基于通道级自动核收缩的高效三维重建模型

罗洲航

摘要

以往基于深度学习的多视点立体图像 (Multi-View Stereo, MVS) 方法几乎都专注于提高重建质量。除了质量，效率也是 MVS 在实际场景中需要的特性。为此，本文引入一种高效的通道级自动核收缩算法 CAKES，用于 MVS 中快速准确的深度估计。在该网络中，我们首先提取视觉图像的深度特征，然后通过可微单应性弯曲在参考摄像机截形上构建三维代价体。接下来，我们应用 3D 卷积来正则化和回归初始深度图，然后与参考图像进行细化以生成最终输出。其中，在 3D 卷积过程中，我们通过将标准的 3D 卷积收缩为一组经济操作 (例如，1D, 2D 卷积) 来实现高效的 3D 学习。与以前的方法不同，CAKES 执行通道的内核收缩，它有以下好处:1) 使部署在每一层的操作都是异构的，这样它们就可以提取不同的和互补的信息，从而有利于学习过程;2) 采用高效、灵活的替换设计，可推广到时空和体积数据。在此基础上，我们提出了一种新的搜索空间，可以自动确定替换配置，从而简化三维网络。在大型室内 DTU 数据集上对所提出的改进 MVSNet 进行了验证，与原 MVSNet 相比，我们实现了高效的 MVS 重建系统，同时保持了可比的模型精度和完整性。

关键词：多目立体视觉；模型压缩；深度学习；神经网络结构搜索

1 引言

多目立体视觉 (MVS) 旨在从一组经过校准的图像中恢复场景的密集三维结构。它是计算机视觉的基础问题之一，由于其在 3D 重建、增强现实、自动驾驶、机器人等方面的广泛应用，已经被广泛研究了几十年。传统方法使用手工制作的相似度量和设计的正则化 (例如，标准化相互关和半全局匹配[1]) 来计算密集对应并恢复 3D 点。虽然这些方法在理想的兰伯情景下显示出了很好的结果，但它们也有一些共同的局限性。例如，场景的低纹理、镜面和反射区域使得密集匹配难以处理，从而导致不完整的重建。据报道，在最近的 MVS 基准中，尽管目前最先进的算法在精度上表现很好，重构完整性仍有很大的改进空间。

近年来卷积神经网络 (CNNs) 研究的成功也引发了人们对改进立体重建的兴趣。从概念上讲，基于学习的方法可以引入镜面先验和反射先验等全局语义信息，从而实现更健壮的匹配。有一些关于双视图立体匹配的尝试，通过替换手工制作的相似性度量或用学习的正则化。他们在立体基准测试中显示出良好的效果，并逐渐超越传统方法。事实上，立体匹配任务非常适合应用基于 CNN 的方法，因为图像对是提前校正的，因此问题就变成了水平像素级的视差估计，而不需要考虑相机参数。

然而，直接将学习到的双视图立体扩展到多视图场景并非简单。虽然可以简单地对所有选定的图像对进行立体匹配，然后将所有的成对重建合并到全局点云，但这种方法不能充分利用多视图信息，导致结果的准确性不高。与立体匹配不同，MVS 的输入图像可以是任意的相机几何形状，这对学习方法的使用提出了一个棘手的问题。只有少数作品承认这一问题，并尝试将 CNN 应用于 MVS 重建:SurfaceNet[2]提前构造了 Colored Voxel Cubes (CVC)，它将所有图像像素颜色和相机信息组合到一个单一的体积中作为网络的输入。相比之下，学习立体声机 (LSM)[3]直接利用可微投影/反投影来实现

端到端训练/推理。然而，这两种方法都利用了规则网格的体积表示。由于受到 3D 体积巨大内存消耗的限制，它们的网络难以扩展:LSM 只能处理低体积分辨率的合成对象，而 SurfaceNet 采用启发式分治策略，大规模重建耗时较长。目前，现代 MVS 基准的领先板仍被传统方法占据。

为此，我们引用了一种用于深度图推断的端到端深度学习架构[4]，每次计算一张深度图，而不是一次计算整个 3D 场景。与其他基于深度图的 MVS 方法类似，所提出的 MVSNet 网络以一张参考图像和多张源图像作为输入，并推断出参考图像的深度图。这里的关键见解是可微分单应性扭曲操作，它隐式编码网络中的摄像机几何形状，从 2D 图像特征构建 3D 成本卷，并实现端到端训练。为了适应输入中的任意数量的源图像，我们提出了一种基于方差的度量，将多个特征映射到体积中的一个成本特征。然后，该成本体积经过多尺度 3D 卷积并回归初始深度图。最后，利用参考图像对深度图进行细化，提高边界区域的精度。我们的方法与以往的学习方法有两个主要区别。首先，为深度图的目的推论，我们的 3D 成本体积是建立在相机截锥上，而不是常规的欧几里得空间。其次，我们的方法将 MVS 重构解耦为更小的视图深度图估计问题，从而使大规模重建成为可能。

但是，在从 2D 图像特征构建 3D 成本卷，并实现端到端训练中，3D 卷积层通常会导致昂贵的计算，并且由于过拟合问题和缺乏预先训练的权重而存在收敛问题。为了解决这些问题，我们引入了基于通道的自动内核收缩 (CAKES)[5]，作为现有 3D 操作的通用高效替代方案。具体来说，所提出的方法通过采用多样化和经济操作 (例如，1D, 2D 卷积) 的组合简化了传统的 3D 操作，其中这些不同的操作可以提取互补的信息，以在同一层中使用。我们的方法不是为任何特定类型的输入 (例如视频) 量身定制的，而是可以推广到不同类型的数据和骨干架构，以实现细粒度和高效的替换。但是，人工选择替换算子集及其定位需要反复试验。为了进一步提高性能和模型效率，我们引入了一个由计算效率高的候选操作符组成的新的搜索空间，以方便搜索给定骨干架构的最优替换配置。通过我们的搜索空间设计，提出的 CAKES 可以在几个 GPU 天内获得一个好的架构。我们在大规模 DTU 数据集上训练和评估所改进 MVSNet，与基线模型相比，改进 MVSNet 不仅表现出了优异的性能，而且有效地减小了模型尺寸。

2 相关工作

2.1 MVS 重建

根据输出表示形式，MVS 方法可分为 1) 直接点云重建，2) 体积重建和 3) 深度图重建。基于点云的方法直接在三维点上操作，通常依赖于传播策略来逐步强化重建。由于点云的传播是按顺序进行的，这些方法难以完全并行化，处理时间长。基于体积的方法将三维空间划分为规则网格，然后估计每个体素是否粘附在表面上。这种表示的缺点是空间离散化错误和高内存消耗。相比之下，深度图是最灵活的表示法。它将复杂的 MVS 问题解耦为相对较小的逐视图深度图估计问题，每次只关注一个参考图像和几个源图像。此外，深度图可以很容易地融合到点云或体积重建。根据最近的 MVS 基准，目前最好的 MVS 算法都是基于深度图的方法。

2.2 学习多视图立体

相比使用传统的手工图像特征和匹配指标，最近的立体研究应用深度学习技术来更好的成对补丁匹配。Han et al.[6]首先提出了一个深度网络来匹配两个图像补丁。Zbontar et al.[7]和 Luo et al.[8]使用学

习到的特征进行立体匹配，半全局匹配 (semi-global matching, SGM) 进行后处理。除了成对匹配的代价外，学习技术还应用于代价正则化。SGMNet 学习调整 SGM 中使用的参数，而 CNN-CRF[9]集成了网络中的条件随机场优化，进行端到端的立体声学习。目前最先进的方法是 GCNet[10]，该方法应用 3D CNN 对成本体积进行正则化，并通过软 argmin 操作对差异进行回归。据报道，KITTI 的基准，学习型立体声，特别是这些端到端学习算法的性能明显优于传统的立体方法。

2.3 学习型 MVS

对于学习型 MVS 方法的尝试较少。Hartmann et al.[11]提出了学习的多斑块相似度来取代传统的 MVS 重建成本度量。用于 MVS 问题的第一个基于学习的管道是 SurfaceNet[2]，它通过复杂的体素视图选择预先计算成本体积，并使用 3D CNN 正则化和推断表面体素。与我们最相关的方法是 LSM[3]，其中摄像机参数在网络中编码为投影操作，以形成成本体积，3D CNN 用于对体素是否属于表面进行分类。然而，由于体积表示的共同缺点，SurfaceNet[2]和 LSM[3]的网络仅限于小规模重建。它们要么采用分治策略，要么只适用于低分辨率输入的合成数据。相比之下，我们的网络专注于每次为一个参考图像生成深度图，这允许我们直接自适应重建一个大场景。

2.4 高效的 3D 卷积神经网络

尽管 3D CNN 取得了巨大的进步，现有的 3D 网络通常需要大量的计算预算。此外，由于缺乏预先训练的权重，3D CNN 的训练也不稳定。这些事实促使研究人员寻找 3D 卷积的有效替代品。例如，Luo 和 Yuille[12]提出, Tran et al.[13]将群卷积和深度卷积应用于 3D 网络以获得资源高效模型。另一种方法建议将每个 3D 卷积层替换为 2D 和 1D 卷积层的结构化组合，以获得更好的性能，同时更高效。例如，Qiu, Yao 和 Mei[3]和 Tran et al.[14]提出使用 2D 空间卷积层，然后使用 1D 时间卷积层来取代标准的 3D 卷积层。此外，Xie et al.[15]证明了 3D 卷积并非处处都需要，其中一些可以被 2D 卷积所取代。类似的尝试也发生在医学成像领域。例如，Gonda et al.[16]尝试通过连续的 2D 卷积层，然后是 1D 卷积层来替换连续的 3D 卷积层。

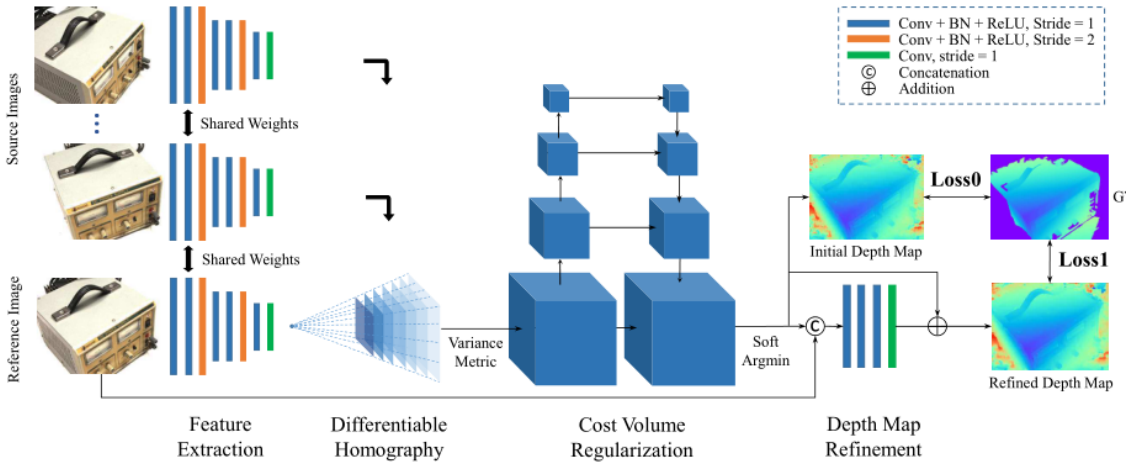

图 1: MVSNet 的网络设计

3 本文方法

在本文中，3D 重建的主体框架采用了的 Yao et al. 的 MVSNet 模型，模型如 1所示，在此基础上，为了减小模型的参数量与计算量以实现模型压缩，且相对集中在原始模型中的三维卷积层中。因此，在三维卷积层中采用了 Yu et al. 的通道级自动核收缩算法 CAKES 来代替传统三维卷积。除此之外，与

原始模型无异，模型主要部分介绍如下。

3.1 特征提取

MVSNet 的第一步是提取 N 张输入图像的深度特征进行密集匹配。采用 8 层二维 CNN，将第 3 层和第 6 层的步长设为 2，将特征塔划分为 3 个尺度。在每个尺度中，应用两个卷积层来提取更高层次的图像表示。除最后一层外，每个卷积层后面是批处理归一化 (BN) 层和整流线性单元 (ReLU)。此外，与普通匹配任务类似，参数在所有特征塔之间共享，以实现高效学习。二维网络的输出是 N 个 32 通道的特征图，与输入图像相比，每个维度缩小了 4 倍。值得注意的是，虽然特征提取后图像帧缩小，但每个剩余像素的原始相邻信息已经被编码到 32 通道像素描述符中，从而防止了密集匹配丢失有用的上下文信息。与简单地对原始图像进行密集匹配相比，提取的特征映射显著提高了重建质量。

3.2 成本体积正规化

从图像特征计算的原始成本体积可能受到噪声污染 (例如，由于存在非兰伯表面或物体遮挡)，应该与平滑度约束结合起来以推断深度图。我们的正则化步骤旨在细化上述成本量 C，以生成用于深度推理的概率量 P。受最近基于学习的立体和 MVS 方法的启发，我们应用多尺度三维 CNN 进行成本体积正则化。这里的四尺度网络类似于 3D 版本的 UNet，它使用编码器-解码器结构，以相对较低的内存和计算成本聚合来自大接收域的邻近信息。为了进一步降低计算量，我们将第一个三维卷积层后的 32 通道成本体积减少到 8 通道，并将每个尺度内的卷积从 3 层改为 2 层。最后一个卷积层输出一个 1 通道的卷。最后应用沿深度方向的 softmax 运算进行概率归一化。

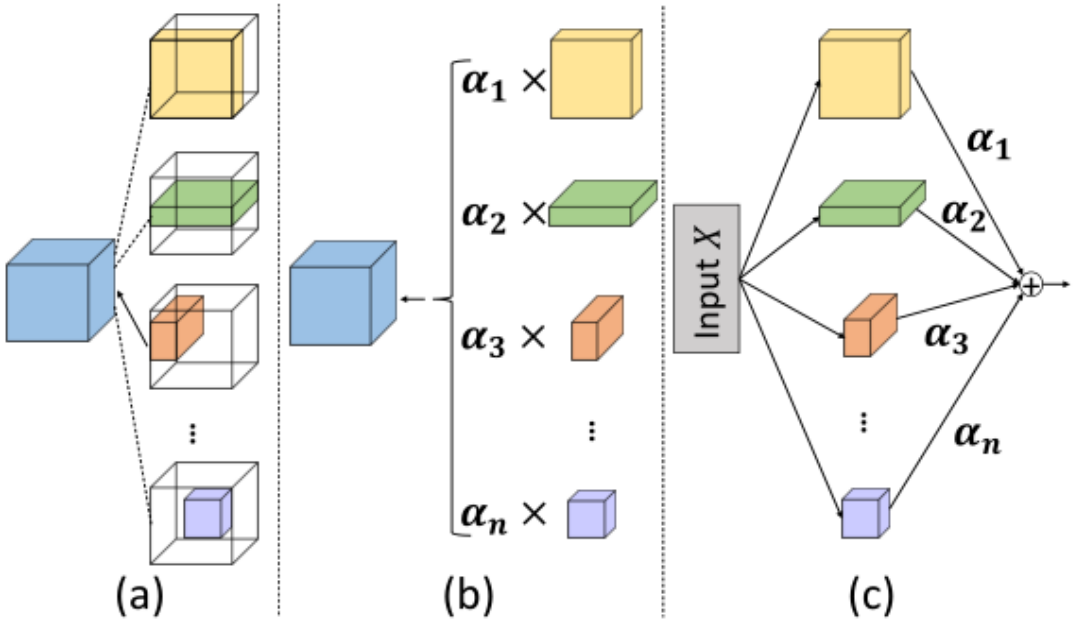

图 2: (a) 同一 3D 内核的 V 个子内核;(b) 三维核表示为子核的加权和;(c) 路径选择

3.3 CAKES 算法

为了解决模型计算量的问题，采用了基于通道的自动内核收缩 (CAKES)，作为 3D 卷积的通用替代方案。其核心思想是将标准 3D 内核压缩成一组更便宜的 1D、2D 和 3D 组件。为了保证设计的灵活性，避免繁琐的手动配置，进一步细化了收缩通道，这样异构核可以像 3D 核一样提取互补信息。此外，还引入了全新的搜索空间，以便自动优化替换配置。

如 2(a) 所示，即使只考虑不同的内核大小，3D 内核也有 $k_d \times k_h \times k_w$ 的子内核选项，通过人工

设计找到最优的子内核是不现实的。因此，我们将这个问题表述为路径级选择，即将子核编码为一个多路径超级网络，并在其中选择最优路径 (2(c))。这个问题可以用可微的方式来解决。我们可以将 3D 内核表示的一般替换如下 (2(b)):

$$W_C^{k_d \times k_h \times k_w} \leftarrow \{\alpha_i W_C^{k_d^i \times k_h^i \times k_w^i}\}. \tag{1}$$

其中 α_i 为第 i 个子核的权值。用这个公式，寻找 $W_C^{k_d \times k_h \times k_w}$ 的最优子核的问题可以近似为找到 $\{\alpha_i\}$ 的最优设置，然后保持 αi 最大的子核。

如 3所示，以通道方式收缩内核可以生成异构操作，在同一层内提取不同且互补的信息，从而产生更细粒度且更有效的替换 3(d))，而之前使用分层替换的方法 (3(a),(b),(c))。与之前的分层替换不同，我们的核心思想是在每个通道分别替换 3D 核，因此目标是找到最优的子核作为第 c 个输出通道 3D 核的替代品。得益于通道收缩，该方法提供了一个更通用和更灵活的设计来取代三维卷积比以前的方法，通过将这些操作集成到候选子内核集合中，它也可以很容易地简化为任意替代 (例如，2D, P3D)。在图 3 中可以找到一个示例。

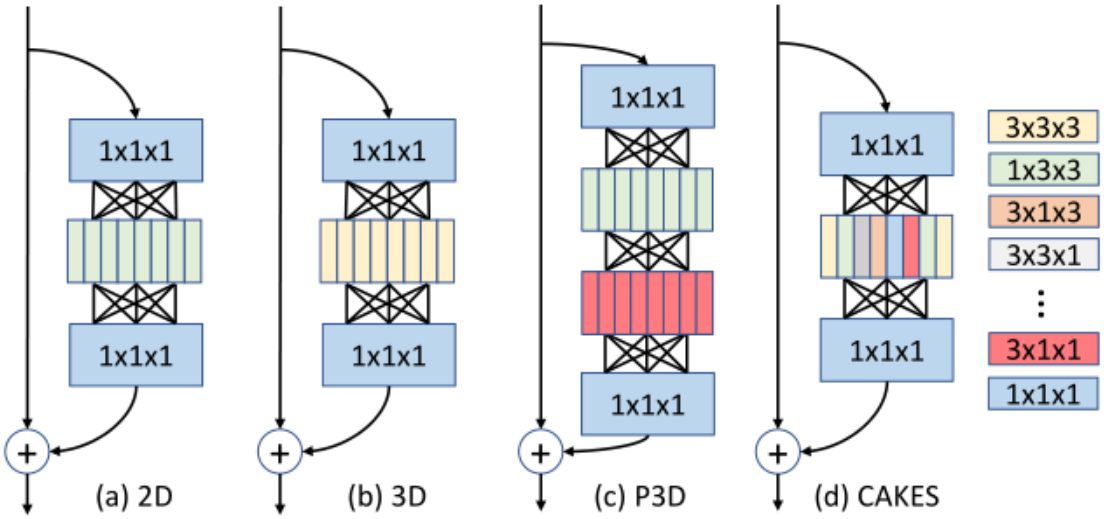

图 3: 残差块中不同类型卷积比较的说明性示例.(a) 二维卷积;(b) 三维卷积;(c) P3D 卷积;(d)CAKES

为了实现这一目标，我们引入了两种不同的搜索策略，分别以性能优先级搜索和成本优先级搜索为目标。性能优先搜索，目的是通过在给定的主干架构下找到最优的子内核来最大化性能，在前面的介绍中，我们通过引入 α 来评定该替换操作对于整体模型的重要程度，换言之，α 越大，该操作对于整体模型预测结果越重要。因此可以直接选择保留 α 最大的路径操作，通过剪枝剪去相对不重要的操作以此实现性能优先搜索。成本优先搜索，目的是获得更紧凑的模型，我们以修剪的方式搜索模型，对昂贵的操作进行惩罚，引入了一个“成本意识”的惩罚项，将许多路径权值推到接近零的值。因此总的损失函数更改为：

$$L = \varepsilon + \lambda \sum_i \beta_i |\alpha_i|. \tag{2}$$

其中 β_i 是一个“成本意识”项，以平衡惩罚项，这与子内核的参数或 FLOPs 成本成正比，λ 为惩罚项的系数，ε 为常规训练损失 (如交叉熵损失结合权值衰减等正则化项)。

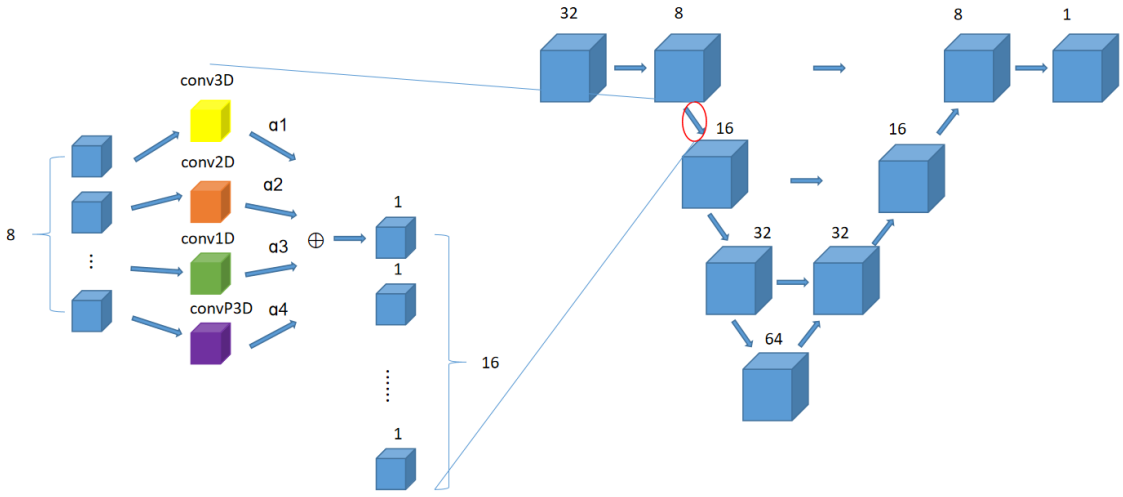

图 4: 应用 CAKES 算法改进后的成本体积正则化模型

4 复现细节

4.1 与已有开源代码对比

本工作的模型框架基本引用了 MVSNet 的一个非官方 Pytorch 版本的代码 (官方版本为 Tensorflow 框架)，并在此基础上根据 CAKES 算法，将 MVSNet 代码中的成本体积正规化模型，更改为 CAKES 算法改进后的三维卷积正则化模型。该部分模型如 4所示。其中，应用 CAKES 算法改进后的成本体积正则化模型伪代码如下：

Procedure 1 Cost Volume Regularization by CAKES.

Input: cost volume X
Output: feature map Y
for i ***in*** *Encoders* **do**
  for j ***in*** CFG_i **do**
    for z ***in*** *SuperNets* **do**
      $out_j \mathrel{+}= \alpha_i \times \text{SuperBlocks}[z](X)$
    end
    $X = \text{Concat}(\text{out}_j)$
  end
end
for i ***in*** *Decoders* **do**
  X =ConvTranspose(Concat(encoder,X))
end

4.2 实验环境搭建

批样本数数，输入视图个数，图像宽度，高度和深度样本个数分别设置为 B=4，N = 5, W = 1600, H = 1184, D =192 。在 4 块 Titan XP 上对 DTU 数据集，首先搜索 1 epoch，然后将搜索后的模型的多余部分剪枝构建新的模型。再将新模型训练 8 epochs，得到最终模型。在 DTU 数据集的 22 个评估扫描上评估方法。

4.3 创新点

相比于原始 MVSNet，本工作通过引用 CAKES 算法，使用神经网络结构搜索的方式在 1D 卷积、2D 卷积、P3D 操作中寻找可以替代 3D 卷积的操作（或者保留原来 3D 卷积），并按照原文提供的两种搜索策略，以此实现在保证模型精度稳定的同时降低模型的计算量与参数量来压缩模型。

5 实验结果分析

量化结果见 1。其中，MVSNet 表示原始模型，CAKES 表示引用了模型压缩方法后的紧凑 MVSNet 模型。上标 P 代表使用了性能优先搜索策略，C 代表使用了成本优先搜索策略；下标代表搜索空间中加入了几种卷积方式。从结果可以看出，本次创新在一定程度上对 MVSNet 进行了压缩，在 $CAKES^{P}_{1,2D}$ 中由于只使用了 1D 和 2D 卷积，使得模型压缩幅度更大，但是相比 $CAKES^{C}_{1,2,3,PD}$，在模型精度上要略负一筹。

模型	参数量 (K)	FLOPs(G)	Loss	MAE
MVSNet	338.129	69.005	8.804	9.227
$CAKES^{P}_{1,2D}$	177.041	42.882	10.869	11.315
$CAKES^{C}_{1,2,3,PD}$	279.731	63.766	10.51	10.94

表 1: 在 DTU 数据集上两种搜索策略与搜索空间的结果

定性结果见 5和 6，本工作中引用 CAKES 算法压缩后的紧凑 MVSNet 在重建方面，虽然在边缘细节部分相当于原始 MVSNet 有一定的误差，但是总体上差距并不大，依然能够很好地完成重建任务。

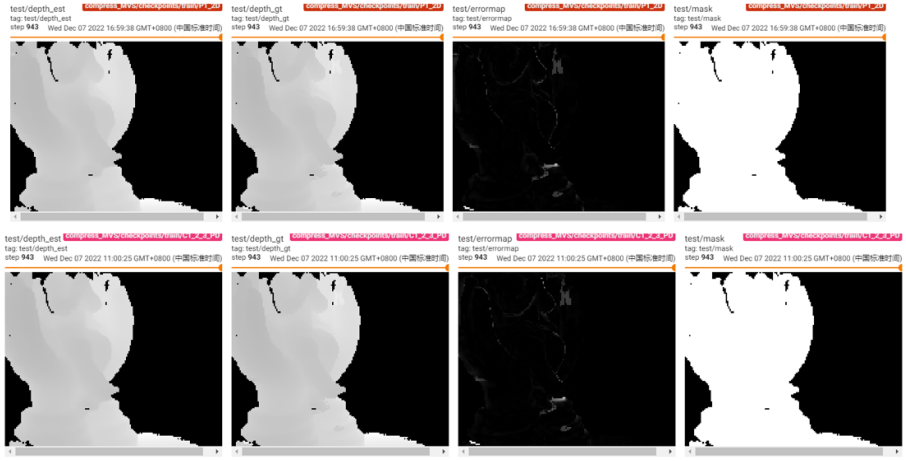

图 5: 两种搜索策略下深度图的结果。上：$CAKES^{P}_{1,2D}$；下：$CAKES^{C}_{1,2,3,PD}$

图 6: DTU 数据集点云重建的定性结果。左：MVSNet；中：$CAKES^{P}_{1,2D}$；下：$CAKES^{C}_{1,2,3,PD}$

6 总结与展望

三维网络作为各种三维视觉应用的重要解决方案，仍然存在参数化过度和计算量大的问题。如何设计出有效的 3D 操作替代方案仍然是一个悬而未决的问题。尤其，作为近年来研究较为火热的三维重建技术来说，设计一个性能优秀且模型紧凑的模型并不容易。因此，为了能使大规模的三维重建模型能够得到更好地应用，诸如移植到便携设备或者小型设备上，使用模型压缩方式对三维重建模型进行压缩是一个不错的策略。因此，本工作将重点放在压缩三维卷积操作上，引入了 CAKES 算法，对大型三维重建模型 MVSNet 进行压缩。实验结果证明，我们的压缩方式在保证了原始模型的精度的同时，很好的完成了降低模型计算量与参数量的任务。今后的研究中，将会继续研究三维重建模型的模型压缩策略与轻量化模型设计。

参考文献

[1] HIRSCHMULLER H. Stereo processing by semiglobal matching and mutual information[J]. IEEE Transactions on pattern analysis and machine intelligence, 2007, 30(2): 328-341.

[2] JI M, GALL J, ZHENG H, et al. SurfaceNet: An End-to-End 3D Neural Network for Multiview Stereopsis[J]. IEEE Computer Society, 2017.

[3] QIU Z, YAO T, MEI T. Learning Spatio-Temporal Representation with Pseudo-3D Residual Networks [J]. IEEE, 2017.

[4] 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). 2018: 767-783.

[5] YU Q, LI Y, MEI J, et al. CAKES: Channel-wise Automatic KErnel Shrinking for Efficient 3D Networks [C]//National Conference on Artificial Intelligence. 2020.

[6] HAN X, LEUNG T, JIA Y, et al. MatchNet: Unifying feature and metric learning for patch-based matching[C]//Computer Vision & Pattern Recognition. 2015.

[7] ŽBONTAR J, LECUN Y. Stereo Matching by Training a Convolutional Neural Network to Compare Image Patches[Z]. 2016.

[8] LUO W, SCHWING A G, URTASUN R. Efficient Deep Learning for Stereo Matching[C]//IEEE Conference on Computer Vision & Pattern Recognition. 2016: 5695-5703.

[9] KNOBELREITER P, REINBACHER C, SHEKHOVTSOV A, et al. End-to-End Training of Hybrid CNN-CRF Models for Stereo[C]//2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2017.

[10] KENDALL A, MARTIROSYAN H, DASGUPTA S, et al. End-to-End Learning of Geometry and Context for Deep Stereo Regression[Z]. 2017.

[11] HARTMANN W, GALLIANI S, HAVLENA M, et al. Learned Multi-patch Similarity[C]//2017 IEEE International Conference on Computer Vision (ICCV). 2017.

[12] LUO C, YUILLE A. Grouped Spatial-Temporal Aggregation for Efficient Action Recognition[J]. IEEE, 2019.

[13] TRAN D, BOURDEV L, FERGUS R, et al. Learning Spatiotemporal Features with 3D Convolutional Networks[Z]. 2014.

[14] DU T, WANG H, TORRESANI L, et al. A Closer Look at Spatiotemporal Convolutions for Action Recognition[J].,

[15] XIE L, YUILLE A. Genetic CNN[Z]. 2017.

[16] GONDA F, WEI D, PARAG T, et al. Parallel Separable 3D Convolution for Video and Volumetric Data Understanding[Z]. 2018.