


图3 视差搜索范围的变化

Fig. 3 Change in disparity search range

2.3 动态注意力机制

提出一个基于动态注意力的多尺度残差模块,引导残差网络更多地关注在当前尺度下空间中那些不准确的区域。假设当前尺度是c(1/2^c原图分辨

率），根据在上一个尺度$c-1$下的视差估计的上采样结果$\hat{s}_{c-1,\mathrm{up}}$，可以根据对右图$I_{c,\mathrm{R}}$进行采样得到合成的左图$\tilde{I}_{c,\mathrm{L}}$，如式(5)所示：

$$\tilde{I}_{c,L}(x,y)=I_{c,R}(x+\hat{s}_{c-1,up}(x,y),y) \tag{5}$$

根据合成的左图和真实左图可以得到误差图，为

$$E_{c,\mathrm{L}}=|\tilde{I}_{c,\mathrm{L}}-I_{c,\mathrm{L}}| \tag{6}$$

一个3层的基于2D卷积的网络被用于接收由误差图、粗粒度视差图和左右视图拼接而成的输入$f_{c,\mathrm{a}}$。

来自解码器网络的上卷积特征映射和来自编码器网络的对应尺度的特征$f_{c,\mathrm{in}}$会和注意力模块的输入相拼接，然后与注意力分布图在空间维度上相乘后形成残差特征，为

$$f_{c,\mathrm{ar}}=T\{f_{c,\mathrm{in}},f_{c,\mathrm{a}}\}\otimes\sigma(f_{c,\mathrm{a}}) \tag{7}$$

$\sigma(\cdot)$代表sigmoid函数。该残差特征会输入二维卷积堆叠网络，输出尺度c下的残差视差r_c，最终预测视差结果$\hat{s}_c$定义为

$$\hat{s}_c=\hat{s}_{c-1,\mathrm{up}}+r_{\mathrm{c}} \tag{8}$$

如图4所示，发现注意力图在误差图输入的引导下自适应地关注不同类型的区域。

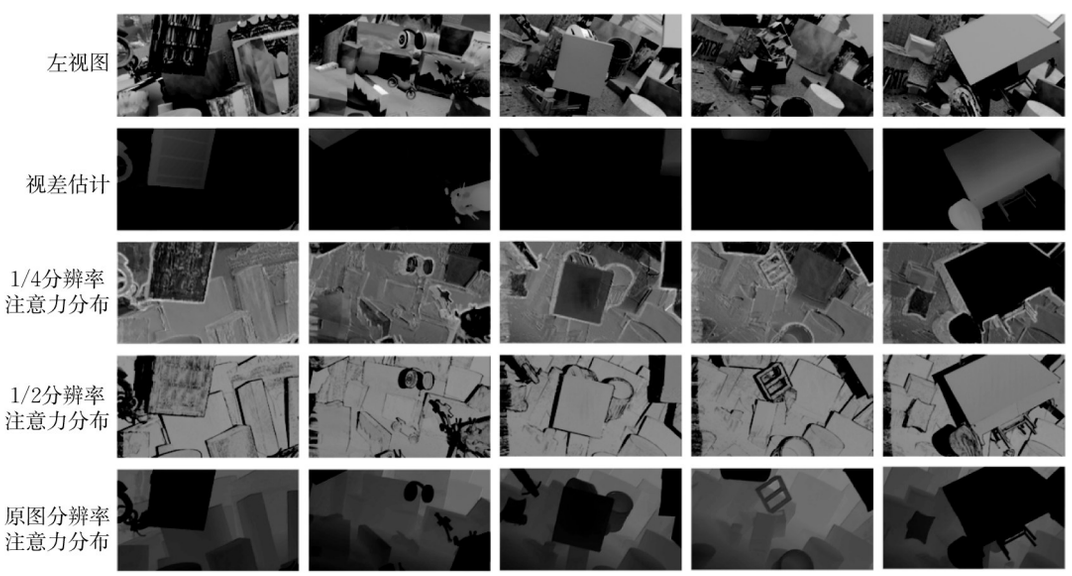


图4　注意力分布图在不同尺度上的可视化

Fig. 4　Attention map in different scales

2.4　损失函数

对于不同尺度s下的视差估计结果，使用逐像素点的L_1平滑损失去计算，为

$$L_c(s_c,\hat{s}_c)=\frac{1}{P}\sum_{i=1}^{P}L_1(s_{c,i}-\hat{s}_{,i}) \tag{9}$$

式中：P为视差图中的像素点数量；$\hat{s}_{c,i}$为尺度c下的视差估计结果的第i个像素点的视差值；s_c为视差真值。

最终的视差损失函数是基准视差损失和多尺度视差损失的结合，为

$$l=\sum_{c=-1}^{C}\lambda_c L_c(s_c,\hat{s}_c) \tag{10}$$

式中：λ_c为在尺度c下的损失权重。

3　性能验证

3.1　数据集和评估指标

4个公共数据集用于训练和测试本文提出的EDNet++。

SceneFlow：SceneFlow数据集[5]由39 824对合成立体RGB图像（5 454个图像对用于训练，4 370个图像对用于测试）组成，全分辨率为960×540。它有很多训练数据，所以它经常被用作预训练数据集。

Kitti：Kitti 2012和Kitti 2015都是真实场景的数据集，全分辨率为1 242×375。这2个数据集的深度真值由激光雷达产生，因此只有稀疏的真值可用。总共有400对图像用于训练，400对用于评估。

Middlebury：Middlebury数据集是在真实场景中由高分辨率的结构化照明捕获的。它只有15对训练用的双目图片和15对测试用的双目图片，所以在训练时很可能会引起过拟合。

ETH3D：ETH3D数据集包括27个训练帧和20个测试帧，涵盖各种室内和室外场景。

在Scene Flow数据集上用像素点平均误差（EPE）、1像素误差和3像素误差来评估模型。像素

点平均误差计算像素的平均视差误差,1像素误差和3像素误差分别测量EPE大于1像素和大于3像素的平均百分比。官方指标(例如D1-all)报告用于KITTI 2012和KITTI 2015数据集的评估。

3.2 训练细节

用PyTorch[26]实现EDNet++,并用Adam(动量为0.9,二阶动量为0.999)作为优化器训练模型。对于Scene Flow数据集,原始图像被随机裁剪到320×640作为输入。网络模型在2个NVIDIA RTX 2080ti GPU上进行训练,共70个轮次,批处理大小为8个(每个GPU 4个)。初始学习率设置为0.001,在第20个轮次之后每10个轮次降低一半学习率。损失权重设置为$\lambda^d=0.3$,$\lambda^0=1.0$,$\lambda^1=\lambda^2=0.8$,$\lambda^3=0.6$。使用了预训练的Sceneflow模型,然后在KITTI 2012和KITTI2015上进行微调,首先在2个数据集上混合训练1 000个轮次,再在2个数据集上分别进行了400个阶段的训练,以获得各自的提交结果。对于KITTI数据集,使用恒定的学习速率0.000 1。

3.3 消融实验

两阶段的代价矩阵构建方法让EDNet++相对于EDNet[27]而言提高了模型的泛化能力和推理速度。通过对比实验验证了两阶段构建方法对泛化效果的改善。在Sceneflow对EDNet[27]和EDNet++进行了70个轮次的训练,并在KITTI2015训练集、Middlebury 2014测试集和ETH3D测试集上进行了测试。如表1所示,该方法可以显著提高EDNet[27]在所有数据集上对所有指标的泛化能力。B_2和B_4分别为预测值和实际值相差2像素和4像素的百分比,E为像素误差百分比。

表1 二阶段构建方式在泛化性能上的表现

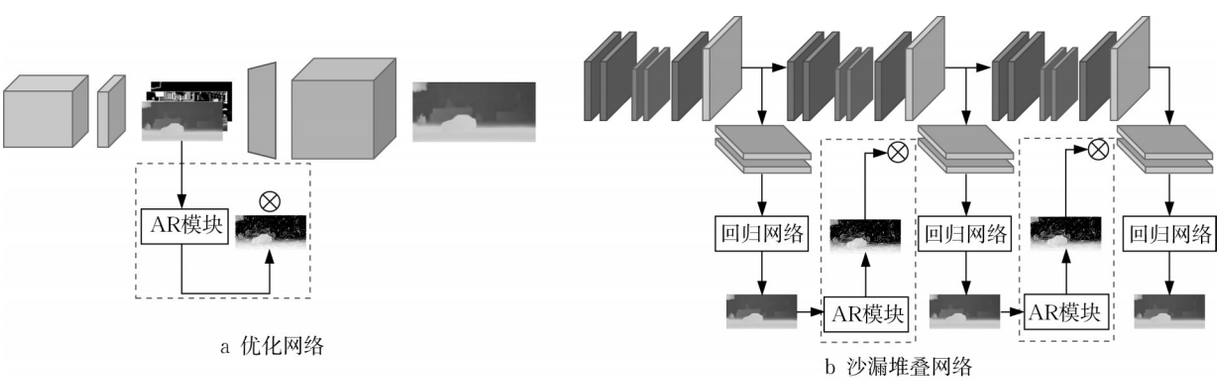
Tab. 1 Improvement in generalization

网络结构	Middlebury			KITTI2012		KITTI2015		ETH3D	
	B_2	B_4	E	像素误差百分比大于3	E	像素误差百分比大于3	E	B_2	B_4
一阶段代价矩阵	54.00	41.20	20.40	43.70	8.00	48.30	10.78	10.31	6.23
二阶段代价矩阵	45.10	30.80	11.80	15.70	1.90	14.90	2.29	6.54	2.96

在Middlebury数据集上最显著的改善是平均像素误差,下降了42%。EDNet++在KITTI 2015上测试时,在误差大于3像素和EPE这2个指标方面的性能分别提高了66%和77%,在KITTI 2012上测试时相同指标的性能分别提高了62%和76%。

3.4 模型迁移

主要的立体匹配聚合网络可分为优化网络(如Steronet[18]、StereoDRnet[28])和堆叠沙漏网络(如PSM[2]、FADNet[29])2种类型。这2种网络结构有一个共同点,就是它们在上采样过程中不可避免地产生误差。将基于注意力的残差模块应用于这2种网络中,通过实验证明其有效性,基于注意力的残差模块的优化网络和堆叠沙漏网络框架如图5所示,基于注意力的残差模块用虚线标记。将粗预测结果和相应的误差图作为输入并输出相同尺寸的注意图,将注意图与特征在空间维度上作点乘。



a 优化网络 b 沙漏堆叠网络

图5 基于注意力的残差模块在两类聚合网络中的应用

Fig. 5 Application of attention-based residual module into two kinds of aggregation networks

表2揭示了基于注意力的残差模块和组合的代价矩阵在Sceneflow测试集的效果。所有网络都在SceneFlow训练集上按照文献中提到的默认批大小和学习率训练10个轮次后在测试集上进行测试,运行时间指的是在显卡rtx2080上推理一对1 390×950像素图像所需要的时间。

表2　基于注意力的残差模块和组合代价矩阵的在不同网络结构上的迁移实验

Tab. 2　Portability of attention-based spatial residual module and cost volume combination module

网络框架	代价矩阵		残差结构		指标结果	
	原始	组合代价矩阵	原始	动态注意力	EPE	时间/s
StereoNet[18]	√		√		1.55	0.169
StereoNet－A	√			√	1.37	0.172
StereoDRNet[27]	√		√		1.33	0.171
StereoDRNet－A	√			√	1.23	0.175
PSMNet[2]	√		√		1.09	0.338
PSMNet－A	√			√	1.05	0.343
AANet[30]	√		√		1.60	0.143
AANet－A	√			√	1.50	0.150
AANet－CA		√		√	1.27	0.281
DispNet[5]	√		√		2.11	0.056
DispNet－A	√			√	2.11	0.068
DispNet－CA		√		√	2.00	0.073
FADNet[29]	√		√		1.74	0.077
FADNet－A	√			√	1.63	0.086
FADNet－CA		√		√	1.61	0.092

可以看到所有的网络在通过基于注意力的残差网络迁移后精度都得到了提高，相应地在时间复杂度上也有些许上升。改善最大的是AANet[30]，EPE下降了20%。虽然所有的网络模型受限于训练成本没有训练到各自的最优状态，但是也足以说明该模块的有效性和可移植性。

3.5　实验结果

将提出的算法模型与现有的最先进的方法在推理速度、内存消耗和准确率方面进行比较，使用的数据集有Sceneflow、KITTI 2015和KITTI 2012。

如表3所示，EDNet＋＋由于实验波动误差比EDNet[27]少0.01，但是运行速度快了14%，相较于其他先进算法而言，运行速度最快，并且能够保持较低的EPE：本文算法在运行时间上比PSMNet[2]快9倍、比GwcNet[3]快5倍、比CFNet[4]快3倍，并且在算法精度上优于这些方法。本文算法虽然在算法精度上没有达到最高（相较于ACVNet[31]，EPE上升了0.26），但是在运行时间上大幅度优于ACVNet（运行速度比它快了4倍）。

在Kitti数据集上，本文将对比实验分为2组，如表4所示。表中N和A分别表示非被遮挡和所有区域的平均离群点占真实像素值的百分比；f和a分别表示前景和所有真值平均离群值的百分比。EDNet[27]和EDNet＋＋在Kitti上迁移学习之后有着相似的结果。首先，与其他性能最好的基于3D卷积的方法相比，EDNet＋＋在评估前景像素的视差时

表3　几种方法在Sceneflow数据集上的EPE和运行时间

Tab. 3　EPE values and running time on scene flow dataset of several state-of-the-art methods

方法	EPE	时间/s
PSMNe	1.09	0.453
CFNet	0.97	0.180
GwcNet	0.76	0.254
Bi3D	0.73	内存溢出
ACVNet	0.43	0.225
AANet＋＋	0.72	0.068
EDNet	0.68	0.059
EDNet＋＋	0.69	0.051

注：推理时间是在单个NVIDIA 2080ti GPU上测试的结果，分辨率为576×960。

仍然取得了具有竞争力的测评结果，并且运行速度相对更快。将本文方法与基于2D卷积的实时模型进行比较。表5显示了本文方法和其他最新方法相比具有竞争性。为了强调本文方法的效率之高，在实验阶段与一些主流的基于3D卷积的模型比较了计算复杂度、内存消耗和推理速度。表5显示，本文模型需要更少的内存消耗，表中所有结果都在单个NVIDIA RTX 2080Ti GPU上进行了测试，分辨率为1 248×384。

表4　KITTI 2015测试集的基准测试结果

Tab. 4　Benchmark results on KITTI 2015 test sets

方法	N/%		A/%		时间/s
	f	a	f	a	
GANet[4]	3.37	**1.73**	3.82	1.93	0.360
GCNet[1]	5.58	2.61	6.16	2.87	0.900
MC－CNN[32]	7.64	3.33	8.88	3.89	67.000
HD[33]	3.43	1.87	3.63	2.02	0.140
CSN[16]	3.55	<u>1.78</u>	4.03	**1.59**	0.600
DeepPruner－B[34]	3.18	1.95	3.56	2.15	0.180
Bi3D[35]	**3.11**	1.79	<u>3.48</u>	1.95	0.480
ACVNet[32]			**3.07**	<u>1.65</u>	0.200
CFNet[17]		1.73		1.88	0.180
EDNet[27]（本文模型）	3.33	2.31	3.88	2.53	<u>0.050</u>
EDNet＋＋（本文模型）	<u>3.14</u>	2.62	3.69	2.83	**0.045**
AANet[30]	4.93	<u>2.32</u>	5.39	<u>2.55</u>	0.075
DeepPruner－B[31]	3.18	**1.95**	<u>3.56</u>	2.59	0.060
FADNet[29]	**3.07**	2.59	**3.50**	2.82	0.050
RLStereo[36]	4.76	2.38	5.38	2.64	<u>0.038</u>
MADNet[37]	8.41	4.27	9.20	4.66	**0.020**
EDNet[27]（本文模型）	3.33	2.31	3.88	**2.53**	0.050
EDNet＋＋（本文模型）	<u>3.14</u>	2.62	3.69	2.83	0.045

注：加粗表示最佳结果；加下划线表示次优。

图6展示了本文方法和KITTI2015数据集上其他最新方法的视差和误差对比。

表5 运行时间、占用内存量和计算成本的比较

Tab. 5 Comparisons of running time, running memory, and computational costs

方法	内存/GB	GFLOPs	时间/s
PSMNet[6]	4.83	937.9	0.393
GANet[4]	6.53	1936.98	2.43
GwcNet[3]	4.27	899.99	0.272
Bi3D[35]	10.74	4212.05	0.899
EDNet[27](本文模型)	**2.52**	**162.92**	0.053
EDNet++(本文模型)	2.78	167.59	**0.047**

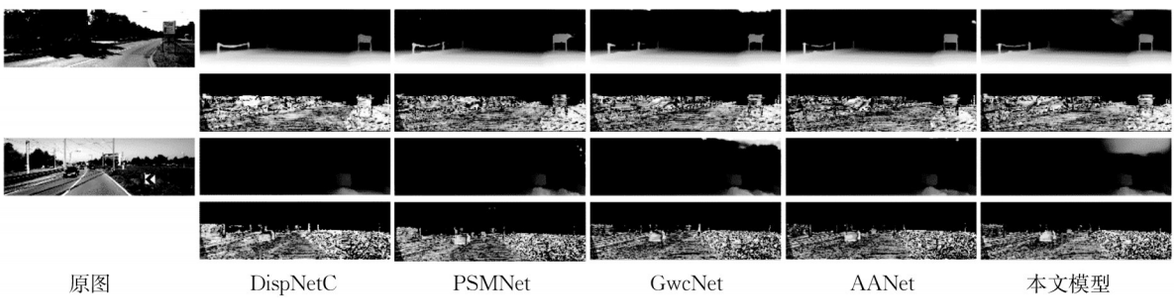


图6 KITTI 2015测试集的视差预测结果

Fig, 6 Results of disparity prediction for KITTI 2015 testing data

在Middlebury数据集上测试了模型的泛化能力。所有模型都在Sceneflow上进行训练。由图7可见,EDNet++可以在低纹理区域产生更平滑、更连续的视差估计,并生成更清晰的边缘。

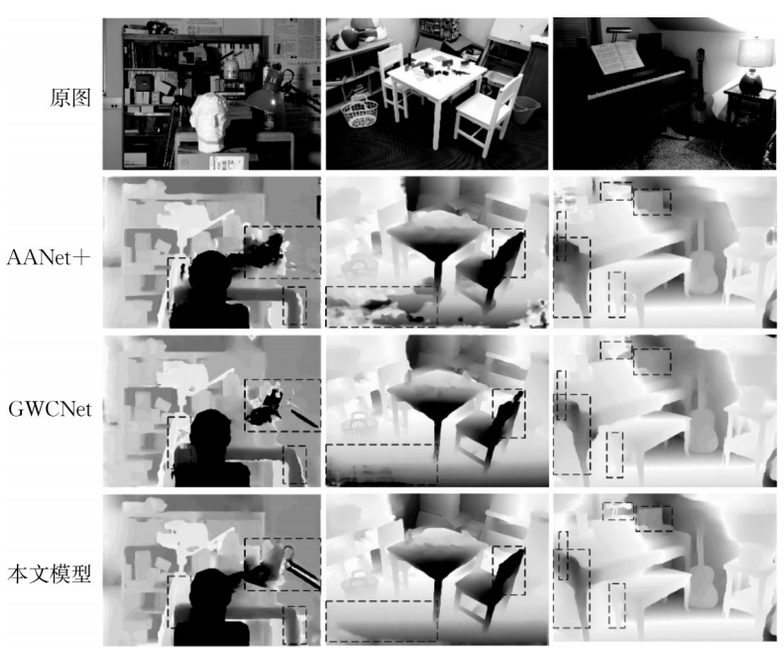


图7 Middlebury 2014数据集的模型泛化能力可视化

Fig. 7 Evaluation of model generalization on Middlebury 2014 dataset

4 结语

提出了一个高效的双目立体匹配网络框架,其中包括两阶段的组合代价矩阵和多尺度动态注意力残差模块。采用两阶段的理论将相似度矩阵和基于特征拼接的4D代价矩阵相结合,使模型的学习能力提高的同时,大幅提高算法的泛化能力,同时还生成了具有更少参数的代价矩阵。此外,多尺度动态注意力模块的使用大幅提高了残差学习框架的效率,并通过大量迁移实验证明了该模块在其他由粗到细

粒度的框架的网络中也具有良好表现。以KITTI和Sceneflow数据集证明了该方法的优越性。

作者贡献声明：

王志成：论文撰写、深度神经网络设计。

王泽灏：论文撰写、深度神经网络设计、实验实施。

参考文献：

[1] 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). Venice: IEEE, 2017: 66-75.

[2] CHANG J, CHEN Y. Pyramid stereo matching network[C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018: 5410-5418.

[3] GUO X, YANG K, YANG W, *et al*. Group-wise correlation stereo network [C]//2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Long Beach: IEEE, 2019: 3268-3277.

[4] ZHANG F, PRISACARIU V, YANG R, *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.

[5] MAYER N, ILG E, H¨AUSSER 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.

[6] LIANG Z, FENG Y, GUO Y, *et al*. Learning for disparity estimation through feature constancy [C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018: 2811-2820.

[7] YANG G, ZHAO H, SHI J, *et al*. Segstereo: Exploiting semantic information for disparity estimation [C]//Proceedings of the European Conference on Computer Vision (ECCV). Munich: IEEE, 2018: 636-651.

[8] PANG J, SUN W, 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). Venice: IEEE, 2017: 878-886.

[9] SCHARSTEIN D, SZELISKI R. A taxonomy and evaluation of dense two-frame stereo correspondence algorithms [J]. International Journal of Computer Vision, 2002, 47: 7.

[10] ZBONTAR J, LECUN Y. Stereo matching by training a convolutional neural network to compare image patches [J]. Journal of Machine Learning Research, 2016, 17(65): 1.

[11] DOSOVITSKIY A, FISCHER P, ILG E, *et al*. Flownet: Learning optical flow with convolutional networks [C]//2015 IEEE International Conference on Computer Vision (ICCV). Santiago: IEEE, 2015: 2758-2766.

[12] CHEN Z, SUN X, WANG L, *et al*. A deep visual correspondence embedding model for stereo matching costs [C]//2015 IEEE International Conference on Computer Vision (ICCV). Santiago: IEEE, 2015: 972-980.

[13] FENG Y, LIANG Z, LIU H. Efficient deep learning for stereo matching with larger image patches [C]//2017 10th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI). Shanghai: IEEE, 2017: 1-5.

[14] NIE G, CHENG M, LIU Y, *et al*. Multi-level context ultra-aggregation for stereo matching [C]//2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Long Beach: IEEE, 2019: 3278-3286.

[15] HUANG G, LIU Z, VAN DER Maaten L, *et al*. Densely connected convolutional networks [C]//2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Honolulu: IEEE, 2017: 2261-2269.

[16] GU X, FAN Z, ZHU S, *et al*. Cascade cost volume for high-resolution multi-view stereo and stereo matching [C]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle: IEEE, 2020: 2492-2501.

[17] SHEN Z, DAI Y, RAO Z. Cfnet: Cascade and fused cost volume for robust stereo matching [C]//2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Nashville: IEEE, 2021: 13901-13910.

[18] KHAMIS S, FANELLO S, RHEMANN C, *et al*. Stereonet: Guided hierarchical refinement for real-time edge-aware depth prediction [C]//Proceedings of the European Conference on Computer Vision (ECCV). Venice: IEEE, 2018: 573-590.

[19] GIDARIS S, KOMODAKIS N. Detect , replace, refine: Deep structured prediction for pixel wise labeling [C]//2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Honolulu: IEEE, 2017: 7187-7196.

[20] JIE Z, WANG P, LING Y, *et al*. Left-right comparative recurrent model for stereo matching [C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018: 3838-3846.

[21] YU A, GUO W, LIU B, *et al*. Attention aware cost volume pyramid based multi-view stereo network for 3d reconstruction [J]. ISPRS Journal of Photogrammetry and Remote Sensing. Netherlands: Elsevier, 2021, 175: 448.

[22] STUCKER C, SCHINDLER K. Resdepth: Learned residual stereo reconstruction [C]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). Seattle: IEEE, 2020: 707-716.

[23] RONNEBERGER O, FISCHER P, BROX T. U-net: convolutional networks for biomedical image segmentation [C]//2015 Medical Image Computing and Computer-Assisted Intervention (MICCAI). Munich: Springer Nature, 2015: 234-241.

[24] WANG Y, LAI Z, HUANG G, *et al*. Anytime stereo image depth estimation on mobile devices [C]//2019 International Conference on Robotics and Automation (ICRA). Montreal:

IEEE, 2019: 5893-5900.

[25] CHENG X, ZHONG Y, HARANDI M, *et al*. Hierarchical neural architecture search for deep stereo matching [J]. Advances in Neural Information Processing Systems. Vancouver: IEEE, 2020, 33: 22158.

[26] PASZKE A, GROSS S, MASSA F, *et al*. Pytorch: An imperative style, high-performance deep learning library [C]// WALLACH H, LAROCHELLE H, BEYGELZIMER A, et al. Advances in Neural Information Processing Systems: volume 32. Curran Associates, Inc., Vancouver: IEEE, 2019: 8026-8037.

[27] ZHANG S, WANG Z, WANG Q, *et al*. Ednet: Efficient disparity estimation with cost volume combination and attention-based spatial residual [C]//2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Nashville: IEEE, 2021: 5433-5442

[28] CHABRA R, STRAUB J, SWEENY C, *et al*. Stereodrnet: Dilated residual stereo net[J]. 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Long Beach: IEEE, 2019: 11786-11795.

[29] WANG Q, SHI S, ZHENG S, *et al*. FADNet: A fast and accurate network for disparity estimation [C]//2020 IEEE/ International Conference on Robotics and Automation (ICRA). Paris: IEEE, 2020: 101-107.

[30] XU H, ZHANG J, *et al*. Aanet: Adaptive aggregation network for efficient stereo matching [C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle: IEEE, 2020: 1956-1965.

[31] XU G, CHENG J, GUO P, *et al*. Attention concatenation volume for accurate and efficient stereo matching [C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans: IEEE, 2022: 12981-12990.

[32] ZBONTAR J, LECUN Y. Computing the stereo matching cost with a convolutional neural network [C]//2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Boston: IEEE, 2015: 1592-1599.

[33] YIN Z, DARRELL T, YU F. Hierarchical discrete distribution decomposition for match density estimation [C]// 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Long Beach: IEEE, 2019: 6037-6046.

[34] DUGGAL S, WANG S, MA W, *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.

[35] BADKI A, TROCCOLI A, KIM K, *et al*. Bi3d: Stereo depth estimation via binary classifications [C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle: IEEE, 2020: 1597-1605.

[36] YANG M, WU F, LI W. Rlstereo: Real-time stereo matching based on reinforcement learning [J]. IEEE Transactions on Image Processing, 2021, 30: 9442.

[37] TONIONI A, TOSI F, POGGI M, *et al*. Real-time self-adaptive deep stereo [C]//2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Long Beach: IEEE, 2019: 195-204.