计算机科学
COMPUTER SCIENCE

视频识别深度学习网络综述

钱文祥, 衣杨

引用本文

钱文祥，衣杨. 视频识别深度学习网络综述[J]. 计算机科学, 2022, 49(11A): 211200025-10.

QIAN Wen-xiang, YI Yang. Survey of Deep Learning Networks for Video Recognition[J]. Computer Science, 2022, 49(11A): 211200025-10.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.211200025

视频识别深度学习网络综述

钱文祥[1,3]　衣　杨[1,2,3]
1 中山大学计算机学院　广州 510275
2 广州新华学院信息科学学院　广州 510520
3 广东省大数据分析与处理重点实验室　广州 510275
(qianwx3@mail2.sysu.edu.cn)

摘　要　视频识别是计算机视觉领域中最重要的任务之一,受到了研究者的广泛关注。视频识别指从视频片段中提取特征,并依据特征识别视频动作。相比于静态图片,视频的各帧间存在较大的关联性。如何高效地使用来自时空等不同维度的特征信息准确地识别视频,是当前研究的重点。以视频识别技术为研究对象,首先介绍了视频识别研究的背景信息及常用数据集。然后,详细地梳理了视频识别方法的演变过程;回顾了基于时空兴趣点、密集轨迹、改进的密集轨迹等传统的视频识别方法,以及近年来提出的可用于视频识别的深度学习网络框架。其中,分别介绍了基于2D卷积神经网络的视频识别框架、基于3D卷积神经网络的视频框架、伪3D卷积神经网络,以及基于Transformer结构的网络,介绍了这些框架的演变,并总结了它们的实现细节及特点;评测了各网络在不同视频识别数据集上的表现情况,分析了各网络的适用场景。最后,展望了视频识别网络框架未来的研究趋势。视频识别任务可以自动、高效地识别出视频所属的类别,基于深度学习的视频识别具有广泛的实用价值。

关键词:视频识别;改进的密集轨迹;深度学习;双流网络;卷积神经网络;深度自注意力网络
中图法分类号　TP183

Survey of Deep Learning Networks for Video Recognition

QIAN Wen-xiang[1,3] and YI Yang[1,2,3]
1 School of Computer Science and Engineering,Sun Yat-sen University,Guangzhou 510275,China
2 School of Information Science,Guangzhou Xinhua University,Guangzhou 510520,China
3 Guangdong Key Laboratory of Big Data Analysis and Processing,Guangzhou 510275,China

Abstract　Video recognition is one of the most important tasks in computer vision research,which is concerned by many researchers. Video recognition refers to extracting the key features from different video clips,analyzing these features,and classification of the video. Compared to a single,static picture,there are many significant differences between frames of a video clip. How to tell the differences through the dimension of spatial-temporal information from video clips are well concerned by researchers. Taking video recognition technology as the target of the research,first,this paper introduces the basic concepts of video recognition and challenges in this area,together with some of the most frequently used datasets in video recognition tasks. Then,the classic video recognition methods based on spatio-temporal interest points,dense trajectories,and improved dense trajectories are reviewed. Also,the deep learning network frameworks for video recognition proposed in recent years are then summarized. They are summarized according to the time order of their proposal and grouped by the different architecture of their network. Among them,the video recognition framework based on 2D convolution neural network is introduced,including two-stream convolutional network architecture,long short-term memory network,and long-term recurrent convolutional network. Then,a framework based on a 3D convolutional neural network is introduced,including Slowfast Network,X3D(eXpand 3D) Network. Following that,the pseudo-3D convolutional neural network is introduced,including R(2+1)d network,Pseudo-3D residual network,and a set of light-weight networks based on building models on temporal information. At last,a Transformer-based network is introduced,including Timesformer,video vision Transformer,shifted window Transformer(Swin Transformer). The evolution of these deep learning frameworks,their implementation details and characteristics are analyzed. The performance of each network on different datasets is evaluated,and the applicable scenarios of each network are analyzed. In the end,the future research trend of video recognition network framework is prospected. Video recognition task can automatically and efficiently recognize the category to which the video belongs,and video recognition based on deep learning has a wide range of practical value.

Keywords　Video recognition,Improved dense trajectory,Deep learning,Two-stream network,Convolutional neural network,

基金项目:广州市科技计划项目(202002030273,202102080656);广州新华学院重点学科项目(2020XZD02)
This work was supported by the Guangzhou Science and Technology Project(202002030273,202102080656) and Key Discipline Project of Guangzhou Xinhua University(2020XZD02).
通信作者:衣杨(issyy@mail.sysu.edu.cn)

Deep self-attention network

1 引言

视频识别是从视频序列的各帧中提取特征信息，并识别出视频动作的任务。它是计算机视觉研究领域的基础任务之一。近年来，视频的创建量随着手机、安防、汽车等具有摄像功能设备的普及而迅速增加。同时，基于深度学习网络的视频识别技术不断发展，视频识别的准确度和速度也得到了显著的提高，并在智能安防、自动驾驶、智慧医疗、体育直播等方面发挥了重要作用。

视频识别的流程主要分为 3 个环节：特征提取、特征编码、特征分类。图 1 给出了这 3 个环节。特征提取环节一般分为全局特征提取和局部特征提取两类。全局特征指视频中对象的轮廓等整体特征；而局部特征则指视频中的局部兴趣点，如灰度信息产生剧烈变化的局部时空区域。分类的过程要求对视频中的特征进行处理，并提取出可以表征视频的关键信息的归一化特征向量。获取视频特征的归一化向量后，视频识别问题转变为分类问题。特征会作为最终识别的依据，因此特征的质量对视频识别的效果起到决定性的作用。

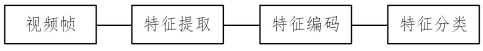


图 1 视频识别过程

Fig. 1 Procedures of video recognition

传统的用于视频识别的算法研究集中于获取物体外观的编码特征以及提取运动信息的特征，并通过使用手工制作的特征来描述局部形状，存在较强的主观性。常用的特征包括 Dalal 等[1]提出的定向梯度直方图(Histograms of Oriented Gradients，HOG)、Chaudhry 等[2]提出的光流直方图(Histograms of Oriented Optical Flow，HOF)和 Wang 等[3]提出的运动边界直方图(Motion Boundary Histogram，MBH)。传统算法通常计算视频光流的不同特征，并通过词袋模型(Bag of Words，BoW)[4]或基于 Fisher 向量的编码[5]对获取到的特征生成全局视频级描述符。在特征提取任务完成后，通常会使用支持向量机(Support Vector Machine，SVM)等常用的模式分类算法对检测出的目标进行分类识别工作。传统方法在一些动作识别类的任务(如行人检测、人体检测)上取得了一定的效果，但对复杂场景下的视频识别任务效果有限。

近几年来，随着网络算法的优化和硬件算力的增强，深度学习在对象识别和分类领域上表现出色[6-8]。深度学习应用到视频识别任务后，视频识别的准确率与速度都得到了很大的提升，但仍然面临着一些问题和挑战。

1.1 视频识别任务的问题与挑战

真实世界的复杂性为视频识别任务带来了许多问题与挑战。

(1)复杂环境因素的干扰

从现实世界采集到的运动数据有十分复杂多样的环境和背景信息，这与实验室环境所训练出的模式可能不一样。光照条件、背景信息带来的噪声都会导致提取可靠的特征变得困难，会对识别结果产生很大的影响。另一个挑战是对信源拍摄到的模糊对象的识别，包括高速移动的物体被低帧率摄像设备采集后产生的运动模糊效果。另外，物体距离摄像机距离较远时，呈现在视频中的物体像素少，难以提取到有效特征。这需要视频识别算法能够有效地提取特征，并表现出较强的鲁棒性。

(2)相似类别间的模式差异

一些类别有相似的模式，例如运动的卡车和火车、步行和慢跑、踢球和跑步、坐下和起立。这些类别的相似度较高，在识别这些物体或动作时，要将其准确分辨，需要算法可以在时空维度提取到更精细、准确的特征。

(3)特殊场景下高实时性的要求

伴随着视频输入源的分辨率和帧率的提升，一些对实时性有高要求的场景，如足球比赛实时分析、汽车违章监控等，都对识别和推理的实时性提出了更高的要求。视频识别使用的网络和模型需要向着准确度与速度提升、延迟下降的方向发展。

1.2 视频识别任务常用的数据集

目前国内外有很多公共的数据集可供广大科研人员下载用来评估和训练网络。

(1) ImageNet 数据集[9]

ImageNet 数据集包含超过 1 400 万张图片，共有 2.1 万个类别。ImageNet 的每张图片都经过严格的人工筛选与标记，其中约 100 万张图片还具有边界框信息。ImageNet 可用于预训练与视频识别任务相关的网络，为模型提供先验知识。

(2)Sports-1M 数据集[10]

该数据集包含 1 113 158 段与体育相关的视频，共 487 个类别，平均每个类别包含 1 000 至 3 000 个视频，其中约 5%的视频为多类别视频。数据集的标签由分析数据源的元信息得到，因此部分标签会与视频内容不符，或者标签无法描述视频的所有帧。例如，标签为“足球”的视频片段，也会包含采访、记分牌的镜头。

(3)Something-Something-v2 数据集[11]

该数据集包含人为主体，对日常物品执行基本操作的视频。该数据集是由大量的群组工作人员创建的，它包含 174 类，共 220 847 个视频片段。其中训练集包含 168 913 个视频片段，验证集包含 24 777 个视频片段，测试集包括 27 157 个视频片段。

(4)UCF101 数据集[12]

该数据集由美国中佛罗里达大学(University of Central Florida)采集，包含 101 个不同的动作类别，因此被称为 UCF101。它是目前评估网络性能时常用的数据集之一。该数据集共包含 13 320 个从 YouTube 收集的视频片段。数据集包含如理发、冲浪、剃须、爬行等各种来自真实世界的动作。

(5)DeepMind Kinetics 人类行为数据集[13]

Kinetics 数据集取自 Youtube，为一组大规模、高质量的数据集。常用的 Kinetics 数据集包括 Kinetics-400，Kinetics-600 和 Kinetics-700 3 种，分别包括 400，600，700 个人体动作类别。数据集中每个视频的时长大约为 10 s，由人工标记动作类别。该数据集中包含常见且多样的人类与物体交互的动作，例如乐器演奏、人与人之间的交互的动作，如握手和拥抱。

Kinetics-400 数据集包含大约 30 万个视频片段，每类动作至少有 400 个视频片段。Kinetics-600 数据集则包含约 50 万个

视频片段,每个类别至少 600 个视频片段。Kinetics-700 则包含约 65 万个视频片段,每类动作至少包含 600 个视频片段。

在 2020 年,Smria 等还发布了一个 Kinetics-700-2020,它是 Kinetics-700 数据集的补充和扩展。在这个新版本中,700 个动作类别中的每个类至少有 700 个视频片段。

(6)AVA 数据集[14]

AVA 动作数据集(AVA Actions Dataset)由 430 个来自电影的视频片段组成。每个视频片段的长度为 15 min,并包含 80 个原子动作的注释。AVA 数据集共标记 162 万个动作标签,各动作标签都多次出现在视频中。

AVA-Kinetics 数据集[15]是 AVA 动作数据集和 Kinetics 人体行为数据集的交叉,包含超过 23 万个视频片段。该数据集使用 AVA 注释协议,为 Kinetics-700 中的视频提供了动作标签,并为在视频的关键帧中出现的人体动作进行标注,共包含 80 个 AVA 动作类。

(7)HMDB51 数据集[16]

该数据集由电影、公共的视频收集整理而来,是常用的数据集之一。该数据集包含 6849 个视频片段,共有 51 个动作类别,如大笑、喝水、梳头、翻跟头、跳跃等类别。每个类别至少包含 101 个视频片段。

(8)Charades 数据集[17]

Charades 数据集包含 9848 个视频片段,包括 157 种不同的室内活动片段。该数据集具有 66500 个时间注释信息,以及 27847 个文本描述。Charades 数据集的更新版本 Charades-Ego 数据集[18]同样由室内活动剪辑组成,视频类别数仍为 157 种,共包含 7860 个由第一视角和第三视角拍摄的视频片段。该数据集具有 68536 个时间注释,以及视频对应的文本描述,可以用于视频分类、目标定位、字幕生成等视觉任务。

(9)EPIC-Kitchen 数据集[19-21]

EPIC-Kitchen 是由拍摄者使用头戴式摄像头,在厨房拍摄的各种行为的第一视角视频组成的数据集,分为 EPIC-Kitchen-55 和 EPIC-Kitchen-100。EPIC-Kitchen-55 包括 55 h 共 1150 万帧的高清视频,数据集包括 39594 个动作片段以及 454255 个物体边框。其扩展数据集 EPIC-Kitchen-100 由 100 h 共 2000 万帧的高清视频组成,共包含 9 万个动作片段,2 万个描述信息,可以用于动作识别、行为检测、行为预测等视觉任务中。

(10)Youtube-8M 数据集[22]

Youtube-8M 数据集是由谷歌 AI 团队创建的大型视频数据库。该数据集由 700 多万个共 4716 个类别的 YouTube 视频片段组成。视频总长度为 50 万小时,涵盖体育、游戏、艺术等 24 个主题。YouTube-8M 数据集也有一个扩展性质的 YouTube-8M 细分数据集,它带有人工验证的细分注释。

该数据集还包括对视频中的实体进行时间定位。该细分数据集共包含整理自 Youtube-8M 验证集的 1000 个类别共 23.7 万段的人工验证标签。Youtube-8M 数据集可以作为大规模分类数据集、时间定位数据集被用在视频分类、视频理解任务上。

本文第 2 节介绍传统的视频识别算法及其研究进展;第 3 节介绍多个用于视频和图像识别的深度学习网络框架,针对其的特点,分别介绍这些框架适用的任务,并在第 4 节将各个框架进行横向对比;最后对视频识别的现状进行总结,对其未来研究方向进行展望。

2 传统的视频识别算法概述

传统的视频识别算法常基于局部特征描述符来完成。识别视频时,一般先进行局部时空区域的特征提取,并对这些特征进行分类以获得识别结果。Antipov 等[23]认为传统的特征提取算法一般依赖人工设计的特征。HOG 是一种在图像场中计算的空间特征,它通过区块的梯度方向来分别统计累计的梯度强度,并以此作为该区块的特征。HOF 和 MBH 在光流图像上计算得到的视频特征,属于时间特征。HOF 计算各帧中光流方向得到直方图信息,它既能表征时域动作信息,又对图像的尺度和运动信息方向不敏感。通过归一化直方图,可以实现 HOF 特征的尺度不变性。MBH 计算方便快捷,它将光流图像分解为 x 方向和 y 方向上的灰度图像,并提取它们的梯度直方图,通过提取运动物体的边界信息来获取视频描述符。为了进一步增加 HOG 特征的有效性,Kläser 等[24]将 HOG 特征扩展到三维,形成 HOG3D,可以在多尺度下对时空块进行快速采样。在分类步骤,则一般使用线性 SVM 分类器。

传统算法中表现较好的是 Wang 等[25-26]提出的密集轨迹(Dense Trajectory,DT)和改进的密集轨迹(improved Dense Trajectory,iDT)方法。DT 分析光流信息获取视频的运动轨迹,并沿着运动轨迹提取特征。iDT 方法主要使用 HOG,HOF 和 MBH 3 种主要特征,抽取并描述密度轨迹的过程如图 2 所示。

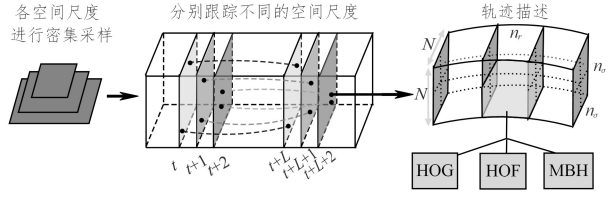


图 2 iDT 方法抽取并描述密度轨迹的流程

Fig. 2 Process of extracting and describing dense trajectory using iDT method

为了检测人的运动,iDT 也对相机运动进行了补偿,用额外的检测器来检测人的动作。iDT 方法存在特征维度高,导致特征可能比原始视频还要大,且识别速度慢的问题。但在深度学习的方法成熟以前,iDT 方法是效果最好的经典方法。根据 Tran 等[27]的测试结果,在 UCF101 数据集上,将 iDT 与 BoW 配合使用,通过线性 SVM 分类,可取得 76.2%的准确率;iDT 与 Fisher 向量配合,可以取得 87.9%的准确率。而即使是基于深度神经网络的视频识别方法,在和 iDT 结合之后,仍然能得到一定程度的效果提升[28]。

3 基于深度学习的视频识别框架概述

Zhang 等[29]认为深度神经网络已经被证实在识别图像、物体检测等计算机视觉任务上可以取得接近甚至超过人类识别水平的成果。Hinton[30]提出的基于受限玻尔兹曼机(Restricted Boltzmann Machine,RBM)的深度置信网络(Deep Belief Networks,DBN)也可用于视频识别任务。在靠近输入层的一端使用贝叶斯网络,而在远端堆叠使用 RBM,便可以得到 DBN。该网络自底向上进行无监督学习,再自顶向下进行监督学习以微调模型的参数。Taylor 等[31]将受限玻尔兹曼机在时间维度上进行扩展,把早期的可见层与当前时刻的

隐含层建立连接，提出条件受限玻尔兹曼机[31-32](Conditional Restricted Boltzmann Machines)，并将其用于人体行为识别任务中。但其计算代价高昂，可扩展性不强。为将 DBN 应用于日常任务中，Chen[33]提出基于卷积受限玻尔兹曼机(Convolutional Restricted Boltzmann Machines，CRBM)的时空深度置信网络(Space-Time Deep Belief Network，ST-DBN)。它结合了 DBN 与卷积网络的特点，可以通过汇集时空信息，提取到视频数据的不变特征，并可以在 KTH 数据集上取得 91%的识别准确率。

在视频识别任务中，基于最新的卷积神经网络或 Transformer 的模型通常比 DBN 表现得更好[34]，最近的研究[35]也显示基于 Transformer 的网络模型通常可以取得最高的识别准确度。Socher 等[36]认为在样本足够多的情况下，相比于传统方法，基于深度学习的网络可以很好地适应各种不同视频内容，并学习到隐含在视频中的特征。

本文按照框架网络特性，把常见的视频识别网络分为四大部分进行梳理：1)基于 2D 卷积神经网络的视频识别框架；2)基于 3D 卷积神经网络的视频框架；3)改进的 2D 卷积神经网络；4)基于 Transformer 结构的网络框架。

3.1 基于 2D 卷积神经网络的视频识别框架

在图像识别任务中，CNN 可以学习更高级的表征，而无需无监督的预训练。因此 CNN 早期基于逐帧识别的 2D 卷积神经网络[37]简单地在视频的各个帧上进行识别，这样的方法速度快，但无法有效利用时间信息，无法准确识别具有先后顺序的动作，例如难以对下蹲和站起这两个依赖时序信息的动作进行分类。

Hochreiter 等[38]提出了卷积神经网络与长短期时间记忆单元(Long Short Term Memory，LSTM)相结合的网络。LSTM 适合用于处理时间上有序的序列，并预测时间序列中间隔和延迟较长的关键事件。为了处理长时间的视频，基于 LSTM，Ng 等[39]提出了 5 种卷积时间特征池架构：1)卷积-池化(Conv Pooling)，可以保留时空信息；2)滞后池化(Late Pooling)，可以组合高维抽象特征的时间信息；3)慢速池化(Slow Pooling)，在组合高维的抽象特征前，先组合局部的时间运动信息；4)局部池化(Local Pooling)，不组合全局运动信息，减少了时间信息的丢失；5)时域卷积结构(Time-Domain Convolution)，增加了时间卷积层，可以在最大池化层之前组合更短时间的局部时间信息。使用包含 LSTM 的 RNN，可以在输入较长时间的训练视频后，取得更好的训练结果。测试结果表明，在 Sports-1M 上，此方案相比同期最佳模型提升显著，取得了 73.1%的准确率；在 UCF101 上也有所提高，取得了 88.6%的准确率。

Donahue 等[40]在 LSTM 的基础上，提出了长期往复卷积网络(Long-term Recurrent Convolutional Network，LRCN)。LRCN 将卷积神经网络和 LSTM 组合用于视频识别、视频描述的工作中。2D 卷积只能按单帧处理数据，LSTM 可以将 2D 卷积神经网络处理后的单帧信息进行融合。该网络将视频在每一帧传递给卷积神经网络，并使用卷积神经网络的输出作为 LSTM 的输入，将 LSTM 的输出作为最终网络输出。卷积神经网络与 LSTM 的参数沿着时间共享。Carreira 等[41]也认为使用 LSTM 进行融合一般可以取得更好的效果。

为了有效利用时空信息，Simonyan 等[42]提出了一种包含空间和时间网络的双流卷积网络结构，通过将光流信息作为输入，提取视频的时序上的运动信息，可以在多帧密集光流上训练并取得较好的性能。通过提取到的时序信息，可以帮助模型更好地理解视频数据。双流网络需要两个独立分离的主干网络，其结构如图 3 所示。相比于单主干网络，其计算量增多，导致训练速度变慢，但在 UCF101 数据集上表现良好，可以达到 88.0%的准确率。

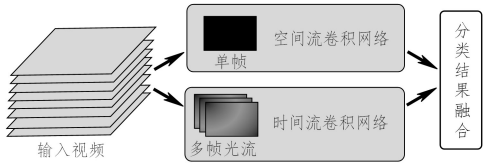


图 3 双流卷积网络结构

Fig. 3 Structure of two-stream convolutional network

Feichtenhofer 等[43]提出的双流网络融合策略，可进一步提高双流网络的效果，时空卷积网络均采用 VGG16[44]作为主干网络时，在 UCF101 上准确率最高达到 92.5%。若与 iDT 相结合，则可进一步提升至 93.6%。表 1 对比了本节中提到的方法在 UCF101 上的准确率。

表 1 基于 2D CNN 的网络模型在 UCF101 上的对比

Table 1 Comparisons between 2D CNN-based models on UCF101 dataset

Method	Average Accuracy/ %
LRCN[40]	82.9
LSTM[38]	84.3
Two-Stream Network[42]	88.0
LSTM[38] Trained on Long Videos	88.6
Two-Stream Network Fusion[43]	92.5
Two-Stream Network Fusion with iDT[43]	93.6

3.2 基于 3D 卷积神经网络的视频识别框架

2D 卷积神经网络虽然速度快，但在一些较大的视频数据集如 Kinetic-400 上却表现不佳。Lin 等[45]认为这是因为 2D 卷积神经网络不能很好地捕获时序上的信息。为了解决这个问题，Ji 等[46]提出了一种 3D 卷积神经网络(3D CNN)模型，并将其用于动作识别。该 3D 卷积神经网络通过从时空两个维度中提取特征，捕获多个相邻帧中的运动信息。

Tran 等[27]提出了卷积的 3D 网络(Convolutional 3D，C3D)，用于在较大尺度上进行视频分类和视频识别。通过使用卷积、池化、分类 3 个流程，可以取得较高的识别率。3D 卷积的理论对后续的网络发展也产生了很大的影响。

Carreira 等[41]提出了基于双流网络的 3D 卷积模型(Two-Stream Inflated 3D ConvNet，I3D)，同时也提出了大规模的视频数据集 Kinetics。通过该数据集对网络进行预训练，可以使 3D 卷积模型在不同的数据集上取得更好的效果。为了有效地捕捉信息，使用双流网络效果更好。相比于双流网络，LSTM 方法在处理时时序信息有损失，捕捉信息的能力并不十分理想，且在反向传播步骤训练时间也十分长。通过 Kinetics 数据集进行预训练，再在 HMDB51 和 UCF101 数据集上进行训练之后，I3D 可以在 HMDB51 上取得最高 66.4%的准确率，在 UCF101 上取得最高 93.4%的准确率。

基于视频中通常存在变化很慢或不变的背景，以及发生变化的动态区域的事实，Feichtenhofer 等[47]提出了快慢网络(SlowFast)，将空间信息和时序信息分开处理。作者对同一个视频片段应用两个平行的卷积神经网络，其中一个是慢速

高分辨率卷积神经网络(Slow 通道)，用于分析视频中的静态内容；另一个是快速低分辨率卷积神经网络(Fast 通道)，用于分辨视频中的动态内容。它们都使用 3D 残差网络模型，在捕获若干帧之后立即运行 3D 卷积操作。来自 Fast 通道的数据通过侧向连接被送入 Slow 通道，并被 Slow 通道用作获取 Fast 通道处理结果的信道。SlowFast 有多种不同的配置，将慢路径采样的帧数表示为 T，原始片段长度为 $T\times\tau$ 帧，可以有 SlowFast 4×16，8×8，16×8 等配置。相比于 C3D 网络，SlowFast 分开处理空间信息与时序信息，且对动作幅度、速度更快的动作类别识别效果提升显著，测试结果如表 2 所列。

表 2　基于 3D CNN 的网络模型在 Kinetics-400 上的对比

Table 2　Comparisons between 3D CNN-based models on Kinetics-400 dataset

Method	Backbone	Average Accuracy/%
C3D[27]	LSTM+RGB	53.9
C3D[27,46]	3DConv+RGB	56.1
C3D[27]	Two-stream Network, RGB with Optical Flow	62.8
I3D[41]	—	71.1
X3D[48]	—	79.1
SlowFast 16×8[47]	ResNet-101	79.8

3D 卷积神经网络的网络参数量多，导致其计算开销大，训练的时间更长，且推理时的延迟明显高于 2D 卷积神经网络。Feichtenhofer 等[48]也提出了扩展 3D 网络(eXpand 3D，X3D)。X3D 家族包含 XS，S，M，L，XL，XXL 等不同尺寸的网络，网络参数、计算量、识别准确率都随着尺寸增大而提高。X3D 网络解决了 3D 卷积神经网络参数量多的问题；它使用了一种逐步扩展网络方法，可以沿着多个网络轴，在时间、帧率、空间、宽度、瓶颈宽度和深度上逐步扩展并对小规模的 2D 图像进行分类。

X3D 取得了与之前工作类似的精度，但大幅减少了网络训练时的运算量。相对于基于 2D 卷积神经网络的视频识别框架，X3D 在 Kinetics 之类的大规模情景数据集的识别任务上往往会取得更好的效果。在 Kinetics-400 上，X3D 可以得到最高 79.1%的 Top-1 准确率和 93.9%的 Top-5 准确率，这与同时期表现最好的 SlowFast 的 79.8%的 Top-1 准确率与 93.9%的 Top-5 准确率不相上下。但 X3D 网络不能很好地学习到时序上的信息变化，因此在类似 Something-Something 这类对时序信息比较敏感的视频数据集上，并不能取得非常好的结果。Lee 等[49]的测试结果表明在 Something-Something-v1 上，X3D 网络的 Top-1 准确率为 48.4%，经过 Kinetics-400 预训练的网络准确率为 52.6%。表 2 列举了本节中各方法在 Kinetics-400 上的评估结果。

3.3　改进的 2D 卷积神经网络

为了解决 3D 卷积神经网络的训练速度问题，Du 等[50]提出一种伪 3D 卷积神经网络 $R(2+1)d$。这是一个用 2D 卷积神经网络模拟的伪 3D 神经网络，使用了混合型卷积(Mixed Convolution)的模型。该模型在浅层使用三维卷积，但在深层使用二维卷积进行连接。同时也提出了 2+1 维卷积模块，这个模块把三维的卷积操作分解成二维空间卷积与一维的时间卷积这两个连续的子卷积操作。将 3D 卷积核分解为单独的空间和时间分量，可以显著提升准确度。在残差学习的框架下，三维卷积神经网络相对于二维卷积神经网络仍具有准确性优势。其提出的 $R(2+1)d$ 网络可以在 Kinetics-400 和 HMDB51 等数据集上取得和当时最先进的 3D 卷积神经网络方法相同或更佳的效果。

Qiu 等[51]也提出了用二维模拟三维神经网络的伪 3D 残差网络(Pseudo-3D Residual Network，P3D)。作者将 2D 卷积扩展为 3D 卷积，并使用卷积核大小为(1，3，3)的空间卷积，以及卷积核大小为(3，1，1)的时间卷积来近似替代卷积核大小为(3，3，3)的 3D 卷积，实现的 P3D 网络可以利用 3D 结构来提取视频的空间时序信息。P3D 泛化能力强，识别准确率相较同期网络也有改善。其在 Sports-1M 上达到了比基于 3D 卷积网络方法高 5.3%的准确率，比基于帧的 2D 卷积网络方法高 1.8%的准确率。

Wang 等[52-56]也提出了一组基于轻量级时序建模方法的网络：时间段网络(Temporal Segment Network，TSN)、时间增强和交互网络(Temporal Enhancement and Interaction Network，TEINet)、时间激发和聚集网络(Temporal Excitation and Aggregation，TEA)、时间自适应网络(Temporal Adaptive Module，TAM)、时间差分网络(Temporal Difference Network，TDN)。

TSN[52]为视频中的动作设计了有效的卷积网络模型，并提出基于段的采样，通过聚合模块对远程时间结构进行建模。它使用整个视频中的动作有效地学习动作模型。TSN 在 HMDB51 和 UCF101 等多个数据集上取得了良好的效果，如表 3 所列。另外，TSN 对视频的处理效率高，可有效处理高达 340 帧每秒的高帧率视频。

TEINet[53]的提出也是为了解决 3D 卷积神经网络中参数量大导致计算时间长，以及 2D 卷积神经网络无法对时间进行卷积的问题。其提出了时间增强和交互(Temporal Enhancement and Interaction，TEI)模块。TEI 模块由动作增强模块(Motion Enchanced Module，MEM)与时间交互模块(Temporal Interaction Module，TIM)构成。MEM 模块利用注意力模型提取全局时序信息，增强了动作相关的特征。TIM 模块提取相邻帧的时序信息，增强了时序上下文信息。TEINet 在 Something-Something 数据集上可以减少计算量，并取得较好的结果。

TEA[54]网络由动作激发(Motion Excitation，ME)和多时间聚合(Multiple Temporal Aggregation，MTA)模块组成。ME 模块用于提取动态信息，通过残差连接来防止背景信息被抑制；同时对运动信道进行激励，以增强运动信息的捕获。MTA 模块则通过串联多个片段的信息，来捕获不同时间范围的时空表示。作者验证了这种方法的效果优于典型方法中使用单个局部卷积得到的局部时间表示的方案。

为了提升速度，TAM[55]只针对时间信息进行建模，而空间信息则由 2D 卷积神经网络学习。将 TAM 应用于 2D 卷积神经网络中时，只需要增加少许计算成本，就可得到一个性能优异的视频网络结构。在 Kinetics-400 上，TAM 的准确率可以达到 76.9%，比 TSM[45]高 2.2%；在 Something-Something-v2 数据集上准确率为 64.3%，只比 TSM 稍高 0.9%。

TDN 网络[56]显式地提取时序上运动的变化，并将其加入到网络中。其提出了长期时间差分模块(Long-term Temporal Difference Module，L-TDM)和短期时间差分模块(Short-term Temporal Difference Module，S-TDM)，通过差分操作提取时序信息中运动的变化，可以用较少的计算量进行

高效快速的时序建模。表 3 总结了本节中各网络在 Kinetics-400 数据集上，经不同配置训练得到的准确率。

表 3 改进的 2D CNN 网络模型在 Kinetics-400 上的对比

Table 3 Comparisons between optimized 2D CNN-based models on Kinetics-400 dataset

Method	Backbone	Pretrained on	Average Accuracy /%
$R(2+1)d$[50]	ResNet-34	Sports-1M	74.3
TSN[52]	Inception v3	ImageNet	72.5
TEINet[53]	ResNet-50	ImageNet	74.9
TEA[54]	ResNet-50	—	76.1
TAM[55]	ResNet-50	—	73.5
TDN[56]	ResNet-50	—	78.4
TDN[56]	ResNet-101	—	79.4
TSN[52]	Inception v3	ImageNet	72.5

3.4 基于 Transformer 的网络框架

传统的 3D 卷积神经网络需要在视频的所有时空位置上大量应用滤波器，因此计算成本较高。另外，3D 卷积滤波器虽然可以有效地捕获局部时空区域内较短时间段下的模式，但却无法对超出其接受域的时空依赖关系进行建模。Transformer 是另一种解决其实际问题的方案。

注意力(Attention)机制被广泛地运用在深度学习的各个领域中。在视觉图像处理方向上，注意力机制常被用于捕捉图像上的感受野。Devlin 等[57]提出的用于生成词向量的 BERT 算法使 NLP 任务获得了显著的效果提升。Vaswani 等[58]认为，BERT 算法中最重要的便是 Transformer 的概念。Transformer 架构由编码组件、解码组件和它们之间的连接构成，如图 4 所示。

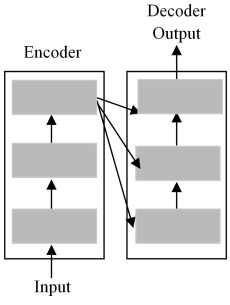


图 4 Transformer 结构的基本组成单位

Fig. 4 Basic unit of transformer

Transformer 结构不使用传统的 CNN 与 RNN，其网络全部由注意力机制构成，如图 5 所示。通过堆叠 Transformer 网络，也可以形成更为复杂的神经网络。Ruan 等[59]认为，纯 Transformer 架构在主要的视频识别基准上都已经达到了较高的准确度。

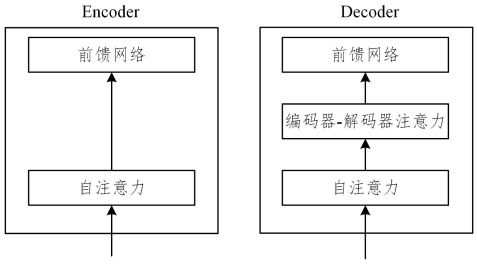


图 5 编码器和解码器内部一层的组成

Fig. 5 Composition of inner layer of encoder and decoder

Girdhar 等[60]提出了 Video Action Transformer Network，通过一种改进过的动作 Transformer(Action Transformer)结构，可以对视频帧中出现的人物动作进行检测和分类。基于动作的判断通常也和人周围的物体有关，基于改进的动作 Transformer 结构，也可以充分识别到关联的物体信息。其结合了 I3D 模型和区域生成网络(Reginal Proposal Network，RPN)模型来进行特征提取和采样。该模型的注意力机制主要关注手部与脸部的信息，作者认为这可以很好地区分动作。通过这些改进与先验知识，只使用原始 RGB 帧作为输入，在 AVA 数据集上训练并在同数据集上进行测试，其 mAP 可以达到最高 24.93%。

为了能更准确地识别视频特征，基于 CNN 的方法主要集中于对特征聚合方法进行优化[61]。一种简单的方法是通过全局平均池化来生成特征，但它无法捕捉到只在特定的帧或区域中出现的重要特征。为了捕捉此类特征，经常使用注意力模型[62]。Seong 等[63]提出一种由注意力模块和特征融合模块组成的网络模型，通过注意力机制从 CNN 的表示中捕获特征。该网络使用预先训练的 CNN 来编码视频中各帧的特征，并使用 Transformer 来对视频的整体特征进行建模。该模型对场景和动作特征的提取能力较强，但在特征重要性得分上表现略弱。

Bertasius 等[64]提出的完全基于 Transformer 的 TimeSformer 网络结构展现出纯 Transformer 架构在视频识别任务上的潜力。此网络结构使用了分割空间和时间注意力的方案。该方案将时间注意力和空间注意力应用到视频片段中，并依此得到视频的一组图像块，然后采用自注意力机制对这组图像块进行比较，在这个过程中，网络可以获取到整个视频的时空依赖关系，可以更好地理解视频内容。与 3D 卷积神经网络 I3D 和 SlowFast 相比，尽管 TimeSformer 的参数量是 I3D 的 4.3 倍(121.4 M 对比 28.0 M)，是 SlowFast 的 3.5 倍(121.4M 对比 34.6 M)。但在 Kinetics-400 上，要达到相近的识别准确度，TimeSformer 的训练速度大约是 I3D 的 3.5 倍(416 h 对比 1 440 h)，是 SlowFast 的 9.2 倍(416 h 对比 3 840 h)。虽然训练速度慢很多，但 TimeSformer 在计算速度上具有明显的优势。在实际应用中，此架构更适合那些视频长度长的片段，并且可以基于长的视频片段训练出更大的模型。

Arnab 等[65]提出了视频视觉变换器(Video Vision Transformer，ViViT)网络结构，使用纯 Transformer 方案实现了视频分类。它通过构建时空 Token，将一个视频变换为一组序列，作为其 Transformer 的输入。ViViT 包含两种方案，分别是统一帧采样(Uniform Frame Sampling)与 Tubelet 绑定(Tubelet Embedding)。统一帧采样是提取 2D 图像特征的过程，它对视频的各帧进行提取 Token 的操作，并将获取到的 Token 连接起来作为后续操作的输入。Tubelet 绑定方法用于提取立方体特征，此方法每隔一段时间提取特征，并取每一帧相同位置组成输入。

同时，文章还提出 4 种不同的 Transformer 结构。1)朴素结构，即将提取到的时空 Token 作为 Transformer 的输入，用普通的 Transformer 结构得到最终的结果。2)分解式编码器(Factorized Encoder)，使用两个 Transformer，第一个称为空间转换器(Spatial Transformer)，第二个是时间转换器(Temporal Transformer)。这两个 Transformer 对视频帧中获取到的 Token 进行编码处理，并通过 MLP 对输出结果进行分类。3)分解式自注意力(Factorized Self-attention)，通过自注意力层将时空数据分开进行处理，空间层在同一帧内

不同的 Token 之间进行注意力操作，时间层对不同帧的同一位置的 Token 进行注意力操作，其时间和空间的处理是串行进行的。4）分解式点积注意力（Factorized Dot-product Attention），与分解式自注意力类似，但分解式点积注意力中时间和空间的处理是并行进行的。ViViT 在 Kinetics，Sports-1M 和 Something-Something 数据集上都取得了领先的结果。

针对 Transformer 在被识别的视觉实体有较大变化的情况下性能下降的问题，以及高分辨率图像导致 Transformer 基于全局的自注意力的计算方法产生较大的计算量的问题，Liu 等[66]提出了滑动窗口变换器（Shifted Window Transformer，Swin Transformer）。这是一种包含滑动窗口操作且具有层级设计的 Transformer。

传统的 Transformer 基于全局来计算注意力，计算复杂度十分高，但 Swin Transformer 将注意力的计算限制在每个窗口内，降低了计算复杂度，减少了计算量。Swin 的算法设计保证了它可以在图像处理任务上取得线性的计算复杂度。因此，在实际的图像分类和密集的图像识别任务中，都可以选用 Swin Transformer 作为其网络的通用主干。基于 Swin Transformer 的网络也在视频分类、目标检测的任务上都取得了当前最佳效果。表 4 列出了本节中各网络在 Kinetics-400 上的测试结果。

表 4　基于 Transformer 的网络在 Kinetics-400 上的对比

Table 4　Comparison between transformer-based models on Kinetics-400 dataset

Method	Average Accuracy/%
TimeSformer-L[64]	80.7
ViViT-H/16×2[65]	84.8
Swin-L，Pretrained on ImageNet-21k[9,66]	84.9

4　不同架构模型的性能对比

基于 2D 卷积的视频识别模型通常需要较少的参数量，计算速度快，但在一些相似的场景或者非常依赖时序信息的视频任务上效果通常不佳。基于 3D 的卷积识别模型则考虑到时序信息，这引入了更多的参数量，也会减慢网络的训练速度和推理速度。

一些基于 Transformer 视频识别模型的准确度已经超越了基于卷积网络的视频识别模型，但在模型训练上，相比于 CNN，从头开始训练 Transformer 需要更多的数据。这是由于 CNN 的实现已经暗含了关于图像的一些先验知识，例如图像的平移同变性（translation equivariance）。但 Transformer 需要在大规模的训练数据集上进行耗时的预训练去学习到这些规则。

Kondratyuk 等[67]将各网络按照视频片段的推理分辨率（Resolution，RES）、视频片段数×每个视频片段的帧数（Frames）和每秒帧数（Frame Per Second，FPS）的维度进行比较，在 Kinetics-600 数据集上的性能对比如表 5 所列。

通过将 X3D 家族中不同尺寸的模型与基于 Transformer 的网络进行比较，可以得出下列结论：1）在 Kinetic-600 数据集上，基于 Transformer 的视频识别模型的准确率高于其他的基于卷积神经网络的模型，体现出其视频识别任务上的准确性；2）基于 Transformer 的视频识别模型通常具有较多的参数数量，对应地也拥有较多的计算量。

表 5　各网络模型在 Kinetics-600 数据集上的运行性能对比

Table 5　Performance comparison of different models on Kinetics-600 dataset

Method	Average Accuracy/ %	GFLOPs	RES	Frames	FPS	Params/M
X3D-XS[48]	72.3	23.3	182	30×4	2	3.8
X3D-S[48]	76.4	76.1	182	30×13	4	3.8
X3D-M[48]	78.8	186	256	30×16	5	3.8
I3D[41]	71.6	216	224	1×250	25	12
X3D-L[48]	80.5	744	356	30×16	5	6.1
ViViT-L/16×2[65]	83.0	3990	320	12×32	12	88.9
TimeSformer-HR[64]	82.4	5110	224	3×8	1.5	120
X3D-XL[48]	81.9	1452	356	10×16	5	11.0
SlowFast-R50[47]	78.8	1080	256	30×16	5	34.4
SlowFast-R101[47]	81.8	7020	256	30×16	5	59.9

Koot 等[68]和 Langerman 等[69]的研究表明，GFLOPs 与准确率有一定关联，但不足以直接体现延迟，对模型推理时延影响较大的是网络结构。

为取得网络的实时性结果，根据训练延迟、批量模式（Batch-Mode，BM）延迟、单实例（Single-Instance，SI）延迟对各网络进行测试[68]，以获取各典型网络在不同推理模式下的延迟情况。结果如表 6 所列。

表 6　各典型网络模型的训练和推理延迟的对比

Table 6　Comparison of model latency in training and inferencing

Method	Training Latency	BM Inference Latency	SI Inference Latency
TSM-R50[45]	2.64	4.41	2.36
X3D-XS[48]	1.64	2.25	1.72
I3D-BERT[41,67]	2.49	3.69	2.44
TimeSformer[64]	10.20	20.10	8.77
ViViT-FE[65]	2.95	5.54	2.83
Swin Transformer[66]	3.64	6.87	3.43

表 6 的结果显示，无论是训练过程还是推理过程，非 Transformer 结构的网络通常都具有较低的延迟。如表 5、表 6 的结果所示，虽然 X3D-XS 具有最低的延迟，但相对地，其准确度也较低。

同时，该结果也表明单纯基于注意力的模型通常会有较高的延迟。TimeSformer 的延迟是所有网络中最高的，而其他基于 Transformer 的架构中，使用分解式编码器的 ViViT-FE 在单实例推理的延迟上表现较好，但批量推理时延迟仍然高于非 Transformer 的网络。

虽然用于视频识别任务的 Transformer 准确性高，但其实时性仍然与基于卷积神经网络的模型存在较大的差距。所以，对基于 Transformer 的网络模型进行实时性的优化也是未来可以进行深入研究的方向之一。

结束语　本文对传统的视频识别框架以及基于深度学习的视频识别框架的方法进行了梳理和总结。传统的视频识别方法非常依赖算法的设计和特征的选择，而且在使用过程中需要人工进行特征设计。而基于深度学习的视频和图像识别框架则可以自动地在有标签的视频数据集上进行训练和学习，相较传统算法，可以取得比较好的识别结果。

DT 和 iDT 方法是传统方法中表现最好的算法，与深度神经网络结合后通常也可以提高网络的性能。

基于 3D 卷积神经网络通常可以取得比基于 2D 的卷积神经网络更好的视频识别结果。但其存在参数量大、训练时间长的问题。因此，一些 2D 卷积神经网络通过模拟 3D 卷积

神经网络的实现，减少了参数量，使其运算速度更快，并且可取得与3D卷积神经网络相近的识别准确度。Transformer于近年被提出，并迅速成为视频识别领域中的一个热门话题。基于Transformer的网络框架灵活运用注意力机制来捕获视频的整体特征，可以高效准确地进行识别任务。

随着研究的深入，相信将来也会有更多的基于深度学习的视频识别和图像识别框架被提出，并有望进一步刷新目前的成绩。但是也可以注意到，现实世界的视频分辨率正逐渐变高，同时各种现实任务对视频识别算法的实时性要求也在提高。另外，一些智能汽车、IoT智能设备也可以使用模型进行推理，但这些设备往往有有限的资源，对视频识别算法的准确率和资源消耗都提出了较为严格的要求。

另外，如何提高跨域的视频识别准确率也是一个具有挑战性的课题。训练视频识别模型时，常常会假设数据全部来自同一个域。虽然目前许多模型结构在数据集上表现良好，但当模型在真实世界中使用时，天气原因、光照变化、拍摄硬件限制、角度、被拍摄对象的动作形变等客观条件，会使输入数据的特征分布发生显著变化，使得模型在开放域上的识别准确度下降。这些问题的解决都有待本领域未来的进一步研究和新的技术方案作为支撑。

参 考 文 献

[1] DALAL N, TRIGGS B. Histograms of oriented gradients for human detection[C]// Institute of Electrical and Electronics Engineering Conference on Computer Vision and Pattern Recognition. San Diego, USA, 2005: 886-893.

[2] CHAUDHRY R, RAVICHANDRAN A, HAGER G, et al. Histograms of oriented optical flow and binet-cauchy kernels on nonlinear dynamical systems for the recognition of human actions[C]// Institute of Electrical and Electronics Engineering Conference on Computer Vision and Pattern Recognition. Miami, USA, 2009: 1932-1939.

[3] WANG H, KLASER A, SCHMID C, et al. Dense Trajectories and Motion Boundary Descriptors for Action Recognition[J]. International Journal of Computer Vision, 2013, 103(1): 61-79.

[4] LAZEBNIK S, SCHMID C, PONCE J. Beyond Bags of Features: Spatial Pyramid Matching for Recognizing Natural Scene Categories[C]// Electrical and Electronics Engineering Computer Society Conference on Computer Vision and Pattern Recognition. New York, USA, 2006: 2169-2178.

[5] YANG M, ZHANG L, FENG X, et al. Sparse representation based fisher discrimination dictionary learning for image classification[J]. International Journal of Computer Vision, 2014, 109(3): 209-232.

[6] HINTON G E. Learning multiple layersof representation[J]. Trends in Cognitive Sciences, 2007, 11(10): 428-434.

[7] DENG L, YU D. Deep learning: methods and applications[J]. Foundations and Trendsr in Signal Processing, 2014, 7(3/4): 197-387.

[8] SCHMIDHUBER J. Deep learning in neural networks: an overview[J]. Neural Networks, 2015, 61(1): 85-117.

[9] KRIZHEVSKY A, SUTSKEVER I, HINTON G E. Imagenet classification with deep convolutional neural networks[J]. Advances in Neural Information Processing Systems, 2012, 25(1): 1097-1105.

[10] KARPATHY A, TODERICI G, SHETTY S, et al. Large-Scale Video Classification with Convolutional Neural Networks[C]// Institute of Electrical and Electronics Engineering Conference on Computer Vision and Pattern Recognition. Columbus, USA, 2014: 1725-1732.

[11] MATERZYNSKA J, XIAO T, HERZIG R, et al. Something-Else: Compositional Action Recognition With Spatial-Temporal Interaction Networks[C]// Institute of Electrical and Electronics Engineering/Computer Vision Foundation Conference on Computer Vision and Pattern Recognition. Virtual, 2020: 1049-1059.

[12] SOOMRO K, ZAMIR A R, SHAH M. UCF101: A Dataset of 101 Human Actions Classes From Videos in The Wild[J]. arXiv: 1212.0402, 2012.

[13] KAY W, CARREIRA J, SIMONYAN K, et al. The Kinetics Human Action Video Dataset[J]. arXiv: 1705.06950, 2017.

[14] GU C, CHEN S, DAVID A R, et al. Ava: A video dataset of spatio-temporally localized atomic visual actions[C]// Institute of Electrical and Electronics Engineering Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA, 2018: 6047-6056.

[15] LI A, THOTAKURI M, ROSS D A, et al. The AVA-Kinetics Localized Human Actions Video Dataset[J]. arXiv: 2005.00214, 2020.

[16] KUEHNE H, JHUANG H, GARROTE E, et al. HMDB: A Large Video Database for Human Motion Recognition[C]// Institute of Electrical and Electronics Engineering International Conference on Computer Vision. Barcelona, Spain, 2011: 2556-2563.

[17] SIGURDSSON G A, VAROL G, WANG X, et al. Hollywood in homes: Crowdsourcing data collection for activity understanding[C]// European Conference on Computer Vision. Cham: Switzerland, 2016: 510-526.

[18] GUNNAR A S, ABHINAV G, CORDELIA S, et al. Charades-ego: A large-scale dataset of paired third and first person videos[J]. arXiv: 1804.09626, 2018.

[19] DAMEN D, DOUGHTY H, FARINELLA G M, et al. Rescaling Egocentric Vision: Collection, Pipeline and Challenges for EPIC-KITCHENS-100[J]. International Journal of Computer Vision, 2022, 130(1): 33-55.

[20] DAMEN D, DOUGHTY H, FARINELLA G M, et al. Scaling egocentric vision: The epic-kitchens dataset[C]// The European Conference on Computer Vision. Munich, Germany, 2018: 720-736.

[21] DAMEN D, DOUGHTY H, FARINELLA G M, et al. The epic-kitchens dataset: Collection, challenges and baselines[J]. Institute of Electrical and Electronics Engineering Transactions on Pattern Analysis & Machine Intelligence, 2020(1): 1-1.

[22] ABU-EL-HAIJA S, KOTHARI N, LEE J, et al. Youtube-8m: A large-scale video classification benchmark[J]. arXiv: 1609.08675, 2016.

[23] ANTIPOV G, BERRANI S A, RUCHAUD N, et al. Learned vs. hand-crafted features for pedestrian gender recognition[C]// 23rd Association for Computing Machinery International Conference on Multimedia. New York, USA, 2015: 1263-1266.

[24] KLASER A, MARSZALEK M, SCHMID C. A spatio-temporal

descriptor based on 3D-gradients[C]// 19th British Machine Vision Conference. Leeds, British, 2008: 1-10.

[25] WANG H, KLASER A, SCHMID C, et al. Action recognition by dense trajectories[C]// 2011 Institute of Electrical and Electronics Engineering Conference on Computer Vision and Pattern Recognition. Colorado Springs, USA, 2011: 316903176.

[26] WANG H, SCHMID C. Action recognition with improved trajectories[C]// 2013 Institute of Electrical and Electronics Engineering International Conference on Computer Vision. Sydney, Australia, 2013: 3551-3558.

[27] TRAN D, BOURDEV L, Fergus R, et al. Learning Spatiotemporal Features with 3D Convolutional Networks[C]// Institute of Electrical and Electronics Engineering International Conference on Computer Vision. Santiago, Chile, 2015: 4489-4497.

[28] HUANG K, DELANY S J, MCKEEVER S. Human Action Recognition in Videos Using Transfer Learning[C]// Irish Machine Vision and Image Processing Conference. Dublin, Ireland, 2019.

[29] ZHANG Z, SEJDIC E. Radiological images and machine learning: trends, perspectives, and prospects[J]. Computersin biology and medicine, 2019, 108(1): 354-370.

[30] HINTON G E. Deep belief networks [J]. Scholarpedia, 2009, 4(5): 5947.

[31] TAYLOR G W, HINTON G E. Factored conditional restricted-Boltzmann machines for modeling motion style[C]// The 26th Annual International Conference on Machine Learning. New York, USA, 2009: 1025-1032.

[32] LAROCHELLE H, BENGIO Y. Classification using discriminative restricted Boltzmann machines[C]// The 25th International Conference on Machine Learning. New York, USA, 2008: 536-543.

[33] CHEN B. Deep learning of invariant spatio-temporal features from video[D]. British Columbia: University of British Columbia, 2010.

[34] YANG T A, SILVER D L. The Disadvantage of CNN versus DBN Image Classification Under Adversarial Conditions[C] // The 34th Canadian Conference on Artificial Intelligence. Vancouver, Canada, 2021.

[35] CHEN M, RADFORD A, CHILD R, et al. Generative pretraining from pixels [C] // International Conference on Machine Learning. Virtual, 2020: 1691-1703.

[36] SOCHER R, HUVAL B, BATH B, et al. Convolutional-recursive deep learning for 3d object classification[J]. Advances in Neural Information Processing Systems, 2012, 25(1): 656-664.

[37] VIJAYANARASIMHAN S, SHLENS J, MONGA R, et al. Deep networks with large output spaces[J]. arXiv: 1412. 7479, 2014.

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

[39] NG Y H, HAUSKNECHT M, VIJAYANARASIMHAN S, et al. Beyond short snippets: Deep networks for video classification[C] // Institute of Electrical and Electronics Engineering Conference on Computer Vision and Pattern Recognition. Boston, USA, 2015: 4694-4702.

[40] DONAHUE J, HENDRICKS L A, ROHRBACH M, et al. Long-term recurrent convolutional networks for visual recognition and description[J]. Institute of Electrical and Electronics Engineering Transactions on Pattern Analysis & Machine Intelligence, 2017, 39(4): 677-691.

[41] CARREIRA J, ZISSERMAN A. Quo vadis, action recognition? a new model and the kinetics dataset[C]// Institute of Electrical and Electronics Engineering Conference on Computer Vision and Pattern Recognition. Honolulu, USA, 2017: 6299-6308.

[42] SIMONYAN K, ZISSERMAN A. Two-stream convolutional networks for action recognition in videos[J]. Advances in Neural Information Processing Systems, 2014, 27: 568-576.

[43] FEICHTENHOFER C, PINZ A, ZISSERMAN A. Convolutional two-stream network fusion for video action recognition[C]// Institute of Electrical and Electronics Engineering conference on Computer Vision and Pattern Recognition. Las Vegas, USA, 2016: 1933-1941.

[44] SIMONYAN K, ZISSERMAN A. Very deepconvolutional networks for large-scale image recognition[J]. arXiv: 1409. 1556, 2014.

[45] LIN J, GAN C, HAN S. Tsm: Temporal shift module for efficient video understanding [C] // Institute of Electrical and Electronics Engineering/Computer Vision Foundation International Conference on Computer Vision. Seoul, Korea, 2019: 7083-7093.

[46] JI S, XU W, YANG M, et al. 3D Convolutional Neural Networks for Human Action Recognition[J]. Institute of Electrical and Electronics Engineering Transactions on Pattern Analysis & Machine Intelligence, 2013, 35(1): 221-231.

[47] FEICHTENHOFER C, FAN H, MALIK J, et al. Slowfast networks for video recognition [C] // Institute of Electrical and Electronics Engineering International Conference on Computer Vision. Seoul, South Korea, 2019: 6202-6211.

[48] FEICHTENHOFER C. X3D: Expanding Architectures for Efficient Video Recognition [C] // Institute of Electrical and Electronics Engineering International Conference on Computer Vision. Virtual, 2020: 203-213.

[49] LEE Y, KIM H I, YUN K, et al. Diverse temporal aggregation and depthwise spatiotemporal factorization for efficient video classification[J]. arXiv: 2012. 00317, 2020.

[50] DU T, WANG H, TORRESANI L, et al. A Closer Look at Spatiotemporal Convolutions for Action Recognition[C]// Institute of Electrical and Electronics Engineering/Computer Vision Foundation Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA, 2018.

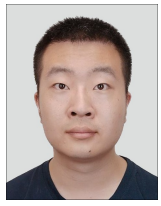
[51] QIU Z, YAO T, MEI T. Learning spatio-temporal representation with pseudo-3D residual networks[C]// Institute of Electrical and Electronics Engineering International Conference on Computer Vision. Venice, Italy, 2017: 5534-5542.

[52] WANG L, XIONG Y, WANG Z, et al. Temporal segment networks: Towards good practices for deep action recognition[C]// European Conference on Computer Vision. Amsterdam, Netherlands, 2016: 20-36.

[53] LIU Z, LUO D, WANG Y, et al. TEINet: Towards an Efficient Architecture for Video Recognition [C] // Association for the Advance of Artificial Intelligence Conference on Artificial Intelligence. New York, USA, 2020: 11669-11676.

[54] LI Y, JI B, SHI X, et al. TEA: Temporal Excitation and Aggregation for Action Recognition [C] // Institute of Electrical and Electronics Engineering/Computer Vision Foundation Confe-

rence on Computer Vision and Pattern Recognition. Virtual, 2020:909-918.
[55] LIU Z, WANG L, WU W, et al. TAM: Temporal adaptive module for video recognition[C]// Institute of Electrical and Electronics Engineering/Computer Vision Foundation International Conference on Computer Vision. Virtual, 2021:13708-13718.
[56] WANG L, TONG Z, JI B, et al. TDN: Temporal Difference Networks for Efficient Action Recognition[C]// Institute of Electrical and Electronics Engineering/Computer Vision Foundation Conference on Computer Vision and Pattern Recognition. Virtual, 2021:1895-1904.
[57] DEVLIN J, CHANG M W, LEE K, et al. BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding [C]// Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1(Long and Short Papers). Minneapolis, USA, 2018:4171-4186.
[58] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is All You Need[J]. arXiv:1706.03762, 2017.
[59] RUAN L, QIN J. Survey: Transformer Based Video-Language Pre-Training[J]. arXiv:2109.09920, 2021.
[60] GIRDHAR R, CARREIRA J, DOERSCH C, et al. Video action transformer network[C]// Institute of Electrical and Electronics Engineering/Computer Vision Foundation Conference on Computer Vision and Pattern Recognition. Long Beach, USA, 2019: 244-253.
[61] HARA K, KATAOKA H, SATOH Y. Can spatiotemporal 3d cnns retrace the history of 2d cnns and imagenet? [C]// Institute of Electrical and Electronics Engineering/Computer Vision Foundation Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA, 2018:6546-6555.
[62] PARK J, JEON S, KIM S, et al. Learning to detect, associate, and recognize human actions and surrounding scenes in untrimmed videos[C]// The 1st Workshop and Challenge on Comprehensive Video Understanding in the Wild. Seoul, Korea, 2018:21-26.
[63] SEONG H, HYUN J, KIM E. Video multitask transformer network[C]// Institute of Electrical and Electronics Engineering/Computer Vision Foundation International Conference on Computer Vision Workshops. Seoul, Korea, 2019.
[64] BERTASIUS G, WANG H, TORRESANI L. Is Space-Time Attention All You Need for Video Understanding[J]. arXiv:2102.05095, 2021.
[65] ARNAB A, DEHGHANI M, HEIGOLD G, et al. ViViT: A Video Vision Transformer[C]// Institute of Electrical and Electronics Engineering/Computer Vision Foundation International Conference on Computer Vision. Virtual, 2021:6836-6846.
[66] LIU Z, LIN Y, CAO Y, et al. Swin Transformer: Hierarchical Vision Transformer Using Shifted Windows[C]// Institute of Electrical and Electronics Engineering/Computer Vision Foundation International Conference on Computer Vision. Virtual, 2021:10012-10022.
[67] KONDRATYUK D, YUAN L, LI Y, et al. Movinets: Mobile video networks for efficient video recognition[C]// Institute of Electrical and Electronics Engineering/Computer Vision Foundation Conference on Computer Vision and Pattern Recognition. Virtual, 2021:16020-16030.
[68] KOOT R, HENNERBICHLER M, LU H. Evaluating Transformers for Lightweight Action Recognition[J]. arXiv: 2111.09641, 2021.
[69] LANGERMAN D, JOHNSONA, BUETTNER K, et al. Beyond Floating-Point Ops: CNN Performance Prediction with Critical Datapath Length[C]// Institute of Electrical and Electronics Engineering High Performance Extreme Computing Conference. Virtual, 2020:1-9.



QIAN Wen-xiang, born in 1992, postgraduate. His main research interests include human body recognition in natural scenes and so on.

YI Yang, born in 1967, Ph.D, associate professor. Her main research interests include human body recognition in natural scenes and so on.