计 算 机 研 究 与 发 展
Journal of Computer Research and Development
https://crad.ict.ac.cn
62(8): 1862−1874, 2025

基于时间卷积神经架构搜索的复杂动作识别

任鹏真[1,2] 梁小丹[1,3] 常晓军[4] 赵子莹[1] 肖 云[5]
[1](鹏城实验室 广东深圳 518000)
[2](暗物智能科技(广州)有限公司 广州 511466)
[3](中山大学智能工程学院 广东深圳 518107)
[4](中国科学技术大学信息科学技术学院 合肥 230026)
[5](西北大学信息科学与技术学院 西安 710127)
(pzhren@foxmail.com)

Neural Architecture Search on Temporal Convolutions for Complex Action Recognition

Ren Pengzhen[1,2], Liang Xiaodan[1,3], Chang Xiaojun[4], Zhao Ziying[1], and Xiao Yun[5]
[1](*Pengcheng Laboratory*, *Shenzhen*, *Guangdong* 518000)
[2](*DarkMatter AI Research*, *Guangzhou* 511466)
[3](*School of Intelligent Engineering*, *Sun Yat-sen University*, *Shenzhen*, *Guangdong* 518107)
[4](*School of Information Science and Technology*, *University of Science and Technology of China*, *Hefei* 230026)
[5](*School of Information Science and Technology*, *Northwest University*, *Xi'an* 710127)

Abstract In the field of complex action recognition in videos, the structural design of the model plays a crucial role in its final performance. However, manually designed network structures often rely heavily on the knowledge and experience of researchers. Therefore, neural architecture search (NAS) has received widespread attention from researchers in the field of image processing because of its automated network structure design. Currently, neural architecture search has achieved tremendous development in the image field. Some NAS methods even reduce the number of graphics processing unit (GPU) days required for automated model design to single digits, and the model structures they search show strong competitive potential. This encourages us to extend automated model structure design to the video domain. But it faces two serious challenges: 1) How to capture the long-range contextual temporal association in video as much as possible; 2) How to reduce the computational surge caused by 3D convolution as much as possible. To address the above challenges, we propose a novel neural architecture search on temporal convolutions for complex action recognition (NAS-TC). NAS-TC is a two-stage framework. In the first stage, we use the classic convolutional neural network (CNN) network as the backbone network to complete the computationally

收稿日期: 2024-01-29; **修回日期**: 2025-02-25
基金项目: 国家科技重大专项基金项目(2020AAA0109704); 中国博士后科学基金项目(2023M734009); 国家自然科学基金项目(62372371); 陕西省国际科技合作计划重点项目(2022KWZ-14); 广东省杰出青年科学基金项目(2021B1515020061); 深圳市科技计划项目(GJHZ20220913142600001); 南沙重点研发计划项目(2022ZD014); 鹏城实验室重大攻关项目(PCL2024AS101); CAAI-华为 MindSpore 开放基金
This work was supported by the Major National Science and Technology Program (2020AAA0109704), the China Postdoctoral Science Foundation (2023M734009), the National Natural Science Foundation of China (62372371), the Key Projects of Shaanxi Province International Science and Technology Cooperation Plan (2022KWZ-14), the Guangdong Provicial Outstanding Youth Foundation (2021B1515020061), the Shenzhen Science and Technology Program (GJHZ20220913142600001), the Nansha Key Research and Development Program (2022ZD014), the Pengcheng Laboratory Major Research Project (PCL2024AS101), and the CAAI-Huawei MindSpore Open Fund.
通信作者: 肖云(yxiao@nwu.edu.cn)

intensive feature extraction task. In the second stage, we propose a neural architecture search layer temporal convolutional layer (NAS-TC) to accomplish relatively lightweight long-range temporal model design and information extraction. This ensures that our method will have a more reasonable parameter allocation and can handle minute-level videos. Finally, the proposed method achieves an average performance gain of 2.3% mAP on three complex action recognition benchmark data sets compared with similar methods, and the number of parameters is reduced by 28.5%.

Key words complex action recognition; neural architecture search; convolution decomposition; video understanding; deep learning

摘 要 在视频的复杂动作识别领域中，模型的结构设计对其最终的性能起着至关重要的作用. 然而，人工设计的网络结构往往严重依赖于研究人员的知识和经验. 因此，神经架构搜索（neural architecture search, NAS）因其自动化的网络结构设计在图像处理领域受到研究人员的广泛关注. 当前，神经架构搜索已经在图像领域获得了巨大的发展，一些 NAS 方法甚至将模型自动化设计所需的 GPU 天数减少到了个位数，并且其搜索的模型结构表现出了强大的竞争潜力，这鼓励将自动化模型结构设计拓展到视频领域. 但它面临 2 个严峻的挑战：1）如何尽可能捕获视频中的长程上下文时间关联；2）如何尽可能降低 3D 卷积所带来的计算激增的问题. 为了应对上述挑战，提出了一个基于时间卷积的神经架构搜索复杂动作识别（neural architecture search on temporal convolutions for complex action recognition, NAS-TC）模型. 具体地，NAS-TC 具有 2 个阶段：在第 1 阶段，采用经典的 CNN 网络作为骨干网络，来完成计算密集型的特征提取任务. 在第 2 阶段，提出了一个神经架构搜索时间卷积层来完成相对轻量级的长程时间模型设计和信息提取. 这确保了提出的方法具有更合理的参数分配并且可以处理分钟级的视频. 最后，提出的方法在 3 个复杂动作识别基准数据集上与同类型方法相比平均获得了 2.3 个百分点的 mAP 的性能增益，并且参数量下降了 28.5%.

关键词 复杂动作识别；神经架构搜索；卷积分解；视频理解；深度学习

中图法分类号 TP391

DOI: 10.7544/issn1000-1239.202440048 **CSTR:** 32373.14.issn1000-1239.202440048

动作识别在计算机视觉领域有着非常重要的研究价值. 考虑如何在未修剪的长视频中识别复杂动作比经过剪切视频中的动作识别更具现实意义. 因此，对此类任务的研究具有重要意义. 然而，深度学习方法在视频领域的应用存在 2 个明显的挑战：1）人工设计的模型过于复杂，并且严重受限于研究人员的先验知识；2）直接使用 3D 卷积来提取长程的时空特征信息通常会导致计算和参数量激增的问题.

针对模型设计的问题，众所周知，网络的模型结构对模型最终的性能起着至关重要的作用[1-4]. 设计一个好的、复杂度适中的网络架构需要研究人员付出巨大的努力，而且要求研究人员具有深刻的洞察力. 考虑如何自动化设计一个符合要求的网络结构对模型结构设计的发展具有重要意义. 近年来，神经架构搜索在图像视觉领域的成功应用，为研究人员实现自动化设计网络架构提供了一个全新的可能[5-8]. 不幸的是，尽管早期基于强化学习[5]的神经架构搜索（neural architecture search, NAS）算法已经取得了非常优异的性能，但计算成本（大约 2 000 个 GPU 天）令大多数小型研究团体望而却步. 此外，采用进化算法的框架搜索的计算成本也达到了惊人的 3 150 GPU 天[6]. 因此，大量优秀的工作都集中在如何减少计算量并加速模型的架构搜索[7-11]. 另外，还有一些基于蒙特卡洛树搜索[12]、序列模型优化[13]、贝叶斯优化[14]的架构搜索方法. 然而，这些方法将架构搜索视为离散域上的黑盒优化问题，这使得优化模型变得极为困难. 幸运的是，DARTS（differentiable architecture search）[15]提出了一种新的可微架构搜索方法. 它首先将离散的搜索空间连续化，然后使用梯度下降来对模型进行架构搜索加速. 通过这种方式，DARTS 可以极大地减少模型架构搜索所需的计算成本，让小型研究团体也能参与到模型架构搜索的相关研究中. 因此，这项研究迅速吸引了大量研究人员的兴趣[7,16-17]，并且在图像领域相比人工设计的模型展现出了强大的竞争潜力.

另外，在视频动作识别领域，先前的工作[5,18-20]主要研究了短程的动作识别任务，如 Kinetics[21], Something-Something[22], UCF101[23], HMDB[24]. 平均地，这些数据

集中的操作往往仅持续 10 s 甚至更短. 正如 Action-Snippets[25] 所发现的, 上述基准数据集往往仅需要识别几帧就能够做出相应的动作识别. 与上述任务不同, 遵循 Timeception[26] 不同, 本文更加关注的是长程复杂动作的识别, 如 Charades[10], BreakfastActions[27], MultiTHUMOS[28]. 如图 1 所示, 在 Charades[10] 数据集中的一个数据实例, 复杂的动作通常持续较长的时间并且包含多个单元动作. 这些单元动作不仅在时间顺序上差异很大, 而且持续时间也不均匀, 并且在时间线上是相互重叠的. 因此, 为了解决视频动作中广泛存在的复杂时空关系并降低计算量, 研究人员提出了许多新颖且复杂的神经架构, 并取得了非常好的性能[25,29]. 例如, 在 Transformer[30] 框架中, TimeSformes[31] 提出了一种无卷积的视频分类方法, 该方法完全基于空间和时间的自我注意力模块. 而 VideoMAE[32] 基于 MAE[33] 的框架提出了一个有高视频掩码率的模型. 这些方法都取得了很好的性能. 不同于上述方法, 本文主要关注基于 CNN 的模型框架. 虽然 3D 卷积神经网络[34] 已经在视频领域取得了不错的性能, 但先前的方法高昂的计算代价使其应用受到了严重的限制. 这激发人们思考如何降低 3D 卷积的计算量. 为此, 研究人员开始考虑分解 3D 卷积. 例如, CoST[35] 提出将 3D 卷积分解为 3 个等效的 2D 卷积, 但这可能不利于时间与空间信息的分离; 此外, S3D[36] 提出使用 2+1D 卷积来替换原始的 3D 卷积, 并对时间与空间 2 种信息进行了分离. 进一步地, Timeception[26] 精心设计了一个具有不同核大小的深度可分离的时间卷积, 以便更加灵活地捕捉长视频中的长程时间信息. 此外, LFB[37] 和 MeMViT[38] 还尝试利用特征库或内存通过端到端训练的方式增强模型在远程上下文建模的能力. 这些方式都显著提升了模型的性能, 但文献 [37−38] 方法仍然面临着模型结构设计复杂对初级研究者不友好的问题. 因此, 考虑如何降低模型的设计门槛变得十分重要.

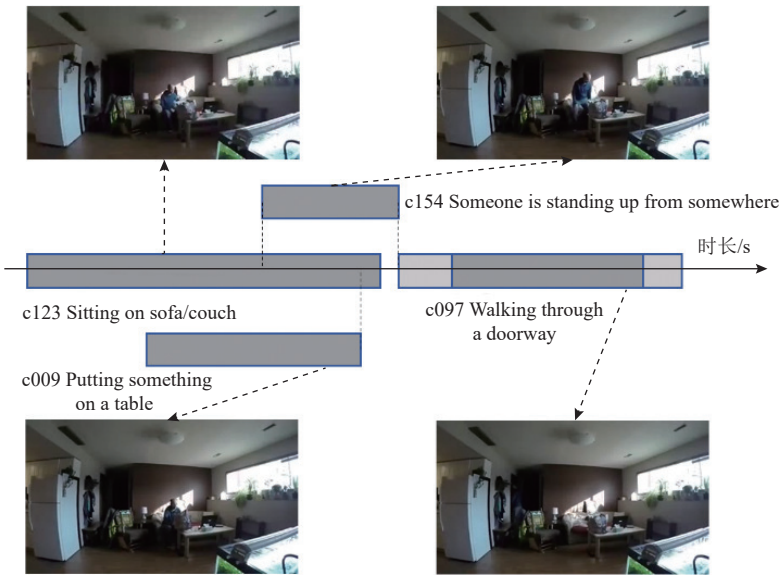


Fig. 1 A long video example of a complex action with four unit actions

图 1 一个包含 4 个单元动作的复杂动作的长视频示例

受上述观察启发, 本文致力于将神经架构搜索自动化模型设计拓展到长视频的复杂动作识别领域. 一种自然的想法是使用 3D 卷积替换 2D 卷积作为基础的操作单元, 直接使用神经架构搜索来实现模型结构的自动化设计. 但它存在 2 个严峻的挑战: 1)如何尽可能捕获视频中的长程上下文时间关联; 2)如何尽可能降低 3D 卷积所带来的计算量激增的问题. 为了解决上述挑战, 本文聚焦于长视频中复杂动作的识别, 提出了一个新颖的基于时间卷积的神经架构搜索的复杂动作识别模型. 具体地, 为了降低计算量并增强模型在时间维度上的信息捕捉能力, 我们为长视频复杂动作识别提出一个新颖的只在时间维度上的搜索空间, 并基于此使用可微架构搜索得到了一个 NAS-Temp (NAS-temporal) 模块, 该模块可以在兼顾模型的效率和时间信息处理能力的前提下实现模型自动化搜索. 我们总结本文的主要贡献有 3 点:

1)为长视频复杂动作识别设计了一个新颖的只有时间维度上的搜索空间, 可以有效地降低模型搜索的计算复杂度, 同时获得更好的长程时间信息学习能力.

2)成功地将神经架构搜索拓展到了长视频复杂动作识别领域，可以有效地降低视频领域模型设计的难度，并为模型设计提供了新的借鉴.

3)在3个常用的基准数据集上进行了广泛的测试，和同类型方法相比，基于时间卷积的神经架构搜索复杂动作识别(neural architecture search on temporal convolutions for complex action recognition, NAS-TC)在复杂动作识别任务上平均获得了2.3个百分点的mAP的性能增益，并且参数量下降了29%.

1 相关工作

1.1 神经架构搜索

神经架构搜索被认为是神经网络结构设计的一场革命，它有望突破人类经验和思维模式的局限性. 因此，神经架构搜索引起了自动化神经架构设计社区研究人员的极大兴趣[7]. 早期的基于强化学习[5,38-39]和进化学习[26,40]的神经架构搜索工作往往由于其庞大的计算量而令人望而却步. 2019年，DARTS[15]的出现极大地改善了这种状况，它使用流行的基于单元网络(cell)[5]的搜索空间，通过搜索2种不同类型的单元(正常单元和还原单元)模块交替堆叠来形成最终的网络架构. 特别地，DARTS通过采用松弛技术实现了搜索空间的连续化，进而得以采用常规的梯度下降方法来解决节点间候选操作的优化选择问题. 这极大地改善了神经架构搜索的搜索效率，使得小型研究团体也能感受来自神经架构搜索的魅力. 随后，研究人员进一步提出了大量与DARTS相关的改进工作[14,16]. 当前，DARTS相关的方法已经大大降低了单元模块(cell)的搜索成本，并且在图像视觉领域取得了显著的成功. 然而，想要直接将DARTS扩展到视频领域仍然面临着计算量激增的问题.

1.2 复杂动作识别

视频任务与图像任务的最大区别在于其需要处理时间维度上的信息以进行额外的建模. 常用的建模方法包括统计池化[41-45]、向量聚合[46-47]、基于神经网络的方法[48-52]. 统计池化和向量聚合忽略了视频任务中的时间线索，而基于神经网络的方法对于时间维度上信息的挖掘也极为有限. 为此，研究人员也提出了一些基于神经网络的短程动作识别模型. 例如，双流卷积网络[53]使用光流作为RGB图像的辅助信息. DynamicImage[54]使用动态图进行动作识别. 进一步地，C3D[35,55]提出使用3D卷积来捕捉时空信息. 此外，为了加速训练，I3D[56]通过扩展2D卷积到3D卷积来快速从视频中学习时间信息. 文献[53−56]方法通常计算成本昂贵，并且无法有效地学习长视频片段中的时间信息. 为此，TRN[57]研究了不同视频帧在多个时间尺度上的时间依赖性；TSN[58-59]将长视频划分为视频片段，并集成不同视频片段的时间模式以进行视频级预测；NL网络[60]引入非局部注意力机制来获取长程时间模式. 类似的相关工作还有很多[61-62]. 然而，文献[57-62]方法在时间信息建模的长短上仍然存在着很大的局限性. 为了解决上述问题，Timeception[26]通过精心设计时间卷积模块，成功建模了1 024个时间步长的视频，获得了很好的性能. 然而，模型的手动设计仍然是一项非常依赖专业和经验、繁琐且具有挑战性的工作. 因此，考虑如何实现长视频复杂动作模型设计的自动化是一个非常具有挑战性和值得关注的问题.

2 动 机

2.1 NAS在视频领域的困境

尽管DARTS[15]能够将模型的搜索成本从数千个GPU天降低到数个GPU天，但要求DARTS使用3D卷积直接从视频数据中学习时空信息仍然是非常困难. 除了要应对计算量激增的问题外，长视频中的时间信息也难以得到充分的学习.

2.2 时间信息上的独立卷积

对于视频任务，3D CNN通常通过3个正交子空间来学习时空特征，即时间(T)、空间(S)和语义通道子空间(C)[26]. 在之前的模型中，时空信息通常是被同时建模的(例如，3D卷积[35,55]，2+1D卷积[36]). 然而，这些方法的时间信息处理能力不足，导致它们无法对分钟级的长视频复杂动作进行有效的识别. 基于此，Timeception[26]的成功在一定程度上说明了独立时间信息卷积的有效性. 这种方式也可以在一定程度上缓解视频处理任务中计算量激增的问题，但仍然面临着模型设计过于复杂且严重受限于研究人员经验的问题.

2.3 NAS-TC的2阶段策略

基于上述观察，本文提出了一个简单直接且有效的2阶段处理策略. 具体地，我们使用3D CNN作为模型的骨干网络来提取低级别的时空特征. 然后使用神经架构搜索方法来实现长视频高级别时间信息提取模型的自动化设计. 这样一来，原本繁重的视频任务被成功分解为2部分：1)计算量大的低级别时空特征提取任务由骨干网络来完成；2)而长视频

的高级别时间特征信息提取任务的轻量级模型自动化设计任务则被分配给了神经架构搜索. 这不仅可以实现了长视频复杂动作识别模型的自动化设计, 而且有效避免模型计算量激增的问题.

3 时间卷积神经架构搜索(NAS-TC)

本节首先简要回顾神经架构搜索的过程, 然后介绍 NAS-TC 的基本组件, 接着重点介绍 NAS-TC 的构建过程和整体的网络结构, 最后对 NAS-TC 的实现细节进行了详细的描述.

3.1 神经架构搜索基础

神经架构搜索通常从一组预定义的操作集O开始, 并利用控制器根据这些操作集创建的搜索空间来获取大量的候选神经网络结构. 然后在训练集上训练候选网络架构, 并根据它们在验证集上的准确性进行排序. 最后将候选神经网络架构的排名信息作为反馈来调整搜索策略, 从而对网络结构进行更新. 当达到终止条件时, 终止搜索过程并选择最优的神经网络结构[7].

图 2 展示了基于单元网络的梯度架构搜索 DARTS[15] 的单元网络的内部结构. 具体的, 每个单元由 2 个输入节点、4 个中间节点和 1 个输出节点组成. 其中$cell_k$的 2 个输入节点为I_{k-1}和I_{k-2}, 它们是前面 2 个单元网络的输出. I_k是$cell_k$的输出节点. 每个单元内部有 4 个中间节点B_0, B_1, B_2, B_3, 用于进行特征提取和处理. 最后将 4 个中间节点的特征串联起来, 得到$cell_k$的输出I_k. 节点B_i, B_j之间的有向边表示特征操作o_{ij}. 在训练阶段, 中间节点B_j的计算方式为:

$$B_j = \sum_{o\in O} \frac{\exp(\alpha_o^{(i,j)})}{\sum_{o'\in O}\exp(\alpha_{o'}^{(i,j)})} o(B_i)\,, \tag{1}$$

其中$\alpha_o^{(i,j)}$表示操作O的参数化后的权重. 通过这种方式, 它将原来离散化的操作O松弛化为连续状态, 然后使用梯度下降优化方法来更新网络的结构, 从而获得最优的网络结构, 即一组连续化的参数$\alpha=\{\alpha^{(i,j)}\}$.

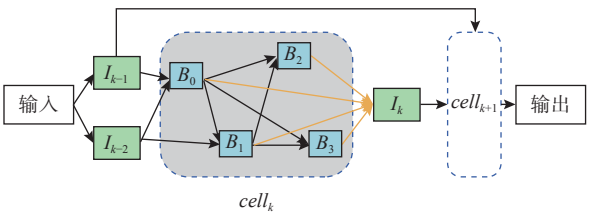


Fig. 2 Structure diagram of neural architecture search network based on cell network stacking

图 2 基于单元网络堆叠的神经架构搜索网络结构图

在测试阶段, 每对节点间只保留关联操作中权重最高操作, 以获得离散化后的网络结构, 计算方式为:

$$o^{(i,j)} = \arg\max_{o\in O} \alpha_o^{(i,j)}\,. \tag{2}$$

通过这种方法, 模型可以在一定程度上降低神经架构搜索的计算量以及对 GPU 显存大小的要求.

3.2 NAS-TC 的基本组件

本文采用经典的基于搜索单元堆叠的神经架构搜索网络结构[63], 它主要由正常单元和压缩单元交替堆叠而成. 因此, 通常只需要搜索 2 个单元的结构即可确定整个网络模型的结构. 接下来, 我们介绍本文设计的时间维度上的候选操作集和单元网络这 2 个基本组件.

1)时间维度上的候选操作集. 为了让神经架构搜索能够在视频领域高效地工作, NAS-TC 为其特别设计了一个基于时间维度的候选操作集. 如图 3(a)所示, $node_i$和 $node_j$分别是单元网络中的第i个节点和第j个节点, 节点表示神经网络中的特征嵌入. 中间的 9 个基础操作是单元网络中节点间的候选操作集O. 候选操作集O中的所有操作都是对$\mathcal{T}$(如果适用)的 1 维操作, 包括卷积核为$3\times1\times1$, $5\times1\times1$, $7\times1\times1$的可分离卷积(SepConv_k), 以及卷积核为$3\times1\times1$和$5\times1\times1$的扩张可分离卷积(DilConv_k)、$2\times1\times1$最大池化(Max_pool_2)、$2\times1\times1$平均池化(Avg_pool_

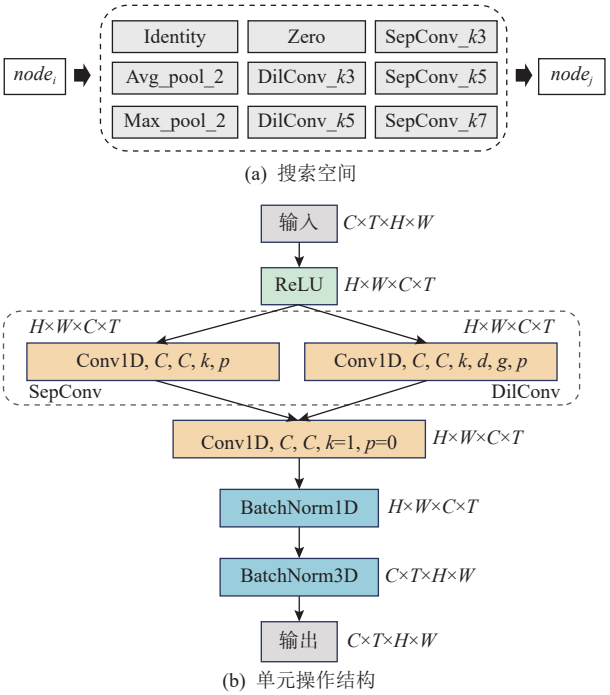


Fig. 3 Schematic diagram of search space and unit operations structure

图 3 搜索空间和单元操作结构示意图

2)、恒等操作(Identity)、零操作(Zero). 所有操作的步幅为 1(即 *stride* = 1, 如果适用). 此外, 我们使用适当的填充来保持特征图的原始尺寸以方便进行特征图拼接. 其中可分离卷积和扩张可分离卷积的结构顺序为 ReLU—Conv—BN, 相应的示意图被展示在图 3(b). 具体地, 单元操作的细节参数被展示在表 1.

Table 1 Parameters of Candidate Operation Sets

表 1 候选操作集的参数

操作集	核大小 k	膨胀率 d	组数 g	填充数 p
Identity	-	-	-	-
Zero	-	-	-	-
Avg pool 2	(2,1,1)	-	-	(1,0,0)
Max pool 2	(2,1,1)	(2,1,1)	-	(1,0,0)
DilConv $k3$	(3,1,1)	2	C	(2,0,0)
DilConv_$k5$	(5,1,1)	2	C	(4,0,0)
SepConv $k3$	(3,1,1)	1	-	(1,0,0)
SepConv $k5$	(5,1,1)	1	-	(2,0,0)
SepConv $k7$	(7,1,1)	1	-	(3,0,0)

注: "-"表示不适用; 括号里的数字分别是 x, y, z 轴上的值.

2)单元网络. 本文采用基于松弛化技术的单元网络结构[15]. 该结构使用梯度优化的方法对松弛化的候选操作集进行优化, 以获得节点间最优的候选操作, 从而得到整个单元网络的结构. 图 2 显示了这种由多个单元堆叠的网络结构示意图. 我们的目标是使用神经架构搜索获得最优结构的单元网络.

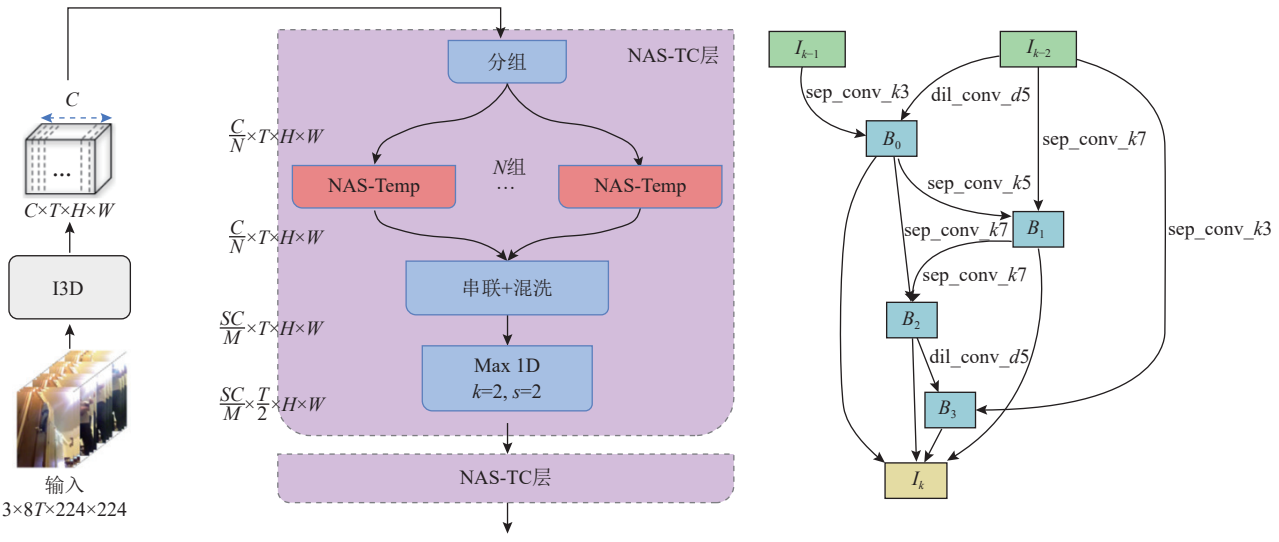
3.3 NAS-TC 总体框架

NAS-TC 的总体框架如图 4(a)所示, 它主要由骨干网络和多个堆叠的 NAS-TC 层组成. 其中我们使用 I3D[56] 作为模型的骨干网络来提取低级别的时空特征. 然后使用神经架构搜索来为长视频高级别时间信息的提取搜索最优的单元网络结构. 具体内容如下:

1)骨干网络. 对于长度为 t 的视频样本, 我们首先将 t 切分为 T 个片段, 然后将 8 个连续帧作为每个段的超级帧. 拼接提取的 T 个超级帧一起输入到骨干网络 I3D[56] 中以提取低级别的时空特征. 我们使用 mixed-5c 的特征图(即 I3D 的最后一个时空卷积层)作为后续网络的输入. 即, 从特征角度上来讲, 原始视频样本 $3\times 8T\times 224\times 224$ 使用 I3D 执行特征提取后获得的特征维度为 $C\times T\times H\times W$.

2)NAS-TC 层. 我们使用图 3(a)中的单元操作为神经架构搜索节点间的候选操作集, 使用骨干网络提取的特征作为 NAS-TC 层的输入. NAS-TC 层采用分组卷积和通道交换[50] 来尽可能降低计算量. 具体地, 在第 i 个 NAS-TC 层中, 我们首先将特征图 $\boldsymbol{X}_{i-1}\in\mathbb{R}^{C\times T\times H\times W}$ 分为 N 个通道组, 并将每个组的特征 $\boldsymbol{x}_j\in\mathbb{R}^{\frac{C}{N}\times T\times H\times W}$ 作为 NAS-Temp 的输入. 然后, NAS-TC 层对获得的 N 个 NAS-Temps 的输出特征在通道维度上进行特征拼接和混洗[63−64], 从而进行跨通道的特征学习. 接着, 在 T 维度上执行 $2\times 1\times 1$ 的最大池化. 因此, 第 i 个 NAS-TC 层的输出为 $\boldsymbol{X}_i\in\mathbb{R}^{\frac{SC}{M}\times\frac{T}{2}\times T\times H\times W}$. 其中 S 是 NAS-Temp 单元网络中的中间节点的数量, $\frac{S}{M}$ 是通道数的缩放比率. 最后, 我们将多个 NAS-TC 层简单地堆叠并串联 MLP 层来完成最终的动作识别任务.

其中核心组件 NAS-Temp 模块通过在时间维度



(a) NAS-TC的整体框架 (b) NAS-Temp

Fig. 4 Network architecture of NAS-TC

图 4 NAS-TC 的网络架构

上的神经架构搜索方法搜索获得,具体的实验设置细节参见 4.2 节.需要注意的是,先前的方法[17,65]通常需要搜索 2 个单元网络:1 个正常单元网络和 1 个压缩单元网络.正常单元网络主要用于进行特征提取和处理,而压缩单元网络则被用于减小特征图的大小.与它们不同的是,NAS-TC 只需要搜索正常单元网络.这主要是因为在骨干网络处理后获得的特征图的空间维度已降低到相对较小的水平.因此,随后的 NAS-TC 层上特征图的空间维度保持不变即可.

4 实 验

4.1 数据集

为了验证模型在长视频中复杂动作上的识别性能,我们选择在数据集 Charades[10], MultiTHUMOS[28], BreakfastActions[27] 上测试模型的性能.那些视频长度小于 30 s 甚至更短的数据集不包含在本文的范围内.这些数据集的详细信息如下:

1) Charades 是一个大型多标签视频分类数据集.它包含了大量的人类日常活动.该数据集有 9 848 个视频,共有 157 个类别,训练集和验证集分别包括 8 000 和 1 800 个视频.每个视频约 30 s,平均包含 6.8 个单元动作,可满足长期视频复杂动作识别任务的要求.因此,这是一个非常具有挑战性的基准数据集.

2) MultiTHUMOS 是从原始数据集 THUMOS14[66] 扩展而来的.它是一个未经修剪的人类运动的多标签数据集,包含 400 个视频和 65 个动作类别,每个视频平均包含 11 个单元动作,总计 30 h.与 Charades 数据集相比,该数据集的每个视频中的动作更加复杂.

3) BreakfastActions 是迄今为止最大的完全注释的长视频动作数据集之一.该数据集包含与早餐准备相关的 10 个动作,由 52 个人在 18 个不同的厨房中记录.它总共包含 1 712 个视频,训练集和测试集的大小分别为 1 357 和 335,视频的平均长度为 2.3 min.每个视频代表由一系列单元动作组成的活动,共有 48 种类型的单元动作.

表 2 总结了这 3 个数据集的统计信息.

Table 2 Statistics of the Datasets

表 2 数据集的统计信息

数据集	视频数	总时长/h	平均时长/min	每个视频的动作数	种类数
Charades	9 848	82	0.5	6	157
MultiTHUMOS	400	30	4.5	11	65
BreakfastActions	1 712	66	2.3	7	48

4.2 实现细节

在神经架构搜索的搜索阶段,遵循 DARTS[15],我们冻结骨干网络模型的参数,首先在 Charades 数据集的训练集上对 NAS-Temp 进行搜索,设置 epoch 为 50,并通过其在测试集上的性能来选择最优的 NAS-Temp 结构,从而构建最终的 NAS-TC 网络.具体地,在数据集 Charades 上,设置 NAS-TC 网络的 T=32,然后将 I3D 作为骨干网络进行特征提取,提取后的特征 $\boldsymbol{X}\in\mathbb{R}^{32\times H\times W\times C}$,其中 $C=1\,024, H=W=7$.在 NAS-TC 层中,$M=3$,中间节点 $S=4$.如图 4 (b)展示了通过 NAS-TC 得到的 NAS-Temp 单元的结构.在神经架构搜索结束后,遵循 Timeception[26],我们冻结骨干网络参数,仅训练插入的多个 NAS-TC 和 MLP 层.我们使用 Adam 作为优化器,其中 $lr=0.01$ 和 $eps=1\mathrm{e}-4$,并以 $batch=18$ 进行了 300 个 epoch 的训练.对于 TC 层数的设置,为了方便比较,我们与 Timeception 的设置保持一致.遵循先前的工作[26,60],我们使用单标签动作准确率(Acc)或多标签单元动作平均精度均值(mAP)作为评估指标.

4.3 在 Charades 上的实验

4.3.1 长程时间依赖性学习能力比较

我们首先在不同时间步长输入下将 NAS-TC 与基线 Timeception[26] 进行性能比较.在表 3 中,我们通过改变输入时间的步数来观察模型对长程时间依赖性的学习能力.我们使用 I3D[56] 作为 Timeception 和 NAS-TC 的公共骨干网络,在 I3D 之上插入多个 TC 层(即 Timeception 或 NAS-TC 层)以对它们进行公平的比较.其中 CNN_Steps 是 CNN(即 I3D)的输入时间步长,而 TC_Steps 是第 1 个 TC 层的输入时间步长;输入到骨干网络的时间步长 $CNN_Steps\in\{256,512,1\,024\}$;因此,输入到第 1 个 TC 层的时间步长 $TC_Steps\in\{32,64,128\}$.

从表 3 可以看出,当 *CNN_Steps* 逐渐增加时,Timeception[26] 和 NAS-TC 的性能单调递增.这表明 NAS-TC 和 Timeception 方法已经能够在一定程度上学习长程时间信息.然而,相比之下,NAS-TC 使用更少的

Table 3 Performance Comparison of Models with Different Time Step Inputs

表 3 不同时间步长输入下模型的性能对比

I3D	*CNN_Steps* ⇒ *TC_Steps*	Timeception[26]		NAS-TC (本文)	
		参数量	mAP/%	参数量	mAP/%
+3 TC	256 ⇒ 32	2.0×10^6	33.9	1.5×10^6	34.5 (0.6 ↑)
+3 TC	512 ⇒ 64	2.0×10^6	35.5	1.5×10^6	37.2 (1.7 ↑)
+4 TC	1 024 ⇒ 128	2.8×10^6	37.2	2.0×10^6	39.3 (2.1 ↑)

注:括号内表示相对 Timeception 提升的数值;"↑"表示上升的百分点.

参数并实现更好的识别性能. 随着 *CNN_Steps* 的增加, 这种性能增益也变得更加明显. 这表明通过时间维度上的神经架构搜索学习到的 NAS- Temp 结构能够更有效地处理长程时间信息.

4.3.2 性能比较

为了进一步测试模型的有效性, 我们将 NAS-TC 与其他相关的方法进行比较. 这些方法包括 Two-stream[61], Two-stream+LSTM[61], ActionVLAD[67], Temporal Fields[61], Temporal Relations[57], ResNet-152[61], ResNet-152+Timeception[26], 3D ResNet-101[60], I3D[56], I3D+ActionVLAD[67], I3D+Timeception[26], I3D-NL[60], I3D+VideoGraph[68], EvaNet[69] 等方法[70-79]. 相应的性能比较结果被展示在表 4. 正如实验结果所示, I3D+NAS-TC 可以更有效地利用原始骨干网络学到的特征, 相比其他方法获得了最优的结果. 此外, 与 NAS-TC 最相关的人工设计的 Timeception 相比, I3D+NAS-TC 仍然获得了 2.1 个百分点的性能增益, 这证明表明了 NAS-TC 的有效性.

Table 4 Performance Comparison of Different Methods on Charades Dataset

表 4 在 Charades 数据集上不同方法的性能比较

方法	数据模态	mAP/%
R-C3D[70]	RGB + Flow	12.7
Two-stream[61]	RGB + Flow	18.6
Two-stream+LSTM[61]	RGB + Flow	17.8
ActionVLAD[67]	RGB + iDT	21.0
TemporalFields[61]	RGB + Flow	22.4
TemporalRelations[57]	RGB	25.2
ResNet-152[61]	RGB	22.8
ResNet-152+Timeception[26]	RGB	31.6
3D ResNet-101[60]	RGB	35.5
I3D+TAN[71]	RGB + Flow	17.6
I3D+Super-events[72]	RGB + Flow	19.4
I3D+TGM[73]	RGB + Flow	21.5
I3D+MLAD[74]	RGB + Flow	23.7
I3D+PDAN[75]	RGB + Flow	26.5
I3D+CTRN[76]	RGB + Flow	27.8
I3D+GCAN[77]	RGB + Flow	27.7
I3D[56]	RGB	32.9
I3D+ActionVLAD[67]	RGB	35.4
I3D+Timeception[26]	RGB	37.2
I3D-NL[60]	RGB	37.5
I3D+VideoGraph[68]	RGB	37.8
I3D+EvaNet[69]	RGB	38.1
I3D+GHRM[78]	RGB	38.3
I3D+UGPT[79]	RGB	38.8
I3D+NAS-TC（本文）	RGB	39.3 (2.1↑)

注：括号内表示相对于 I3D+Timeception，I3D+NAS-TC 获得的性能增益；“↑”表示上升的百分点.

4.3.3 搜索时间

我们使用提出的在时间维度上的神经架构搜索在 Charades 训练集上进行 NAS-Temp 模块的搜索. 50 个 epoch 在一个 GeForce RTX 3090 Ti 24 G 上需要 2.1 个 GPU 天, 这与 DARTS 在图像领域实现的搜索时间相似. 这表明视频领域中仅在时间维度上进行卷积搜索来实现模型的自动化设计并减少计算量是十分有效的.

4.4 在 MultiTHUMOS 上的实验

为了进一步验证模型的有效性, 我们还在 MultiTHUMOS 数据集上测试了模型的性能. 我们使用与 Charades 数据集类似的处理方法, 将 NAS-TC+I3D 与 Timeception 和其他相关模型的性能进行了比较. 遵循 Timeception, 最终的实验结果被展示在表 5 中. 从实验结果可以看出, 多尺度卷积明显优于固定的单尺度卷积, 因为它更灵活, 也可以更容易捕捉到复杂动作中不断变化的时间信息. 另外, 虽然 I3D+Timeception 和 I3D+NAS-TC 都采用了多尺度卷积操作, 相比之下 I3D+NAS-TC 仍然获得了 2.04 个百分点的性能增益. 这表明 I3D+NAS-TC 方法通过 NAS 自动地搜索单元网络的结构, 具有更灵活的时间信息捕获能力, 从而使 I3D+NAS-TC 获得更好的识别精度.

Table 5 Performance Comparison of Different Methods on MultiTHUMOS Dataset

表 5 在 MultiTHUMOS 数据集上不同方法的性能比较

方法	核大小k	膨胀率d	mAP/%
Two-stream[61]	-	-	27.60
Two-stream+LSTM[61]	-	-	28.10
Muti-LSTM[61]	-	-	29.60
I3D+super-events[72]	-	-	36.40
I3D+super-events+TGMs[73]	-	-	46.40
I3D[56]	-	-	72.43
I3D+Timeception[26]	3	1	72.83
I3D+Timeception[26]	3	1,2,3	74.52
I3D+Timeception[26]	1,3,5,7	1	74.79
I3D+NAS-TC（本文）	3,5,7	1,2	76.83 (2.04↑)

注：“-”表示不适用；括号内表示相比 I3D+Timeception，I3D+NAS-TC 获得的性能增益；“↑”表示上升的百分点.

4.5 在 BreakfastActions 上的实验

我们还在 BreakfastActions 数据集上测试了模型

的性能. 为了便于比较, 遵循先前的工作[25,77]中相同的设置, 从每个视频中均匀采样 64 个片段, 每个片段包含 8 个连续帧, 然后使用 NAS-TC 层的 3 个堆叠层的模型来对时间维度进行建模, 展示了模型的性能. 如表 6 所示, I3D+NAS-TC 相比于 I3D+Timeception 在单个动作和多个单元动作上的 Acc 和 mAP 分别提升了 4.82 个百分点和 2.82 个百分点, I3D+NAS-TC 仍然保持着明显的优势.

Table 6 Performance Comparison of Different Methods on BreakfastActions Dataset

表 6 不同方法在 BreakfastActions 数据集上的性能比较

方法	Activities Acc/%	Actions mAP/%
ResNet-152[61]	41.13	32.65
ResNet-152+ActionVLAD[67]	55.49	47.12
ResNet-152+Timeception[26]	57.75	48.47
ResNet-152+VideoGraph[68]	59.12	49.38
I3D[56]	58.61	47.05
I3D+ActionVLAD[67]	65.48	60.20
I3D+Timeception[26]	67.07	61.82
I3D+VideoGraph[68]	69.45	63.14
I3D+NAS-TC (本文)	71.89 (4.82 ↑)	64.64 (2.82 ↑)

注: 括号内表示相对于 I3D+Timeception, I3D+NAS-TC 获得的性能增益; "↑"表示上升的百分点.

4.6 模型对比与分析

在本节, 我们重点讨论了 NAS-TC 和 Timeception[26]的模型参数量对比和参数分配的合理性.

4.6.1 参数量比较

由于 NAS-TC 的架构与 Timeception 最为相似, 因此本节主要比较随 TC 层数的增加, 它们参数量的变化. 如图 5 展示了 NAS-TC 和 Timeception 的参数量随着 TC 层数增加的变化曲线. 具体地, 当 TC 层数为 0 时, 模型只有顶层的全连接层, 参数个数为 0.61×10^6. 随着 TC 层数的增加, NAS-TC 模型在参数量方面上的优势逐渐显示出来并变得更加明显.

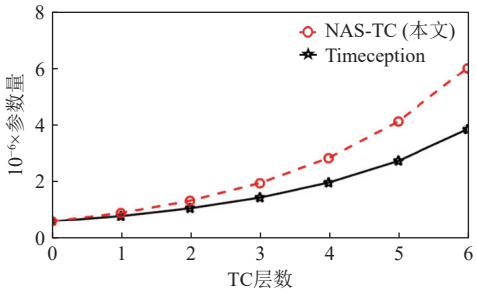


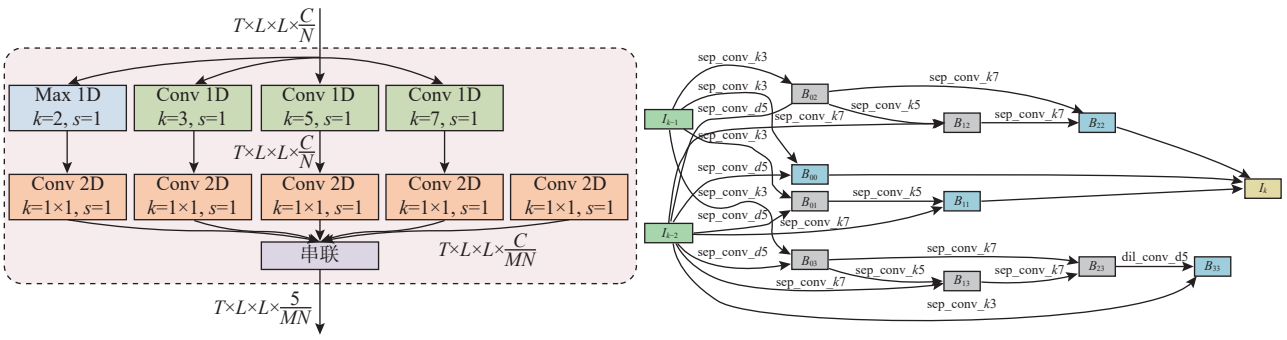
Fig. 5 Variation curve of model parameter quantity with increasing TC layers

图 5 模型参数量随着 TC 层数增加的变化曲线

在性能方面, 从表 3 可以看出 NAS-TC 表现出了明显的优势. 理论上, TC 层也可以堆叠成更多的层. 当 TC 层数为 8 时, NAS-TC 在时间维度上可以处理 4 096 个时间步, 这意味着 NAS-TC 可以处理 2.3 min 的长视频. 此时, NAS-TC 的优势(即需要较少的参数)将得到进一步地体现. 这里需要指出的是, 这并不意味着 TC 的层数越多就越好. TC 的层数越多意味着更大的帧样本. 由于视频长度是固定的, 过大的帧采样可能会导致严重的信息冗余问题.

4.6.2 参数分配的合理性分析

具体地, 图 6 展示了 Timeception 中时间卷积模块和 NAS-TC 中 NAS-Temp 的结构扩展视图的对比, 图 6 (b)是图 4(b)的展开图, 图 4(b)中的 4 个中间节点 B_0, B_1, B_2, B_3 分别与图 6(b)中的 4 个节点 $B_{00}, B_{11}, B_{22}, B_{33}$ 相对应. 对于节点 B_{ij}, i 表示 NAS-Temp 中的第 i 个中间节点, j 表示该节点与第 j 个节点相关联. 输入节点 I_{k-2}、I_{k-1} 和输出节点 I_k 与图 4 (b)中保持一致. 如图 6(a)显示了 Timeception 中的一个核心组件, 即时间卷积模块中的一个核心组件——时间卷积模块. 该模块是一个有多尺度卷积操作的模块, 它有 5 个分



(a) 时间卷积模块[26] (b) 图4(b)中NAS-Temp的展开视图

Fig. 6 Structure comparison of temporal convolutional module and NAS-Temp

图 6 时间卷积模块和 NAS-Temp 的结构对比

支，第 5 个分支仅在空间维度上使用 1×1 的 2D 卷积对输入进行降维操作，而其余 4 个分支则在时间维度上执行多尺度卷积操作或池化操作. 更多详细信息，请参阅 Timeception[26].

在时间卷积模块中，1×1 2D 卷积似乎占据了较小的参数预算，但事实并非如此. 实际上，1×1 2D 卷积在每个分支中都需要大量参数. 在时间卷积模块中，1×1 2D 卷积占据了 5 个分支中参数量的 57%，这些参数主要用于映射减少特征的通道数量. 同时，这些卷积操作也用于增加模型的复杂性，提高了其学习潜力. 然而，实际上只有 3 个分支中在时间维度上的卷积操作（即核大小为{3, 5, 7}）真正用于学习时间维度上的高级别的特征. 因此，基于以上分析，我们认为这个参数分配的合理性是存疑的.

相比之下，在 NAS-Temp 中，我们仅对 2 个输入执行必要的1×1的 2D 卷积，即I_{k-1}和I_{k-2}. 2 个输入的通道数通过1×1的 2D 卷积被压缩到I_{k-1}的1/3. 在 4 个中间节点 (B_0, B_1, B_2, B_3)中，通道数保持不变. 从图 6(b)中 NAS-Temp 的展开视图中，可以清楚地看到这些中间节点如何被重复使用的. 这种方式使每个中间节点的生成拥有足够的复杂度. 例如，为了生成节点$B_{00}, B_{11}, B_{22}, B_{33}$而执行的单元操作的数量分别为2,4,6,8. 与时间卷积模块相比，这种连接方式显著降低了整个 NAS-Temp 层中1×1 2D 卷积操作的比例，从而有效降低了模型的参数. 此外，由于 NAS-Temp 对不同的卷积操作进行了堆叠和拼接，它在时间维度上学习到的信息也更加复杂. 这种策略使 NAS-Temp 有可能从时间维度上学习到更多的信息. 因此，NAS-TC 模型更善于将有限的参数分配给在维度T上的信息学习，从而获得更好的性能.

5 总结和展望

本文提出了一个在时间维度上的时间卷积神经架构搜索（NAS-TC）方法. 它为视频动作识别设计了一个 2 个阶段的框架，将高复杂度的动作识别任务拆分成了特征提取和时间信息捕捉 2 个部分. 时间信息的捕捉对于视频任务的识别是至关重要的，而相应的模型设计却有很大的难度. 因此，本文为时间信息捕捉这一轻量级的任务设计了一个特殊的时间维度上的搜索空间，成功地为复杂动作识别提出了一个模型自动化设计的解决方案. 此外，在 NAS-Temp 中，中间节点的重复使用使模型有足够的能力学习长程的时间依赖关系，这使得 NAS-TC 可以使用相对较少的参数实现良好的性能. 并且，NAS-TC 还保持了 NAS 灵活自动化的优势，这为解决其他领域的类似问题提供了很好的参考.

作者贡献声明：任鹏真提出了模型设计和实验方案，并进行论文撰写；梁小丹和常晓军对模型和实验设计提出了指导意见；赵子莹对实验合理性提出指导意见；肖云提供课题和实验设备并对论文实验的组织和修改提出指导意见.

参 考 文 献

[1] Chang Xiaojun, Yu Yaoliang, Yang Yi, et al. They are not equally reliable: Semantic event search using differentiated concept classifiers[C]//Proc of the 29th IEEE Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2016: 1884−1893

[2] Ji Yuzhu, Zhang Haijun, Zhang Zhao, et al. CNN-based encoder-decoder networks for salient object detection: A comprehensive review and recent advances[J]. Information Sciences, 2021, 546: 835−857

[3] Ren Jiahuan, Zhang Zhao, Hong Richang, et al. Robust low-rank convolution network for image denoising[C]//Proc of the 30th ACM Int Conf on Multimedia. New York: ACM, 2022: 6211−6219

[4] Wu Zhihao, Zhao Zhang, Fan Jicong. Graph convolutional kernel machine versus graph convolutional networks[C/OL]//Proc of the 37th Advances in Neural Information Processing Systems. 2023[2024-08-01]. https://proceedings.neurips.cc/paper_files/paper/2023/hash/3ec6c6fc9065aa57785eb05dffe7c3db-Abstract-Conference.html

[5] Zoph B, Vasudevan V, Shlens J, et al. Learning transferable architectures for scalable image recognition[C]//Proc of the 31st IEEE Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2018: 8697−8710

[6] Real E, Aggarwal A, Huang Y, et al. Regularized evolution for image classifier architecture search[C]//Proc of the 23rd AAAI Conf on Artificial Intelligence. Palo Alto, CA: AAAI, 2019: 4780−4789

[7] Ren Pengzhen, Xiao Yun, Chang Xiaojun, et al. A comprehensive survey of neural architecture search: Challenges and solutions[J]. ACM Computing Surveys, 2021, 54(4): 1−34

[8] Zhang Xingwu, Ma Rui, Zhao Yu, et al. Differentiable sampling based efficient architecture search for automatic fault diagnosis[J]. Engineering Applications of Artificial Intelligence, 2024, 127(1): 107−214

[9] Meng Ziyao, Gu Xue, Liang Yanchun, et al. Deep neural architecture search: A survey[J]. Journal of Computer Research and Development, 2021, 58(1): 22−33 (in Chinese)
(孟子尧, 谷雪, 梁艳春, 等. 深度神经架构搜索综述[J]. 计算机研究与发展, 2021, 58(1): 22−33)

[10] Sigurdsson G A, Varol G, Wang Xiaolong, et al. Hollywood in homes: Crowdsourcing data collection for activity understanding[C]//Proc of

the 14th European Conf on Computer Vision. Berlin: Springer, 2016: 510−526

[11] Cai Han, Chen Tianyao, Zhang Weinan, et al. Efficient architecture search by network transformation[C]//Proc of the 22nd AAAI Conf on Artificial Intelligence. Palo Alto, CA: AAAI, 2018: 2787−2794

[12] Negrinho R, Geoff G. Deeparchitect: Automatically designing and training deep architectures[J]. arXiv preprint, arXiv: 1704.08792, 2017

[13] Liu Chenxi, Zoph B, Neumann M, et al. Progressive neural architecture search[C]//Proc of the 15th European Conf on Computer Vision. Berlin: Springer, 2018: 19−34

[14] Kandasamy K, Neiswanger W, Schneider J, et al. Neural architecture search with Bayesian optimisation and optimal transport[C/OL]//Proc of the 32nd Advances in Neural Information Processing Systems. 2018[2023-10-01]. https://proceedings.neurips.cc/paper_files/paper/2018/hash/f33ba15effa5c10e873bf3842afb46a6-Abstract.html

[15] Liu Hanxiao, Simonyan K, Yang Yiming. DARTS: Differentiable architecture search[J]. arXiv preprint, arXiv: 1806.09055, 2018

[16] Poliakov E, Hung Weijie, Huang Chingchun. Efficient constraint-aware neural architecture search for object detection[C]//Proc of the 15th Asia Pacific Signal and Information Processing Association Annual Summit and Conf. Piscataway, NJ: IEEE, 2023: 733−740

[17] Ozaeta M A A, Fajardo A C, Brazas F P, et al. Seagrass classification using differentiable architecture search[C]//Proc of the 26th Int Joint Conf on Computer Science and Software Engineering (JCSSE). Piscataway, NJ: IEEE, 2023: 123−128

[18] Howard Andrew G, Zhu Menglong, Chen Bo, et al. Mobilenets: Efficient convolutional neural networks for mobile vision applications[J]. arXiv preprint, arXiv: 1704.04861, 2017

[19] Zhang Xiangyu, Zhou Xinyu, Lin Mengxiao, et al. ShuffleNet: An extremely efficient convolutional neural network for mobile devices[C]//Proc of the 31st IEEE Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2018: 6848−6856

[20] Tong Zhan, Song Yibing, Wang Jue, et al. Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training[C/OL]//Proc of the 36th Advances in Neural Information Processing Systems. 2022[2023-08-17]. https://proceedings.neurips.cc/paper_files/paper/2022/hash/416f9cb3276121c42eebb86352a4354a-Abstract-Conference.html

[21] Zisserman A, Carreira J, Simonyan K, et al. The kinetics human action video dataset[J]. arXiv preprint, arXiv: 1705.06950, 2017

[22] Goyal R, Ebrahimi Kahou S, Michalski V, et al. The “Something Something” video database for learning and evaluating visual common sense[C]//Proc of the 30th IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE. 2017: 5842−5850

[23] Soomro K, Zamair A R, Shah M. UCF101: A dataset of 101 human actions classes from videos in the wild[J]. arXiv preprint, arXiv: 1212.0402, 2012

[24] Kuehne H, Jhuang H, Garrote E, et al. HMDB: A large video database for human motion recognition[C]//Proc of the 24th IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2011: 2556−2563

[25] Schindler K, Van Gool L. Action snippets: How many frames does human action recognition require?[C/OL]//Proc of the 21st IEEE/CVF Conf on Computer Vision and Pattern Recognition. 2008[2023-10-15]. https://ieeexplore.ieee.org/document/4587730

[26] Hussein N, Gavves E, Smeulders A W. Timeception for complex action recognition[C]//Proc of the 32nd IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2019: 254−263

[27] Hilde K, Ali A, Thomas S. The language of actions: Recovering the syntax and semantics of goal-directed human activities[C]//Proc of the 27th IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2014: 780−787

[28] Yeung S, Russakovsky O, Jin N, et al. Every moment counts: Dense detailed labeling of actions in complex videos[J]. International Journal of Computer Vision, 2018, 126: 375−389

[29] Zhou Jiaming, Li Hanjun, Lin Kunyu, et al. Adafocus: Towards end-to-end weakly supervised learning for long-video action understanding[J]. arXiv preprint, arXiv: 2311.17118, 2023

[30] Vaswani A, Shazeer N, Parmar N. Attention is all you need[C/OL]//Proc of the 31st Advances in Neural Information Processing Systems. 2017[2023-10-15]. https://www.semanticscholar.org/reader/204e3073870fae3d05bcbc2f6a8e263d9b72e776

[31] Bertasius G, Wang H, Torresani L. Is space-time attention all you need for video understanding?[C/OL]//Proc of the 38th Int Conf on Machine Learning. 2021[2024-01-01]. https://proceedings.mlr.press/v139/bertasius21a/bertasius21a-supp.pdf

[32] Tong Zhan, Song Yibing, Wang Jue, et al. VideoMAE: Masked autoencoders are data-efficient learners for self-supervised video pre-training[C]//Proc of the 36th Advances in Neural Information Processing Systems. 2022[2023-12-10]. https://proceedings.neurips.cc/paper_files/paper/2022/hash/416f9cb3276121c42eebb86352a4354a-Abstract-Conference.html

[33] He Kaiming, Chen Xinlei, Xie Saining, et al. Masked autoencoders are scalable vision learners[C]//Proc of the 35th IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2022: 16000−16009

[34] Ji Shuiwang, Xu Wei, Yang Ming, et al. 3D convolutional neural networks for human action recognition[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2012, 35(1): 221−231

[35] Li Chao, Zhong Qiaoyong, Xie Di, et al. Collaborative spatiotemporal feature learning for video action recognition[C]//Proc of the 32nd IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2019: 7872−7881

[36] Xie Saining, Sun Chen, Huang Jonathan, et al. Rethinking spatiotemporal feature learning: Speed-accuracy trade-offs in video classification[C]//Proc of the 15th European Conf on Computer Vision. Berlin: Springer, 2018: 305−321

[37] Wu Chaoyuan, Feichtenhofer C, Fan Haoqi, et al. Long-term feature banks for detailed video understanding[C]//Proc of the 32nd IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2019: 284−293

[38] Wu Chaoyuan, Li Yanghao, Mangalam K, et al. MeMViT: Memory-augmented multiscale vision transformer for efficient long-term video Recognition[C]//Proc of the 35th IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2022: 13587−13597

[39] Zoph B. Neural architecture search with reinforcement learning[J]. arXiv preprint, arXiv: 1611.01578, 2016

[40] Cai Han, Yang Jiacheng, Zhang Weinan, et al. Path-level network transformation for efficient architecture search[C/OL]//Proc of the 35th Int Conf on Machine Learning. 2018[2023-11-14]. https://proceedings.mlr.press/v80/cai18a.html

[41] Real E, Moore S, Selle A, et al. Large-scale evolution of image classifiers[C/OL]//Proc of the 34th Int Conf on Machine Learning. 2017[2023-10-08]. https://proceedings.mlr.press/v70/real17a.html

[42] Hussein N, Gavves E, Smeulders A W. Unified embedding and metric learning for zero-exemplar event detection[C]//Proc of the 30th IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2017: 1096−1105

[43] Habibian A, Mensink T, Snoek C G M. Video2vec embeddings recognize events when examples are scarce[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2016, 39(10): 2089−2103

[44] Girdhar R, Ramanan D. Attentional pooling for action recognition[C/OL]//Proc of the 31st Advances in Neural Information Processing Systems. 2017[2023-10-13]. https://proceedings.neurips.cc/paper/2017/hash/67c6a1e7ce56d3d6fa748ab6d9af3fd7-Abstract.html

[45] Fernando B, Gavves E, Oramas J, et al. Rank pooling for action recognition[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2016, 39(4): 773−787

[46] Oneata D, Verbeek J, Schmid C. Action and event recognition with fisher vectors on a compact feature set[C]//Proc of the 26th IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2013: 1817−1824

[47] Cosmin D I, Ionescu B, Aizawa K, et al. Spatio-temporal vector of locally max pooled features for action recognition in videos[C]//Proc of the 30th IEEE Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2017: 3097−3106

[48] Donahue J, Anne H L, Guadarrama S, et al. Long-term recurrent convolutional networks for visual recognition and description[C]//Proc of the 28th IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2015: 2625−2634

[49] Ghodrati A, Gavves E, Snoek C G M. Video time: Properties, encoders and evaluation[J]. arXiv preprint, arXiv: 1807.06980, 2018

[50] Huang Gao, Liu Zhuang, Van D M L, et al. Densely connected convolutional networks[C]//Proc of the 30th IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2017: 4700−4708

[51] Shi Wuzhen, Liu Shaohui, Jiang Feng, et al. Video compressed sensing using a convolutional neural network[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2020, 31(2): 425−438

[52] Song Xue, Xu Baohan, Jiang Yugang, et al. Predicting content similarity via multimodal modeling for video-in-video advertising[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2020, 31(2): 569−581

[53] Simonyan K, Zisserman A. Two-stream convolutional networks for action recognition in videos[C/OL]//Proc of the 28th Advances in Neural Information Processing Systems. 2014[2024-01-19]. https://proceedings.neurips.cc/paper_files/paper/2014/hash/00ec53c4682d36f5c4359f4ae7bd7ba1-Abstract.html

[54] Bilen H, Fernando B, Gavves E, et al. Action recognition with dynamic image networks[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 40(12): 2799−2813

[55] Tran D, Bourdev L, Fergus R, et al. Learning spatiotemporal features with 3D convolutional networks[C]//Proc of the 28th IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE. 2015: 4489−4497

[56] Carreira J, Zisserman A. Quo vadis, action recognition? A new model and the kinetics dataset[C]//Proc of the 30th IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2017: 6299−6308

[57] Zhou Bolei, Andonian A, Oliva A, et al. Temporal relational reasoning in videos[C]//Proc of the 15th European Conf on Computer Vision. Berlin: Springer, 2018: 803−818

[58] Wang Limin, Xiong Yuanjun, Wang, Zhe, et al. Temporal segment networks: Towards good practices for deep action recognition[C]//Proc of the 14th European Conf on Computer Vision. Berlin: Springer, 2016: 20−36

[59] Wang Limin, Xiong Yuanjun, Wang Zhe, et al. Temporal segment networks for action recognition in videos[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2018, 41(11): 2740−2755

[60] Wang Xiaolong, Girshick R, Gupta A, et al. Non-local neural networks[C]//Proc of the 31st IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2018: 7794−7803

[61] Sigurdsson G A, Divvala S, Farhadi A, et al. Asynchronous temporal fields for action recognition[C]//Proc of the 30th IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2017: 585−594

[62] Varol G, Ivan L, Cordelia S, et al. Long-term temporal convolutions for action recognition[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 40(6): 1510−1517

[63] Zoph B, Vasudevan V, Shlens J, et al. Learning transferable architectures for scalable image recognition[C]//Proc of the 31st IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2018: 8697−8710

[64] Zhang Xiangyu, Zhou Xinyu, Lin Mengxiao, et al. ShuffleNet: An extremely efficient convolutional neural network for mobile devices[C]//Proc of the 31st IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2018: 6848−6856

[65] Li Guihong, Duc H, Kartikeya B, et al. Zero-Shot neural architecture search: Challenges, solutions, and opportunities[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2024, 46(12): 7618−7635

[66] Haroon I, Amir R Z, Jiang Yugang, et al. The thumos challenge on action recognition for videos "in the wild"[J]. Computer Vision and Image Understanding, 2017, 155(1): 1−23

[67] Girdhar R, Ramanan D, Gupta A, et al. ActionVLAD: Learning spatio-temporal aggregation for action classification[C]//Proc of the 30th IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2017: 971−980

[68] Hussein N, Gavves E, Smeulders A W M. VideoGraph: Recognizing minutes-long human activities in videos[J]. arXiv preprint, arXiv: 1905.05143, 2019

[69] Piergiovanni A, Angelova A, Toshev A, et al. Evolving space-time neural architectures for videos[C]//Proc of the 17th IEEE/CVF Int Conf on Computer Vision. Piscataway, NJ: IEEE, 2019: 1793−1802

[70] Xu Huijuan, Das A, Saenko K. R-C3D: Region convolutional 3D network for temporal activity detection[C]//Proc of the 30th IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2017: 5783−5792

[71] Dai Xiyang, Singh B, Ng J Y H, et al. Tan: Temporal aggregation network for dense multi-label action recognition[C]//Proc of the 6th IEEE Winter Conf on Applications of Computer Vision. Piscataway, NJ: IEEE, 2019: 151−160

[72] Piergiovanni A, Ryoo M S. Learning latent super-events to detect multiple activities in videos[C]//Proc of the 31st IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2018: 5304−5313

[73] Piergiovanni A, Ryoo M. Temporal Gaussian mixture layer for videos[C/OL]//Proc of the 36th Int Conf on Machine Learning. 2019[2023-12-16]. https://proceedings.mlr.press/v97/piergiovanni19a.html

[74] Tirupattur P, Duarte K, Rawat Y S, et al. Modeling multi-label action dependencies for temporal action localization[C]//Proc of the 34th IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2021: 1460−1470

[75] Dai R, Das S, Minciullo L, et al. Pdan: Pyramid dilated attention network for action detection[C]//Proc of the 8th IEEE/CVF Winter Conf on Applications of Computer Vision. Piscataway, NJ: IEEE, 2021: 2970−2979

[76] Dai Rui, Das S, Bremond F. Ctrn: Class-temporal relational network for action detection[J]. arXiv preprint, arXiv: 2110.13473, 2021

[77] Wu Yuankai, Su Xin, Salihu D, et al. Modeling action spatiotemporal relationships using graph-based class-level attention network for long-term action detection[C]//Proc of the 36th IEEE/RSJ Int Conf on Intelligent Robots and Systems (IROS). Piscataway, NJ: IEEE, 2023: 6719−6726

[78] Zhou Jiaming, Lin Kunyu, Li Haoxin, et al. Graph-based high-order relation modeling for long-term action recognition[C]//Proc of the 34th IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2021: 8984−8993

[79] Guo Hongji, Wang Hanjing, Ji Qiang. Uncertainty-guided probabilistic transformer for complex action recognition[C]//Proc of the 35th IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2022: 20052−20061



Ren Pengzhen, born in 1993. PhD, engineer. His main research interests include multi-modal representation learning, visual language pre-training, and model automation design.

任鹏真, 1993 年生. 博士, 工程师. 主要研究方向为多模态表示学习、视觉语言预训练、模型自动化设计.



Liang Xiaodan, born in 1991. PhD, professor, PhD supervisor. Member of CCF. Her main research interests include computer vision, natural language understanding, and smart driving.

梁小丹, 1991 年生. 博士, 教授, 博士生导师. CCF 会员. 主要研究方向为计算机视觉、自然语言理解、智能驾驶.



Chang Xiaojun, born in 1986. PhD, professor, PhD supervisor. His main research interests include multi-modal learning, computer vision, and green artificial intelligence.

常晓军, 1986 年生. 博士, 教授, 博士生导师. 主要研究方向为多模态学习、计算机视觉、绿色人工智能.



Zhao Ziying, born in 1985. PhD, senior engineer, PhD supervisor. Her main research interests include artificial intelligence and spatial analysis, and the application of large model technology.

赵子莹, 1985 年. 博士, 高级工程师, 博士生导师. 主要研究方向为人工智能与空间分析、大模型技术应用.



Xiao Yun, born in 1978. PhD, professor, PhD supervisor. Senior member of CCF. Her main research interests include data mining, machine learning, and artificial intelligence algorithm research and applications.

肖　云, 1978 年生. 博士, 教授, 博士生导师. CCF 高级会员. 主要研究方向为数据挖掘、机器学习、人工智能算法研究及应用.