DOI:10.19652/j.cnki.femt.2204394

融合注意力和多尺度的优化立体匹配算法研究*

谢 鑫 张 博 张美灵 朱 磊
(西安工程大学电子信息学院 西安 710048)

摘 要:当前基于卷积神经网络的立体匹配方法未充分利用图像中各个层级的特征图信息,造成图像在不适定区域的特征提取能力较差,因此,提出了一种基于PSMNet改进的优化立体匹配算法。在特征提取阶段,全新的特征金字塔模块(SPP)能更好的聚合不同尺度和不同位置的环境信息构建代价体,从而充分利用全局环境信息;在构建匹配代价体时,提出组相关的策略来充分地利用特征中的全局和局部信息;在代价聚合阶段,优化沙漏结构并引入通道注意力机制以便网络来提取具有高表示能力和高质量通道注意力向量的信息特征;为了进一步优化视差图,设计视差优化网络来改善初始的视差估计。在Scene Flow、KITTI 2012和KITTI 2015立体数据集上评估,所提模型在Scene Flow数据集上平均预测误差EPE降低到0.71 pixels,在KITTI 2012和KITTI 2015立体数据集上的误匹配率分别下降到1.20%和1.86%,在实验结果表明,方法取得了较优越的性能。

关键词:立体匹配;深度学习;注意力机制;卷积神经网络;分组相关量;视差优化
中图分类号:TP391 **文献标识码**:A **国家标准学科分类代码**:510.70

Research on optimal stereo matching algorithm combining attention and multi-scale

Xie Xin Zhang Bo Zhang Meiling Zhu Lei
(School of Electronics and Information, Xi′an Polytechnic University, Xi′an 710048, China)

Abstract: This paper presents an improved stereo matching algorithm based on PSMNet. In the feature extraction stage, the new SPP feature pyramid module can better aggregate the environmental information of different scales and different locations to construct cost volume, in order to make full use of the global environmental information. When constructing the matching cost volume, the group correlation strategy is proposed to make full use of the global and local information in features. In the cost aggregation stage, the hourglass structure is optimized and the channel attention mechanism is introduced so that the network can extract the information features with high representation ability and high quality channel attention vector. In order to further optimize the disparity map, a disparity optimization network is designed to improve the initial disparity estimation. The method in this paper is evaluated on Scene Flow, KITTI 2012 and KITTI 2015 stereo datasets, and the average prediction error EPE of the proposed model on Scene Flow dataset is reduced to 0.71 pixels. The mismatching rates on KITTI 2012 and KITTI 2015 stereo datasets decreased to 1.20% and 1.86%, respectively. The experimental results show that the proposed method achieves superior performance.

Keywords: stereo matching; deep learning; attention mechanism; convolution neural network; group correlation quantity; parallax optimization

0 引 言

不同视点影像之间的同名像点信息的搜索可以通过立体匹配来实现,立体匹配已成为双目立体视觉中最核心、最重要的一个环节,尤其在基于影像的三维重建[1-2]技术中起着至关重要的作用。其寻找的信息促进了多个领

收稿日期:2022-10-07
* **基金项目**:国家自然科学基金(61971339)、陕西省重点研发计划(2019GY-113)、陕西省自然科学基础研究计划(2019JQ-361)项目资助

域的应用发展，例如基于低纹理匹配的视觉任务[3]、数字表面模型及数字高程模型的制作、实景三维模型重建[4]、自动驾驶[5]、多视影像的结构恢复、智能机器人、生物医学[6]等。传统的立体匹配算法主要围绕损失计算和视差优化展开研究。在计算匹配损失时，主要通过设计良好的度量函数来优化匹配损失；但是，传统算法均采用人工设计的浅函数，在一些病态区域（如反射表面、弱纹理、反光等）处效果不佳，存在大量孔洞和误匹配，在复杂场景下难以实现应用，无论是速度还是精度方面，传统算法的发展都不尽人意。随着深度学习在语义分割、目标检测与识别等高级视觉技术方面取得了阶段性的进展，立体匹配任务也可以通过卷积神经网络（convolutional neural networks，CNN）来实现。卷积神经网络由于拥有较强的特征提取能力，通过简单的非线性模型从原始图像中提取出更加抽象的特征，并且在整个过程中只需少量的人工参与，从而取代了传统手工制作特征描述符的方法。将端对端的神经网络进行立体匹配时，神经网络在多层次、多方面挖掘全局特征信息，在速度和精度方面相比传统算法都发生了大的跳跃，算法具有很强的鲁棒性。

深度学习在各个研究领域都得到了广泛的应用，尤其是卷积神经网络，不仅提高了图像识别和分类的准确性，还提升了在线运算效率，立体匹配可以借助深度学习方法完成。文献[7-8]提出了 MC-CNN，通过从矫正后的图像对中提取深度信息，训练一个卷积神经网络来预测两个图像块的匹配程度并计算立体匹配代价，一举超越了传统算法。文献[9]巧妙地在 MC-CNN 网络中引入 SGM 算法的思想，提出了匹配置信度融合方法。文献[10]构建了双目视觉标准数据集 FlyingThings3D，标志着基于端到端的深度学习开始广泛地被应用到立体匹配算法当中。文献[11]将全卷积神经网络（fully convolutional networks，FCN）应用在深度学习任务语义中取得了阶段性的效果，提出了在光流估计和视差估计当中引入了端到端的神经网络 DispNet，在新建立的数据集上成功的训练了网络，当时在 KITT12012 在线排行榜上排名第 1，错误率达到了 1.75%，匹配速度达到了 0.06 s。文献[12]在 DispNet 的基础上引入多尺度进行残差学习并且添加了上采样来获得更精细化的视差图。文献[13]所提出的 GC-Net 将立体匹配问题转化成了回归问题，利用了图像对的上下文的邻域信息和场景的几何信息，从核线立体像对中端到端的直接输出视差。文献[14]提出 iResNet 网络，在端对端的立体匹配网络中整合传统立体匹配的 4 个步骤，将得到的初始视差图与源图像空间中的特征进行对比反馈优化，输出最终的视差图。文献[15]设计的一个金字塔网络 PSMNet 则是在立体匹配过程中利用全局上下文环境信息，将像素级特征扩展至不同尺度的区域级特征，是立体匹配发展历程中最经典的端对端立体匹配网络，具有很高的研究意义。文献[16]设计的 GWC-Net 提出组相关的策略来构建匹配代价体，很大可能的保留了更多代价信息。文献[17]提出的 HSM-Net 采用编码解码器的结构处理输入的特征图，先对立体像对的多尺度特征进行编码操作，再利用金字塔结构由粗到精的对多尺度代价体进行解码操作，得到高分辨率视差图。文献[18]提出的 GA-Net 为了减少内存消耗和计算复杂度，引入了半全局聚合层与局部引导聚合层来代替广泛采用的 3D 卷积层。文献[19]提出基于稀疏点的高效、灵活的代价聚合方式来解决视差不连续处的边缘变粗的问题。文献[20]提出了一种稀疏成本体积网络，在构建 3D 代价体时不使用批处理维度，用 2 D 卷积层处理匹配代价体，提升了匹配速度。文献[21]将深度学习的密集匹配方法应用于航空遥感影像，取得了优于传统方法的效果。文献[22]在用于多视角立体像对的密集匹配中提出了一种循环编码－解码器的结构，提高了匹配的精度。文献[23]提出了一种由粗到精的立体匹配算法，可以获得更精细化的视差。但这些方法都存在网络结构复杂，消耗高的问题。

为了更好地利用全局上下文信息来进行立体匹配，本文提出了一种新颖的卷积神经网络。利用卷积神经网络训练左右图像的特征表示，并计算出立体匹配的匹配代价。本文所选的基准网络是经典端对端立体匹配网络 PSMNet，在特征提取模块，为了避免其因池化操作会导致像素信息丢失的问题，引入全卷积和注意力相结合的策略；在代价计算阶段，为了提供更好的相似性度量，在 PSMNet 串联代价体的基础上提出分组相关性来构建代价量；在代价聚合阶段，引入通道注意力机制来识别高质量的通道特征信息，并且引入多个 1×1×1 的卷积来进行跳跃连接，此操作优化了 PSMNet 原来的沙漏结构；然后，通过双线性插值和视差回归得到初始的视差。为了进一步优化视差计算得到的视差图，还设计了一个有效的视差优化网络来进一步优化最终的视差估计，视差优化网络引入了多个不同膨胀率的空洞卷积来扩大感受野。

1 基于深度学习的立体匹配

基于深度学习的端对端双目立体匹配是从右图像中找到左图像对应的同名点，不需要任何后处理的密集视差图的输出过程。将深度学习的理念用于立体匹配的步骤一般如下：首先，双目图像首先经过权重共享的卷积神经网络进行特征提取，大部分都是用 RESNet 作为基础网络。接下来，对左右特征图进行代价计算，通过融合左右特征图来构建匹配成本特征体。最后，3DCNN 网络对输入的成本特征体进行代价聚合和视差计算，以扩展在成本量中的上下文信息的区域支持。本文以经典端对端立体匹配网络 PSMNet 作为基准网络，其网络结构如图 1 所示。

本文改进后的结构如图 2 所示，该网络使用具有共享权重的 CNN 处理左右视图，分别用于特征提取。改进的 SPP 模块连接不同尺度的子区域以更好的使用全局信息，代价计算主要采用分组相关性来构建代价量，在代价聚合

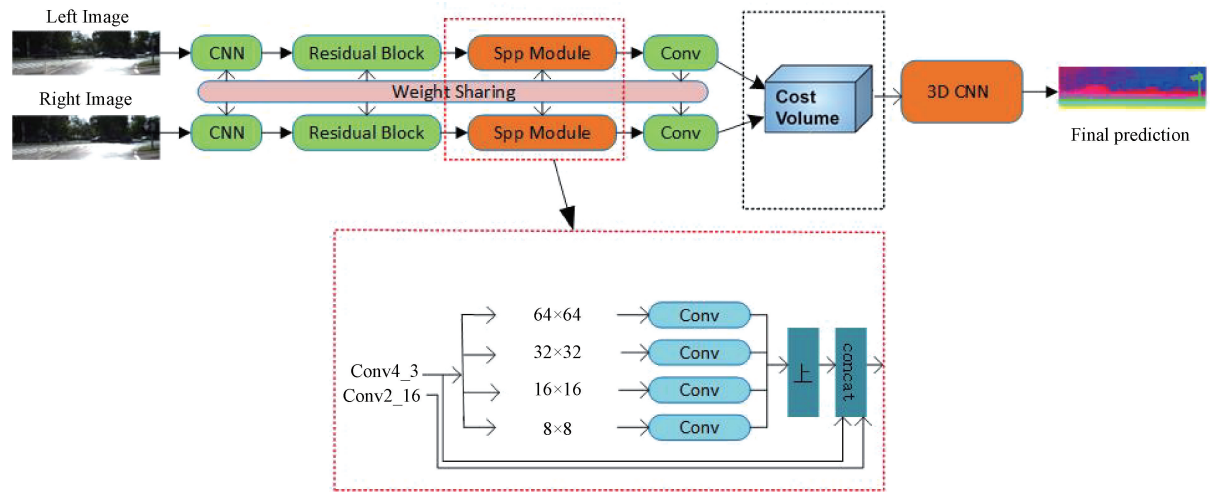


图 1　原 PSM 网络结构

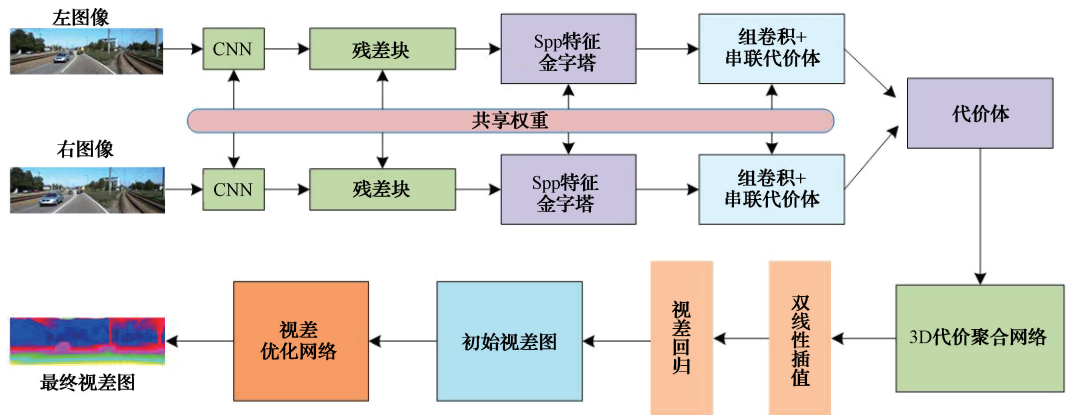


图 2　改进的网络结构

阶段引入通道注意力机制来获得高质量的特征信息，并且优化了原有的沙漏结构，然后通过双线性插值和视差回归得到初始的视差，最后，还引入了一个有效的视差细化网络来进一步改进初始视差，得到最终的视差估计。

1.1　改进的 SPP 特征金字塔模块网络结构

SPP 模块简单来说就是几个不同尺度的池化，再通过上采样串联起来，是整合多尺度信息的过程，立体匹配的效果得利于结合不同级别的特征来完善，为了进一步扩大感受野，使像素级特征扩展至多尺度区域特征，能够更加有效的结合全局环境和一个重叠的沙漏模块去用于匹配代价聚合。由于池化操作会导致像素信息的丢失，因此，本文主要依赖于全卷积和注意力相结合的设计思想提出一种新的 SPP 特征金字塔模块网络结构，改进后的 SPP 特征金字塔模块网络结构如图 3 所示。

经过 CNN、残差(Basicblock)特征提取操作后特征图的大小变成了原来的 1/4，使用全卷积和注意力机制相结合的策略去避免池化操作丢失像素信息的问题，以输出 1/4 特征图作为此模块的输入，设计的 3 个尺度各自包含两个连续的卷积层，前一个尺度卷积的输出作为后一个尺度卷积的输入，以此类推，然后将不同尺度的输出相加。模块的输入经过 1×1 的卷积操作后与 3 个不同的尺度进行求和操作后再进行相乘，金字塔模块的输入还经过了全局池化，然后通过一个 1×1 的卷积和上采样(upsample)操作后与通道金字塔特征模块相加作为最终特征提取模块的输出。

1.2　代价计算方式的优化

在 PSMNet 中，必须通过 3D 聚合网络从头开始学习级联特征的匹配代价，这通常需要更多的参数和计算成本。相比之下，完全相关则提出一种依赖于点积测量特征相似性的有效方法来计算匹配代价。然而，由于它在每个视差水平产生的相关图是单通道的，导致在操作进程中丢失信息较多。级联量不包含有关特征相似性的信息，因此在聚合网络中需要更多的参数来从头开始学习相似性测量功能。本研究通过结合级联量和相关量的优点，在构建匹配代价体时提出组相关的策略，从而避免在计算匹配代价时丢失大量信息。其原理是对多通道的特征图沿着通

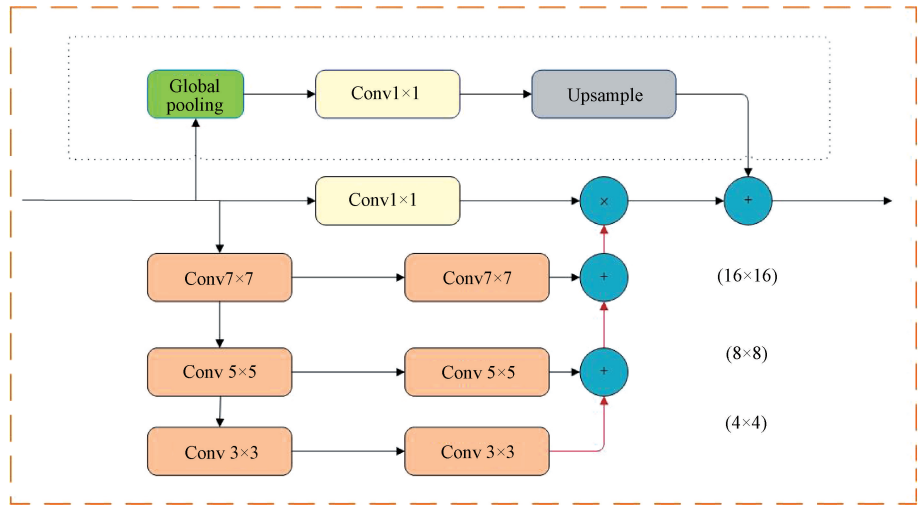


图 3　SPP 特征金字塔模块网络结构

道进行分组，左右特征组间的相似度关系依据向量内积的计算方式来实现。此进程的优点在于代价信息能够极大程度的保留下来，分组相关性克服了上述这两个缺点，从而为相似性度量提供良好的特征支持。

组相关的基本思想如下：首先将输入特征分割为若干个组，然后通过分组来计算相关图。将一元特征的通道表示为 N_c，所有的通道都沿着通道维度被平均分为 N_g 个组，因此每个特征组就有 N_c/N_g 个通道，（本文设定了 320/40 个通道），第 g 个特征组 f_l^g、f_r^g 由原始特征 f_l、f_r 的第 $g\frac{N_c}{N_g}$，$g\frac{N_c}{N_g}+1$，…，$g\frac{N_c}{N_g}+\left(\frac{N_c}{N_g}-1\right)$ 个通道构成。组相关计算过程如下：

$$C_{gwc}(d,x,y,g)=\frac{1}{N_c/N_g}\langle f_l^g(x,y),f_r^g(x-d,y)\rangle \tag{1}$$

其中，$\langle\cdot,\cdot\rangle$是两个特征向量的内积。计算所有分组 g 的相关，而且是在所有的视差层级 d，视差的搜索和特征图要统一。

将所有的相关图打包为一个匹配代价量，形状 $[D_{max}/4,H/4,W/4,N_g]$，D_{max} 表示最大视差，$D_{max}/4$ 对应着该特征的最大视差。当 $N_g=1$ 时，组相关即成为完全相关。本研究所用的代价体的构建方式是将组卷积代价体和串联代价体相结合的策略，具体结构如图 4 所示。

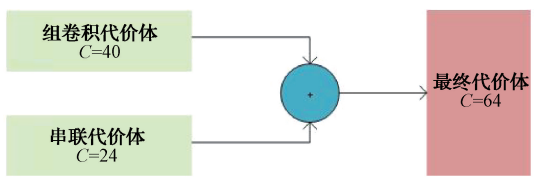


图 4　代价计算方式结构

1.3　代价聚合模块

在代价聚合阶段，本文对堆叠的 3 个沙漏网络进行了改进，通过 4 种相同的编码解码器结构融合多尺度的代价量，以扩大接收域和捕获全局信息，并细化遮挡部分和低纹理模糊。为了提高性能，在不增加大量计算成本的情况下，每个沙漏模块采用 1×1×1 的 3D 卷积添加快捷连接。每个沙漏都通过双线性插值和视差回归输出初始视差，将最终的匹配代价体作为第 1 个沙漏结构的输入，第 1 个沙漏结构的输出作为第 2 个沙漏结构的输入，依次类推，本文设计的简化沙漏结构原理如图 5 所示。

在代价聚合阶段加入通道注意力机制来聚合上下文关系。注意力机制可以在卷积神经网络中达到简单而有效的部署，不仅没有增加很多参数，而且可以有效增强网络结构的表达力，进而在重建过程中能够更加精确地重建出弱纹理区域，本研究在代价聚合模块引入的通道注意力机制结构原理如图 6 所示。

在代价聚合模块使用沙漏结构，其输入为代价计算模块提出的组相关构建的代价体，提出的组相关代价体由组卷积代价体和串联代价体构成。其中，串联代价体由左特征图和右特征图构成，各自的通道数是 12，串联起来最终的串联代价体的通道数就是 24，组卷积代价体的通道数是 40，两者相加通道数为 64，因此，这里输入的特征图的尺寸为[$C=64$，$1/4H$，$1/4W$]，然后，设计了两个卷积块，每个卷积块由两个 3×3×3 的卷积堆叠构成，4 个卷积的大小相同，步长为 1，不进行下采样操作。并且在两个卷积块之间通过跳跃连接来聚合上下文信息，第 1 个卷积块的第 1 个卷积对通道进行了下降到 32 的操作，除了最后的一个卷积没有用 ReLU 进行激活处理，其余的 3 个卷积后面都使用了 ReLU 对数据进行归一化处理。最后将通道为 32 的特征图作为沙漏结构的输入。

对于沙漏结构，本文首先使用两个步长为 2 的 3D 卷积进行下采样，再接入一个步长为 1 的 3D 卷积，最后采用两个步长为 2 的 3D 反卷积进行上采样，并且在这之间加入多个 1×1×1 的 3D 卷积进行跳跃连接来聚合多尺度信

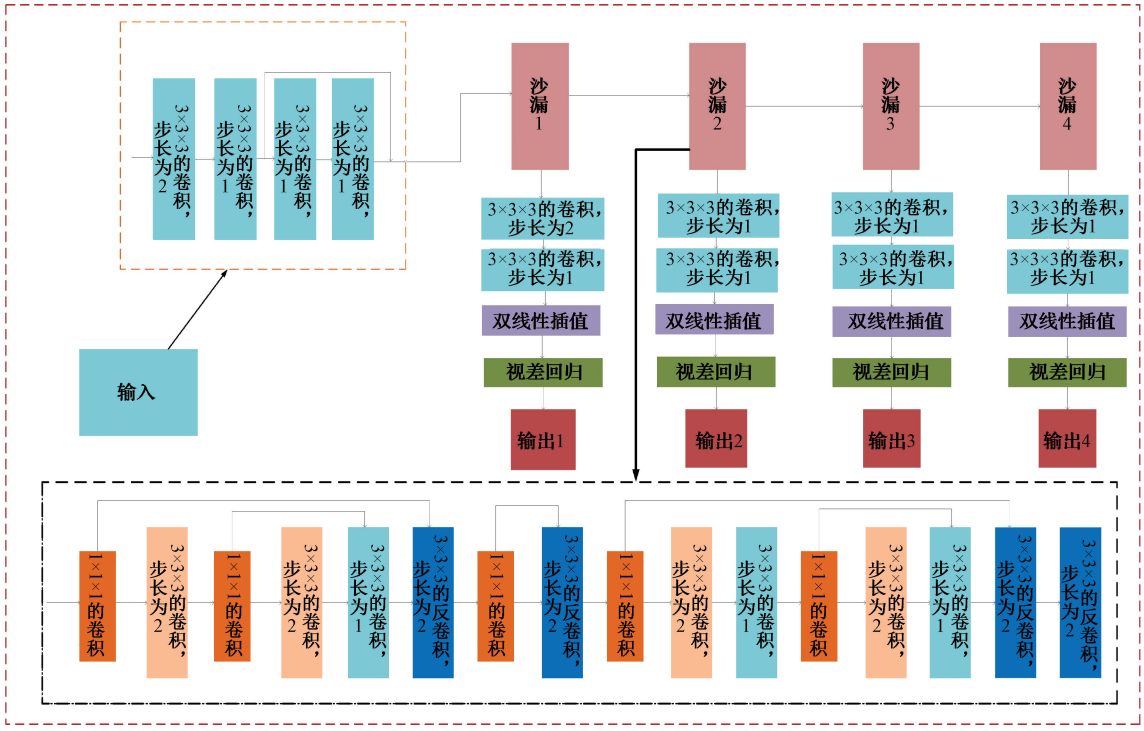


图 5　代价聚合模块结构

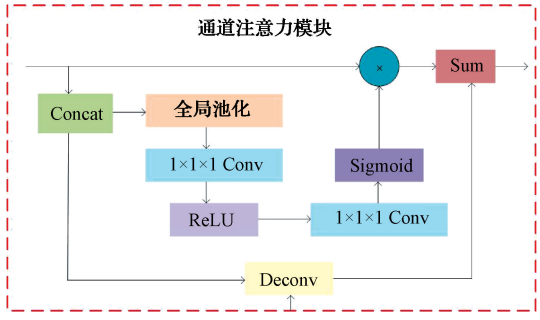


图 6　通道注意力机制结构原理

息，形成一次编解码操作，再重复一次上述编解码的过程，同样在这之间加入多个 1×1×1 的 3D 卷积进行跳跃连接来聚合多尺度信息。其中，上采样和下采样卷积的尺寸大小均设置为 3×3×3。这样，编码-解码-编码-解码的过程就形成了一个沙漏结构。沙漏结构对特征图的编解码原理如图 7 所示。

1.4　视差优化网络

为了进一步优化代价聚合得到的视差图，设计了一个视差优化网络来进一步提升视差精度。本文设计的视差优化网络结构如图 8 所示。

视差优化网络的输入由沙漏（hourglass）结构输出的最后一个视差图、左特征图和右特征图构成。首先，通过右特征图来推理左特征图，然后用右图推理出来的左图跟真实的左图做误差操作，最后，将初始视差图和重构误差在特征通道维度上连接起来形成组合特征图。先对组合

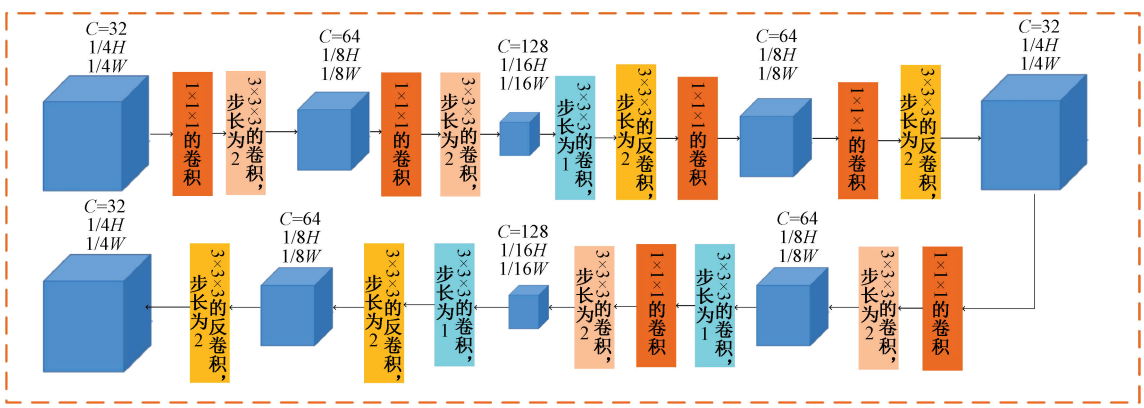


图 7　沙漏结构原理

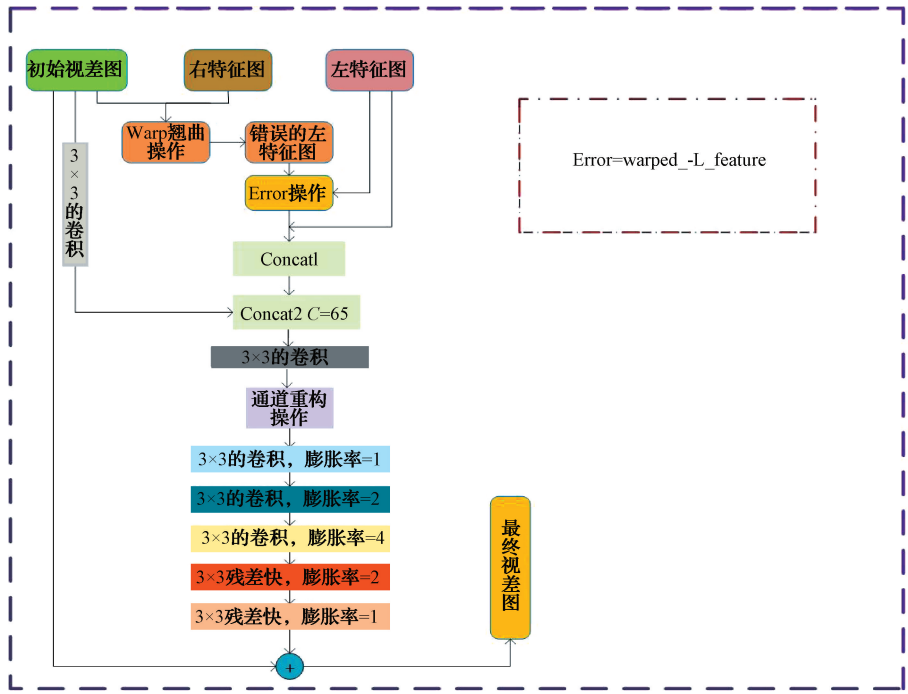


图 8　视差优化网络的体系结构

的特征图进行通道重构操作，再经过一系列的卷积操作。本文设置了 3 个不同膨胀率的卷积来有效的优化低纹理区域，卷积的膨胀率自顶向底依次设定为 1、2、4，同时还设计了两个不同膨胀率的残差块来增强场景的理解能力，残差块的膨胀率自顶向底依次设定为 2、1。

2　视差回归和损失函数

本文采用视差回归的方式来估算连续的视差图。根据 Softmax 操作得到预测代价，d 来计算每一个视差值 d 的可能性。预测视差值 $\hat{d}$ 如下：

$$\hat{d}=\sum_{d=0}^{D_{\max}} d\times\sigma(-c_d) \tag{2}$$

其中，$\sigma(.)$ 表示 Softmax 操作，Softmax 函数将正则化后的代价体中的各个像素值转换为该视差下的概率值，对应点在其相对应匹配图像中位置的最大偏移量用最大视差 $D_{\max}$ 来表示。由于代价越小的概率会越大，因此出现了负号；代价越大，意味着视差的权重越小，最终视差是对网络对应视差的加权平均。视差值回归比基于分类的立体匹配方法鲁棒性更强。

在视差回归中采用平滑的 L1 损失函数来训练网络。平滑的 L1 被广泛应用于物体检测的边缘箱型回归任务里，因为它比 L2 损失函数具有更高的鲁棒性，同时可以避免梯度爆炸问题，损失函数定义如下：

$$L(d,\hat{d})=\frac{1}{N}\sum_{i=1}^{N} smooth_{L1}(d_i-\hat{d}_i) \tag{3}$$

其中：

$$smooth_{L1}(x)=\begin{cases}0.5x^2, & |x|<1\\ |x|-0.5, & \text{其他}\end{cases} \tag{4}$$

式中：N 是标记的像素的数量；d 是真实视差值；$\hat{d}$ 是预测的视差值。为了更深层次全面的学习整个网络体系，每一个 hourglass 结构后面都会进行视差回归操作。

3　实　验

3.1　数据集及其评价指标介绍

1）Scene Flow

SceneFlow 数据集是一个大型的合成数据集，包含 35 454 对训练图像和 4 370 对测试图像，整个数据集均不是在真实场景下拍摄而来的，并且提供了精细且密集的真实视差图。包含 FlyingThing3D、Monkaa 和 Driving 3 个子集，图像分辨率为（$H=540$）·（$W=960$）。对于 SceneFlow 数据集，通常使用终点误差（EPE）作为评估度量的标准，即以像素为单位的平均视差误差。

2）KITTI 2012

该数据集，包含 194 个由 LiDAR 获得的具有稀疏真实视差的训练立体对和另外 195 个没有真实视差的测试立体对。立体图像对和真实视差的大小均为（376 × 1 280 pixels）。评价标准主要是非遮挡（Noc）和所有（all）像素的错误像素以及平均终点误差的百分比。

3）KITTI 2015

这是一个真实的数据集，从具有动态视图的汽车的角度来看，包括城市、乡村和高速公路。包含 200 个由 LiDAR 获得的具有稀疏真实视差的训练立体对和另外 200 个没有真实视差的测试立体对。立体图像对和真实视差的大小均为（376×1 280 pixels）评价标准主要是预测异常值的百分比差距 D1，在标有“Non-occluded Areas”统计了

非遮挡区域的错误率，标有“All pixels”则会将整张图片中的像素点统计到错误率的计算中。两种情况都统计了“bg”、“fg”和“all” 3 种细分区域，在标有“bg”的细分区域是统计背景区域的错误率；在标有“fg”的细分区域则是统计场景中前景区域的错误率；在标有“all”的细分部分，则会将前景和背景的错误率都考虑进去。

3.2 实验细节

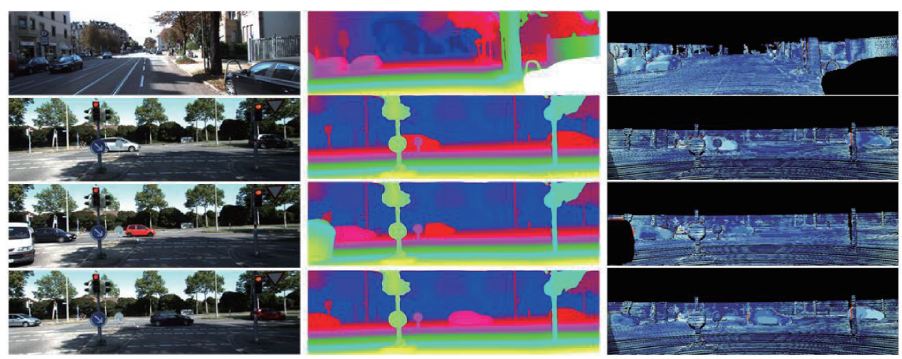
研究的体系结构是基于 PyTorch 实现的。所有模型均采用 Adam 端到端训练（$\beta_1 = 0.9, \beta_2 = 0.999$）。对整个数据集进行颜色归一化及数据预处理，在训练过程中，将图像裁剪到大小 $H=256, W=512$。最大视差 D 大小设置为 192。为了防止小数据集被大数据集淹没，研究采用 3 阶段训练策略：1）预训练阶段，模型在 Sceneflow 训练集上进行预训练，设定 epoch 为 36，batch size 为 12，前 16 个阶段的学习率设置为 0.001，在之后的过程中，设定每 6 个 epoch 学习率减少 1/2；2）联合训练阶段，使用 KITTI 数据集对在 Scene Flow 数据集预训练完成的模型进行微调，最终选择验证效果最好的模型，在 KITTI 2012 和 2015 训练集上进行联合训练，设定 epoch 为 600，训练初期的学习率设置为 0.001，在 300 epochs 后下降到起初的 10%；3）单独训练阶段，继续在 KITTI 2012 和 2015 训练集上进行单独训练，设定要完成的 epoth 为 400，训练的初始学习率设定为 0.001，在 200 epochs 后下降到 0.000 1。

因为训练过程需要计算损失，本文采用的是分别计算每个输出视差图的损失，主要包括代价聚合阶段 4 个沙漏结构的 4 个初始输出视差图和视差优化网络的一个最终输出视差，最后进行加权和，加权和的分配比例设置为 [0.5, 0.5, 0.7, 1.0, 1.3]，在测试阶段，设置视差优化网络最终输出的一个视差图作为网络模型的最终视差输出。

3.3 实验结果

1）KITTI 2015 的基准结果

所提模型在 KITTI 2015 数据集上进行测试，结果如图 9 所示。将改进的网络可视化预测视差图结果上传至 KITTI 官方网站进行 KITTI 2015 指标的评估对比，在线排行榜主要根据“All Pixels”的 D1-all 误差对所有方法进行排名，将反馈的结果和现有的相关结果进行对比，结果如表 1 所示。



(a) 左图像　(b) 视差图　(c) 误差图

图 9　KITTI 2015 测试数据定性结果

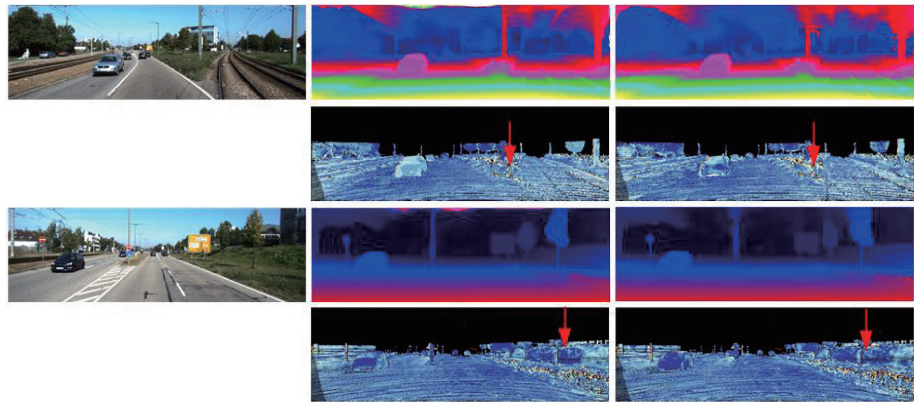
表 1　各算法在 KITTI 2015 数据集上的视差对比

方法	All plxels			Non-Occluded			运行时间/s
	D1-bg	D1-fg	**D1-all**	D1-bg	D1-fg	D1-all	
iResNet[14]	2.35	**3.23**	2.50	2.15	**2.55**	2.22	**0.12**
GC-Net[13]	2.21	6.16	2.87	2.02	5.58	2.61	0.9
PSMNet[15]	1.86	4.62	2.32	1.71	4.31	2.14	0.41
CFPNet[24]	1.90	4.39	2.31	1.73	3.92	2.09	0.95
CRL[23]	2.48	3.59	2.67	2.32	3.12	2.45	0.47
本文	**1.47**	3.80	**1.86**	**1.34**	3.38	**1.68**	0.95

所提模型与原网络模型在 KITTI 2015 数据集上进行可视化及误差对比，如图 10 所示，错误估计视差点主要在视差误差图的黄色区域显示，图 9 和 10 表明，所提方法可以为精细和复杂重叠的对象获得更精确的视差图，比如茂盛的树叶遮挡区域、红绿灯、指示牌等区域，比原网络 PSMNet 获得了更稳健的结果（图 10 红色箭头）。

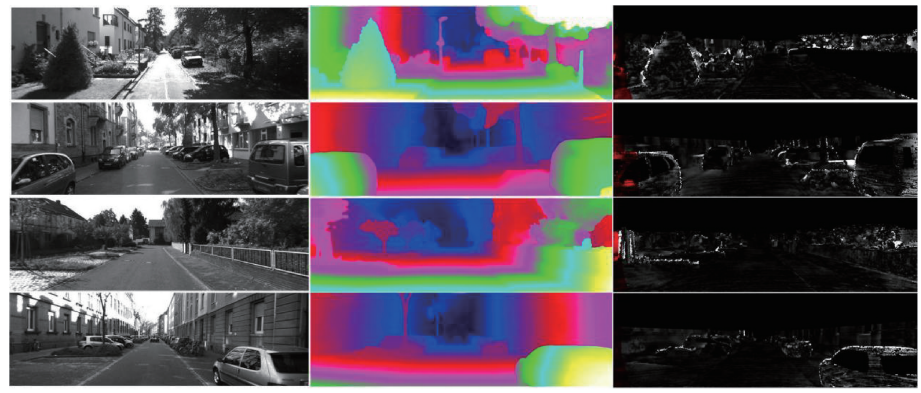
2）KITTI 2012 的基准结果

所提模型在 KITTI 2012 数据集上进行测试，结果如图 11 所示。将 KITTI 2012 生成的预测视差图提交至官方网站中，并将返回的结果同现有的相关算法进行性能的比较，各项指标结果对比如表 2 所示。



(a) 左图像　(b) 所提出模型的结果　(c) PSMNet网络模型的结果

图 10　在 KITTI 2015 测试数据集上与原网络进行比较



(a) 左图像　(b) 视差图　(c) 误差图

图 11　KITTI 2012 测试数据定性结果

表 2　各算法在 KITTI 2012 数据集上的视差对比

模型	>2 pixel		>3 pixel		>5 pixel		Mean error	
	Noc	All	Noc	All	Noc	All	Noc	All
GC-Net[13]	2.71	3.46	1.77	2.30	1.12	1.46	0.6	0.7
PSMNet[15]	2.44	3.01	1.49	1.89	0.90	1.15	0.5	0.6
SGM-Net[25]	3.60	5.15	2.29	3.50	1.60	2.36	0.7	0.9
L-RESMatch[26]	3.64	5.06	2.27	3.40	1.50	2.26	0.7	1.0
GWCNet-gc[27]	2.16	2.71	1.32	1.70	0.80	1.03	0.5	0.5
本文	**1.96**	**2.48**	**1.20**	**1.57**	**0.71**	**0.93**	**0.4**	**0.5**

所提方法与原网络在 KITTI 2012 数据集上进行可视化对比，错误估计视差的点主要集中在视差误差图的红色区域，可见，在纹理质量低、遮挡和反射的区域，所提方法得到的结果更加稳健，尤其是在汽车窗、墙壁区域和模糊区域，模糊区域如左图像的左侧黑色区域，图 11 和 12 表明，所提方法可以实现更高精度的匹配(图 12 黑色方框)。

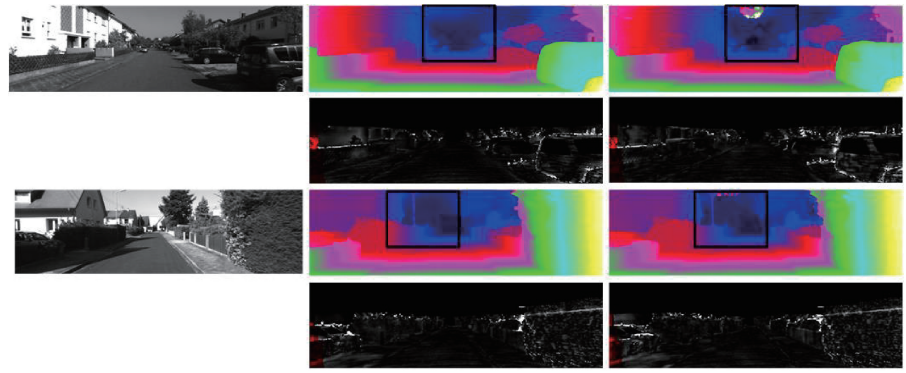
3)Sceneflow 的基准结果

在相同的实验环境配置下，本文对 Sceneflow 测试数据集进行可视化调试，将相应的视差图结果与原网络以及真实视差图进行对比，如图 13 所示。结果表明，所提算法为精细且复杂重叠的对象获得了更准确的视差图(图 13 橙色方框)。

所提算法在 Sceneflow 测试集上最终实现可视化效果，端点错误率为 0.705，相比 PSMNet、GC-Net 和 DispNetC 分别提升了 0.38、1.8 和 0.97。网络的 EPE 性能对比总结如表 3 所示。

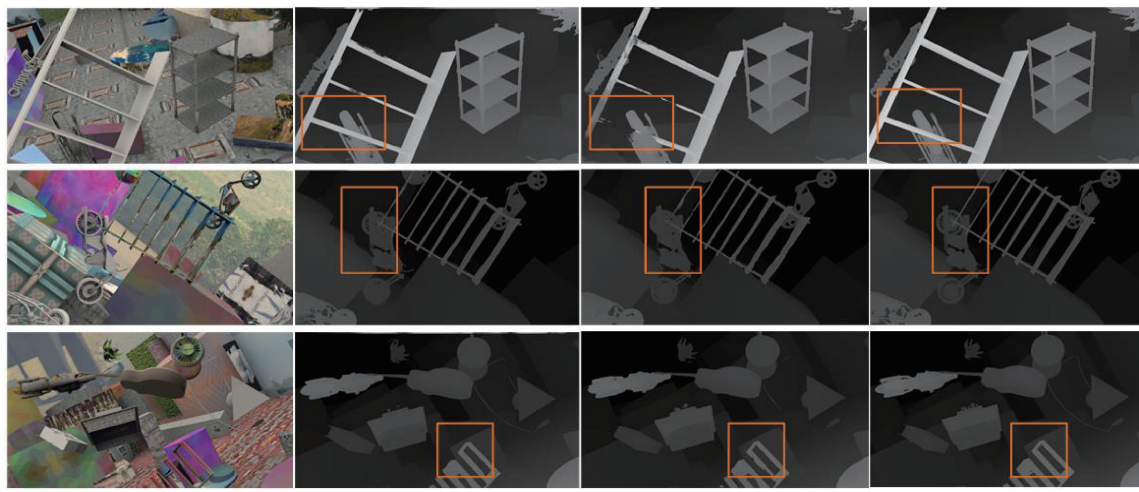
4)消融实验

所提方法在 KITTI 2015 数据集上进行消融实验，以 All plxels 的 D1-all 作为评估标准，SPP attention Network、3D aggregation Network 和 Network Optimization



(a) 左图像　(b) 所提出模型的结果　(c) PSMNet网络模型的结果

图 12　在 KITTI 2012 测试数据集上与原网络进行比较



(a) 左图像　(b) 所提出的模型的结果　(c) PSMNet网络模型的结果　(d) 真实视差图

图 13　所提出模型和 PSMNet 在场景流测试数据集上的结果

表 3　与场景流测试集性能比较 EPE

模型	PSMNet[15]	GC-Net[13]	DispNetC[12]	本文
EPE	1.09	2.51	1.68	0.71

分别代表所提网络的特征提取、代价聚合和视差优化模块，可见，实验指标呈现优化趋势，进而验证了所提网络每个模块的有效性。实验结果如表 4 所示。

表 4　不同网络设置在 KITTI 2015-test 上的结果

Model settings				All plxels (D1-all)
Basic	Spp attention Network	3D aggregation Network	Network Optimization	
√	√			2.45
√	√	√		2.28
√	√	√	√	1.86

4　结　论

本文提出了一种基于 PSMNet 改进的不需要任何后处理的端对端立体匹配算法，为了详细获取像素级的全局上下文信息，在特征提取阶段，提出了一种全新的 SPP 特征金字塔模块，该模块通过多尺度融合注意力来解决因池化操作引起的像素信息丢失的问题；在代价计算阶段，提出分组相关以建立代价量，为 3D 聚合网络提供了良好的匹配特性，提高了性能，降低了聚合网络的参数要求；在代价聚合阶段，对沙漏结构进行优化，引入通道注意力机制以便网络来提取具有高表示能力和高质量通道注意力向量的信息特征；为了进一步优化视差计算得到的视差图，设计了一个有效的视差优化网络来提升视差准度。所提算法无需任何额外的后处理或正则化，相比之下，在整体精度上取得了较好性能，特别是与基准方法 PSMNet 相比，提高了立体匹配的精度，尤其是在弱纹理和无纹理等病态区域具有很高的鲁棒性。

参 考 文 献

[1]　董方新，蔡军，解杨敏. 立体视觉和三维激光系统的联合标定方法[J]. 仪器仪表学报，2017，38(10)：

2589-2596.

[2] 吴渊凯,卞新高. 计算机视觉中摄像机标定的实验分析[J].电子测量技术, 2016, 39(11): 95 99.

[3] 王鑫,王向军,冯登超,等. 特征一致红外弱小目标匹配与定位研究[J]. 电子测量与仪器学报, 2016, 30(9): 1405 1410.

[4] XUE T, XU L, WANG Q, et al. A 3-D reconstruction method of dense bubbly plume based on laser scanning [J]. IEEE Transactions on Instrumentation and Measurement, 2020, 69(5): 2145-2154.

[5] 王昊, 刘雍翡. 基于双目视觉计算的车辆跟驰状态实时感知系统[J]. 中国公路学报, 2019, 32(12): 88-97,105.

[6] 侯亭亭,肖秦琨,杨永侠. 基于动态贝叶斯网络的手势识别[J]. 国外电子测量技术,2015, 34(1):36-39.

[7] ZBONTAR J, LECUN Y. Computing the stereo matching cost with a convolutional neural etwork[C]. Proceedings of the IEEE Computer Society Conference on Computer Vision and I Pattern Recognition, 2015: 1592-1599.

[8] ZBONTAR J, LECUN Y. Stereo matching by training a convolutional neural network to compare image patches [J]. Journal of Machine Learning Research, 2016,17(2): 1-32.

[9] SEKI A, POLLEFEYS M. Patch based confidence prediction for dense disparity map [C]. British Machine Vision Conference,2016, 2(3): 4-12.

[10] MAYER N, ILG E, HAUSSCR P, et al. A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation[C]. IEEE Conference on Computer Vision and Pattern Recognition, 2016: 4040-4048.

[11] LONG J, SHELHAMER E, DARRELL T. Fully convolutional networks for semantic segmentation[C]. IEEE Conference on Computer Vision and Pattern Recognition,2016: 3431-3440.

[12] PANG J, SUN W, REN S Q, et al. Cascade residual learning: A two-stage convolutional neural network for stereo matching [C]. IEEE International Conference on Computer Vision, 2017: 878-886.

[13] KENDALL A, MARTIROSYAN H, DASGUPTA S, et al. End-to-end learning of geometry and context for deep stereo regression[C]. IEEE International Conference on Computer Vision, 2017: 66-75.

[14] LIANG Z, FENG Y, GUO Y, et al. Learning for disparity estimation through feature constancy[C]. IEEE Conference on Computer Vision and Pattern Recognition, 2018: 2811-2820.

[15] ZHANG Z. Microsoft kinect sensor and its effect[J]. IEEE Transactions on Multimedia, 2017, 19(2): 4-10.

[16] GUO X Y, YANG K, YANG W K, et al. Group-wise correlation stereo network [C]. IEEE Conference on Computer Vision and Pattern Recognition, 2019: 3273-3282.

[17] YANG G S, MANELA J, HAPPOLD M, et al. Hierarchical deep stereo matching on high-resolution images [C]. IEEE Conference on Computer Vision and Pattern Recognition, 2019.

[18] ZHANG F H, PRISACARIU V, YANG R G,et al. Torr, GA-Net: Guided aggregation net for end-to-end stereo matching[C]. Proceedings of the IEEE/CVF Conferenceon Computer Virsion and Pattern Recognition, 2019:185-194.

[19] XU H, ZHANG J. AANet: Adaptive aggregation network for efficient stereo matching [C]. Proceedings of the IEEE/CVF Conference on Conference on Computer Vision and Pattern Recognition, 2020:1959-1968.

[20] LU C, UCHIYAMA H, THOMAS D, et al. Sparse cost volume for efficient stereo matching [J]. Remote Sensing, 2018, DOI: 10. 3390/rs10111844.

[21] LIU J, JI S P. Deep learning based dense matching for aerial remote sensing images[J]. Acta Geodaetica et Cartographica Sinica, 2019, 48(9): 1141-1150.

[22] LIU J, JI S P. A novel recurrent EncoderDecoder structure for large-scale multi-view stereo reconstruction from an open aerial dataset[C]. IEEE Conference on Computer Vision and Pattern Recognition, 2020.

[23] XIA L Y, XIAO J, LIN L Q. Segment-based stereo matching using edge dynamic programming [J]. Geomatics and Information Science of Wuhan University, 2011, 36(7): 767-770.

[24] ZHU Z, HE M, DAI Y, et al. Multi-scale cross-form pyramid network for stereo matching[C]. 2019, CoRR abs/1904. 11309.

[25] SEKI A, POLLEFEYS M. SGM-Nets: Semi-global matching with neural networks[C]. IEEE Conference on Computer Vision and Pattern Recognition, 2017: 231-240.

[26] SHAKED A, WOLF L. Improved stereo matching with constant highway networks and reflflective confifidence learning [C]. IEEE Conference on Computer Vision and Pattern Recognition, 2017: 4641-4650.

[27] GUO X Y, YANG K, YANG W K, et al. Group-wise correlation stereo network[C]. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019:3273-3282.

作 者 简 介

谢鑫,硕士研究生,主要研究方向为数字图像处理与计算机视觉。
E-mail:1733747797@qq.com

张博,博士,副教授,主要研究方向为信号检测与信息处理、三维重建等。
E-mail:bozhang@xpu.edu.cn

张美灵,硕士研究生,主要研究方向为数字图像处理与计算机视觉等。
E-mail:996400429@qq.com

朱磊,博士,教授,主要研究方向为 SAR 图像去噪。
E-mail:zhulei791014@163.com