基于注意力机制的立体匹配网络研究

程鸣洋[1,2]，盖绍彦[1,2]，达飞鹏[1,2,3*]

[1]东南大学自动化学院，江苏 南京 210096；

[2]东南大学复杂工程系统测量与控制教育部重点实验室，江苏 南京 210096；

[3]东南大学深圳研究院，广东 深圳 518063

摘要 为了提高基于双目视觉中立体匹配在弱纹理场景下的精准性，提出了一种基于注意力机制特征提取的三维重建算法。利用卷积神经网络(CNN)训练左右图像的特征表示，计算出立体匹配的匹配代价。在 CNN 特征提取阶段，加入图像注意力机制模块和通道注意力机制模块，得到特征图各个像素点之间的联系，使网络可以更好地捕获图像上下文信息，进而在重建过程中能够更加精确地重建出弱纹理区域。对于网络损失函数，集成了语义编码损失，最终将损失函数定义为语义编码损失和重建损失的加权和，有效提升了弱纹理区域下的重建精度。使用 KITTI 和 Sceneflow 数据集对算法进行验证，实验结果证明，相比目前国内外先进方法，本文算法在精度方面有较大提升，尤其体现在弱纹理区域。

关键词 机器视觉；立体匹配；双目视觉；卷积神经网络；注意力机制

中图分类号 TP391.41 **文献标志码** A **doi**：10.3788/AOS202040.1415001

A Stereo-Matching Neural Network Based on Attention Mechanism

Cheng Mingyang[1,2], Gai Shaoyan[1,2], Da Feipeng[1,2,3*]

[1] *School of Automation, Southeast University, Nanjing, Jiangsu* 210096, *China*;

[2] *Key Laboratory of Measurement and Control of Complex Systems of Engineering, Ministry of Education, Southeast University, Nanjing, Jiangsu* 210096, *China*;

[3] *Shenzhen Research Institute, Southeast University, Shenzhen, Guangdong* 518063, *China*

Abstract To improve the accuracy of stereo matching based on binocular vision applied to weak texture scenes, this study proposes a 3D reconstruction algorithm based on feature extraction using an attention mechanism. The proposed model uses convolutional neural network (CNN) to train feature representation of left and right images and calculates the matching cost of stereo matching. First, during the CNN feature extraction stage, attention mechanism module and channel attention mechanism module are summed to obtain the connection of each pixel in the feature image, enabling the network to capture the context information better and reconstruct weak texture areas more accurately in the reconstruction process. Second, we integrate the semantic coding loss in our neural network. The final loss function is defined as the weighted sum of the semantic coding loss and the reconstruction loss, which can effectively improve the reconstruction accuracy of a region with weak texture. We use KITTI and Sceneflow datasets to validate the algorithm. Experimental results show that the proposed method yields good improvements in accuracy, particularly in areas with weak textures.

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

OCIS codes 150.0155; 330.1400; 200.4260

1 引 言

从计算机视觉发展以来，三维重建技术一直是研究领域备受关注的热点话题。深度信息计算的目的是获取双目图像中像素点的视差值，而视差值取决于空间景物在左右视图中的对应关系，这种寻找左右图像平面之间对应点的过程称为立体匹配[1]。三维重建技术在医疗影像、人机交互、远程教学、自动驾驶等方向都有着潜在的应用前景。未来，随着三维重建技术向着更加精细化、实时化、高效化方向

收稿日期：2020-03-04；**修回日期**：2020-03-22；**录用日期**：2020-04-13

基金项目：国家自然科学基金(51475092，61462072)、江苏省自然科学基金(BK20181269)

* **E-mail**：1095528214@qq.com

发展，将会有越来越多的领域涉及三维重建技术。

近年来随着深度学习的流行，卷积神经网络(CNN)也逐渐应用于立体匹配领域中。早期的深度学习方法主要是处理两张图对应点的匹配问题，并通过网络计算出两点的相似度。目前，深度学习在匹配代价聚合、视差计算、视差调整方面皆有应用，可通过形成一套端到端的网络系统来进行深度图估计。然而，同样地，CNN 在处理阴影、低纹理区域、重复结构上仍有很大的提升空间。所以当前基于 CNN 立体匹配的一个主要问题是如何有效地利用上下文并结合语义信息来很大程度地改善代价聚合和深度回归。

最早的由 Žbontar 和 LeCun 提出的 MC-CNN[2]，是 CNN 应用于立体匹配的开端，MC-CNN 是基于 Siamese network，基于 patch 的提取与比较，学习其相似性，得到一个匹配代价，并将正确匹配的 patch 定义为正样本，未正确匹配的 patch 定义为负样本。随后提出了很多网络后处理方法，如亚像素增强、多次迭代、插值等。Cao 等[3]提出的 GC-Net 使用多尺度特征融合的编解码器体系结构正则化匹配代价，对 Displet 进行了改进。Seki 等[4]提出的 SGM-Nets 基于传统半全局匹配(SGM)设计了一个学习的方式，对其中重要的惩罚因子进行估计，估计出的物体结构形状可明确区分。Chang 等[5]提出的 PSM-Net 使用一个新的金字塔池化模块(SPP)来利用全局上下文信息，最终使用 encoder-decoder 获取特征，并且在汇聚匹配代价时提出堆叠残差网络，使重建精度有了显著的提升。Khamis 等[6]提出的 StereoNet 使用 Siamese 网络从左右图像中提取特征用于实施立体匹配的端到端深度架构，在 NVidia Titan X 上以 60 frame/s 运行，可生成高质量、边缘保留、无量化的视差图。Zhang 等[7]使用 GA-Net 提出了高效的引导匹配损失聚集(GA)策略，包括半全局聚集(SGA)以及局部引导聚集(LGA)层的端到端匹配。GA 层主要意义在于提高如遮挡、无纹理/反光区域和小结构等具有挑战的区域的估计精度。GA 层能够用来替换三维(3D)卷积，不仅可以减小其计算消耗，还能得到较高的精度。

低纹理场景如墙壁、桌面、道路等，其特征信息不明显，像素点之间差异很小，在立体匹配问题中很难找到他们的对应点。目前传统的针对弱纹理场景的三维重建方法有以下几种。文献[8]中为了在相邻区域获得更好的匹配关系，充分利用重叠块之间的冗余信息，获得稠密匹配的结果，该方法加快了稠密匹配的处理速度，但问题在于得不到平滑的结果，反而产生了高度不连续的匹配结果。Xiao 等[9]采用多尺度扩张与收缩迭代策略来约束邻近匹配，通过由粗到精的方式逐步修正了匹配不连续问题。Jiang 等[10]计算图像的局部特征，并增加了导向滤波模块，通过局部逼近马尔可夫随机场获得了平滑对应场。Fu 等[11]采用 Loopy 置信传播方法进行推理，提出了自上而下的策略，完成了密集对应关系的分层匹配。Si 等[12]采用双目视觉稠密视差鲁棒估计方法进行弱纹理物体重建，重建后的物体形状发生了较大畸变。目前利用卷积神经网络提取的弱纹理场景的特征大多仍然停留在特征金字塔上[5]。

综上所述，立体匹配的主要问题是在特征不明显的区域的重建效果不够理想，因此本文在特征提取过程中加入具有上下文含义的局部、边缘等细微特征，在构成代价卷后，基于深度估计可以在边缘、重复纹理区域得到更精确的结果。本文将语义特征嵌入特征图中，并使用规则化的语义信息作为损失项，改善了视差学习效果，获得了比较好的结果。

2　基于注意力机制的立体匹配网络

2.1　金字塔结构存在的问题

最近比较热门的特征提取网络多基于特征金字塔模块，通过不同尺度的卷积来扩大感受野，如：PSPNet 采用 PSP(Pyramid Scene Parsing)模块将特征图池化为不同尺寸，再作连接上采样[13]；DeepLab 采用 ASPP (Atrous Spatial Pyramid Pooling)模块，并行使用大扩张率卷积来扩大感受野[14]。

这几种方法都是通过大小不同的卷积层来获得图片特征，但是这对上下文的特征表示都不够明确，因为捕获上下文信息不完全等价于增加感受野大小，图像上下文信息还包括像素点之间反映特征的联系紧密程度。如果能够先捕获到图像上下文信息(例如这是车子)，不仅可以提供许多相关目标的信息(例如车子有轮胎、方向盘配件，车子的大概形状)，还可以动态地减小搜索区域[15]。加入一个场景的先验知识，可使图片中像素分类更有目的性。依照这个思路，本研究在三维重建任务中扩展了用于特征提取的注意力机制，基于两种注意力机制模块来捕捉丰富的上下文关系，得到特征图各个像素点的联系，此方式有助于更好地进行类内紧凑特征表示。综合实验结果验证了本文方法的有效性。

2.2　基于注意力机制的立体匹配网络

给定用于三维重建的对应的左右图像，确保该图像在大小、照明和视野上都是多样的。经典方法中自卷积操作会导致具有对应关系的像素由于局部的感受野不同，所产生的特征也有较大的差异[16]。这些差异导致了立体匹配的不一致性，从而影响重建精度。为了解决这个问题，本研究探索全局上下文信息，通过在特征之间建立联系机制，自适应地聚集一定范围的上下文信息，从而改进立体匹配的特征表示。

算法框图如图 1 所示。第一步将左右图像输入到两个权重共享的以 RESNET 为基础的卷积神经网络中计算特征图，接着对于特征图使用改进的注意力机制模块进行特征获取；第二步使用一个权值共享的特征融合的卷积层串联不同通道的特征，得到由左右图像特征构成的匹配代价卷；最后通过一个 3D 卷积神经网络来完成代价聚合和视差计算。

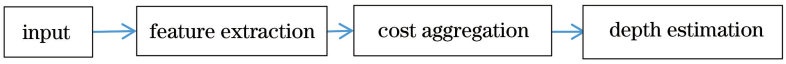

图 1　总算法框图

Fig. 1　Algorithm block diagram

具体的网络架构如图 2 所示，加入两种注意力机制模块，对残差网络生成的特征进行加工，从而获得更好的全局和局部特征来进行重建工作。采用预训练好的扩张残差网络作为骨干网络。在最后两个残差块删掉下采样操作，并加入扩张卷积[17]，从而将最终特征地图的大小放大到输入图像的 1/16。最终特征地图在不添加额外参数的情况下保留了更多图片的特征信息。然后将扩展后的残差网络的特征输入到两个平行的注意力机制模块。

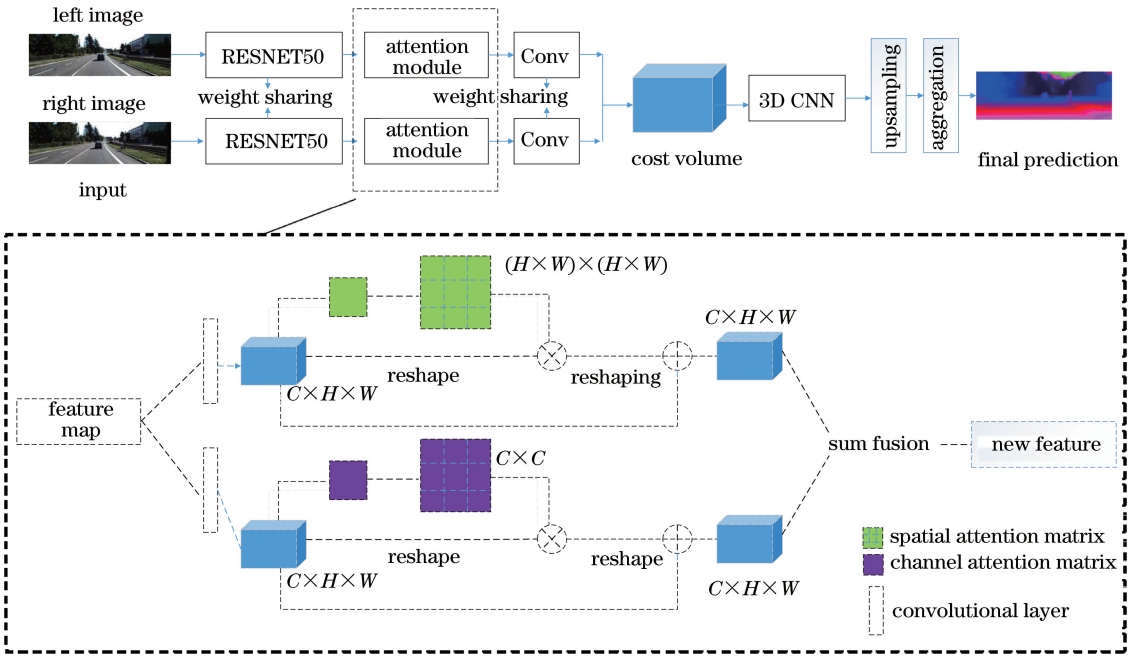

图 2　改进的网络结构

Fig. 2　Structure of our proposed network

本研究通过加入注意力机制来捕获上下文信息，即加入注意力机制模块来整合全局特征和局部特征。不采用级联操作是因为它需要更多的图形处理器(GPU)内存。而注意力机制模块很简单，可以直接插入到现有的网络框架中，不会增加太多参数，但可以有效地增强特征表示。

2.3　注意力机制模块

2.3.1　空间注意力机制

对于三维重建工作的特征提取，图像中不同区域的特征对任务的贡献程度相差很大。空间注意力模型就是寻找网络中特征间的联系，并把最重要的特征放在优先级较高的位置。

而神经网络中池化的方法过于单一，直接将信息合并会导致关键信息无法提取。为解决这个问题，空间转换器[18]模块将图片中的空间域信息作对应的空间变换，从而将关键的信息提取出来。空间转换器其实就是注意力机制的实现，因为训练出的空间转换器能够找出图片信息中需要被关注的区

域，同时这个转换器又具有旋转、缩放变换的功能，这样图片局部的重要信息能够通过变换而被提取出来。

空间注意力机制模型见图 3。

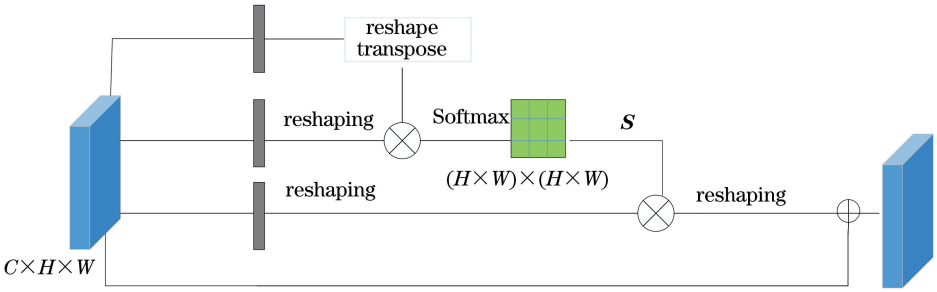

图 3　空间注意力机制结构

Fig. 3　Structure of spatial attention mechanism

具体来说，对于一张大小为 $H_1\times W_1$ 的图片，残差网络提取到特征 $\boldsymbol{F}\in\mathbf{R}^{C\times H\times W}$，其中 $H=H_1/4$，$W=W_1/4$，C 为总通道数，将这个特征送入两个同尺度的卷积层得到两个新的特征图 $\boldsymbol{B}\in\mathbf{R}^{C\times N}$，$\boldsymbol{D}\in\mathbf{R}^{N\times C}$，其中 $N=H\times W$ 为特征图像素数量，接着将 $\boldsymbol{B}$，$\boldsymbol{D}$ 进行矩阵乘法运算，然后作一个 Softmax 操作得到空间特征图 $\boldsymbol{S}\in\mathbf{R}^{N\times N}$，即

$$\boldsymbol{S}_{ji}=\frac{\exp(\boldsymbol{B}_i\times\boldsymbol{C}_j)}{\sum_{i=1}^{N}\exp(\boldsymbol{B}_i\times\boldsymbol{C}_j)},\tag{1}$$

式中：$\boldsymbol{S}_{ji}$ 表示第 i 个像素点对第 j 个像素点的影响，这两个点表达的特征越相像，他们的联系就越紧密。与此同时，将一开始的特征向量 $\boldsymbol{F}$ 作一个卷积得到 $\boldsymbol{F}_1\in\mathbf{R}^{C\times N}$，将 $\boldsymbol{F}_1\in\mathbf{R}^{C\times N}$ 与 $\boldsymbol{S}\in\mathbf{R}^{N\times N}$ 作矩阵乘法，并将得到的结果维度改成 $\boldsymbol{F}_2\in\mathbf{R}^{C\times H\times W}$，其中 $\boldsymbol{F}_2$ 可以表征像素点之间的联系。由(1)式可知，将通过空间注意力机制的特征与 $\boldsymbol{F}_2$ 紧密联系，可以有效表达像素点间的特征表达相似度，而这种特征既包含局部信息，又包含整体信息，可以确保空间注意力机制特征能够更好地获得上下文语义信息。

为了进一步提升特征表示，定义 $\alpha\in[0,1]$ 为自注意力参数，与一开始的特征 $\boldsymbol{F}$ 作一个加权求和得到特征 $\boldsymbol{E}\in\mathbf{R}^{C\times H\times W}$，最终将 $\boldsymbol{E}$ 作为最后的空间注意机制特征图，即

$$\boldsymbol{E}=\alpha\boldsymbol{F}+(1-\alpha)\boldsymbol{F}_2,\tag{2}$$

其中 α 的初始值设为 1，表示训练前只使用特征 $\boldsymbol{F}$。通过不断地增加注意力机制处理特征的比重，使损失函数逐步减小，即重建效果越来越接近预期结果。

2.3.2　通道注意力机制

在残差网络 Resnet 得到的二维特征图中，一个维度表征图像的尺度空间(长宽)，另一个维度表征通道。基于通道的注意力模型，对各个特征通道的重要程度进行建模，然后针对不同的任务增强或者抑制不同的通道，同时，通过通道注意力机制有效地获得不同通道间特征的联系，从而更好地优化特征，最终获得更精确的重建结果。

通道注意力机制模型见图 4。

与空间注意机制不同，直接从残差网络提取的特征 $\boldsymbol{F}\in\mathbf{R}^{C\times H\times W}$ 计算出通道特征图 $\boldsymbol{A}\in\mathbf{R}^{C\times C}$，首先对 $\boldsymbol{F}\in\mathbf{R}^{C\times H\times W}$ 进行降维得到 $\boldsymbol{D}\in\mathbf{R}^{C\times N}$，$N=H\times W$，接着将 $\boldsymbol{D}$ 与 $\boldsymbol{D}$ 的转置作点乘，最后加入一个 Softmax 层得到最终的通道特征图 $\boldsymbol{A}\in\mathbf{R}^{C\times C}$，即

$$\boldsymbol{A}_{ji}=\frac{\exp(\boldsymbol{A}_i\times\boldsymbol{A}_j)}{\sum_{i=1}^{C}\exp(\boldsymbol{A}_i\times\boldsymbol{A}_j)},\tag{3}$$

式中：$\boldsymbol{A}_{ji}$ 表示第 i 个像通道对第 j 个通道的影响，这两个通道表达的特征越相像，它们的联系就越紧密。接着将 $\boldsymbol{A}\in\mathbf{R}^{C\times C}$ 与一开始的特征向量 $\boldsymbol{F}\in\mathbf{R}^{C\times H\times W}$ 作矩阵乘法，得到向量 $\boldsymbol{Z}\in\mathbf{R}^{C\times H\times W}$，其中 $\boldsymbol{Z}$ 表示各个通道之间的联系。最后定义 $\beta\in[0,1]$ 为自注意力参数，与一开始的特征 $\boldsymbol{F}$ 作一个加权求和得到特征 $\boldsymbol{G}\in\mathbf{R}^{C\times H\times W}$，并将 $\boldsymbol{G}$ 作为最后的通道注意机制特征图，即

$$\boldsymbol{G}=\beta\boldsymbol{F}+(1-\beta)\boldsymbol{Z},\tag{4}$$

式中：β 的初始值设为 1，表示训练前只使用特征 $\boldsymbol{F}$。通过逐步增加注意力机制处理特征的比重，可以有效表达通道特征表达的相似度，而这种特征加上原有特征，既包含局部信息，又包含整体信息，对于深度回归的网络函数，主要由两部分组成。这保证了通道注意力机制的特征能够更有效地获得上下文语义信息。

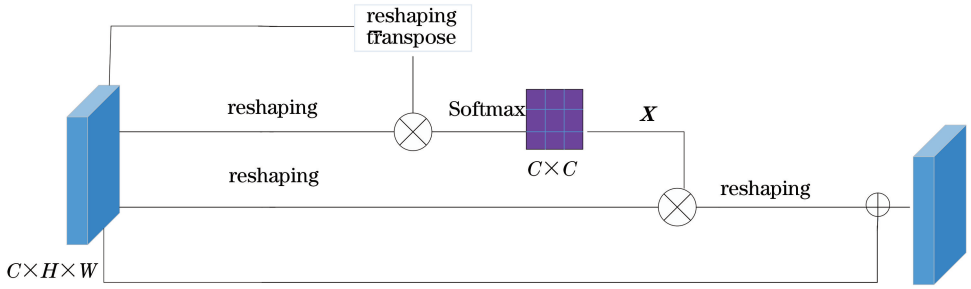

图 4　通道注意力机制结构

Fig. 4　Structure of channel attention mechanism

2.4　损失函数

改进网络的损失函数由两部分组成。

1)语义编码损失(SE-loss)[19]。在现有的标准的训练过程中使用的损失函数是像素距离损失函数,不强调场景的全局信息。而引入语义编码损失,虽然增加了少量计算量,但可进一步规范网络训练,让网络预测能够预测到场景中对象类别的存在,强化网络学习上下文语义。与逐像素的损失不同,SE-loss 对于大小不同的物体有相同的贡献,在实践中能够改善重建弱纹理区域的精度。这里提出的上下文编码模块、语义编码损失和现存的全卷积网络(FCN)方法是兼容的。

将加权求和的图像特征 $\boldsymbol{G}\in\mathbf{R}^{C\times H\times W}$ 转化为一组维度为 C 的输入特征 $\boldsymbol{X}=\{\boldsymbol{x}_1,\cdots,\boldsymbol{x}_N\}$,其中 N 是特征的总个数,即 $H\times W$,与同样方法检测出的局部特征 $\boldsymbol{D}=\{\boldsymbol{d}_1,\cdots,\boldsymbol{d}_C\}$ 作差,将结果相加进行归一化,获得了一个像素位置相对于局部目标的信息 $\boldsymbol{e}_{ik}$,然后将这 N 个结果求和获得最终的 $\boldsymbol{e}_k$,即为整张图像相对于第 k 个局部特征的信息。

$$\boldsymbol{e}_{ik}=\frac{\exp(-\|\boldsymbol{r}_{ik}\|^2)}{\sum_{j=1}^{C}(-\|\boldsymbol{r}_{ij}\|^2)}\boldsymbol{r}_{ik},\tag{5}$$

$$\boldsymbol{r}_{ik}=\boldsymbol{x}_i-\boldsymbol{d}_k,\tag{6}$$

式中:$\boldsymbol{x}_i$ 表示输入特征 $\boldsymbol{X}$;$\boldsymbol{d}_k$ 表示局部特征。

$$\boldsymbol{e}_k=\sum_{i=1}^{N}\boldsymbol{e}_{ik},\tag{7}$$

$$L_{\text{context}}=\sum_{k=1}^{C}\Psi(\boldsymbol{e}_k),\tag{8}$$

式中:$\Psi(\cdot)$表示带 ReLU 的批量归一化(BN)层将 k 个 $\boldsymbol{e}_k$ 融合到一起,避免 k 个独立编码器被排序,并降低了特征表示的维度。此时获得整张图像相对于局部特征图像的全部损失,即语义编码损失 L_{context}。

2)与 GCNET[3] 中损失类似,使用视差回归的方式来估算连续的视差图。由于 L1 损失相对 L2 损失更适用于边界框回归,因此本研究选取平滑 L1 损失函数,并在弱纹理区域赋予更大的权重,来引导网络加强对目标区域的特征提取,进而生成更加精准的目标区域视差。视差估计的目标损失函数计算公式为

$$L_{\text{disparity}}=\beta_1\frac{1}{N_{\text{F}}}\sum_{i=1}^{N_{\text{F}}}\text{smooth}_{\text{L1}}(\boldsymbol{d}_i-\boldsymbol{d}_{i\text{predict}})+\beta_2\frac{1}{N_{\text{B}}}\sum_{i=1}^{N_{\text{B}}}\text{smooth}_{\text{L1}}(\boldsymbol{d}_i-\boldsymbol{d}_{i\text{predict}}),\tag{9}$$

$$\text{smooth}_{\text{L1}}(\boldsymbol{x})=\begin{cases}0.5\boldsymbol{x}^2,\text{if}\,|\boldsymbol{x}|<1\\|\boldsymbol{x}|-0.5,\text{otherwise}\end{cases},\tag{10}$$

式中:$\boldsymbol{d}_i$ 表示实际深度值,$\boldsymbol{d}_{i\text{predict}}$ 表示预测深度值;N_{F} 代表目标区域中拥有视差真值点的个数;N_{B} 代表背景区域中拥有视差真值点的个数;β_1 与 β_2 代表权重。

最后损失函数为这两部分损失函数的加权和,即

$$L=\alpha_1L_{\text{context}}+\alpha_2L_{\text{disparity}},\tag{11}$$

式中:α_1 和 α_2 分别为语义编码损失 L_{context} 和深度损失 $L_{\text{disparity}}$ 的权重,通过实验综合评估,本研究最后取 $\alpha_1=0.2$,$\alpha_2=0.8$。

3　实　　验

3.1　训　　练

在 SceneFlow 和 KITTI 两个数据集上评估改进的方法。SceneFlow 数据集是一个包含 35454 张训练图像和 4370 个测试图像对的综合数据集,在训练和测试阶段都提供了稠密的视差图。另外,这个数据集足够大,可以用来测试实验各个方面的性能。KITTI 数据集包括两个子集,即 KITTI2012 和 KITTI2015。KITTI 2012 数据集包括 194 个训练图像对和 195 个测试图像对,KITTI 2015 数据集由 200 个训练图像对和 200 个测试图像对组成。这些图像记录了不同天气条件下的真实场景。两个数据集都为训练图像提供了稀疏的 LIDAR ground

truth 视差图。对于 SceneFlow 数据集，采用终点误差(EPE)作为评估度量，即以像素为单位的平均视差误差。对于 KITTI，针对背景，采用该前景和所有像素评估视差异常值 D_1 的百分比作为评估度量。视差误差大于 3 pixel 的像素值被定义为异常值。最终将本文方法与 KITTI 数据集上的最新方法作比较。

实验在平台 Pytorch1.0，2 块 GPU1080TI 上实现，批量大小固定为 2。所有模型均采用 Adam 法进行优化，其中参数 $\beta_1=0.9$，$\beta_2=0.999$ 作为优化率，用于训练，最大深度取 192。采取变学习率的训练方法，初始设置学习率至 10^{-4}，在第 20，35，50 轮时学习率减少一半，训练在第 65 次停止迭代。为了进一步优化模型，在另一轮中重复迭代，学习率降低至 10^{-5}，训练在第 120 次停止迭代。将 KITTI 数据集 200 对数据随机以 4∶1的比例进行分割，即 160 对训练集，40 对用于验证集。对图像进行预处理，即减去平均值，除以像素强度标准差，并将其归一化为零均值，将单位标准差的图像作为网络的输入。

在 Scene Flow，KITTI 2012 和 KITTI 2015 数据集上进行测试，结果如图 5～7 所示。实验结果表明，本文方法在多种道路场景和模拟综合场景下均能得到光滑稠密的视差图，特别是在弱纹理以及目标物边缘区域，能够较为明显地保留原目标的信息，匹配效果较好。其中误差图是指视差图与基准视差各个像素点的差距以图像像素的形式展现出来的图像。

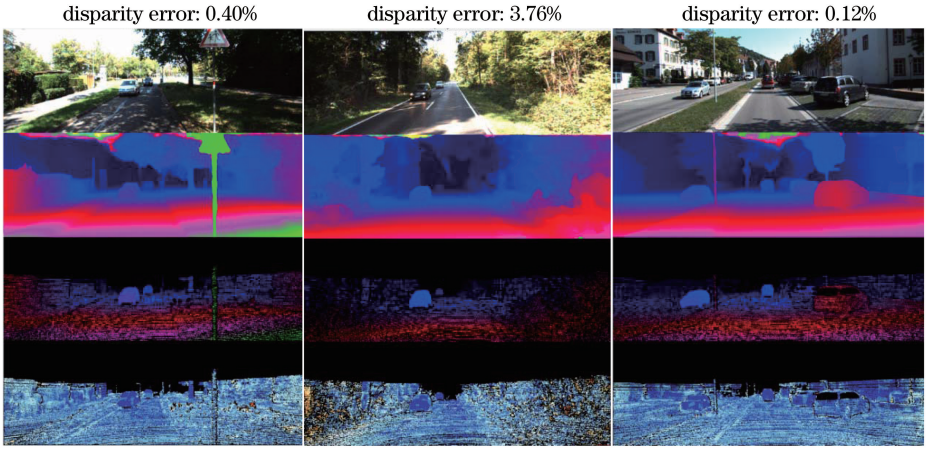

图 5　在 KITTI2015 评价结果，从上到下依次为左输入、预测视差图、实际视差图、误差图

Fig. 5　Results on KITTI2015, from the top to the bottom: the left input, predicted disparity map, actual disparity map, error map

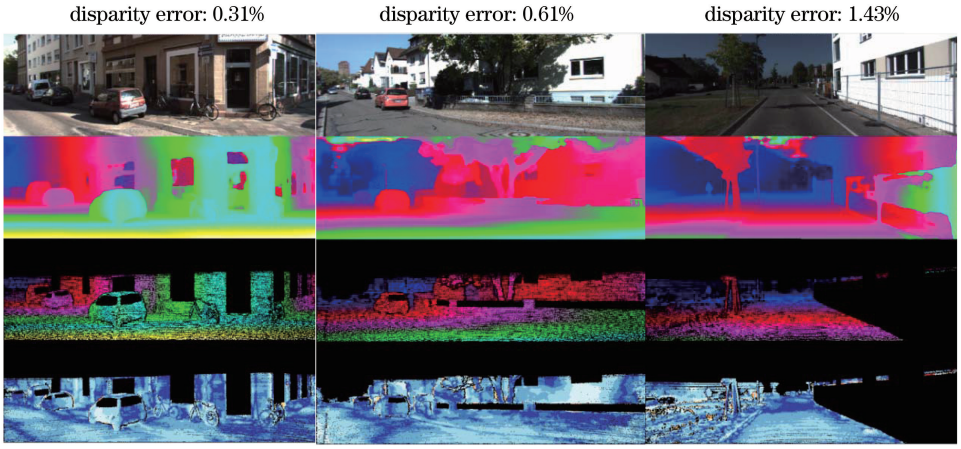

图 6　在 KITTI2012 评价结果，从上到下依次为左输入、预测视差图、实际视差图、误差图

Fig. 6　Results on KITTI2012, from the top to the bottom: the left input, predicted disparity map, actual disparity map, error map

为了进一步验证本文视差估计算法的有效性，在 KITTI 立体匹配数据集中，与改进的基准 PSM-Net 和 KITTI 榜上排名靠前的 GWC-Net 进行对比分析，结果如图8所示。可以看出，在一些弱纹理、

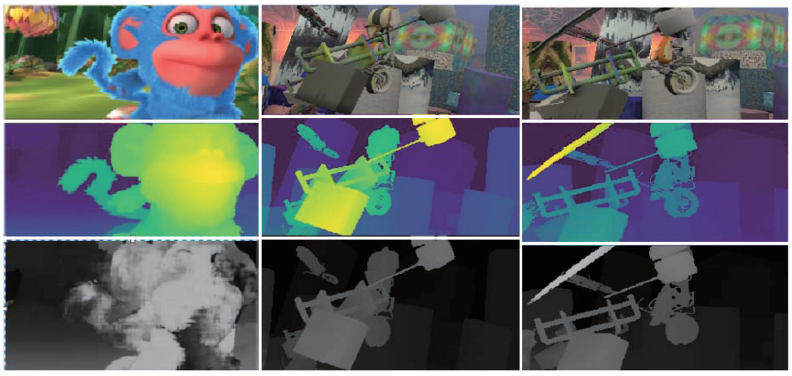

图 7　在 Sceneflow 评价结果，从上到下依次为左输入、实际视差图、预测视差图

Fig. 7　Results on Sceneflow, from the top to bottom: the left input, actual disparity map, predicted disparity map

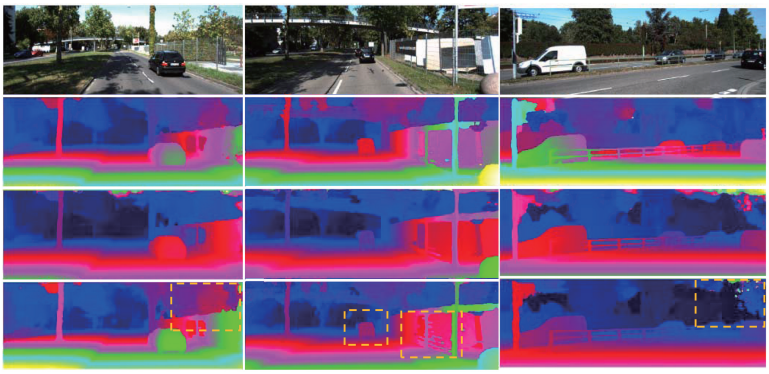

图 8　与其他重建结果比较，从上到下依次为 PSM-Net 预测视差图、GWC-Net 预测视差图、本文得到的视差图、方框框出来的为改进的部位

Fig. 8　Comparison with other algorithms, from the top to the bottom: the PSM-Net results, the GWC-Net results, our results, our improvement results for the parts framed

前后区分度不高的区域，如蓝天、路面、栅栏、车辆，本文算法仍能够生成更加精准的视差信息。

3.2　实验对比

本文算法是基于原始的 PSM-Net 作了改进，将本文算法与近些年神经网络三维重建方法的客观误差指标结果作比较，比较结果如表 1 所示。对于 KITTI2012 数据集，将其分成 180 个图像对的训练集和14个图像对的验证集，最后在训练150轮后提交结果。评价标准采用非遮挡区域中错误像素的比例(Out-Noc)和所有区域中错误像素的比例(Out-All)。本文算法的结果在误差不超过 2，3，4 pixel 的标准下都超过基准 PSM-Net，在误差不超过 3 pixel 的判断标准下超过基准 4.69%。其中原始 PSM-Net 版本每对图像的运行速度为 28 s，本文算法运行时间为 30 s，在速度上基本持平，但在主观和客观视差结果指标上都有一定程度的改善。表中数

表 1　各算法视差对比

Table 1　Comparison with other algorithms

Method	Sceneflow			KITTI2012				KITTI2015	
	EPE	2 pixel		3 pixel		4 pixel		ALL	DOC
		Out-Noc	Out-All	Out-Noc	Out-All	Out-Noc	Out-All	D_1-all	D_1-all
MC-CNN (Žbontar et al., 2016)[2]	3.79	3.90	5.45	2.43	3.63	1.90	2.85	3.88	3.33
GC-Net (Cao et al., 2019)[3]	2.51	2.71	3.46	1.77	2.30	1.36	1.77	2.67	2.45
iResNet-i2 (Liang et al., 2018)[20]	1.40	2.69	3.34	1.71	2.16	1.30	1.63	2.44	2.19
PSM-Net (Chang et al., 2018)[5]	1.09	2.44	3.01	1.49	1.89	1.12	1.42	2.32	2.14
SegStereo (Yang et al., 2018)[14]	1.45	2.66	3.19	1.68	2.03	1.25	1.52	2.25	2.08
GA-Net (Zhang et al., 2019)[7]	0.84	2.18	2.79	1.36	1.80	1.03	1.37	1.93	1.73
Ours	0.95	2.33	2.98	1.42	1.76	0.92	1.21	2.22	2.07

据以 2 pixel 的像素标准作为阈值，数据表明，相比 PSM-Net，本文的算法误差降低了约 11%，相比 MC-CNN[2]方法降低了约 14%，并且在与经典算法对比中，本文网络结构在 KITTI2012 和 KITTI2015 数据集上的错误率均较低。

4 结　论

提出一种基于注意力机制的立体匹配方法，在特征之间建立联系机制，自适应地聚集一定范围内的上下文信息，从而改进立体匹配的特征表示，并在损失函数中引入语义编码损失，使得重建结果更加精确。在 KITTI2012，KITTI2015，Sceneflow 三个数据集上进行验证，与几种典型的基于深度学习的方法相比，所提算法在整体精度上取得了最优性能，特别是与基准方法相比，提高了立体匹配的精度，尤其是在弱纹理区域。但算法时间开销相比基准方法较大，仍无法满足实时性要求，为得到精度高、轻量的、运行效率高的立体匹配算法，仍需要对算法的网络结构进行更深入的研究。同时随着无监督网络的广泛应用，立体匹配可以逐步减少对大规模、带有真实值数据集的依赖，这不仅使网络的训练更加容易，还有助于提高网络的泛化能力。立体匹配已成为我们下一步研究的目标。

参　考　文　献

[1] Scharstein D, Szeliski R. A taxonomy and evaluation of dense two-frame stereo correspondence algorithms [J]. International Journal of Computer Vision, 2002, 47: 7-42.

[2] Žbontar J, LeCun Y. Stereo matching by training a convolutional neural network to compare image patches[EB/OL]. (2016-05-18)[2020-02-25]. https://arxiv.org/abs/1510.05970.

[3] Cao Y, Xu J R, Lin A, et al. GC-Net: non-local networks meet squeeze-excitation networks and beyond [EB/OL]. (2019-04-25) [2020-02-25]. https://arxiv.org/abs/1904.11492v1.

[4] Seki A, Pollefeys M. SGM-Nets: semi-global matching with neural networks [C] // 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), July 21-26, 2017, Honolulu, HI, USA. New York: IEEE Press, 2017: 6640-6649.

[5] Chang J R, Chen Y S. Pyramid stereo matching network [C] // 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 18-23, 2018, Salt Lake City, UT, USA. New York: IEEE Press, 2018: 5410-5418.

[6] Khamis S, Fanello S, Rhemann C, et al. StereoNet: guided hierarchical refinement for real-time edge-aware depth prediction[M] // Ferrari V, Hebert M, Sminchisescu C, et al. Computer Vision-ECCV 2018. Cham: Springer International Publishing, 2018: 596-613.

[7] Zhang F H, Prisacariu V, Yang R G, et al. GA-net: guided aggregation net for end-to-end stereo matching [EB/OL]. (2019-04-13) [22020-02-25]. https://arxiv.org/abs/1904.06587.

[8] Sarder P, Nehorai A. Deconvolution methods for 3-D fluorescence microscopy images [J]. IEEE Signal Processing Magazine, 2006, 23(3): 32-45.

[9] Xiao J S, Tian H, Zou W T, et al. Stereo matching based on convolutional neural network [J]. Acta Optica Sinica, 2018, 38(8): 0815017.
肖进胜, 田红, 邹文涛, 等. 基于深度卷积神经网络的双目立体视觉匹配算法[J]. 光学学报, 2018, 38(8): 0815017.

[10] Jiang H Q, Cai Y, Zhang J S, et al. Research on 3D reconstruction algorithm based on improved SFM[J]. Computer Technology and Its Applications, 2019, 45(2): 88-92.
蒋华强, 蔡勇, 张建生, 等. 基于改进 SFM 的三维重建算法研究[J]. 电子技术应用, 2019, 45(2): 88-92.

[11] 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.

[12] Si S H, Hu F Y, Gu Y J, et al. Improved denoising algorithm based on non-regular area Gaussian filtering [J]. Computer Science, 2014, 41(11): 313-316.
姒绍辉, 胡伏原, 顾亚军, 等. 一种基于不规则区域的高斯滤波去噪算法[J]. 计算机科学, 2014, 41(11): 313-316.

[13] Zhao H S, Shi J P, Qi X J, et al. Pyramid scene parsing network [J]. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), July 21-26, 2017, Honolulu, HI, USA. New York: IEEE Press, 2017: 6230-6239.

[14] Yang M K, Yu K, Zhang C, et al. Dense ASPP for semantic segmentation in street scenes [C] // 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 18-23, 2018, Salt Lake City, UT, USA. New York: IEEE Press, 2018:3684-3692.

[15] He K M, Zhang X Y, Ren S Q, et al. Deep residual learning for image recognition [J]. 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.

[16] Wang J Q, Li J S, Zhou X W, et al. Improved SSD algorithm and its performance analysis of small target detection in remote sensing images[J]. Acta Optica Sinica, 2019, 39(6): 0628005.
王俊强, 李建胜, 周学文, 等. 改进的SSD算法及其对遥感影像小目标检测性能的分析[J]. 光学学报, 2019, 39(6): 0628005.

[17] Yu F, Koltun V, Funkhouser T. Dilated residual networks[C]//2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), July 21-26, 2017, Honolulu, HI, USA. New York: IEEE Press, 2017: 636-644.

[18] Jaderberg M, Simonyan K, Zisserman A, et al. Spatial transformer networks[EB/OL]. (2016-02-04) [2019-02-25]. https://arxiv.org/abs/1506.02025.

[19] Zhang H, Dana K, Shi J P, et al. Context encoding for semantic segmentation[J]. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 18-23, 2018, Salt Lake City, UT, USA. New York: IEEE Press, 2018: 7151-7160.

[20] Liang Z F, Feng Y L, Guo Y L, et al. Learning for disparity estimation through feature constancy[C]// 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition(CVPR), June 18-23, 2018, Salt Lake City, UT, USA. New York: IEEE Press, 2018: 18347645.