应用光学
Journal of Applied Optics
SCOPUS,等来源期刊
中文核心期刊、中国科技核心期刊

基于双重三池化注意力机制的PSMNet算法

刘腾飞 林冬云 兰维瑶 陈岳航

PSMNet algorithm based on dual three-pooling attention mechanism

LIU Tengfei, LIN Dongyun, LAN Weiyao, CHEN Yuehang

引用本文:
刘腾飞, 林冬云, 兰维瑶, 等. 基于双重三池化注意力机制的PSMNet算法[J]. 应用光学, 2025, 46(2): 327-335. DOI: 10.5768/JAO202546.0202005
LIU Tengfei, LIN Dongyun, LAN Weiyao, et al. PSMNet algorithm based on dual three-pooling attention mechanism[J]. Journal of Applied Optics, 2025, 46(2): 327-335. DOI: 10.5768/JAO202546.0202005

在线阅读 View online: https://doi.org/10.5768/JAO202546.0202005

文章编号:1002-2082 (2025) 02-0327-09
引用格式:刘腾飞, 林冬云, 兰维瑶, 等. 基于双重三池化注意力机制的 PSMNet 算法 [J]. 应用光学, 2025, 46(2): 327-335.
LIU Tengfei, LIN Dongyun, LAN Weiyao, et al. PSMNet algorithm based on dual three-pooling attention mechanism[J]. Journal of Applied Optics, 2025, 46(2): 327-335.

在线阅读

基于双重三池化注意力机制的 PSMNet 算法

刘腾飞,林冬云,兰维瑶,陈岳航

(厦门大学 航空航天学院,福建 厦门 361102)

摘　要:为解决小型纹理类物体的视差计算及三维重建问题,提出了基于双重三池化注意力机制的 PSMNet-ECSA 算法。通过在残差网络主干中嵌入通道和空间注意力两个维度,每个维度以平均、最大、混合池化的方式进行特征维度融合,在一定程度上防止了过拟合现象,从而增强网络信息提取能力和泛化能力。在实验环境和数据集一致的条件下,经过 SceneFlow、KITTI2015 和真实场景实验分析,相对比原始 PSMNet 算法,本文算法在平均绝对误差、阈值误差等指标取得了 10% 的提升;将该算法应用于鲍鱼重建三维点云模型,长、宽、呼吸孔等距离测量平均相对误差在 3% 以内,能够以自动化的方式测量并记录小型海洋类生物的生长情况,具有良好的实际应用价值。

关键词:视差图;立体匹配;堆叠卷积;注意力机制;三池化
中图分类号:TN201; TP391.4　　**文献标志码**:A　　**DOI**: 10.5768/JAO202546.0202005

PSMNet algorithm based on dual three-pooling attention mechanism

LIU Tengfei, LIN Dongyun, LAN Weiyao, CHEN Yuehang

(School of Aerospace Engineering, Xiamen University, Xiamen 361102, China)

Abstract: To address the disparity calculation and 3D reconstruction problems for small textured objects, a PSMNet-ECSA algorithm based on a dual three-pooling attention mechanism was proposed. By embedding two dimensions of channel and spatial attention mechanisms into the backbone of the residual network, each dimension was fused using average pooling, maximum pooling, and mixed pooling techniques to merge feature dimensions, which prevented overfitting to some extent, thereby enhancing the network information extraction and generalization capabilities. Under the conditions of consistent experimental environments and datasets, an analysis was conducted through experiments on SceneFlow, KITTI2015, and real-world scenes. Compared to the original PSMNet algorithm, the proposed algorithm achieves a 10% improvement in metrics such as mean absolute errors and threshold errors. Applying this algorithm to reconstruct three-dimensional point cloud models of abalone, the average relative errors in distance measurements such as length, width, and breathing holes are within 3%, which can automatically measure and record the growth status of small marine organisms, and has promising practical application values.

Key words: disparity map; stereo matching; stacked convolution; attention mechanism; three-pooling

引言

随着近几十年来立体匹配技术的不断发展,各种匹配算法精度的不断提高,应用场景也越来越广泛,例如三维重建[1]、无人驾驶[2]、人脸识别[3]、

收稿日期:2024-02-01;　**修回日期**:2024-06-10
基金项目:国家自然科学基金(62273285)
作者简介:刘腾飞(1997—),男,硕士,主要从事图像处理、立体匹配、三维重建研究。E-mail: ltf_autos@126.com
通信作者:林冬云(1971—),女,博士,副教授,主要从事图像处理、三维测量研究。E-mail: dylin@xmu.edu.cn

增强现实[4]等。为了解决局部传统立体匹配算法匹配的单像素局限性，HIRSCHMULLER H 等人[5]提出了一种半全局匹配算法（SGM），通过一维路径聚合方法替代全局算法中最小化二维方法，有效降低了时间成本。Patchmatch-CrossWindow[6]阐述了一种基于双纹理阈值的十字动态臂搜索结构，动态计算当前像素的聚合窗口，在低纹理区域取得了不错的匹配效果，同时提升了算法时间效率。经过 2D 与 3D 视觉的技术突破，基于深度学习的立体匹配网络引起了社会的广泛关注，层出不穷的网络结构与方法使得立体匹配的性能大幅提升，立体匹配网络主要分为两种类型，第一类方法是通过确定学习如何匹配输入图像像素的方式来模仿传统的立体匹配算法：ZAGORUYKO S 等人[7]等人通过使用上层的决策网络计算对应特征向量的相似度得分来生成像素代价体；MC-CNN[8]通过用 5 个全连接层组成的决策网络取代了传统特征相似度度量方法，提高了匹配精度，在保留分辨率且不增加计算时间的条件下，扩大了网络上下文的感受野；YE X 等人[9]在不同网络层采用不同窗口大小的多步池化，并通过连接各层输出生成特征图。第二类方法是使用端到端的可训练管道模型解决立体匹配问题：FlowNet-Simple[10]和 DispNetS[11]使用单个编码器-解码器网络，将左右图像堆叠形成 6D 特征体回归视差图；PSMNet[12]使用了空间金字塔池化结构来提取和聚合多尺度的特征图；YANG G 等人[13]通过保留特征维度形成 4D 代价特征体，使用顶层网络学习相似性度量；为了生成高分辨视差图，KENDALL A 等人[14]构建了一种分层 4D 特征体，并使用 3D 卷积从粗到细的方式进行处理；LEAStereo[15]通过神经架构搜索（NAS）优化 PSMNet 的架构，提出了一种高效且准确的立体匹配网络；HITNet[16]提出一种分层迭代瓷砖细化网络，通过逐步细化视差图来提高匹配精度。

根据人类对复杂场景快速、高效地提取重要信息的方式，注意力机制[17]应运而生，其基于输入图像特征动态地调整权重。近十年以来注意力机制在众多视觉任务中获得了巨大成功，例如图像生成[18]、语义分割[19]、位姿估计[20]、多模态任务[21]等。MNIH V 等人[22]开创性地同时引入卷积神经网络和空间注意力机制，使用策略梯度对网络进行端到端的循环预测和更新，有效降低了网络计算成本，同时提高了图像分类精度。JADERBERG M 等人[23]使用了一种子神经网络，用来预测和选择输入中重要区域的仿射变换，提供了具有平移不变性的深度网络的空间注意力机制（STN）。SENet[24]提出了一种新型的通道注意力网络，该网络隐式且自适应地预测了潜在的关键特征。

目前传统的立体匹配算法计算方式单一，时间效率偏低，从算法精度和实时性上均劣于基于深度学习的立体匹配网络。立体匹配网络大多针对大范围场景的研究，对于小型特定场景的计算精度有待提高，而注意力机制对于细节信息的保留与提取更具优势，因此本文基于通道和空间注意力机制模块提出 PSMNet-ECSA 算法，有效利用不同特征尺度和空间位置信息生成特征层，同时结合 3 种池化方式对各特征层进行多维度融合，在一定程度上降低了网络过拟合现象。通过 SceneFlow[11]、KITTI2015[25]和真实场景对比实验，可知本文算法在多项评价指标均有所提升，同时验证了该算法具有一定的实用价值。

1 PSMNet-ECSA 算法

1.1 网络结构

本文提出的 PSMNet-ECSA 整体网络流程如图 1 所示，网络结构和结构参数分别如图 2、表 1 所示。该网络主干结构为残差结构，包含 2D/3D 卷积、多池化、双线性插值上采样、空洞卷积和 3D 反卷积，其中 ECSA(efficient channel spatial attention) 为双重三池化注意力机制模块，具体结构在下一节说明。SPP(spatial pyramid pooling) 为空间金字塔池化模块[26]，使用 4 个卷积核（64×64、32×32、16×16、8×8）的自适应平均池化操作，分别产生 4 个尺度的特征图，后接1×1的一维卷积和双线性插值上采样，将 4 个尺度的特征级联生成最终的特征体。3DCNN（3D-convolutional neural networks）为 3D 堆叠卷积模块[12]，该模块包含了 12 个3×3×3的卷积层，涉及 1/4、1/8、1/16 这 3 个不同尺度的特征聚合，其中每个尺度输出一组视差图和损失值。本文整体网络结构如图 2 所示，将两种三池化通道注意力结构和空间注意力结构采用串联方式，然后与普通卷积特征图形成残差结构运算，输出特征图。空间金字塔池化模块进一步提取图像上下文信息，经过左右特征图级联计算 4D 代价体，然后采用 3D 堆叠卷积和视差回归获得最终的视差图。

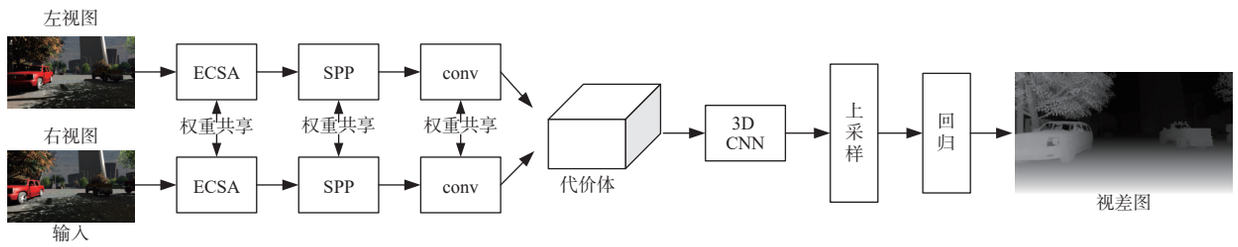


图 1　PSMNet-ECSA 流程图

Fig. 1　Flow chart of PSMNet-ECSA

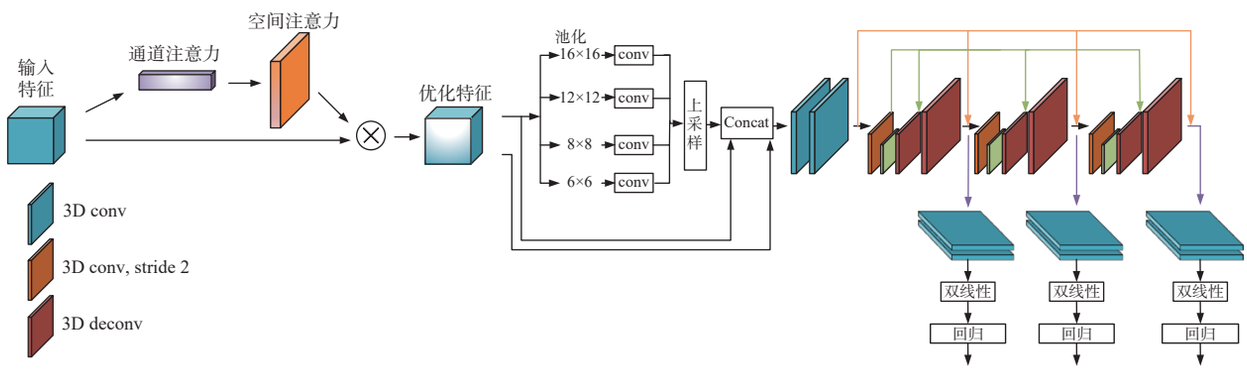


图 2　PSMNet-ECSA 网络结构

Fig. 2　Structure diagram of PSMNet-ECSA network

表 1　PSMNet-ECSA 网络结构参数

Table 1　Structure parameters of PSMNet-ECSA network

结构	参数设置	输出维度
输入		$H\times W\times 3$
双重三池化注意力		
conv0_x	$[3\times 3]\times 3,32$	$\frac{1}{2}H\times\frac{1}{2}W\times 32$
conv1_x	$\begin{bmatrix}3\times 3,128\\3\times 3,128\end{bmatrix}\times 16$,dila=2	$\frac{1}{4}H\times\frac{1}{4}W\times 128$
ca	[maxpool,avgpool,mixpool], gamma=2, b=1	$\frac{1}{4}H\times\frac{1}{4}W\times 128$
sa	[maxpool,avgpool,mixpool], k=3	$\frac{1}{4}H\times\frac{1}{4}W\times 128$
空间金字塔池化		
branch_x	[64, 32, 16, 8] pool, $3\times 3,32$, 双线性插值	$\frac{1}{4}H\times\frac{1}{4}W\times 32$
concat[conv1_2,conv1_4,branch_1, branch_2,branch_3,branch_4]		$\frac{1}{4}H\times\frac{1}{4}W\times 320$
fusion	$\begin{matrix}3\times 3,128\\1\times 1,32\end{matrix}$	$\frac{1}{4}H\times\frac{1}{4}W\times 32$
代价体		
左右特征级联		$\frac{1}{4}D\times\frac{1}{4}H\times\frac{1}{4}W\times 64$
3D 堆叠卷积		
3Dconv0	$\begin{matrix}3\times 3\times 3,32\\3\times 3\times 3,32\end{matrix}$	$\frac{1}{4}D\times\frac{1}{4}H\times\frac{1}{4}W\times 32$
3Dstack1_x	$\begin{bmatrix}\text{conv}3\times 3\times 3,64\\\text{deconv}3\times 3\times 3,32\end{bmatrix}\times 4$	$\frac{1}{4}D\times\frac{1}{4}H\times\frac{1}{4}W\times 32$
output_x	$\begin{bmatrix}3\times 3\times 3,32\\3\times 3\times 3,1\end{bmatrix}\times 3$	$\frac{1}{4}D\times\frac{1}{4}H\times\frac{1}{4}W\times 1$
output[output_1,output_2,output_3]		
上采样	双线性插值	$D\times H\times W$
视差回归		$H\times W$

1.2 双重三池化注意力机制

无论是在图像处理、语音识别还是自然语言处理等方面都有注意力机制的应用。本文提出的双重三池化注意力机制模块，主体分为通道注意力和空间注意力，在特征提取网络中嵌入通道注意力机制、空间注意力机制，减少了过拟合现象，提高了网络特征提取和泛化能力。

通道注意力机制旨在提出一种基于多个特征图通道的显式关联模型，利用神经网络训练的方法，自动提取各通道的影响因子，并在此基础上对各通道进行权值分配，既突出关键特征又抑制次要特征。如图 3 所示，通道注意力分为以下 4 个步骤。

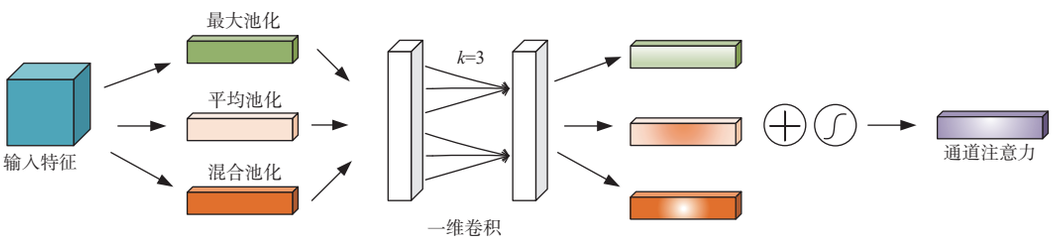


图 3 通道注意力结构

Fig. 3 Structure diagram of channel attention

1）从单张图像开始，提取图像特征，当前特征层X的特征图维度为$[C,H,W]$。

2）特征图在$[H,W]$维度进行平均池化、最大池化、混合池化，池化过后的特征图大小从 $[C,H,W]\rightarrow[C,1,1]$，即$\boldsymbol{F}^c_{\text{avg}}$、$\boldsymbol{F}^c_{\text{max}}$、$\boldsymbol{F}^c_{\text{mix}}$。

$$\boldsymbol{F}^c_{\text{avg}}=\text{GAVGP}^s(X) \tag{1}$$

$$\boldsymbol{F}^c_{\text{max}}=\text{GMAXP}^s(X) \tag{2}$$

$$\boldsymbol{F}^c_{\text{mix}}=\text{GMIXP}^s(X) \tag{3}$$

3）特征$[C,1,1]$可以认为是从每个通道自身提取的权重，这些权重系数反映了每个通道的重要程度，全局池化后的向量通过进行一维卷积操作，卷积核大小为k，生成通道权重。其中卷积核的大小k是根据通道维度C的自适应计算获取，经过网络训练之后获取通道融合权重s_c，W_1、W_2为模型权重参数，σ、δ为网络模型超参数。

$$s_c=\sigma\left(W_2\delta\left(W_1\boldsymbol{F}^c_{\text{avg}}\right)+W_2\delta\left(W_1\boldsymbol{F}^c_{\text{max}}\right)+W_2\delta\left(W_1\boldsymbol{F}^c_{\text{mix}}\right)\right) \tag{4}$$

4）对 3 种不同池化输出结果进行加权累加，然后通过 Sigmoid 运算，将最终的通道权重$[C,1,1]$和原始的特征图$[C,H,W]$进行乘积，最终输出内嵌通道注意力的特征图$M_c(X)$。

$$M_c(X)=s_cX \tag{5}$$

空间注意力旨在提高重点区域的特征表示能力，其实质是通过空间转换来实现对图像信息的有效传递，同时保持重要信息。在此基础上，通过产生一个权值掩膜图像，并对各点进行加权，以此提高某一特定区域的重要性，而使无关背景区域被削弱。如图 4 所示，空间注意力分为以下 4 个步骤。

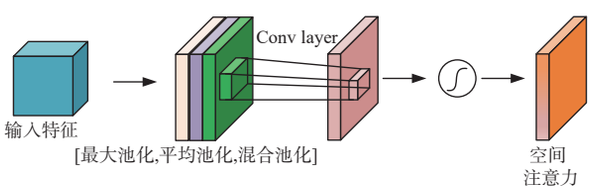


图 4 空间注意力结构

Fig. 4 Structure diagram of spatial attention

1）特征图X沿C维度分别经过平均池化、最大池化、混合池化，形成 3 个$[1,H,W]$的权重向量，即$\boldsymbol{F}^s_{\text{avg}}$、$\boldsymbol{F}^s_{\text{max}}$、$\boldsymbol{F}^s_{\text{mix}}$，特征图$X$的维度从原始的$[C,H,W]$降低为 3 个不同的$[1,H,W]$。

$$\boldsymbol{F}^s_{\text{avg}}=\text{GAVGP}^c(X) \tag{6}$$

$$\boldsymbol{F}^s_{\text{max}}=\text{GMAXP}^c(X) \tag{7}$$

$$\boldsymbol{F}^s_{\text{mix}}=\text{GMIXP}^c(X) \tag{8}$$

2）经过 3 种池化获取的 3 个不同的特征图$\boldsymbol{F}^s_{\text{avg}}$、$\boldsymbol{F}^s_{\text{max}}$、$\boldsymbol{F}^s_{\text{mix}}$，然后在其通道维度进行叠加，继而生成一个$[3,H,W]$的融合特征图。

3）将上述的$[3,H,W]$的融合特征图进行卷积运算和非线性激活响应，特征图的维度再次从$[3,H,W]$降低为$[1,H,W]$，此时获取的特征图$[1,H,W]$的各点数值反映了其图像信息层面的重要性。

$$s_s=\sigma\left(\text{Conv}\left(\boldsymbol{F}^s_{\text{avg}},\boldsymbol{F}^s_{\text{max}},\boldsymbol{F}^s_{\text{mix}}\right)\right) \tag{9}$$

4）最后用所求的空间权重$[1,H,W]$乘以原来的特征图$[C,H,W]$，也就是给每一个点赋权，最终输出内嵌空间注意力的特征图$M_s(X)$。

$$M_s(X)=s_sX \tag{10}$$

将上述通道注意力和空间注意力结合, 输出双重三池化注意力机制的特征图Y。

$$X' = M_c(X) \tag{11}$$

$$Y = M_s(X') \tag{12}$$

2　视差回归

一般而言, 立体匹配算法会生成一个匹配代价体, 通过在代价体视差维度进行 argmin 运算获得最终的视差预测结果, 但这种方式只能产生整数精度视差且过程不可微, 无法进行后向传播计算。为了克服上述的限制, KENDALL A 等人[14]提出了一种可微分且回归预测平滑视差的 soft argmin 运算$\sigma(\cdot)$, 根据代价值c_d和 softmax 算子$\sigma(\cdot)$计算每个视差d的可能性, 最终的预测视差$\hat{d}$由所有视差的加权值求和获得, 如式(13)所示, 其中$D_{\max}$为设置的最大视差。

$$\hat{d} = \sum_{d=0}^{D_{\max}} d \times \sigma(-c_d) \tag{13}$$

本文从随机初始化权重参数开始训练整个有监督网络模型[27], 选择smooth_{L_1}作为损失函数[28], 如式(14)、式(15)所示。

$$\text{smooth}_{L_1}(x) = \begin{cases} 0.5x^2, & \text{if } |x| < 1 \\ |x| - 0.5, & \text{othersize} \end{cases} \tag{14}$$

$$L\left(d, \hat{d}\right) = \frac{1}{N} \sum_{i=1}^{N} \text{smooth}_{L_1}\left(d_i - \hat{d}_i\right) \tag{15}$$

式中: N为像素个数; d为真实视差; $\hat{d}$为预测视差。

3　实验与分析

本文实验系统环境为 Ubuntu16.04 LTS (GNU/Linux 4.15.0-142-generic x86_64), GeForce GTX 1080 Ti 的图形处理器, 显存大小为 12 Gb, 编程语言为 Python, 集成开发环境为 Visual Studio Code, 神经网络训练框架为 Pytorch, 计算平台为 CUDA9.2, torch 版本为 1.6.0+cu92。

3.1　SceneFlow 数据集实验

为了验证本文提出的双重三池化方式增强型注意力机制模块的有效性, 实验采用目前规模最大的双目立体视觉公开数据集 SceneFlow 进行对比分析, 考虑到原 SceneFlow 数据集过于庞大以及实验硬件环境的限制, 在保留训练集和测试集类别的同时, 选取部分数据建立了子数据集并进行实验分析, 子数据集中训练集仍然包含了 Monkaa、Driving、Flying 3 种类型的图像共计 9257 对, 测试集共计 1490 对, 图像尺寸保持一致, H=540, W=960, 算法精度评价指标为端点误差(end point error, EPE), 衡量了图像中每一对同名像素点的视差估计误差, 本文算法训练参数如表 2 所示。

表 2　SceneFlow 训练参数

Table 2　SceneFlow training parameters

参数类型	参数值
训练周期/epoch	10
训练时间/h	6
训练批次大小	6
测试批次大小	4
梯度下降优化器	Adam($\beta_1 = 0.9, \beta_2 = 0.999$)
固定学习率	0.001
最大视差	192

本文实验将双重三池化注意力机制模块分解为单通道双池化模块(PSMNet-CA)、单空间双池化模块(PSMNet-SA)、双重最大池化模块(PSMNet-MAX)、双重平均池化模块(PSMNet-AVG)、双重混合池化模块(PSMNet-MIX)、双重双池化模块(PSMNet-CSA)和双重三池化模块(PSMNet-ECSA), 通道注意力中不同通道特征融合采用自适应一维卷积核(gamma=2, b=1), 模型中通道维度分别为 32、64、128、128, 其对应的具体卷积核尺寸为 3、3、5、5, 空间注意力中非区域特征融合的二维卷积核尺寸设置为 7, 池化操作分别在对应的通道或空间维度中进行。实验结果如表 3 所示。

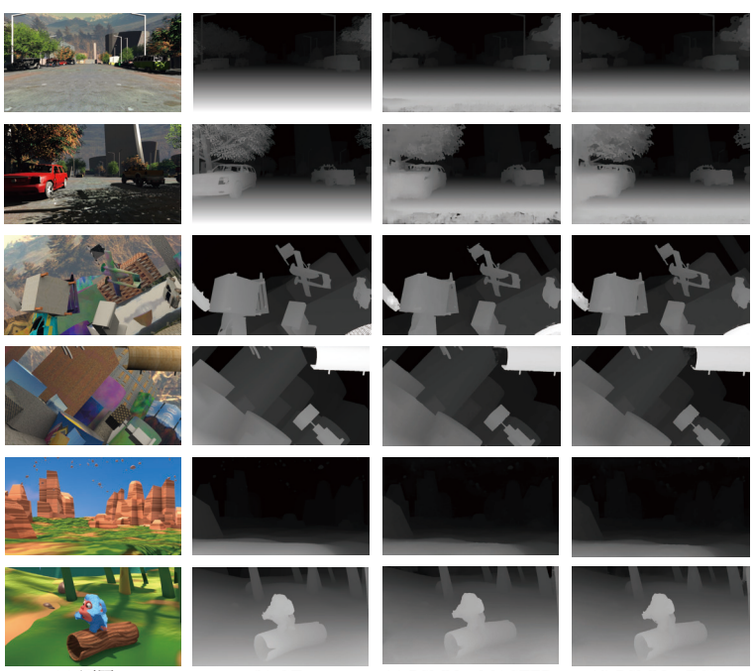
表 3　SceneFlow 数据集消融实验结果

Table 3　Ablation experiment results of SceneFlow dataset

算法类型	模块结构						端点误差
	通道注意力			空间注意力			
	平均池化	最大池化	混合池化	平均池化	最大池化	混合池化	
MASTER							1.466
CA	√	√					1.435
SA				√	√		1.428
AVG	√			√			1.397
MAX		√			√		1.425
MIX			√			√	1.381
CSA	√	√		√	√		1.375
ECSA	√	√	√	√	√	√	1.349

根据表 3 可知, 对于整张图像的平均绝对视差误差来说, CA 和 SA 模块分别降低到了 1.435 和 1.428, 但是单重注意力机制在网络特征提取能力的提升较为有限, 而在双重注意力结构下的不同

池化模块取得了更优的结果。最大池化保留了特征图中响应最强烈的部分，虽然使得网络易于优化，但摒弃了诸多细节信息，而平均池化更多地保留了背景信息，所以混合池化优化了池化调节规则，同时保留前景纹理信息和背景细节特征，有助于防止网络过拟合现象。例如 PSMNet-MIX 算法，在 alpha=0.8 的条件下，平均绝对误差为 1.381，相较于单一池化类型有了明显的提升。而最终的 PSMNet-ECSA 算法，采用了通道和空间双重注意力结构，同时使用了混合池化、最大池化和平均池化的特征维度，形成了多维度特征融合的特征图，评价指标 EPE 为最优的 1.349，算法精度提升了 8%。此外，PSMNet-MASTER 算法模型参数为 5 224 768，上述各算法模型参数增量都在 10 k 以内，最优的模型 PSMNet-ECSA 算法模型参数为 5 228 630，网络模型参数增长量不足 1%，所以双重三池化注意力模块是一个轻量且有效的模型。SceneFlow 数据集视差图对比如图 5 所示。



(a) 原图 (b) GT (c) PSMNet-MASTER (d) PSMNet-ECSA

图 5 SceneFlow 数据集视差图对比

Fig. 5 Disparity map comparison of SceneFlow dataset

本文在实验室环境下重新训练了其他的立体匹配网络[29]，例如 GCNet[14]、DeepPruner[30]，通过在相同测试集的精度对比（结果见表 4），证明了 PSMNet-ECSA 在视差计算上具有一定优势。

表 4 SceneFlow 测试集精度对比

Table 4 Comparison of accuracy on SceneFlow test set

算法	EPE
PSMNet-ECSA	1.349
DeepPruner	1.673
GCNet	2.424

3.2 KITTI2015 数据集实验

为了进一步验证双重三池化方式的增强型注意力机制模块的有效性，实验采用双目立体视觉公开数据集 KITTI2015[25] 进行对比分析，数据集总共包含 200 对双目图像，图像尺寸为 H=376 pixel，W=1 240 pixel，该实验将前 160 对图像划分为训练集，后 40 对图像划分为测试集，算法精度评价指标为 Error Threshold=3 pixel 以内的误差像素百分比，本文神经网络模型预加载了上一节模型权重，训练参数如表 5 所示。

表 5 KITTI2015 训练参数

Table 5 KITTI2015 training parameters

参数类型	参数值
训练周期/epoch	300
训练时间/h	5
训练批次大小	6
测试批次大小	4
初始学习率（前200个周期）	0.001
后期学习率（后100个周期）	0.000 1
最大视差/pixel	192

表 6 分别表示了 4 种模块在 KITTI2015 数据集上的表现, 模型共训练了 300 个周期, 每个周期会分别进行一次训练和测试, 前者每次进行 27 次迭代, 后者为 10 次迭代。评价指标 Test error 含义是在当前图像视差范围 0～192 pixel 的所有像素点, 选取视差估计误差在 3 个 pixel 以内的像素百分比。PSMNet-MASTER 的 3px-Err 为 2.304%, PSMNet-CSA、PSMNet-MIX 和 PSMNet-ECSA 分别降低了 6.25%、16.10%、16.79%, 可知在误差容忍度更高的阈值误差指标下, 双重双池化模块虽然在特征提取中有所提升, 但由于单一的选择策略容易产生较大的误差, 而混合池化的平衡调节方式能有限地缓和绝对或过于平均的问题, 能够较大幅度地降低误差像素百分比。KITTI2015 数据集视差图对比如图 6 所示。

表 6　KITTI2015 数据集消融实验结果

Table 6　Ablation experiment results of KITTI2015 dataset

注意力类型	模块结构						3px-Err/%
	通道注意力			空间注意力			
	平均池化	最大池化	混合池化	平均池化	最大池化	混合池化	
MASTER							2.304
CSA	√	√		√	√		2.160
MIX			√			√	1.933
ECSA	√	√	√	√	√	√	1.917

(a) 原图　(b) GT　(c) PSMNet-MASTER　(d) PSMNet-ECSA

图 6　KITTI2015 数据集视差图对比

Fig. 6　Disparity map comparison of KITTI2015 dataset

3.3　真实场景实验

为了验证双重三池化注意力机制的 PSMNet-ECSA 模型在真实场景的匹配精度和实时性, 实验采用鲍鱼壳的小场景进行视差计算及三维重建, 证明了该算法在 50 cm 以内的拍摄距离, 对小型带纹理物体的三维重建具有一定优势。本文实验使用 PSMNet-ECSA 模型经过 SceneFlow 数据集 10 个周期和 KITTI2015 数据集 300 个周期的训练, 使用双目相机在鲍鱼正上方距离 30 cm 左右的位置, 自上而下的方式, 拍摄的左右双目鲍鱼图像 E582、0524, 图像尺寸为 640×360 像素, 当前算法单张视差图的计算时间在 0.5 s 左右, 具有良好的实时性。经过立体校正之后的标准双目图像及左目预测视差图如图 7 所示, 可以看到左、右目图像中的鲍鱼分别偏向另一侧, 并且基本达到极线校正的标准, 并且在左目视差图鲍鱼区域的灰度值高于桌面区域, 表示视差越大也就是距离相机越近的客观事实, 不过由于鲍鱼自身高度较小, 所以与背景区域的灰度差异不够明显, 下面将通过三维点云重建来展示效果。

(a) 左目图像　(b) 右目图像　(c) 左视差图

图 7　左、右目图像和左视差图

Fig. 7　Left and right stereo images, and left disparity map

基于三角测量原理, 双目视觉中视差与深度的转换关系为 $z = fB/d$, 其中 z 为当前像素的深度, f 为相机焦距, B 为双目基线长度, d 为像素视差, 经过前面的双目相机标定获得了 f = 459.5 mm, B = 47.8 mm, 视差 d 由视差图提供, 至此可以计算鲍鱼的深度信息, 并调用 PCL 库保存 XYZRGB 类型点云文件如图 8 所示, 结果保留了鲍鱼三维信息和颜色信息。

图 8 鲍鱼三维重建点云模型

Fig. 8 Abalone three-dimensional reconstruction point cloud model

为了验证 PSMNet-ECSA 的重建精度，本文基于真实鲍鱼进行手工测量获取真实长、宽、呼吸孔间距等数据，同时分别利用 PSMNet-MASTER 和 PSMNet-ECSA 算法重建点云模型之后，利用三维模型测量工具 MeshLab 获取重建三维模型中的上述数据进行分析对比。从表 7 可知，PSMNet-MASTER 整体各项数据的平均相对误差为 5.5%，而 PSMNet-ECSA 则是降到了 2.8%，平均相对误差指标基本降低了一倍。此外，对于鲍鱼长宽数据而言，基础数据本身偏大，相对误差对两者而言都更小，但是对于呼吸孔这类狭小区域，双重三池化模块在处理边缘和细节特征更具优势，呼吸孔的平均相对误差由 7.2% 降为了 3.5%。

表 7 鲍鱼重建点云精度对比

Table 7 Comparison of point cloud reconstruction accuracy for abalone

图像	类型	长度	宽度	呼吸孔12	呼吸孔23	呼吸孔34
E582	真实值/mm	94	67	15	13	16
	PSMNet-MASTER测量值/相对误差/(mm/%)	91.2/3.0	64.4/3.9	13.8/8.0	14.2/9.2	16.8/5.0
	PSMNet-ECSA测量值/相对误差/(mm/%)	92.9/1.2	65.5/2.2	14.2/5.3	12.6/3.1	15.5/3.1
0524	真实值/mm	79	55	11	12	–
	PSMNet-MASTER测量值/相对误差/(mm/%)	76.6/3.0	52.9/3.8	10.1/8.2	11.3/5.8	–
	PSMNet-ECSA测量值/相对误差/(mm/%)	77.6/1.8	53.5/2.7	10.5/4.5	11.8/1.7	–

4 结论

为了解决小型纹理类物体的视差计算及三维重建测量问题，本文提出了一种基于双重三池化注意力机制的 PSMNet-ECSA 算法，在实验环境和数据集一致的条件下，分别在人工合成森林场景、街道场景、鲍鱼场景，基于原算法 PSMNet 对上述不同模块进行消融实验对比分析，验证各个模块在 EPE 和 3px-Err 评价指标的不同优化表现，验证了算法在不同场景下均具有一定的适应性，算法的各项评价指标也体现了良好的泛化能力。PSMNet-ECSA 算法在 SceneFlow 数据集上平均绝对误差为 1.349, EPE 精度提升了 8%, 在 KITTI2015 数据集上, 3px-Err 误差指标为 1.917, 精度提升了 16.79%, 同时分析了时间效率、模型参数量等因素，验证了双重三池化模块增强图像特征提取能力和泛化能力，证明了 PSMNet-ECSA 算法的优势。最后将 PSMNet-ECSA 应用于鲍鱼重建三维点云模型，长、宽、呼吸孔等距离测量平均相对误差在 3% 以内，说明了该算法满足实时性要求和，在精细场景中具有良好的实际应用价值。

参考文献：

[1] KAKADE A, DESHPANDE M, SARDESHPANDE S, et al. 3D modelling using sequential and convolutional generative adversarial networks[C]// 2021 International Conference on Artificial Intelligence and Machine Vision, New York: IEEE, 2021: 1-4.

[2] MIN K, HAN S, LEE D, et al. SAE Level 3 Autonomous driving technology of the ETRI[C]// 2019 International Conference on Information and Communication Technology Convergence. New York: IEEE, 2019: 464-466.

[3] YANG J L, REN P R, ZHANG D Q, et al. Neural aggregation network for video face recognition[C]// IEEE Conference on Computer Vision and Pattern Recognition. New York: IEEE, 2017: 5216–5225.

[4] ZENATI N, ZERHOUNI N. Dense stereo matching with application to augmented reality[C]// 2007 IEEE International Conference on Signal Processing and Communications. New York: IEEE, 2007: 1503-1506.

[5] HIRSCHMULLER H. Accurate and efficient stereo processing by semi-global matching and mutual information[J]. 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2005(2): 807-814.

[6] LIU T F, LIN D Y, LAN W Y. PatchMatch stereo - cross dynamic windows based on textured bithreshold rule[C]// Chinese Control Conference. New York: IEEE, 2023: 7382-7387.

[7] ZAGORUYKO S, KOMODAKIS N. Learning to com-

pare image patches via convolutional neural networks[C]// IEEE Computer Vision and Pattern Recognition. New York: IEEE, 2015: 4353-4361.

[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] YE X, LI J, WANG H, et al. Efficient stereo matching leveraging deep local and context information[J]. IEEE Access, 2017(5): 18745-18755.

[10] DOSOVITSKIY A, FISCHER P, ILG E, et al. FlowNet: learning optical flow with convolutional networks. [C]// IEEE International Conference on Computer Vision. New York: IEEE, 2015: 2758-2766.

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

[12] CHANG J, CHEN Y. Pyramid stereo matching network[C]// IEEE Computer Vision and Pattern Recognition. New York: IEEE, 2018: 5410-5418.

[13] YANG G, MANELA J, HAPPOLD M, et al. Hierarchical deep stereo matching on high-resolution images[C]// IEEE Computer Vision and Pattern Recognition. New York: IEEE, 2019: 5515-5524.

[14] 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. New York: IEEE, 2017: 66-75.

[15] CHENG X, ZHONG Y, HARAKEH A, et al. Learning stereo matching network with convolutional spatial propagation network[C]// IEEE Conference on Computer Vision and Pattern Recognition. New York: IEEE, 2020: 156–165.

[16] TANKOVICH V, KAR A, HANE C, et al. Hitnet: hierarchical iterative tile refinement network for real-time stereo matching [C]//IEEE/CVF Conference on Computer Vision and Pattern Recognition . New York: IEEE, 2021: 14362-14372.

[17] GUO M H, XU T X, LIU J J, et al. Attention mechanisms in computer vision: A survey[J]. Computational Visual Media, 2022, 8(3): 331-368.

[18] GREGOR K, DANIHELKA I, GRAVES A, et al. DRAW: a recurrent neural network for image generation[C]// 32nd International Conference on Machine Learning. New York: Association of Computing Machinery, 2015: 1462–1471.

[19] FU J, LIU J, TIAN H J, et al. Dual attention network for scene segmentation[C]// IEEE/CVF Conference on Computer Vision and Pattern Recognition.New York: IEEE, 2019: 3141-3149.

[20] CHU X, YANG W, OUYANG W L, et al. Multi-context attention for human pose estimation[C]// IEEE Conference on Computer Vision and Pattern Recognition. New York: IEEE, 2017: 5669-5678.

[21] XU T, ZHANG P C, HUANG Q Y, et al. AttnGAN: fine-grained text to image generation with attentional generative adversarial networks[C]// IEEE/CVF Conference on Computer Vision and Pattern Recognition. New York: IEEE, 2018: 1316-1324.

[22] MNIH V, HEESS N, GRAVES A, et al. Recurrent models of visual attention[J]. 27th International Conference on Neural Information Processing Systems, 2014(2): 2204-2212.

[23] JADERBERG M, SIMONYAN K, ZISSERMAN A, et al. Spatial transformer networks[J]. 28th International Conference on Neural Information Processing Systems, 2015(2): 2017-2025.

[24] 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.

[25] MENZE M, GEIGER A. Object scene flow for autonomous vehicles[C]// IEEE Computer Society Conference on Computer Vision and Pattern Recognition.New York: IEEE, 2015: 3061-3070.

[26] HE K, ZHANG X, REN S, et al. Spatial pyramid pooling in deep convolutional networks for visual recognition[C]// European Conference on Computer Vision. Berlin: Springer, 2014: 346-361.

[27] GOODFELLOW I, BENGIO Y, COURVILLE A, et al. Deep Learning[M]. Cambridge: MIT Press, 2016.

[28] GIRSHICK R. Fast R-CNN[C]// IEEE International Conference on Computer Vision. New York: IEEE, 2015: 1440-1448.

[29] LAGA H, JOSPIN L V, BOUSSAID F, et al. A survey on deep learning techniques for stereo-based depth estimation[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2022, 44(4): 1738-1764.

[30] DUGGAL S, WANG S, MA W C, et al. DeepPruner: Learning efficient stereo matching via differentiable PatchMatch[C]// IEEE/CVF International Conference on Computer Vision. New York: IEEE, 2019: 4383-4392.