【摘　要】

利用深度图像进行手部关键点三维位置的估计是具有广泛应用价值的计算机视觉技术，至今已有几十年的研究历史。近来，有研究者开始尝试将深度图像转换为点云，利用近些年开始涌现的免预处理点云分析算法得到关键点位置。这些算法通常需要根据输出结果对其本身进行繁杂的修正处理以及不可微分的最远点采样算法。本文将“压缩-激活”模块引入点云分析网络，并指出其能将修正处理引入网络，使得修正模块加入端到端的训练。同时指出保留点的压缩可以在加入最大池化后成为可微分的点采样模块。我们对“压缩-激活”模块在点云分析网络中的重要特性给出了理论证明，并分析了其与其它模块功能是否具有互补性。最后我们在 MSRA 数据集上探究了其表现。

【关键词】 深度图像，手部关键点三维位置估计，压缩-激活模型，点云处理

[ABSTRACT]

Depth-based hand pose estimation is a computer vision technology with a wide range of applications. It is an research area with a history of decades. Recently, some researchers try to convert depth images into point clouds, using pre-processing-free point cloud analysis algorithms that have emerged in recent years to obtain hand pose. These algorithms usually need to perform complex refinement and non-differentiable farthest-point-sampling . This thesis introduces the ”squeeze-excitation” module into point cloud analysis network for depth based hand pose estimation and points out that it can introduce refinement into the network, so that the refinement module can be added to end-to-end training. At the same time, we point out that the ”squeeze-excitation” module reserving points is essentially a differentiable sampling operation when a max pooling layer is inserted. We give a theoretical proof of its important characteristics in the point cloud analysis network and analyze whether it is complementary to other modules. Finally, we explore performance of our model on the public MSRA dataset.

[Keywords] Depth-based hand pose estimation,Squeeze-excitation,Point cloud analysis

目录

插图目录

表格目录

一、诸论

1.1 选题背景与意义

精确的手部关键点 3 维位置估计具有广泛的应用价值，是实现人机交互、增强现实、虚拟现实、机械臂抓取等所需的关键技术之一。随着微软 Kinect 和英特尔 RealSense 等深度相机的普及，基于深度图像的手部关键点 3 维位置估计变得越来越重要。近些年深度学习[1] 的发展进一步推动了该研究方向的发展[2;3;4;5;6;7;8;9;10;11;12]，然而由于手部铰链连接非常复杂且不同的手指之间会造成严重的遮挡现象，这仍然是一个十分具有挑战性的问题。

对于深度图像，最直接的处理方式是利用 2 维卷积网络[3;13;14;10;15]。然而让 2 维的卷积核学习到有效的 3 维特征是十分困难的。于是，人们尝试首先将深度图像转化为体素表示，然后利用 3 维卷积对其进行处理[7;16]。该方案最大的缺陷在于 3 维卷积巨大的计算量或为了降低计算量以很低的输入分辨率作为妥协而引起细节特征的丢失。

受到 poinet，pointnet++[17;18] 的启发，Ge 等人[8;9] 尝试将深度图像转化为点云表示，并直接利用 pointnet++ 结构对其进行分析处理。这种方法不仅有利于网络学到有效的 3 维特征，同时比基于 3 维卷积的算法高效。然而，该算法需要根据输出关键点的位置重新修正结果，这需要花费很大的计算量。

Hu 等人[19] 通过实验说明了“压缩-激活”机制在卷积神经网络中是有效的，他们认为是注意力机制与门控机制起了主要作用。本文指出“压缩-激活”也有“初估计-修正”的功能，并将其用于点云处理。并且我们指出保留点维度的“压缩-激活”具有可微分降采样的功能，并讨论其非点置换不变性对其性能可能产生的影响。[19] 中指出“压缩-激活”的功能有可能被隐含在卷积核中，然而我们通过理论说明在某些情况下点云网络中“压缩-激活”无法被隐含在全连接层中。同时，我们探讨了“压缩-激活”模块是否有可能与图卷积模块或全局信息融合模块产生互补效果。最后，我们通过在 MSRA 数据集上的实验对结论进行验证。

1.2 国内外研究现状和相关工作

1.2.1 “压缩-激活”机制和“初估计-修正”

受到早期在信号处理中引入注意力机制研究[20;21;22;23;24]以及门控激活研究[25;26]的启发，Hu 等人[19]提出了卷积神经网络中的”压缩-激活”机制。“压缩”是指将每个通道的特征图池化得到概括每个通道的信息，然后把池化结果通过全连接层得到每个通道在特征分辨中的重要程度，通过该重要程度“激活”原始通道。

手部关键点的估计通常会包括初估计和修正两个步骤[11;8;3]，首先用模型得到手部关键点的初步估计，然后利用该初步估计送回原始模型“指导”模型再次进行估计。如果该“估计-修正”循环多次重复，精确度通常能得到提高，然而这样的循环算法十分耗时。

Hu 等人认为，“压缩-激活”是具备注意力和门控的效果。本文提出另外一个角度，“压缩-激活”还具备了“初估计-修正”的效果。Hu 等人认为如果不加入“压缩-激活”模块，卷积核仍有可能学到相似的信息，引入该模块仅仅是为了显式地学到所需注意力。本文指出在全连接地点云处理网络中，该观点不再适用，严格证明了在全连接网络中，“压缩-激活”机制不能被隐含在全连接权重中。

1.2.2 点云处理算法

随着近年来 3D 传感器的普及，点云数据成为常见的视觉数据表示方式。研究者希望不经过数据预处理，直接对点云进行建模，其动机是避免基于预处理的点云分析算法的缺陷：直接将 3 维点投射到不同的平面并栅格化后对于每个投射面采取 2 维卷积的方法[27;28]通常会丢失有用的 3 维信息，直接将 3 维信息栅格化再利用 3 维卷积的方法[29;30]计算量庞大，只能采取小分辨率输入从而丢失细节信息。尽管近些年来也出现了计算效率非常高的 3D 数据处理算法：基于八叉树的算法[31;32]、高效的 4 维空间-时序卷积网络 MinkowskiNet[33]、子流形稀疏卷积[34]，但直接处理点云的算法在 3 维数据的分析算法中仍有很强的竞争力。

点云处理算法的开山之作为 pointnet[17]所有的点通过一系列共享的全连接层后通过最大池化得到全局特征，这样处理的优点在于该特征提取方式是输入点特征的对称函数，即任意改变输入点的顺序输出结果都不会改变，直观上这个要求也是十分合理的。Mo-Net[35]为了更好地捕捉物体几何结构将 pointnet 的输入加入位置的高阶矩。由于 pointnet 仅通过一步最大池化来获取全局特征，其无法准确的利用局部的信息，因此 Qi 等人提出 pointnet++[18]引入层次点云信息粗获取模块获取较细粒度的局部特征。层次点云信息粗获取模块包含了点云降采样层、点云邻

近点分组层、点云信息学习层。基于 pointnet++ 对层次点云信息粗获取模块中三个层进行优化的算法层出不穷：PATs[36] 将近邻点相对中心点的位置作为特征，并将采样算法改进为可微分的 Gumbel 降采样。PointWeb[37] 利用自适应特征调整模块改进点云信息学习层。SRN[5] 通过全连接层把不同位置的局部结构信息关联起来。PointANSL[38] 用自适应采样算法改进点云降采样层，然后利用全局信息融合改进点云信息学习层。

目前仍没有考察“压缩-激活”[39] 在手部点云信息学习中有效性的工作。受[8] 启发，为探究“压缩-激活”机制在手部点云信息学习中的效果，可先将手部的深度图像可先转化为点云表示。本文将理论证明该结构在 pointnet 结构下不可被全连接层隐含表示，通过实验结果说明其有效性。

1.2.3 基于深度图像的手部三维关键点估计

该问题目前的解决方案分三大类：判别式模型[40;41;7;8]，生成式模型[42;43;44] 以及混合模型[45;46;47;15;48;11;49]。判别式模型直接处理输入图像得到结果，尽管算法较为高效，但在处理一些自遮挡程度严重的手部姿态时无法准确捕捉手部的物理约束。生成式模型通常需要引入一个人手模型，通过优化手部模型参数使得人手模型在二维平面投影后与输入图像相近进而得到最终结果，其优化过程十分耗时。混合模型试图在两个方案之间找到折中，利用判别式模型所得结果为生成式模型提供优化时的初始化参数，这本质上也是在进行“初估计-修正”。考虑到判别模型的高效性，以及边缘计算对计算量的苛刻需求，本文基于判别式方法设计模型。

判别式方法可以进一步细分成三类：

1) 基于回归的模型：[10;11;7;8;2;13] 基于回归的模型用神经网络直接回归得到手部关键点位置向量。Oberweger 等人[11;10] 考虑到由于物理约束，手的自由度不会太高，于是在输出层前加入低维全连接层，意图在不修改损失函数的情况下构成隐式的约束。Guo 等人[13] 结合特征图中不同区域所得特征联合估计关键点三维坐标。Ge 等人[7] 认为二维的卷积核的难以有效学到三维结构信息，他们先将深度图像转化为三维的体素表示，再利用三维卷积得到关键点位置估计。受到[17;18] 的启发，Ge 等人[8;9] 舍弃卷积，将深度图像转化为点云后通过 pointnet++ 组件得到关键点位置估计。
2) 基于检测的模型：基于检测的模型[6;16;9;12;50] 中网络输出热图，其中包括二维热图、体素热图、点云热图，通常来说一个关键点对应一个热图，热图可表示关键点位置出现概率或其到真实关键点的向量。Tompson 等人[50] 利用卷积网络得到二维热图后基于逆向运动学模型得到三维坐标，Ge 等人[6] 在在

卷积网络得到多视角的二维热图利用基于优化的后处理得到三维坐标。考虑到从二维热图得到三维坐标过程繁杂，研究者考虑直接获取体素热图或点云热图。Moon 等人[16] 将深度图像转换为体素表示后利用三维卷积降采样再上采样得到体素热图。将热图中的元素可以表示为到关键点的向量，但 Wan 等人[12] 和 Ge 等人[9] 注意到将向量热图转换为单位向量热图与距离热图的组合更利于神经网络进行学习，区别主要在于前者使用的是二维热图表示，后者使用点云热图表示。基于判别的模型往往比基于回归的模型有更高的精确度，然而由于该类模型的架构常出现的下采样后再上采样恢复分辨率模块导致模型参数量过大，以及需要繁杂的后处理算法：如从热图中推断最终三维位置，该算法往往非常耗时，有时难以达到实际应用需求。本文提出的算法在精度上逊于基于检测的模型，但在速度方面表现的更好。

3) 层次模型：考虑到让网络同时学习估计所有关键点的位置或许比较困难，[14;3;4] 将关键点进行分组，对每组关键点分别进行估计。
4) 结构模型：结构模型[10;11;14;47] 显式或者隐式地将物理约束加入模型。Oberweger 等人[10;11] 试图通过低维全连接层建立隐式物理约束，zhou 等人[47] 试图通过在损失函数中添加新的项引入显式约束。

基于模型高效性的考量，本文提出的模型为基于回归的模型。本文认为“压缩-激活”模型在“压缩”的过程中已经将全局信息进行了深度整合，在“激活”时隐式地将某些对于手部物理约束有关键作用的通道增加了权重，因此可以将其理解为隐式的结构化模型。

尽管层次模型考虑对每部分关键点分别估计，但会影响到算法的高效性。而且，本文为基于 Pointnet++ 设计的模型，其中很重要的模块为层次化的信息融合模块，本文认为该模块相较于传统卷积网络的卷积模块具有更大的感受野以及感受野灵活度（不规则的感受野），因此无需额外的层次化估计。

1.3 本文的论文结构与章节安排

本文共分为六章，各章节内容安排如下：

第一章诸论。简单说明了研究将“压缩-激活”机制引入点云处理网络分析并用于基于深度图像手部 3 维关键点估计任务的背景与意义。

第二章对基于“压缩-激活”的深度图像手势识别给出理论解释和直观分析。2.1 节进行问题的阐述与模型结构概述。2.2 节回顾 HandPointnet 的模型架构并指出本文将对其作出的改进。2.3 节回顾“压缩-激活”的基本思路并简述本文将如何

对其作出理论解释以及应用推广。2.4 节从新的角度解释“压缩-激活”模块。2.5 节推广“压缩-激活”机制。2.6 节对压缩机或机制的具体实现作简要阐述并讨论保留点维度的压缩的点置换不变性。2.7 节从理论上说明“压缩-激活”功能的必要性。2.8 节探讨“压缩-激活”和其它模块的互补性。

第三章实验结果。3.1 节对数据集进行简要描述。3.2 节叙述了实验参数设置。3.3 节探究“压缩-激活”模块的结构配置。3.4 节进一步分析“压缩-激活”模块与其它模块的互补性。

第四章对全文作简要总结。

二、基于“压缩-激活”的深度图像手势识别

2.1 问题阐述与模型概述

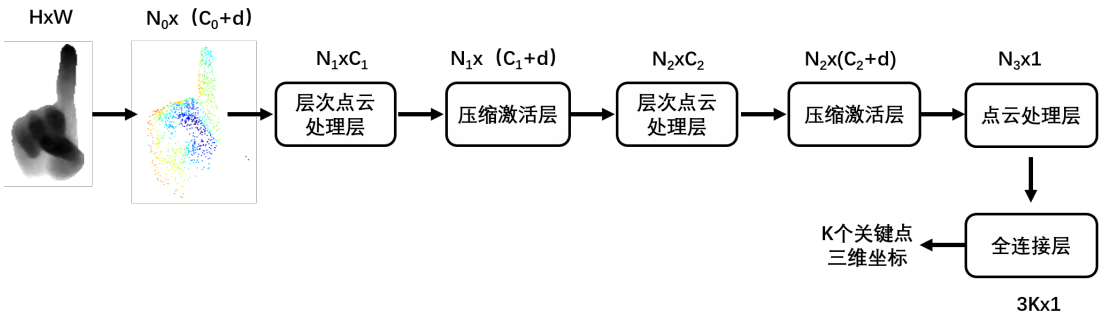


图 2-1 **模型架构**。深度图像先被转化为定向的点云，在经过两次层次点预处理与压缩激活层后经过点云处理层与全连接层最终得到 K 个关键点的 3 维坐标。

输入为深度图像 $I \in \mathbf{R}^{H\times W}$，$I$ 中每个元素表示手部在摄像机坐标系中的深度，我们的目标是根据 I 估计 K 个手部关键点在摄像机坐标系下的三维坐标 $\{x_i, y_i, z_i\}_{i=1}^{K}$。

下面概述模型架构，除了“压缩-激活”层外，其它结构的设计与 Handpointnet 一致，将在 2.2 节作详细介绍。如图 2-1 所示，模型根据摄像机参数将深度图像转化为点云，从点云中我们选取固定数目的 N 个点 $p_i \in \mathbf{R}^3 (i = 1, ..., N)$，然后利用主成分分析的方法将其规范化到一个定向框中。然后这些点送入进入层次点云处理模块[18]，输入模块时通道维度为 $N_0 + d$，这里 $C_0 = 3$ 为表面法向，$d = 3$ 为三维坐标，输出时根据最远距离采样算法降采样得到 N_1 个点的信息。然后进入本文所重点叙述的点云“压缩-激活层”，这里的“压缩-激活”可以是图 2-2(a) 中的压缩点，保留通道，也可以是图 2-2(b) 中的压缩通道，保留点。“压缩-激活”的位置也由两种选择，可以在拼接三维位置前压缩，也可以在拼接三维位置后压缩。再次经过层次点云处理层和“压缩-激活”层后得到 N_2 个点的信息，最后接连进入点云处理层[17] 和全连接层得到 K 个关键点的三维坐标。

本节的剩下部分将如下述安排：

1) 2.2 节简要叙述 HandPointnet[8] 的模型结构以及本文将对其改进的方向。
2) 2.3 节简要介绍“压缩-激活”[19] 的主要思想以及本文对其进行应用推广和理论证明的简要思路。
3) 2.4 节站在与[19] 不同的角度阐述“压缩-激活”的作用。
4) 2.5 节推广[19] 提出的“压缩-激活”机制，并阐述推广后可能的其它应用场景。

5) 2.6 节给出“压缩-激活”机制的具体实现参数。
6) 2.7 节从理论上严格论证全连接 + “压缩-激活”机制在点云处理网络中不能被单一全连接替代的原因。
7) 2.8 节探究“压缩-激活”模块与其它模块进行信息或功能互补的可能性。

2.2 HandPointnet 模型及其缺陷

HandPointnet 的主要思想是把深度图像根据深度相机的内参数转化为点云后利用 Pointnet[18] 对其进行处理得到粗估计关键点 3 维坐标。将该粗估计指尖坐标附近的关键点再次输入类似结构的点云处理网络得到指尖的修正估计。

其中点云处理主要包括降采样，分组，各组信息融合三步。利用最远距离采样算法进行点云降采样，将降采样中每个点最近的 K 个点作为一组。在信息融合时，每组中各个点的特征分别通过相同参数的全连接层后进行最大池化得到该组的特征。在最后一次进行该操作后将所有组的特征分别通过相同参数的全连接层后再进行一次最大池化得到点云的全局特征，将该特征输入若干个全连接层后得到粗估计 3 维关键点坐标。

以上算法主要有几点缺陷。第一，修正模块未加入端到端的训练，会导致最后训练出的网络参数是次最优的，而且这样的修正非常耗时。第二，最远距离采样算法对极端值非常敏感，而且最远距离采样是不可微分的，这同样导致了其无法加入端到端的训练。我们将在 2.4 节说明“压缩-激活”可以嵌入点云处理网络作为中间修正加入端到端的训练，在 2.6 节指出“压缩-激活”模块在进行合适的改变后可以作为可微分降采样模块。

2.3 “压缩-激活”机制简述

Hu 等人[19] 提出的“压缩-激活”主要思想是根据二维卷积网络中的特征图的重要性对其进行加权，使得网络能关注到对输出结果更有利用价值的特征图。

二维卷积时的压缩激活流程如下所述：通过对特征图进行最大池化——压缩，然后把压缩所得向量依次通过全连接层、RELU 激活函数、全连接层、sigmoid 激活函数后就得到用来对特征图进行激活的权重向量，最后利用该权重向量的各个分量与对应通道相乘就完成了通道激活。

关于“压缩-激活”有几个问题目前仍未得到解答。第一，“压缩-激活”是否隐式地被其它模块表示从而没有作用？尽管 Hu 等人[19] 的实验结果不支持这种解释，但仍未有严格的理论解释，我们将在 2.7 节中对某种特殊情况给出否定的答

案。第二，如果其在二维卷积情形下有效，那么它在点云处理中有效么？我们将在 2.5 节与 2.6 节给出答案。第三，除了注意力机制外能否对“压缩-激活”作出其它解释？我们将在 2.4 节中分析其修正作用。第四，它是否能作为一种可微分的降采样方式？我们将在 2.6 节给出具体的实现方案。第五，从形式上看，压缩激活兼具滤波与全局信息融合的功能，那么它与其它滤波模块或全局信息融合模块在点云处理网络中孰优孰劣？若互有优劣，它们能进行功能上的互补么？我们将在 2.8 节对该问题作出具体分析。

2.4 作为修正手段的“压缩-激活”

Hu 等人[19] 指出“压缩-激活”机制本质是注意力机制，一个很自然的问题是：“压缩-激活”机制是否在本质上还有其它的功能，或者说该机制如果应用于如点云回归问题，其本质上实现了另外的功能？

本文认为：“压缩-激活”机制实现了根据不太准确的输出，而对特征进行修正的功能。先把目光暂时移向卷积网络，在卷积网络中，在浅层网络引入侧枝构建损失函数新项以帮助训练是屡见不鲜的手段，这主要是因为浅层网络层输出的特征图已经有了一些基本的必要信息。一个自然的想法是用这些特征图大概地“猜测”一下最后的结果，然后通过加权的方式根据该“猜测”对特征图进行修正。

一个简单的例子是分类卷积神经网络：输入猫的图片，从中间某些特征图大概知道这是个动物，那么就对眼睛，鼻子等生物特征有关通道进行加权。如果没有进行这样的加权，可能出现的情况是：尽管网络不会把这只猫认为是一辆汽车，但有可能因为对生物特征考虑的欠缺猫会被错认成狗。“压缩-激活”起到的作用就是把各通道特征压缩再经过全连接层后得到一个代表“生物”的向量，此即“粗估计”，利用该“粗估计”与通道特征交互进而实现修正。

对于利用 pointnet++ 结构估计深度图像的关键点位置估计而言，“压缩”实质上实现了根据浅层全连接层输出的一些信息大致估计出关键点分布情况并将其降维至一个向量中，根据该粗略猜测的位置所对应的低维向量进行关键信息“激活”。总结来说，本文认为“压缩-激活”可以被认为是“猜测-修正”。

在[11;8;3] 等工作中都采取了修正模块，根据初步预测的关键点位置提取更有用的关键点信息，他们采取了显式的修正，计算量和参数量的增加都十分显著。而本文提出的“压缩-激活”事实上是采取了隐式的修正，这对参数量和计算量的增加很小。

2.5 广义的“压缩-激活”机制

Hu 等人[19] 指出卷积神经的 “压缩-激活” 机制，下面在给出点云网络中的 “压缩-激活” 机制前，我们先给出 “压缩-激活” 机制更普遍的形式。

“压缩-激活” 映射为 $F : X \to \tilde{X}$。其中 X 和 $\tilde{X}$ 为 $N_1 \times N_2 \times ... \times N_D$ 的 D 维张量。插入 2 维卷积神经网络的 “压缩-激活” $D = 3$。插入点云处理网络 pointnet++ “压缩-激活” $D = 2$。

选取一个压缩后保留的维度序号 k。“压缩” 操作将 X 压缩成为 N_k 维向量 X^{sq}，其中:

$$X_i^{sq} = f_{sq}(X_i) \tag{2.1}$$

X_i 为 X 在第 k 维上的第 i 个元素。f_{sq} 可以是任何一个将 $N_1 \times N_2 ... \times N_{k-1} \times N_{k+1} ... \times N_D$ 张量转化为标量的映射，为了后文叙述方便，也记 $X^{sq} = f_{sq}(X)$。Hu 等人[19] 用平均池化作为 f_{sq}，本文同样采取平均池化。平均池化对于二维或一维的张量来说是常用的降采样操作，但如果在较浅的网络层进行全局平均池化，或是对于更高维的张量（如视频为 3 维）有可能会丢失有用的信息，此时选取细粒度的 f_{sq} 可以更好的保留信号，对于细粒度 f_{sq} 的设计留给未来的工作完成。

事实上，保留维度可以是小于 D 的任意正整数。但本文认为，有两个原因使得在 Hu 等人[19] 的工作中仅仅保留通道这一个维度。第一，我们有时仅关注属于某一个维度的元素的特征修正，对于卷积网络，我们更关心在每个通道上修正元素特征。第二，通道、高、宽这三个维度并非是各维度同性或者说各向同性的。对于图像处理，模型上基本上会假设高和宽维度是同性的，因此我们不在这两个维度上做文章。但如果对于某些图像的分类来说，有用位置信息成明显的纵向分布或是明显的横向分布，如不同类型树的精细分类和不同品牌汽车的精细分类，对应保留高和保留宽或许会是不错的想法。对于视频处理，通道、高、宽、时间，可任选少于三个维度进行保留，对于时间维度的保留能让我们关注到时序内更加有用的信息。第三，保留一维信息后可直接利用全连接层进行 “激活” 操作。保留多维信息的 “压缩-激活” 设计留给未来的工作去完成。本文重点讨论点云网络 pointnet++ 上的一维保留，压缩时保留点列维度和保留通道维度。

“激活” 操作与[19] 一样。首先得到权重向量 $s \in \mathbf{R}^{N_1}$，再利用权重向量 s 对 X 第 1 维上的元素进行加权，从而达到隐式修正的效果:

$$s = f_{ex}(X^{sq}, W_1, W_2) = \sigma(W_2 \delta(W_1 X^{sq})) \tag{2.2}$$

$$\tilde{X}_i = s_i \cdot X_i \tag{2.3}$$

若采取残差连接，

$$\tilde{X}_i = (s_i + 1) \cdot X_i \tag{2.4}$$

其中 $W_1 \in \mathbf{R}^{\frac{N_k}{r} \times N_k}$ $W_2 \in \mathbf{R}^{N_k \times \frac{N_k}{r}}$，$r$ 为通道压缩倍率，δ 为 RELU 激活函数，σ 为 sigmoid 激活函数

2.6 压缩-激活机制的具体实现

2.6.1 卷积网络中的“压缩-激活”机制

如图 2-2(c), 取 $D = 3, N_1 = C, N_2 = H, N_3 = W, k = 1$，其中 C 为通道数，H 为特征图高，W 为特征图宽。并取 f_{sq} 为平均池化。即得到[19] 中卷积神经网络中的“压缩-激活”结构。图例中白色的通道表示激活层度高，黑色的通道表示激活程度低。激活后，无用的通道信息被滤去，有用的通道信息得以保留。

2.6.2 点云网络中的“压缩-激活”机制

如图 2-2(a)(b)，取 $D = 2, N_1 = N, N_2 = C$，其中 C 为通道数，N 为经过层次处理模块后降采样得到的点的个数。

取保留维度为 $k = 1$ 即得到如图 2-2(b) 所示的点激活模块。取保留维度为 $k = 2$ 即得到如图 2-2(a) 所示的通道激活模块。

如果我们对激活向量进行核大小为 n 的最大池化并记录下对应的最大元素的所在序号，将 $X \in \mathbf{R}^{N \times C}$ 的对应序号行进行激活。那么此时点激活模块便起到了可微分降采样的效果，把采样过程放到训练中，使得采样过程可进入端到端训练，这与[36] 的本质思路是一致的，但参数量更小，模型更轻便。但是点激活有其不可避免地缺点：它不是点置换不变的，此即下面的定理。（关于点置换不变的直观叙述，请读者参考 pointnet[17]）

定理 2.1 (**点激活不是点置换不变的**) *存在 $W_1 \in \mathbf{R}^{\frac{N}{r} \times N}$，$W_2 \in \mathbf{R}^{N \times \frac{N}{r}}$，$N$ 阶置换矩阵 $P \in \mathbf{R}^{N \times N}$，$X \in \mathbf{R}^{N \times C}$ 使得*

$$\tilde{Y} \neq P\tilde{X} \tag{2.5}$$

其中 $\tilde{Y}$，$\tilde{X}$ 由下面的式子确定：

$$s = f_{ex}(f_{sq}(X)) = \sigma(W_2\delta(W_1 f_{sq}(X))) \tag{2.6}$$

$$\tilde{X}_i = s_i \cdot X_i \tag{2.7}$$

$$Y = PX \tag{2.8}$$

$$s_Y = f_{ex}(f_{sq}(Y)) = \sigma(W_2\delta(W_1 f_{sq}(Y))) \tag{2.9}$$

$$\tilde{Y}_i = (s_Y)_i \cdot Y_i \tag{2.10}$$

f_{ex}，δ，σ 如前所述，f_{sq} 为如前所述保留点的压缩。

证明 取 P 为 N 阶单位矩阵互换第 1，2 行所得矩阵，再取 W_1，W_2，X 左上角元素为 1 其余元素为 0。容易验证这样的取法满足条件。 □

值得指出，我们无法确定网络是否有可能在学习的过程中学习到有点置换不变性，或近似能达到点置换不变效果的 W_1, W_2。尽管如此，我们无法确信网络能学到这点，后面的实验表明点激活可能确实无法很好学到这点。因此本文认为通道激活或许在点云网络的“压缩-激活”机制中是更合适的。

2.7 “压缩-激活”的功能必要性的理论证明

工作[19] 指出“压缩-激活”的功能被隐式地包含在卷积核中，言外之意是卷积与“压缩-激活”两层组合的功能有可能会被单一的卷积操作取代。这里的取代指近似或完全等同。他们认为采取“压缩-激活”的目的仅仅是显式地保证网络能注意到某些特征，用本文地话说是能够做到“猜测-修正”。而本文的其中一个目的是说明在 $D = 2$ 的情况下，即点云网络中的“压缩-激活”是无法被隐式地包含在全连接层中的，这就是下面的定理。为了方便叙述证明的主要思路，这里忽略全连接层后的批归一化层与激活层以及偏置。下面先给出本节的主要结果 (不失一般性，给出保留通道压缩情况下的叙述)。

定理 2.2 ($D = 2$ 时“压缩-激活”的不可隐式表示性) 不存在 $W_{sub} \in \mathbf{R}^{C\times C}$ 使得 $\forall X \in \mathbf{R}^{C\times N}$ 成立

$$\tilde{X} \equiv W_{sub}X \tag{2.11}$$

其中 $\tilde{X}$ 由下面的式子确定:

$$s = f_{ex}(f_{sq}(W_{fc}X)) = \sigma(W_2\delta(W_1 f_{sq}(W_{fc}X))) \tag{2.12}$$

$$\tilde{X}_i = s_i \cdot (W_{fc}X)_i \tag{2.13}$$

f_{ex}，δ，σ 如前所述，f_{sq} 为如前所述保留通道的压缩，$W_{fc} \in \mathbf{R}^{C\times C}$ 为任意给定的满秩矩阵，$W_1 \in \mathbf{R}^{\frac{C}{r}\times C}$ 为任意给定的满秩矩阵，$W_2 \in \mathbf{R}^{C\times\frac{C}{r}}$ 为任意给定的非零矩阵，$C \leq N$，角标 i 表示第 i 行元素。

定理中要求 W_{fc}，W_1 满秩，这是非常合理且容易达到的要求，这是因为它们的行向量表示了全连接层将原来的 C 维特征转化为更深层特征的其中一个元素时的权重。所有的行向量必须线性无关才能使得更深层特征相互区别开。类比于卷积网络，不同的卷积核必须做到无法被其它卷积核线性表示，否则这个卷积核应该直接去掉——我们没必要花更多的计算量和参数量计算一个完全无用的卷积核。W_2 不为 0 矩阵的要求就更为宽松了。下面先给出几个引理，最后再给出主要定理的证明。

引理 2.1　若 $W_2 \in \mathbf{R}^{C\times\frac{C}{r}}$ 第 i 行不为 0 向量，$W_1 \in \mathbf{R}^{\frac{C}{r}\times C}$ 满秩，则 $\exists x \in \mathbf{R}^C$ 使得:

$$(W_2\delta(W_1 x))_i \neq 0 \tag{2.14}$$

下标 i 仍如前文一样表示第 i 行，δ 表示逐元素的 RELU 激活函数。

证明 设 W_2 第 i 行第 j 列的元素非 0，则令 $y \in \mathbf{R}^{\frac{C}{r}}$ 为第 j 行为 1，其它元素为 0 的向量，则

$$(W_2 y)_i \neq 0 \tag{2.15}$$

下面只需说明 $\exists x \in \mathbf{R}^C$ 使得

$$\delta(W_1 x) = y \tag{2.16}$$

由于 δ 为 RELU 函数，我们只要找到 x 使得

$$W_1 x = y \tag{2.17}$$

由于 $r(W_1) = \frac{C}{r} < dim(x) = C$ 且根据 W_1 满秩，有 $r([W_1, y]) = r(W_1)$。故上述关于 x 的线性方程组有解，引理得证。 □

引理 2.2　$\forall x \in \mathbf{R}^C$，如果 $W_{fc} \in \mathbf{R}^{C\times C}$ 满秩，$C \leq N$，则存在 $X \in \mathbf{R}^{C\times N}$ 满

秩使得

$$f_{sq}(W_{fc}X) = x \tag{2.18}$$

证明 如果我们找到这样一个 $Y \in \mathbf{R}^{C\times N}$ 满秩使得

$$f_{sq}(Y) = x \tag{2.19}$$

那么由于 W_{fc} 为满秩方阵，我们可以直接令

$$X = W_{fc}^{-1}Y \tag{2.20}$$

显然这样的 X 满足条件。下面我们来找出这样的 Y。事实上，由于 f_{sq} 为平均池化函数，我们只需令 $Y_{ii} = N \cdot x_i (0 < i \leq C)$，$Y$ 的其它元素为 0，容易验证这样的 Y 满足条件。 □

引理 2.3 若 f_{ex}，f_{sq}，δ，σ 如前所述，$W_{fc} \in \mathbf{R}^{C\times C}$ 满秩，$W_1 \in \mathbf{R}^{\frac{C}{r}\times C}$ 满秩，$W_2 \in \mathbf{R}^{C\times \frac{C}{r}}$ 第 i 行不为 0 向量，$C \leq N$，则存在 $X \in \mathbf{R}^{C\times N}$ 满秩且使得

$$s_i \neq s_i^{'} \tag{2.21}$$

其中角标 i 表示第 i 行元素，s 与 $s^{'}$ 由下面两式确定

$$s = f_{ex}(f_{sq}(2W_{fc}X)) = \sigma(W_2\delta(W_1 f_{sq}(W_{fc}X))) \tag{2.22}$$

$$s^{'} = f_{ex}(f_{sq}(W_{fc}X)) = \sigma(W_2\delta(W_1 f_{sq}(2W_{fc}X))) \tag{2.23}$$

证明 由引理 2.1，引理 2.2，我们可找出这样一个满秩的 X 使得

$$W_2\delta(W_1 f_{sq}(W_{fc}X))_i \neq 0 \tag{2.24}$$

再根据 RELU 和平均池化映射的性质，故有：

$$\begin{aligned}
&W_2\delta(W_1 f_{sq}(2W_{fc}X))_i \\
&= W_2\delta(2W_1 f_{sq}(W_{fc}X))_i \\
&= 2W_2\delta(W_1 f_{sq}(W_{fc}X))_i \\
&\neq W_2\delta(W_1 f_{sq}(W_{fc}X))_i
\end{aligned}$$

再根据 sigmoid 函数 σ 的严格单调性便有

$$s_i \neq s_i^{'} \tag{2.25}$$

□

下面正式给出本节主要定理的证明:

证明 如果这样的 W_{sub} 确实存在，我们下面要通过说明 W_{fc} 有一行为 0 从而不满秩来导出矛盾。不失一般性，令 W_2 的第 i 行不是 0 向量。由于满足引理 2.3 的条件，我们可以取一个使得 $s_i \neq s_i^{'}$ 的 X。用 X_i 表示 X 第 i 行，由 2.11 与 2.13 式有:

$$(W_{sub}X)_i = s_i \cdot (W_{fc}X)_i \tag{2.26}$$

$$(W_{sub})_i X = s_i \cdot (W_{fc})_i X \tag{2.27}$$

$$(W_{sub} - s_i \cdot (W_{fc}))_i X = 0 \tag{2.28}$$

由于 X 的秩为 C，故关于 w 的方程组，$wX = 0$ 只有零解 $w = 0$，即

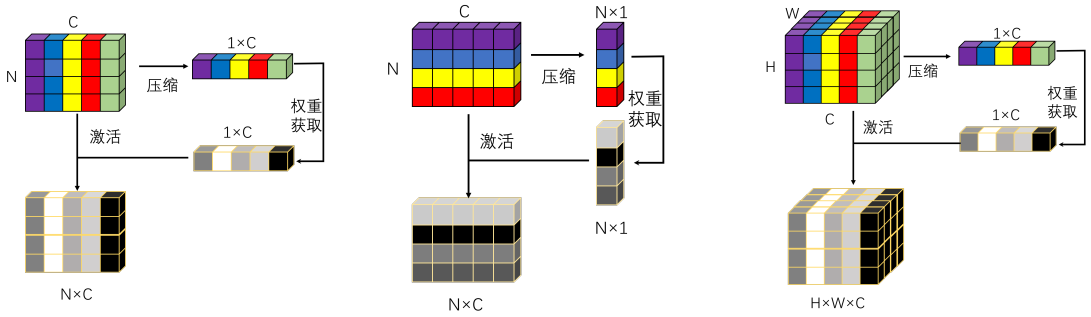
$$(W_{sub})_i = s_i \cdot (W_{fc})_i \tag{2.29}$$

在 2.12 式中将 X 替换成 $2X$，对应激活权重变为 $s_i^{'}$，通过类似的推导可以得到

$$(W_{sub})_i = s_i^{'} \cdot (W_{fc})_i \tag{2.30}$$

由于 $s_i \neq s_i^{'}$，再由 2.29 与 2.30 式立刻得到 $(W_{fc})_i = 0$ 这与 W_{fc} 满秩矛盾，故不存在这样的 W_{sub}。 □

由此我们说明了引入“压缩-激活”模块一个很重要的原因是其不能被隐含在单一的全连接层中。



(a) 点云网络中保留通道维度的“压缩-激活”。(b) 点云网络中保留点维度的“压缩-激活”。(c) 卷积神经网络中的“压缩-激活”。

图 2-2 **“压缩-激活”的不同实现方式。**

2.8 “压缩-激活”与其它模块

在 HandPointnet[8] 中，作者进行了非常细致的超参数选择，我们在其原始的模型上维持了超参的不变。这有一个很大的隐患，即模型本身已处于过拟合边缘，且没有明确的实验结果说明残差连接[39] 对于全连接的点云处理网络也有明显效果，因此在插入组件后很有可能导致过拟合。因此将“压缩-激活”模型直接与其它现有模型对比是不公平的。故在 HandPointnet 中插入图卷积[51] 或全局信息融合模块[52] 的一个重要原因是给“压缩-激活”模块提供较为公平的比对。

另外，“压缩-激活”模块只能根据全局信息对输入信息在通道方向或点方向进行线性的修正，一个自然的问题是：引入能提供较高非线性的信息融合模块能否与“压缩-激活”模块产生互补的效果？

同时，“压缩-激活”模块尽管根据全局信息对每点信息进行了改变，但这是非常隐式的。尽管层次点云处理模块提供了显式的局部信息融合功能，其方式是邻近点通过共享的全连接后进行最大池化，但我们希望知道其它局部信息融合的方式是否会对层次点云处理模块提供的功能有所补充或改进。

2.8.1 “压缩-激活”与图卷积网络能否功能互补？

图卷积网络常用于处理非栅格化数据，主要分为频域上的图卷积[53;54;51] 与空间域图卷积[51;55;56]。kipf 等人提出的图卷积[51] 可作为频域上任意图卷积的一阶近似，其表达形式也能解释为空间图卷积，本文采取该种算法。选取该种图卷积的方式是处于以下考虑：首先，点云的获取与降采样不可避免存在一些噪声，从频域上过滤掉这些噪声十分必要。其次，由于层次点云处理层已能融合部分的局部信息，我们尽管需要，但没必要过分强调图卷积层在空间域上融合局部信息的功能。

我们希望了解图卷积的降噪效果与根据关联性融合局部信息的功能是否能给

“压缩-激活”模块的功能提供补充。

下面仅给出模块的信息处理方式，其它设计考虑请读者参考[51]。该模块接受 $X \in \mathbf{R}^{N\times C}$ 作为输入，输出 $\tilde{X} \in \mathbf{R}^{N\times C}$ 由下面式子给出：

$$\tilde{X} = \tilde{D}^{-\frac{1}{2}}\tilde{A}\tilde{D}^{-\frac{1}{2}}X\Theta \tag{2.31}$$

其中 $\tilde{A} \in \mathbf{R}^{N\times N}$ 为单位矩阵与邻接矩阵之和，邻接矩阵第 i 个点与第 j 个点连接度由 $a_{ij} = e^{-d(x_i,x_j)}$ 确定，$\tilde{D} \in \mathbf{R}^{N\times N}$ 为对角矩阵，$\tilde{d}_{ii} = \sum_j \tilde{a}_{ij}$，$\Theta \in \mathbf{R}^{C\times C}$ 为待学习参数在本文实验中该模块仅设置 1 层至 2 层，不必太担心梯度爆炸或梯度消失问题，可尝试忽略规范化操作：

$$\tilde{X} = \tilde{A}X\Theta \tag{2.32}$$

2.8.2 “压缩-激活”与全局信息融合模块能否功能互补？

全局信息融合模块[52] 在视频分类，目标检测与点云识别[38] 任务上取得了很好的效果，其作者认为这是补充了卷积核不能迅速融合全局信息的缺陷。

尽管“压缩-激活”也利用到了全局信息，但其仅利用其进行单个信号增强与减弱，我们试图用全局信息融合模块进行信号的显式融合，希望了解这样的方式能否补充“压缩-激活”模块的功能。

下面仅给出模块的信息处理方式，其它设计考虑请读者参考[52]。接受 $X \in \mathbf{R}^{C\times N}$ 作为输入，输出 $\tilde{X} \in \mathbf{R}^{C\times N}$ 由下面式子给出

$$\tilde{X} = X + (W_{out}(W_g X + B_g)\,\text{softmax}(X^T W_\theta^T W_\phi X) + B_{out}) \tag{2.33}$$

其中 $W_\theta, W_\phi, W_g \in \mathbf{R}^{\frac{C}{2}\times C}, W_{out} \in \mathbf{R}^{C\times\frac{C}{2}}, B_g \in \mathbf{R}^{\frac{C}{2}\times N}$ 列向量相同，$B_{out} \in \mathbf{R}^{C\times N}$ 列向量相同，都为待学习参数

在[52] 中没加入规范化操作，后面实验将表明这样的方式会增加网络学习的困难，很大程度降低模型表现。受到[57] 启发，我们进行规范化

$$\tilde{X} = X + (W_{out}(W_g X + B_g)\,\text{softmax}(\frac{X^T W_\theta^T W_\phi X}{\sqrt{\frac{C}{2}}}) + B_{out}) \tag{2.34}$$

三、实验结果

3.1 数据集

我们在 MSRA[41] 数据集上进行实验。该数据集包含了 7.6 万多张手部深度图像，所有深度图像来自 9 位志愿者，每个志愿者都会做出 17 个不同的手势，图3-1展示了部分图像。数据集中已提供手部边框，并给出 21 个关键点的位置。我们仅在 0 号个体上进行测试，在 1-8 号个体上进行训练。

(a) 不同个体的同一手势。

(b) 同一个体的不同手势。

图 3-1 **MSRA 数据集中的部分深度图像可视化**。颜色越深表示在相机参照系下深度越深。

3.2 实验参数设置

每个点的法向通过对该点的近邻点三维坐标做主成分分析并选取最小奇异值对应的特征向量近似获得，初始采样点 1024 个，第一次层次处理降采样至 512 个点，第二次层次处理降采样至 128 个点。层次点云处理层第一和二层的球半径分别设为 0.015 和 0.4，K 邻近点为 64 个。取平方偏差作为损失函数。

训练时共进行 90 代训练，批大小设置为 32。选择 Adam[58] 作为优化算法，$\beta_1 = 0.5$,$\beta_2 = 0.999$，采取 0.001 作为学习率，每隔 50 代训练学习率衰减至原来的 0.1 倍。网络参数量和乘加操作数 (MACs) 由开源代码①计算得到。模型利用

① https://github.com/Lyken17/pytorch-OpCounter

pytorch 框架在一张 GPU 上实现。对下文所述的不同模型设置进行实验所需的每次完整训练时间大致为 2 天，因具体模型设置以及所用 GPU 不同有一定区别。

3.3 “压缩-激活”模块自身对比试验

3.3.1 点压缩与通道压缩选择

表3.1展示了两种不同压缩方式对最终结果的影响。对点压缩的方式比不加“压缩-激活”模块的误差更大，我们认为可能的一个解释是对点压缩不是具有点置换不变性的压缩方式，这给网络的学习和推断都带来了困难，这降低了模型表现，然而“压缩-激活”本身是对点信息的修正是有效的，两方面综合作用的结果是精确度仅下降了 0.04mm。对通道压缩的方式有效地降低了 0.07mm 的误差，又由于对通道压缩是点置换不变的，因此这可以被看作“压缩-激活”的净提升，对点压缩因置换变化的影响被降低了 0.11mm 的精度。可以看出两种方式对参数量的增加都较小，对运算量都增加微乎其微。

3.3.2 是否在位置信息合并前压缩

表3.2展示了压缩激活模块与拼接原始点位置信息的模块先后顺序对最终结果的影响。可以看到在位置信息合并前压缩能够有效降低平均误差，然而在位置合并后压缩会增加 0.19mm 的误差。我们认为一个可能的解释如下：将全连接层直接置于位置压缩之后能最大限度地将位置信息与其它通道信息进行融合，然而在经过压缩激活模块后，由于无法保证“压缩-激活”模块能很好地对位置通道的信息进行放缩（即“压缩-激活”模块在浅层网络中对权重预估的不准确），全连接层对位置信息与其它通道信息融合的效率因此受到不利影响。

。

“压缩-激活”方式	平均误差（mm）	参数量 (M)	MACs(G)
对点压缩	8.18	2.4425	1.0878
对通道压缩	**8.07**	2.4176	1.0878
基准模型	8.14	2.4070	1.0877

表 3.1 **通道压缩方式对比**。通道缩减倍率均为 16，并在位置信息合并前压缩。模型在 MSRA 数据集的 1-8 号训练，在 0 号上测试。

“压缩-激活”模块插入位置	平均误差（mm）	参数量 (M)	MACs(G)
在位置信息合并前压缩	**8.07**	2.4176	1.0878
在位置信息合并后压缩	8.33	2.4178	1.0878
基准模型	8.14	2.4070	1.0877

表 3.2 **模块插入位置影响**。通道缩减倍率均为 16，采用对通道压缩的方式。模型在 MSRA 数据集的 1-8 号训练，在 0 号上测试。

“压缩”后通道缩减倍率	平均误差（mm）	参数量 (M)	MACs(G)
x16 通道缩减	8.33	2.4178	1.0878
x8 通道缩减	8.32	2.4282	1.0878
x4 通道缩减	8.62	2.4490	1.0878
基准模型	**8.14**	2.4070	1.0877

表 3.3 **“压缩”后全连接层通道缩减倍率影响**。在信息合并后压缩，采用对通道压缩的方式。模型在 MSRA 数据集的 1-8 号训练，在 0 号上测试。

3.3.3 模块中通道缩减倍率影响

表3.3展示了在“激活”操作时通道缩减率对位置信息的影响。这里所用的是对点压缩的方式，误差均比基准模型要大，这说明置换不变性确实对提升模型的表现非常重要。误差最低的是 8 倍的通道衰减率，这与[19] 所报告的利用卷积网络进行图片分类的实验的结果是一样的。有可能的一个解释是 4 倍缩减率时模型出现了过拟合。

第一层插入结构	第二层插入结构	平均误差（mm）	参数量 (M)	MACs(G)
	”压缩-激活”	**8.07**	2.4176	1.0878
”压缩-激活”	图卷积	9.35	2.4749	1.0961
	全局信息融合	19.63	2.5408	1.1045
	”压缩-激活”	10.21	2.4319	1.0961
图卷积	图卷积	10.90	2.4893	1.1044
	全局信息融合	8.99	2.5552	1.1128
	”压缩-激活”	13.95	2.4485	1.1045
全局信息融合	图卷积	15.48	2.4485	1.1045
	全局信息融合	15.01	2.5718	1.1212

表 3.4 **“压缩-激活”模块与其它模块的配合效果**。采用保留通道维度的激活方式，图卷积和全局信息融合采取规范化处理。模型在 MSRA 数据集的 1-8 号训练，在 0 号上测试。

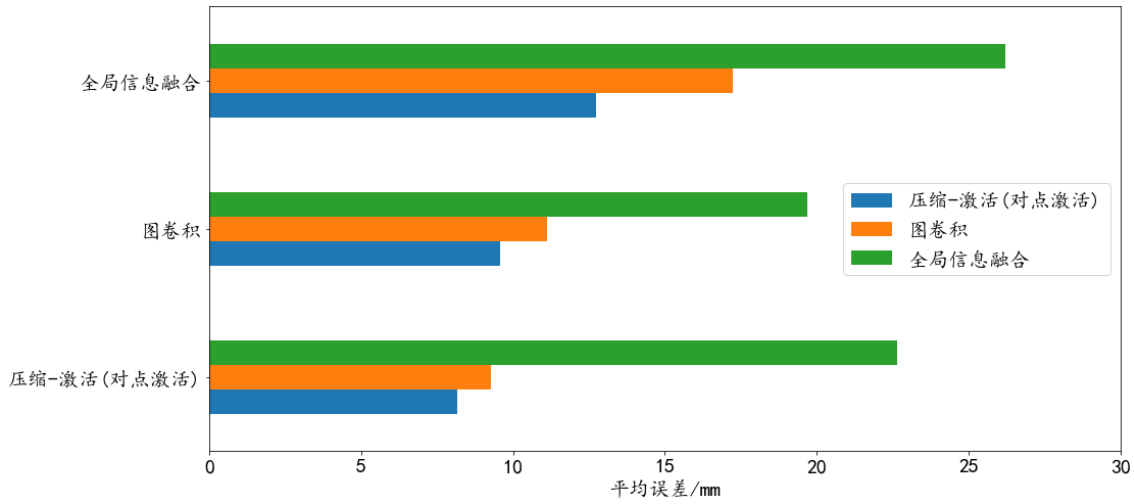


图 3-2 **“压缩-激活”模块与其它模块的配合效果**。采用保留点维度的激活方式，图卷积和全局信息融合不采取规范化处理。模型在 MSRA 数据集的 1-8 号训练，在 0 号上测试。

3.4 “压缩-激活”模块与其它模块的对比与配合

3.4.1 压缩激活与图卷积

从表3.4与图3-2，3-3可以看出，图卷积并不能与压缩激活模块产生很好的互补效果。一个可能的解释是压缩激活模块通过权重分配进行降噪的效果比图卷积试图用一阶近似直接在频域上过滤信号的效果好。并且压缩激活模块能根据全局的信息进行降噪，但图卷积仅根据邻域的信息进行降噪，这种降噪方式很可能受到邻域中异常点的影响，故模型的鲁棒性无法得到保证。有趣的是先压缩激活再进行图卷积的效果比先进行图卷积再压缩激活的效果好，一个可能的解释是在点云信息经过压缩激活后已经得到了有效的降噪处理，第二层加入的图卷积承担的更多是融合局部信息的功能，而不是频域上进行滤波的一阶近似。

表3.5表示了规范化操作对图卷积的影响，可以看出尽管两层不采取规范化处理直观上并不会带来很严重的梯度爆炸或梯度消失的情况但确实会增加误差。对比图3-2，3-3，可以看出图卷积是否进行规范化仅对其与全局信息融合的模型表现会产生影响，但对压缩激活模型几乎在配合上没有产生太大影响。一个可能的解释是全局信息融合模块本身受是否规范化影响极大，因此未规范化图卷积会对梯度规模产生的影响会很大程度降低全局信息融合模块的表现。然而压缩激活模型可以调整通道激活量的大小从而调节梯度大小，以抵消未规范化图卷积带来的梯度爆炸或梯度消失的现象。

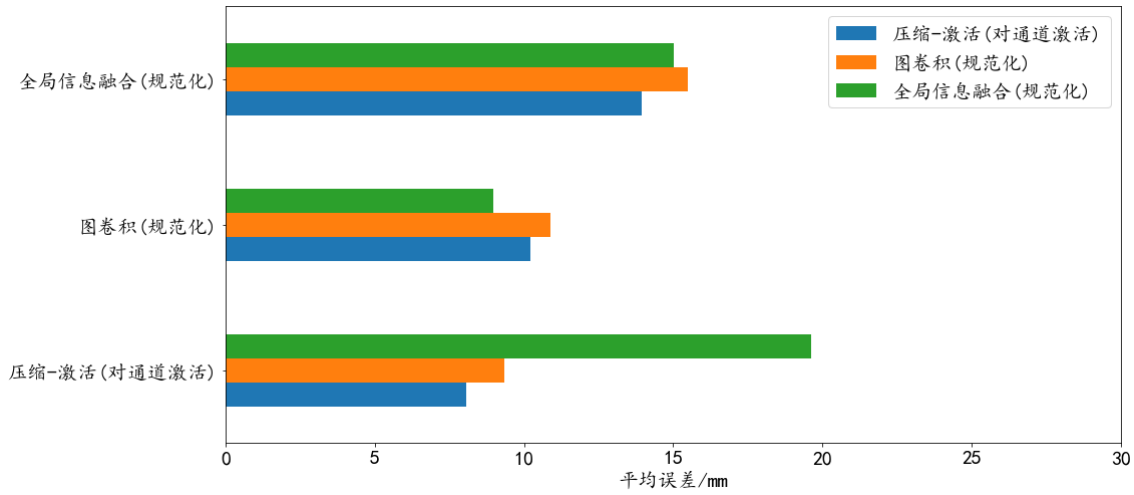


图 3-3 **“压缩-激活”模块与其它模块的配合效果**。采用保留通道维度的激活方式，图卷积和全局信息融合采取规范化处理。模型在 MSRA 数据集的 1-8 号训练，在 0 号上测试。

是否采取规范化	模型描述	平均误差（mm）
是	$\tilde{X}=\tilde{D}^{-\frac{1}{2}}\tilde{A}\tilde{D}^{-\frac{1}{2}}X\Theta$	**10.90**
否	$\tilde{X}=\tilde{A}X\Theta$	11.14

表 3.5 **规范化对图卷积模块的影响**。输出通道与输入通道相同，采用距离衰减邻接矩阵，两个层次点云处理模块后均采用图卷积。模型在 MSRA 数据集的 1-8 号训练，在 0 号上测试。

3.4.2 压缩激活与全局信息融合

从表3.4与图3-2，3-3可以看出，全局信息融合并不能与压缩激活模块产生很好的互补效果。且全局信息融合本身的表现即使在规范化后也不尽理想。这与[38]所报告的在点云分类上全局信息融合的表现不尽相同。一个可能的解释是我们当前的目标是得到手部各个位置关键点的信息，全局信息对于局部关键点的回归并没有太大帮助，反而增加了模型复杂度导致过拟合。尽管压缩激活模块通过全局信息获取通道或点的权重，但其参数量较低，过拟合的程度不如全局信息融合模块严重。表3.6表示了规范化对全局信息融合模型表现的影响，可以看出不采取规范化会严重降低模型表现。但即使在采取规范化后，其仍然难以与“压缩-激活”模块进行很好的互补作用。这说明对于手部关键点回归来说，压缩激活模块更适合作为全局信息的捕获方式。

是否采取规范化	模型描述	平均误差（mm）
是	$\tilde{X}=X+(W_{out}(W_gX+B_g)\,\mathrm{softmax}(\frac{X^TW_\theta^TW_\phi X}{\sqrt{\frac{C}{2}}})+B_{out})$	**15.01**
否	$\tilde{X}=X+(W_{out}(W_gX+B_g)\,\mathrm{softmax}(X^TW_\theta^TW_\phi X)+B_{out})$	26.22

表 3.6 **规范化对全局信息融合的影响**。输出通道与输入通道相同，两个层次点云处理模块后均采用全局信息融合模块。模型在 MSRA 数据集的 1-8 号训练，在 0 号上测试。

四、结论

本文主要研究从手部深度图像估计其 3 维关键点坐标，将“压缩-激活”机制引入点云处理网络。文章推广了“压缩-激活”机制，说明其在保留点维度的情况下是一种可微分的降采样手段，并说明其在保留点维度情况下非点置换不变。理论证明了其在点云处理网络中不能被全连接层隐含表示。通过实验说明了“压缩-激活”在保留通道维度情况下对模型表现有一定提升以及“压缩-激活”与图卷积和全局信息融合模块无法很好地相互补充。

参考文献

[1] KRIZHEVSKY A, SUTSKEVER I, HINTON G E. Imagenet classification with deep convolutional neural networks[J]. Advances in neural information processing systems, 2012, 25 : 1097 – 1105.

[2] CHEN X, WANG G, ZHANG C, et al. Shpr-net: Deep semantic hand pose regression from point clouds[J]. IEEE Access, 2018, 6 : 43425 – 43439.

[3] CHEN X, WANG G, GUO H, et al. Pose guided structured region ensemble network for cascaded hand pose estimation[J]. Neurocomputing, 2020, 395 : 138 – 149.

[4] DU K, LIN X, SUN Y, et al. Crossinfonet: Multi-task information sharing based hand pose estimation[C] // Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019 : 9896 – 9905.

[5] DUAN Y, ZHENG Y, LU J, et al. Structural relational reasoning of point clouds[C] // Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019 : 949 – 958.

[6] GE L, LIANG H, YUAN J, et al. Robust 3d hand pose estimation in single depth images: from single-view cnn to multi-view cnns[C] // Proceedings of the IEEE conference on computer vision and pattern recognition. 2016 : 3593 – 3601.

[7] GE L, LIANG H, YUAN J, et al. 3d convolutional neural networks for efficient and robust hand pose estimation from single depth images[C] // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017 : 1991 – 2000.

[8] GE L, CAI Y, WENG J, et al. Hand pointnet: 3d hand pose estimation using point sets[C] // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018 : 8417 – 8426.

[9] GE L, REN Z, YUAN J. Point-to-point regression pointnet for 3d hand pose estimation[C] // Proceedings of the European Conference on Computer Vision (ECCV). 2018 : 475 – 491.

[10] OBERWEGER M, WOHLHART P, LEPETIT V. Hands deep in deep learning for hand pose estimation[J]. arXiv preprint arXiv:1502.06807, 2015.

[11] OBERWEGER M, LEPETIT V. Deepprior++: Improving fast and accurate 3d hand pose estimation[C] // Proceedings of the IEEE international conference on computer vision Workshops. 2017 : 585 – 594.

[12] WAN C, PROBST T, VAN GOOL L, et al. Dense 3d regression for hand pose estimation[C] // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018 : 5147 – 5156.

[13] GUO H, WANG G, CHEN X, et al. Towards good practices for deep 3d hand pose estimation[J]. arXiv preprint arXiv:1707.07248, 2017.

[14] MADADI M, ESCALERA S, BARÓ X, et al. End-to-end global to local cnn learning for hand pose recovery in depth data[J]. arXiv preprint arXiv:1705.09606, 2017.

[15] YE Q, YUAN S, KIM T-K. Spatial attention deep net with partial pso for hierarchical hybrid hand pose estimation[C] // European conference on computer vision. 2016 : 346 – 361.

[16] MOON G, CHANG J Y, LEE K M. V2v-posenet: Voxel-to-voxel prediction network for accurate 3d hand and human pose estimation from a single depth map[C] // Proceedings of the IEEE conference on computer vision and pattern Recognition. 2018 : 5079 – 5088.

[17] QI C R, SU H, MO K, et al. Pointnet: Deep learning on point sets for 3d classification and segmentation[C] // Proceedings of the IEEE conference on computer vision and pattern recognition. 2017 : 652 – 660.

[18] QI C R, YI L, SU H, et al. Pointnet++: Deep hierarchical feature learning on point sets in a metric space[J]. arXiv preprint arXiv:1706.02413, 2017.

[19] HU J, SHEN L, SUN G. Squeeze-and-Excitation Networks[C] // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2018.

[20] ITTI L, KOCH C. Computational modelling of visual attention[J]. Nature reviews neuroscience, 2001, 2(3) : 194 – 203.

[21] ITTI L, KOCH C, NIEBUR E. A model of saliency-based visual attention for rapid scene analysis[J]. IEEE Transactions on pattern analysis and machine intelligence, 1998, 20(11) : 1254 – 1259.

[22] OLSHAUSEN B A, ANDERSON C H, VAN ESSEN D C. A neurobiological model of visual attention and invariant pattern recognition based on dynamic routing of information[J]. Journal of Neuroscience, 1993, 13(11) : 4700 – 4719.

[23] MNIH V, HEESS N, GRAVES A, et al. Recurrent models of visual attention[J]. arXiv preprint arXiv:1406.6247, 2014.

[24] LAROCHELLE H, HINTON G E. Learning to combine foveal glimpses with a third-order boltzmann machine[J]. Advances in neural information processing systems, 2010, 23 : 1243 – 1251.

[25] HOCHREITER S, SCHMIDHUBER J. Long short-term memory[J]. Neural computation, 1997, 9(8): 1735–1780.

[26] STOLLENGA M, MASCI J, GOMEZ F, et al. Deep networks with internal selective attention through feedback connections[J]. arXiv preprint arXiv:1407.3068, 2014.

[27] 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. 2015: 945–953.

[28] 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. 2016: 5648–5656.

[29] 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. 2015: 1912–1920.

[30] MATURANA D, SCHERER S. Voxnet: A 3d convolutional neural network for real-time object recognition[C] // 2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). 2015: 922–928.

[31] 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): 1–11.

[32] RIEGLER G, OSMAN ULUSOY A, GEIGER A. Octnet: Learning deep 3d representations at high resolutions[C] // Proceedings of the IEEE conference on computer vision and pattern recognition. 2017: 3577–3586.

[33] CHOY C, GWAK J, SAVARESE S. 4d spatio-temporal convnets: Minkowski convolutional neural networks[C] // Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019: 3075–3084.

[34] GRAHAM B, ENGELCKE M, VAN DER MAATEN L. 3d semantic segmentation with submanifold sparse convolutional networks[C] // Proceedings of the IEEE conference on computer vision and pattern recognition. 2018: 9224–9232.

[35] JOSEPH-RIVLIN M, ZVIRIN A, KIMMEL R. Momen (e) t: Flavor the moments in learning to classify shapes[C] // Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops. 2019: 0–0.

[36] YANG J, ZHANG Q, NI B, et al. Modeling Point Clouds With Self-Attention and Gumbel Subset Sampling[C] // Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2019.

[37] ZHAO H, JIANG L, FU C-W, et al. Pointweb: Enhancing local neighborhood features for point cloud processing[C] // Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019 : 5565 – 5573.

[38] YAN X, ZHENG C, LI Z, et al. PointASNL: Robust Point Clouds Processing Using Nonlocal Neural Networks With Adaptive Sampling[C] // Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2020.

[39] 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 (CVPR). 2016.

[40] WAN C, YAO A, VAN GOOL L. Hand pose estimation from local surface normals[C] // European conference on computer vision. 2016 : 554 – 569.

[41] SUN X, WEI Y, LIANG S, et al. Cascaded hand pose regression[C] // Proceedings of the IEEE conference on computer vision and pattern recognition. 2015 : 824 – 832.

[42] TZIONAS D, BALLAN L, SRIKANTHA A, et al. Capturing hands in action using discriminative salient points and physics simulation[J]. International Journal of Computer Vision, 2016, 118(2) : 172 – 193.

[43] KHAMIS S, TAYLOR J, SHOTTON J, et al. Learning an efficient model of hand shape variation from depth images[C] // Proceedings of the IEEE conference on computer vision and pattern recognition. 2015 : 2540 – 2548.

[44] REMELLI E, TKACH A, TAGLIASACCHI A, et al. Low-dimensionality calibration through local anisotropic scaling for robust hand model personalization[C] // Proceedings of the IEEE International Conference on Computer Vision. 2017 : 2535 – 2543.

[45] TANG D, TAYLOR J, KOHLI P, et al. Opening the black box: Hierarchical sampling optimization for estimating human hand pose[C] // Proceedings of the IEEE international conference on computer vision. 2015 : 3325 – 3333.

[46] OBERWEGER M, WOHLHART P, LEPETIT V. Training a feedback loop for hand pose estimation[C] // Proceedings of the IEEE international conference on computer vision. 2015 : 3316 – 3324.

[47] ZHOU X, WAN Q, ZHANG W, et al. Model-based deep hand pose estimation[J]. arXiv preprint arXiv:1606.06854, 2016.

[48] WAN C, PROBST T, VAN GOOL L, et al. Crossing nets: Combining gans and vaes with a shared latent space for hand pose estimation[C] // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017 : 680 – 689.

[49] OBERWEGER M, WOHLHART P, LEPETIT V. Generalized feedback loop for joint hand-object pose estimation[J]. IEEE transactions on pattern analysis and machine intelligence, 2019, 42(8): 1898 – 1912.

[50] TOMPSON J, STEIN M, LECUN Y, et al. Real-time continuous pose recovery of human hands using convolutional networks[J]. ACM Transactions on Graphics (ToG), 2014, 33(5): 1 – 10.

[51] KIPF T N, WELLING M. Semi-supervised classification with graph convolutional networks[J]. arXiv preprint arXiv:1609.02907, 2016.

[52] WANG X, GIRSHICK R, GUPTA A, et al. Non-Local Neural Networks[C] // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2018.

[53] SHUMAN D I, NARANG S K, FROSSARD P, et al. The emerging field of signal processing on graphs: Extending high-dimensional data analysis to networks and other irregular domains[J]. IEEE signal processing magazine, 2013, 30(3): 83 – 98.

[54] SANDRYHAILA A, MOURA J M. Discrete signal processing on graphs[J]. IEEE transactions on signal processing, 2013, 61(7): 1644 – 1656.

[55] GILMER J, SCHOENHOLZ S S, RILEY P F, et al. Neural message passing for quantum chemistry[C] // International Conference on Machine Learning. 2017: 1263 – 1272.

[56] XU K, HU W, LESKOVEC J, et al. How powerful are graph neural networks?[J]. arXiv preprint arXiv:1810.00826, 2018.

[57] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[J]. arXiv preprint arXiv:1706.03762, 2017.

[58] KINGMA D P, BA J. Adam: A method for stochastic optimization[J]. arXiv preprint arXiv:1412.6980, 2014.