多维注意力特征聚合立体匹配算法

张亚茹 [1] 孔雅婷 [2] 刘 彬 [1]

摘 要 现有基于深度学习的立体匹配算法在学习推理过程中缺乏有效信息交互, 而特征提取和代价聚合两个子模块的特征维度存在差异, 导致注意力方法在立体匹配网络中应用较少、方式单一. 针对上述问题, 本文提出了一种多维注意力特征聚合立体匹配算法. 设计 2D 注意力残差模块, 通过在原始残差网络中引入无降维自适应 2D 注意力残差单元, 局部跨通道交互并提取显著信息, 为匹配代价计算提供丰富有效的特征. 构建 3D 注意力沙漏聚合模块, 以堆叠沙漏结构为骨干设计 3D 注意力沙漏单元, 捕获多尺度几何上下文信息, 进一步扩展多维注意力机制, 自适应聚合和重新校准来自不同网络深度的代价体. 在三大标准数据集上进行评估, 并与相关算法对比, 实验结果表明所提算法具有更高的预测视差精度, 且在无遮挡的显著对象上效果更佳.

关键词 深度学习, 立体匹配, 多维注意力机制, 信息交互

引用格式 张亚茹, 孔雅婷, 刘彬. 多维注意力特征聚合立体匹配算法. 自动化学报, 2022, **48**(7): 1805−1815

DOI 10.16383/j.aas.c200778

Multi-dimensional Attention Feature Aggregation Stereo Matching Algorithm

ZHANG Ya-Ru[1] KONG Ya-Ting[2] LIU Bin[1]

Abstract Existing deep learning-based stereo matching algorithms lack effective information interaction in the learning and reasoning process, and there is difference in feature dimension between feature extraction and cost aggregation, resulting in less and single application of attention methods in stereo matching networks. In order to solve these problems, a multi-dimensional attention feature aggregation stereo matching algorithm was proposed. The two-dimensional (2D) attention residual module is designed by introducing the adaptive 2D attention residual unit without dimensionality reduction into the original residual network. Local cross-channel interaction and extraction of salient information provide abundant and effective features for matching cost calculation. The three-dimensional (3D) attention hourglass aggregation module is constructed by designing a 3D attention hourglass unit with a stacked hourglass structure as the backbone. It captures multi-scale geometric context information and expand the multi-dimensional attention mechanism, adaptively aggregating and recalibrating cost volumes from different network depths. The proposed algorithm is evaluated on three standard datasets and compared with related algorithms. The experimental results show that the proposed algorithm has higher accuracy in predicting disparity and has better effect on unobstructed salient objects.

Key words Deep learning, stereo matching, multi-dimensional attention mechanism, information interaction

Citation Zhang Ya-Ru, Kong Ya-Ting, Liu Bin. Multi-dimensional attention feature aggregation stereo matching algorithm. *Acta Automatica Sinica*, 2022, **48**(7): 1805−1815

计算两个输入图像上对应像素的相对水平立体匹配对于理解或重建 3D 场景至关重要, 广泛应用于自动驾驶[1]、无人机[2]、医学成像和机器人智能控制等领域. 通常, 给定一对校正后的图像, 立体匹配的目标是位移, 即视差.

近年来, 基于深度学习的立体匹配算法研究已取得重大进展, 相比传统方法[3−4], 可从原始数据理解语义信息, 在精度和速度方面有着显著优势. 早期基于深度学习的方法[5−6] 是经卷积神经网络 (Convolutional neural network, CNN) 获得一维特征相关性度量之后, 采用一系列传统的后处理操作预测最终视差, 无法端到端网络训练. 随着全卷积神经网络 (Fully convolutional networks, FCN) 的发展[7], 研究者们提出了将端到端网络整合到立体匹配模型中[8−16]. 对于全卷积深度学习立体匹配网络, PSMNet[17] 提出一种空间金字塔池化模块, 扩大深层特征感受野, 提取不同尺度和位置的上下文信息. CFPNet[18]

收稿日期 2020-09-23 录用日期 2020-12-01
Manuscript received September 23, 2020; accepted December 1, 2020
河北省自然科学基金 (F2019203320) 资助
Supported by Natural Science Foundation of Hebei Province (F2019203320)
本文责任编委 吴建鑫
Recommended by Associate Editor WU Jian-Xin
1. 燕山大学信息科学与工程学院 秦皇岛 066004 2. 燕山大学电气工程学院 秦皇岛 066004
1. School of Information Science and Engineering, Yanshan University, Qinhuangdao 066004 2. School of Electrical Engineering, Yanshan University, Qinhuangdao 066004

在金字塔模块基础上引入扩张卷积和全局平均池化扩大感受野, 使其更有效地感知全局上下文信息. MSFNet[19] 利用多尺度特征模块, 通过级联融合不同层级的特征捕获细节和语义特征. 除了以上对特征提取网络的研究之外, 在代价聚合中, 第一个端到端视差估计网络 DispNet[20] 提出沿视差方向计算一维相关性的匹配代价计算方法. 由于仅沿着一个维度计算相关性, 损失了其余多个维度的有效信息, 因此为了更好地利用多维度的上下文语义特征, Kendall 等[21] 提出了 GC-Net, 通过采用 3D 编解码结构在三个维度上理解全局语义信息. 受 GC-Net 启发, 众多学者提出了多种变体来正则化代价体, 建模匹配过程, 例如, 结合 2D 和 3D 卷积运算的多维聚合子网络[22]、多尺度残差 3D 卷积模块[23]、堆叠 3D 沙漏结构[17, 24] 等. 尽管上述方法在视差估计中已取得长足进步, 但在网络学习推理过程中, 图像特征和代价体特征的多层级多模块交互利用仍存在不足, 缺乏全局网络信息的长距离依赖, 导致网络不具有敏锐的鉴别性能, 准确估计视差依然极具挑战性.

随着注意力机制在多种研究任务, 如语义分割[25]、自然语言处理[26]、超分辨率[27] 等方面的广泛应用, 注意力机制在立体匹配网络中引起了关注[28−30]. 其中, 基于 SE-Net[31] 的扩张空间金字塔注意力模块[28] 虽然采用降维减小了计算成本, 但是降维的同时导致特征通道与其权重之间的对应是间接的, 降低了通道注意力的学习能力. MRDA-Net[30] 只在 2D 特征提取网络末端和 3D 编解码网络末端引入单一池化 3D 注意力模块来整合全局信息, 无法做到多模块信息交互, 导致网络获取显著信息不充分. 综上, 由于 2D 图像特征为 3D 张量, 3D 代价体特征为四维张量, 两者之间的维度差异使常规注意力方法无法同时应用于特征提取与代价聚合这两个子模块中, 注意力机制在立体匹配网络中应用较少、方式单一, 从而整个立体匹配网络缺乏有效协同的注意力机制, 对长距离上下文信息无法做到多模块多层级关注.

考虑上述问题, 本文在 Gwc-Net[24] 的基础上提出一种多维注意力特征聚合立体匹配算法, 通过对特征提取和代价聚合两个子模块设计不同的注意力方法, 从多模块多层级的角度去理解关注整个网络传输过程中的上下文信息. 设计 2D 注意力残差模块, 使用无降维自适应 2D 通道注意力, 逐像素提取和融合更全面有效的信息特征, 学习局部跨通道间的相关性, 自适应关注通道间的区别信息. 提出 3D 注意力沙漏聚合模块, 利用 3D 平均池化和 3D 最大池化构建 3D 通道注意力, 将其嵌入多个子编解码块的末端, 重新校准来自不同模块的多个代价体, 整合多模块输出单元计算匹配代价.

1 多维注意力特征聚合立体匹配算法

所提算法主要包括 2D 注意力残差模块, 联合代价体, 3D 注意力沙漏聚合模块. 算法网络结构如图 1 所示. 2D 注意力残差模块对输入左图像 $I_{\rm l}$ 和右图像 $I_{\rm r}$ 进行特征提取, 将提取的特征用于构建联合代价体, 采用 3D 注意力沙漏聚合模块计算匹配代价, 最终通过视差回归函数输出预测视差.

1.1 2D 注意力残差模块

为保留网络的低级结构特征以提取左右图像的细节信息, 首先构建 3 个卷积核尺寸为 3×3 的滤波器获取浅层特征, 输出特征图尺寸为 $1/2H \times 1/2W \times 32$. 然后, 采用基本残差块 conv1_x, conv2_x, conv3_x 和 conv4_x 逐像素提取深层语义信息. 其中, conv1_x, conv2_x, conv3_x 和 conv4_x 包含的基本残差单元个数分别为 3, 16, 3 和 3. 每个残差块由两个卷积核尺寸为 3 × 3 的 2D 卷积、批归一化 (Batch normalization, BN) 层和线性整流 (Rectified linear unit, ReLU) 激活层组成[17]. 级联 conv2_x, conv3_x 和 conv4_x, 融合低级结构信

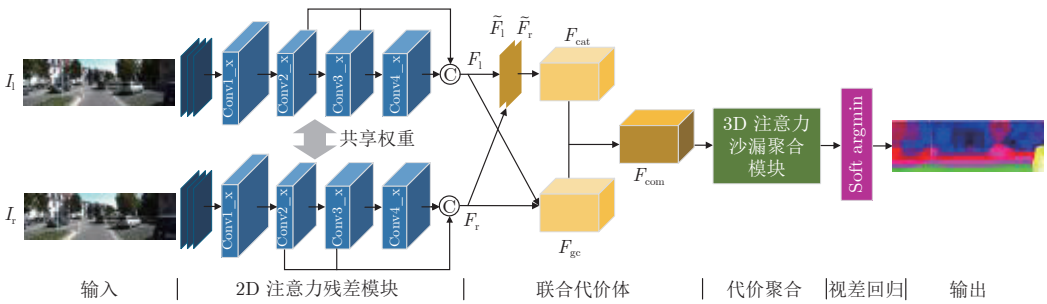


图 1 算法网络结构图

Fig. 1 Architecture overview of proposed algorithm

息和高级语义信息, 构建尺寸为$1/4H \times 1/4W \times 320$的特征表示. 该模块共 53 层, 输出左特征图 $F_{\rm l}$ 和右特征图 $F_{\rm r}$ 的尺寸均为$1/4H \times 1/4W \times 320$, 具体参数设置如表 1 所示.

表 1　2D 注意力残差单元和联合代价体的参数设置 (D 表示最大视差, 默认步长为 1)

Table 1　Parameter setting of the 2D attention residual unit and combined cost volume (D represents the maximum disparity. The default stride is 1)

层级名称	层级设置	输出维度
$F_{\rm l}/F_{\rm r}$	卷积核尺寸, 通道数, 步长	$H\times W\times 3$
	2D 注意力残差模块	
Conv0_1	3×3, 32, 步长 = 2	$1/2H \times 1/2W \times 32$
Conv0_2	3×3, 32,	$1/2H \times 1/2W \times 32$
Conv0_3	3×3, 32,	$1/2H \times 1/2W \times 32$
Conv1_x	$\begin{bmatrix}3\times 3, 32\\ 3\times 3, 32\end{bmatrix}\times 3$	$1/2H \times 1/2W \times 32$
Conv2_x	$\begin{bmatrix}3\times 3, 32\\ 3\times 3, 32\end{bmatrix}\times 16$, 步长 = 2	$1/4H \times 1/4W \times 64$
Conv3_x	$\begin{bmatrix}3\times 3, 32\\ 3\times 3, 32\end{bmatrix}\times 3$	$1/4H \times 1/4W \times 128$
Conv4_x	$\begin{bmatrix}3\times 3, 32\\ 3\times 3, 32\end{bmatrix}\times 3$	$1/4H \times 1/4W \times 128$
$F_{\rm l}$ / $F_{\rm r}$	级联: Conv2_x, Conv3_x, Conv4_x	$1/4H \times 1/4W \times 320$
	联合代价体	
$F_{\rm gc}$	—	$1/4D \times 1/4H \times 1/4W \times 40$
$\tilde{F}_{\rm l}$ / $\tilde{F}_{\rm r}$	$\begin{bmatrix}3\times 3, 128\\ 1\times 1, 12\end{bmatrix}$	$1/4H \times 1/4W \times 12$
$F_{\rm cat}$	—	$1/4D \times 1/4H \times 1/4W \times 24$
$F_{\rm com}$	级联: $F_{\rm gc}$, $F_{\rm cat}$	$1/4D \times 1/4H \times 1/4W \times 64$

PSMNet[17] 在特征提取过程中只采用单路径卷积方式, 没有对提取的特征进一步整合和交互, 缺乏信息之间的长距离依赖. 为自适应地增强特征表示, 在残差块中引入通道注意力, 强调重要特征并抑制不必要特征. 这种机制对每一通道赋予从 0 到 1 的不同权值, 代表各个通道的重要程度, 使得网络可以区别不同对象的特征图. 在之前的工作中, 通道注意力大多采用 SE-Net, 通过两次全连接层缩放所有特征图的通道维度. 然而, 缩放特征通道数量虽然在整合信息过程中大大减小了计算量, 但是降维的同时导致特征通道与其权重之间的对应是间接的, 降低了通道注意力的学习能力.

因此在 SE-Net[31] 的基础上, 设计无降维的注意力, 去除缩放通道[32]. 鉴于无降维会增加计算复杂度, 且通道特征具有一定的局部周期性, 本文构造无降维局部跨通道注意力, 在无降维的基础上, 通过局部约束计算通道之间的依赖关系, 并将其注意力嵌入每个残差块. 参照文献 [30], 通道注意力是针对 2D 图像特征的长和宽进行的 2D 卷积滤波操作, 因此命名为 2D 注意力. 2D 注意力残差单元结构如图 2 所示.

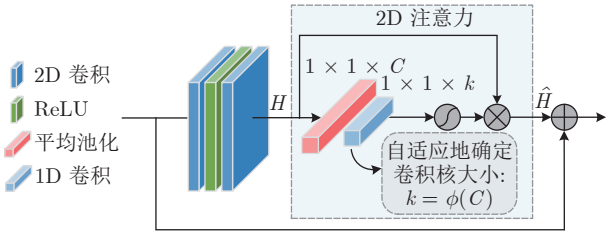


图 2　2D 注意力残差单元结构图

Fig. 2　2D attention residual unit architecture

设 2D 注意力输入特征图为 $H \in \mathbf{R}^{H\times W\times C}$, 在不降低通道维数的情况下, 执行全局平均池化和卷积核尺寸为 $1\times 1\times k$ 的一维 (One-dimensional, 1D) 卷积. 其中卷积核大小为 k 的 1D 卷积用来计算每个通道与其 k 个邻域间的相互作用, k 表示局部跨通道间的覆盖范围, 即有多少邻域参与一个通道的注意力预测. k 的大小可以通过一个与特征图通道个数 C 相关的函数自适应确定. 为进一步降低模型复杂度, 所有通道共享相同的权值, 该过程通过快速的 1D 卷积实现. 上述过程可表述为

$$\boldsymbol{s} = f_{\rm 1D}(z_{\rm avg}) \tag{1}$$

$$k = \phi(C) = \left|\frac{\log_2(C)+1}{2}\right|_{\rm odd} \tag{2}$$

式中, $z_{\rm avg}$ 是经 2D 全局平均池化生成的特征图, $z_{\rm avg} \in \mathbf{R}^{1\times 1\times C}$; $f_{\rm 1D}$ 是卷积核尺寸为 $1\times 1\times k$ 的 1D 卷积; s 表示使用 1D 卷积为各通道权重赋值后的张量, $s \in \mathbf{R}^{1\times 1\times C}$; ϕ 表示 k 和 C 之间的映射关系; $|p|_{\rm odd}$ 表示 p 的相邻最近奇数.

将具有不同通道权重的特征张量通过 sigmoid 激活函数归一化处理, 并与输入特征图的通道对应乘积, 实现对特征图自适应地重新校准

$$\hat{H} = H \times \sigma(s) \tag{3}$$

式中, σ 表示 sigmoid 激活函数; $\hat{H}$ 表示 2D 注意力的输出特征图, $\hat{H} \in \mathbf{R}^{H\times W\times C}$.

1.2　联合代价体

代价计算通常是计算 1D 相似性或者通过移位级联构建代价体, 前者损耗信息多, 后者计算成本高. 因此, 构建联合代价体 $F_{\rm com}$, 由分组关联代价体

分量 F_{gc} 和降维级联代价体分量 F_{cat} 构成, 其中 F_{gc} 和 F_{cat} 分别提供一维相似性度量和丰富的空间语义信息. 联合代价体结构如图 3 所示. 首先, 将包含 320 个通道的左特征图 F_l 和右特征图 F_r 沿特征维度等分为 n 组, 即每个特征组有 $320/n$ 个通道. 根据 Gwc-Net[24] 表明网络的性能随着组数的增加而增加, 且考虑到内存使用量和计算成本, 故设置为 $n = 40$.

$$F_{gc} = \frac{1}{\frac{320}{n}} \langle F_l^i, F_r^i \rangle \tag{4}$$

式中, $\langle \cdot, \cdot \rangle$ 表示点积运算; F_l^i 和 F_r^i 分别表示第 i 组的左特征图和右特征图.

针对代价体相比图像特征具有更多维度这一属性, 对提取的左右特征图降维以减少内存占用. 分别对输出尺寸均为 $1/4H \times 1/4W \times 320$ 的左特征图 F_l 和右特征图 F_r 依次执行卷积核尺寸为 3×3 和 1×1 的 2D 卷积操作, 得到具有 12 个特征维度的左特征图 $\tilde{F}_l$ 和右特征图 $\tilde{F}_r$, 进而级联获得降维级联代价体分量, 其通道维度为 24.

$$F_{cat} = \left\{ \tilde{F}_l, \tilde{F}_r \right\} \tag{5}$$

式中, $\{\cdot, \cdot\}$ 表示级联操作.

最后, 将 F_{gc} 和 F_{cat} 沿着通道维度堆叠形成尺寸为 $1/4D \times 1/4H \times 1/4W \times 64$ 的 F_{com}. 联合代价体的参数设置如表 1 所示.

1.3 3D 注意力沙漏聚合模块

堆叠 3D 沙漏模块与 Gwc-Net[24] 相同, 包含 1 个预处理结构和 3 个 3D 注意力沙漏结构, 以捕获不同尺度的上下文信息. 3D 注意力沙漏聚合模块结构如图 4 所示. 其中, 预处理结构由 4 个卷积核

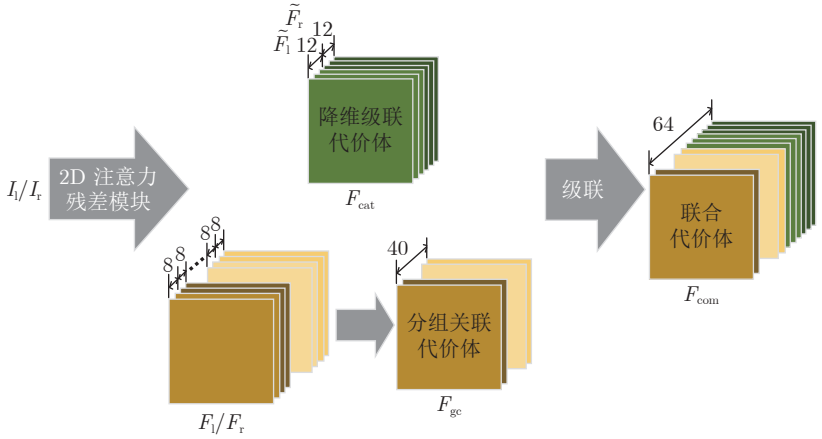


图 3 联合代价体结构图

Fig. 3 Combined cost volume architecture

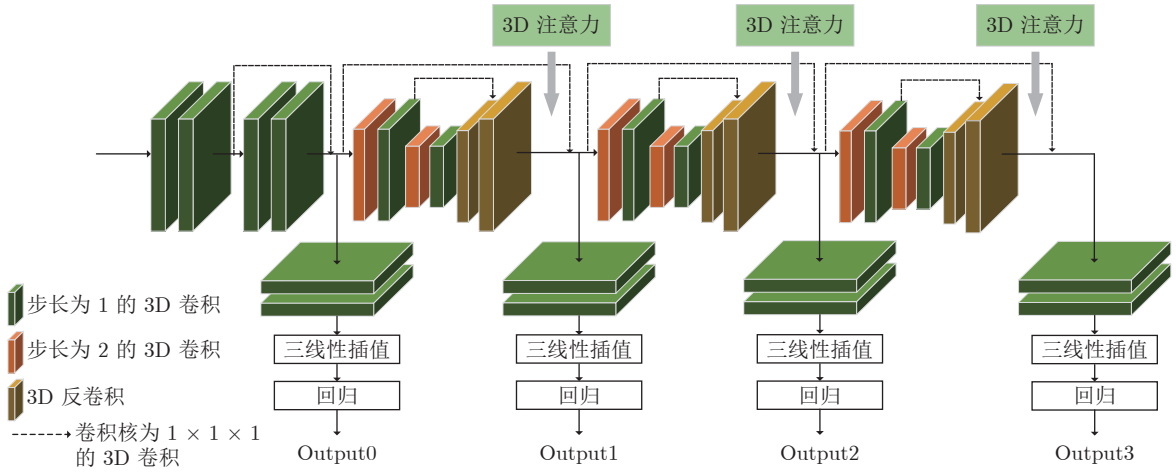


图 4 3D 注意力沙漏聚合模块结构图

Fig. 4 3D attention hourglass aggregation module architecture

为 3×3×3 的 3D 卷积层组成用于提取低级特征, 并为最终视差预测提供细节信息, 作为视差图的几何约束. 对于沙漏结构, 在编码部分执行 4 次卷积核为 3×3×3 的 3D 卷积, 在每一个步长为 1 的 3D 卷积层后紧接一个步长为 2 的 3D 卷积层进行下采样操作, 降低特征图分辨率的同时将通道数翻倍. 由于编码部分共两次下采样, 在解码部分相应执行两次上采样即两次卷积核为 3×3×3 的 3D 反卷积操作以恢复分辨率, 同时特征通道数减半, 并将第 2 个反卷积层的输出与编码器中同分辨率的特征级联. 此外, 使用卷积核为 1×1×1 的 3D 卷积将预处理结构和沙漏结构直连, 减少网络计算参数. 网络包括 Output0, Output1, Output2 和 Output3 共 4 个输出单元, 每一个输出单元执行两次卷积核为 3×3×3 的 3D 卷积, 并应用三线性插值恢复与输入图像大小相同的分辨率 $H \times W \times D$.

以往基于 CNN 的代价聚合算法[17, 24] 并未对代价体的通道信息进行多模块多层级关注, 无法有效利用和整合传输信息中的有效特征, 导致网络缺乏选择性鉴别信息特征和关注显著特征的能力. 此外, 代价聚合与特征提取中的特征维度存在差异, 使用特征提取模块中的无降维注意力增大代价体聚合的计算成本. 因此, 我们针对代价体特征的属性, 扩展通道注意力机制, 在堆叠 3D 沙漏结构[24] 的基础上, 对 3D 代价体特征的长、宽和深度共 3 个维度进行 3D 卷积滤波, 计算不同通道之间的相互依赖性. 为了区分特征提取模块中的 2D 注意力, 我们命名为 3D 注意力. 3D 注意力沙漏单元结构如图 5 所示, 沿着通道维度推断 3D 注意力特征图, 与输入代价体相乘, 细化代价体特征.

由于 3D 卷积滤波器具有感受局部视野的特性, 难以有效利用局部区域以外的上下文信息, 因此采用 3D 全局平均池化整合全局空间信息. 与文献 [30] 不同, 本文不仅使用 3D 全局平均池化, 而且使用 3D 最大池化编码特征图, 通过两种池化方式进一步区别对象的显著特征. 设 3D 注意力单元输入代价体为 $X \in \mathbf{R}^{D\times W\times H\times C}$, 首先在同一层级分别执行 3D 平均池化和 3D 最大池化获得两个尺寸为 $D\times 1\times 1\times C$ 的代价体特征图. 其次, 相比 2D 图像特征对应的 3D 张量, 代价体为四维张量, 故为了减少参数数量, 降低计算负担, 采用卷积核尺寸为 1×1×1 的 3D 卷积来整合所有通道间的信息, 压缩特征维度为 $C/16$, 再次执行同卷积核尺寸的 3D 卷积, 将特征维度恢复至 C. 上述过程可表示为

$$s_{\text{avg}} = f''_{1\times1\times1}(f'_{1\times1\times1}(u_{\text{avg}})) \tag{6}$$

$$s_{\max} = f''_{1\times1\times1}(f'_{1\times1\times1}(u_{\max})) \tag{7}$$

式中, u_{avg} 和 $u_{\max}$ 表示分别经 3D 平均池化和 3D 最大池化生成的特征图, u_{avg}, $u_{\max} \in \mathbf{R}^{D\times1\times1\times C}$; $f'_{1\times1\times1}$ 和 $f''_{1\times1\times1}$ 分别表示用于降维和升维的卷积核为 1×1×1 的 3D 卷积; s_{avg} 和 $s_{\max}$ 分别表示赋予通道不同权值的特征图, s_{avg}, $s_{\max} \in \mathbf{R}^{D\times1\times1\times C}$.

将 s_{avg} 和 $s_{\max}$ 逐像素相加, 采用 sigmoid 激活函数, 得到最终的 3D 注意力特征图

$$\hat{X} = X \times \sigma(s_{\text{avg}} + s_{\max}) \tag{8}$$

式中, $\hat{X}$ 表示 3D 注意力单元的输出特征图, $\hat{X} \in \mathbf{R}^{D\times H\times W\times C}$.

1.4　损失函数

所提算法的输出分别为 Output0, Output1, Output2 和 Output3, 对应的损失为 Loss0, Loss1, Loss2 和 Loss3. 在训练阶段, 总损失为 4 个损失的加权总和. 在测试阶段, 最终输出为 Output3, 损失为 Loss3. 视差估计采用 GC-Net[17] 提出的 soft-

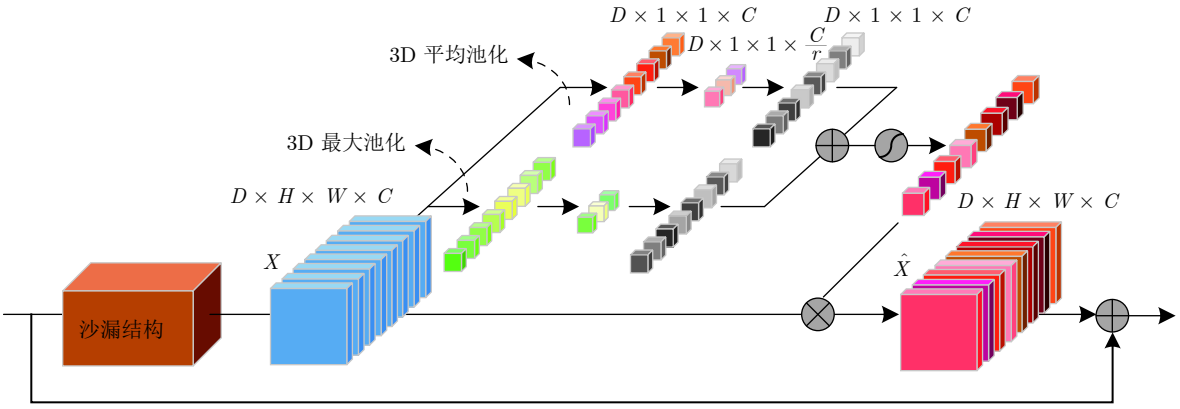


图 5　3D 注意力沙漏单元结构图

Fig. 5　3D attention hourglass unit architecture

argmin 方法, 将每个像素 i 的视差值与相应的概率 p_i 乘积求和获得预测视差 $\tilde{d}$

$$\tilde{d} = \sum_{i=0}^{D_{\max}-1} i \times p_i \tag{9}$$

式中, $D_{\max}$ 表示特征图的最大视差值.

最终的损失函数定义为

$$L(\tilde{d}, d) = \frac{1}{N}\sum_{k=0}^{3} \lambda_k \times L_1(\tilde{d}_i - d_i) \tag{10}$$

式中, N 表示标签像素的数量, $\tilde{d}_i$ 表示预测视差图, d_i 表示真值视差图. 平滑 L_1 损失函数表示为

$$L_1(\tilde{d}_i - d_i) = \begin{cases} 0.5(\tilde{d}_i - d_i)^2, & 若 \left|\tilde{d}_i - d_i\right| < 1 \\ \left|\tilde{d}_i - d_i\right| - 0.5, & 否则 \end{cases} \tag{11}$$

2 实验及结果分析

本文在公开数据集 SceneFlow[20], KITTI2015[33] 和 KITTI2012[34] 上进行实验分析, 并使用 EPE (End-point-error) 和 D1 等评价指标对所提算法进行评估. 其中, EPE 表示估计的视差值与真实值之间的平均欧氏距离; D1 表示以左图为参考图像预测的视差错误像素百分比.

2.1 数据集与实验细节

所提算法应用 PyTorch 深度学习框架实现, 在单个 Nvidia 2080Ti GPU 上进行训练和测试, 且设置批次大小为 2. 采用 Adam 优化器且设置 $\beta_1 = 0.9$, $\beta_2 = 0.999$. 在训练阶段, 随机将图像大小裁剪为 256×512. 使用的数据集如下:

1) SceneFlow 数据集. 是一个具有室内和室外场景的大型综合数据集, 包含 35 454 幅训练图像对和 4 370 幅测试图像对. 本文使用 SceneFlow 数据集的子数据集 Flyingthings3D, 其中, 训练图像对提供精细且密集的真值图. 图像的分辨率为 540×960, 最大视差为 192. 使用完整的数据集从头训练该模型, 以学习率 0.001 训练 10 个周期. 训练过程大约花费 56 小时, 训练的模型直接用于测试.

2) KITTI2015 数据集. 是一个室外驾驶真实场景数据集. 该数据集较小, 包含 200 幅训练图像对和 200 幅测试图像对. 其中, 训练图像对提供 LiDAR 获得的稀疏真值视差图, 测试图像对不提供真值视差图. 图像的分辨率为 375×1 242 像素, 最大视差为 128. 整个训练图像对被随机分成训练集 (80%) 和验证集 (20%). 使用 SceneFlow 数据集预训练的模型在 KITTI2015 上微调 300 个周期, 设置恒定学习率为 1×10^{-4}, 微调过程大约花费 20 小时.

3) KITTI2012 数据集. 与 KITTI2015 类似, KITTI2012 数据集只具有室外驾驶场景, 包含 194 幅训练图像对和 195 幅测试图像对, 提供训练图像对的稀疏真值视差图. 其微调过程与 KITTI2015 数据集一致.

2.2 超参数分析

本文分别在 KITTI2012 和 KITTI2015 数据集上对立体匹配网络进行消融实验, 定量评估 2D 注意力残差模块、3D 注意力沙漏聚合模块、联合代价体以及损失函数权重对立体匹配性能的影响.

1) 验证 2D 注意力残差模块的有效性

本文在不含 3D 注意力单元的情况下, 比较 4 种 2D 注意力的变体: 无 2D 注意力的残差网络, 具有最大池化层的降维 2D 注意力, 具有平均池化层的降维 2D 注意力[31] 和无降维自适应 2D 注意力. 表 2 给出了在 KITTI2015 数据集上 2D 注意力残差模块在不同设置下的性能评估结果, 其中 "> [n] px" 表示 EPE 大于 n 时的像素百分比, "✓"表示模块使用该结构. 由表 2 可知, 未添加 2D 注意力时 EPE 值仅为 0.631, 错误率明显高于其他 3 种方法, 无降维自适应 2D 注意力 EPE 值可达 0.615, 性能优于分别具有最大池化层和平均池化层的降维 2D 注意力. 实验结果表明, 所提出的 2D 注意力残差模块性能最优, 通过保持维度一致和局部跨通道间的信息交互, 有效提高了网络注意力, 有助于立体匹配任务降低预测视差误差.

表 2 2D 注意力残差模块在不同设置下的性能评估
Table 2 Performance evaluation of 2D attention residual module with different settings

网络设置	KITTI2015			
2D 注意力单元	> 1 px (%)	> 2 px (%)	> 3 px (%)	EPE (px)
—	13.6	3.49	1.79	0.631
最大池化 + 降维	12.9	3.20	1.69	0.623
平均池化 + 降维	12.7	3.26	1.64	0.620
✓	12.4	3.12	1.61	0.615

2) 验证 3D 注意力沙漏聚合模块的有效性

本文在 2D 注意力残差模块的基础上, 比较 4 种 3D 注意力的变体: 无 3D 注意力的原始沙漏聚合模块, 具有 3D 最大池化层的 3D 注意力, 具有 3D 平均池化层的 3D 注意力和同时使用两种池化方式的 3D 注意力. 表 3 给出了在 KITTI2012 和 KITTI2015 数据集上 3D 注意力沙漏聚合模块在不同设置下的性能评估. 由表 3 可知, 加入 3D 注意力后, 算法的 D1-all 和 EPE 值都明显降低, 证明具有 3D

表 3　联合代价体和 3D 注意力沙漏聚合模块在不同设置下的性能评估

Table 3　Evaluation of 3D attention hourglass aggregation module and combined cost volume with different settings

网络设置			KITTI2012		KITTI2015	
联合代价体	3D 注意力单元		EPE (px)	D1-all (%)	EPE (px)	D1-all (%)
	3D 最大池化	3D 平均池化				
✓	—	—	0.804	2.57	0.615	1.94
✓	✓	—	0.722	2.36	0.610	1.70
✓	—	✓	0.703	2.33	0.607	1.68
PSMNet[17]	✓	✓	0.867	2.65	0.652	2.03
✓	✓	✓	0.654	2.13	0.589	1.43

注意力的沙漏聚合模块优于原始沙漏聚合模块. 具有两种池化方式的 3D 注意力沙漏聚合模块在 KITTI2012 和 KITTI2015 数据集上 EPE 值分别达到 0.654 和 0.589, 其性能明显优于仅含单一池化的 3D 注意力沙漏聚合模块. 实验结果表明, 本文不能忽略 3D 最大池化的重要性, 其与 3D 平均池化一样有意义, 将两种池化方式结合可帮助立体匹配任务更多样地获取上下文信息.

3) 验证联合代价体的有效性

此外, 鉴于联合代价体是特征提取与代价聚合之间的枢纽, 本文将联合代价体与 PSMNet[17] 的级联代价体进行对比, 表 3 给出了在 KITTI2012 和 KITTI2015 数据集上联合代价体在不同设置下的性能评估. 从表中可以看出, 联合代价体相比 PSMNet[17] 的级联代价体, 增加了相关代价体分量的引导, 对于多维注意力聚合特征的性能产生了积极的作用, 整个网络结构相辅相成.

4) 验证不同损失函数权重对网络的影响

由于 3D 聚合网络有 4 个输出单元, 因此损失函数权重对网络的影响也至关重要. 本文将损失权重以 λ_1, λ_2, λ_3, λ_4 的顺序设置, 如图 6 所示. 当 $\lambda_1 = 0.5$, $\lambda_2 = 0.5$, $\lambda_3 = 0.7$, $\lambda_4 = 1.0$ 时, 越接近网络末端的损失计算对网络训练越重要, 同时网络其余子编码块的输出也对网络的性能起着辅助训练的作用, 使整个网络从前到后都能得到有效的误差回传, 多个子编码块的输出保证了网络的均衡训练.

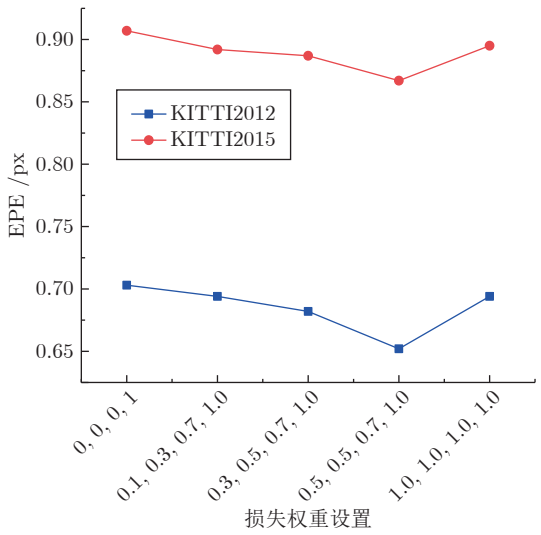


图 6　损失函数权重对网络的影响

Fig. 6　The influence of the weight of loss function on network performance

2.3　与其他方法的性能比较与分析

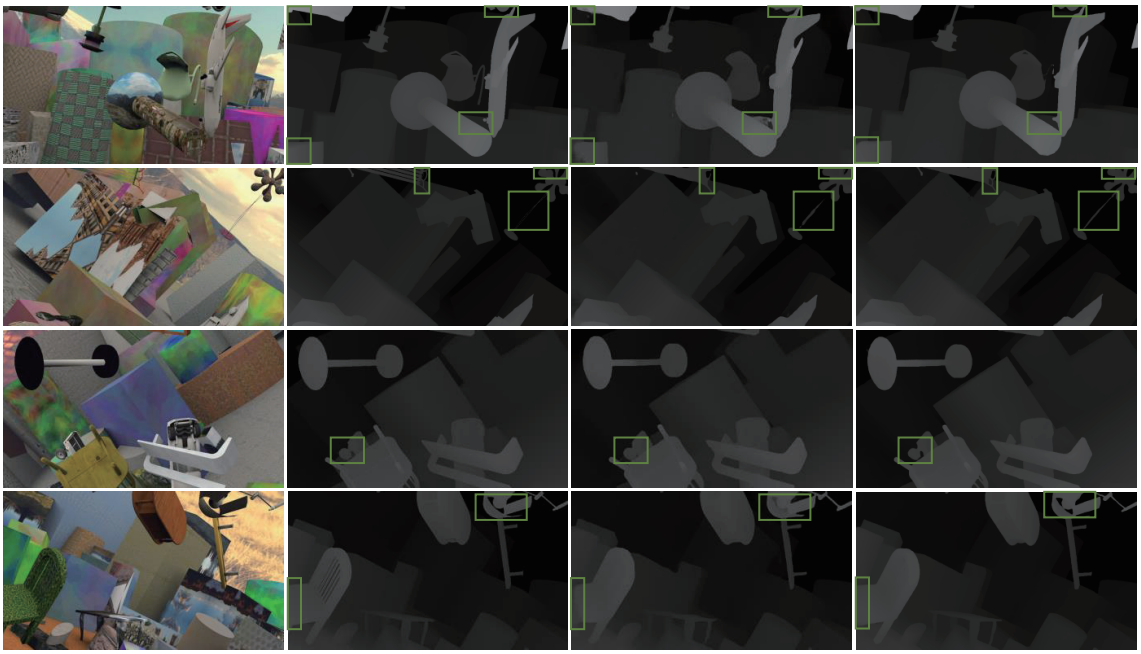
为进一步验证算法有效性, 在 SceneFlow 数据集上将所提方法与其他方法进行比较, 包括 Gwc-Net[24], PSMNet[17], MCA-Net[29], CRL[35] 和 GC-Net[21]. 定量评估结果如表 4 所示, 其中 px 表示像素. 所提算法在 SceneFlow 数据集上的 EPE 值均低于其他 5 种算法, 其中, 与 Gwc-Net 相比 EPE 降低了 0.055, 与 PSMNet 相比 EPE 降低了 0.28. 此外, 图 7 显示了在 SceneFlow 数据集上的视差评估结果, 其中图 7(c) Gwc-Net 表示不包含多维注意力的算法. 由图 7 中标注的小方框可看出, 具有多维注意力的算法能更有效地提取不同对象的显著变化特征. 因此所提算法能够为主体对象的显著特征分配更高的响应值, 提高网络的学习推理能力, 实现比未添加注意力时更精细的视差图.

表 4　不同算法在 SceneFlow 数据集上的性能评估

Table 4　Performance evaluation of different methods on the SceneFlow dataset

算法	EPE (px)
本文算法	0.71
Gwc-Net[24]	0.765
PSMNet[17]	1.09
MCA-Net[29]	1.30
CRL[35]	1.32
GC-Net[21]	2.51

表 5 反映了在 KITTI2015 数据集上本文算法



(a) 左图像 (a) Left image　(b) 真值图 (b) Ground truth map　(c) Gwc-Net 视差图 (c) Disparity map of Gwc-Net　(d) 所提算法视差图 (d) Our disparity map

图 7 SceneFlow 视差估计结果

Fig. 7 Results of disparity estimation on SceneFlow dataset

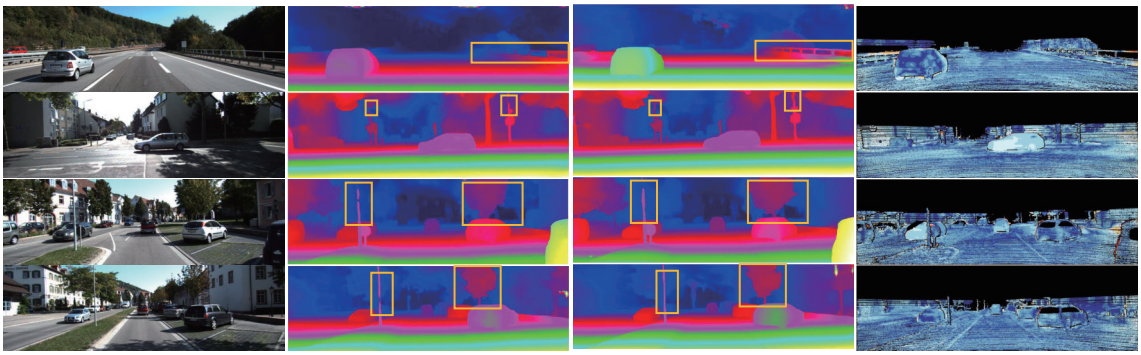
与 DispNetC[20], MC-CNN-art[36], CRL[35], PDSNet[37], GC-Net[21] 和 PSMNet[17] 的定量评估结果. 其中, "All"表示所有区域像素, "Noc"表示仅非遮挡区域的像素. 本文分别在背景区域 (bg)、前景区域 (fg) 和所有区域内 (all) 计算评价指标 D1 值. 由表 5 可知, 所提算法在所有区域和非遮挡区域的 D1-bg、D1-fg 和 D1-all 值都低于其他方法. 特别是在非遮挡部分, 相比于 PSMNet[17], 前景区域 D1-fg 值减小了 0.23, 远远大于背景区域 D1-bg 的减小值 0.07, 而且, 所有区域 D1-all 值也减小了 0.08, 说明所提算法对于非遮挡区域显著对象的预测精度具有明显提升, 在整体视差预测方面具有优越的性能. 图 8 显示了在 KITTI2015 数据集上视差估计结果. 由图可知, 在处理重复图案区域中 (如栅栏、道路) 匹配效果较好, 而且保留了对象的显著信息 (如车辆、电线杆和树干的边缘区域). 大的弱纹理区域 (如天空) 和被遮挡区域中由于没有可用于正确匹配的特征, 将存在很多噪声. 实验结果表明, 多维注意力通过聚集丰富的匹配信息可有效鉴别不同对象的显著特征, 提取更全面有效的特征, 降低匹配误差.

表 5 不同算法在 KITTI2015 上的性能评估 (%)

Table 5 Performance evaluation of different methods on the KITTI2015 dataset (%)

算法	All			Noc		
	D1-bg	D1-fg	D1-all	D1-bg	D1-fg	D1-all
DispNetC[20]	4.32	4.41	4.34	4.11	3.72	4.05
MC-CNN-art[36]	2.89	8.88	3.88	2.48	7.64	3.33
CRL[35]	2.48	3.59	2.67	2.32	3.12	2.45
PDSNet[37]	2.29	4.05	2.58	2.09	3.68	2.36
GC-Net[21]	2.21	6.16	2.87	2.02	5.58	2.61
PSMNet[17]	1.86	4.62	2.32	1.71	4.31	2.14
本文算法	1.72	4.53	2.30	1.64	4.08	2.06

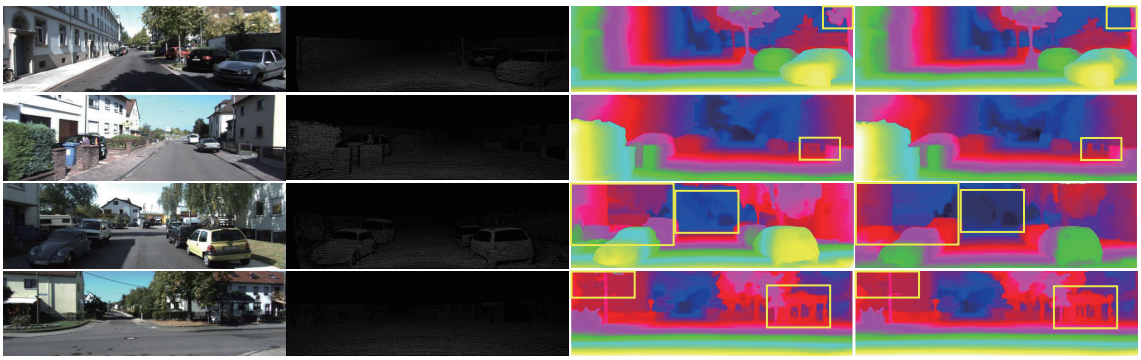
在 KITTI2012 数据集上的性能评估与 KITTI2015 类似. 表 6 反映了在 KITTI2012 数据集上本文算法与 DispNetC[20]、MC-CNN-acrt[36]、GC-Net[21]、SegStereo[11] 和 PSMNet[17] 的定量评估结果. 由表 6 可知, 本文算法与其他几种算法相比, 在无遮挡区域和所有区域中, 大于 3 像素和大于 5 像素的误差值均最低, 其中, 无遮挡区域分别为 1.46 和 0.81; 所有区域分别为 1.73 和 0.90, 再次证明基于多维注意力特征聚合的立体匹配网络在无遮挡区域和所有区域中视差预测的有效性和可行性. 图 9 显示了在 KITTI2012 数据集上视差估计结果, 由图中标注的方框可看出本文算法在显著对象 (如围栏, 车辆) 方面视差预测结果较好, 且不受光线变化的影响. 此外, 从图 9 第 1 行最后一列的黄框可以看出, 尤其是对于墙壁这类很显著平滑的对象, 虽然 KITTI2012 数据集中真值视差图的稀疏性导致训练的网络模型对于树木的视差预测精度不高, 但相



(a) 左图像 (a) Left image　(b) Gwc-Net 视差图 (b) Disparity map of Gwc-Net　(c) 所提算法视差图 (c) Our disparity map　(d) 所提算法误差图 (d) Our error map

图 8　KITTI2015 视差估计结果

Fig. 8　Results of disparity estimation on KITTI2015 dataset



(a) 左图像 (a) Left image　(b) 真值图 (b) Ground truth map　(c) Gwc-Net 视差图 (c) Disparity map of Gwc-Net　(d) 所提算法视差图 (d) Our disparity map

图 9　KITTI2012 视差估计结果

Fig. 9　Results of disparity estimation on KITTI2012 dataset

表 6　不同算法在 KITTI2012 上的性能评估 (%)

Table 6　Performance evaluation of different methods on the KITTI2012 dataset (%)

算法	> 2 px		> 3 px		> 5 px		平均误差	
	Noc	All	Noc	All	Noc	All	Noc	All
DispNetC[20]	7.38	8.11	4.11	4.65	2.05	2.39	0.9	1.0
MC-CNN-acrt[36]	3.90	5.45	2.43	3.63	1.64	2.39	0.7	0.9
GC-Net[21]	2.71	3.46	1.77	2.30	1.12	1.46	0.6	0.7
SegStereo[11]	2.66	3.19	1.68	2.03	1.00	1.21	0.5	0.6
PSMNet[17]	2.44	3.01	1.49	1.89	0.90	1.15	0.5	0.6
本文算法	3.01	3.60	1.46	1.73	0.81	0.90	0.5	0.6

比 Gwc-Net, 本文算法预测的墙壁没有和树木混淆, 符合墙壁的属性. 实验结果表明, 本文算法具有良好的泛化性, 多维注意力的引入提高了网络的学习能力, 可有效鉴别无遮挡对象的显著特征, 提高视差预测精度.

3　结束语

本文提出了一种多维注意力特征聚合立体匹配算法, 以多模块及多层级的嵌入方式协同两种不同维度的注意力单元. 2D 注意力残差模块在原始残差网络基础上引入自适应无降维 2D 通道注意力, 理解局部跨通道间的相互依赖性, 保留显著细节特征, 为代价聚合过程提供了全面有效的相似性度量. 3D 注意力沙漏聚合模块在多个沙漏结构的基础上嵌入双重池化 3D 注意力单元, 捕获多尺度上下文信息, 有效提高了网络的聚合能力. 2D 注意力和 3D 注意力之间相辅相成, 对整个网络的权重修正, 误差回传都起到了积极的作用. 在 3 个公开数据集上的实验结果表明, 所提算法不仅具有较高的预测精度, 而且可以敏锐地鉴别推理无遮挡区域中主体对象的显著性特性.

References

1　Feng D, Rosenbaum L, Dietmayer K. Towards safe autonomous

driving: capture uncertainty in the deep neural network for lidar 3D vehicle detection. In: Proceedings of the 21st International Conference on Intelligent Transportation Systems. Maui, HI, USA: IEEE, 2018. 3266−3273

2 Schmid K, Tomic T, Ruess F, Hirschmüller H, Suppa M. Stereo vision based indoor/outdoor navigation for flying robots. In: Proceedings of the 2013 IEEE/RSJ International Conference on Intelligent Robots and Systems. Tokyo, Japan: IEEE, 2013. 3955−3962

3 Li Pei-Xuan, Liu Peng-Fei, Cao Fei-Dao, Zhao Huai-Ci. Weight-adaptive cross-scale algorithm for stereo matching. *Acta Optica Sinica*, 2018, **38**(12): 248−253
(李培玄, 刘鹏飞, 曹飞道, 赵怀慈. 自适应权值的跨尺度立体匹配算法. 光学学报, 2018, **38**(12): 248−253)

4 Han Xian-Jun, Liu Yan-Li, Yang Hong-Yu. A stereo matching algorithm guided by multiple linear regression. *Journal of Computer-Aided Design and Computer Graphics*, 2019, **31**(1): 84−93
(韩先君, 刘艳丽, 杨红雨. 多元线性回归引导的立体匹配算法. 计算机辅助设计与图形学学报, 2019, **31**(1): 84−93)

5 Zagoruyko S, Komodakis N. Learning to compare image patches via convolutional neural networks. In: Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition. Boston, MA, USA: IEEE, 2015. 4353−4361

6 Luo W, Schwing A G, Urtasun R. Efficient deep learning for stereo matching. In: Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, NV, USA: IEEE, 2016. 5695−5703

7 Long J, Shelhamer E, Darrell T. Fully convolutional networks for semantic segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017, **39**(4): 640−651

8 Mayer N, Ilg E, Hausser P, Fischer P, Cremers D, Dosovitskiy A, Brox T. A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation. In: Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, NV, USA: IEEE, 2016. 4040−4048

9 Song X, Zhao X, Fang L J, Hu H W. Edgestereo: An effective multi-task learning network for stereo matching and edge detection. *International Journal of Computer Vision*, 2020, **128**(4): 910−930

10 Song X, Zhao X, Hu H W, Fang L J. Edgestereo: A context integrated residual pyramid network for stereo matching. In: Proceedings of the 14th Asian Conference on Computer Vision. Springer, Cham, 2018. **11365**: 20−35

11 Yang G R, Zhao H S, Shi J P, Deng Z D, Jia J Y. Segstereo: Exploiting semantic information for disparity estimation. In: Proceedings of the 15th European Conference on Computer Vision. Springer Verlag: 2018. **11211**: 660−676

12 Zhang J M, Skinner K A, Vasudevan R, Johnson-Roberson M. Dispsegnet: Leveraging semantics for end-to-end learning of disparity estimation from stereo imagery. *IEEE Robotics and Automation Letters*, 2019, **4**(2): 1162−1169

13 Jie Z Q, Wang P F, Ling Y G, Zhao B, Wei Y C, Feng J S, Liu W. Left-right comparative recurrent model for stereo matching. In: Proceedings of the 2018 IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE, 2018.3838−3846

14 Liang Z F, Feng Y L, Guo Y L, Liu H Z, Chen W, Qiao L B, Zhou L, Zhang J F. Learning for disparity estimation through feature constancy. In: Proceedings of the 2018 IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE, 2018. 2811−2820

15 Cheng Ming-Yang, Gai Shao-Yan, Da Fei-Peng. A stereo-matching neural network based on attention mechanism. *Acta Optica Sinica*, 2020, **40**(14): 144−152
(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. 光学学报, 2020, **40**(14): 144−152)

16 Wang Yu-Feng, Wang Hong-Wei, Yu Guang, Yang Ming-Quan, Yuan Yu-Wei, Quan Ji-Cheng. Stereo matching based on 3D convolutional neural network. *Acta Optica Sinica*, 2019, **39**(11): 227−234
(王玉锋, 王宏伟, 于光, 杨明权, 袁昱纬, 全吉成. 基于三维卷积神经网络的立体匹配算法. 光学学报, 2019, **39**(11): 227−234)

17 Chang J R, Chen Y S. Pyramid stereo matching network. In: Proceedings of the 2018 IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE, 2018. 5410−5418

18 Zhu Z D, He M Y, Dai Y C, Rao Z B, Li B. Multi-scale cross-form pyramid network for stereo matching. In: Proceedings of the 14th IEEE Conference on Industrial Electronics and Applications. Xi′an, China: IEEE, 2019. 1789−1794

19 Zhang L, Wang Q H, Lu H H, Zhao Y. End-to-end learning of multi-scale convolutional neural network for stereo matching. In: Proceedings of Asian Conference on Machine Learning. 2018. 81−96

20 Mayer N, Ilg E, Hausser P, Fischer P. A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation. In: Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, NV, USA: IEEE, 2016. 4040−4048

21 Kendall A, Martirosyan H, Dasgupta S, Henry P, Kennedy R, Bachrach A, Bry A. End-to-end learning of geometry and context for deep stereo regression. In: Proceedings of the 2017 IEEE International Conference on Computer Vision. Venice, Italy: IEEE, 2017. 66−75

22 Lu H H, Xu H, Zhang L, Ma Y B, Zhao Y. Cascaded multi-scale and multi-dimension convolutional neural network for stereo matching. In: Proceedings of the 2018 IEEE Visual Communications and Image Processing. Taichung, China: IEEE, 2018. 1−4

23 Rao Z B, He M Y, Dai Y C, Zhu Z D, Li B, He R J. MSDC-Net: Multi-scale dense and contextual networks for automated disparity map for stereo matching. arXiv PreprintarXiv: 1904.12658, 2019.

24 Guo X Y, Yang K, Yang W K, Wang X G, Li H S. Group-wise correlation stereo network. In: Proceedings of the 2019 IEEE Conference on Computer Vision and Pattern Recognition. Long Beach, USA: IEEE, 2019. 3273−3282

25 Liu M Y, Yin H J. Cross attention network for semantic segmentation. In: Proceedings of the 2019 IEEE International Conference on Image Processing. Taipei, China: IEEE, 2019. 2434−2438

26 Wang Ya-Shen, Huang He-Yan, Feng Chong, Zhou Qiang. Conceptual sentence embeddings based on attention mechanism. *Acta Automatica Sinica*, 2020, **46**(7): 1390−1400
(王亚珅, 黄河燕, 冯冲, 周强. 基于注意力机制的概念化句嵌入研究. 自动化学报, 2020, **46**(7): 1390−1400)

27 Kim J H, Choi J H, Cheon M, Lee J S. RAM: Residual attention module for single image super-resolution. arXiv PreprintarXiv: 1811.12043, 2018.

28 Jeon S, Kim S, Sohn K. Convolutional feature pyramid fusion via attention network. In: Proceedings of the 2017 IEEE Inter-

national Conference on Image Processing. Beijing, China: IEEE, 2017. 1007−1011

29 Sang H W, Wang Q H, Zhao Y. Multi-scale context attention network for stereo matching. *IEEE Access*, 2019, **7**: 15152−15161

30 Zhang G H, Zhu D G, Shi W J, Ye X Q, Li J M, Zhang X L. Multi-dimensional residual dense attention network for stereo matching. *IEEE Access*, 2019, **7**: 51681−51690

31 Hu J, Shen L, Albanie S, Sun G, Wu E H. Squeeze-and-excitation networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019, **42**(8): 2011−2023

32 Wang Q L, Wu B G, Zhu P F, Li P H, Zuo W M, Hu Q H. ECA-Net: Efficient channel attention for deep convolutional neural networks. In: Proceedings of the 2020 IEEE Conference on Computer Vision and Pattern Recognition. Seattle, WA, USA: IEEE, 2020. 11531−11539

33 Menze M, Geiger A. Object scene flow for autonomous vehicles. In: Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition. Boston, MA, USA: IEEE, 2015. 3061−3070

34 Geiger A, Lenz P, Urtasun R. Are we ready for autonomous driving? The KITTI vision benchmark suite. In: Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition. Providence, RI, USA: IEEE, 2012. 3354−3361

35 Pang J H, Sun W X, Ren J S, Yang C X, Yan Q. Cascade residual learning: A two-stage convolutional neural network for stereo matching. In: Proceedings of the 2017 IEEE International Conference on Computer Vision Workshops. Venice, Italy: IEEE, 2017. 878−886

36 Žbontar J, Lecun Y. Stereo matching by training a convolutional neural network to compare image patches. *Journal of Machine Learning Research*, 2016, **17**: 1−32

37 Tulyakov S, Ivanov A, Fleuret F. Practical deep stereo (PDS): Toward applications-friendly deep stereo matching. In: Proceedings of the 32nd Conference on Neural Information Processing Systems. Montreal, Canada: 2018.

张亚茹 燕山大学博士研究生. 主要研究方向为人工智能, 计算机视觉和计算机图形学.
E-mail: yrzhang1014@163.com
(**ZHANG Ya-Ru** Ph.D. candidate at the School of Information Science and Engineering, Yanshan University. Her research interest covers artificial intelligence, computer vision, and computer graphics.)

孔雅婷 燕山大学硕士研究生. 主要研究方向为人工智能, 计算机视觉和计算机图形学.
E-mail: kongyt10@163.com
(**KONG Ya-Ting** Master student at the School of Electrical Engineering, Yanshan University. Her research interest covers artificial intelligence, computer vision, and computer graphics.)

刘 彬 燕山大学信息科学与工程学院教授. 主要研究方向为人工智能和计算机视觉. 本文通信作者.
E-mail: liubin@ysu.edu.cn
(**LIU Bin** Professor at the School of Information Science and Engineering, Yanshan University. His research interest covers artificial intelligence and computer vision. Corresponding author of this paper.)