基于体素特征重组网络的三维物体识别

路 强[1,2]， 张春元[1]， 陈 超[1]， 余 烨[1,2]， YUAN Xiao-hui[3]

(1. 合肥工业大学计算机与信息学院 VCC 研究室，安徽 合肥 230601；
2. 工业安全与应急技术安徽省重点实验室(合肥工业大学)，安徽 合肥 230009；
3. 北德克萨斯大学计算机科学与工程学院，德克萨斯 丹顿 76201)

摘 要：三维物体识别是计算机视觉领域近年来的研究热点，其在自动驾驶、医学影像处理等方面具有重要的应用前景。针对三维物体的体素表达形式，特征重组卷积神经网络 VFRN 使用了直接连接同一单元中不相邻的卷积层的短连接结构。网络通过独特的特征重组方式，复用并融合多维特征，提高特征表达能力，以充分提取物体结构特征。同时，网络的短连接结构有利于梯度信息的传播，加之小卷积核和全局均值池化的使用，进一步提高了网络的泛化能力，降低了网络模型的参数量和训练难度。ModelNet 数据集上的实验表明，VFRN 克服了体素数据分辨率低和纹理缺失的问题，使用较少的参数取得了优于现有方法的识别准确率。

关 键 词：物体识别；体素；卷积神经网络；特征重组；短连接
中图分类号： TP 391 **DOI：**10.11996/JG.j.2095-302X.2019020240
文献标识码： A **文 章 编 号：**2095-302X(2019)02-0240-08

3D Object Recognition Based on Voxel Features Reorganization Network

LU Qiang[1,2], ZHANG Chun-yuan[1], CHEN Chao[1], YU Ye[1,2], YUAN Xiao-hui[3]

(1. VCC Division, School of Computer and Information, Hefei University of Technology, Hefei Anhui 230601, China;
2. Anhui Province Key Laboratory of Industry Safety and Emergency Technology (Hefei University of Technology), Hefei Anhui 230009, China;
3. Department of Computer Science and Engineering, University of North Texas, Denton TX 76201, United States)

Abstract: 3D object recognition is a research focus in the field of computer vision and has significant application prospect in automatic driving, medical image processing, etc. Aiming at voxel expression form of 3D object, VFRN (voxel features reorganization network), using short connection structure, directly connects non-adjacent convolutional layers in the same unit. Through unique feature recombination, the network reuses and integrates multi-dimensional features to improve the feature expression ability to fully extract the structural features of objects. At the same time, the short connection structure of the network is conducive to the spread of gradient information. Additionally, employing small convolution kernel and global average pooling not only enhances generalization capacity of network, but also reduces the parameters in network models and the training difficulty. The experiment on ModelNet data set indicates that VFRN overcomes problems including low resolution ratio in voxel data and texture deletion, and achieves better recognition accuracy rate using less parameter.

Keywords: object recognition; voxel; convolution neural network; feature reorganization; short connection

收稿日期：2018-09-03；**定稿日期：**2018-09-12
基金项目：安徽省自然科学基金项目(1708085MF158)；国家自然科学基金项目(61602146)；国家留学基金项目(201706695044)；合肥工业大学智能制造技术研究院科技成果转化及产业化重点项目(IMICZ2017010)
第一作者：路 强(1978–)，男，安徽合肥人，副教授，博士，硕士生导师。主要研究方向为信息可视化、可视分析。E-mail：luqiang@hfut.edu.cn

三维数据采集设备的普及和建模工具的简易化，使得三维模型的数量一直在快速增长。如何快速有效的识别这些三维形状，成为了计算机视觉和图形学领域，尤其是在医学影像、自动驾驶及 CAD 等应用场景下的一个重要问题。常见的三维物体描述方式，包括点云[1]、流形网格[2]、体素[3]和深度图[4]等。点云和流形网格作为一种不规则的数据组织形式，难以利用高性能的学习方法进行处理。深度图作为一种间接表现三维物体的形式，难以直观展现物体的三维结构，同时也由于遮挡问题缺失了很多信息。而体素数据能够完整地描述物体的空间占用情况，其以体素作为基本单位，数据组织形式规则，可以很好地适用现有的学习方法。

近年来，卷积神经网络(convolution neural network，CNN)被广泛地应用在分析和理解二维图像的任务中，包括图像分类[5]、物体检测[6]、语义分割[7]等。其独特的设计结构可以很好地提取图像的特征，在复杂的任务场景中具有良好的鲁棒性，表现出相较于传统方法的独特优势。鉴于体素与图像在数据组织形式上的相似性，使用 CNN 处理三维体素数据成为研究热点。相较于二维图像数据，三维体素数据由于增加了一个维度，空间开销更大，容易导致维度灾难(curse of dimensionality)[8]，其限制了体素模型的分辨率。而且体素的表现方式抛弃了物体本身的纹理信息。低分辨率和纹理缺失是使用三维体素数据训练 CNN 必然要面对的问题，要求网络能够从有限的信息中，充分提取具有代表性的物体特征。

本文针对三维体素模型识别问题，设计并搭建了一个三维 CNN VFRN(voxel features reorganization networks)。VFRN 针对现有三维体素 CNN 难以充分学习物体结构信息、参数量大、训练困难等问题，采用多维特征重组方法，融合复用多维特征提取物体特征，并通过大小为 1 的卷积核降维以减少网络参数。VFRN 使用短连接方式，减少参数量，缩短特征传递路径，降低训练难度，并加入全局均值池化[9]的方法，进一步减少了网络参数并降低了过拟合的风险。

1　相关工作

在二维图像领域，CNN 的应用已经较为成熟。2012 年 AlexNet[10]提出了 ReLU 和 Dropout 的概念，有效抑制了过拟合现象。之后，CNN 的架构逐步更新，如 VGG Net[11]，GoogLeNet[12]，Res-Net[13]等。这些网络在增加网络深度的同时，使用了不同的方法提高网络泛化能力，减小过拟合，如 GoogLeNet 中的 Inception 结构，ResNet 中的残差结构等。文献[14]提出了 DenseNet，通过密集连接方式，复用了低维特征，在增加网络深度的同时，保证了参数量的线性增长，取得了很好的效果。

目前，CNN 是提取二维图像特征最有效的方法之一。而三维形状领域发展较晚，主要进展大多在近三年内。最先使用三维数据进行深度学习实验的，WU 等[15]提出的 3D ShapeNet。该网络是一个 5 层卷积深度置信网络(convolution depth confidence network，CDCN)，输入为 30^3 分辨率的体素数据，完成识别三维物体的任务。为了进行实验，该研究构建了一个标签好的公开三维模型数据集 ModelNet[15]，此后，大量研究都在该数据集上进行了实验。作为三维工作的开端，3D ShapeNet 模型简单，识别准确率较低。鉴于 CNN 在图像应用上的优良表现以及体素与图像在数据组织形式上的相似性，文献[16]提出了 VoxNet[16]，将基本的二维 CNN 架构拓展到三维，该网络输入的是分辨率为 32^3 的体素，采用了三维卷积层和池化层，最后使用全连接层生成特征向量。虽然相较于 3D ShapeNet，VoxNet 识别效果有了较大的提升，证明了 CNN 同样适合处理三维数据，但该网络仅仅使用了普通的卷积和池化操作，并没有在分辨率限制和纹理缺失的前提下，更加充分的提取物体的三维结构特征。考虑到二维 CNN 使用的许多新结构能够提高网络表现，BROCK 等[17]提出了 VRN，该网络借鉴 GoogLeNet 中的 Inception 结构和 ResNet 中的残差结构，设计了针对三维数据的 Voxception 结构和 VRB 结构，以替换传统的卷积层和池化层。这两种结构增加了网络的支路，并融合了多尺度特征。VRN 通过对体素数据的增广和预处理，以及多个网络的联合使用，大大提高了识别准确率，但结构的复杂和多网络的联合使用，造成整个模型参数量巨大，训练困难。文献[18]构建了 3 个不同的结构网络，两个基于体素的网络和一个基于多视图的网络，通过加权综合 3 个网络的特征向量构成 FusionNet，也获得了较好的识别效果。但通过分析 FusionNet 各子网络的效果发现，两个 V-CNN 网络准确率并不高，对于提升两个网络的准确率的作用有限。而且同 VRN 一样，多网络的联合使用在训练和部署方面开销巨大，实时性较差。针对三维体素数据识别问题，SU 等[19]在分析比较了基于体素

的方法(3D ShapeNet)和基于多视图的方法(MVCNN)后，提出了 SubVolume 和 AniProbing 两种网络结构。文献[20]认为现有的三维卷积网络未能充分挖掘三维形状信息，所以在 SubVolume 网络中引入了使用局部数据来预测整体的子任务，减少过拟合的同时，也能更好地提取细节特征。AniProbing 网络则是另一种思路，使用长各向异性卷积核(long anisotropic kernels)来提取长距离特征。在网络的具体实现上，长各向异性卷积核将三维体素数据处理成二维特征图，之后使用 NIN[9]进行识别，两种网络均取得了很好的效果。由于三维 CNN 相较二维增加了一个维度，网络参数的数量也成倍增长，过多的参数量导致网络模型具有很高的计算成本，难以应用在实时领域。ZHI 等[21]提出 LightNet，使用单一模型，通过精简网络结构，大大减少了参数量，以满足实时任务的需要，缺点是牺牲了识别的准确率。

除体素数据以外，近年来也出现了一些使用点云和视图进行三维物体识别的研究。点云方面，QI 等[22]提出的 PointNet 和 PointNet++[23]，在点云数据上使用多层感知器学习一个描述点云的全局特征向量，用于识别等任务。但这两种网络受限于点云数据无序、不规则的特点，并没有考虑到一个邻域范围内的物体结构特征信息。针对上述问题，LI 等[24]搭建了 PointCNN，使用 X-Conv 操作对点云进行 X 变换，在变换后的特征上进行典型的卷积操作，一定程度上解决了将无序、不规则的数据形式映射成有序、规则形式的问题。然而 LI 等[24]也指出了网络所学习到的 X 变换远不理想，无法保证变换结果与原始点云分布的等价性。视图方面，SU 等[19]提出的 MVCNN 将三维模型数据在多个视角下渲染成一组二维图像，作为二维 CNN 的训练数据。网络中间添加 View Pooling 层用于综合多角度视图信息，得到了很好的识别效果。相似地，冯元力等[25]将三维物体绘制成多角度球面全景深度图，代替普通的多视角图像，采用同样的网络结构完成识别任务。但多视图的方式不仅需要对三维数据进行二次处理，而且对于视图的视角较为敏感。由于采用了图像作为网络输入，三维图形识别问题通过转换简化为了二维图像识别问题。

此外，还有许多针对其他三维物体表现形式的研究。如 O-CNN[26]使用八叉树方式组织三维数据并进行卷积操作，FPNN[27]使用 3D 距离场描述三维数据，3D-A-NET[28]使用三维深度形状描述符，联合训练 CNN、RNN 和敌对鉴别器。这些工作也给三维视觉领域的研究带来了新思路，但相对的，在当前环境下通用性不强。

综上，本文重点研究使用 CNN 进行三维体素数据的识别任务。目前，三维体素数据存在分辨率低，纹理缺失等问题。简单的卷积结构难以充分捕捉物体的特征信息，需要增加卷积核数量和网络深度来提取更多的高维特征，然而这会导致网络参数过多，造成网络训练困难并且容易过拟合。当前针对三维体素的 CNN，往往难以兼顾充分提取三维体素特征和控制参数数量避免过拟合这两方面的问题。本文提出了一种新的三维 CNN，用于提取三维体素数据结构特征，该网络在增加网络深度的同时，控制了参数的数量，并融合多维度特征进行卷积操作，以充分提取三维结构信息。此外，网络的短连接结构有利于梯度的反向传播，加快了训练速度，相对较少的参数有效抑制了过拟合，在三维物体识别任务上取得了很好的效果。

2 本文方法

针对三维体素识别问题，本文借鉴 DenseNet 的设计思想，提出一种全新的三维 CNN VFRN。该网络通过密集连接结构综合复用多维特征，网络参数量随深度增加线性增长，避免了参数过多导致显存不够的问题，也大大降低了训练难度。此外，网络使用了残差结构[13]，在不增加参数的前提下，进一步融合相邻维度的特征。这两种短连接的结构，有效避免了增加网络深度时可能出现的梯度消失问题。网络中对于特征通道的复用较多，考虑到卷积层对于每个特征通道的关注度会随着层数的加深而有所变化，本文使用特征重标定技术[29]对每个特征通道赋予一个权值，将加权处理后的特征通道输入卷积层进行特征提取，降低冗余特征对卷积操作的影响。本文网络结构如图 1 所示，包含两个主要模块，特征重组模块(features reorganization module，FRM)和下采样(downsample)模块。

2.1 特征重组模块(FRM)

FRM 是基于 DenseNet 的网络结构，针对三维体素识别任务的需要所设计的三维网络模块，如图 2 所示，每个 FRM 内部的特征尺寸大小保持不变。FRM 是一个多层结构，每层都包含一个连接层(Link)和一个卷积层(Conv)，输出与后面层直接相连。每层的输入都由上层的输入和输出组成，可以表示为

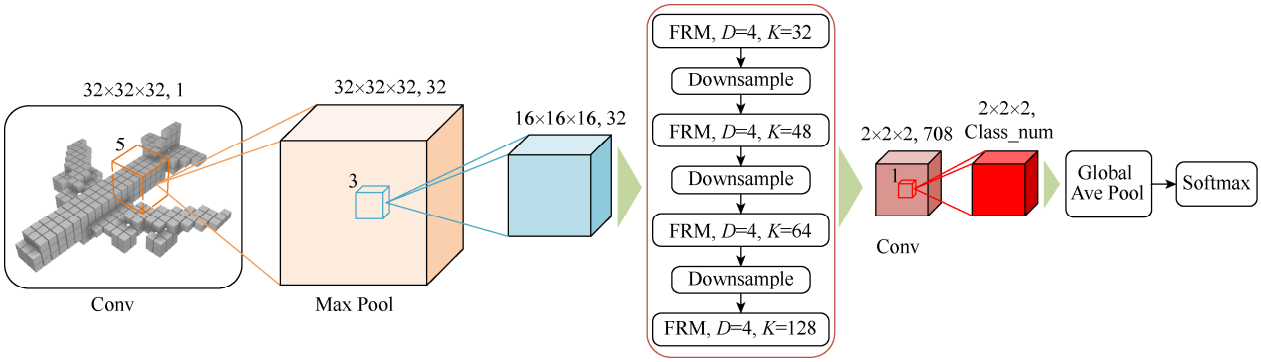


图 1　VFRN 结构

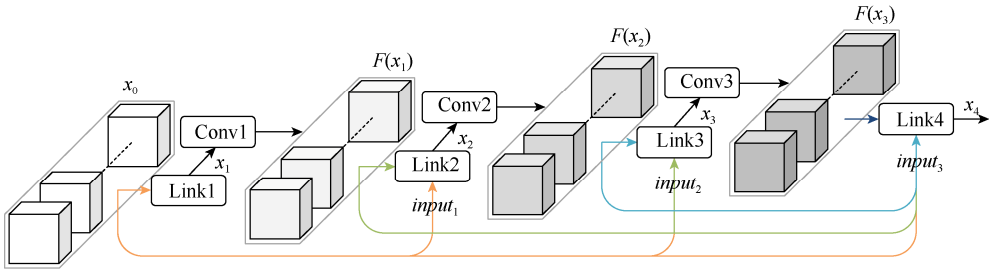


图 2　FRM 结构

$$input_l = F(x_{l-1}) + input_{l-1} \tag{1}$$

其中，$F(x)$为一个非线性变换；l 为层的编号；x_l 为第 l 层中卷积层的输入；$F(x_0)$为 x_0，$input_0$ 为空。这样每一层与损失函数都有一条短路径相连，在反向传播过程中梯度信息能够轻松地传递到每个卷积层，从而构建更深的网络以获得更好的效果。此外，FRM 的另一个特点是在同样深度下，相比其他卷积结构，参数更少。因为 FRM 中超参数 Ki 限定了第 i 个 FRM 中每一个卷积层输出的特征数量。并且卷积层的输入先通过一个 1×1×1 的卷积操作降维，减少特征通道的数量，并融合多个通道的信息。FRM 的特征复用方式，能够充分提取目标的结构特征，并保证随着深度增加，参数量线性增长。

2.1.1　连接层

连接层用于组合上层网络的输出和输入，并赋予特征通道权值。连接层的设计结构如图 2 所示，其中 l 表示连接层的序号，$F(x_{l-1})$是前一层的输出，$input_{l-1}$ 是前一层的输入。本层输入 $input_l$ 分为 $F(x_{l-1})$和 $input_{l-1}$ 两部分，首先通过 H_1 进行矩阵间对应元素相加的操作。由于 FRM 的跨层连接结构，随着 l 的增大，$input_{l-1}$ 的特征通道数 c_{l-1} 会越来越大，即

$$c_l = K_i \times l \tag{2}$$

但 $F(x_{l-1})$的通道数量受超参数 K_i 的限制，固定为 K_i。鉴于两个输入 $input_{l-1}$ 和 $F(x_{l-1})$的特征通道数不同，本文选择在 $F(x_{l-1})$与 $input_{l-1}$ 中的最后 K_i 个通道间进行对应元素求和操作，得到新的特征 f_1。之后，$F(x_{l-1})$与融合后的特征 f_1，由 H_2 完成通道维度的连接操作，即将 $F(x_{l-1})$连接到 f_1 的最后，得到特征 f_2。根据式(1)，$input_{l-1}$ 最后 K_i 个通道实际上就是 $F(x_{l-2})$，求和操作实际上是在相邻两层的输出上进行的，因此 H_1 实现了相邻层间特征的融合。而 H_2 的通道连接操作，复用了前层的低维特征，保证本层能够全局感知多维特征信息。H_1 和 H_2 两种连接结构，满足了本文在网络设计思路中，对于充分提取三维体素数据特征和融合多维度特征进行学习的要求。而且此结构也能在参数量开销较少的前提下，进一步提高网络的泛化能力。

上层网络的输入和输出组合而成的特征 f_2，包含着多个维度的特征通道，为了保证卷积层尽可能的集中注意力在其更关心的通道上，本文对各通道进行了加权操作。如图 3 所示，一个全局均值池化层将融合连接后的特征 f_2，映射为一个维度等同于 f_2 通道数的向量。以该向量作为输入，通过两个全连接层来学习一个权重向量 $\boldsymbol{w}$，中间添加 Dropout 层，Dropout 率为 0.5。第一个全连接层的神经元数

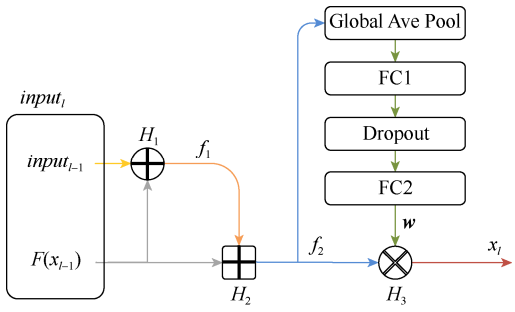


图 3　连接层结构

量设置为 f_2 通道数的 1/8，第二个全连接层的神经元数量与 f_2 的通道数相同。H_3 使用学习到的权重向量 $\boldsymbol{w}$ 来重标定 f_2 的各个通道，即将每个特征通道乘以其对应的权重，以此来增强卷积层感兴趣的特征，抑制冗余特征，综上，连接层的输出 x_l 为

$$x_l = \left\{\left[input_{l-1} \oplus F(x_{l-1})\right] + F(x_{l-1})\right\} \times \boldsymbol{w} \quad (3)$$

之后，卷积层以 x_l 作为输入，进行特征提取。

2.1.2 卷积层

输入 x_l 经过卷积层，得到输出 $F(x_l)$。如图 4 所示，卷积层由两个卷积操作和两个 dropout 操作构成。1×1×1 卷积作为一个通道数限制瓶颈，根据超参数 K_i 将通道数超过 $2K_i$ 的输入 x_l 降维到 $2K_i$，避免随着层数加深，参数量爆炸式增长，同时也能起到融合多通道特征的作用。三维卷积操作的参数量为

$$n_p = c_i \times kernel\_size^3 \times c_o \quad (4)$$

其中，n_p 为参数量；c_i 为输入的特征通道数；c_o 为输出的特征通道数；$kernel\_size$ 为卷积核的大小。在相同的输入、输出通道下，卷积参数量正比于卷积核大小的三次方。本文网络对于特征的复用重组，使得输入的通道数随着深度增加也在快速增长，所以先使用大小为 1 的卷积核降低通道数，再使用大小为 3 的卷积核，可以有效减少参数量。3×3×3 卷积用于提取邻域结构特征，输出 K_i 个特征通道。考虑到特征的复用，本文并没有使用更大的卷积核，因为文献[11]中证明多个小卷积核连接使用，可以得到等同于大卷积核的效果。而且相较大卷积核，小卷积核能够减小参数量和计算开销。在两个卷积之后，均使用了 Dropout 来保证网络的泛化能力，避免过拟合。此外，卷积操作的步长均为 1，以保持同一模块内特征的尺寸不变，便于连接层融合多维度特征。

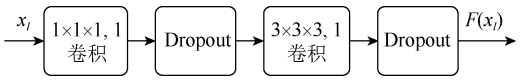


图 4 卷积层结构

2.2 下采样模块

下采样模块用于连接相邻的 FRM，由卷积和池化两步操作完成。虽然池化操作并不需要额外的参数，但考虑到输入包含多个维度的特征通道，需要同卷积层一样进行多通道特征的融合。同卷积层一样，使用了 1×1×1 的卷积来融合多通道特征并降低输入特征通道数到原先的一半。不同于通常的 CNN 中池化层的输入是同一维度的不同特征，本文网络中池化层的输入融合了多个维度的不同特征，常用的最大池化操作不能较好地采样出可以代表局部特征的信息，本文采用了平均池化操作来综合邻域信息进行下采样。

2.3 三维体素特征重组网络结构

本文网络总体结构如图 1 所示，输入为 32^3 分辨率的体素数据。网络先对输入进行步长为 1，卷积核大小为 5 的卷积操作，和步长为 2，窗口大小为 3 的最大池化操作。5×5×5 的卷积输出 32 个特征，配合最大重叠池化，初步提取目标的基本结构特征，并将体素尺寸从 32^3 降低到 16^3。之后，4 个 FRM 通过 3 个下采样层连接，用于充分提取目标特征。最后使用全局均值池化得到一个维度等同于目标类别数量的特征向量，输入 Softmax 层获得识别结果。由于网络特征通道数量一般远大于目标类别数量，所以在最后一个 FRM 和全局均值池化之间，加入一个 1×1×1 的卷积操作，输出数目等同于类别数量的特征。

网络中每次卷积操作前都使用 Batch Normalize[30]对输入进行规范化处理，并采用 ReLU 激活函数完成特征映射。

3 实验及结果分析

相比于传统面向三维体素的 CNN，本文网络不再严格按照从低维到高维的顺序进行卷积操作，而是连接重组前层多维特征，通过卷积操作提取特征，多次复用低维特征，更充分地捕捉结构特征。与二维卷积网络结构相似，高维的特征更加丰富，需要增加卷积核的数量来提取不同特征，所以 FRM 中的超参数 K_i，随着 i 的增加而增大，使得网络能够捕捉到更多高维特征，得到更高的识别精度。

3.1 实验数据集

ModelNet 是一个大型三维数据集，其中包括 662 类共 127 915 个三维模型。通常使用其中的两个子集，ModelNet10 和 ModelNet40 进行实验。ModelNet10 包含 10 类共 4 899 个三维模型，其中 908 个作为测试集，剩余 3 991 个作为训练集。ModelNet40 包含 40 类共 12 311 个三维模型，其中 2 468 个作为测试集，剩余 9 843 个作为训练集。数据集部分模型如图 5 上半部分所示。

本文将 ModelNet 数据集转换为分辨率为 32^3 的二值体素数据，部分转换实例如图 5 下半部分所示。可以看出，在 32^3 的分辨率下，对于形状特征较为突出的物体，如飞机、桌子等，体素转换可以较好地还原物体的三维轮廓结构，而对于汽车这类

结构较为简单的物体，体素转换对于轮廓的还原较为模糊。基于上述情况，且求网络对于物体的细微特征的敏感程度要更高，要能够充分提取具有代表性的物体特征。通常二值体素数据以 1 代表该位置的空间被物体占据，0 表示没有占据。为鼓励网络更关注物体占据的部分，本文使用{0，5}二值数据代替{0，1}二值数据。实验证明，加大非 0 值有利于提高识别准确率[17]。此外，为进一步提高网络的泛化能力，本文将体素数据在垂直方向上旋转 12 个角度来增广数据集，训练及测试时分别使用未增广的单角度数据和增广后的多角度数据进行实验。

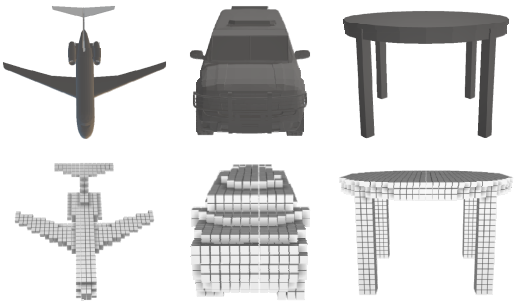


图 5　ModelNet 部分模型(上)及体素转换实例(下)

3.2　实验环境及设置

本文网络模型使用 Tensorflow 1.2 实现，cuda 版本为 8.0。硬件配置为 Intel Core i7-7700K 处理器和 Nvidia GTX1080 显卡，搭配 16 G 内存。

网络的训练阶段设置 batch size 为 32，Dropout 率为 0.2，采用交叉熵损失函数，优化策略选用 Adam 算法。初始学习率设置为 1e-4，每 30 次迭代降低为当前学习率的 20%，整个训练过程迭代 90 次，故学习率降低 2 次。

3.3　VFRN 在 ModelNet 上的性能评估

表 1 中展示了本文提出的 VFRN 与现有面向三维物体的深度学习方法在 ModelNet40 数据集上的相关性能指标。可以看出，本文提出的 VFRN 在仅使用单角度数据的情况下，就达到了较高的识别准确率，与使用多视图的 MVCNN 和使用深度全景图的全景识别网络相比，仍有明显优势，证明了 VFRN 能够充分提取物体结构特征，并且具有良好的泛化能力。相对于最早的 3D ShapeNet，VFRN 识别准确率提高了 18%，且参数量大大减少。VoxNet 由基本的 CNN 结构组成，参数较少，对于三维物体难以充分提取其特征，识别准确率较低。识别准确率较高的 FusionNet，采用的是 3 个网络组合的方式完成识别任务，其中的多视图子网络使用了 ImageNet 进行预训练。多网络组合导致整个模型参数量巨大，达到了 118 M，而单网络的 VFRN 相较 FusionNet 参数减少了 90%，并且在识别结果上有明显的提升。LightNet 的参数量最少，但识别准确率并不突出。文献[20]中提出的 SubVolume 和 AniProbing 网络，采用了比较特殊的网络结构，但在参数量和识别准确率两方面并没有明显优势。VFRN 相比于使用点云数据的 PointNet 和 PointNet++，在识别准确率上也有明显的提升。另与目前使用体素达到最好识别效果的 VRN 对比，VFRN 的参数量减少了一半，并且在单网络的前提下，另识别效果要比 VRN 略好。

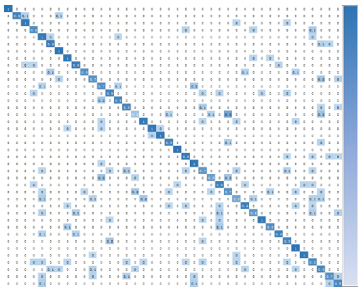
VRN Ensemble 训练了 5 个网络进行识别任务，然后依据这 5 个网络的结果进行投票，按照少数服从多数的规则确定识别结果。多网络投票的方式使

表 1　ModelNet 上多种方法识别性能比较

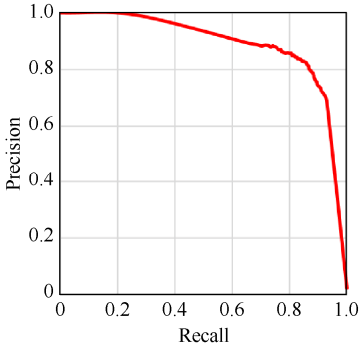
网络模型	数据表现形式	预训练	数据增广方式	参数量(M)	识别准确率(%)	
					ModelNet40	ModelNet10
本文 VFRN	体素	-	单角度	约 8.50	89.30	92.41
本文 VFRN	体素	-	垂直方向，12 个角度	约 8.50	91.41	93.79
3D ShapeNets	体素	ModelNet40	垂直方向，12 个角度	约 38.00	77.32	83.54
VoxNet	体素	-	垂直方向，12 个角度	约 0.92	83.00	92.00
VRN	体素	ModelNet40	单角度	约 18.00	88.98	-
VRN	体素	ModelNet40	垂直方向，24 个角度	约 18.00	91.33	93.61
VRN Ensemble	体素	ModelNet40	垂直方向，24 个角度	约 90.00	95.54	97.14
FusionNet	体素+多视图	ModelNet40 ImageNet 1K	垂直和水平方向，60 个角度	约 118.00	90.80	93.11
SubVolume	体素	-	垂直和水平方向，60 个角度	约 16.00	87.20	-
AniProbing	体素	-	垂直和水平方向，60 个角度	-	85.90	-
LightNet	体素	ModelNet10	垂直方向，12 个角度	约 0.30	86.90	93.39
MVCNN	多视图	ModelNet40 ImageNet 1K	80 个角度	-	90.10	-
PointNet	点云	ModelNet40	单角度	约 3.50	89.20	-
PointNet++	点云	ModelNet40	单角度	-	90.70	-
全景识别网络	深度全景图	-	8 个角度	-	86.18	89.98

得准确率得到了显著提升，因为初始状态的随机性，每个网络的拟合结果并不完全相同，结合使用弥补了单个网络识别效果的不足，但模型的参数量也成倍增长。由于策略的较大差异，VRN Ensemble 和 VFRN 之间并不具有可比性。而且针对单一数据集训练的多网络集合，很容易导致模型泛用性较差，文献[17]也指出这一结果不具有普适性。

针对数据增广方式，相比于 VRN 在垂直方向上 24 个角度的旋转，以及 FusionNet 等在垂直和水平方向上 60 个角度的旋转，本文仅做了垂直方向 12 个角度的旋转。通常数据集的增广可以带来网络泛化能力的提升，尤其是在网络参数过多的情况下，增大数据集有助于抑制过拟合现象，从而提高网络效果。表 1 中 VFRN 和 VRN 在多角度数据集上识别准确率相较于单角度数据集的提升，也进一步说明了数据增广对于网络效果的正面作用。但考虑到更大的数据集容易造成训练困难，对于学习率等参数的调整也更为敏感，并且 VFRN 的目标在于精简参数以降低训练难度的同时提高网络的性能，因此本文并没有选择更多角度旋转的方式增广数据集。得益于参数量的控制，相比 VRN，本文在其规模一半大小的数据集上训练 VFRN 仍然得到了更好的识别准确率。此外，图 6 是 VFRN 在 ModelNet40 上进行测试的混淆矩阵和 PR 曲线，反映出 VFRN 网络的稳定性和可靠性，进一步佐证了 VFRN 在提取特征和抑制过拟合方面的优势。



(a) 混淆矩阵



(b) PR 曲线

图 6 混淆矩阵和 PR 曲线

表 2 为部分网络在 ModelNet40 上的训练用时及硬件设备情况。由于网络参数量过大，模型复杂，VRN 的训练需要约 6 天，远远大于其他网络模型。而 VFRN 在保证识别效果的情况下，训练时间仅需 8 h 左右，与参数较少的 LightNet 和 VoxNet 相近。说明 VFRN 的短连接结构，使得梯度能够更好地传递到各层，有效加快了网络的训练速度。

表 2 4 种方法训练参数对比

网络模型	GPU	训练时长(h)
VoxNet	K40	约 6~12
VRN Ensemble	Titan X	约 144
LightNet	GTX1080	约 6~7
本文 VFRN	GTX1080	约 8

综合上述分析，本文提出的 VFRN 能够从体素数据中，充分提取三维物体的结构特征，并表现出良好的泛化能力。VFRN 较好地平衡了参数量和识别准确率，独特的网络结构降低了训练难度，与现有前沿方法相比具有明显的优势。

4 结 束 语

针对计算机视觉领域中三维物体的识别任务，本文设计实现了一个基于体素数据的三维 CNN VFRN，以充分提取物体的结构特征，提高目标识别的准确率。VFRN 通过短连接结构，实现了多维特征的复用和重组，弥补了传统三维体素 CNN 中特征利用率低的缺陷。同时特征复用的特性保证网络中参数量随深度增加线性增长，相比现有网络参数更少，较好地解决了三维数据空间开销过大的问题，一定程度上抑制了过拟合的问题。实验结果表明，VFRN 的识别准确率高于其他方法，并且在识别效果和参数开销两方面达成了很好的平衡。考虑到多角度数据对于识别结果的提升，后续研究将针对网络自适应变换对齐体素数据，在不添加额外训练数据的情况下进一步提升网络效果来进行。

参 考 文 献

[1] 张爱武, 李文宁, 段乙好, 等. 结合点特征直方图的点云分类方法[J]. 计算机辅助设计与图形学学报, 2016, 28(5): 795-801.

[2] 徐敬华, 盛红升, 张树有, 等. 基于邻接拓扑的流形网格模型层切多连通域构建方法[J]. 计算机辅助设计与图形学学报, 2018, 30(1): 180-190.

[3] 吴晓军, 刘伟军, 王天然, 等. 改进的基于欧氏距离测度网格模型体素化算法[J]. 计算机辅助设计与图形学学报, 2004, 16(4): 592-597.

[4] 范涵奇, 孔德星, 李晋宏, 等. 从含噪采样重建稀疏表达的高分辨率深度图[J]. 计算机辅助设计与图形学学报, 2016, 28(2): 260-270.

[5] 吕刚, 郝平, 盛建荣. 一种改进的深度神经网络在小图像分类中的应用研究[J]. 计算机应用与软件, 2014, 31(4): 182-184, 213.

[6] REN S, HE K, GIRSHICK R, et al. Faster R-CNN: Towards real-time object detection with region proposal networks [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 39(6): 1137-1149.

[7] 李琳辉, 钱波, 连静, 等. 基于卷积神经网络的交通场景语义分割方法研究[J]. 通信学报, 2018, 39(4): 123-130.

[8] BELLMAN R E. Dynamic programming [M]. Princeton: Princeton University Press, 1957.

[9] LIN M, CHEN Q, YAN S. Network in network [EB/OL]. (2013-12-16). [2014-03-04]. http://arvix.org/abs/1312.4400.

[10] KRIZHEVSKY A, SUTSKEVER I, HINTON G E. ImageNet classification with deep convolutional neural networks [C]//Proceedings of International Conference on Neural Information Processing Systems. New York: CAM Press, 2012: 1097-1105.

[11] SIMONYAN K, ZISSERMAN A. Very deep convolutional networks for large-scale image recognition [EB/OL]. (2014-09-04). [2015-04-10]. https://arxiv.org/abs/1409.1556.

[12] SZEGEDY C, LIU W, JIA Y, et al. Going deeper with convolutions [C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Tokyo: IEEE Computer Society Press, 2015: 1-9.

[13] HE K, ZHANG X, REN S, et al. Deep residual learning for image recognition [C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2016: 770-778.

[14] HUANG G, LIU Z, WEINBERGER K Q, et al. Densely connected convolutional networks [C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2017: 243.

[15] 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. Los Alamitos: IEEE Computer Society Press, 2015: 1912-1920.

[16] MATURANA D, SCHERER S. Voxnet: A 3D convolutional neural network for real-time object recognition [C]//Proceedings of the Intelligent Robots and Systems (IROS), 2015 IEEE/RSJ International Conference on. Los Alamitos: IEEE Computer Society Press, 2015: 922-928.

[17] BROCK A, LIM T, RITCHIE J M, et al. Generative and discriminative voxel modeling with convolutional neural networks [EB/OL]. (2016-08-15). [2016-08-16]. https://arxiv.org/abs/1608.04236.

[18] HEGDE V, ZADEH R. Fusionnet: 3D object classification using multiple data representations [EB/OL]. (2016-07-19). [2016-11-27]. https://arxiv.org/abs/1607.05695.

[19] SU H, MAJI S, KALOGERAKIS E, et al. Multi-view convolutional neural networks for 3D shape recognition [C]//Proceedings of the IEEE International Conference on Computer Vision. New York: IEEE Press, 2015: 945-953.

[20] QI C R, SU H, NIESSNER M, et al. Volumetric and multi-view cnns for object classification on 3d data [C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2016: 5648-5656.

[21] ZHI S F, LIU Y X, LI X, et al. Lightnet: A lightweight 3D convolutional neural network for real-time 3D object recognition [C]//Proceedings of Eurographics Workshop on 3D Object Retrieval. Goslar: Eurographics Association Press, 2017: 9-16.

[22] QI C R, SU H, MO K, et al. Pointnet: Deep learning on point sets for 3d classification and segmentation [J]. ProcEedings of the IEEE Conference on Computer Vision and Pattern Recognition. Washington, DC: IEEE Computer Society Press, 2017: 77-85.

[23] QI C R, YI L, SU H, et al. Pointnet++: Deep hierarchical feature learning on point sets in a metric space [C]//Proceedings of Advances in Neural Information Processing Systems. Heidelberg: Springer, 2017: 5105-5114.

[24] LI Y, BU R, SUN M, et al. PointCNN [EB/OL]. (2018-06-23). [2018-11-05]. https://arxiv.org/abs/1801.07791.

[25] 冯元力, 夏梦, 季鹏磊, 等. 球面深度全景图表示下的三维形状识别[J]. 计算机辅助设计与图形学学报, 2017, 29(9): 1689-1695.

[26] WANG P S, LIU Y, GUO Y X, et al. O-cnn: Octree-based convolutional neural networks for 3d shape analysis [J]. ACM Transactions on Graphics (TOG), 2017, 36(4): 72.

[27] LI Y Y, PIRK S, SU H, et al. Fpnn: Field probing neural networks for 3d data [C]//Proceedings of Advances in Neural Information Processing Systems. New York: Curran Associates Inc. 2016: 307-315.

[28] REN M, NIU L, FANG Y. 3D-A-Nets: 3D deep dense descriptor for volumetric shapes with adversarial networks [EB/OL]. (2017-11-28). [2017-11-28]. https://arxiv.org/abs/1711.10108.

[29] HU J, SHEN L, SUN G. Squeeze-and-excitation networks [EB/OL]. (2017-09-05). [2018-10-25]. https://arxiv.org/abs/1709.01507.

[30] IOFFE S, SZEGEDY C. Batch normalization: Accelerating deep network training by reducing internal covariate shift [EB/OL]. (2015-02-11). [2015-03-02]. https://arxiv.org/abs/1502.03167.