激光与光电子学进展
先进成像

基于多重注意力机制的立体匹配算法

陈其博[1,2]，葛宝臻[1,2*]，李云鹏[1,2]，权佳宁[1,2]
[1]天津大学精密仪器与光电子工程学院，天津 300072；
[2]天津大学光电信息技术教育部重点实验室，天津 300072

摘要 现有的立体匹配算法在弱纹理、阴影等病态区域的匹配效果较差，为了充分利用场景上下文信息来提高视差匹配精度，提出一种有效的多重注意力立体匹配算法(MAnet)。特征提取阶段，通过多种注意力机制，即位置通道注意力和多头十字交叉注意力(MCA)，调整特征通道并有选择性地聚合任意范围内的上下文信息，为匹配代价计算提供更有辨识性的特征。将MCA扩展到3D卷积中，扩大网络感受野，使聚合的匹配代价更准确。对于网络的损失函数，通过加权超过误差阈值部分的损失来提高网络对困难区域的学习能力。在KITTI数据集中对算法进行实验验证，MAnet对KITTI2015测试集的误差仅为2.06%，实验结果表明，与基准算法相比，MAnet提高了视差精度，有助于改善病态区域的匹配效果。

关键词 机器视觉；立体匹配；卷积神经网络；双目视觉；注意力机制
中图分类号 TP391.4 **文献标志码** A **DOI**：10.3788/LOP202259.1633001

Stereo Matching Algorithm Based on Multi-Attention Mechanism

Chen Qibo[1,2], Ge Baozhen[1,2*], Li Yunpeng[1,2], Quan Jianing[1,2]
[1]*School of Precision Instrument and Opto-Electronics Engineering, Tianjin University, Tianjin* 300072, *China*;
[2]*Key Laboratory of Opto-Electronics Information Technology, Ministry of Education, Tianjin University, Tianjin* 300072, *China*

Abstract Stereo matching algorithms used currently are ineffective at matching weak textures, shadows, and other pathological regions. To make full use of scene context information to improve disparity matching accuracy, this study proposes an effective multiple attention stereo matching algorithm (MAnet). At the feature extraction stage, according to multiple attention mechanisms, such as location channel attention and multiheads crisscross attention (MCA), we adjust the feature channels and selectively aggregate contextual information in any range to provide more discriminative features for matching cost calculation. Extending MCA to 3D convolution expands the network perceptual region to accumulate more precise matching cost. The learning ability of challenging regions is enhanced for the networks' loss function by weighting the loss outside the error threshold. The algorithms are experimentally validated on the KITTI dataset, and the error of MAnet for the KITTI2015 test set is 2.06%. The experimental findings demonstrate that compared to the benchmark algorithm, the MAnet enhances disparity accuracy and improves the matching performance in the pathological region.

Key words machine vision; stereo matching; convolutional neural network; binocular vision; attention mechanism

1 引 言

立体匹配作为立体视觉的核心任务，通过寻找两幅图像间像素级的对应关系，从一对立体校正后的图像中匹配视差。传统算法将匹配过程分为代价计算、代价聚合、视差计算和视差求精四个步骤[1]。得益于深度学习的不断发展，卷积神经网络(CNN)展现出强大的特征提取能力，许多学者提出基于CNN的立体匹配算法。

最早，Žbontar等[2]将CNN用于提取匹配特征，但仍然需要结合传统算法中的代价聚合及滤波等后处理步骤。Mayer等[3]提出一种沿视差方向进行一维匹配代价计算的端到端网络。Kendall等[4]通过3D卷积实现代价聚合、视差计算，为后续算法提供了思路。Chang等[5]在3D卷积中引入了编码解码机制的堆叠

收稿日期：2021-06-09；**修回日期**：2021-06-17；**录用日期**：2021-06-27
通信作者：*gebz@tju.edu.cn

3D沙漏结构，提高算法精度。刘建国等[6]通过构建局部稠密代价卷对得到的视差进行进一步优化。王玉锋等[7]通过加入辅助任务，网络有效学习边缘和特征的一致性信息，并采用循环迭代的方式更新视差图。

基于CNN的立体匹配算法在精度和速度上都取得明显进步，但在弱纹理、反光等病态区域仍然存在大量错误的匹配结果，主要原因是卷积网络的感受野[8]受限于卷积核大小，忽略了远距离像素对当前区域的贡献。而模拟人类视觉的注意力机制通过学习不同特征之间的关联性来突出重要特征，很好地弥补了卷积操作的局限性。CNN中的注意力机制可分为三类，分别是通道注意力，如挤压激发网络(SE-net)[9]通过计算不同通道的重要程度，对特征通道进行调整；空间注意力，如卷积块注意力模组(CBAM)通过池化压缩通道，计算特征图不同位置的重要程度；非局部注意力，如非局部神经网络(Non-local)[11]通过计算相似性的方式建立像素间的关联性。因此也有学者将注意力机制引入立体匹配算法中，程鸣洋等[12]在特征提取阶段嵌入非局部的空间注意力和通道注意力，通过建立所有像素之间的相关性来捕获场景信息。张亚茹等[13]在特征提取和3D卷积中使用全局池化的通道注意力进行信息交互，所提出的方法具有计算开销小的优点。张文等[14]在视差优化模块中引入CBAM中的通道和空间注意力级联机制，使网络自主选择视差优化区域。而黄继辉等[15]在特征提取阶段将通道和空间注意力串联，增强对特征的表达能力。

本文在金字塔立体匹配网络(PSMnet)基础上，通过多层嵌入多种注意力模块的方式，提出了一种多重注意力立体匹配算法(MAnet)。嵌入的位置通道注意力模块通过混合编码生成通道权重的方式进行特征通道调整。设计多头十字交叉注意力(MCA)模块，通过集成多个独立的相关计算结果，在不同特征子空间中建立行、列方向上的像素间联系，进而捕获不同方面的场景信息，这些场景信息使病态区域处聚合得到的匹配代价更加精准。最后在损失函数部分，通过加权超过误差阈值部分的损失来强化网络的学习能力。实验结果表明，MAnet在病态区域获得较好的匹配效果，匹配精度得到一定提高。

2　基于多重注意力的立体匹配算法

2.1　网络结构

以PSMnet作为基准，网络结构如图1所示。对于左右图像，通过共享的CNN模块和金字塔池化(SPP)模块提取图像特征，其中CNN模块共使用3个3×3的2D卷积层和4个残差连接层，每个残差连接层分别包含3，16，3，3数量的残差基础块，而每个残差基础块又由2个3×3大小的2D卷积层、批归一化层、非线性激活函数组成。SPP模块由不同步长和大小的池化核组成，可以提取不同尺度的信息。CNN模块和SPP模块的输出通过2D卷积层融合，并以级联的方式构建匹配代价卷(cost-volume)[16]。而得到的匹配代价卷由3D卷积层和3个堆叠的沙漏模块(hourglass)[13]进行代价聚合。每个沙漏模块包含2个步长为2、大小为3×3×3的3D卷积层，2个步长为1、大小为3×3×3的3D卷积层及2个大小为3×3×3的3D转置卷积。最后将匹配代价卷上采样到原始图像大小，通过Softmax函数生成各个视差级对应的概率，将概率和视差值加权，得到输出视差图。

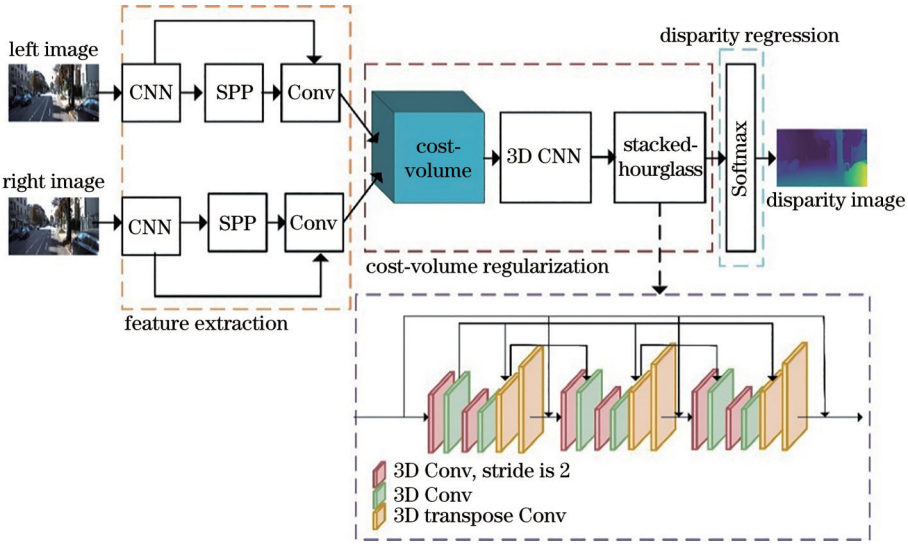

图1　PSMnet结构

Fig. 1　Architecture overview of PSMnet

所提算法网络结构如图2所示。MAnet使用Attention-CNN、MCA、SPP三个模块来提取左右图像特征。其中Attention-CNN模块由3个3×3的2D卷积层和残差基础块数量分别为3、9、3、3的4个残差连

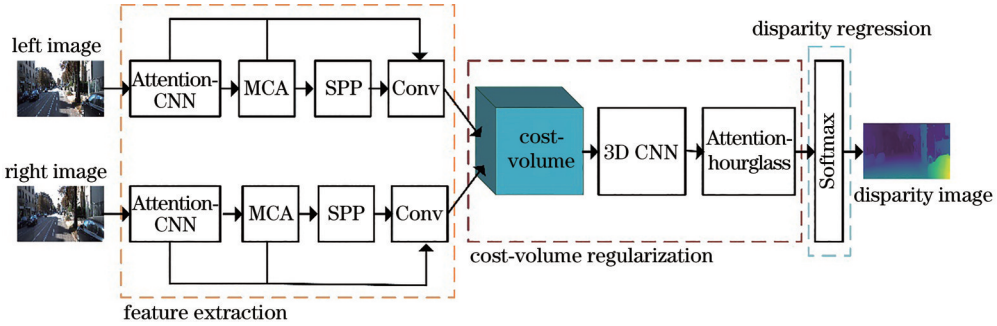

图 2　所提算法网络结构

Fig. 2　Architecture overview of proposed algorithm

接层组成，位置通道注意力模块添加在最后两层的残差块中，实现特征通道的动态调整。而 MCA 模块被添加在 Attention-CNN 和 SPP 模块之间，对 Attention-CNN 提取的特征计算相关性，获得更多全局以及局部的上下文信息；然后将 Attention-CNN、MCA、SPP 输出的特征融合，构建匹配代价卷。在匹配代价卷正则化阶段[5]，通过注意力沙漏模块（Attention-hourglass）完成匹配代价聚合。最后将匹配代价卷上采样到原始图像大小，通过概率加权方式得到输出视差图。

2.2　位置通道注意力模块

通道注意力能根据编码得到的通道权重对特征通道进行调整，提高网络对图像信息的利用能力。而多数立体匹配算法中的通道注意力采用全局平均池化[9]的编码方式，这种编码方式将图像空间信息压缩到一个通道相关性描述符中，对不同空间位置的特征通道产生相同的响应，不适合需要捕捉空间结构信息的像素级任务，所以本文在 CNN 模块中嵌入一种沿行、列方向分别进行池化编码的位置通道注意力模块。由于平均池化只能保存池化范围内的整体信息，因此与文献[17]不同的是，所提算法不仅使用平均池化（Avg pool）还使用最大池化（Max pool）混合编码，以保留池化范围内的纹理信息。并且为了充分利用平均池化和最大池化获得的特征，本文对二者特征进行融合之后没有采用通道降维的方式减少计算量。

所提位置通道注意力结构如图 3 所示，大小为（B, C, H, W）的输入特征图通过大小为（H, 1）以及（1, W）的池化核进行平均池化和最大池化编码，其中 B 代表批大小，C 表示通道数，H 代表特征图高，W 代表特征图宽。对不同方向的最大池化和平均池化编码结果通过级联和 1×1 卷积的方式分别进行融合，表达式分别为

$$\boldsymbol{p}^{w}=\sigma\left(F\left(\left[\boldsymbol{p}_{\mathrm{avg}}^{y},\boldsymbol{p}_{\mathrm{max}}^{y}\right]\right)\right),\tag{1}$$

$$\boldsymbol{p}^{h}=\sigma\left(F\left(\left[\boldsymbol{p}_{\mathrm{avg}}^{x},\boldsymbol{p}_{\mathrm{max}}^{x}\right]\right)\right),\tag{2}$$

式中：[,]表示级联操作；F 表示 1×1 的卷积融合操作；σ 表示非线性激活函数；$\boldsymbol{p}_{\mathrm{avg}}$ 表示平均池化编码；$\boldsymbol{p}_{\mathrm{max}}$ 表示最大池化编码；上标 x, y 分别表示水平和垂直的编码方向；$\boldsymbol{p}^{h}$ 为（B, C, H, 1）大小的水平方向编码输出；$\boldsymbol{p}^{w}$ 为（B, C, 1, W）大小的垂直方向编码输出。为了能高效地利用式（1）和式（2）中的位置信息，将 $\boldsymbol{p}^{h}$ 和 $\boldsymbol{p}^{w}$ 在空间维度上级联，使用一个 1×1 卷积进行特征通道间信息交互：

$$\boldsymbol{p}=F_{1}\left(\left[\boldsymbol{p}^{h},\boldsymbol{p}^{w}\right]\right),\tag{3}$$

式中：F_1 为 1×1 卷积；输出 $\boldsymbol{p}\in\mathbf{R}^{B\times C\times 1\times(H+W)}$。

随后沿空间维度拆分 $\boldsymbol{p}$，重新得到 $\boldsymbol{p}^{h}\in\mathbf{R}^{B\times C\times H\times 1}$ 和 $\boldsymbol{p}^{w}\in\mathbf{R}^{B\times C\times 1\times W}$。对 $\boldsymbol{p}^{h}$ 和 $\boldsymbol{p}^{w}$ 使用 Sigmoid 激活函数，得到[0, 1]区间的通道权重，将两个方向的通道权重和输入特征图相乘，调整特征通道，得到输出特征图。整

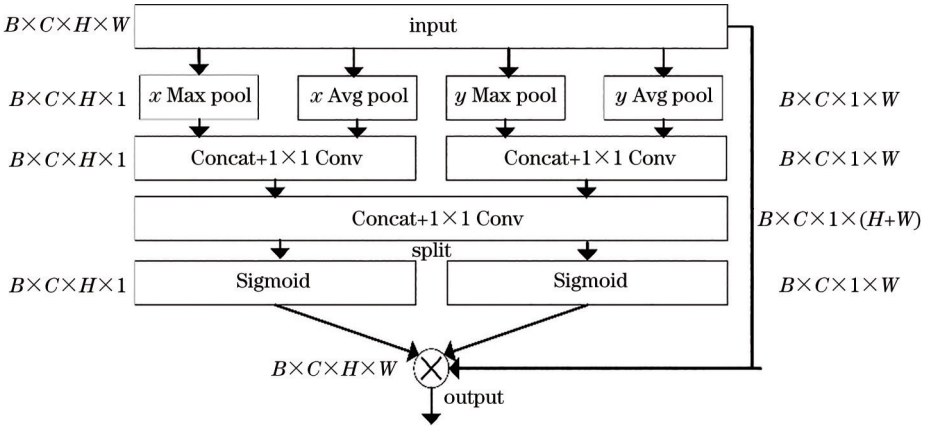

图 3　基于位置的通道注意力模块

Fig. 3　Location-based channel attention module

个位置通道注意力模块可描述为

$$\boldsymbol{y}(i,j)=\boldsymbol{x}(i,j)*\sigma\left[\boldsymbol{p}^{h}(i)\right]*\sigma\left[\boldsymbol{p}^{w}(j)\right],\tag{4}$$

式中：$\boldsymbol{y}(i,j)\in\mathbf{R}^{B\times C\times H\times W}$表示通道加权后的输出特征图；$\boldsymbol{x}(i,j)\in\mathbf{R}^{B\times C\times H\times W}$为输入特征图；$i\in(1,\cdots,H)$，$j\in(1,\cdots,W)$；激活函数$\sigma$为Sigmoid函数；*表示矩阵点乘。加入的位置通道注意力根据输入特征图在不同位置产生不同的通道响应，这种像素级通道调整在建立通道间相互依赖关系的同时还能融合空间位置信息，使网络能准确定位感兴趣的区域，增强网络区分不同像素的能力。

2.3　多头十字交叉注意力模块

立体匹配过程中不同位置的特征对匹配任务的重要程度是不同的，而非局部注意力仅考虑当前像素和其余像素之间的相关性，使不同距离上的像素有相同的表达机会，从而捕获丰富的上下文信息。然而受计算复杂度$O(HW\times HW)$限制，非局部注意力在输入图像较大时会明显拖慢网络的计算速度。考虑立体匹配沿水平方向寻找匹配点这一几何约束以及垂直方向的像素能提供上下文信息这些特点，将自然语言处理中的多头机制[18]迁移到语义分割中的十字交叉[19]注意力中，从而设计适用于立体匹配任务的MCA模块。所提MCA模块中的多头机制集成多个独立的十字交叉注意力计算结果，使网络计算像素相关性的同时关注不同方面的信息，提高了网络捕获场景上下文信息的能力。

MCA模块如图4所示。对于输入的特征图$\boldsymbol{X}\in\mathbf{R}^{B\times C\times H\times W}$，根据多头数$s$，将通道均分为$s$份，从而得到$s$个子特征图$\boldsymbol{X}^{h}\in\mathbf{R}^{B\times C/s\times H\times W}$；之后对每个子特征图使用三个不同的1×1卷积，生成查询矩阵$\boldsymbol{Q}\in\mathbf{R}^{B\times C/s\times H\times W}$，键矩阵$\boldsymbol{K}\in\mathbf{R}^{B\times C/s\times H\times W}$，值矩阵$\boldsymbol{V}\in\mathbf{R}^{B\times C/s\times H\times W}$；下一步通过查询矩阵和键矩阵的affinity操作，得到相关矩阵$\boldsymbol{Z}\in\mathbf{R}^{B\times(H+W-1)\times(H\times W)}$。取$\boldsymbol{Q}$矩阵中一点$i$，记为$\boldsymbol{Q}_i\in\mathbf{R}^{B\times C/s}$，同时在$\boldsymbol{K}$矩阵中按相应行列取得向量$\boldsymbol{K}_i\in\mathbf{R}^{B\times(H+W-1)\times C/s}$，$\boldsymbol{K}_{i,j}$定义为$\boldsymbol{K}_i$中第$j$个元素，$j\in\{1,\cdots,H+W-1\}$。图中affinity操作定义为

$$\boldsymbol{t}_{i,j}=\boldsymbol{Q}_i\boldsymbol{K}_{i,j}^{\mathrm{T}},\tag{5}$$

式中：$\boldsymbol{t}_{i,j}$表示$\boldsymbol{Q}_i$和$\boldsymbol{K}_{i,j}$间矩阵乘法得到的相关结果。

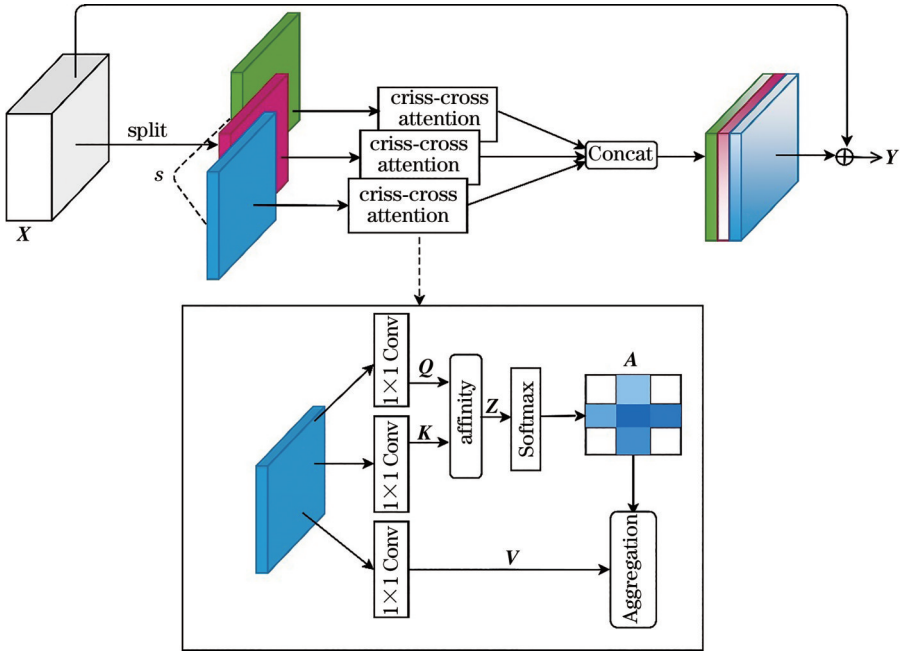

图4　多头十字交叉注意力模块

Fig. 4　Multi-cross-attention module

计算$\boldsymbol{Q}$中每点的affinity，得到$\boldsymbol{Z}\in\mathbf{R}^{B\times(H+W-1)\times(H\times W)}$。之后在$\boldsymbol{Z}$的$(H+W-1)$维度上使用Softmax函数，得到注意力响应图$\boldsymbol{A}\in\mathbf{R}^{B\times(H+W-1)\times(H\times W)}$。为了聚合信息，将得到的注意力图$\boldsymbol{A}$和值矩阵$\boldsymbol{V}$通过Aggregation操作重新加权特征。值矩阵$\boldsymbol{V}$在i点空间位置按行、列方向取得值向量$\boldsymbol{V}_i\in\mathbf{R}^{B\times(H+W-1)\times C/s}$，图中Aggregation定义为

$$\boldsymbol{X}_i^{h\prime}=\boldsymbol{V}_i^{\mathrm{T}}\boldsymbol{A}_i,\tag{6}$$

式中：$\boldsymbol{X}_i^{h\prime}$表示输出子特征图$\boldsymbol{X}^{h\prime}\in\mathbf{R}^{B\times C/s\times H\times W}$在$i$点处的特征向量；$\boldsymbol{A}_i$表示在$i$点对应行列范围内的注意力权值。最后将不同特征子空间中的输出级联后与输入相加融合，使网络通过学习特征残差的方式降低学习难度：

$$\boldsymbol{Y}=\left[\boldsymbol{X}^{1\prime},\cdots,\boldsymbol{X}^{h\prime}\right]+\boldsymbol{X},\tag{7}$$

式中：[，]表示级联操作；$\boldsymbol{Y}$为最终输出特征图；$\boldsymbol{X}$为输入特征图。所提MCA模块中的多头机制通过切分通道使网络同时考虑不同特征子空间的信息，并根据得到的注意力图$\boldsymbol{A}$有选择地聚合上下文信息，得到

鲁棒的匹配结果。所提MCA模块的计算复杂度仅为$O[(H+W-1)\times HW]$，能轻便地嵌入特征提取模块中。

为了在匹配代价卷中捕获更多上下文信息，将MCA模块扩展到沙漏模块中。如图5所示，相比基准的沙漏模块，本文在两个转置卷积之前插入扩展的MCA模块（3D MCA）和一个残差连接构成注意力沙漏模块。相比于MCA，3D MCA在特征子图中扩展了视差维度$\boldsymbol{X}^h\in\mathbf{R}^{B\times C/s\times D\times H\times W}$，并通过1×1×1卷积得到相应的查询矩阵$\boldsymbol{Q}\in\mathbf{R}^{B\times C/s\times D\times H\times W}$，键矩阵$\boldsymbol{K}\in\mathbf{R}^{B\times C/s\times D\times H\times W}$，值矩阵$\boldsymbol{V}\in\mathbf{R}^{B\times C/s\times D\times H\times W}$，affinity操作生成的注意力图为$\boldsymbol{A}\in\mathbf{R}^{B\times(D+H+W-2)\times(D\times H\times W)}$。MCA模块的输出为

$$\boldsymbol{Y}=\left[\boldsymbol{X}^{1\prime},\cdots,\boldsymbol{X}^{h\prime}\right]+f(\boldsymbol{X}),\tag{8}$$

式中：相加操作是为了使网络能通过学习特征残差的方式降低学习难度；f表示一个不带激活函数的1×1×1卷积。

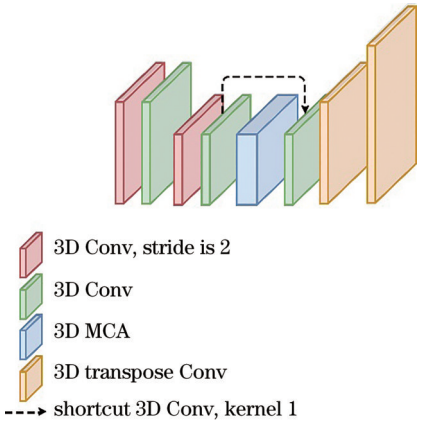

图5 注意力沙漏模块

Fig. 5 Attention hourglass module

2.4 损失函数

视差回归采用GC-Net[4]设计的soft-argmin回归方法，将每个像素的视差值与经过Softmax函数后得到的概率相乘，求和获得预测视差d'。

$$d'=\sum_{d=0}^{D}d\times p_d,\tag{9}$$

式中：D为最大视差级；p_d为视差级对应的概率。

采用平滑的L1损失函数作为基础视差损失函数：

$$L_1(d,d')=\frac{1}{N}\sum_{i=1}^{N}\mathrm{smooth}_{\mathrm{L1}}(d,d'),\tag{10}$$

$$\mathrm{smooth}_{\mathrm{L1}}(x)=\begin{cases}0.5x^2, & |x|<1\\ |x|-0.5, & |x|\geqslant 1\end{cases},\tag{11}$$

式中：N表示存在真实视差点的个数；d表示对应的真实视差。

为了使网络更加关注预测错误的区域，引入限制损失函数：

$$e_{\mathrm{abs}}=|d-d'|,\tag{12}$$

$$N'=\left(e_{\mathrm{abs}}>\delta\right),\tag{13}$$

式中：e_{abs}表示绝对误差；N'为误差大于阈值δ的点的个数；δ为人工设置的参数。得到的输出损失为

$$L(d,d')=\frac{1}{N}\sum_{i=1}^{N}\mathrm{smooth}_{\mathrm{L1}}(d,d')+\frac{\gamma}{N'}\sum_{i=1}^{N'}\mathrm{smooth}_{\mathrm{L1}}(d,d'),\tag{14}$$

式中：γ为超过误差阈值部分的损失的权重。堆叠的沙漏结构之间使用中间监督，因此总共产生3个输出，测试时仅使用最后一个沙漏模块的输出，总的损失函数为

$$L_{\mathrm{total}}(d,d')=\sum_{k=1}^{3}\omega_k L(d,d'),\tag{15}$$

式中：ω_k为不同输出的损失权重。

3 实 验

为了公平地验证和评估算法的性能，在两个常用的公开数据集KITTI2012和KITTI2015中进行实验验证。

3.1 数据集与评价标准

KITTI2012数据集为不同天气条件情况下记录真实车辆驾驶场景的数据集，其中包含194对训练图像对和195对测试图像对，同时提供由激光雷达采集的稀疏数据作为真值视差图。测试环节采用t-pixels误差以及EPE（end-point-error）误差作为算法评价指标。EPE误差表示算法预测视差和真实视差的平均欧氏距离，t-pixels误差表示预测视差图中误差大于t像素的像素数量占总数的百分比。

KITTI2015数据集与KITTI2012数据集类似，都是室外真实场景下的数据集。在建立真实视差图时不仅使用激光雷达还将3D CAD模型拟合到单独移动的车辆上，因此KITTI2015数据集相比于KITTI2012数据集有更加密集的视差监督。KITTI2015提供200对训练图像以及200对测试图像，采用D1指标作为误差评价标准。D1指标将预测误差同时大于3像素和5%真实视差值的像素作为异常点，并计算异常数量占总体数量的百分比。

3.2 训练细节

算法通过Pytorch 1.7.1框架实现，并在一块3090GPU上训练。训练采用Adam优化器，设置参数$\beta_1=0.9$，$\beta_2=0.999$用于梯度下降，最大视差D设置为192，初始批大小B为6。MCA模块多头数s设置为4，损失函数的超参数设为$\delta=0.3$，$\gamma=0.5$，$\omega_1=0.5$，$\omega_2=0.7$，$\omega_3=1.0$。为节约训练时间，加载部分PSMnet的Sceneflow数据集权重作为MAnet预训练权重。后续使用KITTI2012和KITTI2015总共394对训练图像联合训练所提算法，在训练之前将图像随机裁剪为256×512的图像块进行数据增强。受文献[20]启

发将训练分为两个过程，一阶段采用ReLU激活函数训练600个周期，前400个周期学习率设置为0.001，之后每50个周期学习率衰减一半。二阶段采用Mish激活函数[21]，如图6所示，Mish函数相比于ReLU激活函数更加平滑。二阶段训练400个周期，前200个周期学习率为0.001，之后每50个周期学习率衰减一半。Mish函数会额外增加显存占用，因此二阶段批大小设置为4。

$$\mathrm{ReLU}(x)=\max(0,x), \tag{16}$$

$$\mathrm{Mish}(x)=x\times\tanh\left\{\ln\left[1+\exp(x)\right]\right\}。 \tag{17}$$

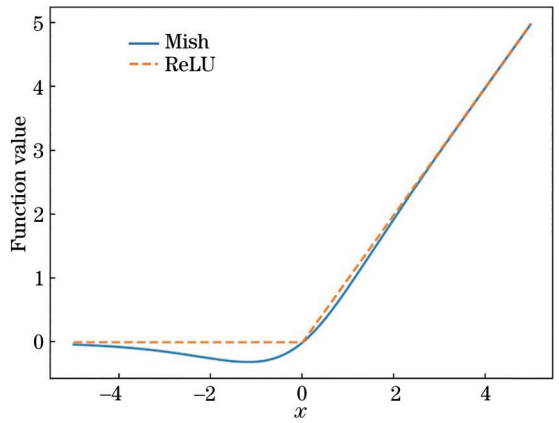

图6　ReLU函数和Mish函数的对比

Fig. 6　Comparison of ReLU function and Mish function

3.3　实验结果与对比

使用联合训练后的网络对KITTI2015测试集中200对图像进行预测，由于KITTI测试集不提供视差真值，将结果提交到KITTI在线测试网站上。提交后网站仅提供前20张测试图像的误差图作为参考，选择病态区域较多的图像3和图像13进行展示，如图7所示。图7展示了PSMnet和所提MAnet的预测视差图，其中视差图颜色从黄到蓝表示视差值从大到小的变化趋势，误差图显示预测视差与真实视差之间的差距，蓝色点表示正确匹配点，黄色点表示误匹配点，黑色点表示忽略的点。从图中标注的方框和对应的视差局部放大图可以看出，所提MAnet在栏杆、树干附近的匹配结果更加清晰完整，同时误差局部放大图显示在这些弱纹理、空洞区域，所提MAnet的匹配结果更加准确，表明多重注意力能改善弱纹理等病态区域的匹配结果。

为了定量比较所提MAnet与近年来基于神经网络的立体匹配算法的匹配性能，在KITTI2012和KITTI2015两个测试集中进行预测并将结果提交至在线测试网站进行评估。所提MAnet和其他主流算法在KITTI2015测试集的客观指标对比结果如表1所示，所提MAnet和其他主流算法在KITTI2012测试集的客观指标对比结果如表2所示，其中“Noc”表示计算非遮挡区域像素，“All”表示计算所有区域像素。

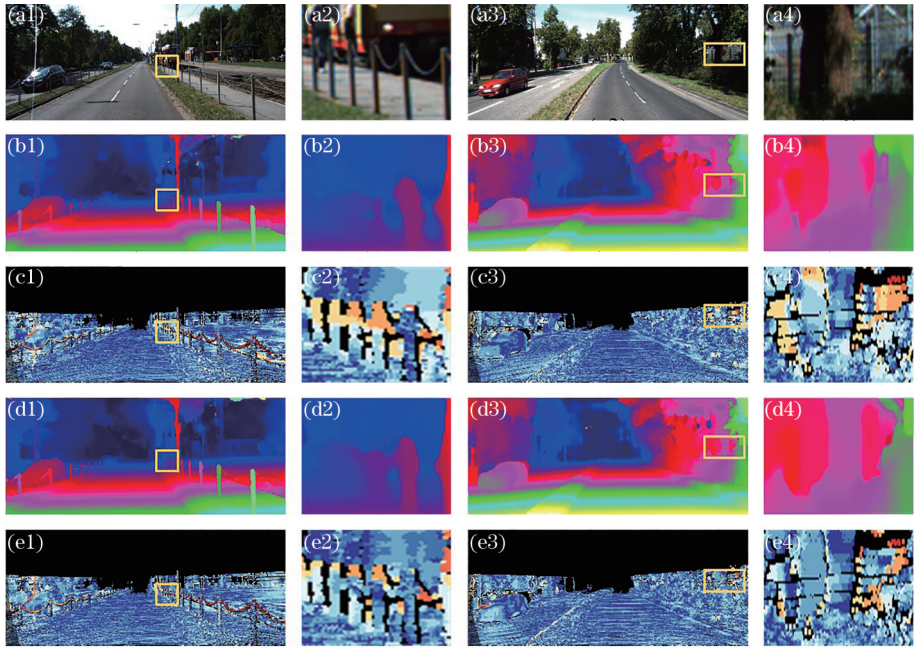

图7　KITTI2015测试集的视差评估结果。(a1)(a3)左图；(a2)(a4)左图的局部放大图；(b1)(b3) PSMnet预测的视差图；(b2)(b4)视差图的局部放大图；(c1)(c3)对应误差图；(c2)(c4)对应误差图的局部放大图；(d1)(d3)MAnet预测的视差图；(d2)(d4)视差图的局部放大图；(e1)(e3)对应误差图；(e2)(e4)对应误差图的局部放大图

Fig. 7　Parallax evaluation results on KITTI2015 test set. (a1) (a3) Left images; (a2) (a4) partial enlargement of left images; (b1) (b3) disparity maps predicted by PSMnet; (b2) (b4) partial enlargement of disparity maps; (c1) (c3) the corresponding error maps; (c2) (c4) local enlargement of the corresponding error maps; (d1) (d3) disparity maps predicted by MAnet; (d2) (d4) partial enlargement of disparity maps; (e1) (e3) the corresponding error maps; (e2) (e4) partial enlargement of the corresponding error maps

KITTI2015中还将图像细分为前景(fg)、背景(bg)、所有区域(All)。

由表1可知,相比基准PSMnet,所提MAnet的D1-bg、D1-fg、D1-all指标的误差降幅均超过10%。在KITTI2015排行榜排序采用的D1-all指标上,所提MAnet的误差为2.06%,与基准相比降低11.21%,而与同期基于PSMnet的改进算法CTFnet相比,误差降低6.36%。为公平地测试算法运行时间,在3090GPU上对PSMnet和MAnet进行测试,PSMnet测试100张图像的平均运行时间为0.396 s,MAnet测试100张图像的平均运行时间为0.363 s,降低了8.3%,表明所提算法的高效性。

在表2 KITTI2012测试集评估中,所提MAnet在误差不超过2 pixel,3 pixel的标准下均领先PSMnet,其中在2-pixels评价标准下误差降低了4.10%。然而对比表1中的误差降低幅度可以发现,MAnet在KITTI2015数据集的精度提升得更加明显。由于KITTI2015数据集中使用3D CAD模型拟合真值视差图,车辆上病态区域(车窗等)的视差监督更加稠密,因此KITTI2015数据集将更多的病态区域纳入测试范围。这一现象从侧面说明,所提算法是通过改善病态区域的匹配效果来提高整体匹配精度的。

表1 不同算法在KITTI 2015测试集上的性能评价

Table 1 Performance evaluation of different algorithms on the KITTI2015 test set

Algorithm	D1-bg /%		D1-fg /%		D1-all /%	
	Noc	All	Noc	All	Noc	All
GC-NET[4]	2.02	2.21	5.58	6.16	2.61	2.87
SegStereo[22]	1.76	1.88	3.70	4.07	2.08	2.25
PSMnet[5]	1.71	1.86	4.31	4.62	2.14	2.32
Bi3D[23]	1.79	1.95	3.11	3.48	2.01	2.21
CTFnet[6]	1.68	1.84	3.69	4.03	2.01	2.20
DeepPruner-Best[24]	1.71	1.87	3.18	3.56	1.95	2.15
GWCnet[25]	1.61	1.74	3.49	3.93	1.92	2.11
SSPCVNet[26]	1.61	1.75	3.40	3.89	1.91	2.11
MAnet	1.51	1.65	3.88	4.13	1.90	2.06

表2 不同算法在KITTI 2012测试集上的性能评价

Table 2 Performance evaluation of different algorithms on the KITTI2012 test set

Algorithm	2-pixels /%		3-pixels /%		4-pixels /%		EPE /pixel	
	Noc	All	Noc	All	Noc	All	Noc	All
GC-NET[4]	2.71	3.46	1.77	2.30	1.36	1.77	0.6	0.7
SegStereo[22]	2.66	3.19	1.68	2.03	1.25	1.52	0.5	0.6
PSMnet[5]	2.44	3.01	1.49	1.89	1.12	1.42	0.5	0.6
SSPCVNet[26]	2.47	3.09	1.47	1.90	1.08	1.41	0.5	0.6
MAnet	2.34	2.97	1.44	1.87	1.09	1.42	0.5	0.6

4 结 论

针对立体匹配算法在病态区域匹配效果不佳的问题,提出了一种基于多重注意力机制的立体匹配算法,以多种注意力模块在多层嵌入的方式增强网络利用图像信息的能力。不同层嵌入的位置通道注意力模块通过池化编码产生通道响应的方式实现特征通道的调整,并由MCA和3D MCA模块建立特征间的联系,充分利用不同范围内的场景上下文信息获得更精确的匹配特征。同时通过加权损失函数,网络更关注错误区域,提高了整体匹配精度。在KITTI2012和KITTI2015这两个室外的公开数据集上对算法进行验证,与近年来基于神经网络的经典算法相比,所提算法取得最优的整体精度,特别是与基准方法相比,有效改善了病态区域的匹配效果,在精度上有一定提升。

相较于基准方法,所提算法的运行速度虽然有所提升,但在小型设备中仍无法满足实时性的需求。为得到运行效率更高的立体匹配算法,在未来的研究中需要对网络结构进行进一步优化。

参 考 文 献

[1] 陈炎, 杨丽丽, 王振鹏. 双目视觉的匹配算法综述[J]. 图学学报, 2020, 41(5): 702-708.
Chen Y, Yang L L, Wang Z P. Literature survey on stereo vision matching algorithms[J]. Journal of Graphics, 2020, 41(5): 702-708.

[2] Žbontar J, LeCun Y. Computing the stereo matching cost with a convolutional neural network[C]//2015 IEEE Conference on Computer Vision and Pattern Recognition, June 7-12, 2015, Boston, MA, USA. New York: IEEE Press, 2015: 1592-1599.

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

[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, October 22-29, 2017, Venice, Italy. New York: IEEE Press, 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, USA. New York: IEEE Press, 2018: 5410-5418.

[6] 刘建国, 纪郭, 颜伏伍, 等. 基于视差优化的立体匹配网络[J]. 计算机工程, 2022, 48(3): 220-228.
Liu J G, Ji G, Yan F W, et al. Stereo matching network based on disparity optimization[J]. Computer Engineering, 2022, 48(3): 220-228.

[7] 王玉锋, 王宏伟, 刘宇, 等. 基于多任务学习的立体匹

配算法[J]. 激光与光电子学进展, 2021, 58(4): 0415010.
Wang Y F, Wang H W, Liu Y, et al. Algorithm for stereo matching based on multi-task learning[J]. Laser & Optoelectronics Progress, 2021, 58(4): 0415010.

[8] 王玉锋, 王宏伟, 刘宇, 等. 渐进细化的实时立体匹配算法[J]. 光学学报, 2020, 40(9): 0915002.
Wang Y F, Wang H W, Liu Y, et al. Real-time stereo matching algorithm with hierarchical refinement[J]. Acta Optica Sinica, 2020, 40(9): 0915002.

[9] Hu J, Shen L, Sun G. Squeeze-and-excitation networks [C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, June 18-23, 2018, Salt Lake City, UT, USA. New York: IEEE Press, 2018: 7132-7141.

[10] Woo S, Park J, Lee J Y, et al. CBAM: convolutional block attention module[M]//Ferrari V, Hebert M, Sminchisescu C, et al, Computer vision-ECCV 2018. Lecture notes in computer science. Cham: Springer, 2018, 11211: 3-19.

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

[12] 程鸣洋, 盖绍彦, 达飞鹏. 基于注意力机制的立体匹配网络研究[J]. 光学学报, 2020, 40(14): 1415001.
Cheng M Y, Gai S Y, Da F P. A stereo-matching neural network based on attention mechanism[J]. Acta Optica Sinica, 2020, 40(14): 1415001.

[13] 张亚茹, 孔雅婷, 刘彬. 多维注意力特征聚合立体匹配算法[J/OL]. 自动化学报: 1-12[2021-05-03]. https://doi.org/10.16383/j.aas.c200778.
Zhang Y R, Kong Y T, Liu B. Multi dimensional attention feature aggregation stereo matching algorithm[J/OL]. Acta Automatica Sinica: 1-12[2021-05-03]. https://doi.org/10.16383/j.aas.c200778.

[14] 张文, 邵小桃, 杨维, 等. 基于卷积神经网络的高效精准立体匹配算法[J]. 计算机辅助设计与图形学学报, 2020, 32(1): 45-53.
Zhang W, Shao X T, Yang W, et al. An efficient and accurate stereo matching algorithm based on convolutional neural network[J]. Journal of Computer-Aided Design & Computer Graphics, 2020, 32(1): 45-53.

[15] 黄继辉, 张荣芬, 刘宇红, 等. 一种优化的深度学习立体匹配算法[J]. 激光与光电子学进展, 2021, 58(24): 2433002.
Huang J H, Zhang R F, Liu Y H, et al. Optimized deep learning stereo matching algorithm[J]. Laser & Optoelectronics Progress, 2021, 58(24): 2433002.

[16] 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): 161501.
Gong W, Qin L, Ren G F, et al. Binocular stereo matching algorithm based on multidimensional feature fusion[J]. Laser & Optoelectronics Progress, 2020, 57(16): 161501.

[17] Hou Q B, Zhou D Q, Feng J S. Coordinate attention for efficient mobile network design[C]//2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 20-25, 2021, Nashville, TN, USA. New York: IEEE Press, 2021: 13708-13717.

[18] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[EB/OL]. (2017-06-12)[2021-05-03]. https://arxiv.org/abs/1706.03762v3.

[19] Huang Z L, Wang X G, Huang L C, et al. CCNet: criss-cross attention for semantic segmentation[C]//2019 IEEE/CVF International Conference on Computer Vision, October 27- November 2, 2019, Seoul, Korea(South). New York: IEEE Press, 2019: 19398730.

[20] Shen Z L, Dai Y C, Rao Z B. MSMD-net: deep stereo matching with multi-scale and multi-dimension cost volume[EB/OL]. (2020-06-23)[2021-05-03]. https://arxiv.org/abs/2006.12797.

[21] Misra D. Mish: a self regularized non-monotonic neural activation function[EB/OL]. (2019-08-23)[2021-05-03]. https://arxiv.org/abs/1908.08681v2.

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

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

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

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

[26] Wu Z Y, Wu X Y, Zhang X P, et al. Semantic stereo matching with pyramid cost volumes[C]//2019 IEEE/CVF International Conference on Computer Vision (ICCV), October 27-November 2, 2019, Seoul, Korea (South). New York: IEEE Press, 2019: 7483-7492.