浙江科技学院学报，第 33 卷第 5 期，2021 年 10 月
Journal of Zhejiang University of Science and Technology
Vol. 33 No. 5, Oct. 2021

doi：10.3969/j.issn.1671-8798.2021.05.005

一种改进的多尺度引导聚合立体匹配网络研究

唐家辉[a]，赵　芸[a]，徐　兴[b]
（浙江科技学院 a.信息与电子工程学院；b.机械与能源工程学院，杭州 310023）

摘　要：为了解决现阶段大多数深度学习的立体匹配算法无法优化视差图的边缘结构问题，提出一种多尺度引导聚合立体匹配网络并改进了损失函数。对输入特征进行 4 种不同尺度的空间金字塔特征提取以形成四维代价空间，然后采用半全局代价聚合层（semi-global aggregation layer，SGA）与局部引导聚合层（local guided aggregation layer，LGA）优化代价聚合步骤以生成高精度视差图；为了提高收敛速度并获取更好的初始参数，在微调时引入了 L_2 损失函数。试验结果显示，在 KITTI 2012 数据集以 3 像素为阈值的端点误差评估中取得 98.31% 的准确率，在 KITTI 2015 整体评估中取得了 97.95%的准确率，从而有效地提高了整体的视差精度。研究结果可为双目测距在智创交通中的应用提供一定的参考。

关键词：立体匹配；视差图；多尺度；代价聚合
中图分类号：TP391.41　　**文献标志码**：A　　**文章编号**：1671-8798(2021)05-0378-08

Research on an improved multi-scale guided aggregation stereo matching network

TANG Jiahui[a], ZHAO Yun[a], XU Xing[b]
(a. School of Information and Electronic Engineering; b. School of Mechanical and Energy Engineering, Zhejiang University of Science and Technology, Hangzhou 310023, Zhejiang, China)

Abstract: In order to solve the current problem that most deep learning stereo matching algorithms fail to optimize the edge structure of the disparity map, an improved multi-scale guided aggregation stereo matching network was proposed, accordingly improving the loss function. The network used four different scales of spatial pyramid pooling (SPP) for feature extraction to generate a 4D cost volume. Then the cost aggregation was optimized by virtue of a semi-global aggregation layer (SGA) and a local guided aggregation layer (LGA) to generate the high-precision disparity map. In order to fasten the convergence speed and obtain better

收稿日期：2020-10-28
基金项目：国家自然科学基金项目(61605173)；浙江省自然科学基金项目(LY16C130003)
通信作者：赵　芸(1981—　)，女，浙江省杭州人，教授，博士，主要从事机器智能与视觉研究。E-mail：zy_super0201@163.com。

initial parameters, the L_2 loss function was introduced during fine-tuning. The results show that the network achieves 98.31% accuracy in three-pixel endpoint error evaluation in KITTI 2012 dataset, and 97.95% accuracy in overall evaluation of KITTI 2015, thus having effectively improved the overall disparity accuracy. The results can provide a certain reference for binocular ranging application in intelligent transportation.

Keywords: stereo matching; disparity map; multi-scale; cost aggregation

在过去的几十年中，双目立体匹配广泛应用于无人机[1]、自动驾驶[2]和医疗成像[3]等三维领域，尤其在对三维成像获取物体的深度有着极高精确度要求的医疗与自动驾驶领域。双目立体匹配的原理是通过模拟人眼的视觉感知，采用两个在同一水平线上的传感器获取相同场景的图像，通过两张图片对应同一个像素之间的位置差与相机本身参数之间的关系重建三维场景信息。传统的立体匹配流程通常包括四个步骤：代价计算、代价聚合、视差估计及视差细化[4-5]，但其无法在反射、无纹理、昏暗及物体细小的情况下优化视差图的边缘结构以生成精确的视差图。

近些年来，随着深度学习和神经网络的发展，传统的立体匹配方法逐渐发展为端到端的深度学习立体匹配方法。早期研究者如 Žbontar 等[6]采用了孪生卷积神经网络来对代价计算进行改进，可以高效地形成代价空间，但是在代价聚合与视差细化中仍采取了传统的方法。直到 Zhang 等[7]提出的引导聚合网络(guided aggregation network，GANet)，才使传统的半全局立体匹配算法(semi-global matching，SGM)具有可微性，构建了端到端的立体匹配神经网络，才最终实现了端到端的深度学习立体匹配网络。Guo 等[8]通过增加对低分辨率和不连续深度的超分辨率感知，提出了可学习的高效立体匹配网络(efficient stereo matching network，ESMNet)。Zhang 等[9]提出的领域不变的立体匹配网络(domain-invariant stereo matching networks，DSMNet)包含两个可以提高网络鲁棒性的网络层。为了克服当前无法从真实世界中大量获取立体匹配数据集的缺点，Mayer 等[10]提出了全新的大型虚拟合成数据集场景流，可以有效地提升微调时网络的稳定性。Zou 等[11]提出了一种通过对视频序列中的场景深度及场景切换的差异进行视差估计的无监督的网络，以获取更稳定的视差结果。现阶段，大部分研究都采用不同的特征提取方法来获取精确的代价空间以进行代价聚合。Duggal 等[12]采用可微块匹配(patchmatch)[13]对局部有效区域进行代价计算来提升视差图的生成速度。Guo 等[14]提出了分组相关的立体匹配网络(group-wise correlation stereo network，GWC-Net)，通过将左右特征图分成多个相关的特征组进行相互映射，再打包各组之间的关系图进行代价聚合。自从 Kendall 等[15]提出几何上下文信息网络(geometry and context network，GC-Net)后，人们在改进代价聚合三维卷积的过程中提出了更多的方法。Liu 等[16]提出了新的代价计算方法并采用了自适应的形状引导滤波器进行代价聚合以降低大面积无纹理区域的匹配错误率。Chang 等[17]提出的金字塔立体匹配网络(pyramid stereo matching network，PSMNet)采用了空间金字塔池化(spatial pyramid pooling，SPP)[18]进行特征提取，并且在代价聚合过程中采用三维的堆叠沙漏，有效地提高了生成视差图的整体精度。但是，沙漏堆叠网络中多次上采样与下采样的操作会造成代价空间中包含的原图像轮廓信息丢失。

在上述研究的基础上，我们提出了一种多尺度引导聚合网络(multi-scale guided aggregation network，MSGANet)，在特征提取阶段采用了 4 个大小不同的空间金字塔池化层用以消除卷积层固定大小的约束，形成代价空间，在代价聚合过程中加入了半全局聚合层，可以在代价空间的四个方向上聚合最佳代价，以减少三维卷积层在上下采样过程中带来的三维信息的损失。为了进一步提升网络的性能，还采用了 L_2 损失函数进行预训练，获取更加收敛的初始值参数进行微调，以提升视差图的精度。

1　数据集与试验设备

1.1　数据集

本研究使用了卡尔斯鲁厄和丰田技术研究所(Karlsruhe Institute of Technology and Toyota Tech-

nological Institute,KITTI)提供的数据集来训练 MSGANet。大型合成数据集场景流广泛用于对真实数据集进行微调之前,可以有效地解决因数据集过小而导致的泛化性差问题。场景流数据集由开源的三维建模软件 Blender 获取,它采用单个双目虚拟成像传感器以左右两个视角拍摄图像来获取动态场景的深度信息。该数据集采用的虚拟传感器尺寸为高 32.0 mm、宽 18.0 mm,以 35.0 mm 的焦距获取分辨率为 540×960 像素的图像。数据集还区分为纯净版本和最终版本,纯净版本包含发光和阴影的特殊场景,最终版本包含动态模糊和散焦模糊的场景。数据集共包含 40 024 对视图,由 26 760 对包含悬浮物体场景的 FlyingTing3D,8 864 对包含静态立体卡通猴子场景的 Monkaa,以及 4 400 对包含合成立体交通场景的 Driving 3 个子数据集组成。KITTI 数据集是一个包含真实世界道路场景的立体匹配数据集。该数据集由安装在车辆中的 2 只彩色摄像机和 2 只灰度摄像机 PointGray Flea2(频率为 10 Hz,分辨率为 1 392×512 像素,开度角为 90°×35°)来获取道路场景的立体视频,并采用三维激光传感器 Velodyne HDL(频率为 10 Hz,64 个激光束,范围为 100 m)来获取稀疏度为 50%的真实光流与视差值。KITTI 2012 数据集于 2012 年提出,在 2015 年扩展成为 KITTI 2015,它采用不同的评价指标对特定的测试集进行在线视差图评估。与其他数据集相比,KITTI 数据集中包含乡村、城市及高速公路等真实场景,数据集涵盖了遮挡、不连续、反射、重复纹理和无纹理等场景,以供各种立体匹配网络进行全面评估。

1.2 设 备

本试验使用的测试平台设置如下:处理器为 Intel(R) Core(TM) i7-9700,处理器主频为 3.0 GHz,内存为 32 GB,显卡为 11 GB Nvidia GTX 2080Ti,操作系统为 Win10;相关的支持软件有 Anaconda、Python3.6、Cuda10.0、Pytorch1.3.0 和 GCC5.3 等。训练时,所有的模型都采用了 Adam 优化器,以 $\beta_1=0.9$,$\beta_2=0.999$ 的参数进行优化,并且首先在场景流数据集中进行 10 个周期的预训练,然后在 KITTI 数据集中预训练 300 个周期和微调 50 个周期。考虑到 Nvidia GTX 2080Ti 仅有 11 GB 的显示内存,试验中仅将 batchsize 的参数设置为 2,这样在损失值平稳下降的过程中可以保证具有足够的显示内存进行训练。

2 立体匹配网络

2.1 多尺度引导聚合网络结构

多尺度引导聚合网络(MSGANet)由金字塔池化网络和沙漏堆叠网络组成。输入的样本图像首先经过含有多层卷积和残差模块的二维特征提取网络,获取更大的感受野后采用 4 个不同尺寸的卷积核进行空间金字塔池化以消除卷积神经网络对尺寸的约束。对提取特征后的左右视图进行差异计算以生成用于代价聚合的四维代价空间。MSGANet 用权重相乘的方式来替代权重累加,使半全局立体匹能加入整个神经网络进行回归训练。MSGANet 在代价聚合的过程中在卷积与反卷积步骤中掺杂了多个 SGA 层,SGA 层在 4 个方向上对代价进行评估以优化生成视差图的边缘部分,提高计算效率。虽然代价聚合过程中 SGA 能更准确地定位场景中物体的边缘,但由于网络中多次上下采样,不可避免地造成部分信息的损失。为了进一步解决这个问题,我们采用了具有 3 个引导滤波器的 LGA 层,将滤波数组与代价空间进行加权求和来重新定义最后获得的视差图边缘。在整个网络中,原本需要用户自定义的超参数都被自适应权重所替代,这些参数由二维引导子网络与整个立体匹配网络一起训练生成;子网络中以原本的 RGB 图像作为输入,训练过程中对每个需要的参数进行归一化并将其重塑为 SGA 层与 LGA 层需要的权重再输入下一个训练周期;最后,采用 softmax 得到概率乘上对应的视差以获得最后的真实视差,生成精确的视差图。进行回归训练时,对预训练与微调过程都采用了平滑的 L_1 损失函数,来评估生成视差与真实视差之间的差距,持续进行迭代,直至获取最准确的视差图。MSGANet 结构如图 1 所示。

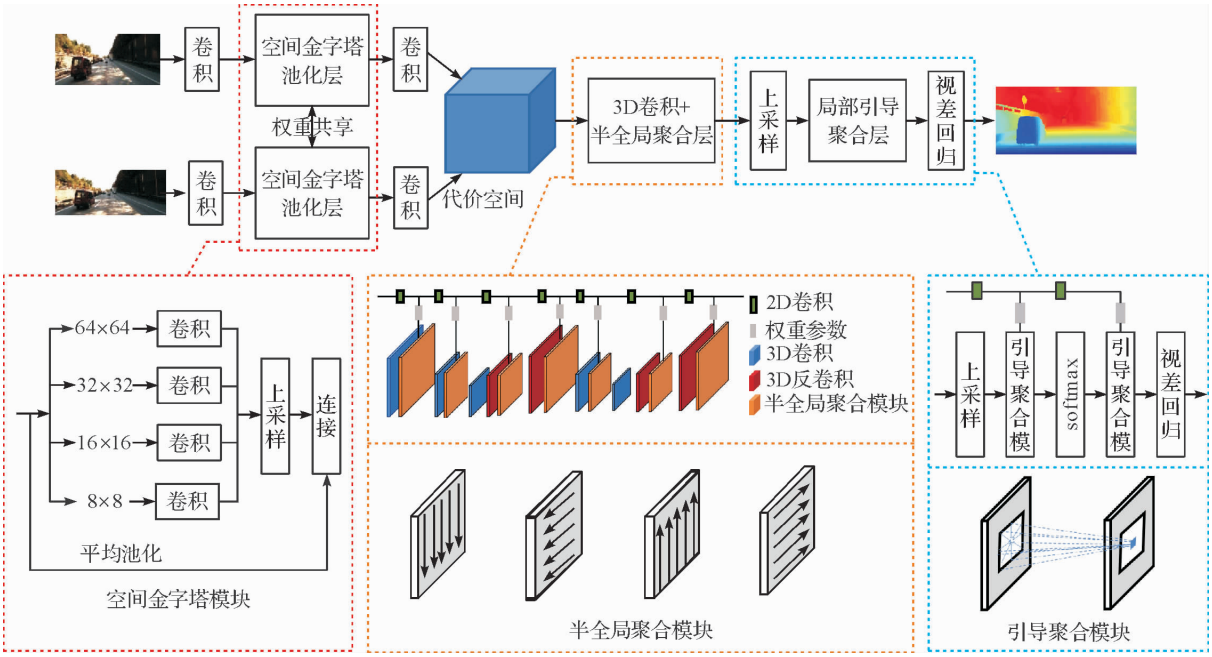

图 1　多尺度代价聚合网络的结构

Fig. 1　Architecture overview of MSGANet

2.2　多尺度引导聚合网络损失策略的改进措施

为了提高 MSGANet 在微调时获取模型的效果，本研究提出了一种改进的损失函数计算策略。现阶段的立体匹配方法，在整个训练过程中采用平滑的 L_1 损失函数对预测值与真实值进行差距评估。由图 2 可知，在预训练过程中预测值与真实值之间的差距越大所对应的 L_2 损失函数的梯度比平滑的 L_1 损失函数就越大，因此 L_2 损失函数具有更快的收敛性。在该过程中 L_2 损失函数采用平方项增大了预测值与真实值之间的差距，容易造成梯度爆炸。但是考虑到立体匹配预训练的任务的目的是获取足够收敛的初始模型参数，并且所采用的大型场景流数据集中的大量样本可以提高训练中损失值的稳定性。结合以上原因，在预训练过程中所采用 L_2 损失函数可以获取更快的收敛速度，并没有梯度爆炸的风险。在图 2 中红圈所示的微调过程中，平滑的 L_1 损失函数相比于 L_2 损失函数具有更小的梯度，能更好地克服微调时离群值对损失值的影响。因此本试验在预训练时采用 L_2 损失函数，在微调过程时再改为平滑的 L_1 损失函数进行计算。改进的损失计算策略如式(1)所示。

$$L = \lambda L_2(d_i, \hat{d_i}) + (1-\lambda)\text{smooth } L_1(d_i, \hat{d_i})。\tag{1}$$

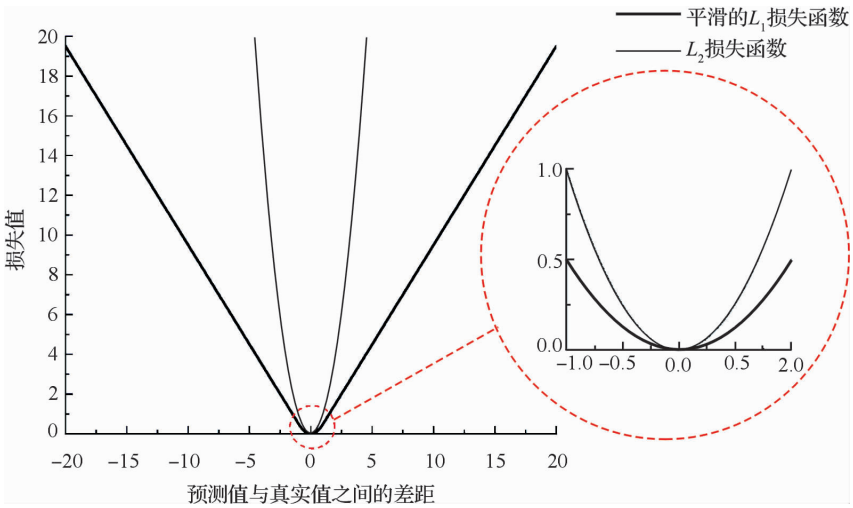

图 2　平滑的 L_1 损失函数与 L_2 损失函数示意图

Fig. 2　Diagram of smooth L_1 loss and L_2 loss function

式(1)中：L_2 为损失计算函数；smooth L_1 为平滑的 L_1 损失计算函数；d_i 为真实的视差；$\hat{d}_i$ 为生成的视差；λ 为超参数用于控制损失函数的改变以匹配当前进行中的任务，在场景流数据集上进行预训练时，λ 的值设置为 1，而在 KITTI 数据集上进行微调时将 λ 的值设置为 0。

2.3 评价指标

MSGANet 模型在 KITTI 数据集上采用如式(2)所示的端点误差(endpoint-error，EPE)进行性能评估，以获取整体的像素平均错误率。

$$e_{\mathrm{PE}}=\frac{1}{N}\sum_{i=1}^{N}(\|D_i-\hat{D}_i\|\geqslant t)。\tag{2}$$

式(2)中：e_{PE}为端点误差；D_i 为真实视差图；$\hat{D}_i$ 为生成的视差图；N 为需要评估的整张图像中所有的像素值；t 为划分的阈值，当得到的 e_{PE}值大于 t 像素即认为该像素生成的位置是错误的。该公式通过计算生成的视差值与真实视差值之间的平均欧几里得距离获取视差的准确表达式，并且还可以在多个像素阈值之间进行评估。在 KITTI 2012 数据集中分别对 EPE 值大于 2 像素、3 像素和 5 像素的未遮挡区域和所有区域进行误差评估。在 KITTI 2015 数据集中分别对 EPE 值大于 3 像素的未遮挡区域与所有区域的前景与背景进行评估。

3 试验结果及讨论

3.1 MSGANet 训练模型的评估结果

为了证明我们的改进是有效的，将我们提出的网络与同样采用了空间金字塔池化模块但没有采用 SGA 与 LGA 的 PSMNet 网络进行对比试验。两个对比的网络在场景流数据集上进行预训练时，图像会随机剪裁为高 240 像素和宽 512 像素的尺寸以达到图像增强的目的，视差搜索的最大值 $D_{\max}$设置为 192 像素。KITTI 数据集仅具有 200 对训练集与 200 对测试集，如果直接用于训练容易出现过拟合现象导致模型的泛化性差。因此，我们将大型数据集场景流划分为 35 454 对训练集、1 000 对验证集和 3 570 对测试集进行 0.01 的学习率的 10 个周期预训练。KITTI 数据集我们采用 0.01 的学习率训练了 300 个周期再以 0.001 的学习率训练 50 个周期，以确保训练后的模型可以获取最小的损失值。将 PSMNet 与 MSGANet 生成的视差图结果上传至 KITTI 官方网站进行 KITTI 2012 与 KITTI 2015 指标的评估对比，结果见表 1。由表可知，在代价聚合中增加了半全局引导聚合层与局部引导聚合层的 MSGANet 的精度后，要比原本采用三维堆叠沙漏卷积的 PSMNet 在 KITTI2012 以 3 像素为阈值的所有测评结果提高 0.18%，在 KITTI2015 的所有区域整体评估中增加了 0.32%的准确率。这表明半全局引导聚合层与局部引导聚合层通过考虑当前像素周围的代价与最后的滤波操作，解决了三维堆叠沙漏网络上下采样带来的精度损失问题。与表 1 对应的可视化对比结果如图 3 所示，在可视化对比的 3 组中，挑选出了场景中包含细小物体、无纹理区域及背光区域的场景做对比。由图 3(b)可知，PSMNet 无法保证在这些场景中物体的完整性，产生了大量的误匹配点；由图 3(c)可知，MSGANet 由于在训练过程中充分考虑了全局信息，因此可以克服场景中包含物体不完整的缺点。

表 1 PSMNet 和 MSGANet 在 KITTI 2012 与 KITTI 2015 上的测评结果

Table 1 Evaluation results of PSMNet and MSGANet on KITTI 2012 and KITTI 2015

网络结构	在 KITTI 2012 上的测评结果							
	≥2 像素百分比/%		≥3 像素百分比/%		≥5 像素百分比/%		平均误差像素/pixel	
	无遮挡	所有	无遮挡	所有	无遮挡	所有	无遮挡	所有
PSMNet[17]	2.41	3.00	1.47	1.88	0.89	1.15	0.5	0.5
MSGANet	1.82	2.54	1.17	1.70	0.75	1.11	0.4	0.5

网络结构	在 KITTI 2015 上的测评结果					
	所有区域评估的百分比/%			无遮挡区域评估的百分比/%		
	背景	前景	整体	背景	前景	整体
PSMNet[17]	1.93	4.84	2.43	1.75	4.57	2.23
MSGANet	1.80	3.70	2.11	1.59	3.26	1.87

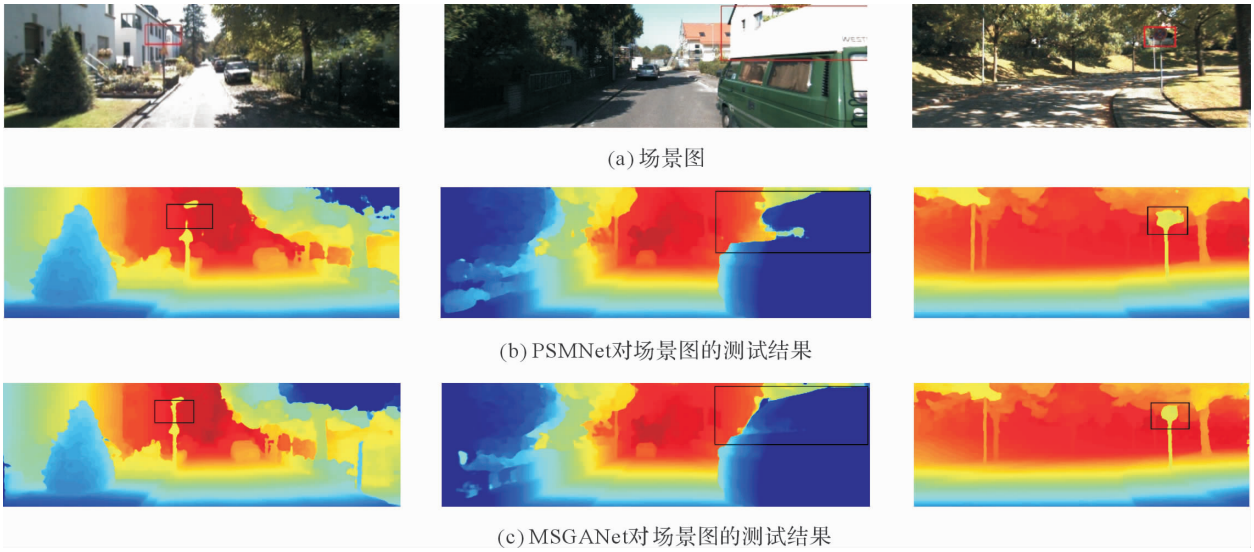

图 3　KITTI 数据集采用 PSMNet 与 MSGANet 获取的可视化结果

Fig. 3　Visual comparison of PSMNet and MSGANet on KITTI dataset

3.2　改进损失函数后的结果

本试验将改进损失函数的多尺度引导聚合网络(multi-scale guided aggregation network changed，MSGANet-C)与 MC-CNN-art、ESMNet、GC-Net、DFNet、DSMNet、DispletC、PSMNet、Deeppruner 及 GWC-Net 等网络进行了比较。图 4 为 MSGANet-C 与在 PSMNet、Deeppruner 及 GWC-Net 的可视化对比。从图4中的第1列中具有反射区域场景的对比结果中可以看出，大部分深度学习立体匹配依然会

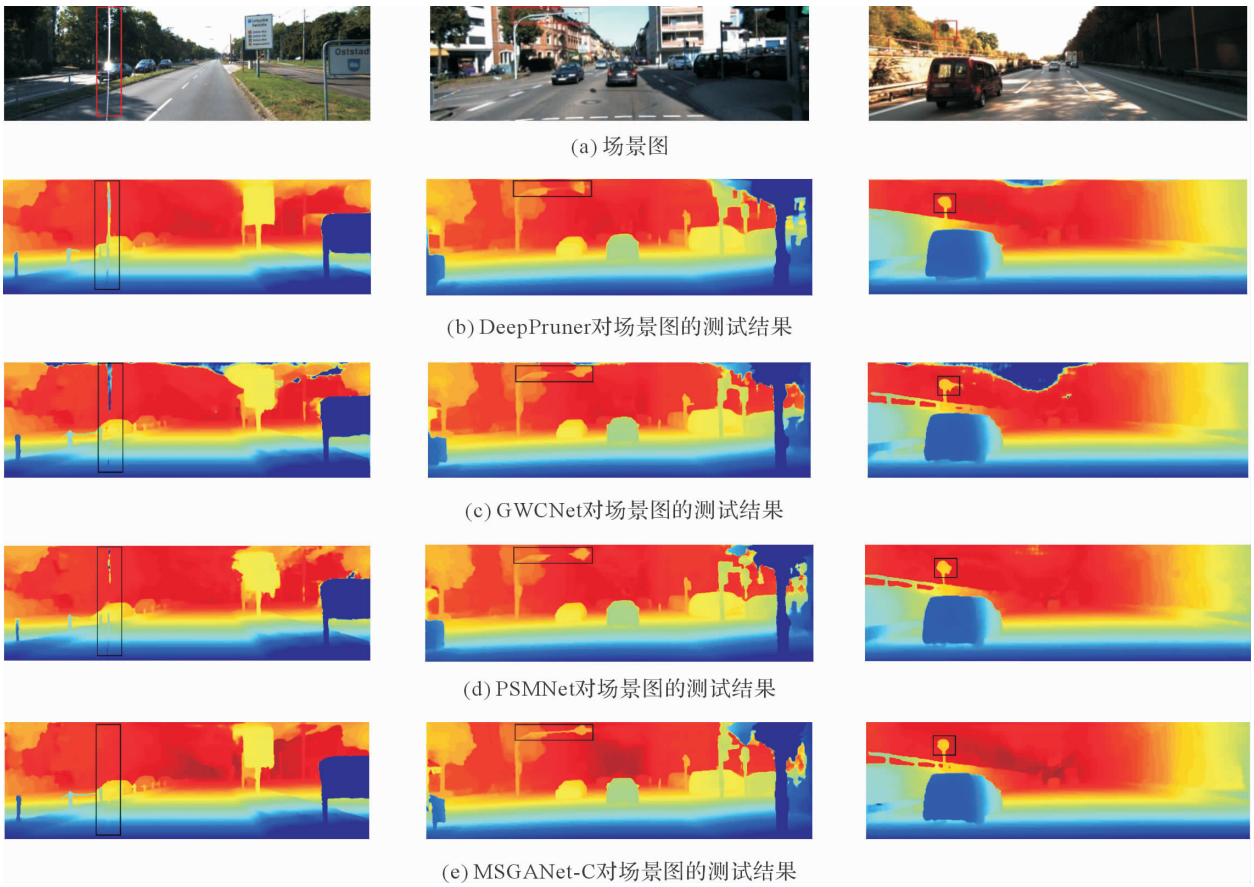

图 4　4 种网络结构在 KITTI 数据集上的可视化对比结果

Fig. 4　Visual comparison of four networks of KITTI dataset

产生很多的误匹配点，这是由于在反光区域中没有明显的特征可以用来提取，因此在代价聚合期间会导致视差图在这部分产生很多的噪声点。为了解决这类问题，采用多尺度空间金字塔池化层，能有效利用不同尺度的上下文信息使获取的特征图更具鲁棒性。在图 4 的第二列包含细小物体和第三列具有背光物体的场景中，由于很难将前景物体与背景物体区分开，很容易导致生成的视差不连续，因此在特征提取之后，我们采用的 SGA 在代价聚合过程中考虑了多个方向上的代价，可以有效地抑制离群值在其他方向对代价计算的干扰以获取更为清晰的视差图中物体的轮廓。

表 2 为 MSGANet-C 与 9 种网络对 KITTI 2012 与 KITTI 2015 的评估结果。MSGANet-C 在以 3 像素为阈值的 KITTI 2012 数据集与所有区域进行整体评估的 KITTI 2015 数据集中，比首次采用三维卷积的端到端网络 GC-Net 精度要高 0.61%与 0.82%。并且在 KITTI 2015 的评估结果中，MSGANet-C 对背景中的评估结果与近些年来提出的 DFNet、DSMNet、GWC-Net 的结果相似，但是对前景的深度估计比这 3 种网络有更大的优势，因此整体评估精度分别比这 3 种网络高 0.10%、0.09%与 0.16%。这是因为采用了半全局聚合层考虑了多个方向的代价，可以更好地区分前景的目标边缘，从而有效地改善前景的精确度。

表 2 MSGANet-C 与 9 种网络对 KITTI 2012 与 KITTI 2015 的评估结果

Table 2 MSGANet-C evaluation results with 9 types of networks on KITTI 2012 and KITTI 2015

网络结构	在 KITTI 2012 上的测评结果							
	≥2 像素百分比/%		≥3 像素百分比/%		≥5 像素百分比/%		平均误差/pixel	
	无遮挡	所有	无遮挡	所有	无遮挡	所有	无遮挡	所有
MC-CNN-art[6]	3.90	5.45	2.43	3.63	1.64	2.39	0.7	0.9
ESMNet[8]	3.65	4.30	2.08	2.53	1.11	1.41	0.6	0.7
GC-Net[15]	2.71	3.46	1.77	2.30	1.12	1.46	0.6	0.7
DFNet[11]	2.31	2.91	3.23	4.12	0.85	1.10	0.5	0.5
DSMNet[9]	2.25	2.83	1.39	1.79	0.83	1.07	0.5	0.5
DispNetC[10]	7.38	8.11	4.11	4.65	2.05	2.39	0.9	0.7
PSMNet[17]	2.41	3.00	1.47	1.88	0.89	1.15	0.5	0.5
Deeppruner[12]	2.41	2.94	1.49	1.86	1.11	1.38	0.5	0.5
GWC-Net[14]	2.20	2.79	1.40	1.83	0.87	1.14	0.5	0.5
MSGANet-C	1.82	2.54	1.16	1.69	0.75	1.10	0.4	0.5

网络结构	在 KITTI 2015 上的测评结果					
	所有区域评估的百分比/%			无遮挡区域评估的百分比/%		
	背景	前景	整体	背景	前景	整体
MC-CNN-art[6]	2.89	8.88	3.89	2.48	7.64	3.33
ESMNet[8]	2.57	4.86	2.95	2.41	4.30	2.72
GC-Net[15]	2.21	6.16	2.87	2.02	5.58	2.61
DFNet[11]	1.78	4.03	2.15	1.62	3.65	1.95
DSMNet[9]	1.78	3.97	2.14	1.64	3.53	1.95
DispNetC[10]	4.32	4.41	4.34	4.11	3.72	4.05
PSMNet[17]	1.93	4.84	2.43	1.75	4.57	2.23
Deeppruner[12]	2.10	3.68	2.36	1.95	3.33	2.18
GWC-Net[14]	1.73	4.60	2.21	1.59	4.12	2.01
MSGANet-C	1.74	3.65	2.05	1.54	3.26	1.82

4 结 语

为了立体匹配能在双目成像领域获取更好的视差效果，我们提出了一种视差生成网络 MSGANet。该网络采用多尺度空间“金字塔”池化进行特征提取，生成四维代价空间，并用半全局聚合层(SGA)和局部引导聚合(LGA)进行代价聚合。为了证实我们的改进是有效的，我们在大型合成数据集场景流数据集上进行预训练，并在 KITTI 2012 和 KITTI 2015 上进行微调。试验结果表明，MSGANet-C 将 KITTI 2012

的准确性提升至 98.31%，KITTI 2015 准确性提升至 97.95%。在未来的研究工作中，我们将侧重于提升网络的运行速度，在检测场景中车辆的同时获取物体与车辆间的距离，以适应复杂的道路状况。

参考文献：

[1] ZENG J, XUE W, XU B, et al. Research on robot positioning and grasping technology based on binocular vision[J]. Modular Machine Tool & Automatic Manufacturing Technique, 2019(1): 131.

[2] YANG L, LI M, SONG X, et al. Vehicle speed measurement based on binocular stereovision system[J]. IEEE Access, 2019(7): 106628.

[3] LAI X, XU X, ZHANG J, et al. An efficient implementation of a census-based stereo matching and its applications in medical imaging[J]. Journal of Medical Imaging and Health Informatics, 2019, 9(6): 1152.

[4] HIRSCHMULLER H. Stereo processing by semiglobal matching and mutual information[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2007, 30(2): 328.

[5] GONG M, YANG R, WANG L, et al. A performance study on different cost aggregation approaches used in real-time stereo matching[J]. International Journal of Computer Vision, 2007, 75(2): 283.

[6] 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.

[7] ZHANG F, PRISACARIU V, YANG R, et al. Ga-net: guided aggregation net for end-to-end stereo matching[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2019: 185.

[8] GUO C, CHEN D, HUANG Z. Learning efficient stereo matching network with depth discontinuity aware super-resolution[J]. IEEE Access, 2019(7): 159712.

[9] ZHANG F, QI X, YANG R, et al. Domain-invariant stereo matching networks[C]//Proceedings of the European Conference on Computer Vision. Cham: Springer, 2020: 420.

[10] MAYER N, ILG E, HAUSSER P, et al. A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas: IEEE, 2016: 4040.

[11] ZOU Y, LUO Z, HUANG J B. Df-net: unsupervised joint learning of depth and flow using cross-task consistency [C]//European Conference on Computer Vision. Cham: Springer, 2018: 38.

[12] DUAGGAL S, WANG S, MA W C, et al. Deeppruner: learning efficient stereo matching via differentiable patchmatch[C]//Proceedings of the IEEE International Conference on Computer Vision. Seoul: IEEE, 2019: 4384.

[13] BARNES C, SHECHTMAN E, FINKELSTEIN A, et al. Patchmatch: a randomized correspondence algorithm for structural image editing[J]. ACM Transactions on Graphics, 2009, 28(3): 1.

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

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

[16] LIU H, WANG R, XIA Y, et al. Improved cost computation and adaptive shape guided filter for local stereo matching of low texture stereo images[J]. Applied Sciences, 2020, 10(5): 1869.

[17] CHANG J R, CHEN Y S. Pyramid stereo matching network[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018: 5410.

[18] HE K, ZHANG X, REN S, et al. Spatial pyramid pooling in deep convolutional networks for visual recognition[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2015, 37(9): 1904.