激光与光电子学进展

# 基于多尺度注意力特征融合的双目深度估计算法

杨蕙同，雷亮*，林永春
广东工业大学物理与光电工程学院，广东 广州 510006

**摘要** 针对目前端到端的立体匹配算法在具有挑战性的复杂场景中出现的误匹配现象，提出一种多尺度注意力特征融合立体匹配算法(MGNet)。设计了一个轻量级的组相关注意力模块，该模块采用组相关融合单元来有效地结合空间注意机制与通道注意力机制，同时捕获丰富的全局上下文信息和远距离通道依赖关系。设计了多尺度卷积全局注意力模块，它能够在多个尺度下处理局部和全局信息，在全局特征处理阶段引入非局部操作，可以同时捕获多尺度上下文与全局上下文，提供丰富的语义信息。在代价聚合阶段引入通道注意力，抑制具有歧义的匹配信息，提取有区别性的特征。使用三大数据集评估了所提算法的有效性，由实验结果可知，所提算法在薄结构、反射区域、弱纹理、重复纹理等复杂场景均表现优异。

**关键词** 深度估计；立体匹配；深度学习；注意力机制
**中图分类号** TP391.41 **文献标志码** A **DOI**：10.3788/LOP202259.1815005

# Binocular Depth Estimation Algorithm Based on Multi-Scale Attention Feature Fusion

Yang Huitong, Lei Liang*, Lin Yongchun
*School of Physics & Optoelectronic Engineering, Guangdong University of Technology, Guangzhou* 510006, *Guangdong, China*

**Abstract** This research proposes a multi-scale attention feature fusion stereo matching algorithm (MGNet) to address the mismatching phenomenon of the current end-to-end stereo matching algorithm in challenging and complex scenes. A lightweight group-related attention module was designed. This module uses group-related fusion units to effectively combine the spatial and channel attention mechanisms while capturing rich global context information and long-distance channel dependencies. The designed multi-scale convolutional global attention module can process local information and global information at multiple scales, add non-local operations in the global feature processing stage. The module captures multi-scale and global contexts simultaneously, providing rich semantic information. In the cost aggregation stage, channel attention was introduced to suppress ambiguous matching information and extract differentiated information. Three datasets were used to analyze the proposed algorithm's effectiveness. The results indicate that the proposed algorithm performs effectively in morbid areas like thin structures, reflective areas, weak textures, and repeating textures.

**Key words** depth estimation; stereo matching; deep learning; attention mechanism

## 1 引 言

真实世界场景的深度估计在计算机视觉、场景理解、图像和视频增强、自动驾驶、三维重建等领域有着广泛的应用[1-4]。例如，精确的深度估计可产生清晰的前景-背景分割，从而将场景中感兴趣的前景(近)对象与背景(远)对象分离。前景背景分割可用于目标检测、跟踪和语义分割[5]。传统立体匹配算法包括4个步骤：匹配代价计算、代价聚合、视差优化和后处理[6]。最近的研究提出了许多不同的方法[7-9]来实现与相邻像素的匹配代价计算。例如，Zabih等[10]将局部变换引入到匹配代价计算中，并提出了Census算法，其主要思想是利用局部区域中像素值的相对顺序进行统计。

**收稿日期**：2021-07-15；**修回日期**：2021-07-18；**录用日期**：2021-07-20
**基金项目**：国家自然科学基金(61675050)
**通信作者**：*leiliang@gdut.edu.cn

近年来，卷积神经网络（CNN）表现出较强的特征理解能力。Žbontar 等[11]首先将 CNN 应用于立体匹配，计算匹配代价。CNN 从图像中提取特征，并计算像素块之间的相似度得分。匹配代价体由代价聚合模块和半全局匹配模块处理。由于 CNN 可以显著提升立体匹配任务的效果，许多基于神经网络的算法被提出，但大多数算法只是利用 CNN 来解决相似度计算问题[11-12]。最近的研究[13-19]表明，基于深度学习的端到端立体匹配算法极大地提高匹配的精度和速度。DispNet[13]将传统算法思想引入到端到端的立体匹配网络中，对匹配特征进行编码。GC-Net[14]将 3DCNN 引入到立体匹配网络中，以聚合代价体。GC-Net 利用 3DCNN 构建了一种堆叠的编解码结构，以更好地利用上下文信息。PSMNet[15]使用空间金字塔池化模块来提取上下文信息。在文献[20]中，提出了一种新的亚像素卷积方法，用于补偿传统的插值上采样造成的特征信息损失。尽管基于 CNN 的算法在处理立体匹配任务时性能得到很大提高，但由于网络并不能捕获充分的上下文信息和多尺度信息，在复杂场景中进行像素的视差估计仍然存在一些困难。

上下文信息可以理解为目标对象与其周围像素之间的关系，充分地利用它可以更好地估计复杂场景像素的视差。因此，为了进行更精确的匹配，需要引入全局上下文信息，为此本文设计了组相关注意力融合模块，该模块将信道的维度划分为多组子特征。对于每组子特征，该模块同时构建通道注意力和空间注意力机制。该模块设计了一个注意力掩模，以抑制图像中的噪声，并突出显著的语义特征区域。来自不同尺度的特征融合可以弥补深层网络的信息损失。

高层特征具有丰富的语义信息，但分辨率较低，对细节的感知能力较差，因此关键是将高层次特征恢复到高分辨率，并与低层次特征进行融合。本文设计了多尺度卷积全局注意力模块，它包含全局和局部两个多尺度特征提取阶段和一个多尺度融合模块，每个特征提取阶段有不同大小的卷积核，除了扩大接收感受野，该模块还引入非局部操作，以捕获不同级别特征的全局上下文信息。综上所述，本文的贡献主要表现为几个方面：提出了一种轻量级的组相关注意力融合模块，该模块可以同时捕获全局上下文信息与远距离依赖关系，并跨通道交互全局语义信息；设计了多尺度卷积全局注意力模块，该模块可以捕获从局部到全局不同级别的上下文信息；采用 3D 注意力聚合模块替代常规的 3D 卷积，不增加额外计算量的同时，重新校准来自不同通道的匹配信息；MGNet 轻量而高效，可以在具有挑战性的复杂场景中预测准确的视差值。MGNet 在 KITTI2015 排行榜中所有区域的像素误差率（D1-all）为 2.01%。

## 2　多尺度注意力卷积网络架构

整体网络结构如图 1 所示，包括三个阶段，即特征提取、三维代价聚合和视差回归。首先采用类似于 PSMNet 的残差结构提取特征，再将特征分 16 组，每组特征分别通过并行的空间注意力层与通道注意力层，然后对 16 组具有不同语义信息的特征进行融合。此外，引入多尺度卷积全局注意力模块，对特征图同步应用局部和全局两组具有不同空间分辨率和深度的卷积操作，在全局卷积操作中引入自注意力机制。通过级联的方式联系左右图像的特征，构建代价体，并采用由 3D 卷积构成的编解码结构对代价体进行聚合，最后采用视差回归的方式生成视差图。

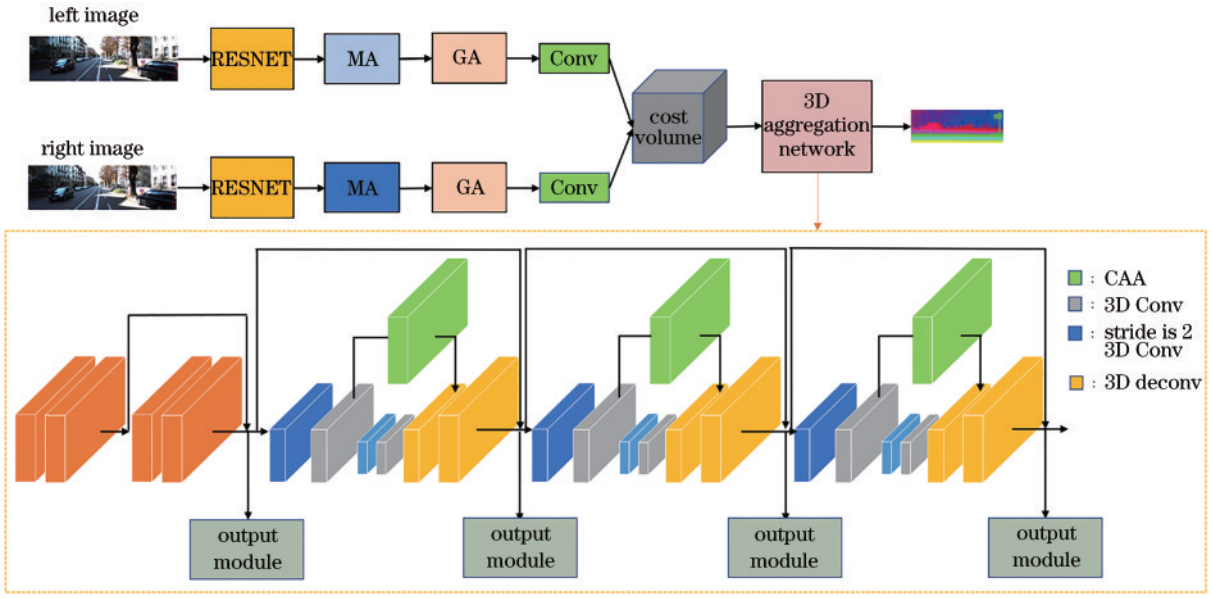


图 1　多尺度注意力融合网络总体结构

Fig. 1　Overall structure of multi-scale attention fusion network

### 2.1　组相关注意力融合模块

首先介绍组相关注意力融合模块(GA)的构建过程。该模块将输入特征图分组,并使用注意力融合模块将通道注意力和空间注意力整合为每组一个单元;然后,对所有单元的特征进行聚合,并利用通道融合操作实现不同单元的特征之间的信息交互。GA 的总体架构如图 2 所示。

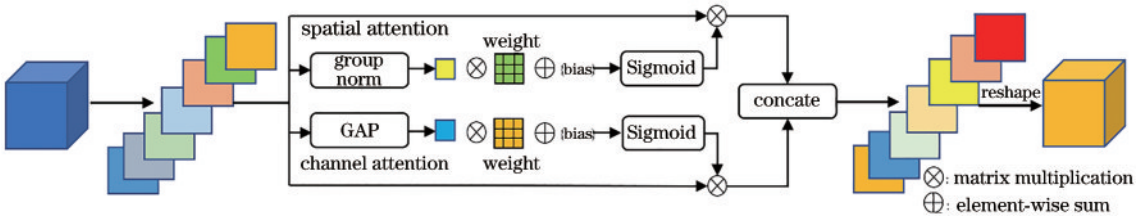


图 2　组相关注意力融合模块

Fig. 2　Group-related attention fusion module

#### 2.1.1　特征分组

对于给定的特征图 $\boldsymbol{A}\in\mathbf{R}^{C\times H\times W}$,其中 $C$、$H$ 和 $W$ 分别表示特征图通道数量、特征图高度和宽度,GA 首先沿着通道维度将 $\boldsymbol{A}$ 分为 $G$ 组($G$=16),$\boldsymbol{A}=[\boldsymbol{A}_1,\cdots,\boldsymbol{A}_G]$,$\boldsymbol{A}_n\in\mathbf{R}^{C/G\times H\times W}$,每组特征 $\boldsymbol{A}_n$ 沿着通道维度被分为两个分支,$\boldsymbol{A}_{n1}$ 和 $\boldsymbol{A}_{n2}\in\mathbf{R}^{C/2G\times H\times W}$,其中一个分支通过通道注意力机制捕获特征的长距离依赖关系,另一个分支通过空间注意力机制捕获全局上下文信息。

#### 2.1.2　通道注意力机制

不同于需要消耗大计算量的 SENet,ECANet 使用自适应卷积核大小为 $k$ 的 1D 卷积构建了一个更轻量级的通道注意力模块,在速度和精度方面均有提升,但当输入特征图的通道数量更多时,卷积核 $k$ 和网络计算量也会随之变大。因此针对这个问题,改进了 ECANet,提出的通道注意力模块更适用于立体匹配任务。首先通过全局平均池化(GAP)来嵌入全局信息,通过在 $H$,$W$ 维度缩放 $\boldsymbol{A}_{n1}$,生成形如 $\boldsymbol{b}\in\mathbf{R}^{C/2G\times H\times W}$ 的特征向量:

$$\boldsymbol{b}=\frac{1}{H\times W}\sum_{i=1}^{H}\sum_{j=1}^{W}\boldsymbol{A}_{n1}(i,j)。\tag{1}$$

引入自适应调优操作,不断调整每组特征的权重及偏置量,调优后的特征经过 Sigmoid 操作形成通道注意力机制:

$$\boldsymbol{A}_{n1}'=\sigma\left(\boldsymbol{W}_1\boldsymbol{b}+\boldsymbol{b}_{\mathrm{ia}1}\right)\cdot\boldsymbol{A}_{n1}\ ,\tag{2}$$

式中:$\boldsymbol{W}_1\in\mathbf{R}^{2G\times1\times1}$ 和 $\boldsymbol{b}_{\mathrm{ia}1}\in\mathbf{R}^{2G\times1\times1}$ 分别是对每组组内特征调优的权重和偏置量;$\sigma$ 是 Sigmiod 函数操作。

#### 2.1.3　空间注意力机制

与通道注意力不同,空间注意力机制的作用是捕获全局上下文,它是通道注意力机制的互补操作。首先对第二分支的特征引入 GroupNorm 操作,归一化每组特征,生成具有组内全局相关性的注意力掩模。与通道注意力操作类似,对空间维度的特征同样采用自适应调优操作。空间注意力最终的输出表示为

$$\boldsymbol{A}_{n2}'=\sigma\left\{\left[\boldsymbol{W}_2\mathrm{GroupNorm}(\boldsymbol{A}_{n2})\right]+\boldsymbol{b}_{\mathrm{ia}2}\right\}\cdot\boldsymbol{A}_{n2}\ ,\tag{3}$$

式中:$\boldsymbol{W}_2$ 与 $\boldsymbol{b}_{\mathrm{ia}2}$ 分别是对每组特征自适应调优的权重和偏置量,随着训练过程进行,不断调整空间注意力权值。

将两个注意力模块的输出逐像素相加,使输出特征图同时具有远距离依赖关系和全局上下文信息,融合后的特征与输入的分组特征尺寸一致,$\boldsymbol{A}_n'=[\boldsymbol{A}_n',\boldsymbol{A}_{n2}']\in\mathbf{R}^{C/G\times H\times W}$。

最后执行类似于 ShuffleNetv2 的通道操作,使 16 组不同的语义信息在通道维度上能够跨群组交互传播。类似于 DANet[21],GA 也采用两个不同功能的注意力机制,融合具有不同语义信息的特征,然后再通过通道注意力模块为融合特征生成对应的权重,从而达到联系全局特征的效果。

### 2.2　多尺度卷积全局注意力模块

提出的多尺度卷积全局注意力模块(MA)框架如图 3 所示,主要由三部分构成,即局部多尺度卷积操作(LP)、全局多尺度卷积操作(GP)、全局-局部融合操作(M L-G)。MA 能够处理特征细节,同时考虑多尺度上下文信息,在多个层级上处理局部和全局信息。

#### 2.2.1　局部多尺度卷积操作

LP 是一个局部多尺度上下文聚合模块,应用了具有不同空间大小和深度的卷积层,主要处理图像中较小的对象以及在多个尺度上捕获局部精细结构,如图 3 所示。首先应用 1×1 卷积层将特征图通道数量降低至 64,然后执行四层局部卷积操作,在卷积核为 9×9、7×7、5×5、3×3 四个尺度的卷积层上捕获不同层级的局部细节。此外,由于采用不同的分组数量(1,4,8,16),四个卷积核采用不同的连接方式。最后,利用一个 1×1 卷积层对多尺度信息进行融合。

#### 2.2.2　全局多尺度卷积操作

GP 是一个多尺度全局聚合模块,可以捕获场景中的全局信息,并处理图像中的大尺寸目标。GP 的组成部分如图 3 所示。由于训练图像和测试图像大小不一致,为了确保 GP 可以捕获到完整的全局信息,首先采用一种自适应的平均池化层,将特征图的尺寸降低至 9×18(参考 KITTI2015 训练图像等比例缩放)。然后采用 1×1 卷积层将特征图通道数量降低至 64。采用

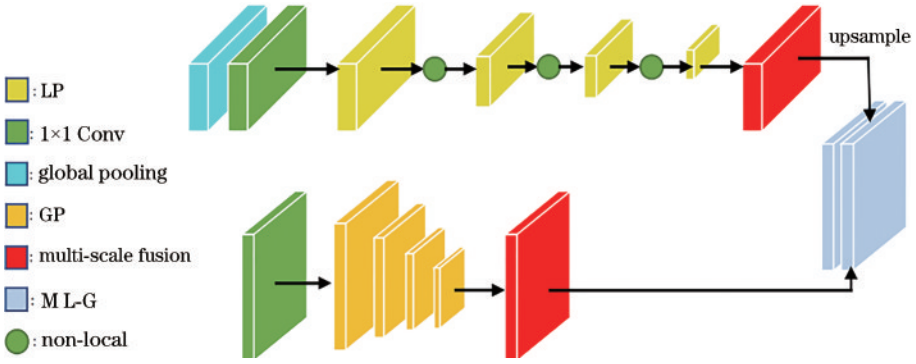


图 3　多尺度卷积全局注意力模块

Fig. 3　Multi-scale convolution global attention module

卷积核为 9×18、7×14、5×10、3×6 的卷积层完全覆盖每层特征图从而捕获完整的全局信息，此外在每层卷积操作后，采用全局空间注意力机制，捕获特征图任意两个位置之间的空间依赖性，为 GP 引入丰富的全局上下文信息。全局空间注意力机制可表示为

$$A_{ji}=\frac{\exp(B_i\cdot C_j)}{\sum_{i=1}^{P}\exp(B_i\cdot C_j)},\tag{4}$$

式中：$[\boldsymbol{B},\boldsymbol{C}]\in\mathbf{R}^{C\times H\times W}$ 为每个全局卷积层的输出特征图；$i$ 和 $j$ 代表图像中不同位置的像素；$P$ 为图像中像素总数。最后采用 1×1 卷积对不同尺度的信息进行融合，并使用双线性插值将特征图上采样到池化操作之前的尺寸。

#### 2.2.3　全局-局部融合操作

由于 GP 与 LP 已经捕获了所有尺度的上下文信息，M L-G 的作用仅是集中合并来自不同尺度的上下文信息。首先执行一个 3×3 的卷积操作将融合的特征通道维度从 256 降低至 128，然后采用 1×1 卷积细化特征并适当扩大感受野。如图 3 所示，所提框架能够捕获多尺度局部和全局信息，增强特征的表示能力。

### 2.3　3D 通道注意力聚合模块

在代价聚合阶段引入 3D 通道注意力聚合模块（CAA），以获取高质量的特征信息，模块结构如图 4 所示。

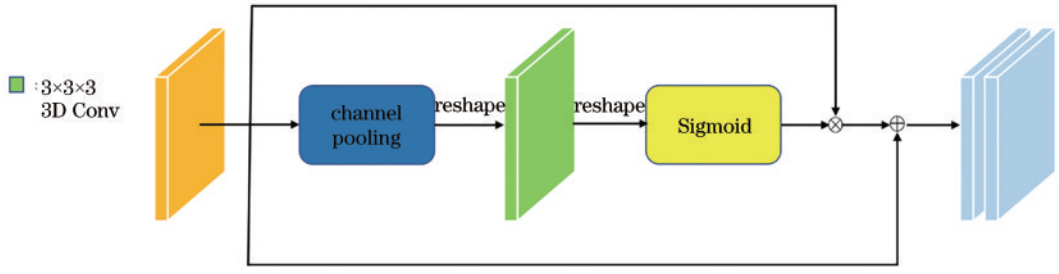


图 4　3D 通道注意力聚合模块

Fig. 4　3D channel attention aggregation module

首先，CAA 生成一个注意力掩模 $\boldsymbol{M}\in\mathbf{R}^{N\times D\times 1\times H\times W}$，用于在所有通道维度与输入特征 $\boldsymbol{X}\in\mathbf{R}^{N\times D\times C\times H\times W}$ 之间进行逐元素乘积。如图 4 所示，首先在通道维度对输入特征 $\boldsymbol{X}$ 进行平均池化操作，得到一个全局的通道注意力张量 $\boldsymbol{F}\in\mathbf{R}^{N\times D\times 1\times H\times W}$。然后将 $\boldsymbol{F}$ 转换为 $\boldsymbol{F}'\in\mathbf{R}^{N\times D\times 1\times H\times W}$ 并将其输入到一个 3×3×3 的 3D 卷积层 $K$ 中，将 $K$ 输出的结果转换回原来的维度 $\boldsymbol{F}''\in\mathbf{R}^{N\times D\times 1\times H\times W}$，并将其送入 Sigmoid 函数，生成通道注意力掩模 $\boldsymbol{M}$。整个过程可以表示为

$$\boldsymbol{Y}=\boldsymbol{X}+\boldsymbol{X}\cdot\boldsymbol{M}\ 。\tag{5}$$

与常规的 3D 卷积层相比，CAA 先对特征进行通道平均池化，然后将结果输入 3D 卷积层，其运算量大大减少，每个通道都可以感知视差维度特征信息的重要性，提升每个阶段的特征表示能力，从而获得更有效的图像特征匹配，以进行视差估计。

### 2.4　代价体构建与代价聚合

采用级联的方式，在每个视差级别上连接左特征图和对应的右特征图，形成一个 4D 代价体（$H\times W\times D\times S$），其中 $S$ 为特征尺寸；同时参考 GwcNet[22] 组相关的形式，将左侧特征和右侧特征沿通道维度划分为若干组，计算每组之间的相关映射，获得多个匹配成本代价；然后将这些匹配成本代价压缩成一个成本量，与上述基于级联的代价体共同构成本文的匹配代价体。为了捕获更丰富的上下文信息，使用堆叠的具有编-解码器结构的 3D 沙漏网络聚合代价体。3D 卷积的引入可以为立体匹配任务带来更准确的视差预测。

### 2.5　视差回归与损失函数

引入在 PSMNet[15] 中提出的视差回归来预测视差

图。输出的特征图大小为$(H, W, D+1)$，$D$表示最大视差。使用Softmax操作$\sigma(\cdot)$，可以从预测的成本$C_d$中计算出每个视差$d$的概率。另外，使用每个视差$d$的概率加权和来计算预测视差。视差回归定义为

$$\text{soft}\arg\min = \sum_{d=0}^{D} d \times \sigma(-C_d), \tag{6}$$

使用具有真实距离信息的点云数据作为Ground Truth来训练模型。由于点云标签是稀疏的，因此对像素损失值取平均。采用Smooth L1损失函数来训练网络，因为它不易受异常视差值的影响。定义损失函数为

$$L(d, \hat{d}) = \frac{1}{N'} \sum_{i=1}^{N'} \text{smooth L1}(d_i - \hat{d}_i), \tag{7}$$

式中：$d$为真值视差；$\hat{d}$为预测视差；$N'$为所有标记像素的总数。

## 3 实　　验

将设计模型生成的视差图可视化，并对所提算法与其他先进方法在SceneFlow数据集[23]、KITTI stereo2015数据集[24]和KITTI stereo2012数据集[25]上的实验结果进行比较，以证明所提算法的优越性。此外，将对设计的模块进行多次消融实验。

### 3.1 数据集与实验细节

使用三个公共数据集来训练和测试网络。

1) SceneFlow

SceneFlow合成立体数据集由35454个训练立体图像对和4370个测试立体图像对组成。图像尺寸为$H=540$，$W=960$。该数据集提供了复杂而密集的Ground Truth视差图。如果视差大于实验中设定的上限，将在损失计算中舍弃视差较大的像素。对于SceneFlow数据集，使用端点误差(EPE)来评估各个模块，EPE是预测视差和视差真值之间的欧氏距离度量，以像素为单位。

2) KITTI2015

该数据集采集自真实的世界场景，源于一辆行驶中的汽车捕捉到的动态街景。它通过在线排行榜提供200个具有稀疏视差真值的立体图像对用于训练，200个没有视差真值的立体图像对用于测试。数据集的图像分辨率为$H=1242$，$W=374$。在该数据集中，对背景、前景以及所有像素的视差异常值百分比进行评估。

3) KITTI2012

该数据集源于真实的街道场景，包括194个训练立体图像对和195个没有视差真值的测试立体图像对，大小均为$H=376$，$W=1280$。为了提高网络的性能，将全部194个图像对作为训练集。对于该数据集，采用视差误差大于$t$个像素的误差像素占比($>t$ pixel)评估模型。

在两张Nvidia 3090 GPU上对模型进行训练，使用PyTorch深度学习框架。采用Adam作为优化器($\beta_1=0.9$，$\beta_2=0.99$)，在训练过程中将图像随机裁剪为$H=256$，$W=512$。对于SceneFlow，共训练16个周期，初始学习率为0.001，之后每两个周期学习率降低一半。对于KITTI数据集，对包括训练集、验证集的所有图像对训练1000个周期，初始学习率为0.001，每100个周期学习率降低一半。

### 3.2 SceneFlow 实验

#### 3.2.1 消融实验

在相同的实验条件下对设计的各个模块进行消融实验。分别使用>1 pixel、>2 pixel、>3 pixel、全部区域的像素误差率(D1-all)、EPE评估每个模块的有效性，实验结果如表1所示，GA+MA+CAA模块组合相比MA在SceneFlow数据集上表现更好，EPE从0.757%降低至0.662%，此外GA+MA+CAA具有更小的D1-all，为0.0226。

表1 SceneFlow数据集上的消融实验结果

Table 1 Ablation study results on SceneFlow dataset

| Module | | | >1 pixel | >2 pixel | >3 pixel | D1-all | EPE / % |
|---|---|---|---|---|---|---|---|
| GA | MA | CAA | | | | | |
| | √ | | 0.0809 | 0.0438 | 0.0319 | 0.0260 | 0.757 |
| √ | √ | | 0.07780 | 0.0429 | 0.0316 | 0.0258 | 0.746 |
| √ | √ | √ | 0.0702 | 0.0384 | 0.0281 | 0.0226 | 0.662 |

#### 3.2.2 定量评估

所提模型在梯子缝隙和摩托车轮胎处可以产生稠密和清晰的视差图，如图5方框所示。并且相比于基准网络PSMNet，所提算法产生的视差图更接近于Ground Truth。这证明所提MGNet在重复纹理和薄结构区域可以预测精准的视差，可以通过关联全局上下文对具有不同视差值的像素进行分离，显著突出目标物体的轮廓细节。

表2展示了MGNet的EPE与其他先进算法误差值的对比情况，MGNet的EPE比PSMNet降低0.428个百分点，相比SegStereo降低了0.788个百分点，结果表明所提算法有效地降低了立体匹配任务的误匹配率。

### 3.3 KITTI2015数据集实验

#### 3.3.1 消融实验

取消验证集，对全部200张图像进行训练。同时分别用GA、GA+MA、GA+MA+Gwc、GA+MA+Gwc+CAA构成的模型对KITTI2015评估网站提供的200张测试图像进行预测，并将最终视差结果提交网站。如表3所示，分别对比了四个模块的>3 pixel，由测评网站最终的结果表明，所提GA+MA+Gwc+CAA的>3 pixel比仅含MA的基础模型降低了0.19个百分点。

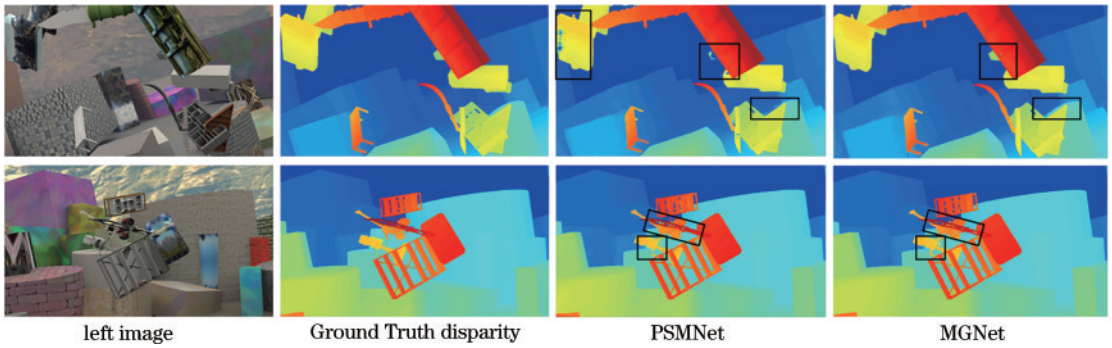


图 5　不同算法在 SceneFlow 数据集中得到的视差图

Fig. 5　Parallax maps obtained by different algorithms on SceneFlow dataset

表 2　MGNet 与其他方法的 EPE 对比

Table 2　Comparison of EPE between MGNet and other methods

| Parameter | MCCNN | GCNet | iResNeti2 | CRL | PSMNet | EdgeStereo | SegStereo | MGNet |
|---|---|---|---|---|---|---|---|---|
| EPE /% | 3.79 | 1.84 | 1.40 | 1.32 | 1.09 | 1.11 | 1.45 | 0.662 |

表 3　设计模块在 KITTI2015 数据集上的测评结果

Table 3　Benchmark results of designed module on KITTI2015 dataset

| GA | MA | Gwc | CAA | >3 pixel /% |
|---|---|---|---|---|
| √ | | | | 2.20 |
| √ | √ | | | 2.18 |
| √ | √ | √ | | 2.06 |
| √ | √ | √ | √ | 2.01 |

将 KITTI2015 的部分测试集的视差预测结果可视化。如图 6 方框区域所示，最佳模型为树干和光滑的车窗提供了稠密又清晰的视差图，同时整体视差图效果优于仅包含其余的模块组合。GA 的引入，有助于提高对小目标的视差预测精度，同时为弱纹理区域提供充分的上下文信息。CAA 关联各通道的视差相关性，为光滑的物体表面提供精准的匹配特征信息。由此可以证明所设计模块的有效性，MGNet 在图像的薄结构细节区域和无纹理区域均表现出优越的性能。在误差图中，蓝色表示正确的视差估计，黄色表示错误的估计。

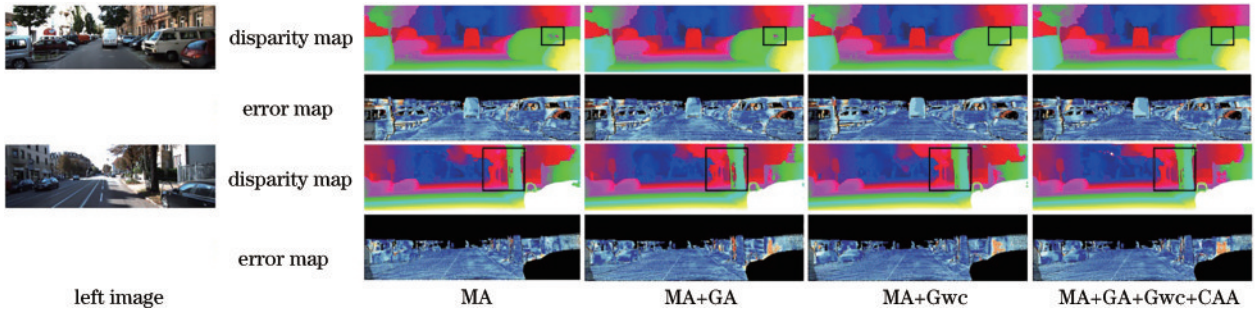


图 6　KITTI2015 测试集上消融实验可视化结果

Fig. 6　Visualization results of ablation experiment on KITTI2015 test set

### 3.3.2　定量评估

比较了 MGNet 与 DispNetC[13]、PDSNet[26]、GC-Net[14]、CRL[27]、PSMNet[15]、EdgeStereo[28]、Big3D[29]、AANet[30]等先进算法的立体匹配效果，表 4 提供了所有模型的性能评估结果，其中评价指标包括所有区域(ALL)和非遮挡区域(Noc)的背景像素的 3 像素误差(D1-bg)、前景像素的 3 像素误差(D1-fg)、所有区域的 3 像素误差(D1-all)。在所有比较方法中，除了前景像素(D1-fg)指标，MGNet 在其他所有误差测评指标中均取得了最佳性能；此外所提算法的 D1-all 为 2.01%，相比 Big3D 减小 0.2 个百分点，相比 AANet 减小 0.54 个百分点，这充分证明了 MGNet 可以利用全局上下文抑制具有歧义性视差信息的特征，突出显著性匹配信息，为各种场景提供精准的视差预测。

如图 7 方框区域所示，由于 MGNet 引入了 MA 模块，该模块可以结合不同尺度相似目标的信息，利用不同尺度卷积核提取不同层次的语义信息，使网络可以完好地保留图像中不同尺度标识牌的轮廓结构；此外还引入了 GA，该模块可以同时对图像中的长距离依赖关系以及全局上下文进行建模，赋予特征更丰富的语义信息；CAA 模块使网络模型更加有效地将正确的匹配信息聚合到具有挑战性的无纹理或反射的区域中，从而获得精确的视差估计。相对于 PSMNet，MGNet 几乎可以完全滤除掉汽车玻璃的反光现象。

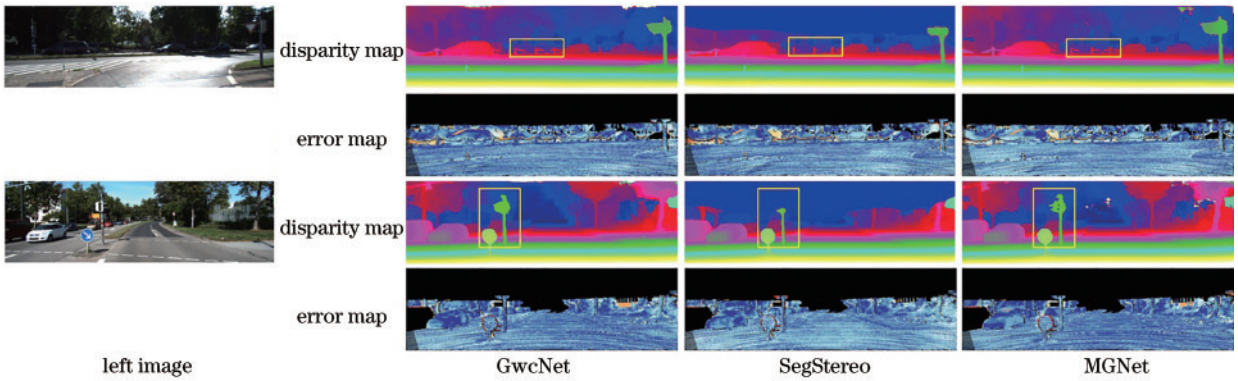


图 7　不同网络在 KITTI2015 数据集上的定量评估结果

Fig. 7　Qualitative evaluation results of different networks on KITTI2015 dataset

表 4　不同网络在 KITTI2015 数据集上的对比结果

Table 4　Comparison of different networks on KITTI2015 dataset　　unit: %

| Network | ALL | | | Noc | | |
|---|---|---|---|---|---|---|
| | D1-bg | D1-fg | D1-all | D1-bg | D1-fg | D1-all |
| DispNetC | 4.32 | 4.41 | 4.34 | 4.11 | 3.72 | 4.05 |
| CRL | 2.48 | 3.59 | 2.67 | 2.32 | 3.12 | 2.45 |
| PDSNet | 2.29 | 4.05 | 2.58 | 2.09 | 3.68 | 2.36 |
| GCNet | 2.21 | 6.16 | 2.87 | 2.02 | 5.58 | 2.61 |
| PSMNet | 1.86 | 4.62 | 2.32 | 1.71 | 4.31 | 2.14 |
| AANet | 1.99 | 5.39 | 2.55 | 1.80 | 4.93 | 2.32 |
| EdgeStereo | 2.27 | 4.18 | 2.59 | 2.12 | 3.85 | 2.40 |
| Big3D | 1.95 | 3.48 | 2.21 | 1.79 | 3.11 | 2.01 |
| MGNet | 1.65 | 3.84 | 2.01 | 1.51 | 3.49 | 1.84 |

实验结果表明 MGNet 在反射、复杂纹理、弱纹理区域表现良好，非常适用于立体匹配任务。

### 3.4　KITTI2012 数据集实验

对于 KITTI2012，将全部 194 张图像用于训练。将最佳模型的测试结果提交给在线评估网站。对测试集的评估结果如表 5 所示，MGNet 在所有区域 ALL 中的＞3 pixel 指标相比 SegStereo[31] 减小 0.27 个百分点，而且 MGNet 在非遮挡区域 Noc 均达到了最佳性能。

如图 8 方框所示，MGNet 预测的视差图比 SegStereo 更清晰平滑，即使在细小结构边缘与反射表面等挑战性区域，MGNet 仍然提供了高水平的性能，并显著优于其他先进算法。可视化结果表明，MGNet 可以显著降低小目标边缘（电线杆和围墙）的匹配错误率，并且由于 GA 可以为多组特征提取空间注意力掩模，空间注意力的长距离建模大大增强，使 MGNet 能很好地处理范围较大的无纹理区域（商店玻璃窗），这表明了引入全局语义信息和多尺度上下文信息的重要

表 5　不同网络在 KITTI2015 数据集上的对比结果

Table 5　Comparison of different networks on KITTI2012 dataset　　unit: %

| Network | ＞2 pixel | | ＞3 pixel | | ＞4 pixel | | ＞5 pixel | |
|---|---|---|---|---|---|---|---|---|
| | Noc | ALL | Noc | ALL | Noc | ALL | Noc | ALL |
| DispNetC | 7.38 | 8.11 | 4.11 | 4.65 | 2.77 | 3.20 | 2.05 | 2.39 |
| PDSNet | 3.82 | 4.65 | 1.92 | 2.53 | 1.38 | 1.85 | 1.12 | 1.51 |
| GCNet | 2.71 | 3.46 | 1.77 | 2.30 | 1.36 | 1.77 | 1.12 | 1.46 |
| PSMNet | 2.44 | 3.01 | 1.49 | 1.89 | 1.12 | 1.42 | 0.90 | 1.15 |
| Edgestereo | 2.79 | 2.43 | 1.73 | 2.18 | 1.30 | 1.64 | 1.04 | 1.32 |
| SegStereo | 2.66 | 3.19 | 1.68 | 2.03 | 1.25 | 1.52 | 1.00 | 1.21 |
| SSPCVNET | 2.47 | 3.09 | 1.47 | 1.90 | 1.08 | 1.41 | 0.87 | 1.14 |
| EdgestereoV2 | 2.32 | 2.88 | 1.46 | 1.83 | 1.07 | 1.34 | 0.83 | 1.04 |
| AANet | 2.30 | 2.96 | 1.55 | 2.04 | 1.20 | 1.58 | 0.98 | 1.30 |
| MGNet | 2.12 | 2.71 | 1.34 | 1.76 | 1.01 | 1.34 | 0.82 | 1.08 |

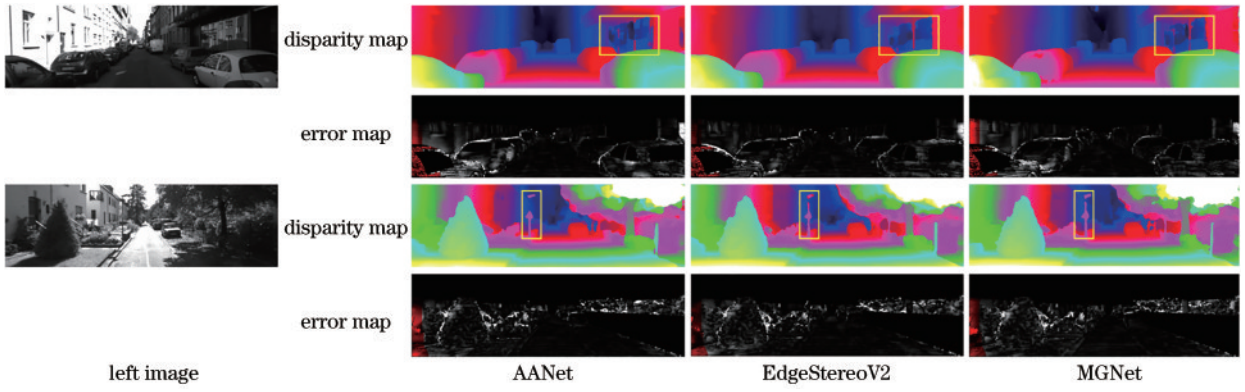


图 8　不同网络在 KITTI2012 数据集上的定量评估结果

Fig. 8　Qualitative evaluation results of different networks on KITTI2012 dataset

性。在误差图中，红色像素表示遮挡区域的错误估计，白色像素表示在非遮挡区域的错误估计。总之，MGNet在立体匹配任务中表现出了强大的性能。

### 3.5 Middlebury-v3 实验

为了进一步验证MGNet的鲁棒性，使用在SceneFlow数据集中训练过16个周期的模型，对Middlebury-v3的部分测试集生成视差图，该数据集采集自真实世界场景，分辨率为715×492。如图9方框所示，与PSMNet和EdgeStereo的视差图相比，MGNet涉及GA、MA和CAA，通过整合全局语义信息，在全局场景信息和目标细节方面保留了更完整的结构。

经实验证明，在一张3090GPU上，PSMNet对一组图像的预测时间为0.49 s，MGNet的预测时间为0.53 s，在提升精度的同时引入了一定的计算量。

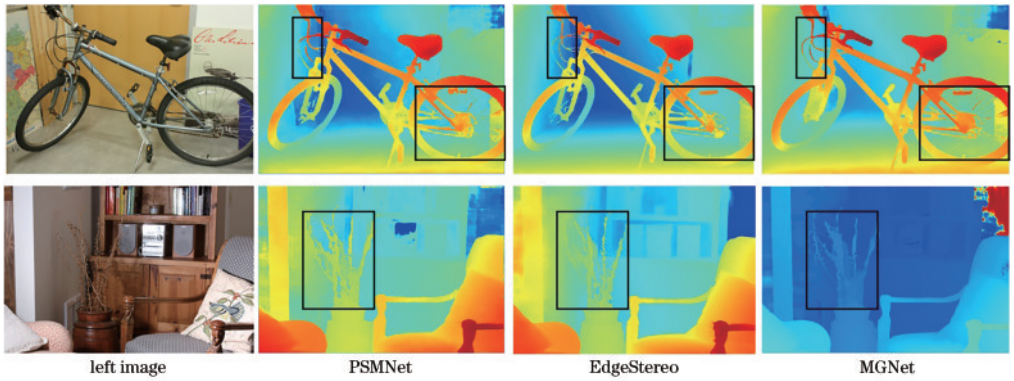


图9 不同网络在Middlebury-v3数据集上的定量评估结果

Fig. 9 Qualitative evaluation results of different networks on Middlebury-v3 dataset

## 4 结 论

提出一种多尺度注意力特征融合立体匹配算法，采用极轻量级的组相关注意力融合模块，捕获空间和通道维度的特征依赖关系。引入多尺度卷积全局注意力模块，捕获图像中不同层次的细节，同时非局部操作的设计为每个层级的特征嵌入丰富的全局上下文信息。相比于国内外先进算法，所提算法在代价聚合阶段使用一种可以感知时空信息的3D通道注意力机制，为代价体嵌入更丰富的跨通道交互信息。所提算法在挑战性区域提供了更准确的深度预测。

## 参 考 文 献

[1] Kerl C, Sturm J, Cremers D. Robust odometry estimation for RGB-D cameras[C]//2013 IEEE International Conference on Robotics and Automation, May 6-10, 2013, Karlsruhe, Germany. New York: IEEE Press, 2013: 3748-3754.

[2] Shotton J, Girshick R, Fitzgibbon A, et al. Efficient human pose estimation from single depth images[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2013, 35(12): 2821-2840.

[3] Gupta S, Girshick R, Arbeláez P, et al. Learning rich features from RGB-D images for object detection and segmentation[M]//Fleet D, Pajdla T, Schiele B, et al. Computer vision-ECCV 2014. Lecture notes in computer science. Cham: Springer, 2014, 8695: 345-360.

[4] Li H Y, Dai B, Shi S S, et al. Feature intertwiner for object detection[EB/OL]. (2019-03-28) [2021-02-06]. https://arxiv.org/abs/1903.11851.

[5] Zhang H, Zhang H, Wang C G, et al. Co-occurrent features in semantic segmentation[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: 548-557.

[6] Scharstein D, Szeliski R, Zabih R. A taxonomy and evaluation of dense two-frame stereo correspondence algorithms[C]//Proceedings IEEE Workshop on Stereo and Multi-Baseline Vision (SMBV 2001), December 9-10, 2001, Kauai, HI, USA. New York: IEEE Press, 2001: 131-140.

[7] Mei X, Sun X, Dong W M, et al. Segment-tree based cost aggregation for stereo matching[C]//2013 IEEE Conference on Computer Vision and Pattern Recognition, June 23-28, 2013, Portland, OR, USA. New York: IEEE Press, 2013: 313-320.

[8] Yang Q X. A non-local cost aggregation method for stereo matching[C]//2012 IEEE Conference on Computer Vision and Pattern Recognition, June 16-21, 2012, Providence, RI, USA. New York: IEEE Press, 2012: 1402-1409.

[9] Zhang K, Lu J B, Lafruit G. Cross-based local stereo matching using orthogonal integral images[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2009, 19(7): 1073-1079.

[10] Zabih R, Woodfill J. Non-parametric local transforms for computing visual correspondence[M]//Eklundh J O. Computer vision-ECCV '94. Lecture notes in computer science. Heidelberg: Springer, 1994, 801: 151-158.

[11] Žbontar J, LeCun Y. Stereo matching by training a convolutional neural network to compare image patches [EB/OL]. (2015-10-20)[2021-03-02]. https://arxiv.org/

abs/1510.05970.

[12] Shaked A, Wolf L. Improved stereo matching with constant highway networks and reflective confidence learning[C]//2017 IEEE Conference on Computer Vision and Pattern Recognition, July 21-26, 2017, Honolulu, HI, USA. New York: IEEE Press, 2017: 6901-6910.

[13] Güney F, Geiger A. Displets: resolving stereo ambiguities using object knowledge[C]//2015 IEEE Conference on Computer Vision and Pattern Recognition, June 7-12, 2015, Boston, MA, USA. New York: IEEE Press, 2015: 4165-4175.

[14] 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.

[15] 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.

[16] 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.

[17] 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.

[18] 陈其博, 葛宝臻, 李云鹏, 等. 基于多重注意力机制的立体匹配算法[J]. 激光与光电子学进展, 2022, 59(16): 1633001.
Chen Q B, Ge B Z, Li Y P, et al. Stereo matching algorithm based on multi attention mechanism[J]. Laser & Optoelectronics Progress, 2022, 59(16): 1633001.

[19] Wang Q, Liu X C, Liu W, et al. MetaSearch: incremental product search via deep meta-learning[J]. IEEE Transactions on Image Processing, 2020, 29: 7549-7564.

[20] Shi W Z, Caballero J, Huszár F, et al. Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network[C]//2016 IEEE Conference on Computer Vision and Pattern Recognition, June 27-30, 2016, Las Vegas, NV, USA. New York: IEEE Press, 2016: 1874-1883.

[21] Fu J, Liu J, Tian H J, et al. Dual attention network for scene segmentation[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: 3141-3149.

[22] 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.

[23] 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.

[24] Geiger A, Lenz P, Urtasun R. Are we ready for autonomous driving? The KITTI vision benchmark suite [C]//2012 IEEE Conference on Computer Vision and Pattern Recognition, June 16-21, 2012, Providence, RI, USA. New York: IEEE Press, 2012: 3354-3361.

[25] Menze M, Geiger A. Object scene flow for autonomous vehicles[C]//2015 IEEE Conference on Computer Vision and Pattern Recognition, June 7-12, 2015, Boston, MA, USA. New York: IEEE Press, 2015: 3061-3070.

[26] Tulyakov S, Ivanov A, Fleuret F. Practical Deep Stereo (PDS): toward applications-friendly deep stereo matching [EB/OL]. (2018-06-05)[2021-02-03]. https://arxiv.org/abs/1806.01677.

[27] 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. New York: IEEE Press, 2017: 878-886.

[28] Song X, Zhao X, Hu H W, et al. EdgeStereo: a context integrated residual pyramid network for stereo matching [M]//Jawahar C V, Li H D, Mori G, et al. Computer vision-ACCV 2018. Lecture notes in computer science. Cham: Springer, 2019, 11365: 20-35.

[29] 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.

[30] Xu H F, Zhang J Y. AANet: adaptive aggregation network for efficient stereo matching[C]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 13-19, 2020, Seattle, WA, USA. New York: IEEE Press, 2020: 1956-1965.

[31] 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.