doi:10.3969/j.issn.1672-1292.2023.01.007

基于胶囊卷积网络的多视图三维重建

胡荣林,付浩志,何旭琴,张新新,陆文豪

(淮阴工学院计算机与软件工程学院,江苏 淮安 223003)

[**摘要**] 从深度神经网络对重建效果影响的角度,提出了基于胶囊卷积网络的多视图三维重建模型 Caps-MVSNet,包括特征提取、构建代价体、代价体正则化、回归深度图和细化深度图 5 个阶段. 提出了 FENet-T 特征提取网络和 3D-CapsCNN 网络,并分别应用于模型的特征提取阶段和代价体正则化阶段. 其中,FENet-T 利用高效的 Block 计数比率以及大尺度空洞卷积和分组卷积提高网络的特征提取效率. 3D-CapsCNN 使用比卷积神经网络更强空间表示能力的 3D 胶囊网络来正则化代价体. Caps-MVSNet 在 DTU 数据集上完成了效果测试,结果表明,与先前主流重建方法相比该模型在完整性上达到了最优结果,在准确性、整体性上均取得较大提升. 另外,与基准模型 MVSNet 相比,该模型在准确性、整体性和完整性上分别提高 3.3%、4.9%和 8.2%,参数量减少 3.3%.

[**关键词**] 特征提取网络,3D 胶囊网络,空洞卷积,分组卷积,多视立体匹配
[**中图分类号**]TP391　[**文献标志码**]A　[**文章编号**]1672-1292(2023)01-0046-10

Multi-View 3D Reconstruction Based on Capsule Convolution Network

Hu Ronglin, Fu Haozhi, He Xuqin, Zhang Xinxin, Lu Wenhao

(Faculty of Computer and Software Engineering, Huaiyin Institute of Technology, Huai'an 223003, China)

Abstract: By exploring the influence of deep neural networks on the reconstruction effect, the paper proposes a multi-view 3D reconstruction model Caps-MVSNet based on a capsule convolutional network. Caps-MVSNet includes five stages: feature extraction, construction cost volume, cost volume regularization, regression depth map and refinement depth map. This paper focuses on the FENet-T feature extraction network and the 3D-CapsCNN network, which are used for the feature extraction stage and the cost volume regularization stage of the model, respectively. Among which, FENet-T uses an efficient block counting ratio, large-scale dilated convolutions and group convolutions to improve the feature extraction efficiency of the network. 3D-CapsCNN uses 3D capsule networks with a stronger spatial representation than convolutional neural networks to regularize the cost volume. Caps-MVSNet has completed the effect test with the DTU datasets. The results show that compared with the previous mainstream reconstruction methods (Colmap, Tola, Camp, Gipuma, Furu, SurfaceNet), the model proposed by this study achieves the optimum of the current reconstruction method in terms of integrity, and significantly improves the accuracy and completeness. Furthermore, it shows that compared to the model of MVSNet as benchmark, the accuracy, completeness and overall of the proposed model are improved by 3.3%, 4.9% and 8.2%, respectively, the number of parameters is reduced by 3.3%.

Key words: feature extraction network, 3D capsules network, dilated convolution, group convolution, multi-view stereo matching

从 RGB 图像中重建三维几何图形是一个经典的计算机视觉问题. 基于多视角的三维重建算法近年来应用于诸多领域,如:物体识别、医疗诊断、机器人导航、场景理解和文物修复等.

多视图立体[1](multi-view stereo,MVS)是多视角三维重建算法的重要组成部分,从二维图像中恢复丢失的维度是经典 MVS 算法的目标. 经典 MVS 方法注重于从数学的角度理解和形式化 3D 到 2D 的投影过程. 从略有不同的视角中捕获图像,并在图像间进行特征匹配,然后通过几何原理恢复图像像素的 3D 坐标[2]. 这些方法在准确度方面已取得了优异成果,但过渡依赖于精良设备. 若要实现高质量的三维重建,就必须依赖标定良好的高精度相机拍摄的高质量图像. 然而,在实际应用中并未能提供高质量图像. 弱光

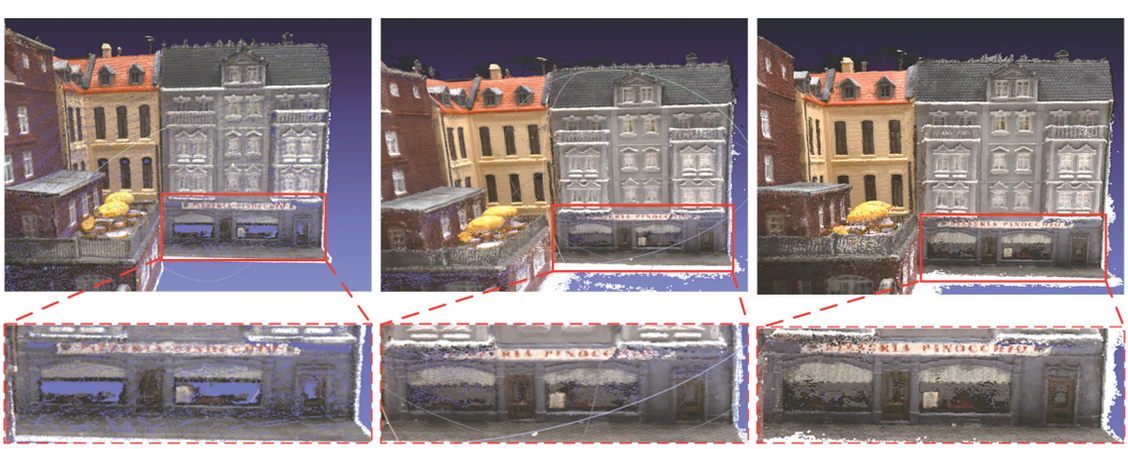
收稿日期:2022-09-15.
基金项目:江苏省研究生实践创新计划项目(SJCX22-1676).
通讯作者:胡荣林,博士,副教授,研究方向:人机交互技术. E-mail:huronglin@ hyit.edu.cn

照条件,目标外形较复杂,存在遮挡或物体表面纹理特征较弱都会影响图像质量进而影响重建效果.

基于深度卷积网络(Convolutional network,CNN)的 MVS 方法可以很好地解决上述问题. 实际上,由于 CNN 强大的特征提取能力[3],不仅提高了匹配精度也提高了重建速度和效果. 因此,将 CNN 应用到 MVS 重建算法上是三维重建的必然发展趋势.

当前,基于深度学习的 MVS 重建算法主要包括 3 种方法:基于点云表示的重建方法[4-5]、基于体素表示的重建方法[6]以及基于深度图表示的重建方法. 其中,基于点云表示的方法直接对点云进行处理生成稀疏点云,然后通过聚类局部和全局特征不断细化稀疏点云得到稠密点云,该方法存在的问题是它们不是规则结构,不适合利用空间规律性的卷积结构. 基于体素表示的方法直接从二维图像回归出 3D 体素网格[7],其主要问题在于这些方法受到计算复杂度和内存的限制,导致产生的体素网格分辨率较低. 由于上述两种表示方法存在低效性[8],基于深度图的表示方法不直接从二维图像重建三维物体,而是将深度图作为中间步骤[9],先对一组图像进行推断得到深度图,再根据深度图构建三维点云模型. 最新 MVS 基准测试显示[10],将深度图作为中间层的表示方法更加高效. 以此方法衍生出的端到端网络模型可以直接从输入的图像实现深度图推断,如 Yao 等[11]提出的 MVSNet、Yu 等[12]提出的 Fast-MVSNet,网络的准确性虽然得到了验证,但其均通过 3D-CNN 回归生成深度图,不仅占用大量内存,且限制了深度图的分辨率.

从视觉角度来看,相比于 3D-CNN 方法,胶囊网络[13-14]不仅有更强的空间表示能力,也有很强的逆渲染性能,即可根据图像反推出图像中物体的空间几何、位姿信息. 为了提高模型的重建效率和重建效果,本文提出了一种基于分组卷积和 3D 胶囊网络进行三维重建的端到端监督学习模型. 该模型实现深度推断的过程如下:(1)提取深度视觉图像特征. 从一组图像提取深度视觉特征,在特征提取网络中使用分组卷积以减少计算量提高计算效率,同时采用大卷积核提高感受野,增加全局特征,从而提高深度推断中远程依赖关系. (2)构建平面扫描代价体. 首先利用可微分单应性变换将源图像投影到参考视角视锥平面上,计算匹配代价并建立代价体. (3)正则化匹配代价体和深度图回归与细化. 首先利用 3D-CapsCNN 模型对代价体进行代价聚合,然后提取全局代价信息以及相邻像素间的依赖关系,最后回归细化得到最终深度图. 在本文设备上实现的 Colmap、MVSNet 以及本文模型的重建效果对比效果如图 1 所示.



(a) Colmap重建效果　(b) MVSNet重建效果　(c) Caps-MVSNet(本文)重建效果

图 1　Colmap、MVSNet、Caps-MVSNet 重建效果对比

Fig. 1　Comparison of reconstruction effects of Colmap, MVSNet, Caps-MVSNet

1　相关工作

三维重建方法可概述为经典 MVS 重建方法和基于深度学习的重建方法,后者首先从提取的视觉特征中构建匹配代价体,然后正则化代价体,接着回归出深度图.

1.1　MVS 三维重建方法

经典的 MVS 重建方法通过视图一致性和光学一致性原则来实现深度场景表示,具有代表性的方法有 Colmap、OpenMVG 等开源方法. Colmap 是 Schonberger 等[15]在 2016 年提出的开源三维重建系统,该方法以 NCC 作为图像间光学一致性的测度,并利用轮廓特征匹配(Patch-based match)[16]进行深度传递并使用

GEM 算法做进一步深度图优化. 在增量式三维重建算法的基础上提出了一种新的 SFM[17] 技术并取得了巨大进步. 由于图像冗余特性不可避免,对于比较稠密的场景,Colmap 在深度图完整性和连续性方面会存在一些问题,推断出的深度图存在较多漏洞,并且该系统运行速度过慢.

Moulon 等[18] 提出的 OpenMVG 是一个关于三维重建的开源库,该开源软件封装了丰富的重建算法并且在摄影测量、计算机视觉和机器人领域有重要应用. OpenMVG 虽然能够准确计算多视立体几何位姿且模型算法稳定,但是开源库中个别模块函数嵌套冗杂、灵活性低,且缺乏大规模 SFM 处理算法,Cernea 提出了开源库 OpenMVS[19] 以解决 OpenMVG 存在的摄影链流问题,旨在提供一套完整算法恢复待重建物体的完整表面. 但 OpenMVS 依然存在弱纹理表面重建效果准确率差的问题.

随着深度学习的不断发展,研究人员将深度学习方法应用于三维重建算法中. 早期研究[20-21] 并没有形成一种端到端的学习方式,只是将深度学习应用于重建算法其中的一个步骤,例如:Han 等[20] 首先提出利用深度网络来实现两个图像的匹配,GCNet[21] 利用 3D-CNN 来正则化代价体,这些方法虽然解决了一些问题,但算法执行比较复杂.

早期端到端的三维重建网络[22-23] 直接通过学习图像一致性和表面结构几何关系进行建模,但建模过程较长,重建完整度不高. DeepMVS[24] 首先构造匹配代价体,然后利用 Encoder-Decoder 结构来推断视差图,由于 DeepMVS 基于 patch 进行特征匹配,在视差估计过程中存在缺失全局语义信息的问题. MVSNet[11] 通过从图像中提取的特征图构造代价体,并利用 3D-CNN 正则化和回归代价体实现深度估计,这使得模型参数量巨大. 另外,由于特征图和代价体尺度固定导致推断出的深度图较为粗糙,影响重建效果. 对于 MVSNet 存在的问题,Yao 等[25] 提出的 RMVSNet 将 3D-CNN 替换成 GRU 来降低模型参数量,精度较 MVSNet 有所降低且训练时间变长,MVSCRF[26] 在代价体正则化阶段利用条件随机场(CRFs)约束深度映射的平滑性. Cascade-MVSNet[27] 利用标准特征金字塔编码几何构建代价体并回归出不同尺度的深度图,以由粗到细的形式生成高分辨率深度图. CVP-MVSNet[28] 与 Point-MVSNet[29] 类似,都是以由粗到细的方式构建代价金字塔,这极大的减少了模型的参数量. Fast-MVSNet[12] 着眼于提高重建效率,该网络通过小规模神经网络编码局部像素间依赖关系并利用高斯-牛顿层来进一步优化深度映射. 由于 MVSNet[11] 中构建代价体的方法非常适合研究特征提取网络和正则化网络对模型整体重建效果的影响,因此,本文采用 MVSNet 中的可微分单应性变换方法来构建代价体.

1.2　视觉特征提取

目前大多数重建方法[11,23-29] 采用经典 CNN 作为特征提取阶段的主干网络,如 ResNet[30]、VGG[31]、U-Net[32] 等. 然而,不同计算机视觉任务对特征提取的要求是不同的,例如:对于图像分类任务,提取全局特征很重要. 对于目标检测任务,提取局部特征很重要. 对于 MVS 三维重建任务,如果是重建目标表面纹理丰富的高频区域,则期望更小的局部感受野. 若表面纹理特征较弱,则需要较大范围局部信息.

鉴于三维重建任务的特殊要求,Liu 等[33] 提出的 ConvNeXt 网络能够很好地适应该要求. ConvNeXt 既在准确性和扩展性方面超越 Vision-Transformer[34],又保持了标准 CNN 简单性和高效性. 在特征提取模块设计上采用 ConNeXt 提出的方法,如采用 1∶1∶3∶1 的 Block 计数比率以及利用更大的卷积核提高感受野的思想. 考虑到特征提取网络轻量化问题,与 ConvNeXt 不同,本文采用大尺度空洞卷积提取图像特征.

1.3　代价体正则化

目前,基于深度学习的重建模型之间的主要区别在于正则化代价体的方式不同,主要三种方式:3D-CNN、RNN 和 Coarse-to-Fine. 许多模型[11,24,27,29] 均采用 3D-CNN 来正则化代价体. 由于 3D-CNN 计算成本高,内存消耗大,一些尝试通过将 3D-CNN 替换为 RNN 来线程化所有深度假设或者采用 Coarse-to-Fine 的策略[35-36]. 相比于利用 RNN 实现正则化的方法,Coarse-to-Fine 能够适应更精细的深度预测,同时它们都减少了内存消耗.

Hinton 等[14] 根据人类对图像的认知方式提出的胶囊网络(Capsules),解决了 CNN 在提取图像特征过程中丢失大量语义信息的问题. 一些尝试将 Capsules 应用于处理具有空间问题的计算机视觉任务并取得了不错的效果,Tran 等[37] 将 Capsule 应用于医疗图像立体分割. Zhao 等[38] 通过 3D 点胶囊网络处理稀疏的三维点云. 但是,MVS 重建算法中还未出现利用 3D 胶囊网络来正则化代价体的相关研究. 由于 Capsules 中独特的动态路由算法不仅具备计算机视觉中的注意力机制特性[39],而且具有类似 RNN 中的时

序特性,另外,Capsules 还具有保留更多图像内空间位置信息的能力. 因此,本文将利用 3D 胶囊网络正则化代价体并实现像素级的深度推断.

2 重建方法

本节主要论述 Caps-MVSNet 模型结构,包括深度视觉图像特征提取、建立匹配代价体、深度图推断以及深度图细化,网络模型如图 2 所示.

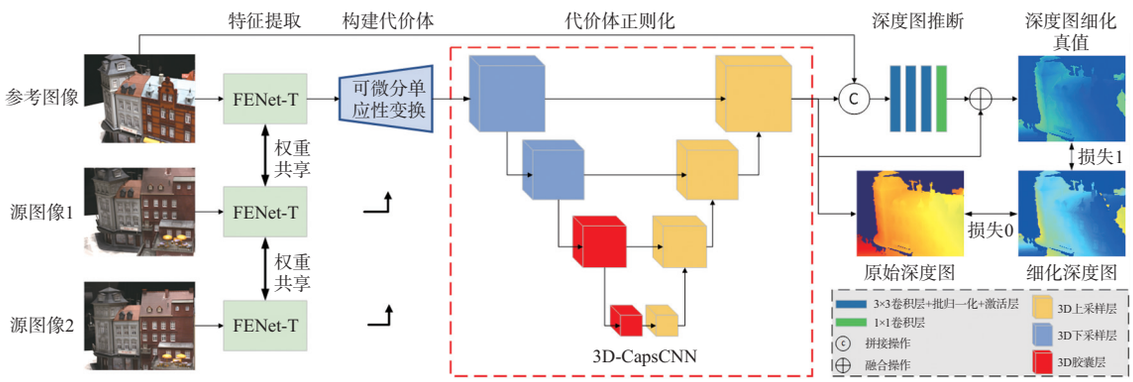


图 2 网络整体框架

Fig. 2 The overall framework of the network

2.1 深度视觉图像特征提取网络

受 ConvNeXt[33] 启发,本文提出 FENet-T 网络. 与 MVSNet[11] 的特征提取模块完全不同,FENet-T 在结构设计上由 1 个 stem 层、3 个降采样层以及 4 个 block 层组成. 降采样层与 block 层交替置于 stem 层后,图像信息经特征编码后输出为 32 通道的特征图. 由于采用了比传统卷积网络更大的卷积核,FENet-T 具有更大的感受野,也捕捉了更多的全局信息,从而解决了图像特征在经过降采样层时丢失过多语义信息的问题. 另外,模型中采用了与 MobileNet 相似的分组卷积并行编码语义信息[40],因此,特征提取网络效率更高. FENet-T 在提取 N 幅图像特征时,采用了权重共享策略,其网络结构如图 3 所示.

以往的 CNN 通过在卷积层后添加 BN 层来加速模型收敛和防止过拟合,由于在数据处理过程中会出现协方差偏移的现象并且 BN 过分依赖 mini-batch 的大小,而 LN 层不依赖与 mini-batch 的大小,因此,在模型设计上将 LN 层代替 BN 层. 与 ConvNeXt 相似,本文提出的 FENet-T 相对于 MVSNet[11] 中的特征提取网络(FE)采用了更少的激活函数 GELU. 另外,每个 Block 之间采用与 Swin-Transformer[34] 相同的计数比率,经过消融实验验证,本文最终确定 4 个 Block 数依次为[1,1,3,1],具体实验细节将在后文中描述. 除此以外,为了增加全局语义信息,每个 Block 中均包含一个 7×7 卷积和两个 1×1 卷积,如图 4 所示.

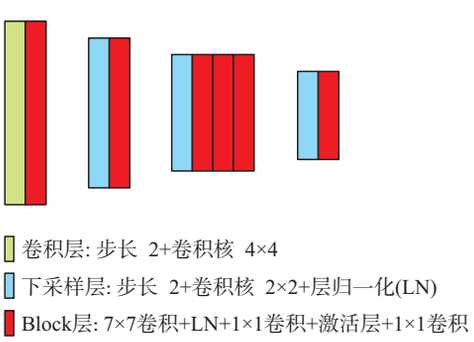


图 3 FENet-T 网络结构图

Fig. 3 The structure network of FENet-T

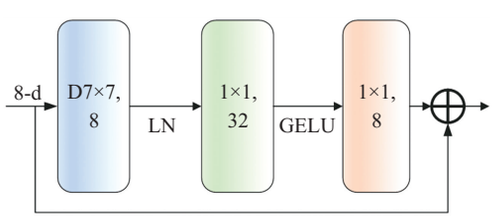


图 4 FENet-T 中 Block 模块

Fig. 4 The block module in FENet-T

如表 1 所示,该模型的输入为 3 通道 RGB 图像,经过 stem 层输出为 8 通道特征图 F_1,将 F_1 输入 Block1 得到 8 通道特征图 F_2,F_2 经下采样 1 和 Block2 后得到 16 通道特征图 F_3,接着 F_3 输入下采样 2 和 Block3 得到特征图 F_4,其中 Block3 包含 3 个 7×7 空洞卷积层. 最后将 F_4 输入下采样 3 和 Block4 输出结果为 32 通道特征图 F_5.

表 1　FE 与 FENet-T 网络对比

Table 1　Comparison of FE and FENet-T

层数	FE	FENet-T	层数	FE	FENet-T
stem	3×3,8,stride 1	4×4,8,stride 4	下采样 2	3×3,16,stride 1	2×2,64,stride 2
Block1	3×3,8,stride 1	$\begin{bmatrix} \text{d7×7} & 8 \\ 1\times1 & 32 \\ 1\times1 & 8 \end{bmatrix}$	Block3	5×5,32,stride 1	$\begin{bmatrix} \text{d7×7} & 64 \\ 1\times1 & 256 \\ 1\times1 & 64 \end{bmatrix}\times3$
下采样 1	5×5,16,stride 2	2×2,16,stride 1	下采样 3	3×3,32,stride 1	2×2,32,stride 1
Block2	3×3,16,stride 1	$\begin{bmatrix} \text{d7×7} & 16 \\ 1\times1 & 32 \\ 1\times1 & 16 \end{bmatrix}$	Block4	3×3,32,stride 1	$\begin{bmatrix} \text{d7×7} & 32 \\ 1\times1 & 128 \\ 1\times1 & 32 \end{bmatrix}$

2.2　构建代价体

通过可微分单应性变换方法来构建代价体. 单应性变换投影与平面扫描算法类似[41]. 不同之处在于文中未直接从图像上采样,而是从特征图上采样. 平面扫描法合并先验在场景中的平面位置,这不仅提高了重建质量也减少了计算时间,尤其针对无纹理表面的物体. 假设一组图像集合 $I_i(i=0,\cdots,N)$,其中 I_0 表示参考图像,$I_i(i=1,\cdots,N)$ 表示一组源图像. 使用 FENet-T 模型从图像集合 $I_i(i=0,\cdots,N)$ 中提取特征得到 32 通道特征图 $F_i(i=0,\cdots,N)$,通过单应性变换将源图像特征图 $F_i(i=1,\cdots,N)$ 依次投影到参考图像所在平面法向量方向上不同深度的平面上. 深度平面的个数按照某一深度间隔 $\nabla \boldsymbol{d}$ 从 $[\boldsymbol{d}_{\min},\boldsymbol{d}_{\max}]$ 依次得到不同深度间隔的相机锥体,深度假设平面个数为:

$$n=\frac{\boldsymbol{d}_{\max}-\boldsymbol{d}_{\min}+1}{\nabla \boldsymbol{d}}. \tag{1}$$

MVSNet 深度假设平面的个数 $n=256$,由于设备内存限制,将深度平面个数设为 $n=192$. 在确定深度假设平面数后,将参考图像特征图 F_0 依次投影到深度假设平面上,并利用插值算法使每幅投影具有相同的高和宽. 通过计算图像一致性,确定源图像在 $[\boldsymbol{d}_{\min},\boldsymbol{d}_{\max}]$ 中对应的深度 $\boldsymbol{d}_i$. 为了保证深度图的平滑性,利用可微的单应性矩阵 $\boldsymbol{H}$,将第 j 个视角特征图投影到深度 $\boldsymbol{d}_j$ 平面上. 单应性矩阵计算公式为:

$$\boldsymbol{H}_i(\boldsymbol{d}_j)=\boldsymbol{K}_i\boldsymbol{R}_i\left(-\left(\frac{\boldsymbol{R}_i^{-1}\boldsymbol{t}_i-\boldsymbol{R}_0^{-1}\boldsymbol{t}_0}{\boldsymbol{d}_j}\right)\boldsymbol{k}^{\mathrm{T}}\boldsymbol{R}_0\right)\boldsymbol{R}_0^{\mathrm{T}}\boldsymbol{K}_0^{\mathrm{T}}. \tag{2}$$

式中,$\boldsymbol{K}_i,\boldsymbol{R}_i(i=0,\cdots,N)$ 分别为对应视角相机的内参矩阵和外参矩阵,$\boldsymbol{k}$ 表示沿参考图像相机主轴方向的法向量,$\boldsymbol{I}$ 为单位矩阵. 定义参考图和源图像的单应性变化分别为:

$$\overline{\boldsymbol{X}}_0=\boldsymbol{H}_0(\boldsymbol{d}_j)\times\boldsymbol{X}_0. \tag{3}$$

2.3　代价体正则化与回归网络

原始代价体中往往存在很多噪声,这导致网络易出现过拟合. 为了进一步过滤代价体中噪声,为了生成用于深度推理的概率体 P,本文利用 3D-CapsCNN 网络正则化代价体. 3D-CapsCNN 的整体架构采用编码器-解码器结构. 其中,在编码器部分使用 3DCNN 网络和 3D 胶囊网络编码语义信息,在解码器结构上,通过 3DCNN 解码深度信息并实现深度推断. 考虑到正则化网络会产生昂贵的计算成本,在编码器中使用 3D 空洞卷积进行降采样,不仅提高了网络对全局代价信息的感受能力,也降低了模型计算量,同时捕捉更多不同尺度的上下文信息. 另外,为了灵活地处理任意数量视图,采用了基于方差的代价指标将多个特征代价融合为一个特征代价体.

如图 5 所示,3D-CapsCNN 主要由代价特征提取模块、代价特征编码器和代价特征解码器 3 个部分组成. 首先,利用代价特征提取模块进行升维操作,将 32 通道的代价体变成 64 通道. 其次,在编码器的前两层设置两个空洞卷积层进行低层级特征编码. 胶囊网络(Capsules)在空间位姿推理时能保留更多的空间信息,相对于 CNN,Capsules 更善于通过捕获局部与整体之间的长程依赖关系来加强特征学习. 故使用两个胶囊网络层编码高层语义特征信息. 胶囊层的输入为 6D 张量,首先压缩输入张量的前两个维度,接着输入胶囊层中的 3D-CNN 网络,然后利用动态路由算法对投票结果进行细化,最后使用 squash 函数进行非线性激活后输出结果,

$$s=\sum c\hat{u}, \tag{4}$$

$$v=\frac{\|s\|^2}{1+\|s\|^2}\cdot\frac{s}{\|s\|}. \tag{5}$$

此外,胶囊层中的每个胶囊都是一个卷积单元,不同胶囊之间共享同一个卷积核. 在特征解码阶段,首先重构特征图的维度,然后将重构后的特征图输入解码器中的上采样层和卷积层,在模型中使用长跳跃连接方式连接编码器和解码器,以保证模型重用模型编码器阶段学习到的图像细粒度细节.

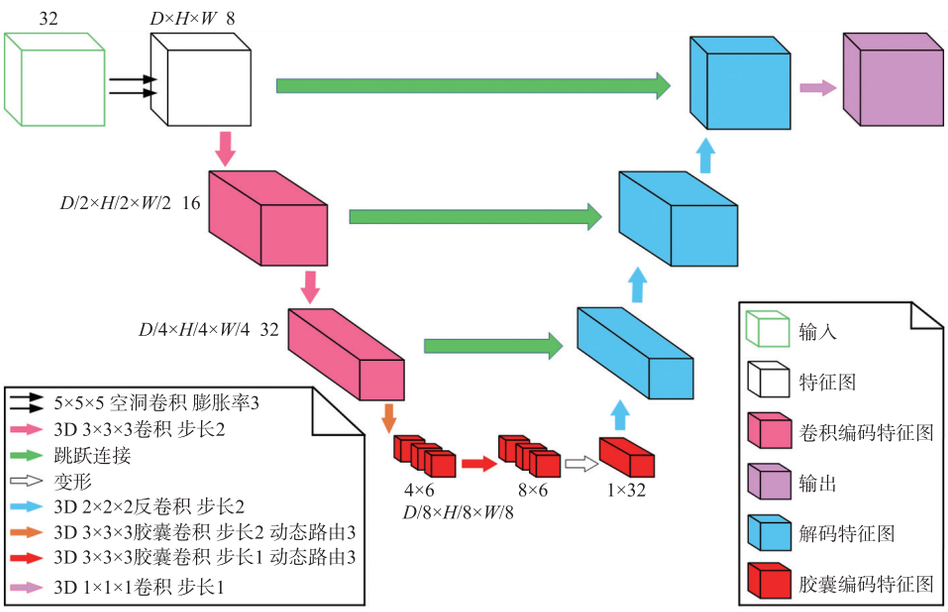


图 5　3D-CapsCNN 网络图

Fig. 5　The network of 3D-CapsCNN

2.4　深度图推断

与以往获取深度图的方式不同,本文模型直接通过神经网络实现深度推断. 实现深度推断最简单的方式是采用“赢家通吃”的策略. 然而,由于“赢家通吃”算法不可微,不能用反向传播进行训练. 另外,该方法不能实现像素级深度推断,还会造成深度图边界不平滑的现象. 采用与 MVSNet 相同的方法,沿着深度方向计算概率体 P 的期望,

$$D=\sum_{d_{\min}}^{d_{\max}} d\times P(d). \tag{6}$$

式中,$P(d)$是在深度 d 上所有像素的置信值.

2.5　深度图细化

初始深度图在重构边界时可能会出现边界过渡平滑的问题,这与图像抠图[42]出现问题类似. 受抠图算法[42]启发,需要对生成的深度图做进一步细化. Caps-MVSNet 使用一个深度细化网络来完成细化任务. 首先,将参考图像与初始深度图拼接为一个通道数为 4 的张量. 然后,输入 3 个 32 通道的 2D-CNN 网络,再输入单通道的卷积层来学习深度残差,从而生成精细化的深度映射. 为了对比实验的公正性,使用了与 MVSNet 相同的损失函数,即 L_1 损失,

$$loss=\sum_{p\in P_{\text{valid}}} d(p)-\hat{d}_i(p)_1+\lambda\cdot d(p)-\hat{d}_r(p)_1. \tag{7}$$

式中,P_{valid}表示有效 Ground Truth 像素集合,$d(p)$表示像素点 p 的 Ground Truth 深度值,$\hat{d}_i(p)$为初始深度估计值,$\hat{d}_r(p)$为精细化深度估计值,系数 λ 设置为 1.

3　实验

3.1　数据集

DTU 数据集[43]是针对 MVS 重建的大型室内场景数据集,共包含 27 097 个训练样本,样本由 124 个不同光照条件的场景组成,场景由 49 个精确的摄像机位置和参考结构光扫描组成,每次扫描包含 7 种不同

光照条件下的 49 幅图像. 该数据集中还包含了法线信息以及不同视角下相机的内参数和外参数,图像分辨率为 1 200×1 600 像素. 为了公平地比较 Caps-MVSNet 与其他重建方法,本文选择与 MVSNet[11] 中相同的训练、验证和测试集.

3.2　实验细节

本文在 DTU 数据集上训练 Caps-MVSNet 网络,并在验证集上进行评估. 采用与 MVSNet[11] 相同的预处理方式,设置视图数分别为 $N=3$,$W=1\ 600$,$H=256$,$D=192$. 其中,视图选择包括 1 幅参考图和 2 幅源图像,共训练 27 097 个样本,测试 7 546 个样本. 设置深度平面假设范围为 $[\boldsymbol{d}_{\min},\boldsymbol{d}_{\max}]=[425\ \text{mm},935\ \text{mm}]$,相邻平面间隔为 $\nabla\boldsymbol{d}=1.06$ mm,同时,选择深度间隔分别为 2 mm、4 mm 和 8 mm 进行深度细化. 本文模型在 Pytorch 上实现,在 1 块 NVIDIA RTX 3090 SUPER 显卡上运行,batch size 设置为 4,初始学习率设置为 0.001,采用 Adam 优化器($\beta_1=0.9$,$\beta_2=0.999$),共训练 16 个 epoch. 在重建质量评估方面采用平均准确率(ACC)、平均完整性(Comp)以及重建整体性(Overall)作为评价指标.

3.3　消融实验

3.3.1　消融实验一

在图像特征提取阶段,设定两组特征提取网络的 Block 数,分别为[1,1,3,1]和[3,3,9,3]. 为了确定网络的最优组合,进行了消融实验和定量分析. 实验保持正则化网络不变,同时改变 FENet-T 中的 Block 数并在 DTU 测试集上进行了 4 组对比测试,测试结果如表 2 所示.

表 2　消融实验一结果对比

Table 2　Comparison of ablation experiments Ⅰ　　mm

方法	Acc	Comp	Overall	模型参数
FENet-T[1,1,3,1]	**0.416**	**0.432**	**0.357**	**327 120**
FENet-T[3,3,9,3]	0.511	0.465	0.439	463 952

表 2 分别展示了 FENet-T 在不同 Block 数时的三维重建评价数据,包括重建准确性误差(Acc)、完整性误差(Comp)以及重建整体性(Overall). 整个过程仅变换 FENet-T,并控制正则化网络(3D-CapsCNN)保持不变. 由表 2 数据可知,Block 数为[1,1,3,1]与 Block 数为[3,3,9,3]的 FENet-T 网络在 Acc、Comp 和 Overall 评价指标上表现相似,但前者模型参数却比后者少 29.4%. 因此,在重建效果相似的前提下,最终选择了更加轻量化的 FENet-T,其 Block 数为[1,1,3,1].

3.3.2　消融实验二

为了进一步证明本文模型有效性,进行了第二组消融实验,用以评估模型总体关键组件的优势. 针对本文中提出的特征提取网络 FENet-T 和正则化网络 3D-CapsCNN,在 DTU 数据集上进行了 4 组消融实验,并使用 DTU 数据集中的评价指标来评估重建效果. 原 MNVNet[11] 论文中的深度平面假设数 $D=256$,由于内存限制,将 MNVNet 论文中的深度平面假设数 $D=256$ 改为 $D=192$,同时使用 Pytorch 版的 MVSNet 进行训练和测试,实验结果如表 3 所示. 其中,FE 表示 MVSNet 中使用的特征提取网络,UNet 表示 MVSNet 中使用的代价体正则化网络.

表 3　消融实验二结果对比

Table 3　Comparison of ablation experiments Ⅱ　　mm

组数	方法	Acc	Comp	Overall	模型参数
1	FE+UNet	0.449	0.481	0.439	338 129
2	FE+3D-CapsCNN	0.455	0.466	0.418	**285 264**
3	FENet-T+UNet	0.422	0.470	0.401	379 985
4	FENet-T+3D-CapsCNN	**0.416**	**0.432**	**0.357**	**327 120**

由表 3 可见,(1)采用更大卷积核以及分组卷积的 FENet-T 性能要比经典卷积 FE 网络好,其重建 Acc、重建 Comp 以及重建 Overall 分别提高了 2.7%、1.1%和 3.8%. 主要原因是利用大卷积核和有效的计算比率提高了 FENet-T 捕捉全局语义信息的能力,进而提高了特征匹配效果. (2)分别对比 1、2 组实验和 3、4 组实验,在控制特征提取网络不变情况下,正则化网络 3D-CapsCNN 比 UNet 重建效果更好. 其中,1、2

组对比实验在 Acc、Comp、Overall 上分别提高 0.6%、1.5%和 2.1%,3、4 组在 Acc、Comp、Overall 上分别提高 0.6%、3.8%和 4.4%. 这充分说明利用 3D 胶囊网络正则化代价体的有效性,也说明胶囊网络中动态路由更新方式解决了卷积网络由于池化操作丢失大量语义信息的问题. 实验结果表明,该方法结合了特征提取网络 FENet-T 和正则化网络 3D-CapsCNN 的优势,实现了更高质量深度推断,得到更加清晰的重建效果,DTU 中不同数据集的重建效果如图 6 所示.

图 6　DTU 数据集中部分数据重建效果图

Fig. 6　Reconstruction of partial data in DTU datasets

3.4　对比实验

通过对比实验进一步验证本文方法的有效性,本文方法与传统重建方法[44-47]和基于深度学习的重建方法[11,23]进行对比实验研究,使用重建 Acc、重建 Comp 以及重建 Overall 进行了评估,对比实验结果如表 4 所示.

可见,本文方法在 Comp 和 Overall 方面取得了最佳结果,与 MVSNet 相比,分别提升了 3.3%、4.9%和 8.2%,模型参数量减少了 3.3%. 与其他方法相比,模型在准确性结果不是最优,但在整体性评价却取得了最佳结果,实验结果充分验证了本文模型的有效性. 图 7 为本文模型在测试数据集上的重建效果图.

表 4　对比实验结果

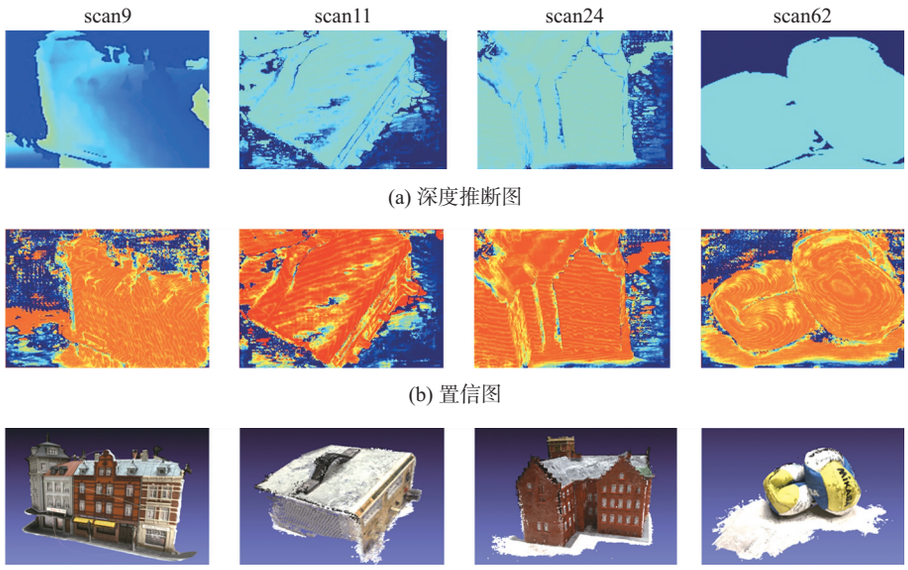
Table 4　Comparative experimental results　　mm

方法	Acc	Comp	Overall	方法	Acc	Comp	Overall
Colmap	0.400	0.664	0.532	SurfaceNet	0.450	1.04	0.745
Tola	0.342	1.190	0.766	MVSNet	0.396	0.527	0.462
Camp	0.835	0.554	0.695	MVSNet *	0.449	0.481	0.439
Gipuma	**0.283**	0.873	0.578	Caps-MVSNet	0.416	**0.432**	**0.357**
Furu	0.613	0.941	0.777				

注:* 表示在本文设备上运行 MVSNet 代码结果.

4　结论

本文提出了面向室内场景的端到端多视角三维重建网络 Caps-MVSNet,包括深度视觉特征提取网络 FENet-T 和代价特征正则化网络 3D-CapsCNN. 实验结果表明,Caps-MVSNet 性能优于现有网络,在准确率、完整性和整体性评价上较 MVSNet 方法分别提高 3.3%、4.9%和 8.2%,达到 41.6%、43.2%和 35.7%,在模型参数量上比 MVSNet 减少 3.3%,充分证明了 FENet-T 和 3D-CapsCNN 的有效性. 可见,本文提出的特征提取网络和正则化网络能够使模型获得更好的重建效果.



(a) 深度推断图

(b) 置信图

(c) 重建图

图 7　部分数据测试效果

Fig. 7　Partial test effect pictures

[参考文献](References)

[1] 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. New York,NY,USA:IEEE,2006,1:519-528.

[2] HARTLEY R,ZISSERMAN A. Multiple view geometry in computer vision[M]. London:Cambridge University Press,2003.

[3] 韩文军,孙小虎,吉根林,等. 基于卷积神经网络的多光谱与全色遥感图像融合算法[J]. 南京师大学报(自然科学版),2021,44(3):123-130.

[4] 杨会君,王瑞萍,王增莹,等. 基于多视角图像的作物果实三维表型重建[J]. 南京师大学报(自然科学版),2021,44(2):92-103.

[5] 张天安,云挺,薛联凤,等. 基于骨架提取的树木主枝干三维重建算法[J]. 南京师范大学学报(工程技术版),2014,14(4):51-57.

[6] WU Z,SONG S,KHOSLA A,et al. 3D shapeNets:A deep representation for volumetric shapes[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Boston,Massachusetts,USA,2015:1912-1920.

[7] TULSIANI S,ZHOU T,EFROS A A,et al. Multi-view supervision for single-view reconstruction via differentiable ray consistency[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Honolulu,Hawaii,USA,2017:2626-2634.

[8] LI K,PHAM T,ZHAN H,et al. Efficient dense point cloud object reconstruction using deformation vector fields[C]//Proceedings of the European Conference on Computer Vision. Munich,Germany,2018:497-513.

[9] SMITH E,FUJIMOTO S,MEGER D. Multi-view silhouette and depth decomposition for high resolution 3D object representation[J]. Advances in Neural Information Processing Systems,2018,31.

[10] ZHU Q,MIN C,WEI Z,et al. Deep learning for multi-view stereo via plane sweep:a survey[J]. arXiv Preprint arXiv:2106.15328,2021.

[11] YAO Y,LUO Z,LI S,et al. Mvsnet:Depth inference for unstructured multi-view stereo[C]//Proceedings of the European Conference on Computer Vision(ECCV). Munich,Germany,2018.

[12] YU Z,GAO S. Fast-Mvsnet:Sparse-to-dense multi-view stereo with learned propagation and gauss-newton refinement[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle,WA,USA,2020.

[13] SABOUR S,FROSST N,HINTON G E. Dynamic routing between capsules[C]//31st Conference on Advances in Neural Information Processing Systems. Long Beach,CA,USA,2017,30.

[14] HINTON G E,SABOUR S,FROSST N. Matrix capsules with EM routing[C]//International conference on learning

representations. Vancouver,Canada,2018.
[15] SCHONBERGER J L,FRAHM J M. Structure-from-motion revisited[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas,Nevada,USA,2016.
[16] 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):24.
[17] ULLMAN S. The interpretation of structure from motion[J]. Proceedings of the Royal Society of London. Series B. Biological Sciences,1979,203(1153):405-426.
[18] MOULON P,MONASSE P,PERROT R,et al. Openmvg:open multiple view geometry[C]//International Workshop on Reproducible Research in Pattern Recognition. Cancún,Mexico,2016:60-74.
[19] GOESELE M,SNAVELY N,CURLESS B,et al. Multi-view stereo for community photo collections[C]//2007 IEEE 11th International Conference on Computer Vision. Rio de Janeiro,Brazil,2007.
[20] HAN X,LEUNG T,JIA Y,et al. Matchnet:Unifying feature and metric learning for patch-based matching[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Boston,MA,USA,2015.
[21] 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. Chengdu,China,2017.
[22] CLARK R,WANG S,WEN H,et al. Vinet:Visual-inertial odometry as a sequence-to-sequence learning problem[C]//Proceedings of the AAAI Conference on Artificial Intelligence. San Francisco,USA,2017.
[23] JI M,GALL J,ZHENG H,et al. Surfacenet:An end-to-end 3D neural network for multiview stereopsis[C]//Proceedings of the IEEE International Conference on Computer Vision. Chengdu,China,2017.
[24] HUANG P H,MATZEN K,KOPF J,et al. Deepmvs:Learning multi-view stereopsis[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City,UT,USA,2018.
[25] YAO Y,LUO Z,LI S,et al. Recurrent mvsnet for high-resolution multi-view stereo depth inference[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach,CA,USA,2019.
[26] XUE Y,CHEN J,WAN W,et al. Mvscrf:Learning multi-view stereo with conditional random fields[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. Seoul,South Kerean,2019.
[27] GU X,FAN Z,ZHU S,et al. Cascade cost volume for high-resolution multi-view stereo and stereo matching[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle,WA,USA,2020.
[28] YANG J,MAO W,ALVAREZ J M,et al. Cost volume pyramid based depth inference for multi-view stereo[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle,WA,USA,2020.
[29] CHEN R,HAN S,XU J,et al. Point-based multi-view stereo network[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. Seoul,South Kerean,2019.
[30] TARG S,ALMEIDA D,LYMAN K. Resnet in resnet:Generalizing residual architectures[J]. arXiv Preprint arXiv:1603.08029,2016.
[31] SIMONYAN K,ZISSERMAN A. Very deep convolutional networks for large-scale image recognition[J]. arXiv Preprint arXiv:1409.1556,2014.
[32] RONNEBERGER O,FISCHER P,BROX T. U-Net:Convolutional networks for biomedical image segmentation[C]//International Conference on Medical Image Computing and Computer-Assisted Intervention. Munich,Germany,2015.
[33] LIU Z,MAO H,WU C Y,et al. A ConvNet for the 2020s[J]. arXiv Preprint arXiv:2201.03545,2022.
[34] LIU Z,LIN Y,CAO Y,et al. Swin transformer:Hierarchical vision transformer using shifted windows[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. Montreal,Canada,2021.
[35] YAN J F,WEI Z Z,YI H W,et al. Dense hybrid recurrent multi-view stereo net with dynamic consistency checking[C]//European Conference on Computer Vision. Springer,Cham,2020:674-689.
[36] CHENG S,XU Z,ZHU S,et al. Deep stereo using adaptive thin volume representation with uncertainty awareness[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle,WA,USA,2020.
[37] TRAN M,VO-HO V K,LE N T H. 3DConvCaps:3DUnet with convolutional capsule encoder for medical image segmentation[J]. arXiv Preprint arXiv:2205.09299,2022.
[38] ZHAO Y,BIRDAL T,DENG H,et al. 3D point capsule networks[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach,CA,USA,2019.

(下转第 92 页)

[7] 倪伟平. 最大度是 6 且不含有弦的小圈的可平面图的边染色[J]. 南京师大学报(自然科学版),2011,34(3):19-24.
[8] 唐保祥,任韩. 几类图完美匹配的数目[J]. 南京师大学报(自然科学版),2010,33(3):1-6.
[9] 陈轩泽,霍静,费峰,等. 基于 PCA 与 ArcGIS 网络分析的图书馆阅览室管理系统[J]. 南京师范大学学报(工程技术版),2012,12(2):57-63.
[10] 李林,封志明,赵品彰,等. 一种基于散射参数测试的人工电源网络校准方法[J]. 南京师范大学学报(工程技术版),2011,11(2):4-8.
[11] 张先迪,李正良. 图论及其应用[M]. 北京:高等教育出版社,2005.
[12] ERDÖS P,RENYI A. On random graphs I[J]. Publicationes Mathematicae,1959(6):290-297.
[13] WATTS D J,STROGATZ S H. Collective dynamics of 'small-world' networks[J]. Nature,1998,393(6684):440-442.
[14] NEWMAN M E J,WATTS D J. Renormalization group analysis of the small-world network model[J]. Physics Letter A,1999,293(4/5/6):341-346.
[15] BARABÁSI A L,ALBERT R. Emergence of scaling in random networks[J]. Science,1999,286:509-512.
[16] 刘胜久,李天瑞,珠杰,等. 具有双峰效应特性的复杂网络模型研究[J]. 复杂系统与复杂性科学,2017,14(1):46-51,102.
[17] 刘胜久,李天瑞,洪西进,等. 基于矩阵运算的复杂网络构建方法[J]. 中国科学(信息科学),2017,46(5):610-626.
[18] 朱大智,吴俊,谭跃进,等. 度秩函数——一个新的复杂网络统计特征[J]. 复杂系统与复杂性科学,2006,3(4):28-34.
[19] WEI D J,LIU Q,ZHANG H X,et al. Box-covering algorithm for fractal dimension of weighted networks[J]. Scientific Reports,2013,3:3049.
[20] LIU J L,YU Z G,ANH V. Topological properties and fractal analysis of a recurrence network constructed from fractional Brownian motions[J]. Physical Review,E. Statistical,Nonlinear,and Soft Matter Physics,2014,89(3):032814.
[21] 刘胜久,李天瑞,刘小伟. 网络维数:一种度量复杂网络的新方法[J]. 计算机科学,2019,46(1):51-56.
[22] LUCE R D,PERRY A D. A method of matrix analysis of group structure[J]. Psychometrika,1949,14(2):95-116.

[责任编辑:严海琳]

(上接第 55 页)

[39] WANG X,GIRSHICK R,GUPTA A,et al. Non-local neural networks[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City,UT,USA,2018.
[40] HOWARD A G,ZHU M L,CHEN B,et al. MobileNets:Efficient convolutional neural networks for mobile vision applications[J]. arXiv Preprint arXiv:1704.04861,2017.
[41] GALLUP D,FRAHM J M,MORDOHAI P,et al. Real-time plane-sweeping stereo with multiple sweeping directions[C]//2007 IEEE Conference on Computer Vision and Pattern Recognition. Minneapolis,Minnesota,USA,2007.
[42] XU N,PRICE B,COHEN S,et al. Deep image matting[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Honolulu,HI,USA,2017.
[43] JENSEN R,DAHL A,VOGIATZIS G,et al. Large scale multi-view stereopsis evaluation[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Columbus,OH,USA,2014.
[44] TOLA E,STRECHA C,FUA P. Efficient large-scale multi-view stereo for ultra high-resolution image sets[J]. Machine Vision and Applications,2012,23(5):903-920.
[45] CAMPBELL N D F,VOGIATZIS G,HERNÁNDEZ C,et al. Using multiple hypotheses to improve depth-maps for multi-view stereo[C]//European Conference on Computer Vision. Marseille,France,2008:766-779.
[46] GALLIANI S,LASINGER K,SCHINDLER K. Gipuma:Massively parallel multi-view stereo reconstruction[J]. Publikationen der Deutschen Gesellschaft für Photogrammetrie,Fernerkundung und Geoinformation,2016,25:361-369.
[47] FURUKAWA Y,PONCE J. Accurate,dense,and robust multiview stereopsis[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence,2009,32(8):1362-1376.

[责任编辑:陈　庆]