第 37 卷 第 1 期 石家庄铁道大学学报(自然科学版) Vol. 37 No. 1
2024 年 3 月 Journal of Shijiazhuang Tiedao University(Natural Science Edition) Mar. 2024

分组卷积编码的单视图三维重建算法研究

白景禧， 刘春宇， 王学军

(石家庄铁道大学 信息科学与技术学院,河北 石家庄 050043)

摘要:为了进一步提升单视图图像三维重建的精度,通过对当前的先进算法进行研究,提出了一种改进的单视图三维重建网络。该网络的编码器通过改进特征提取网络,获取更加丰富完整、深层次的二维特征。在精炼器的网络架构中引入注意力机制,进一步细化三维特征,使其生成更加精细的三维体素模型。另外在网络中添加阈值调整模块,来弥补不同种类图像之间的差异,以达到更好的重建效果。实验结果表明,该网络在公共数据集 ShapeNet 上三维重建的整体 IoU 值达到 0.675,在单图像重建方面取得了更好的效果。

关键词:三维重建;单视图;体素;注意力机制

中图分类号:TP391 **文献标志码**: A **文章编号**: 2095-0373(2024)01-0107-07

近年来,使用三维重建技术恢复一张或多张二维图像的三维几何形状已经成为计算机视觉方向重要的研究内容之一[1]。相比于成本高昂的人工手动建模以及扫描设备建模,使用二维图像重建三维模型操作简单且省时省力,已被广泛应用在医学诊断、城市建设、文物修复、商品展示和无人驾驶等多个领域。在基于单幅图像物体的三维重建方法中,为了将二维卷积神经网络更好地应用于三维领域,人们探索了点云、网格、体素等形状表示方法,并根据输入图像数量的不同,将其划分为单视图三维重建和多视图三维重建[2],本文着重研究单幅图像物体的三维体素重建方法。

自 2015 年开始,3D ShapeNets[3]网络首次成功地将 3D 几何形状通过深度卷积置信网络转化成了 3D 体素的概率分布。在这之后,CHOY et al[4]提出 3D-R2N2,完成了单张或者多张二维图像的三维重建,实现 2D 图像到 3D 体素模型之间的端到端映射,成功解决了缺少纹理、摄像机的拍摄视角变化大等情形中使用传统三维重建方法失效的问题。XIE et al[5-6]提出了 Pix2Vox 和 Pix2Vox++网络架构,Pix2Vox 网络采用生成对抗网络构建了一个生成器和一个判别器,并通过对抗损失函数来优化模型,使其可以处理不同视角下的二维图像,生成高质量的三维模型。基于 Pix2Vox 的改进,出现了 Pix2Vox++网络,该网络采用深度表面几何网络结构来进一步提高生成三维模型的精度和准确性,通过逐层编码和解码的方式,将 2D 图像转化为具有高精度的 3D 网格结构。ZHU et al[7]提出了 GARNet 网络,主要用来完成多视图的三维重建,相比于 Pix2Vox++网络在精度方面有所提升,提出了一种基于全局感知注意力的融合方法,该方法建立了每个分支与全局之间的相关性,为权重推断提供了基础。但 GARNet 对于处理单视图重建的复杂形状和细节部分还不够精确。为了更好地处理单视图重建物体的复杂形状和边缘细节,提出了一个新的网络来提高网络的重建效果。

1 分组卷积编码的多尺寸融合网络

网络结构如图 1 所示,网络编码器引入分组注意力机制来提高特征提取和分类的性能;精炼器中利用注意力机制以及特征拼接的方式来更好地反映对象的形状、纹理和空间关系,从而提升三维模型的重建质量和细节还原能力;阈值调整模块则使用 YOLO 对输入图像的信息进行识别,根据识别结果选择性

收稿日期:2023-08-15 **责任编辑**:车轩玉 DOI:10.13319/j.cnki.sjztddxxbzrb.20230216

基金项目:河北省教育厅科学研究重点项目(ZD2016052)

作者简介:白景禧(1999—),男,硕士研究生,研究方向为三维重建、数字图像处理。E-mail:1455837477@qq.com

白景禧,刘春宇,王学军.分组卷积编码的单视图三维重建算法研究[J].石家庄铁道大学学报(自然科学版),2024,37(1):107-113.

地调整重建时的阈值,以达到最优的重建效果。

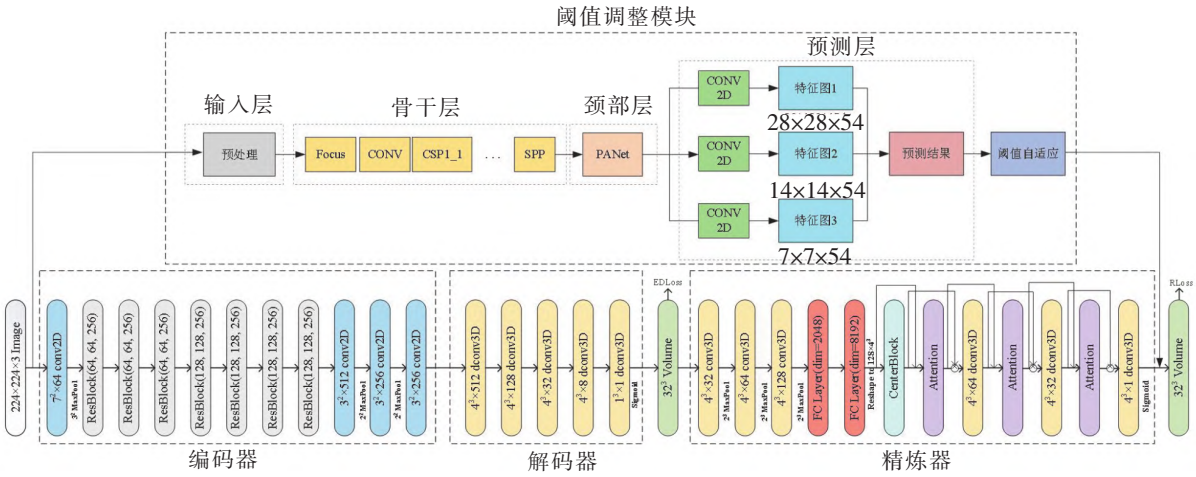

图 1 网络结构

1.1 基于层级表示学习的图像特征提取

传统的注意力机制是根据输入的权重来调整每个位置或特征在模型中的贡献,使模型能够在处理输入时更加集中地关注那些与任务相关的信息。通常使用软注意力的形式,来计算每个位置或特征的注意力权重并给予它们不同的重要性。相比之下,所使用的分组注意力机制通过对输入特征图的通道维度进行注意力加权,自适应地调整不同通道的权重,从而增强对重要特征的提取和利用能力,并有效改善模型的表达能力。通过共享注意力机制的方式,避免了为每个通道学习独立的权重参数,极大地减少模型的参数量,提升模型的轻量化和计算效率。除此之外,SplAtConv2d[8]能够更加灵活地对特征进行非线性调整,使得模型更好地理解数据中的关联和特征重要性,并将其有效应用于建模过程中,实现更准确的预测和推断。模型结构如图 2 所示,该结构可用以下公式表示

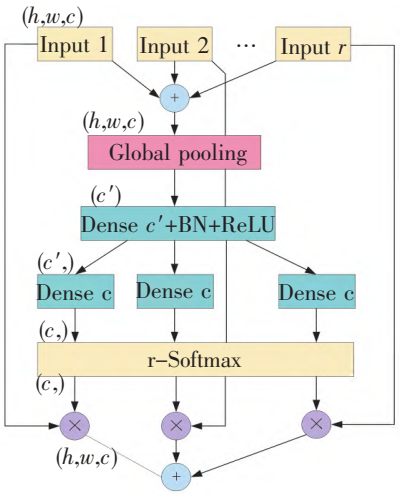

图 2 SplAtConv2d 结构

$$x_1, x_2, \cdots, x_r = \mathrm{split}(x, r) \tag{1}$$

$$m_i = \mathrm{AvgPool}(x_i) \tag{2}$$

$$a = FC_2[FC_1(m_1); FC_1(m_2); \cdots; FC_1(m_r)] \tag{3}$$

$$a_1, a_2, \cdots, a_r = \mathrm{split}(a, r) \tag{4}$$

$$y = \sum_{i=1}^{r}[\mathrm{Conv2D}(x_i, a_i) + b_i] \tag{5}$$

式中,x 为输入特征,$x \in \mathbf{R}^{HWC}$;r 为特征划分的数量;AvgPool 为平均池化操作;x_i 为划分后的特征,$x_i \in \mathbf{R}^{HW\frac{C}{r}}$;$m_i$ 为经过平均池化后的特征,$m_i \in \mathbf{R}^{\frac{C}{r}}$;$FC_1$、$FC_2$ 为全连接操作,$FC_1 \in \mathbf{R}^{C'}$,$FC_2 \in \mathbf{R}^{HW\mathrm{in\_channels}\ r}$,in_channels 为通道数;$a_i$ 为划分后的特征,$a_i \in \mathbf{R}^{HW\mathrm{in\_channels}}$;Conv2D 为 2D 卷积操作;$b_i$ 为第 i 组 2D 卷积层的偏置项;$y \in \mathbf{R}^{HWC}$ 为最终得到与输入特征维度相同的特征。此模块用于分组处理输入数据,并计算每个分组的平均值。然后通过全连接层对输入的平均池化结果进行线性变换,可以理解为对输入特征进行降维或提取更高级别的特征表示,从而帮助网络更好地理解输入数据。将输出进行分割,有助于将不同来源的特征或信息进行组合和整合。通过卷积操作可以在局部区域上提取特征,参数共享机制使得卷积具有平移不变性,可以进一步处理输入数据,捕捉更复杂的特征。

1.2 注意力驱动的三维模型细化

由于经过解码器生成的三维模型比较粗糙,部分物体的边缘细节得不到很好的处理,因此加入了注

意力机制用以改善三维模型的质量和细节,通过在神经网络中添加一个中心块,用于进行特征提取和上采样操作。具体来说,它在网络中间位置对输入的特征图进行处理,以提取更高级的抽象特征,并进行上采样以恢复细节。这对于处理具有复杂结构和丰富纹理的数据,如三维体积数据或图像数据,特别有帮助。其结构可用公式表示为

$$y_1=\mathrm{ReLU}\{\mathrm{BatchNorm}[\mathrm{Conv3D}(x)]\} \tag{6}$$

$$y_2=\mathrm{ReLU}\{\mathrm{BatchNorm}[\mathrm{Conv3D}(y_1)]\} \tag{7}$$

$$y_3=\mathrm{ReLU}\{\mathrm{BatchNorm}[\mathrm{Conv3DTranspose}(y_2)]\} \tag{8}$$

$$y=\mathrm{MaxPool}(y_3) \tag{9}$$

式中,x 为输入特征,$x\in\mathbf{R}^{D'H'W'C''}$;$y_1$、$y_2$ 为经过 3D 卷积操作后的特征,$y_1\in\mathbf{R}^{D'H'W'C}$,$y_2\in\mathbf{R}^{D'H'W'C'}$;;$y_3$ 为经过 3D 转置卷积操作后的特征,$y_3\in\mathbf{R}^{D'H'W'C'}$;$y$ 为最终得到与输入特征维度相同的特征,$y\in\mathbf{R}^{D'H'W'C''}$;$C$、$C'$、$C''$均为输出通道数;Conv3D 为 3D 卷积操作;BatchNorm 为批处理操作;ReLU 为 ReLU 激活函数;Conv3DTranspose 为 3D 转置卷积操作;MaxPool 为最大池化操作。这一部分用于对输入特征进行处理并生成与其维度相同的输出特征。它通过 3D 卷积操作提取输入数据的局部特征,然后通过批处理、ReLU 激活函数、转置卷积和最大池化等操作进行特征加工和维度变换,最终得到丰富、高维度的输出特征。这个网络结构能够有效地捕捉输入数据的空间关系和层级特征,并通过融合和整合操作生成更全面和表达力强的特征表示。

经过强化后的特征将通过不同的卷积和归一化操作进行融合,然后通过 Sigmoid 函数生成权重,最后与原始特征相乘得到增强的特征。这样可以使模型更关注对当前任务而言重要的特征,提高三维模型对边缘细节部分的细化能力。同时在网络进行输入特征融合的过程中,通过 ReLU 激活函数保留了对应位置上 2 个分支的共同关注的特征,并使用 Sigmoid 函数生成的权重来动态调整不同位置的重要性,从而增强对不同种类物体三维重建的自适应性,提高重建物体各个部分的关联性和准确性,模块结构如图 3 所示。

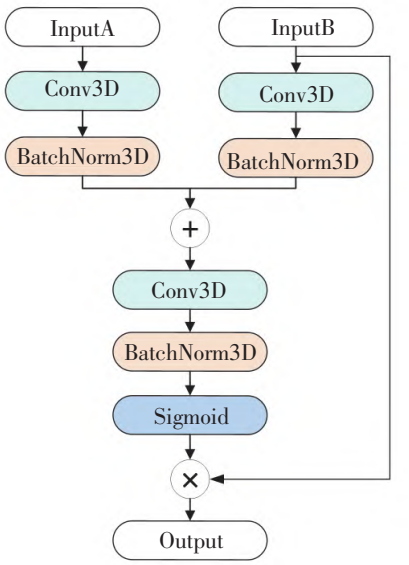

图 3　模块注意力机制

1.3　三维模型重建中的阈值自适应策略

本文提出的阈值调整模块可以动态调整三维重建任务过程中的阈值,以适应不同图像和任务的需求。模块选用 YOLO 作为识别网络可以很好地提高网络的适应性和鲁棒性,并保证物体识别的准确性,使其在不同场景和条件下都能保持较好的性能。通过将检测到的物体信息纳入三维重建过程中,可以更好地还原物体的特征和细节,提供更准确、完整的三维模型。同时可以有效降低系统的误检率,使其在生成三维模型时能够更准确地还原图像的真实形状和结构。它拥有输入层、骨干层、颈部层和预测层 4 个部分[9]。

输入层完成对数据的预处理,公式可表示为

$$a=f_{\mathrm{Anchor}}(x)\in\mathbf{R}^{9\times4} \tag{10}$$

$$x'=f_{\mathrm{Mosaic}}(x)\mathbf{R}^{672\times672\times3} \tag{11}$$

式中,x 为输入图像,$x\in\mathbf{R}^{224\times224\times3}$;$f_{\mathrm{Anchor}}$为锚框尺寸计算;$f_{\mathrm{Mosaic}}$为数据增强操作。这个预处理过程的作用是对输入数据进行有效的预处理和数据增强操作,以增强模型对不同类型物体的检测和识别能力。锚框可以帮助模型根据输入图像中物体的位置和大小进行预测,而数据增强操作可以扩展数据集,提供更多实例来训练模型,同时增加模型的鲁棒性和泛化能力。

骨干层对预处理后的图像进行特征提取获得不同尺寸的特征图,用公式可表示为

$$y_1,y_2,y_3=f_{\mathrm{Backbone}}(x')\in\mathbf{R}^{168\times168\times64},\mathbf{R}^{84\times84\times128},\mathbf{R}^{42\times42\times256} \tag{12}$$

$$z_1,z_2,z_3=f_{\mathrm{Down}}(y_1),f_{\mathrm{Down}}(y_2),f_{\mathrm{Down}}(y_3)\in\mathbf{R}^{28\times28\times256},\mathbf{R}^{14\times14\times512},\mathbf{R}^{7\times7\times1\,024} \tag{13}$$

式中,f_{Backbone}为骨干层函数,它可以将经过数据增强处理后的输入图像 x'转换为 3 个不同尺寸的特征图;

f_{Down}为下采样操作,目的是对预处理后的图像进行特征提取,得到多尺度、多层次的特征表示。

颈部层完成不同尺寸的特征拼接工作,用公式可表示为

$$p=f_{\text{PANet}}(z_1,z_2,z_3)\in\mathbf{R}^{28\times28\times512} \tag{14}$$

式中,f_{PANet}为 PANet 网络函数,它可以将不同尺寸的特征图拼接并进行相关操作,得到新的特征图 p。其作用在于整合不同尺度的特征信息,以提供更全面和丰富的特征表示,提高模型对目标的理解能力。

$$y_{\text{pred}}=f_{\text{predict}}(p,a)\in[0,1]^{3S^2(C+5)} \tag{15}$$

式中,a 为锚框尺寸;S 为特征图的大小;C 为检测中物体类别数量。预测层对图像结果进行预测,它将根据锚框尺寸和特征图大小生成一组锚框,然后为每个锚框预测它所属的物体类别概率。

2 实验结果及分析

实验以 PyTorch 框架为支撑,在 NVIDIA RTX A4000GPU 显卡上运行。实验采用 ShapeNet 3D 公共数据集,该数据集包含 13 个类别、48 235 个模型数据,每个数据包括 1 个三维模型和 24 张图像[10]。在实验中,训练集、测试集和验证集之间的比例为 8∶2∶1,遵循 Pix2Vox++网络的数据分割策略。网络输入数据是三通道 RGB 图像,通过预处理后调整为 224×224 像素尺寸,同时以大小为 32×32×32 的三维体素模型作为输出数据。将交并比(Intersection over Union,IoU)作为评价指标,以 250 次迭代训练作为最终结果,选用二元交叉熵(Binary Cross Entropy,BCE)作为损失函数[11],使用 Adam 优化器更新模型参数,并采用与 Pix2Vox++网络一致的优化器参数设置。其中权重衰减因子设置为 0.000 5,超参数 β_1 和 β_2 分别设置为 0.900 和 0.999。初始学习率为 0.001 0,并在 150 个 Epoch 后衰减至 0.000 1。初始阈值设定为 0.3。

2.1 编码器模块消融实验

为分析改进后的编码器模块在三维重建任务过程中的影响,设计了编码器模块消融实验,实验以 5 种模型的参数量、图像特征提取所需时间以及 IoU 值为参照对比,实验结果如表 1 所示。

表 1 编码器模块消融实验对比

模型	参数量/Mb	时间/s	IoU
MobileVit	11.36	1 692	0.656
MaxVit	16.16	2 916	0.659
VGG16	3.58	324	0.661
ResNet50	5.58	684	0.670
本文网络	5.82	1 152	0.672

根据实验结果分析,本文所提出网络模型在参数量方面明显优于 MobileVit、MaxVit 模型,接近 ResNet50 模型的水平。相较于 ResNet50 模型,本文模型的参数量略高,因此导致输入图像到特征提取完成所花费的时间略长,但在经过 250 个 Epochs 的训练后,本文模型的 IoU 值相较于 ResNet50 模型提升了 0.002,能够有效改善网络的重建效果。

2.2 精炼器模块消融实验

为探究精炼器加入注意力机制对重建任务的提升效果,进行不同网络精炼器模块的消融实验,对其差异进行比较。实验以参数量、三维模型精细化所需时间以及 IoU 值为参照,实验结果如表 2 所示。

表 2 精炼器模块消融实验对比

网络	参数量/Mb	时间/s	IoU
Pix2Vox	90.73	239	0.661
Pix2Vox 与改进精炼器	96.16	370	0.662
Pix2Vox++	92.73	239	0.670
Pix2Vox++与改进精炼器	98.16	368	0.672
本文网络	106.97	240	0.672
本文网络与改进精炼器	113.98	381	0.674

在本文网络的改进方法中,三维重建任务的性能得到了显著提升,同时在保持参数量不明显增长的前提下,对 Pix2Vox 和 Pix2Vox++网络结合改进的精炼器模块进行测试。实验结果显示,尽管重建时间相对较长,但整体而言,该模块对重建任务的提升显著,尤其在本文网络添加改进精炼器模块后,重建精度得到明显提升,整体 IoU 值达到了 0.674,相比于 Pix2Vox++模型提升了 0.002。

2.3　阈值调整模块消融实验

在实验过程中,对 ShapeNet 数据集的 13 个物体随机选取 1 000 张图片进行标注,共计 13 000 张图像样本,训练集和测试集划分比例为 9∶1,选择 YOLOv7m.pt 为初始训练模型,经过 100 次迭代训练生成最终结果。实验以 5 种网络的参数量、训练时间、准确率(Acc)以及 IoU 值为参照,对比结果如表 3 所示。

表 3　阈值调整模块消融实验对比

网络	参数量/Mb	训练时间/h	准确率/%	IoU
LeNet-5	0.06	4.3	63	0.673
AlexNet	0.61	5.7	69	0.672
VGG16	3.58	7.1	72	0.673
ResNet50	5.58	7.8	77	0.674
YOLOv7m	7.19	9.2	88	0.675

结果证明,不同阈值对于不同物体的重建效果存在显著影响。准确率指标较低表示网络的识别准确性较差,这会导致阈值向着错误的方向调整,从而影响三维重建的效果。使用 YOLO 进行训练通常需要较长时间,但其识别准确率远高于其他网络。在本文网络引入该模块后,整体 IoU 值达到了 0.675,三维重建效果得到明显优化。

2.4　实验结果分析

本文网络与多种三维重建网络进行对比实验,对比结果如表 4 所示。本文网络相比于 Pix2Vox++网络在整体评估中取得了 0.005 的 IoU 值提升,将 IoU 值提高至 0.675,并在衣柜、汽车等物品的三维重建任务中达到最高的 IoU 值,重建细节得到进一步提升,重建效果如图 4 所示。重建效果图中物体的边缘细节部分有了显著增强,视觉效果更为接近真实模型,整体效果明显优于改进前的 Pix2Vox、Pix2Vox++网络。因此可以得出结论,本文网络能够生成更加精细的 3D 网络模型,证明了其在提高 3D 模型精度方面的有效性。

表 4　7 种网络在不同重建物体上的 IoU 值对比

物体名称	3D-R2N2	OccNet	Matryoshka	IM-Net	Pix2Vox	Pix2Vox++	本文网络
飞机	0.513	0.532	0.647	**0.702**	0.684	0.674	0.666
长椅	0.421	0.597	0.577	0.564	**0.616**	0.608	0.606
衣柜	0.716	0.674	0.776	0.68	0.792	0.799	**0.804**
汽车	0.798	0.671	0.85	0.756	0.854	0.858	**0.862**
椅子	0.466	0.583	0.547	**0.644**	0.567	0.581	0.581
显示器	0.468	**0.651**	0.532	0.585	0.537	0.548	0.557
台灯	0.381	**0.474**	0.408	0.433	0.443	0.457	0.462
音响	0.662	0.655	0.701	0.683	0.714	0.721	**0.726**
步枪	0.544	0.656	0.616	**0.723**	0.615	0.617	0.629
沙发	0.628	0.669	0.681	0.694	0.709	0.725	**0.731**
桌子	0.513	**0.659**	0.573	0.621	0.601	0.62	0.632
电话	0.661	0.794	0.756	0.762	0.776	0.809	**0.813**
船	0.513	0.579	0.591	0.607	0.594	0.603	**0.617**

注:加粗表示最优值。

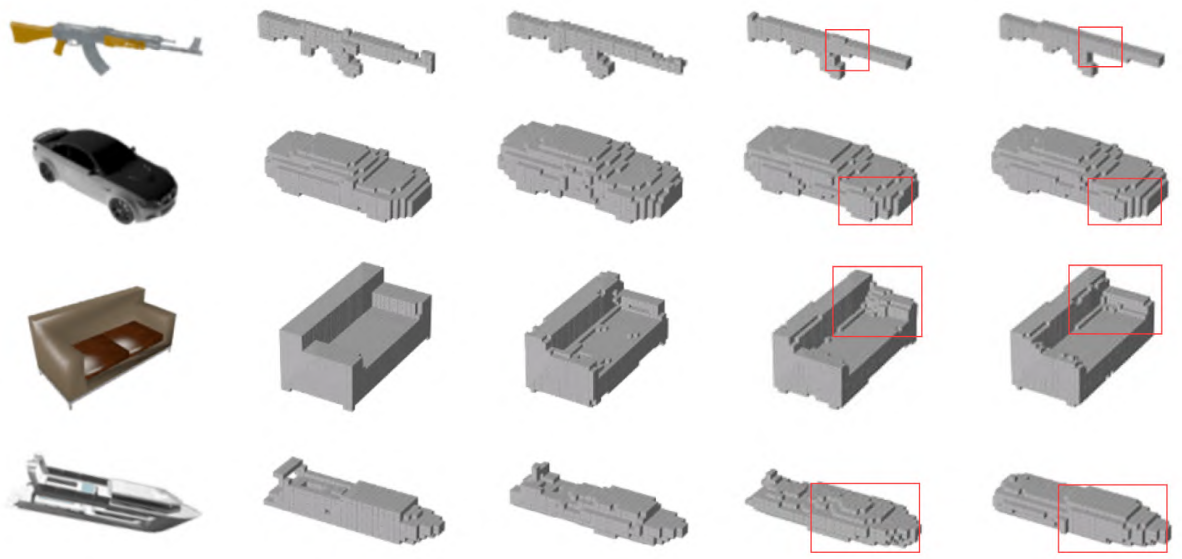

图 4 不同网络重建效果对比

3 结论

本文提出的单视图三维重建网络在公共数据集 ShapeNet 上取得了不错的重建效果,有效提升了三维重建的质量和细节还原能力,为实现更精确、高质量的三维模型重建提供了有价值的思路和方法。在未来的工作中,将研究如何进一步优化网络结构,提高重建效果和速度。此外,可以考虑将该网络应用到虚拟现实和增强现实等领域,通过不断推动三维重建技术的发展,为未来的数字化创新开辟更广阔的前景,并提供更多创新可能性。

参 考 文 献

[1]HAN Xianfeng , HAMID LAGA, MOHAMMED BENNAMOUN. Image-based 3D object reconstruction: State-of-the-art and trends in the deep learning era[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2019, 43(5): 1578-1604.

[2]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.

[3]WU Zhirong , SONG Shuran , ADITYA KHOSLA, et al. 3D shapenets: A deep representation for volumetric shapes [C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE Computer Society, 2015: 1912-1920.

[4]CHOY C B , XU Danfei , JUNYOUNG GWAK, et al. 3D-r2n2: A unified approach for single and multi-view 3D object reconstruction [C]//Computer Vision-ECCV 2016: 14th European Conference, Amsterdam, Netherlands, 2016, Proceedings, Part Ⅷ 14. Amsterdam. Netherlands: Springer International Publishing, 2016: 628-644.

[5]XIE Haozhe , YAO Hongxun , SUN Xiaoshuai , et al. Pix2Vox: Context-aware 3D reconstruction from single and multi-view images [C]// Proceedings of the IEEE/CVF International Conference on Computer Vision. Piscataway, NJ: IEEE Computer Society, 2019: 2690-2698.

[6]XIE Haozhe , YAO Hongxun , ZHANG Shengping , et al. Pix2Vox++: Multi-scale context-aware 3D object reconstruction from single and multiple images [J]. International Journal of Computer Vision, 2020, 128(12): 2919-2935.

[7]ZHU Z, YANG L, LIN X, et al. GARNet: Global-aware multi-view 3D reconstruction network and the cost-performance tradeoff [J]. Pattern Recognition, 2023, 142: 1-10.

[8]ZHANG H, WU C, ZHANG Z, et al. Resnest: Split-attention networks [C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE Computer Society, 2022: 2736-2746.

[9]顾德英,罗聿伦. 基于改进 YOLOv5 算法的复杂场景交通目标检测 [J]. 东北大学学报(自然科学版),2022,43(8):1073-1079.

[10]何鑫睿,李秀梅,孙军梅,等. 基于改进 Pix2Vox 的单图像三维重建网络 [J]. 计算机辅助设计与图形学学报,2022,34(3):364-372.

[11]BAI Zhongxin ,WANG Jianyu ,ZHANG Xiaolei , et al. End-to-end speaker verification via curriculum bipartite ranking weighted binary cross-entropy [J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2022, 30: 1330-1344.

Research on Single-View 3D Reconstruction Algorithm for Grouped Convolutional Coding

BAI Jingxi, LIU Chunyu, WANG Xuejun

(School of Information Science and Technology, Shijiazhuang Tiedao University, Shijiazhuang 050043, China)

Abstract: In order to further improve the accuracy of 3D reconstruction of single view image, an improved single view 3D reconstruction network was proposed by studying the current advanced algorithms. By improving the feature extraction network, the encoder of the network could obtain more abundant, complete and deep two-dimensional features. The attention mechanism was introduced into the network architecture of the refiner to further refine the 3D features and generate a more refined 3D voxel model. In addition, threshold adjustment module was added in the network to make up for the differences between different kinds of images to achieve better reconstruction effect. The experimental results show that the overall IoU value of 3D reconstruction on the public data set ShapeNet reaches 0.675, and the network achieves better results in single image reconstruction.

Key words: three-dimensional reconstruction; single view; voxel; attention mechanism