第 58 卷 第 14 期/2021 年 7 月/激光与光电子学进展 研究论文

激光与光电子学进展 先进成像

基于门控循环单元的立体匹配方法研究

杜弘志，张腾，孙岩标，杨凌辉，朱继贵*

天津大学精密测试技术及仪器国家重点实验室，天津 300072

摘要 基于三维卷积神经网络(3DCNN)结构的深度学习立体匹配方法是目前实现高精度视差结果的重要手段，其核心是以高计算成本来换取匹配的精确性。为了实现低计算成本的立体匹配方法，提出一种基于门控循环单元网络的立体匹配方法。所提方法采用门控循环单元结构来代替三维卷积实现代价聚合，基于循环结构特性来减少网络所需的计算资源。同时，为了保证门控循环单元方法在弱纹理与遮挡区域的视差估计精度，通过“编码-解码”结构来增强网络在三维匹配代价空间中的感受野，以残差连接的方式对多尺度下的匹配代价进行有效融合。在实验验证过程中，利用 KITTI2015 和 Scene Flow 数据集进行方法性能验证。实验结果证明，所提匹配方法的精度与三维卷积立体匹配方法接近，但是显存消耗降低 45%，运行时间减少 18%，极大降低了立体视觉匹配的计算成本。

关键词 机器视觉；立体匹配；深度学习；循环神经网络；门控循环单元

中图分类号 TP391.41 **文献标志码** A **doi**：10.3788/LOP202158.1415001

Stereo Matching Method Based on Gated Recurrent Unit Networks

Du Hongzhi, Zhang Teng, Sun Yanbiao, Yang Linghui, Zhu Jigui*

State Key Laboratory of Precision Measuring Technology and Instruments, Tianjin University, Tianjin 300072, *China*

Abstract Deep learning stereo matching method based on three-dimensional convolutional neural networks (3DCNN) is fundamental to obtain accurate disparity results. The main concern with this approach is the high demand of computational resources for achieving high accuracy. To perform stereo matching method at a low computational cost, a method based on a gated recurrent unit network is proposed herein. The proposed method performs cost aggregation by replacing the 3D convolution with a gated-loop unit structure and reduces the computational resource requirements of the network based on the characteristics of the loop structure. To ensure high disparity estimation accuracy in images with weak textures and occluded areas, the proposed method includes an encoder-decoder architecture to further enlarge the receptive field in the 3D matching cost space and effectively aggregate contextual information of multiscale matching costs using residual connections. The proposed method was evaluated on the KITTI2015 and Scene Flow datasets. Experimental results demonstrate that the accuracy of the proposed stereo matching method is close to that of 3D convolutional stereo matching method while reducing the video memory consumption by 45% and the running time by 18%, greatly alleviating the calculation burden of stereo matching.

Key words machine vision; stereo matching; deep learning; recurrent neural network; gated recurrent unit

OCIS codes 150.0150; 150.6910; 330.1400

1 引 言

立体匹配是获取场景中物体距离信息的重要手段，广泛应用于无人驾驶、机器人引导等视觉任务中。传统的立体匹配方法是基于图像灰度信息判断立体图像中的同名点的，虽然容易实现，但对于弱纹理和受遮挡等病态区域，缺乏有效的计算方法。

在图像语义分割、目标检测等复杂视觉问题中，

收稿日期：2020-09-21；**修回日期**：2020-11-02；**录用日期**：2020-11-14

基金项目：国家自然科学基金(41701522，51975408，51721003)

通信作者：* jiguizhu@tju.edu.cn

深度学习通过卷积神经网络(CNN)提取出带有语义信息的高级图像特征，避免了人工特征中非线性表达能力不足的缺点，为解决上述问题提供了有效手段。在匹配代价的计算中，Žbontar 等[1-2]使用卷积神经网络描述图像块间的相似性，提高了立体匹配算法的精度与鲁棒性。Mayer 等[3]提出端到端的视差学习策略，使用“编码-解码”结构的二维卷积网络融合像素之间的相关性对视差进行估计，解决了匹配代价学习中受区域匹配与优化方法限制的问题。CRL[4]与 FADNet[5]同样基于 2DCNN 对视差进行估计，并通过多个尺度下的细化提高了最终的精度。为充分利用双目相机模型中的几何约束，Kendall 等[6]使用 3DCNN 进行代价聚合，在视差空间中聚合特征向量，通过回归预测视差。Chang 等[7]在此基础上提出了 PSMNet，为了增加网络的感受野，在特征提取与代价聚合部分分别使用了空间金字塔池化和沙漏三维卷积网络堆叠操作，提高了网络在遮挡与弱纹理区域的匹配精度。Zhang 等[8]提出的 GA-Net 是基于传统的立体匹配算法的，将引导聚合层代替部分三维卷积操作，进一步提高了代价聚合的效果，是目前精度领先的公开算法。此外，通过在代价聚合的过程中引入注意力机制或对匹配代价进行自适应滤波，同样可提升这类网络的视差估计精度。

在端到端的立体匹配网络中，使用三维卷积进行代价聚合的方法尽管取得了较高的精度，但增加了显存消耗和运行时间，难以在有限的硬件条件下完成大场景下的立体匹配任务。针对网络计算成本高的问题，Wang 等[9]提出了一种从低分辨率到高分辨率对视差图进行逐级估计的策略，提高了运算速度但精度较低。Guo 等[10]对 PSMNet 进行了改进，通过对图像特征进行分组处理，降低了匹配代价在网络中的维度，但要精度达到相同水平时，运算时间要有少量提升。综上所述，在基于深度学习的立体匹配方法中，视差估计精度与计算成本之间矛盾的问题一直没有得到有效的解决，并且缺乏关于显存占用方面的研究。门控循环单元(GRU)网络[11-12]是一种具有长短期记忆(LSTM)的循环神经网络[13]，由门控单元组成循环体，根据当前状态控制信息传递，具有结构简单、收敛速度快的特点，广泛应用于时序数据的处理，如语音识别、文本生成、视频分类等。在视差维度上传递代价信息时，GRU 网络可通过循环结构以低内存成本实现代价聚合的功能[14-15]。

为实现高精度、低成本的视差估计方法，本文提出了一种基于门控循环单元的立体匹配网络。在实现端到端的视差估计过程中，为代替 3DCNN，设计了一种循环代价聚合模块，有效减小了网络对内存的消耗。针对堆叠 GRU 网络中感受野不足的问题，使用“编码-解码”结构增大匹配代价的计算范围，以残差连接的方式聚合多尺度的代价信息，保证了所提方法在弱纹理与遮挡区域的有效性。

2　基于门控神经网络的立体匹配算法

所提网络结构如图1所示，该网络由特征提取

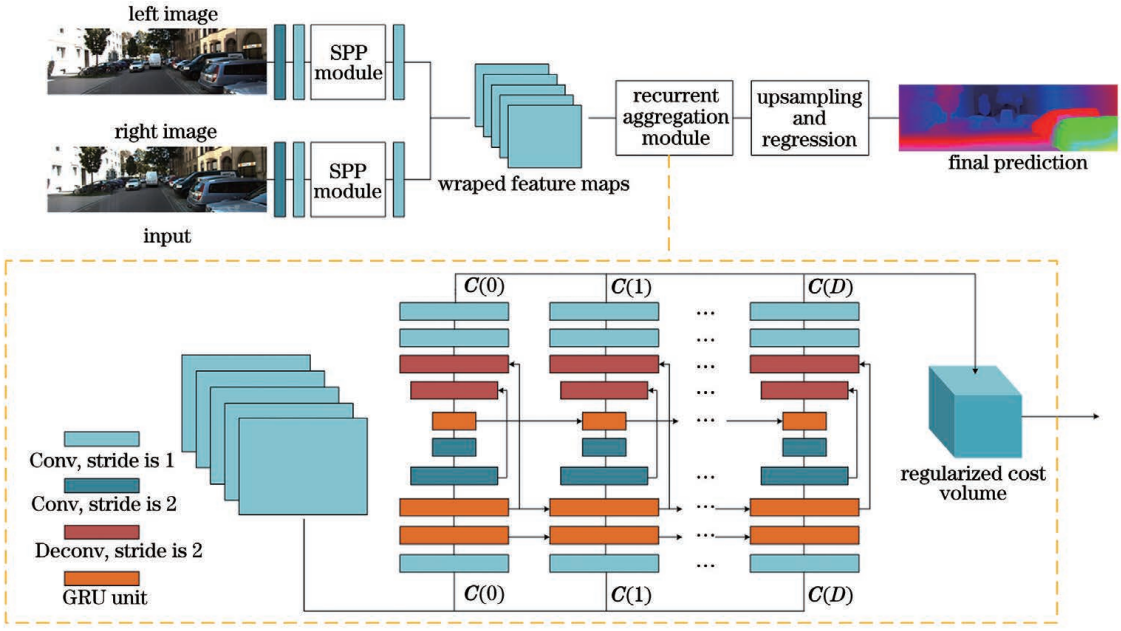

图 1　所提网络结构

Fig. 1　Proposed network structure

网络、循环代价聚合模块和视差回归 3 部分组成。左右视图经特征提取网络后，得到分辨率为原图像分辨率 1/4 的特征图，再使用文献[7]中的空间金字塔池化(SPP)模块融合多尺度下的特征信息，保证提取出的特征具有较大的感受野，满足了弱纹理与遮挡区域匹配的需求。在选定的视差区间$[0,D]$内，对左右视图中提取出的特征向量按行对齐并进行拼接，结果作为参考图像的初始匹配代价序列，记为$\boldsymbol{C}(d)(d=0,1,\cdots,D)$。循环代价聚合模块沿视差方向进行迭代计算，得到最终的匹配代价$\boldsymbol{C}_{\mathrm{r}}(d)(d=1,2,\cdots,D)$，通过上采样将代价图恢复到与原图像一致的尺寸，最终使用 Softmax 函数回归得到视差。

2.1　卷积门控循环单元

在代价聚合中使用卷积 GRU，用于沿视差方向传递匹配代价信息。GRU 在避免循环神经网络中梯度爆炸与消失问题的同时，通过使用更新门，减少了内部状态的数量，解决了 LSTM 中记忆单元存在冗余的问题。GRU 结构如图 2 所示。

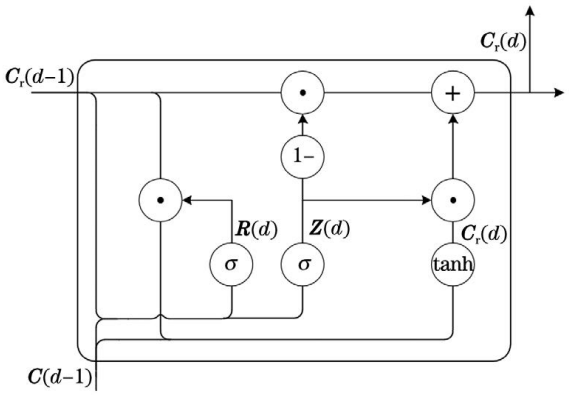

图 2　GRU 结构

Fig. 2　GRU structure

定义“⊙”为矩阵或向量对应元素相乘，“∗”为卷积操作，“[ ]”为向量拼接。在对匹配代价进行聚合时，GRU 的前向传播可表示为

$$\begin{cases}\boldsymbol{R}(d)=\sigma\{\boldsymbol{W}_{\mathrm{r}}*[\boldsymbol{C}_{\mathrm{r}}(d-1),\boldsymbol{C}(d)]+\boldsymbol{b}_{\mathrm{r}}\}\\ \boldsymbol{Z}(d)=\sigma\{\boldsymbol{W}_{\mathrm{z}}*[\boldsymbol{C}_{\mathrm{z}}(d-1),\boldsymbol{C}(d)]+\boldsymbol{b}_{\mathrm{z}}\}\\ \widetilde{\boldsymbol{C}}_{\mathrm{r}}(d)=\tanh\{\boldsymbol{W}_{\tilde{c}}*[\boldsymbol{R}(d)\odot\boldsymbol{C}_{\mathrm{r}}(d-1),\boldsymbol{C}(d)]+\boldsymbol{b}_{\tilde{c}}\}\\ \boldsymbol{C}_{\mathrm{r}}(d)=[\boldsymbol{1}-\boldsymbol{Z}(d)]\odot\boldsymbol{C}_{\mathrm{r}}(d-1)+\boldsymbol{Z}(d)\odot\widetilde{\boldsymbol{C}}_{\mathrm{r}}(d)\end{cases},\tag{1}$$

式中：$\boldsymbol{Z}(d)$为更新门，决定上一状态对当前结果的影响；$\boldsymbol{R}(d)$为重置门，用于控制上一状态的输入；$\boldsymbol{W}_{\mathrm{r}}$、$\boldsymbol{W}_{\mathrm{z}}$、$\boldsymbol{W}_{\tilde{c}}$ 和 $\boldsymbol{b}_{\mathrm{r}}$、$\boldsymbol{b}_{\mathrm{z}}$、$\boldsymbol{b}_{\tilde{c}}$ 分别为卷积核的参数矩阵与偏置向量；$\sigma(\cdot)$为 Sigmoid 函数；$\tanh(\cdot)$为双曲正切函数。

带有卷积的 GRU 可在视差空间上进行匹配代价聚合，相较于 3DCNN 在运算时保留全部的匹配代价信息，该结构仅保留中间状态信息并在循环过程中更新，因此避免了对显存的大量消耗。将 GRU 作为网络的中间层，通过堆叠的方式，可进一步提高代价聚合的性能，结构如图 3 所示。

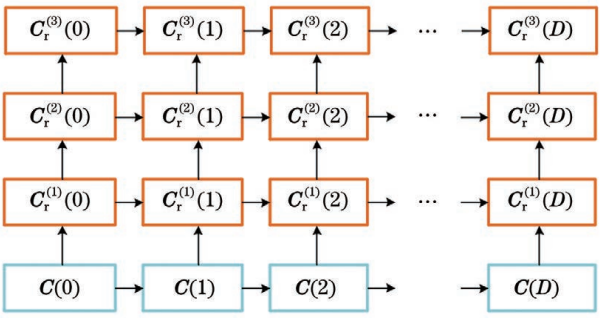

图 3　堆叠 GRU 结构

Fig. 3　Stacked GRU structure

2.2　循环代价聚合模块

代价聚合是一种基于全局约束对初始匹配代价进行更新的过程。通过图像特征计算的匹配代价易受图像噪声、光照变换和遮挡的影响，不能真实地反映像素的相关性，往往需要结合多尺度下的信息对这类区域的视差进行推断。因此，在基于卷积神经网络的代价聚合模块中，应使网络具有较大的感受野以聚合更多的代价信息。感受野是神经网络中输出节点映射到输入的范围，计算公式为

$$l_k=l_{k-1}+(f_k-1)\prod_{t=1}^{k-1}s_t,\tag{2}$$

式中：l_k 为第 k 层的感受野；f_k 为卷积核的大小；s_t 为卷积操作的步长。

由(2)式可知，在网络层数一定的条件下，通过在网络中使用卷积进行降采样操作，能够显著地提高输出的感受野，但同时也会造成网络输出的分辨率降低。为保证序列输出尺寸的一致性，在 GRU 中只能使用步长为 1 的卷积操作，因此网络的最终输出具有较小的感受野。在卷积神经网络中，常采

用“编码-解码”结构增加网络的感受野，进而更好地联系上下文信息，使用卷积，不断降采样“编码”后，通过反卷积进行上采样并使用按通道拼接的方式融合相同尺度下的信息，逐层“解码”至输入尺寸。该结构在使用拼接的方式传递多尺度信息的同时，也会因通道数的增加导致参数量与计算量较大。

基于上述分析，本文设计了一种带有“编码-解码”结构的循环代价聚合模块。相较文献[14]中的堆叠 GRU 结构，所提结构在增加网络感受野的同时，能够有效融合多尺度下的匹配代价信息，提高了算法在病态区域上的计算精度，结构如图 1 虚线框所示。

根据当前指定的视差，对右视图的特征图进行整体平移，再与左视图的特征进行拼接，作为该模块的输入；经过两层的卷积 GRU 后，通过两层步长为 2 的卷积操作实现降采样，在这一尺度下，再使用卷积 GRU 沿视差方向传递信息；在使用反卷积实现上采样的过程中，通过残差连接的方式[16]实现多尺度代价信息的融合，在避免梯度消失的同时减少了网络的计算量；最后，通过卷积操作得到聚合后的匹配代价，沿视差方向对每次循环输出的结果进行排列作为该模块的输出。除输出层外，该模块中每个卷积操作均使用了批处理归一化处理以及 ReLU 函数作为激活函数。具体网络参数如表 1 所示，其中 H 表示输入图像的高度，W 表示输入图像的宽度，K 表示卷积核的大小，C 表示输出通道数，S 表示卷积操作的步长。

表 1　循环代价聚合模块的参数

Table 1　Parameters of recurrent aggregation module

Operation	Layer setting	Output size
input		$1/4H\times1/4W\times64$
GRU_1	$K=3\times3,C=32$	$1/4H\times1/4W\times32$
GRU_2	$K=3\times3,C=32$	$1/4H\times1/4W\times32$
Conv_1	$K=3\times3,C=48,S=2$	$1/8H\times1/8W\times48$
Conv_2	$K=3\times3,C=64,S=2$	$1/16H\times1/16W\times64$
GRU_3	$K=3\times3,C=64$	$1/16H\times1/16W\times64$
Deconv_1	$K=4\times4,C=48,S=2$	$1/8H\times1/8W\times48$
add(Conv_1)	$K=3\times3,C=48,S=1$	$1/8H\times1/8W\times48$
Deconv_2	$K=4\times4,C=32,S=2$	$1/4H\times1/4W\times32$
add(Conv_2)	$K=3\times3,C=32,S=1$	$1/4H\times1/4W\times32$
Conv_3	$K=3\times3,C=8,S=1$	$1/4H\times1/4W\times8$
Conv_4	$K=3\times3,C=1,S=1$	$1/4H\times1/4W\times1$

使用回归的方式进行视差分配。首先通过 Softmax 函数，将在[0，D]范围内的匹配代价大小 $C_r(d)$转换为对应视差的概率：

$$P(d)=\mathrm{Softmax}[-C_r(d)]=\frac{e^{-C_r(d)}}{\sum_{d=0}^{D}e^{-C_r(d)}}。\tag{3}$$

通过计算数学期望，可得到预测的视差结果：

$$\hat{d}=\sum_{d=0}^{D}d\times P(d)。\tag{4}$$

利用视差分配方式，可进行亚像素级的视差估计且计算过程是可导的，因此该方式广泛应用于基于三维卷积的立体匹配网络中。由于它是以数学期望近似计算视差概率最大的取值，在以一个像素为间隔对视差进行均匀采样的前提下，视差 D 越大，匹配代价的差异性越明显，进而可得到越准确的视差结果。由于所提方法中的循环网络结构，增加视差估计范围对运行显存的额外消耗要远小于三维卷积网络，所提方法中的循环网络结构可用于解决大场景下的视差估计任务。

2.3　损失函数

选取平滑的 L1 损失作为基础损失函数，数学描述为

$$L(\hat{d},d)=\frac{1}{N}\sum_{n=1}^{N}l(|\hat{d}-d|),\tag{5}$$

$$l(x)=\begin{cases}x-0.5, & x\geqslant1\\ x^2/2, & x<1\end{cases},\tag{6}$$

式中：$\hat{d}$ 为预测视差值；d 为真实的视差值；N 为带有真实数据的样本数。

为缓解 GRU 网络中梯度消失的问题，使用基于深度监督的训练策略，计算代价聚合过程中第二层 GRU 输出的损失函数。最终的损失函数可表示为

$$L_T=w_1L_1(\hat{d},d)+w_2L_2(\hat{d},d),\tag{7}$$

式中：w_1 和 w_2 分别为中间层损失函数与最终损失函数的权重。

3　实　　验

在 Scene Flow 数据集[3]、KITTI2015 和 KITTI2012[17]三个公开数据集上对方法进行训练与评价。首先，在 Scene Flow 数据集上进行消融实验，分析循环代价聚合模块对最终结果的影响；然后，与几种典型的立体匹配网络进行比较，分析所提方法对运行时间与显存消耗的影响；最后，给出在 KITTI2015 和 KITTI2012 排行榜中与其他方法比

较的结果。

3.1　实验细节

Scene Flow 是一个大型合成数据集，包含约 39000 组 540 pixel×960 pixel 的立体图像数据，其中包含 4370 组测试数据，主要评价指标包括视差的平均绝对误差 E_{ep}、视差估计范围内误差大于 1 个像素的百分比 R_1、视差估计范围内误差大于 3 个像素的百分比 R_3、算法在测试过程中的运行时间 t_{run}、运行显存消耗。KITTI2015 是一个面向自动驾驶的真实场景数据集，包含训练与测试数据各 200 组，每对图像大小为 375 pixel×1242 pixel，由激光雷达提供稀疏的真实视差值，评价指标包括视差估计结果中误差大于 3 个像素(或大于真实值 5%)的百分比 E_{D1}、t_{run}、运行显存消耗。KITTI2012 数据集同样采用激光雷达获取深度信息，分别包含了 194 对训练数据和 195 对测试数据，评价指标包括以不同像素个数为阈值的错误率、整体的平均误差。

在 pytorch 上实现算法，以小批量梯度下降法进行训练，单次迭代样本数为 4，选用延迟参数为(0.9，0.999)的 Adam 优化器[18]，视差范围设置为[0，192]。对于评价的算法，均在 Scene Flow 训练集上以 240 pixel×624 pixel 的随机剪裁尺寸训练 10 个 epoch，学习率设置为 0.01。随后在 KITTI2015 和 KITTI2012 训练集上进行微调，随机剪裁尺寸到 240 pixel×624 pixel，以 0.01 的学习率训练 300 个 epoch。

3.2　模型消融实验

在验证模型有效性的同时对代价聚合模块进行消融分析，使用相同的特征提取网络和训练方式，仅改变循环代价聚合模块中的网络结构。对使用三层结构的堆叠 GRU 网络与所提方法进行比较，其中卷积 GRU 采用尺寸为 3×3 大小的卷积核，通道数均设置为 32，最终经过两层卷积操作输出最终的匹配代价结果。在 Scene Flow 测试集上进行评价，结果如表 2 所示，可以看出，改进后的网络结构能够显著提升视差的估计精度，平均绝对误差由 1.7 pixel 下降至 1.2 pixel，相对下降了约 29%。

表 2　不同代价聚合模块的比较

Table 2　Comparison of different cost aggregation modules

Module	E_{ep}/pixel	R_1/%	R_3/%
Stacked GRU	1.7	16.29	7.79
Proposed method	1.2	12.01	4.99

为直观比较二者的差异，对两种网络结构得到的视差图进行了比较，如图 4 所示。图 4 中标记了堆叠 GRU 结构中匹配精度较低的区域，1、2 列中主要分布在弱纹理表面，第 3 列中因遮挡，匹配错误。相较于堆叠 GRU 结构，所提代价聚合模块具有更大的感受野，在弱纹理等病态区域上具有更高的匹配精度。

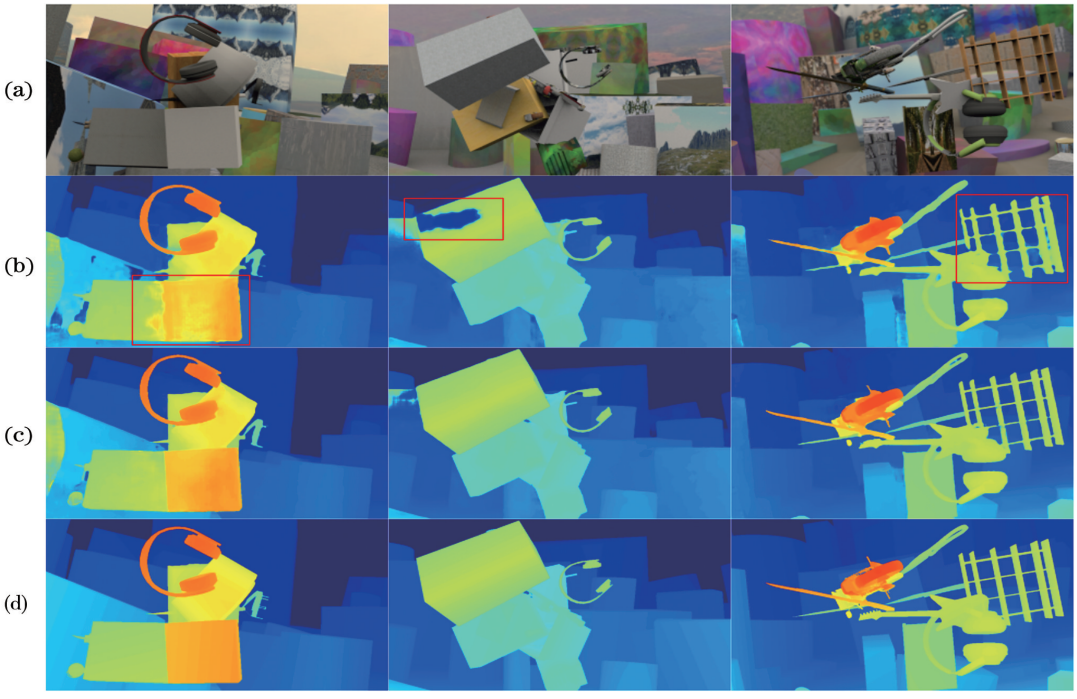

图 4　视差估计结果。(a)左视图；(b)堆叠 GRU；(c)所提方法；(d)真实视差

Fig. 4　Result of the disparity estimation. (a) Left images; (b) stacked GRU; (c) proposed method; (d) ground truth

3.3 算法性能对比

对所提方法与已提出的立体匹配网络进行比较。在 Scene Flow 测试集上对算法性能进行比较，结果如表 3 所示，使用 TITAN RTX 显卡对各方法的显存消耗进行了测试，其中 t_{run} 是在显卡上对 100 组图像进行视差估计的平均运行时间。

表 3　Scene Flow 测试集上不同方法的性能评价

Table 3　Performance evaluation of different methods on Scene Flow test dataset

Method	E_{ep}/pixel	Memory /GB	t_{run}/ms
DispNetC[3]	1.68	1.62	18.7
PSMNet[7]	1.09	4.65	399.3
GANet[8]	0.84	6.65	2251.1
Proposed method	1.22	2.57	326.8

表 3 中，DispNetC 是基于二维卷积的立体匹配网络，因此具有较强的实时性但精度较差；PSMNet 与 GANet 采用三维卷积进行代价处理，通过牺牲显存与运行时间的方式得到了较高的精度；相较于采用相同特征提取网络的 PSMNet，所提方法对显存的需求减少约 45%，精度上仅下降约 12%，运行时间上也有所降低。这些结果可以证明，所提循环代价聚合模块在有效实现 3DCNN 代价聚合功能的同时，能够显著降低显存消耗，可在低成本硬件条件下实现高精度的测量结果。

同其他方法一样，在 KITTI2015 数据集上对模型进行微调后，提交测试结果进行统一评价，评价结果对比如表 4 所示。其中“All”与“Noc”分别表示在评价时考虑全部像素与非遮挡区域像素，$E_{D1\text{-}bg}$、$E_{D1\text{-}fg}$、$E_{D1\text{-}all}$ 分别表示在背景、前景以及所有区域内对 E_{D1} 的计算结果。

表 4　KITTI2015 测试集上的性能评价

Table 4　Evaluation of different method on KITTI2015 test dataset

Method	All			Noc		
	E_{D1-bg}/%	E_{D1-fg}/%	E_{D1-all}/%	E_{D1-bg}/%	E_{D1-fg}/%	E_{D1-all}/%
DispNetC[3]	4.32	4.41	4.34	4.11	3.72	4.05
MADNet[18]	3.75	9.20	4.66	3.45	8.41	4.27
CRL[4]	2.48	3.59	2.67	2.32	3.12	2.45
FADNet[5]	2.68	3.50	2.82	2.49	3.07	2.59
GC-Net[6]	2.21	6.16	2.87	2.02	5.58	2.61
PSMNet[7]	1.86	4.62	2.32	1.71	4.31	2.14
Proposed method	2.20	4.85	2.64	1.82	4.09	2.20

从表 4 可以看出：与二维卷积方法相比，如 DispNetC、MADNet、CRL、FADNet，所提方法具有精度上的优势；在基于三维卷积的方法中，所提方法的精度在 GC-Net 与 PSMNet 之间。为更好比较所提方法与 PSMNet 的差异，在测试结果中选择出误差分布不一致的结果，如图 5 所示。其中对两组图

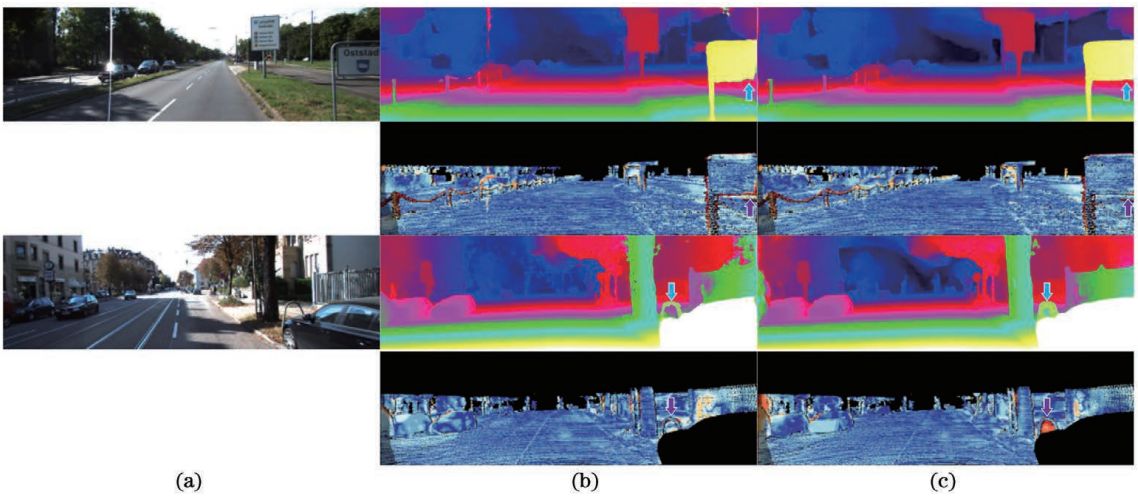

图 5　KITTI2015 测试集上的视差估计。(a)左视图；(b)PSMNet；(c)所提方法

Fig. 5　Result of the disparity estimation on KITTI2015 test dataset. (a) Left image; (b) PSMNet; (c) proposed method

像的视差图与误差图进行了比较，并指出误差分布不一致的区域。可以看出，对于物体表面，所提方法具有较好的重建效果（第一组图片），但也会对某些孔洞区域进行错误的表面重建（第二组图片）。

在 KITTI2012 测试集上的评价结果如表 5 所示，给出了绝对误差大于 2 个像素到大于 5 个像素的百分比、总体的平均误差（Avg. error），并在不考虑遮挡的情况下同样进行了评价。在 KITTI2012 数据集上，所提方法在精度上同样优于基于二维卷积的立体匹配方法（DispNetC 与 FADNet），各项指标的结果与 PSMNet 十分接近，说明所提网络结构可同样实现高精度的视差估计；相比于 GANet，所提方法虽然在精度上表现不足，但可显著减少运行过程中的显存消耗（表 3），更易于在有限的硬件条件下实现。

表 5 KITTI2012 测试集上的性能评价

Table 5 Evaluation of different methods on KITTI2012 test dataset unit: %

Method	>2 pixel		>3 pixel		>4 pixel		>5 pixel		Avg. error	
	Noc	All	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	0.9	1.0
FADNet	3.98	4.63	2.42	2.86	1.73	2.06	1.34	1.62	0.6	0.7
GC-Net	2.71	3.46	1.77	2.30	1.36	1.77	1.12	1.46	0.6	0.7
PSMNet	2.44	3.01	1.49	1.89	1.12	1.42	0.90	1.15	0.5	0.6
GANet	1.89	2.50	1.19	1.60	0.91	1.23	0.76	1.02	0.4	0.5
Proposed method	2.39	3.03	1.48	1.91	1.10	1.43	0.87	1.14	0.5	0.5

4 结 论

设计了一种基于门控循环单元的立体匹配网络，使用 GRU 网络代替 3DCNN 优化匹配代价，避免了显存大量消耗的问题。在循环代价聚合模块中，对堆叠 GRU 结构进行改进，引入带有残差连接的“编码-解码”结构，提高了代价聚合过程中的感受野大小，提升了在弱纹理与遮挡区域的匹配精度。通过在 Scene Flow 和 KITTI2015 数据集上进行评价，所提方法仅消耗了 2.6 G 的显存，与 PSMNet 相比减少了 45%，而且运行时间上也减少了 18%，能够在低成本的硬件设备上实现高精度的视差估计。在代价聚合过程中，GRU 虽然能够实现 3DCNN 的功能，但由于使用迭代的方式进行信息传递，导致误差容易积累，因此在网络层数较少的情况下精度较低。如何得到轻量级、高精度的立体匹配网络将是下一步研究的重点。

参 考 文 献

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

[2] Žbontar J, LeCun Y. Stereo matching by training a convolutional neural network to compare image patches[J]. The Journal of Machine Learning Research, 2016, 17(1): 2287-2318.

[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 (CVPR), June 27-30, 2016, Las Vegas, NV, USA. New York: IEEE Press, 2016: 4040-4048.

[4] 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.

[5] Wang Q, Shi S H, Zheng S Z, et al. FADNet: a fast and accurate network for disparity estimation[C]//2020 IEEE International Conference on Robotics and Automation (ICRA), May 31-August 31, 2020, Paris, France. New York: IEEE Press, 2020: 101-107.

[6] Kendall A, Martirosyan H, Dasgupta S, et al. End-to-end learning of geometry and context for deep stereo regression[C]//2017 IEEE International Conference on Computer Vision (ICCV), October 22-29, 2017, Venice, Italy. New York: IEEE Press,

2017: 66-75.

[7] 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.

[8] Zhang F H, Prisacariu V, Yang R G, et al. GA-net: guided aggregation net for end-to-end stereo matching [C]//2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 15-20, 2019, Long Beach, CA, USA. New York: IEEE Press, 2019: 185-194.

[9] Wang Y, Lai Z H, Huang G, et al. Anytime stereo image depth estimation on mobile devices[C]//2019 International Conference on Robotics and Automation (ICRA), May 20-24, 2019, Montreal, QC, Canada. New York: IEEE Press, 2019: 5893-5900.

[10] 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.

[11] Chung J, Gulcehre C, Cho K, et al. Empirical evaluation of gated recurrent neural networks on sequence modeling[EB/OL]. (2014-12-11)[2020-08-06]. https://arxiv.org/abs/1412.3555v1.

[12] Elman J L. Finding structure in time[J]. Cognitive Science, 1990, 14(2): 179-211.

[13] Cho K, van Merrienboer B, Gulcehre C, et al. Learning phrase representations using RNN encoder-decoder for statistical machine translation [C] // Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), September 3, 2014, Doha, Qatar. Stroudsburg: Association for Computational Linguistics, 2014: 1724-1734.

[14] Yao Y, Luo Z X, Li S W, et al. Recurrent MVSNet for high-resolution multi-view stereo depth inference [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: 5520-5529.

[15] Yao Y, Luo Z X, Li S W, et al. MVSNet: depth inference for unstructured multi-view stereo [M] // Ferrari V, Hebert M, Sminchisescu C, et al. Computer vision-ECCV 2018. Lecture notes in computer science. Cham: Springer, 2018, 11211: 785-801.

[16] He K M, Zhang X Y, Ren S Q, et al. Deep residual learning for image recognition[C]//2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 27-30, 2016, Las Vegas, NV, USA. New York: IEEE Press, 2016: 770-778.

[17] 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.

[18] Kingma D P, Ba J. Adam: a method for stochastic optimization[EB/OL]. (2014-12-22)[2020-08-06]. https://arxiv.org/abs/1412.6980.