光电技术及应用

DOI：10.16818/j.issn1001－5868.2022092004

基于轴向空间注意力和中间融合表示的单图像三维重建模型

周　敏，张俊然*，李南欣

（四川大学 电气工程学院，成都 610065）

摘　要： 提出一种基于注意力和中间融合表示的三维重建模型，旨在重建具有精细化结构的三维模型。该方法利用轴向空间注意力机制学习不同方向的信息，将其嵌入编码器中以捕获局部结构特征；并基于双流网络推测深度图和三维平均形状以设计中间融合表示模块，该模块能够有效地融合可见表面细节信息，从而更好地描绘对象的三维空间结构。实验结果表明：所提出的轴向空间注意力机制和中间融合表示模块增强了特征提取的能力，IoU 和 F-score 比 PixVox＋＋分别提升了 1.3％和 0.4％，三维重建效果更优。

关键词： 深度学习；三维重建；轴向空间注意力；深度图；中间融合表示

中图分类号： TP391.9　**文章编号：** 1001－5868(2023)01－0122－06

A Single-image 3D Reconstruction Model Based on Axial Spatial Attention and Intermediate Fusion Representation

ZHOU Min, ZHANG Junran, LI Nanxin

（1. College of Electrical Engineering, Sichuan University, Chengdu 610065, CHN）

Abstract: A 3D reconstruction model based on attention and intermediate fusion representation is proposed, aiming to reconstruct a 3D model with refined structure. This method used the axial spatial attention mechanism to learn information in different directions, and embeded it in the encoder to capture local structural features. Then, based on the two-stream network, the depth map and 3D average shape could be inferred to design an intermediate fusion representation module which could effectively fuse visible surface details to better describe the 3D spatial structure of the objects. The experimental results show that the axial spatial attention mechanism and the intermediate fusion representation module proposed in this paper enhance the ability of feature extraction. The IoU and F-score are 1.3% and 0.4% higher than PixVox＋＋, respectively, proving that the 3D reconstruction effect is better.

Key words: deep learning; 3D reconstruction; axial spatial attention; depth map; intermediate fusion representation

0　引言

基于多视图几何的三维重建技术借助高精度的图像采集装置获取目标对象的特征，然后利用重建算法对其进行重建，该类算法的计算量较大且易受光照的影响，导致精度不高。近年来，随着深度学习领域的发展，基于单图像实现三维重建的方法广泛应用于考古与文物保护、测绘、VR 和自动驾驶等[1]领域，具有计算量小、受光照影响小等优点。

目前，卷积神经网络、递归神经网络等深度学习

收稿日期：2022－09－20.

基金项目：德阳科技（揭榜）项目（2021JBJZ007）；成都市科技项目重点研发支撑计划（技术创新研发项目：2021-YF05-00916-SN）；华西医院 135 交叉学科创新项目（ZYJC21041）；四川大学“医学＋信息”中心融合创新项目（YGJC012）.

* **通信作者：**张俊然 **E-mail：** junranzhang@scu.edu.cn

模型已成功应用于基于单图像的重建方法中，通过学习单图像数据与三维模型数据之间的映射关系，达到重建的目的[2]。例如：Choy 等[3]提出 3D-R2N2，使用 LSTM 提取结构特征，随着时间推移，LSTM 存在结构信息能力不足的缺陷。Yang 等[4]引入注意力聚合模块对深层次特征进行加权，但对图像的低维特征能力较低。Mescheder[5]提出 Occupancy Networks，利用 CNN 描述三维模型的决策边界，降低了占用内存，能够捕捉较为完整的拓扑结构。Xie 等[6]提出 Pix2Vox，利用上下文编码器并行选择不同的三维模型以重建目标模型；在 Pix2Vox 基础上，Xie 等[7]提出 Pix2Vox＋＋，该模型的多尺度上下文感知融合通过对重建对象不同视图分配权重以融合不同视图的三维信息，提升了重建精度，在当前的三维重建模型中表现出色。

为了进一步提升单图像三维重建精度，令网络捕获更多的局部结构信息，本文对 Pix2Vox＋＋模型进行改进，提出基于轴向空间注意力和中间融合表示的单图像三维重建模型。首先，提出轴向空间注意力机制用于编码器中增强对局部结构信息的细粒度特征提取；然后，利用编码器-解码器生成深度图和三维平均形状两种表示，通过中间表示融合模块将二者结合，以弥补在可见表面细节信息方面的缺失。

1　模型设计

1.1　模型概述

本文方法整体概述如图 1 所示：先通过编码器-解码器为三维模型对应的单图像生成深度图和三维平均形状，中间融合表示模块将生成的深度图和三维平均形状进行加权和，以捕获更多的细节信息；多尺度上下文融合模块将不同层次的三维平均形状进行融合；细分器通过“U”型结构对中间融合表示模块和多尺度上下文融合模块联合作用下的重建结果进行修正，生成最终的三维形状。

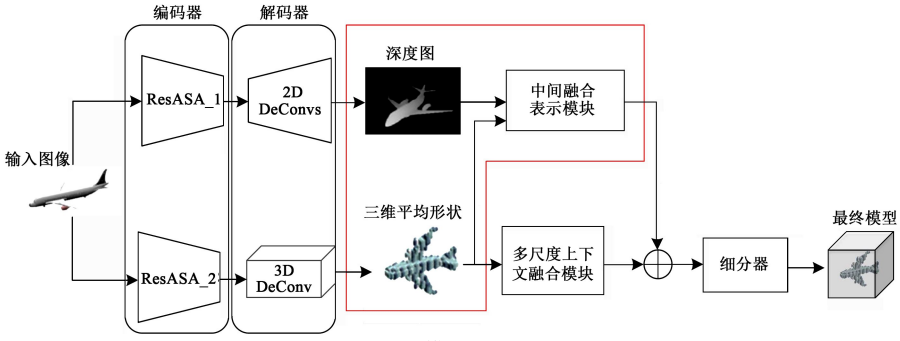


图 1　模型流程图

1.2　基于轴向空间注意力机制的残差单元模块

在编码器-解码器中，编码器由 ResNet-50 网络层和普通卷积层构成，深度图支路的编码器由 1 个 7×7 普通卷积层和 3 个残差模块级联而成，三维平均形状支路的编码器由 1 个 7×7 普通卷积层、2 个残差模块、3 个 3×3 的卷积层和 3 个最大池化层级联而成。2 条支路中，解码器借助 5 个转置卷积层实现特征向量到深度图和三维平均形状的转换。

编码器可以提取目标对象的结构特征，而编码器在利用 CNN 对图像进行特征提取时，主要关注二维图像中目标物体的局部结构信息，忽视了像素稀疏特性对提取形状信息的影响。除此之外，本文中的二维图像像素分布稀疏且集中在中心区域，因此在对三维形状模型实现重建时，普通卷积层易使特征信息产生冗余，并且普通卷积层的计算量较大。为解决上述问题，本文基于稀疏卷积运算[8]提出一种轴向空间注意力机制（Axial Spatial Attention，ASA）嵌入编码器结构中，提升局部结构信息捕获能力。轴向空间注意力模块示意图如图 2 所示。

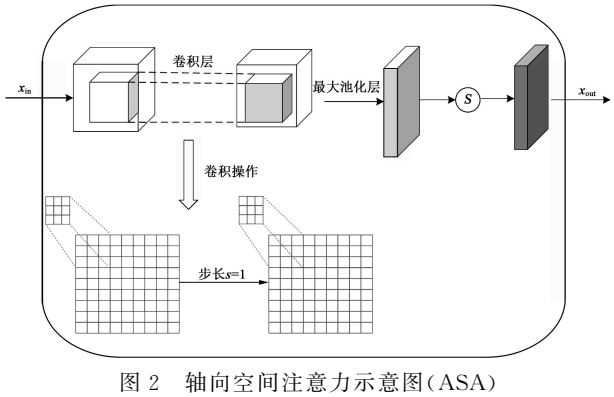


图 2　轴向空间注意力示意图（ASA）

针对图像特征 $k_i \in R^C$，普通卷积对 C 个通道

特征使用卷积核 $X \in R^{C\times C\times S\times S}$，卷积操作可表示为 $\sum_{j\in\Omega(j)} X[\alpha_i-\alpha_j]k_j+b$。其中，$\alpha_i=(x_i,y_j)^{\mathrm{T}}$ 为像素坐标，$\Omega(i)$ 为卷积窗口；b 为偏置，输出特征 $k'_i \in R^{C'}$。本文基于传统卷积操作设计了如下的“十”字交叉型卷积核 $X'=[0,1,0;1,1,1;0,1,0]$，以提高对像素特征的学习能力，改进后的卷积操作表示为 $\sum_{j\in\Omega(j)} X'[\alpha_i-\alpha_j]k_j+b$。

图 2 中给定输入特征图 $x_{\mathrm{in}} \in R^{C\times H\times W}$，“十”字交叉型的卷积核 x_{in} 进行卷积操作，然后将得到的空间注意力因子通过最大池化层、激活层处理，输出 x_{out}：

$$x_{\mathrm{out}}=\sigma(\mathrm{MaxPool}(f^{3\times3}(x_{\mathrm{in}}))) \tag{1}$$

式中，$f^{3\times3}$ 表示卷积核为 3、步长为 1 的卷积运算，C 为通道数，σ 表示 SwitchNorm 激活操作。

轴向空间注意力机制是将原始特征中的空间信息通过稀疏卷积模块做空间变换，同时保留了关键信息，强化了对重要结构特征信息的关注度。将该模块引入到编码器中的残差单元[9]，便于充分挖掘图像中蕴含的空间信息以及关注重建对象的局部结构特征。其具体操作如图 3 所示。

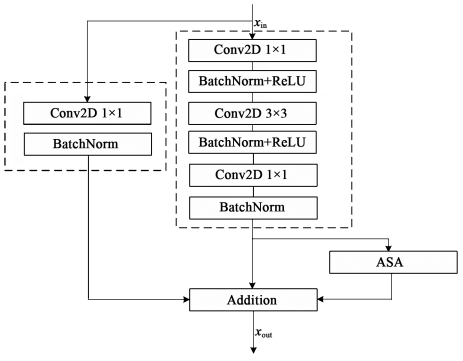


图 3 残差单元模块具体操作

1.3 本文模型的中间融合表示方法

除了引入轴向空间注意力解决细节特征提取能力不足、参数量过大等问题之外，本文将深度图作为目标物体可见表面形状信息的一种中间表示，三维平均形状作为目标物体的局部结构特征的一种中间表示，借助改进的编码器-解码器结构用于生成深度图和三维平均形状，并将二者两种中间表示进行融合以获得更好的物体细节重建效果。

1.3.1 三维平均形状与深度图

三维平均形状是一种三维空间表示形式，蕴含着目标物体的空间结构信息，本文基于轴向空间注意力的编码器和三维解码器构建了一条三维平均生成分支。输入二维图像，经过二维编码器-三维解码器得到 32×32×32 的三维平均形状模型，二者由 reshape 进行连接。

目前 2.5D 草图[10]（深度图、法线贴图和轮廓）作为一种 2D 和 3D 的中间域，将二维图像转化为 2.5D 草图，然后再生成三维模型；该过程与直接通过二维图像生成三维模型相对比，减轻了域传输的负担，并借助 2.5D 草图中对象的内在属性更容易重建三维模型的形状。考虑到深度图估计较为容易，本文将深度图作为另外一种中间表示，将深度图与三维平均形状进行结合，可用于重建高精度的三维形状模型。

1.3.2 中间融合表示模块

本文将三维平均形状和深度图像这两种中间表示进行融合，融合结果用于探索目标物体的空间结构信息与可见表面细节信息，融合网络如图 4 所示。从整体上看，通过输入二维图像初步提取前期结构特征，促进可见表面细节信息和局部结构特征的提取；采用中间融合表示模块提取丰富的形状细节信

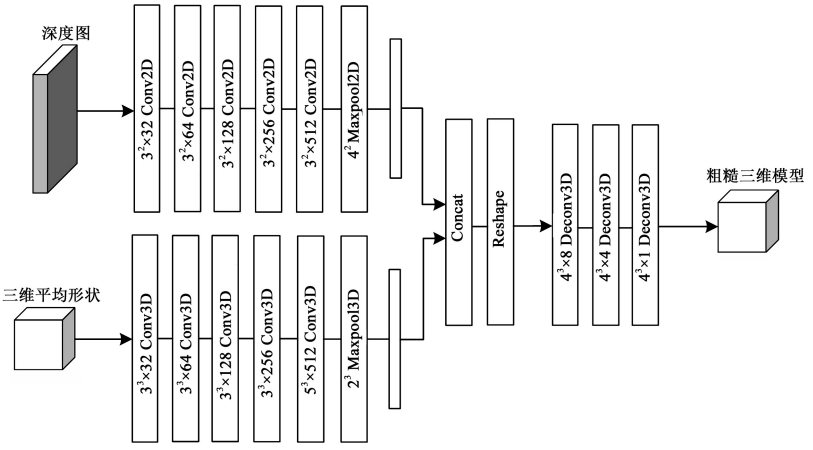


图 4 中间融合表示模块

息，以此作为连接两条分支的关键信息，与原有的三维模型进行融合，对于细节部分有更好的重建效果。

网络结构中，首先将三维平均形状和深度图分别输入通道数为 32，64，126，256 和 512 的卷积层，分别利用一个最大池化层，将特征图转化为 512 维的特征向量；然后分别利用 Concat 拼接和 Reshape 重塑操作将特征向量转化为三维特征；最后利用三个卷积核大小为 4×4×4 大小的反卷积层得到所需的粗糙三维模型。网络将隐式学习的可见表面细节特征嵌入至主干网络中，利于构建高精度的三维重建模型。

1.4　损失函数

损失函数是影响网络收敛的一大重要因素，它用来评价模型的预测值和真实值之间的差异。三维重建网络模型重建出体素表征的模型是一个二值化模型，由此可将其视为一种逻辑回归任务，因此选用二元交叉熵损失函数作为任务的损失函数，其数学表达式如式(2)所示。二元交叉熵损失越小，表示预测的三维模型与真实的三维模型距离越小。

$$loss=\frac{1}{n}\sum_{i=1}^{n}[p_i\log(q_i)+(1-p_i)\log(1-q_i)] \tag{2}$$

式中，n 表示真实三维模型中的体素数量，p_i 表示真实三维模型中体素的存在概率，q_i 表示本文预测三维模型中体素的存在概率。

2　实验与结果分析

2.1　数据集与模型参数

本文基于 Pycharm 开发环境、Pytorch 深度学习框架、NVIDIA RTX1080GPU 显卡硬件条件，完成实验设计。

数据集：本文采用 ShapeNet 的 13 种类别，共有 48 235 份数据，每份数据包含一个三维模型和 24 幅图像，训练集、测试集和验证集的比例为 8∶2∶1。

实验参数设置：输入 224×224 的二维图像，经编码、解码网络得到大小为 32×32×32 的三维体素模型。epoch 为 250，Batch size 为 32，阈值设为 0.4。本文采用 Adam 算法对网络参数进行更新，参数设置与 Pix2Vox＋＋网络选取一致，超参数 β_1，β_2 分别设为 0.9 和 0.999，学习率设置为 0.001。

2.2　实验指标

本文选用 IoU[11] 和 F-score 作为评价指标，上述指标值越高表示重建效果越好，IoU 计算公式如式(3)所示：

$$IoU=\frac{\sum_{i,j,k}I(q(i,j,k)>t)I(gt(i,j,k))}{\sum_{i,j,k}I[(q(i,j,k)>y)+I(gt(i,j,k))]} \tag{3}$$

式中，$I(\cdot)$ 表示指数函数，$q(i,j,k)$ 和 $gt(i,j,k)$ 分别表示在 (i,j,k) 处预测的三维体素模型和真实三维体素模型占用概率值，t 表示体素化阈值。

计算 F-score 时首先应用移动立方体(Marching Cubes)算法[12]将体素三维模型转化为三维网格曲面，从中采样 8 192 个点用于计算 F_1 分数(F-score)。计算公式如式(4)所示：

$$F\text{-score}=\frac{2Pre(d)Rec(d)}{Pre(d)+Rec(d)} \tag{4}$$

式中，d 表示距离阈值，设为 0.01，$Pre(d)$ 和 $Rec(d)$ 分别表示准确率和召回率，计算公式分别如式(5)，(6)所示：

$$Pre(d)=\frac{1}{k_R}\sum_{r\in R}[\min_{r\in G}\|g-r\|<d] \tag{5}$$

$$Rec(d)=\frac{1}{k_G}\sum_{g\in G}[\min_{r\in R}\|r-g\|<d] \tag{6}$$

式中，R，G 分别表示预测点云、真实点云，k 表示 R，G 中点云数。

2.3　消融实验与分析

为分析不同模块针对基于单幅图像实现三维重建时的提升效果，本文设置消融实验以比较和验证轴向空间注意力模块(Axial Spatial Attention，ASA)和中间融合表示模块(Middle Fusion Representation，MFR)的重建效果。消融实验以 Pix2Vox＋＋网络为基准模型，以及添加不同模块的重建效果，实验结果见表 1。

表 1　消融实验结果

网络	ASA	MFR	IoU	F-score
Pix2Vox＋＋	×	×	0.670	0.436
本文网络/ASA	√	×	0.674	0.437
本文网络/MFR	×	√	0.680	0.438
本文网络	√	√	0.683	0.440

从表 1 可以看出，与 Pix2Vox＋＋网络相比，存在以下结果：(1)引入轴向空间注意力模块，便于通过关注更有效的局部结构特征和抑制冗余的特征，重建三维模型的 IoU 值、F-score 值分别提升了 0.4%和 0.1%；(2)引入中间表示融合模块，将深度图和三维平均形状结合以关注目标对象的可见表面几何信息，本文的 IoU 值、F-score 值分别提升了 1.0%和 0.2%；(3)引入空间注意力模块和中间表

示融合模块，重建三维模型的 IoU 值、F-score 值分别提升了 1.3%和 0.4%。实验结果表明，所提出的轴向空间注意力机制和中间表示融合模块结合能够最大程度地提升重建三维模型的精度。

2.4 特征向量不同维度的实验及分析

为分析中间融合模块中特征向量维度对重建效果的影响，在优化本文模型参数的基础上，分别将特征向量的维度变为 128，256 和 1 024，实验结果见表 2。从表 2 中可知：相较于 128，256 和 1 024 维度的特征向量，加入 512 维度的特征向量，重建效果有明显的提升。128 和 256 维的特征向量在降维时丢失了一定的信息，而 1 024 维的特征向量可能有冗余信息，512维的特征向量能够充分捕捉细节信息，从而使得重建模型的精度越高。

表 2 不同维度的实验

向量维度	IoU	F-score	向量维度	IoU	F-score
128	0.675	0.436	1 024	0.670	0.438
256	0.681	0.438	512	0.683	0.440

2.5 对照实验

为了评价本文模型在基于单幅图像的三维模型重建任务中的有效性，选用几种经典三维模型重建方法进行比较，包括 3D-R2N2[3]、Attsets[4]、OccNet[5]、基准模型 Pix2Vox＋＋[7]，对照实验数据为同一个数据集 ShapeNet 中的 13 种类别数据，实验结果如表 3 所示，表中包含了 5 种网络的 IoU 和 F_1-score 值，对比网络数据来源于对应其文献中对应的实验数据。

综合表 3 可知，本文模型有着相对较好的性能，与其他网络模型相比：(1)13 种类别中，本文模型重建的 8 种类别的 IoU 优于其余 4 种网络，5 种类别的 F-score 优于其他 4 种网络；(2)本文模型的平均重建效果较好。通过与不同经典网络的比较，说明了本文模型的有效性。

表 3 5 种网络三维模型重建的 IoU/F-score 值对比

类别	3D-R2N2	Attsets	OccNet	Pix2Vox＋＋	本文模型
飞机	0.513/0.412	0.594/0.489	0.532/0.494	0.674/0.583	0.717/0.598
长凳	0.421/0.345	0.552/0.406	0.597/0.318	0.608/0.478	0.638/0.496
橱柜	0.716/0.327	0.783/0.367	0.674/0.449	0.799/0.408	0.791/0.418
汽车	0.798/0.481	0.884/0.497	0.671/0.315	0.858/0.564	0.869/0.559
椅子	0.466/0.238	0.559/0.334	0.583/0.365	0.581/0.309	0.597/0.324
显示器	0.468/0.227	0.565/0.310	0.651/0.468	0.548/0.296	0.542/0.307
灯	0.381/0.267	0.445/0.315	0.474/0.361	0.457/0.315	0.461/0.322
扬声器	0.662/0.231	0.721/0.211	0.655/0.249	0.721/0.152	0.713/0.285
枪	0.544/0.521	0.601/0.524	0.656/0.219	0.617/0.574	0.635/0.570
沙发	0.628/0.274	0.703/0.334	0.669/0.324	0.725/0.377	0.722/0.389
桌子	0.513/0.340	0.590/0.419	0.659/0.549	0.620/0.406	0.630/0.396
手机	0.661/0.504	0.743/0.469	0.794/0.273	0.809/0.633	0.813/0.620
船	0.513/0.305	0.601/0.315	0.579/0.347	0.603/0.390	0.626/0.396
平均	0.560/0.351	0.642/0.395	0.626/0.393	0.670/0.436	0.683/0.440

2.6 重建效果比较

图 5 给出了若干个示例样本，且所有样本来自测试集，作为基准，提供了单幅图像下 3D-R2N2 网络、Pix2Vox＋＋网络的重建结果。从图 5 可以看到，由本文网络重构出的三维模型具有良好的视觉感受，而且与基准网络相比拥有更为接近 GT (Ground-Truth)的三维体素形状。具体而言，通过 Pix2Vox＋＋网络将二维图像映射为三维模型存在细节特征信息提取不足，而本文所提出的模型增强了对细节部分的特征提取，提升重建效果。第一行图片中，输入飞机的二维图像，利用本文模型所生成的飞机模型相较于 3D-R2N2 网络和 Pix2Vox＋＋网络而言，机身更为平滑，重建效果更为自然。

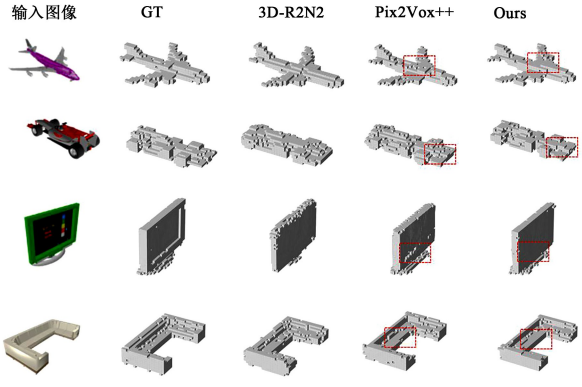


图 5 重建示例

3 结论

针对单幅图像进行三维重建时，由于特征提取

不足导致细节缺失的问题，本文从注意力网络和深度图的中间融合表示这两方面考虑以解决上述问题。首先，本文提出的轴向空间注意力机制有助于充分提取目标物体的局部结构信息；其次，二维图像能通过深度图这一中间表示推测出物体的可见表面细节信息，与三维平均形状结合，对于重建细节方面有较大的提升效果。消融实验结果表明单独使用轴向空间注意力机制或中间融合表示模块都是有效的，且两者可以联合应用来获得更优的效果，能够改善细节特征提取不足的问题。本文方法在公开的ShapeNet数据集上可以取得较好的重建效果，可以为自动驾驶、农业、规划设计、考古与文物保护、医疗等领域的应用提供参考。未来考虑将模型轻量化以及增强模型的迁移性，应用于具体领域如基于单幅图像的文物、人脸和复杂场景的三维重建。

参考文献：

[1] 陈　加，张玉麒，宋　鹏，等. 深度学习在基于单幅图像的物体三维重建中的应用[J]. 自动化学报，2019，45(4)：657-668.
Chen Jia，Zhang Yuqi，Song Peng，et al. Application of deep learning to 3D object reconstruction from a single image[J]. Acta Automatica Sinica，2019，45(4)：657-668.

[2] Fu K，Peng J，He Q，et al. Single image 3D object reconstruction based on deep learning：A review [J]. Multimedia Tools and Applications，2021，80(1)：463-498.

[3] Choy C B，Xu D，Gwak J Y，et al. 3dr2n2：A unified approach for single and multi-view 3d object reconstruction [C]// European Conf. on Computer Vision，2016：628-644.

[4] Yang B，Wang S，Markham A，et al. Robust attentional aggregation of deep feature sets for multi-view 3D reconstruction[J]. Inter. J. of Computer Vision，2020，128(1)：53-73.

[5] Mescheder L，Oechsle M，Niemeyer M，et al. Occupancy networks：Learning 3d reconstruction in function space[C]// Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition，2019：4460-4470.

[6] Xie H，Yao H，Sun X，et al. Pix2vox：Context-aware 3d reconstruction from single and multi-view images[C]// Proc. of the IEEE/CVF Inter. Conf. on Computer Vision，2019：2690-2698.

[7] Xie H，Yao H，Zhang S，et al. Pix2Vox＋＋：multi-scale context-aware 3D object reconstruction from single and multiple images[J]. Inter. J. of Computer Vision，2020，128(12)：2919-2935.

[8] Graham B，Engelcke M，Van Der Maaten L. 3d semantic segmentation with submanifold sparse convolutional networks [C]// Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition，2018：9224-9232.

[9] He K，Zhang X，Ren S，et al. Deep residual learning for image recognition[C]// Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition，2016：770-778.

[10] Wu J，Wang Y，Xue T，et al. Marrnet：3d shape reconstruction via 2.5d sketches [J]. Adv. in Neural Information Processing Systems，2017，30：540-550.

[11] Rahman M A，Wang Y. Optimizing intersection-over-union indeep neural networks for image segmentation[C]// Inter. Symp. on Visual Computing，2016：234-244.

[12] Lorensen W E，Cline H E. Marching cubes：A high resolution 3D surface construction algorithm [J]. ACM Siggraph Computer Graphics，1987，21(4)：163-169.

作者简介：

周　敏(1998—)，女，四川省内江市人，硕士研究生，主要研究方向为深度学习、图像处理；

张俊然(1978—)，男，湖南省常德市人，教授，博士研究生导师，CCF会员(09692M)，主要研究方向包括医学影像与大数据分析、类脑人工智能与模式识别。

(上接第108页)

[12] 玄光男，程润伟. 遗传算法与工程优化[M]. 北京：清华大学出版社，2004.
Xuan Guangnan，Cheng Runwei. Genetic Algorithms and Engineering Optimization[M]. Beijing：Tsinghua University Press，2004.

[13] 林贻翔. 无线激光通信光电跟瞄技术研究[D]. 武汉：武汉大学，2017.
Lin Yixiang. Research of electro-optical tracking system for wireless laser communication[D]. Wuhan：Wuhan University，2017.

作者简介：

林贻翔(1989—)，男，福建省福清市人，博士，高级工程师，分别于2012年、2017年获得武汉大学理学学士、工学博士学位，主要从事无线激光通信、光束控制方面的研究工作；

吴世奇(1983—)，男，四川省乐山市人，硕士，高级工程师，主要研究方向为宽带无线通信系统传输组网技术、激光/射频混合通信技术、高速信号处理技术；承担国家级、省部级项目10多项，获省部级科技进步二等奖1项，获发明专利授权3件。