

引用格式:郑秋梅,温阳,王风华. 基于注意力机制和可分离卷积的双目立体匹配算法[J]. 微电子学与计算机,2021,38(5):42-47. [ZHENG Q M, WEN Y, WANG F H. Stereo matching based on attention mechanism and separable convolution [J]. Microelectronics & Computer,2021,38(5):42-47.]

基于注意力机制和可分离卷积的双目立体匹配算法

郑秋梅,温 阳,王风华

(中国石油大学(华东) 计算机科学与技术学院,山东 青岛 266580)

摘 要: 针对当前基于卷积神经网络的双目立体匹配算法需要较高的特征提取能力且网络的参数量过多的问题,提出一种基于注意力机制的立体匹配网络,在特征提取阶段采用改进后的通道注意力机制根据通道内所含的信息进行特征加权,同时采用改进的空间金字塔结构实现多尺度特征提取,以提高网络的特征提取能力;设计 3D 注意力模块和 3D 可分离卷积进行视差计算,相比于标准卷积不仅可以降低网络的计算参数同时增加了通道维数可以保证匹配精度.最后,在 Scene Flow 数据集、KITTI 2012 数据集和 KITTI 2015 数据集上进行评估,实验表明,本文的网络模型在保证匹配精度的同时有效减少网络的计算参数.

关键词: 双目视觉;立体匹配;注意力机制;可分离卷积

中图分类号: TP391.4

文献标识码: A

文章编号: 1000—7180(2021)05—0042—06

Stereo matching based on attention mechanism and separable convolution

ZHENG Qiu-mei, WEN Yang, WANG Feng-hua

(Department of Computer and Communication Engineering, China University of Petroleum, Qingdao 266580, China)

Abstract: To improve the feature extraction capability of current stereo matching network and reduce the parameter, the multi-scale context attention network for stereo matching network is proposed. In the feature extraction stage, the improved channel-wise attention mechanism is used to weight the features based on the information contained in the channel, and the improved spatial pyramid structure is used to achieve multi-scale feature extraction to improve the network's feature extraction ability. A three-dimensional attention module and a three-dimensional separable convolution for disparity calculation. Compared with the standard convolution, the computational parameters of the network can be reduced and the channel dimension can be increased to ensure the matching accuracy. Experiments on Scene Flow datasets, KITTI 2012 datasets and KITTI 2015 datasets show that the network model in this paper can effectively reduce the calculation parameters of the network while ensuring matching accuracy.

Key words: binocular vision; stereo matching; attention; separable convolution

1 引言

立体匹配作为计算机视觉领域的研究热点,在三维重建,自动驾驶以及移动机器人等方面有着广

泛的应用^[1-2]. 目前,双目立体匹配算法主要分为两类:传统的立体匹配算法和基于深度学习的立体匹配算法. 其中,传统算法主要采用手工设计特征的方式进行立体匹配,因此很难使用有限的特征获得足

够精确的视差结果. 虽然基于卷积神经网络的算法相比于传统算法在精度上得到一定的提高^[3-8]. 但是,提高网络的特征提取能力获得更加丰富的信息,实现在病态区域(如阴影区域,图像中较细的物体等)得到精确的视差图仍有一定的困难. 目前,多数基于卷积神经网络的算法^[5-8]为了保证视差计算的精度常采用 3D 卷积计算匹配代价,导致网络具有大量的计算参数.

本文通过考虑不同通道上特征的信息丰富程度不同以及上下文信息对于提高误匹配区域匹配精度的影响,采用改进后的通道注意力机制和空间金字塔结构构造特征提取单元;考虑 3D 的匹配代价可以得到更加精确的视差图,采用 3D 注意力机制和可分离卷积构造视差计算模块在保证匹配精度的同时降低网络参数.

2 相关理论

2.1 双目立体匹配

立体匹配的实质是将位于同一水平线上不同位置的双目摄像机所获得的图像作为输入,通过双目立体匹配算法为图像中的每个像素计算视差,最后获得图像中的深度信息. 因此,立体匹配算法的关键是如何获得精确的视差图.

通常,基于卷积神经网络的双目立体匹配算法的总体框架:首先,将原始图像对输入两个共享权重的网络中进行特征提取,之后利用提取到的特征构造匹配代价进行视差计算与优化得到最后的视差图. 由此可见,网络的特征提取能力影响后续匹配代价计算进而影响最终的视差结果.

2.2 金字塔立体匹配算法(PSM-Net)

金字塔立体匹配算法在特征提取阶段采用空洞卷积的方式扩大感受野,通过空间金字塔池化实现信息聚合,之后通过堆叠的 3D 沙漏结构实现视差的计算优化. 虽然该算法提高了立体匹配的精度,但是在特征提取阶段未能充分考虑特征在通道维度的信息,因此导致在最终的视差结果中某些细节信息的丢失. 同时,由于编解码结构内部的链接以及采用 3D 卷积计算匹配代价导致网络的具有较多的计算参数.

2.3 注意力机制

注意力机制^[9]根据特征中包含的信息生成权重,含有信息越丰富的特征权重越大. 大多数研究都是利用空间注意力机制提高网络性能. Hu 等^[10]首

先引入了通道注意力机制来提高网络性能. 他们提出了 squeeze-exciation 块(SE block)计算特征的权重. 实验表明,通道注意力机制可以获得更多的细节信息. Li 等^[11]设计了 selective kernel block (SK block),通过在神经网络的每一层设置不同大小的卷积核,使神经网络根据对象的不同尺度调整感受野. 这种机制使得 SK block^[11] 具有较好的分类效果.

Song^[12]等使用注意力机制提取雾天下的图像特征有效的降低了误差率. 在立体匹配任务中,图像的前景区域中通常包含大量的细节信息,如何从图像中更加精确的保留这部分区域的视差成为研究热点. 因此,本文在 SK block^[11]基础上提出了通道特征提取结构.

2.4 可分离卷积

Howard 等^[13]设计了深度可分离卷积,将标准卷积分解为深度卷积和点卷积(1×1 卷积). 例如,如果输入特征图的大小为 $D_f \cdot D_f \cdot M$ (M 为输入通道数),输出特征图的大小为 $D_k \cdot D_k \cdot N$ (N 为输出通道数),则标准卷积的计算量为 $D_f \cdot D_f \cdot M \cdot N \cdot D_k \cdot D_k$. 对于深度可分离卷积,计算量为 $D_f \cdot D_f \cdot M \cdot D_k \cdot D_k + M \cdot N \cdot D_f \cdot D_f$. 与标准卷积相比,深度分离卷积通过解耦通道维度和空间维度可以将计算量减少到标准卷积的 $\frac{1}{N} + \frac{1}{K^2}$.

Liu^[14]等使用可分离卷积代替标准卷积处理遥感图像变化检测问题,可以有效地降低网络的计算参数. 受此启发,本文采用可分离卷积思想设计视差计算模块.

3 基于注意力机制和可分离卷积的立体匹配算法(MSDC-Net)

3.1 网络整体结构

网络采用端到端的方法进行设计:采用孪生网络作为特征提取模块的基础结构,利用三维注意力机制和三维可分离卷积获取几何信息,利用编解码结构计算视差图. 其中,特征提取模块由通道特征提取结构和多尺度信息提取模块组成. 整个网络结构如图 1 所示.

3.2 特征提取模块

本文采用通道注意力机制和残差思想构造通道特征提取结构. 同时,为了使通道注意力机制更加适合求解立体匹配问题,本文对通道注意力模块进行

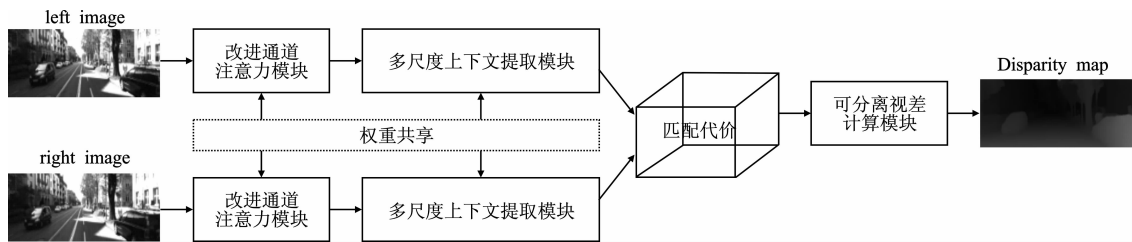


图 1 网络整体结构

了改进. SK block^[11]在全局平均池化模块之后采用全连接层压缩特征,之后利用注意力机制对压缩后的特征进行处理,生成相应的权重.对于立体匹配任务,首先使用全局平均池化得到 16×16 的特征,之后本文使用两个相连的卷积层(3×3)代替第一层全连接层,提高网络的学习能力.同时,由于 PRELU^[15]在负区进行的是线性运算,可以解决 ReLU 函数^[16]在负数区神经元死亡的问题,有利于信息的保留.因此本文采用 PRELU^[15]作为通道特征提取结构的激活函数.通道特征提取结构如图 2 所示.

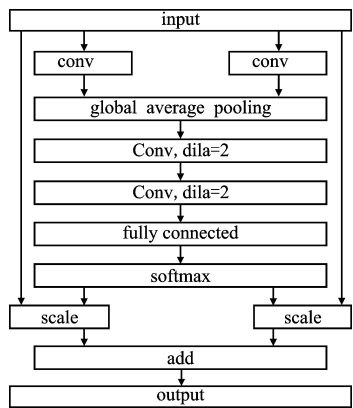


图 2 通道特征提取结构

要获得准确的视差图,不仅需要保留更多的细节信息,还需要增强提取上下文信息的能力.由于文献[7]论述了只通过图像中像素之间点对点的关系无法获得到足够多的上下文信息,因此仅仅通过图像中像素之间的关系来获得特征是不够的.因此,设计基于空洞空间金字塔^[17]的多尺度信息提取模块来提取上下文信息.多尺度特征提取模块如图 3 所示.

3.3 视差计算模块

编解码结构可以有效地从输入特征中提取上下文信息.为了保证视差图的准确性,本文使用三个堆叠的沙漏模块(编解码结构)来计算视差.在第 3.2 节中,本文设计的通道特征提取结构在立体匹配中具有较好的特征提取能力.在本节中,在每个沙漏模

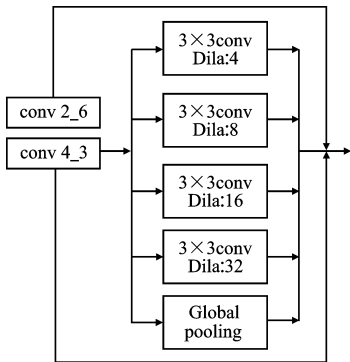


图 3 多尺度信息提取结构

块中添加了三维通道特征提取结构,以增加信息丰富的特征的权重.这可以提高有用信息的利用率,使网络能够保留更多的细节信息.

随着维数和网络层数的增加,导致视差计算模块产生大量的参数.这增加了网络的计算时间.为了减少网络参数同时节省计算时间,本文设计 3D 深度可分离卷积代替标准卷积构造视差计算模块.可分离卷积不仅可以降低网络的计算参数,由于其同时采用不同大小的卷积核进行计算,可以使网络获得更多的上下文信息有利于提高视差计算的精度.此外,本文网络还删除了不同沙漏模块之间的冗余连接以及沙漏模块中各层之间的冗余连接.视差计算模块如图 4 所示.

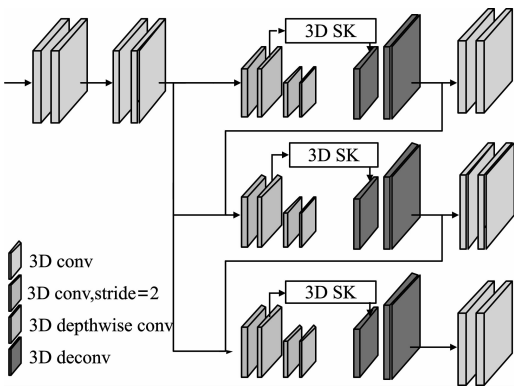


图 4 可分离视差计算结构

4 实验结果及分析

4.1 实验数据集介绍

立体匹配常用的公开数据集有:Scene Flow 数据集^[3]以及 KITTI 数据集^[18]. Scene Flow 数据集分为三个子数据集:Driving, Monkaa, Flying Thing 3D, 总共包含 35 454 张训练图像和 4 370 张测试图像, 并且提供了密集精确的视差真值. KITTI 数据集可以分为 KITTI 2012 数据集和 KITTI 2015 数据集, 都是由真实场景中的不同情境中的车辆图像构成. KITTI 2012 数据集包含 194 对训练图像和 195 对测试图像. KITTI 2015 数据集包含 200 对训练图像和 200 对测试图像, 提供了用 LiDAR 获得的视差图.

4.2 实验结果及分析

本文网络是在 PyTorch 框架上实现的. 所有的训练和测试实验都在 Tesla P100 上进行. 首先使用 Scene Flow 数据集对网络进行预训练, 然后在 KITTI 数据集上调整模型. 网络使用 Adam^[19] ($\beta_1=0.9, \beta_2=0.999$) 并将批处理大小设置为 4. 网络采用多级学习率训练模型: 先以 0.001 的学习率在 Scene Flow 数据集上获取预训练模型, 之后以 0.001 的学习速率在 KITTI 数据集上训练 300 epochs, 以 0.0001 的学习速率训练 150 epochs 对模型进行微调优化.

4.2.1 视差结果分析

(1) Scene Flow 数据集

本文首先在 Scene Flow 数据集上采用端点误差(EPE)对改进后的网络进行评价. EPE 表示检测到的视差图和真实值之间对应像素的欧式距离之和. 其在 Scene Flow 数据集上结果分别如表 1 所示.

表 1 Scene Flow 数据集上算法的性能评价					
算法	MSDC-Net(our)	PSMNet	CRL	DispNetC	GC-Net
EPE/%	0.98	1.09	1.32	1.68	2.51

从表 1 中可以看出, 本文算法相对于其他经典的匹配算法在计算端点误差时, 误差率较小.

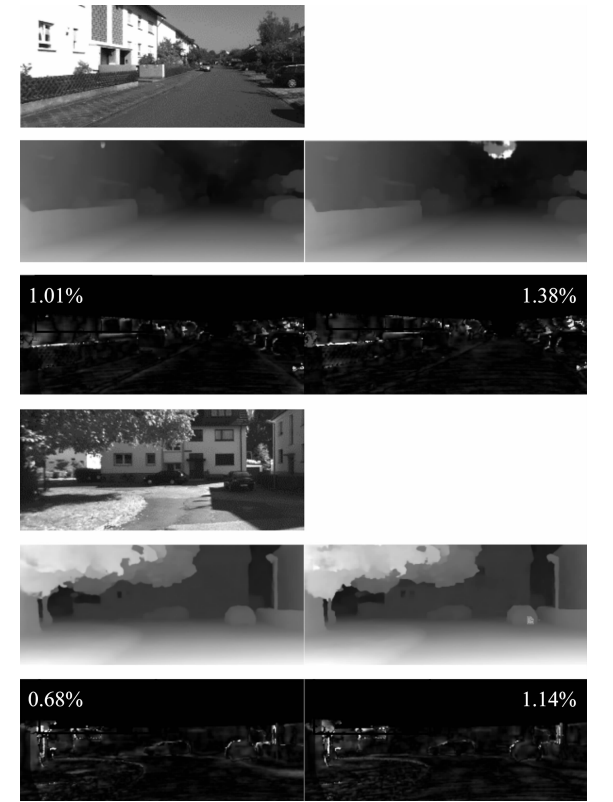
(b) KITTI 数据集

本文将得到的视差图提交到 KITTI 在线平台来评估网络的性能. 表 2 和表 3 展示了本文网络和几个典型网络在 KITTI2012 数据集和 KITTI2015 数据集上的实验结果.

表 2 本文算法与其他网络在 KITTI2012 数据集上的对比

Method	>2px/%		>3px/%	
	Out-Noc	Out-All	Out-Noc	Out-All
DispNetC	7.38	8.11	4.11	4.65
GCNet	2.71	3.46	1.77	2.30
WCLNet	3.57	4.35	2.19	2.74
MSDC-Net(ours)	2.51	3.11	1.52	1.94

“Out-Noc”表示图像中未遮挡区域的误差, “Out-All”表示所有像素的误差. “2px”表示误差数大于两个像素的百分比.



(a) MSDC-Net(ours) (b) PSMNet
图 5 本文算法与其他网络在 KITTI2012 数据集上的可视化对比

由表 2 可以看出, 本文的网络模型在 KITTI2012 数据集上可以获得较好的实验结果. 图 5 展示了本文网络和 PSMNet^[6] 在 KITTI2012 测试集中的可视化结果. 从图 5 可以看出, 本文算法在处理图像中的阴影区域时生成的视差图边缘更加清晰. 误差图中的数据表示的是考虑全部像素时 3px 误差率. 其中, 第 1, 4 行表示原始输入图像; 2, 5 行表示得到的视差图; 3, 6 行表示得到的误差图.

由表 3 可以看出, 本文方法在对前景区域(D1-fg)进行处理时, 可以获得更加准确的结果. 这证明了本文方法可以从给定的图像中获得更多的细节. 在处理背景区域时, 性能相对较弱. 这可能是由于删除了编解码

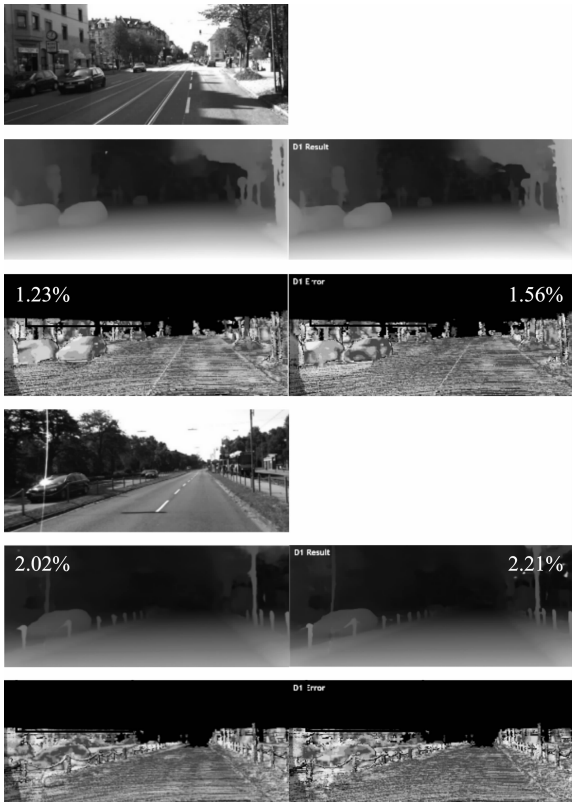
表 3 本文算法与其他网络在 KITTI2015 数据集上的对比

Method	All/%			Noc/%			time/s
	Dl-bg	Dl-fg	Dl-all	Dl-bg	Dl-fg	Dl-all	
GCNet	2.21	6.16	2.87	2.02	5.28	2.61	0.90
CRL	2.48	3.59	2.67	2.32	3.12	2.45	0.47
PSMNet	1.86	4.62	2.32	1.71	4.31	2.14	0.41
WCLNet	2.72	6.30	3.31	2.52	5.48	3.01	0.2
CPNet	3.61	7.14	4.20	3.41	6.59	3.94	\
MSDC-Net(ours)	1.99	4.27	2.37	1.83	4.04	2.20	0.37

表中参数“ALL”表示接下来的估计是基于所有像素的. “Noc”指的是未遮挡区域的像素. “Dl-bg”表示背景区域的误差. “Dl-fg”表示前景区域的错误. 参数“Dl-all”表示所有像素点的误差

模块中的不同沙漏模块之间的链接造成的.

图 6 展示了本文方法与 PSMNet^[6]可视化结果之间的比较. 从图 6 可以看出, 本文方法在处理弱纹理区域(如反射区域)时取得了很好的效果. 本文方法能够提取出更精确的视差图, 特别是在图像中的前景区域. 误差图中的数据表示的是 Dl-all 误差率. 其中, 第 1, 4 行表示原始输入图像; 2, 5 行表示得到的视差图; 3, 6 行表示得到的误差图.



(a)MSDC-Net(ours) (b)PSMNet
图 6 本文算法与其他网络在 KITTI2015 数据集上的可视化对比

4.2.2 模型分析

本文算法最终训练得到的模型体积如表 4 所示.

表 4 算法模型体积表

算法	PSM-Net	CRL	MSDC-Net (ours)
GPU Memory /G	4.4	10.4	3.2

由于本文所提出的网络在特征提取模块采用残差思想中的跳跃链接对特征提取模块进行优化设计以及采用深度可分离卷积构造匹配代价并进行视差计算因此本文的网络参数得到有效的降低.

综合视差计算精度和最终的模型体积可以看出, 本文所提算法均不仅提高了视差的计算精度, 同时降低了网络的计算参数.

5 结束语

本文提出了一种基于注意力机制和可分离卷积的端到端立体匹配网络. 通过设计通道特征提取结构和基于多尺度信息的特征提取模块, 实现了局部特征的提取. 然后利用三维注意力机制和三维混合可分离卷积构建视差计算模块, 不仅增加了网络的宽度提高了网络在病态区域的匹配精度, 同时降低了网络的计算参数. 由于在同一个网络层中采用不同尺寸的卷积核融合不同尺度的感受野, 因此接下来考虑采用在可分离卷积内部采用不同的卷积核进行计算从而进一步提高匹配精度.

参考文献:

[1] GEIGER A, LENZ P, URTASUN R. Are we ready for autonomous driving? The KITTI vision benchmark suite[C]//Proceedings of 2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Providence, RI, USA: IEEE, 2012: 3354-3361. DOI: 10.1109/CVPR.2012.6248074.

[2] CHEN X Z, KUNDU K, ZHU Y K, et al. 3D object proposals for accurate object class detection[J]. Proceedings of the 28th International Conference on Neu-

- ral Information Processing Systems, 2015 (1): 424-432.
- [3] 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 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Las Vegas: IEEE, 2016:4040-4048. DOI:10.1109/CVPR.2016.438.
- [4] PANG J H, SUN W X, REN JS J, et al. Cascade residual learning: a two-stage convolutional neural network for stereo matching[C]//2017 IEEE International Conference on Computer Vision Workshops (ICCVW). Venice, Italy: IEEE, 2017: 878-886. DOI:10.1109/ICCVW.2017.108.
- [5] 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). Venice, Italy: IEEE, 2017: 66-75. DOI: 10.1109/ICCV.2017.17.
- [6] CHANG J R, CHEN Y S. Pyramid stereo matching network[C]//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, UT: IEEE, 2018: 5410-5418. DOI: 10.1109/CVPR.2018.00567.
- [7] NGUYENT P, JEONJ W. Wide context learning network for stereo matching[J]. Signal Processing: Image Communication, 2019(78): 263-273. DOI:10.1016/j.image.2019.07.008.
- [8] KANG J H, CHEN L, DENG F, et al. Context pyramidal network for stereo matching regularized by disparity gradients[J]. ISPRS Journal of Photogrammetry and Remote Sensing, 2019, 157: 201-215. DOI: 10.1016/j.isprsjprs.2019.09.012.
- [9] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[C]//Proceedings of the 31st International Conference on Neural Information Processing Systems. Red Hook: Curran Associates Inc., 2017: 6000-6010.
- [10] HU J, SHEN L, ALBANIE S, et al. Squeeze-and-excitation networks[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2020, 42(8): 2011-2023. DOI:10.1109/TPAMI.2019.2913372.
- [11] LI X, WANG W H, HU X L, et al. Selective kernel networks[C]//2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Long Beach: IEEE, 2019: 510-519. DOI: 10.1109/CVPR.2019.00060.
- [12] SONG T, KIM Y, OH C, et al. Simultaneous deep stereo matching and dehazing with feature attention [J]. International Journal of Computer Vision, 2020, 128(4): 799-817. DOI:10.1007/s11263-020-01294-2.
- [13] HOWARD A G, ZHU M L, CHEN B, et al. MobileNets: efficient convolutional neural networks for mobile vision applications[C]//Proceedings of Computer Vision and Pattern Recognition. 2017.
- [14] LIU R C, JIANG D W, ZHANG LL, et al. Deep depthwise separable convolutional network for change detection in optical aerial images[J]. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 2020 (13): 1109-1118. DOI: 10.1109/JSTARS.2020.2974276.
- [15] HE K M, ZHANG X Y, REN S Q, et al. Delving deep into rectifiers: surpassing human-level performance on ImageNet classification[C]//Proceedings of the 2015 IEEE International Conference on Computer Vision. Santiago: IEEE, 2015: 1026-1034. DOI: 10.1109/ICCV.2015.123.
- [16] GLOROT X, BORDES A, BENGIO Y. Deep sparse rectifier neural networks[C]//Proceedings of the 14th International Conference on Artificial Intelligence and Statistics, 2010: 315-323.
- [17] CHEN L C, PAPANDREOU G, KOKKINOS I. DeepLab: semantic image segmentation with deep convolutional nets, atrousconvolution, and fully connected-CRFs[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2018, 40(4): 834-848. DOI: 10.1109/TPAMI.2017.2699184.
- [18] MENZE M, GEIGER A. Object scene flow for autonomous vehicles[C]//2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Boston: IEEE, 2015: 3061-3070. DOI: 10.1109/CVPR.2015.7298925.
- [19] KINGMA D, BA J. Adam: a method for stochastic optimization[EB/OL]. (2014-12-22) [2019-10-15]. <https://arxiv.org/abs/1412.6980v1>.

作者简介:

郑秋梅 女, (1964-), 硕士, 教授. 研究方向为图形与图像处理.

温 阳(通讯作者) 女, (1996-), 硕士研究生. 研究方向为图形与图像处理电子. E-mail: 2453079161@qq.com.

王风华 男, (1979-), 博士, 讲师. 研究方向为图像处理与模式识别、嵌入式系统设计.