Journal of System Simulation

Volume 36 | Issue 1 Article 8

1-20-2024

Multi-view Depth Estimation Based on Adaptive Space Feature Enhancement

Dong Wei
School of Information Science and Engineering, Shenyang University of Technology, Shenyang 110870, China, dongweisut@sut.edu.cn

Huan Liu
School of Information Science and Engineering, Shenyang University of Technology, Shenyang 110870, China, 1121339441@qq.com

Xiaohan Zhang
School of Information Science and Engineering, Shenyang University of Technology, Shenyang 110870, China

Changkai Li
School of Information Science and Engineering, Shenyang University of Technology, Shenyang 110870, China

See next page for additional authors

Multi-view Depth Estimation Based on Adaptive Space Feature Enhancement

Abstract

Abstract: A multi-view depth estimation algorithm based on adaptive space feature enhancement (ASFE) is presented to improve the multi-view depth estimation accuracy. A multi-scale feature extraction module composed of an improved feature pyramid network (FPN) and ASFE is designed. This module obtains multi-scale feature maps with global context-aware information and coordinate information. The residual learning network is used to optimize the depth map to prevent the problem of blurred reconstructed edges in multiple convolution operations. The proposed algorithm constructs a focal loss function through the idea of classification to enhance the prediction ability of the network model. The experimental results show that on the technical university of denmark (DTU) dataset, compared with the cascade MVSNet (CasMVSNet) method, the proposed method reduces overall accuracy error, running time, and video memory resource occupation by 14.08%, 72.15%, and 4.62%, respectively. The Mean of the model on the Tanks and Temples dataset is superior to other algorithms, which proves the effectiveness of the proposed multi-view depth estimation algorithm based on ASFE.

Keywords

multi-view depth estimation, adaptive space feature enhancement, residual learning network, convolution operation, focal loss function

Authors

Dong Wei, Huan Liu, Xiaohan Zhang, Changkai Li, Tianyi Sun, and Ziyou Zhang

Recommended Citation

Wei Dong, Liu Huan, Zhang Xiaohan, et al. Multi-view Depth Estimation Based on Adaptive Space Feature Enhancement[J]. Journal of System Simulation, 2024, 36(1): 110-119.

第 36 卷第 1 期 系统仿真学报© Vol. 36 No. 1
2024 年 1 月 Journal of System Simulation Jan. 2024

基于自适应空间特征增强的多视图深度估计

魏东，刘欢*，张潇瀚，李昌恺，孙天翼，张子优
(沈阳工业大学信息科学与工程学院，辽宁 沈阳 110870)

摘要：*为了提高多视图深度估计结果精度，提出一种基于自适应空间特征增强的多视图深度估计算法。设计了由改进后的特征金字塔网络(feature pyramid network, FPN)和自适应空间特征增强(adaptive space feature enhancement, ASFE)组成的多尺度特征提取模块，获取到具有全局上下文信息和位置信息的多尺度特征图像。通过残差学习网络对深度图进行优化，防止多次卷积操作出现重建边缘模糊的问题。通过分类的思想构建focal loss 函数增强网络模型的判断能力。*由实验结果可知，该算法在DTU(technical university of denmark)数据集上和CasMVSNet(Cascade MVSNet)算法相比，在整体精度误差、运行时间、显存资源占用上分别降低了14.08%、72.15%、4.62%。在Tanks and Temples数据集整体评价指标Mean上该模型优于其他算法，证明提出的基于自适应空间特征增强的多视图深度估计算法的有效性。

关键词：多视图深度估计；自适应空间特征增强；残差学习网络；卷积操作；focal loss 函数
中图分类号：TP391.4 文献标志码：A 文章编号：1004-731X(2024)01-0110-10
DOI: 10.16182/j.issn1004731x.joss.23-0112
引用格式: 魏东, 刘欢, 张潇瀚, 等. 基于自适应空间特征增强的多视图深度估计[J]. 系统仿真学报, 2024, 36(1): 110-119.
Reference format: Wei Dong, Liu Huan, Zhang Xiaohan, et al. Multi-view Depth Estimation Based on Adaptive Space Feature Enhancement[J]. Journal of System Simulation, 2024, 36(1): 110-119.

Multi-view Depth Estimation Based on Adaptive Space Feature Enhancement

Wei Dong, *Liu Huan**, *Zhang Xiaohan*, *Li Changkai*, *Sun Tianyi*, *Zhang Ziyou*
(School of Information Science and Engineering, Shenyang University of Technology, Shenyang 110870, China)

Abstract: *A multi-view depth estimation algorithm based on adaptive space feature enhancement (ASFE) is presented to improve the multi-view depth estimation accuracy. A multi-scale feature extraction module composed of an improved feature pyramid network (FPN) and ASFE is designed. This module obtains multi-scale feature maps with global context-aware information and coordinate information. The residual learning network is used to optimize the depth map to prevent the problem of blurred reconstructed edges in multiple convolution operations. The proposed algorithm constructs a focal loss function through the idea of classification to enhance the prediction ability of the network model.* The experimental results show that on the technical university of denmark (DTU) dataset, compared with the cascade MVSNet (CasMVSNet) method, the proposed method reduces overall accuracy error, running time, and video memory resource occupation by 14.08%, 72.15%, and 4.62%, respectively. The Mean of the model on the

收稿日期：2023-02-09 修回日期：2023-04-10
基金项目：辽宁省教育厅项目(LJGD2020006)
第一作者：魏东(1968-)，男，副教授，硕士，研究方向为计算机图形学、计算机视觉、虚拟现实等。E-mail：dongweisut@sut.edu.cn
通讯作者：刘欢(1998-)，女，硕士生，研究方向为计算机视觉。E-mail：1121339441@qq.com

Tanks and Temples dataset is superior to other algorithms, which proves the effectiveness of the proposed multi-view depth estimation algorithm based on ASFE.

Keywords: multi-view depth estimation; adaptive space feature enhancement; residual learning network; convolution operation; focal loss function

0 引言

近年来由于自动驾驶、虚拟现实和导航等行业的蓬勃发展，三维重建成为热门的研究方向。而多视图深度估计算法是三维重建中的重要环节，其本质是在已知相机内外参数的条件下进行一对多图像特征匹配任务，其中参考图像的每个像素都沿着单应变换后目标图像中的极线进行搜索，用最低匹配成本生成的代价体估计出最佳深度，恢复出重建场景的3D模型。

传统多视图深度估计算法对理想的朗伯体重建时获得不错的效果，例如Gipuma[1]算法在重建精度方面取得了巨大成功。但是传统的多视图深度估计算法性能依赖人工设计的特征，导致在低纹理区域、反射表面、不同光照强度下提取特征困难，使重建完整性方面表现较差且需要很长时间来建立3D对应关系。

随着深度学习的应用领域逐渐扩大。文献[2-3]用学习到的特征进行立体匹配和半全局匹配的后处理。文献[4]将3D CNN(convolutional neural networks)用于代价体正则化并设计了Soft argmin方法。多视图深度估计研究受以上算法启发，文献[5]提出通过复杂的体素视图选择预先计算代价体的SurfaceNet网络，并使用3D CNN来正则化和推断表面体素。文献[6]利用摄像机参数通过可微分投影操作来形成代价体，使用3D CNN划分体素是否属于物体表面。但是体素的表示方法存在对内存需求大的缺点，仅适用于小规模场景低分辨率重建。为了解决这个问题，文献[7]提出了MVSNet网络。该网络由特征提取、构建代价体、代价体正则化、深度估计4个部分构成，且深度估计结果以深度图的方式输出，并对目标场景进行三维重建。后续的大部分多视图深度估计算法遵循MVSNet网络流程进行设计。

首先，MVSNet方法在处理更高分辨率的图像时，高维的代价体需要占用大量显存资源估计深度。为缓解这个问题，后续的研究分为2种：一种是使用递归结构的网络，例如文献[8-9]提出在代价体正则化部分引入门控循环单元的R-MVSNet网络和采用长短时记忆网络构建的D2HC-RMVSNet网络。采用递归结构的方法用时间来替代空间上的占用，虽然显存资源占用减少，但是网络推理速度变慢；第二种是采用级联结构的网络，例如文献[10]提出三层级联结构的CasMVSNet，文献[11]提出四层级联结构的HighRes-MVSNet算法，不仅降低了对显存的占用，还提高了估计每张深度图像的速度。

其次，特征提取的效果会直接影响深度估计结果精度。所以在特征提取阶段获取具有全局上下文信息的特征在多视图深度估计算法中尤为重要。文献[12]利用多尺度信息提取原始尺度的密集特征获取到全局上下文信息。文献[13]在提取到特征后加入自注意力机制以增强特征信息。同时，由于卷积操作的局部性阻碍了对全局上下文信息的感知。如果感受野足够大，则可以更好地重建低纹理区域。例如MCV-MVSNet [14]方法在特征提取模块加入双UNet网络，增加局部感受以获取特征图中细节信息。上述算法虽然在效率和完整度上取得较大提升，但是网络模型的整体性能依旧具有改善的空间。

最后，网络模型训练时计算出的损失值不断地反向传播去更新模型的各个参数，让预测值向真值接近，所以损失函数的选择会对网络模型的整体预测性能产生影响。

DOI: 10.16182/j.issn1004731x.joss.23-0112

综上所述，基于深度学习的多视图深度估计算法仍有以下问题：①缺少全局信息和合适的感受野来获取更有用的特征信息。②卷积规模太大导致显存占用较大、训练时间更久和可能影响深度结果，从而导致重建边缘较差。③多视图深度估计网络模型训练时，图像的不同复杂区域损失占比相等，这限制了模型的判断力，无法提升模型的性能。因此本文提出一种自适应空间特征增强的多视图深度估计算法。

(1) 在多尺度特征提取阶段不仅改进了特征金字塔网络[15](feature Pyramid network, FPN)，还设计了自适应空间特征增强(adaptive space feature enhancement, ASFE)模块从图像中获取到具有全局上下文信息和位置信息的特征。

(2) 网络模型采用级联结构且添加深度优化模块，利用残差学习网络优化深度图。

(3) 采用 focal loss 作为损失函数，对特征提取困难的区域给予相应较大的损失权重。提升网络训练效果和增强模型鲁棒性。

1 算法分析

本文算法网络结构如图 1 所示。采用级联结构，由多尺度特征提取、深度求精和深度优化模块组成。深度求精模块包括构建代价体、代价体正则化、深度估计 3 个部分。将相机拍摄的 N+1 张不同角度视图作为输入设为 $\boldsymbol{I}_v$，v 的取值范围为 0~N，$\boldsymbol{I}_{v=0}$ 为参考图像，其余 N 张为目标图像，记为 $\boldsymbol{I}_{v=1,2,\cdots,N}$。首先，二维图像进入多尺度特征提取模块，生成多尺度特征图。不同尺度的特征图经过单应变换到参考特征图坐标系下的平行深度平面构建代价体。再使用 3D UNet[16]对代价体进行正则化，把正则化后的代价体通过 Softmax 操作沿深度方向求出概率体 P，最后使用 Winner-take-all 算法得到估计深度图。同时，上一层估计出的深度图引导下一层中单应变换时获取更精细的深度范围和深度间隔。最后，将第三层输出的初始深度图 D^{est} 输入到残差学习网络优化后得到最终深度图 D^{ref}。

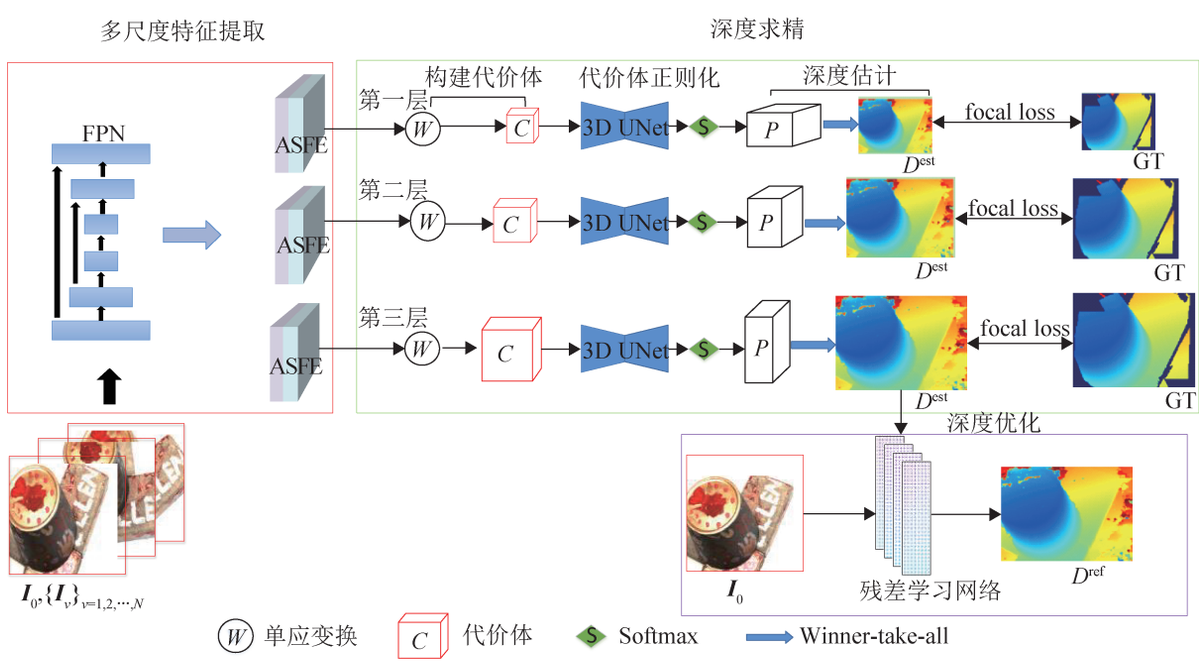

图 1 网路结构图

Fig. 1 Structure of network

1.1 改进的FPN模块

由于从图像中提取到的特征作为后面步骤的输入，所以从图像中获取到具有全局上下文信息和潜在位置信息的特征对多视图深度估计算法显得尤为重要。本文算法在改进的FPN模块后，设计了将全局上下文信息与局部特征进行加权融的ASFE模块以获取到更加可靠的图像特征。

改进的FPN由Conv模块和Out模块组成，结构如表1所示。为了减少显存资源占用，将FPN模块中BN层和ReLu激活函数的组合改为INPLACE-ABN[17] (inplace activated batch normalization)层。FPN模块将分辨率为$W\times H$的图像作为输入对象，Conv0~Conv2模块对输入进行下采样操作，Out0~Out2模块中对输入进行2倍上采样、拼接操作、卷积操作，输出分辨率分别为$W/4\times H/4$、$W/2\times H/2$、$W\times H$，通道分别为32、16、8的N+1组多尺度特征图。

INPLACE-ABN层反向传播中的参数通过翻转前向传播计算时的值得到，因此可以放弃部分中间计算结果，达到减少显存资源占用的目的。

表1 改进的FPN结构
Table 1 Improved structure of FPN

模块	卷积层描述	输出大小
Conv0	3×3 8,Inplace-ABN	$W\times H\times 8$
	3×3 8, Inplace-ABN	$W\times H\times 8$
Conv1	5×5 16, s=2, Inplace-ABN	$W/2\times H/2\times 16$
	3×3 16, Inplace-ABN	$W/2\times H/2\times 16$
	3×3 16, Inplace-ABN	$W/2\times H/2\times 16$
Conv2	5×5 32, s=2, Inplace-ABN	$W/4\times H/4\times 32$
	3×3 32, Inplace-ABN	$W/4\times H/4\times 32$
	3×3 32, Inplace-ABN	$W/4\times H/4\times 32$
Out0	Conv2, 1×1 32, Inplace-ABN	$W/4\times H/4\times 32$
Out1	(2*Out0, Conv1),1×1 16, Inplace-ABN	$W/2\times H/2\times 16$
Out2	(2*Out1, Conv0),1×1 8,Inplace-ABN	$W\times H\times 8$

1.2 ASFE模块

ASFE模块由可变形卷积网络[18](deformable convolution network, DCN)和坐标注意力机制[19](coordinate attention, CA)组成，将改进的FPN模块提取的特征转换为具有全局信息的空间特征，其结构如图2所示。

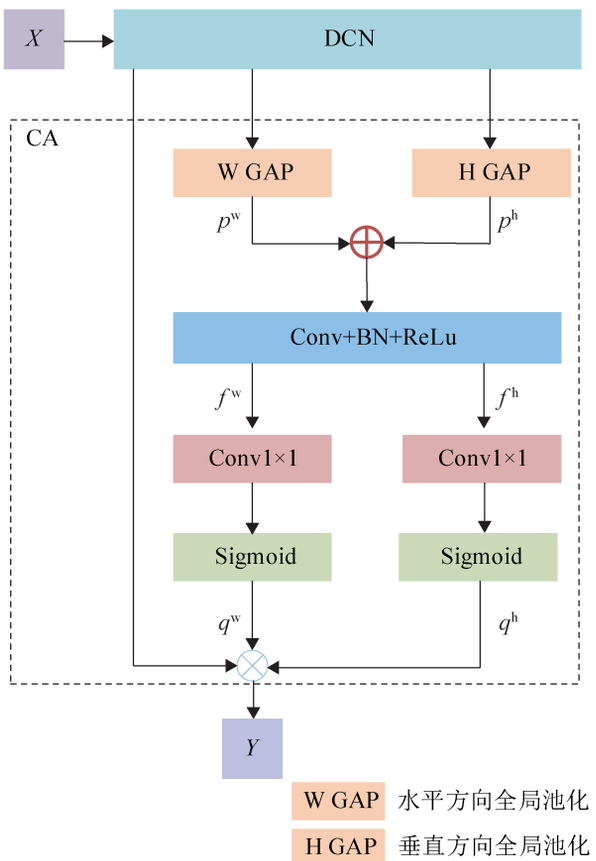

图2 ASFE模块结构图
Fig. 2 Structure of ASFE

改进的FPN模块提取到的是局部上下文信息，而CA通过对输入特征分别沿水平和垂直两个方向获取位置信息，同时沿这两方向进行获取全局感受野的操作，输出的结果是全局上下文信息。改进的FPN和CA在上下文信息获取范围上存在差异，所以加入可以自适应调整特征获取范围的DCN模块作为过渡。DCN对卷积核的每个元素加入方向参数即额外偏移量，从而根据当前输入特征内容自适应地调整采样位置以获取不同感受野。

以往通过对通道之间的相互依赖性进行编码操作就可以获取全局信息，例如具有通道注意力的SENet[20]。但是，对于生成的空间注意力图来讲，SENet忽略了重要的位置信息。所以本文算法采用将位置信息嵌入到通道中的CA模块。CA

DOI: 10.16182/j.issn1004731x.joss.23-0112

模块将输入特征沿水平和垂直两个方向分别用$W\times 1$和$1\times H$的两个全局平均池化层对通道进行编码：

$$\boldsymbol{P}^{\mathrm{w}}(w)=\frac{1}{H}\sum_{j=1}^{H}\boldsymbol{x}(w,j) \tag{1}$$

$$\boldsymbol{P}^{\mathrm{h}}(h)=\frac{1}{W}\sum_{i=1}^{W}\boldsymbol{x}(i,h) \tag{2}$$

式中：$\boldsymbol{x}\in\boldsymbol{X}^{W\times H\times C}$，$W$、$H$、$C$分别表示图像宽、高、通道数；$(w,j)$和$(i,h)$为输入张量，分别表示沿水平和垂直方向的特征坐标位置；$\boldsymbol{p}^{\mathrm{w}}$和$\boldsymbol{p}^{\mathrm{h}}$表示沿水平和垂直两个方向的池化层输出结果。

沿水平和垂直两个方向的池化层输出结果进行拼接操作，将拼接结果通过1×1卷积操作、BN层和非线性激活函数得到中间特征图：

$$\boldsymbol{f}=\delta(F_1[\boldsymbol{p}^{\mathrm{w}},\boldsymbol{p}^{\mathrm{h}}]) \tag{3}$$

式中：$[\cdot,\cdot]$表示拼接操作；F_1为1×1卷积层和BN层操作；δ为ReLu非线性激活函数。$\boldsymbol{f}\in\boldsymbol{R}^{C/r\times(H+W)}$是对输入特征图沿水平和垂直方向的空间信息编码后的中间特征图，r为限制计算复杂性而设置的超参数。

再将中间特征图$\boldsymbol{f}$沿水平方向和垂直方向分为两个单独的张量，分别利用1×1卷积操作，使这两个张量的通道数和输入特征图的通道数相同。两个单独的张量再通过Sigmoid激活函数处理，求出注意力权值，如式(4)(5)所示：

$$q^{\mathrm{w}}=\sigma\left(\boldsymbol{F}_{\mathrm{w}}\left(\boldsymbol{f}^{\mathrm{w}}\right)\right) \tag{4}$$

$$q^{\mathrm{h}}=\sigma\left(\boldsymbol{F}_{\mathrm{h}}\left(\boldsymbol{f}^{\mathrm{h}}\right)\right) \tag{5}$$

式中：$\boldsymbol{f}^{\mathrm{w}}\in\boldsymbol{R}^{C/r\times W\times H}$和$\boldsymbol{f}^{\mathrm{h}}\in\boldsymbol{R}^{C/r\times W\times H}$表示两个单独的张量；$\sigma$为Sigmoid激活函数；$\boldsymbol{F}_{\mathrm{w}}$和$\boldsymbol{F}_{\mathrm{h}}$分别表示沿水平和垂直方向的1×1卷积操作；$q^{\mathrm{w}}$和$q^{\mathrm{h}}$分别为沿水平和垂直方向求出的注意力权值。

最后的输出特征张量：

$$\boldsymbol{y}=q^{\mathrm{w}}(j)\times q^{\mathrm{h}}(i)\times\boldsymbol{x}(i,j),\ \boldsymbol{y}\in\boldsymbol{Y}^{W\times H\times C} \tag{6}$$

1.3 深度优化

过多的卷积操作可能导致估计的初始深度图过于平滑，使重建后的模型边缘出现缺损。因此采用Xu等[21]提出的图像细化方法，将三通道参考图像和初始深度图进行拼接后的张量，输入到残差学习网络。既可以让参考图像作为引导信息细化从概率体中估计出的初始深度图，又融合了特征的多样性。残差学习网络如图3所示，由4个卷积核大小均为3×3、步长为1的卷积层，BN层和ReLu激活函数组成，每层输出通道分别为32、32、32、1，得到的残差值和初始深度图作相加操作输出最终深度图。

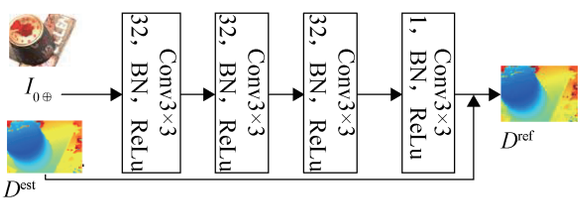

图3 残差学习网络
Fig. 3 Residual learning network

1.4 损失函数

本文算法将多视图深度估计视为一项分类任务，每层多视图深度估计任务采用focal loss[22]作为损失函数：

$$L=\sum_{p\in\{p_v\}}-(1-P^{(\tilde{d})}(p))^{\gamma}\ln(P^{(\tilde{d})}(p)) \tag{7}$$

式中：p_v为有真值的像素子集；$\left(1-P^{(\tilde{d})}(p)\right)^{\gamma}$为focal loss损失函数在不同复杂程度区域调节因子；$p^{(\tilde{d})}(p)$为像素p在深度假设d处的估计概率，$\tilde{d}$为最接近真值的深度值；$\gamma=2$为参数。当估计深度越准确时$\left(1-P^{(\tilde{d})}(p)\right)$越小则损失权重越小，深度估计结果越差的区域$\left(1-P^{(\tilde{d})}(p)\right)$越大则损失权重越大。

具有M层级联结构的多视图深度估计网络，总损失为

$$L_{\mathrm{t}}=\sum_{m=1}^{M}\lambda^m\cdot L^m\sqrt{a^2+b^2} \tag{8}$$

式中：L_m为第m层的损失；λ_m为其相应的损失权重。

综上所述，本文算法首先通过改进后的FPN

网络和ASFE模块得到具有全局上下文信息和位置信息的多尺度特征图，再利用三层级联结构的深度求精模块输出初始深度图，为防止初始深度图过平滑，采用深度优化模块得到最终深度图。最后为了提高网络模型整体性能，构建focal loss函数作为训练损失函数。

2 实验结果与分析

2.1 数据集与评价指标

2.1.1 数据集

采用DTU[23]数据集和Tanks and Temples[24]数据集对本文算法进行训练和测试。DTU数据集属于多视图深度估计的经典数据集，包括124个不同的场景，每个场景图像用7种不同环境光照，通过49或64个相机位置围绕拍摄获取得到分辨率为1 600×1 200的不同视角图像。DTU数据集总共有42 532张图像，其中训练集，验证集和测试集分别有34 000，4 266和4 266张。数据集中的相机参数由Matlab工具箱精准计算得到。

为了检测DTU数据集训练后的网络模型泛化能力，本文算法在Tanks and Temples复杂室外场景的数据集上进行测试。Tanks and Temples数据集是在任意角度和不同的光照条件下拍摄的目标场景图像，其中测试集有2 000张图像，常被用来测试模型的泛化能力。

2.1.2 评价指标

DTU Benchmark中提供Matlab评价代码[25]，计算准确度误差(accuracy error, Acc)、完整度误差(completeness error, Comp)和整体精度误差(overall accuracy error, Overall)。准确度误差和完整度误差利用预测点云结果和真值点云相互投影后之间的绝对距离来计算，整体精度误差是准确度误差和完整度误差的平均值。DTU数据集3个评价指标结果越低表示算法生成点云质量越高。

Tanks and Temples数据集评价方式是由Tanks and Temples Benchmark给出每个场景点云模型的得分和整体性能指标Mean。其中Mean是所有场景得分的平均分。

2.2 实验设置

本网络输入分辨率为864×1 152的图像、相邻匹配图像数量为5、每层深度平面个数为[48，32，8]、深度间隔1.06及每层深度假设比率为[4.0，1.0，0.5]、Batch Size为1、训练轮数为16轮，每层损失权重为[1.0，1.0，1.0]。实验使用的服务器环境为：Intel(R) Xeon(R) CPU E5-2683 v4处理器，32 G内存，NVIDIA RTX 2080Ti×4，Ubuntu20.04操作系统。基于python3.8，pytorch1.9.1，CUDA 10.2的虚拟环境。

2.3 实验分析

2.3.1 消融实验

本文算法在CasMVSNet算法基础上进行改进，为了验证本文算法中CA、DCN、focal loss模块的有效性，在DTU数据集进行了消融实验，结果如表2所示。加入CA和DCN模块后，模型1和模型2的Comp评价指标结果明显下降。本文算法和CasMVSNet相比，Acc、Comp、Overall三个评价指标上均优于CasMVSNet结果，证明了每个模块的有效性。

2.3.2 DTU对比分析

本文算法在DTU数据集上和其他多视图深度估计算法进行对比如表3所示。评价指标值均以毫米为单位(越低越好)，加粗和下划线表示最优和次优结果。可以看出本文算法和其他算法相比在Comp和Overall两指标上取得最优，其余指标均为次优。在Overall指标上本文算法比Acc指标最高的Gipuma[1]、显存资源占用最少的R-MVSNet[8]、运行速度最快的HighRes-MVSNet[11]网络分别提升了47.23%、27.73%、18.23%，证明了本文算法在DTU数据集上定量结果的真实可靠，且模型达到了占用显存小训练速度快的目标。

DOI: 10.16182/j.issn1004731x.joss.23-0112

2.3.3 三维点云模型效果对比

本文算法在DTU的测试数据集上得到全部深度图后，采用Fusibile工具箱通过光度过滤、几何一致性过滤和深度融合3个步骤生成密集点云。为了对比不同算法生成点云效果的差异，本文算法和R-MVSNet[8]、D2HC-MVSNet[9]、真值点云的重建后可视化效果对比如图4所示。在DTU数据集的Scan9中复杂的橱窗部分(放大红色框)和Scan77中边缘手柄的位置(放大绿色框)的点云模型完整度上可以看出其他算法都出现了相对更多的缺损，相比之下本文算法重建的点云模型更加完整。

表2 DTU数据集上消融实验(越低越好)

Table 2 Ablation experiment on DTU dataset (lower value is better)

算法模型	CA	DCN	focal loss	Acc/mm	Comp/mm	Overall/mm
CasMVSNet[10]	×	×	×	0.325	0.385	0.355
模型1	√	×	×	0.343	0.320	0.331
模型2	√	√	×	0.329	0.309	0.319
Ours	√	√	√	0.323	0.287	0.305

表3 不同算法在DTU数据集上比较结果(越低越好)

Table 3 Comparison results of different algorithms on DTU dataset (lower value is better)

算法模型	Acc/mm	Comp/mm	Overall/mm	GPU /MB	Run-time/s
Gipuma[1]	**0.283**	0.873	0.578	—	—
Surfacenet[5]	0.450	1.040	0.745	—	—
MVSNet[7]	0.396	0.527	0.462	22 511	1.210
R-MVSNet[8]	0.385	0.459	0.422	**6 915**	1.28
D2HC-RMVSNet[9]	0.395	0.378	0.386	13 946	2.6
CasMVSNet[10]	0.325	0.385	0.355	9 891	0.492
HighRes-MVSNet[11]	0.354	0.393	0.373	1 119	**0.10**
EPM-RMVSNet[12]	0.468	0.521	0.495	—	—
AACVP-MVSNet[13]	0.357	0.326	0.341	1 048	—
MCV-MVSNet[14]	0.353	0.357	0.355	21 400	3.1
Ours	0.323	**0.287**	**0.305**	9 434	0.137

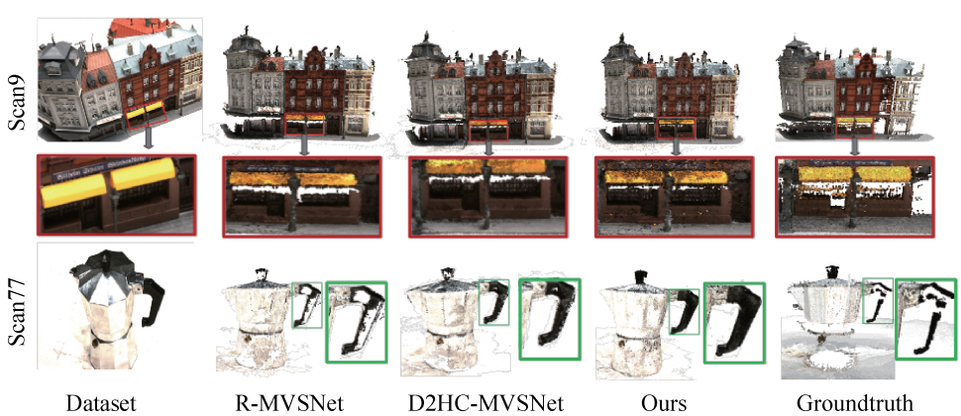

图4 DTU数据集上点云对比结果

Fig. 4 Point cloud comparison results on DTU dataset

2.3.4 Tanks and Temples对比分析

为了测试模型的泛化能力，在Tanks and Temples数据集上和其他算法进行比较如表4所示。评价指标是每个场景的得分和Mean(越高越好)，加粗体数字表示最佳得分、下划线数字表示次优得分。本文算法在Tanks and Temples数据集中的6个场景和Mean上均获得了最高的得分，证明了本文算法的有效性。其中仅在具有高频重复纹理的Panther场景和存在细杆结构的P.G.(Playground)点云场景得分比D2HC-RMVSNet[9]分别低了1.44%和2.69%。在复杂场景“马”上的精度点云结果如图5所示，τ是官方确定的场景相关距离阈值，较暗的点表示与τ相关的较大异常点。本文算法与CasMVSNet[10]网络、D2HC-RMVSNet网络进行比较，在马前腿的边缘和腹部生成较少的深色异常点，重建出的精确点云更多，视觉效果更好。

表4 不同算法在Tanks and Temples数据集对比结果(越高越好)

Table 4 Comparison results of different algorithms on Tanks and Temples benchmark(higher is better)

算法模型	Mean	Family	France	Horse	L.H	M60	Panther	P.G.	Train
MVSNet[7]	43.48	55.99	28.55	25.07	50.79	53.96	50.86	47.90	34.69
R-MVSNet[8]	48.40	69.96	46.65	32.59	42.95	51.88	48.80	52.00	42.38
D2HC-RMVSNet[9]	59.20	74.69	56.04	49.42	60.08	59.81	**59.61**	**60.04**	53.92
CasMVSNet[10]	56.84	76.37	58.45	46.26	55.81	56.11	4.06	58.18	49.51
HighRes-MVSNet[11]	49.81	66.62	44.17	30.84	55.13	53.20	50.32	55.45	42.73
Ours	**61.43**	**78.74**	**64.79**	**53.37**	**60.31**	**61.86**	58.75	58.42	**55.20**

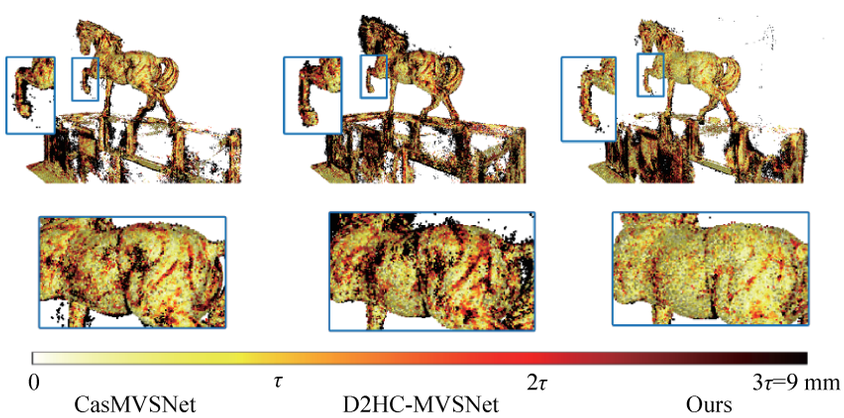

图5 在Tanks and Temples数据集中马场景的精度点云结果

Fig. 5 Point cloud precision results of horse scene on Tanks and Temples dataset

3 结论

针对基于深度学习的多视图深度估计算法中，缺乏全局的特征信息和多视图深度估计模型整体性能有待提高的现状，本文提出一种基于自适应空间特征增强的多视图深度估计算法，在CasMVSNet中加入由可分变形卷积和坐标注意力构成的自适应空间特征增强模块，获取到具有全局上下文信息的有用特征提高模型的准确性和完整性，在深度优化阶段加入残差学习网络减少过多卷积操作对深度图的影响，采用focal loss函数训练提高多视图深度估计模型的判断力和整体性能。本文算法在DTU数据集和Tanks and Temples数据集上的结果优于CasMVSNet算法，获取到比其他算法更好重建效果的同时还减少了对显存资源的占用。本文算法对推断深度假设的数量、深度间隔和深度间隔的衰减因子等超参数敏感，未来需要进一步改进网络结构减少对超参数敏感。

https://dc-china-simulation.researchcommons.org/journal/vol36/iss1/8
DOI: 10.16182/j.issn1004731x.joss.23-0112

参考文献：

[1] Silvano Galliani, Katrin Lasinger, Konrad Schindler. Massively Parallel Multiview Stereopsis by Surface Normal Diffusion[C]//Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV). Piscataway, NJ, USA: IEEE, 2015: 873-881.

[2] Jure Žbontar, LeCun Y. Stereo Matching by Training a Convolutional Neural Network to Compare Image Patches[J]. Journal of Machine Learning Research, 2016, 17(1): 2287-2318.

[3] Luo Wenjie, Alexander G Schwing, Raquel Urtasun. Efficient Deep Learning for Stereo Matching[C]//2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway, NJ, USA: IEEE, 2016: 5695-5703.

[4] 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). Piscataway, NJ, USA: IEEE, 2017: 66-75.

[5] Ji Mengqi, Juergen Gall, Zheng Haitian, et al. SurfaceNet: An End-to-end 3D Neural Network for Multiview Stereopsis[C]//2017 IEEE International Conference on Computer Vision (ICCV). Piscataway, NJ, USA: IEEE, 2017: 2326-2334.

[6] Kar A, Christian Häne, Malik J. Learning a Multi-view Stereo Machine[C]//Proceedings of the 31st International Conference on Neural Information Processing Systems. Red Hook, NY, USA: Curran Associates Inc., 2017: 364-375.

[7] Yao Yao, Luo Zixin, Li Shiwei, et al. MVSNet: Depth Inference for Unstructured Multi-view Stereo[C]// Computer Vision-ECCV 2018. Cham: Springer International Publishing, 2018: 785-801.

[8] Yao Yao, Luo Zixin, Li Shiwei, et al. Recurrent MVSNet for High-resolution Multi-view Stereo Depth Inference [C]//2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway, NJ, USA: IEEE, 2019: 5520-5529.

[9] Yan Jianfeng, Wei Zizhuang, Yi Hongwei, et al. Dense Hybrid Recurrent Multi-view Stereo Net with Dynamic Consistency Checking[C]//Computer Vision-ECCV 2020. Cham: Springer International Publishing, 2020: 674-689.

[10] Gu Xiaodong, Fan Zhiwen, Zhu Siyu, et al. Cascade Cost Volume for High-resolution Multi-view Stereo and Stereo Matching[C]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway, NJ, USA: IEEE, 2020: 2492-2501.

[11] Rafael Weilharter, Friedrich Fraundorfer. HighRes-MVSNet: A Fast Multi-view Stereo Network for Dense 3D Reconstruction from High-resolution Images[J]. IEEE Access, 2021, 9: 11306-11315.

[12] 叶春凯, 万旺根. 基于特征金字塔网络的多视图深度估计[J]. 电子测量技术, 2020, 43(11): 91-95.
Ye Chunkai, Wan Wanggen. Feature Pyramid Network for Multi-view Depth Estimation[J]. Electronic Measurement Technology, 2020, 43(11): 91-95.

[13] Yu Anzhu, Guo Wenyue, Liu Bing, et al. Attention Aware Cost Volume Pyramid Based Multi-view Stereo Network for 3D Reconstruction[J]. ISPRS Journal of Photogrammetry and Remote Sensing, 2021, 175: 448-460.

[14] 刘万军, 王俊恺, 曲海成. 多尺度代价体信息共享的多视角立体重建网络[J]. 中国图象图形学报, 2022, 27 (11): 3331-3342.
Liu Wanjun, Wang Junkai, Qu Haicheng. Multi-scale Cost Volumes Information Sharing Based Multi-view Stereo Reconstructed Model[J]. Journal of Image and Graphics, 2022, 27(11): 3331-3342.

[15] Lin T Y, Piotr Dollár, Girshick R, et al. Feature Pyramid Networks for Object Detection[C]//2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway, NJ, USA: IEEE, 2017: 936-944.

[16] Olaf Ronneberger, Philipp Fischer, Thomas Brox. U-Net: Convolutional Networks for Biomedical Image Segmentation[C]//Medical Image Computing and Computer-Assisted Intervention-MICCAI 2015. Cham: Springer International Publishing, 2015: 234-241.

[17] Samuel Rota Bulò, Lorenzo Porzi, Peter Kontschieder. In-place Activated BatchNorm for Memory-optimized Training of DNNs[C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway, NJ, USA: IEEE, 2018: 5639-5647.

[18] Dai Jifeng, Qi Haozhi, Xiong Yuwen, et al. Deformable Convolutional Networks[C]//2017 IEEE International Conference on Computer Vision (ICCV). Piscataway, NJ, USA: IEEE, 2017: 764-773.

[19] Hou Qibin, Zhou Daquan, Feng Jiashi. Coordinate Attention for Efficient Mobile Network Design[C]//2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway, NJ, USA: IEEE, 2021: 13708-13717.

[20] Hu Jie, Shen Li, Sun Gang. Squeeze-and-excitation Networks[C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway, NJ, USA: IEEE, 2018: 7132-7141.

[21] Xu Ning, Price B, Cohen S, et al. Deep Image Matting [C]//2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway, NJ, USA: IEEE, 2017: 311-320.

[22] Lin T Y, Goyal P, Girshick R, et al. Focal Loss for Dense Object Detection[C]//2017 IEEE International Conference on Computer Vision (ICCV). Piscataway, NJ, USA: IEEE, 2017: 2999-3007.

[23] Aanæs H, Jensen R R, Vogiatzis G, et al. Large-scale Data for Multiple-view Stereopsis[J]. International Journal of Computer Vision, 2016, 120(2): 153-168.

[24] Knapitsch A, Park J, Zhou Qianyi, et al. Tanks and Temples: Benchmarking Large-scale Scene Reconstruction[J]. ACM Transactions on Graphics, 2017, 36(4): 78.

[25] Seitz S M, Curless B, Diebel J, et al. A Comparison and Evaluation of Multi-view Stereo Reconstruction Algorithms[C]//2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06). Piscataway, NJ, USA: IEEE, 2006: 519-528.

DOI: 10.16182/j.issn1004731x.joss.23-0112