Video Action Recognition任务是理解video中人的行为,其中涉及识别,定位,和预测人的行为的任务.具有很多真实场景下的应用:behavior analysis,video retrieval,human-robot interaction,gaming,and entertainment.该领域的发展也是近几年被大力关注,数据集从11年到15年公布的3个,16年到20年就13个了.就具体的发展趋势而言,有 1.双流网络:光流+RGB流.其中RGB流就是连续帧的图像,用来捕获其中的位置,特征等信息;而光流反应的是连续运动帧中运动物体的运动趋势,是行为识别中的运动信息. 这类方法就是结合RGB特征和运动特征的双流网络来进行行为识别.(光流需要预先计算,需要大量的预置存储空间,因此本质上不太实用;另外光流是分析运动情况的,那么如果针对摄像头也是运动的视频场景不知道是不是也有相对改进的方法) 2.3D卷积:既然需要结合图像特征与运动信息,那不如就直接塞到3D卷积黑盒里面去,给个监督标签自己让它学,虽然道理是这么个道理,但是不用细看就知道,肯定有很多需要额外需要考虑的内容,不过目前没有看很多这方面的文章.最近一些研究都在将3D卷积进行优化,就是嫌弃它的计算量过大. 3.围绕计算效率上展开, 使之在更大size的数据集上训练,适用于真实世界场景.

A Comprehensive Study of Deep Video Action Recognition

Yi Zhu, Xinyu Li, Chunhui Liu, Mohammadreza Zolfaghari, Yuanjun Xiong,
Chongruo Wu, Zhi Zhang, Joseph Tighe, R. Manmatha, Mu Li
Amazon Web Services
{yzaws,xxnl,chunhliu,mozolf,yuanjx,chongrwu,zhiz,tighej,manmatha,mli}@amazon.com

关于数据集部分,如今也开源了不少,具体比较常用是Kinetics和something系列.构建数据集的步骤如下(可能会对后面训练理解标签含义有帮助): 1.定义action list, 此部分通过组合之前数据集中的定义, 以及添加新的action类别定义得到. 2.通过匹配网上各种小视频的title和subtitle来爬取相关action list中定义的类别. 3.手动标注一个具体action类别的起始与结束时间, 用以后面的训练. 4.数据清洗, 删除重复数据以及无关噪音数据

Abstract

Video action recognition is one of the representative tasks for video understanding. Over the last decade, we have witnessed great advancements in video action recognition thanks to the emergence of deep learning. But we also encountered new challenges, including modeling long-range temporal information in videos, high computation costs, and incomparable results due to datasets and evaluation protocol variances. In this paper, we provide a comprehensive survey of over 200 existing papers on deep learning for video action recognition. We first introduce the 17 video action recognition datasets that influenced the design of models. Then we present video action recognition models in chronological order: starting with early attempts at adapting deep learning, then to the two-stream networks, followed by the adoption of 3D convolutional kernels, and finally to the recent compute-efficient models. In addition, we benchmark popular methods on several representative datasets and release code for reproducibility. In the end, we discuss open problems and shed light on opportunities for video action recognition to facilitate new research ideas.

视频动作识别时视频理解任务最具有代表性的任务之一.再过去的十年中,由于深度学习技术的发展,视频动作识别任务也取得了巨大的进步.然而我们还是遇到了非常多的挑战包括视频中存在的长时信息,计算需要的高计算量,由于数据集和衡量带来的第可比性.因此,本文对超过200个现存的在视频识别领域使用深度学习的文章进行了总结.目前的这些方法可以分为:双流网络,3D卷积网络,以及最近的一些注重计算效率的模型.我们对一些具有代表新的数据集上同一测试了一些流行的方法.最后,我们讨论了一些开放的问题,并且对视频动作分类任务留下了可能

1. Introduction

One of the most important tasks in video understanding is to understand human actions. It has many real-world applications, including behavior analysis, video retrieval, human-robot interaction, gaming, and entertainment. Human action understanding involves recognizing, localizing, and predicting human behaviors. The task to recognize human actions in a video is called *video action recognition*. In Figure 1, we visualize several video frames with the associated action labels, which are typical human daily activities such as shaking hands and riding a bike.

视频理解任务中最重要的一个任务就是理解人类的动作.理解人类的动作有非常多的运用,包括行为分析,视频检索,动作定位,动作预测.在这么多的视频理解任务中,视频中识别人类的动作的任务被称为视频动作识

Over the last decade, there has been growing research interest in video action recognition with the emergence of high-quality large-scale action recognition datasets. We summarize the statistics of popular action recognition

再过去的十年中,我们对视频动作分类问题的研究导致出现了很多高质量的大规模数据集.我们在图2中可视化出来了

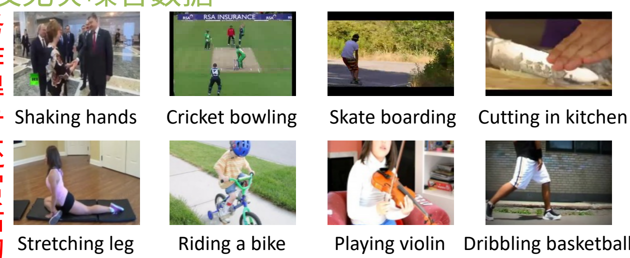
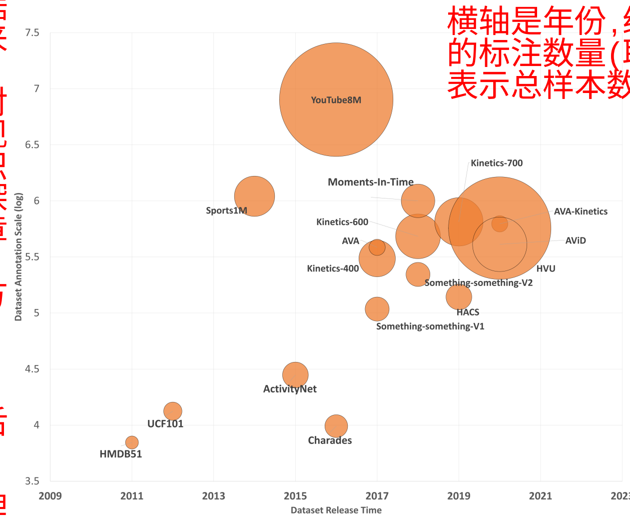


Figure 1. **Visual examples of categories in popular video action datasets.**



横轴是年份,纵轴是数据集的标注数量(取对数),大小表示总样本数量

Figure 2. **Statistics of most popular video action recognition datasets from past 10 years.** The area of an circle represents the scale of each dataset (i.e., number of videos).

过去的十年中,我们不仅见证了数据集规模(视频总数/类别数)的增长,还见证了更多的数据集的开发

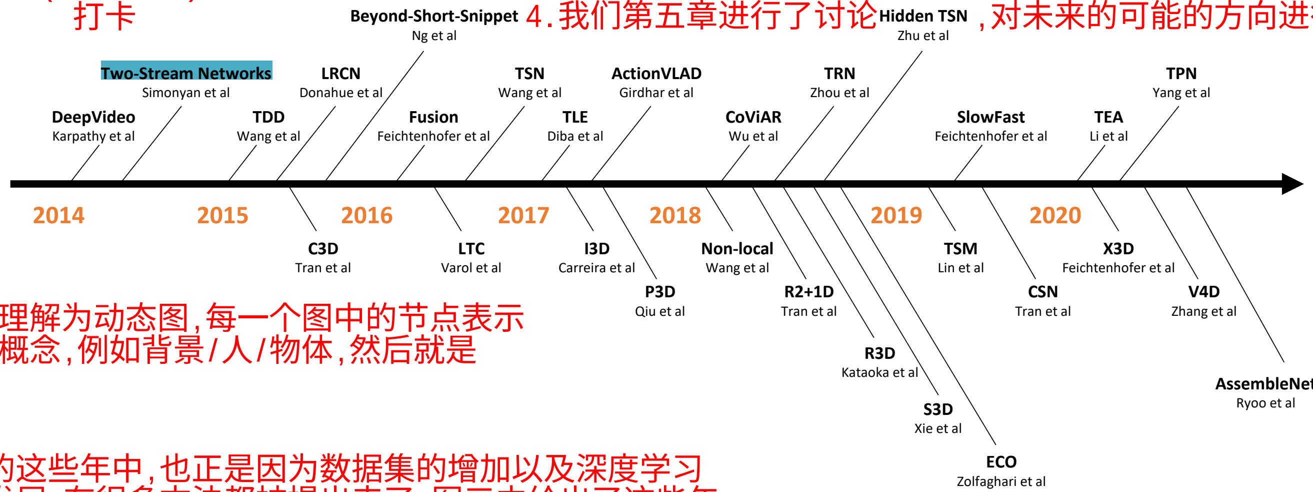
datasets in Figure 2. We see that both the number of videos and classes increase rapidly, e.g, from 7K videos over 51 classes in HMDB51 [109] to 8M videos over 3,862 classes in YouTube8M [1]. Also, the rate at which new datasets are released is increasing: 3 datasets were released from 2011 to 2015 compared to 13 released from 2016 to 2020.

Thanks to both the availability of large-scale datasets and the rapid progress in deep learning, there is also a

模型学到空间知识和时间知识.
那么用模型学到的知识-空间知识=时间知识

一张图像里是有不同的概念的.第一个网络抽取概念(图),第二个网络根据概念间的变换进行分类(图的变换)
打卡

本文剩下的部分将按照如下的方式进行组织:
1.我们在第2章中介绍了用于benchmark的一些常见的数据集以及一些常见的挑战
2.我们在第3章中详细介绍了最近使用深度学习技术来进行视频识别的进展
3.我们在第4章中对现有的广泛流行的方法进行了测试
4.我们第五章进行了讨论,对未来的可能的方向进行了介绍



把视频理解为动态图,每一个图中的节点表示不同的概念,例如背景/人/物体,然后就是

再过去的这些年中,也正是因为数据集的增加以及深度学习技术的发展,有很多方法都被提出来了,图三中给出了这些年具有代表性的方法的一个编年表

Figure 3. **A chronological overview of recent representative work in video action recognition**.

14年的DeepVideo是最早尝试使用卷积网络来处理视频任务的.在这之后的文章大体上可以分为三类:1.双流网络 2.3D卷积核 3.计算高效性

rapid growth in deep learning based models to recognize video actions. In Figure 3, we present a chronological overview of recent representative work. DeepVideo [99] is one of the earliest attempts to apply convolutional neural networks to videos. We observed three trends here. The first trend started by the seminal paper on Two-Stream Networks [187], adds a second path to learn the temporal information in a video by training a convolutional neural network on the optical flow stream. Its great success inspired a large number of follow-up papers, such as TDD [214], LRCN [37], Fusion [50], TSN [218], etc. The second trend was the use of 3D convolutional kernels to model video temporal information, such as I3D [14], R3D [74], S3D [239], Non-local [219], SlowFast [45], etc. Finally, the third trend focused on computational efficiency to scale to even larger datasets so that they could be adopted in real applications. Examples include Hidden TSN [278], TSM [128], X3D [44], TVN [161], etc.

尽管有很多视频识别的网络被提出来了目前仍然没有一个综合性的对这些方法进行介绍的文章.早期的一些视频相关的文章要么是关注于手动设计的特征,要么是关注于视频字母,视频预测,动作检测这类其他的任务

Despite the large number of deep learning based models for video action recognition, there is no comprehensive survey dedicated to these models. Previous survey papers either put more efforts into hand-crafted features [77, 173] or focus on broader topics such as video captioning [236], video prediction [104], video action detection [261] and zero-shot video action recognition [96]. In this paper:

- We comprehensively review over 200 papers on deep learning for video action recognition. We walk the readers through the recent advancements chronologically and systematically, with popular papers explained in detail.
- We benchmark widely adopted methods on the same set of datasets in terms of both accuracy and efficiency. We also release our implementations for full reproducibility[1].
- We elaborate on challenges, open problems, and opportunities in this field to facilitate future research.

最后,本文的贡献如下: 1.我们超过200篇文章进行了总结,我们以编年表的顺序系统而又详细的介绍了近年来深度学习技术在视频识别任务上的进展.对于一些流行的方法进行了详细的介绍. 2.我们对一些流行的方法在同一的数据集下进行了测试,以比较不同方法的准确度和效率 3.我们对现存的一些问题进行了讨论

The rest of the survey is organized as following. We first describe popular datasets used for benchmarking and existing challenges in section 2. Then we present recent advancements using deep learning for video action recognition in section 3, which is the major contribution of this survey. In section 4, we evaluate widely adopted approaches on standard benchmark datasets, and provide discussions and future research opportunities in section 5.

深度学习方法在有大量数据的时候,性能会有极大的提升而对于视频动作理解来说,这就意味着我们需要大量的标注的数据集来学习到有用的模型

2. Datasets and Challenges

2.1. Datasets

Deep learning methods usually improve in accuracy when the volume of the training data grows. In the case of video action recognition, this means we need large-scale annotated datasets to learn effective models.

For the task of video action recognition, datasets are often built by the following process: (1) Define an action list, by combining labels from previous action recognition datasets and adding new categories depending on the use case. (2) Obtain videos from various sources, such as YouTube and movies, by matching the video title/subtitle to the action list. (3) Provide temporal annotations manually to indicate the start and end position of the action, and (4) finally clean up the dataset by de-duplication and filtering

[1]Model zoo in both PyTorch and MXNet: `https://cv.gluon.ai/model_zoo/action_recognition.html`

构建视频的数据集包含下面几步:1.定义动作列表,具体可以根据先前的数据集来增加新的动作 2.从多种渠道获取视频,包括从youtube的视频中进行配对 3.提供时间标注信息来说明动作什么时候开始什么时候结束 4.最后通过去除重复和滤去噪音来清洗数据集

HMDB51数据集在2011年公开.里面的动作主要是从电影,还有一小部分是从Prelinger归档,YouTube和谷歌视频中获取的.数据集一共包含6849个视频,包含51种动作.每个动作至少有101个视频.视频一共提供了三个官方的分割,一般来说,使用这个数据集的人都是提供在split 1上的top 1准确率或者三个split的平均准确率

UCF101数据集在2012年被公开,是先前UCF50数据集的扩展.他从YouTube中得到了13320段视频,一共包含101中动作.数据集的测试方法和HMDB51是一样的

Dataset	Year	# Samples	Ave. Len	# Actions
HMDB51 [109]	2011	7K	$\sim$5s	51
UCF101 [190]	2012	13.3K	$\sim$6s	101
Sports1M [99]	2014	1.1M	$\sim$5.5m	487
ActivityNet [40]	2015	28K	$[5, 10]m$	200
YouTube8M [1]	2016	8M	$229.6s$	3862
Charades [186]	2016	9.8K	$30.1s$	157
Kinetics400 [100]	2017	306K	$10s$	400
Kinetics600 [12]	2018	482K	$10s$	600
Kinetics700 [13]	2019	650K	$10s$	700
Sth-Sth V1 [69]	2017	108.5K	$[2, 6]s$	174
Sth-Sth V2 [69]	2017	220.8K	$[2, 6]s$	174
AVA [70]	2017	385K	$15m$	80
AVA-kinetics [117]	2020	624K	$15m, 10s$	80
MIT [142]	2018	1M	$3s$	339
HACS Clips [267]	2019	1.55M	$2s$	200
HVU [34]	2020	572K	$10s$	739
AViD [165]	2020	450K	$[3, 15]s$	887

Table 1. **A list of popular datasets for video action recognition**

Sports1M是第一个大规模的数据集,在2014年发布.包含从YouTube上获取到的超过一百万和视频,分为487类.每一类动作划分的都非常细致,因此类内的差异就会很小.他的官方提供了10折验证划分

ActivityNet最早在2015年提出,而后ActivityNet又有非常多的版本.最近的版本是ActivityNet200,包含200个日常动作,一共有10024个训练数据,4926个验证数据和5044个测试数据.

out noisy classes/samples. Below we review the most popular large-scale video action recognition datasets in Table 1 and Figure 2.

HMDB51 [109] was introduced in 2011. It was collected mainly from movies, and a small proportion from public databases such as the Prelinger archive, YouTube and Google videos. The dataset contains 6,849 clips divided into 51 action categories, each containing a minimum of 101 clips. The dataset has three official splits. Most previous papers either report the top-1 classification accuracy on split 1 or the average accuracy over three splits.

UCF101 [190] was introduced in 2012 and is an extension of the previous UCF50 dataset. It contains 13,320 videos from YouTube spreading over 101 categories of human actions. The dataset has three official splits similar to HMDB51, and is also evaluated in the same manner.

Sports1M [99] was introduced in 2014 as the first large-scale video action dataset which consisted of more than 1 million YouTube videos annotated with 487 sports classes. The categories are fine-grained which leads to low inter-class variations. It has an official 10-fold cross-validation split for evaluation.

ActivityNet [40] was originally introduced in 2015 and the ActivityNet family has several versions since its initial launch. The most recent ActivityNet 200 (V1.3) contains 200 human daily living actions. It has 10,024 training, 4,926 validation, and 5,044 testing videos. On average there are 137 untrimmed videos per class and 1.41 activity instances per video.

YouTube8M [1] was introduced in 2016 and is by far the largest-scale video dataset that contains 8 million YouTube videos (500K hours of video in total) and annotated with 3,862 action classes. Each video is annotated with one or multiple labels by a YouTube video annotation system. This dataset is split into training, validation and test in the ratio

YouTube8M是2016年被公开的,所有视频由一个或者多个YouTube标注系统标注的.所有的视频按照7:2:1的比例划分的

Charades在2016年公开,主要为了真实生活中的动作理解.它包含9848个长度在30秒左右的视频.数据集包括157个多标注的日常室内活动,有267个不同的人扮演.他官方提供的训练/验证集合有7985个example用于训练,1863个用于验证

70:20:10. The validation set of this dataset is also extended with human-verified segment annotations to provide temporal localization information.

Charades [186] was introduced in 2016 as a dataset for real-life concurrent action understanding. It contains 9,848 videos with an average length of 30 seconds. This dataset includes 157 multi-label daily indoor activities, performed by 267 different people. It has an official train-validation split that has 7,985 videos for training and the remaining 1,863 for validation.

Kinetics Family is now the most widely adopted benchmark. Kinetics400 [100] was introduced in 2017 and it consists of approximately 240k training and 20k validation videos trimmed to 10 seconds from 400 human action categories. The Kinetics family continues to expand, with Kinetics-600 [12] released in 2018 with 480K videos and Kinetics700[13] in 2019 with 650K videos.

Kinect系列是目前使用最广泛的数据集.Kinect400在2017年被公开,包含240k个训练example和20k个验证example.每个视频都被修剪成10秒,一共包含400中人类动作.Kinect系列的数据量还在增加

20BN-Something-Something [69] V1 was introduced in 2017 and V2 was introduced in 2018. This family is another popular benchmark that consists of 174 action classes that describe humans performing basic actions with everyday objects. There are 108,499 videos in V1 and 220,847 videos in V2. Note that the Something-Something dataset requires strong temporal modeling because most activities cannot be inferred based on spatial features alone (e.g. opening something, covering something with something).

SS数据集在2017年发布了第一版,2018年发布了第二版.这一系列的数据集是目前最流行的数据集.V1包含了108499个视频,V2包含了220847个视频.需要注意的是SSV2数据集需要有强大的时空建模能力,因为他的视频都是没有办法通过单张图片推理得到的

AVA [70] was introduced in 2017 as the first large-scale spatio-temporal action detection dataset. It contains 430 15-minute video clips with 80 atomic actions labels (only 60 labels were used for evaluation). The annotations were provided at each key-frame which lead to 214,622 training, 57,472 validation and 120,322 testing samples. The AVA dataset was recently expanded to AVA-Kinetics with 352,091 training, 89,882 validation and 182,457 testing samples [117].

AVA数据集在2017年首次被提出,是第一个提供了时空动作侦测的数据集.它包含了430个15分钟的视频,一共包含80个原子动作标签.其中的60个用于测试性能.只有关键帧才被分配了标注

Moments in Time [142] was introduced in 2018 and it is a large-scale dataset designed for event understanding. It contains one million 3 second video clips, annotated with a dictionary of 339 classes. Different from other datasets designed for human action understanding, Moments in Time dataset involves people, animals, objects and natural phenomena. The dataset was extended to **Multi-Moments in Time (M-MiT)** [143] in 2019 by increasing the number of videos to 1.02 million, pruning vague classes, and increasing the number of labels per video.

HACS [267] was introduced in 2019 as a new large-scale dataset for recognition and localization of human actions collected from Web videos. It consists of two kinds of manual annotations. HACS Clips contains 1.55M 2-second clip annotations on 504K videos, and HACS Segments has 140K complete action segments (from action start to end) on 50K videos. The videos are annotated with the same 200 human action classes used in ActivityNet (V1.3) [40].

MiT是在2018年发布的用于时间理解的数据集.其中包含一百万个超过3秒的视频片段,标注为339类.和其他只包含人类动作数据集不一样,MiT包含了人类,动物,物体和自然现象的视频.在2019年这个数据集通过增加视频总量,删去模糊的类别以及增加label的方式升级到了M-Mit

HVU是2020年为了多任务/标签而发表的新数据集.HVU一共包含3142个标签.官方提供的划分中训练/验证/测试集分别有481K/31K/65K个视频.HVU数据集一共有六个任务目录,分别是:场景,物体,动作,时间,属性以及概念.

AViD则是在2020年发布的用于匿名动作识别的数据集.里面包含410K个训练数据和40K个测试数据.每段视频的时长都在3-15秒之间,分为887个类.在数据集制作期间,作者考虑到了种族和隐私问题.因此利用了多个国家的视频来确保数据集没有bias,并且抹去了人脸

HVU [34] dataset was released in 2020 for multi-label multi-task video understanding. This dataset has 572K videos and 3,142 labels. The official split has 481K, 31K and 65K videos for train, validation, and test respectively. This dataset has six task categories: scene, object, action, event, attribute, and concept. On average, there are about 2,112 samples for each label. The duration of the videos varies with a maximum length of 10 seconds.

AViD [165] was introduced in 2020 as a dataset for anonymized action recognition. It contains 410K videos for training and 40K videos for testing. Each video clip duration is between 3-15 seconds and in total it has 887 action classes. During data collection, the authors tried to collect data from various countries to deal with data bias. They also remove face identities to protect privacy of video makers. Therefore, AViD dataset might not be a proper choice for recognizing face-related actions.

在进入编年史之前,我们先来可视化一下数据集中的几个例子.

Before we dive into the chronological review of methods, we present several visual examples from the above datasets in Figure 4 to show their different characteristics. In the top two rows, we pick action classes from UCF101 [190] and Kinetics400 [100] datasets. Interestingly, we find that these actions can sometimes be determined by the context or scene alone. For example, the model can predict the action riding a bike as long as it recognizes a bike in the video frame. The model may also predict the action cricket bowling if it recognizes the cricket pitch. Hence for these classes, video action recognition may become an object/scene classification problem without the need of reasoning motion/temporal information. In the middle two rows, we pick action classes from Something-Something dataset [69]. This dataset focuses on human-object interaction, thus it is more fine-grained and requires strong temporal modeling. For example, if we only look at the first frame of dropping something and picking something up without looking at other video frames, it is impossible to tell these two actions apart. In the bottom row, we pick action classes from Moments in Time dataset [142]. This dataset is different from most video action recognition datasets, and is designed to have large inter-class and intra-class variation that represent dynamical events at different levels of abstraction. For example, the action climbing can have different actors (person or animal) in different environments (stairs or tree).

最上方是UCF101和Kinetics400的数据,我们发现其实视频的动作只需要背景或者场景就可以确定了

中间的数据集是SSV2的视频,这个数据集关注人物交互,因此这个数据集是一个更加细分的数据集而且需要很好的时空建模能力.

2.2. Challenges

Non-trivial就是具有一定复杂度,需要一定脑力活动、加工过程才能得到的(结果、结论、实现...)

There are several major challenges in developing effective video action recognition algorithms.

In terms of dataset, first, defining the label space for training action recognition models is non-trivial. It's because human actions are usually composite concepts and the hierarchy of these concepts are not well-defined. Second, annotating videos for action recognition are laborious (e.g., need to watch all the video frames) and ambiguous

因为丢东西和捡东西光看前几帧图像是无法分类的

HACS是在2019年发布的用于识别和定位的大型数据集.视频来源于网络,分类和ActivityNet一样,包含200个类别.标注有两种方式,一种是对504k个视频中得到的1.55M个2秒视频的标注,还有从50k个视频中得到140k个完整的动作片段(从开始到结束)

Figure 4. **Visual examples from popular video action datasets.** Top: individual video frames from action classes in UCF101 and Kinetics400. A single frame from these scene-focused datasets often contains enough information to correctly guess the category. Middle: consecutive video frames from classes in Something-Something. The 2nd and 3rd frames are made transparent to indicate the importance of temporal reasoning that we cannot tell these two actions apart by looking at the 1st frame alone. Bottom: individual video frames from classes in Moment in Time. Same action could have different actors in different environments.

(e.g, hard to determine the exact start and end of an action). Third, some popular benchmark datasets (e.g., Kinetics family) only release the video links for users to download and not the actual video, which leads to a situation that methods are evaluated on different data. It is impossible to do fair comparisons between methods and gain insights.

In terms of modeling, first, videos capturing human actions have both strong intra- and inter-class variations. People can perform the same action in different speeds under various viewpoints. Besides, some actions share similar movement patterns that are hard to distinguish. Second, recognizing human actions requires simultaneous understanding of both short-term action-specific motion information and long-range temporal information. We might need a sophisticated model to handle different perspectives rather than using a single convolutional neural network. Third, the computational cost is high for both training and inference, hindering both the development and deployment of action recognition models. In the next section, we will demonstrate how video action recognition methods developed over the last decade to address the aforementioned challenges.

现在的数据集存在的问题:1.类别的划分是非平凡的,因为人类的动作通常是由多个概念的,并且这些概念以不同的层次组合. 2.对视频进行标注是费时费力的. 3.一些数据集只提供了下载链接,因此不同的人可能在不同的数据上进行的测试,因此在不同的数据下训练得到的模型是不具有可比性的

第三类数据例如MiT这类是具有很大的类内和类间差异

因为里面的动力学运动时间是高级别的抽象

对于建模来说,也存在很多的问题,例如:1.人类动作之间存在很多的类内和类间差异. 2.识别人类动作需要同步理解长短期动作信息 3.计算开销很大

在英文及其他很多语言中，单词"奥德赛"(odyssey)现在用来指代一段史诗般的征程.
https://zh.wikipedia.org/wiki/%E5%A5%A5%E5%BE%B7%E8%B5%9B

尽管有的方法试着在视频动作识别使用卷积神经网络，但是在2015年以前，手工设计的特征，尤其是IDT这类特征，它们的高准确率导致了他们是占主导的方法.然而，手工设计的特征需要很大的计算量，并且非常难以部署到大型的数据集合上

3. An Odyssey of Using Deep Learning for Video Action Recognition

In this section, we review deep learning based methods for video action recognition from 2014 to present and introduce the related earlier work in context.

3.1. From hand-crafted features to CNNs

Despite there being some papers using Convolutional Neural Networks (CNNs) for video action recognition, [200, 5, 91], hand-crafted features [209, 210, 158, 112], particularly Improved Dense Trajectories (IDT) [210], dominated the video understanding literature before 2015, due to their high accuracy and good robustness. However, hand-crafted features have heavy computational cost [244], and are hard to scale and deploy.

随着深度学习技术的发展，越来越多的研究源开始把卷积网络运用到视频识别问题.最早的DeepVideo对每一帧图像使用2D卷积网络，然后探究了不同的时间模式的来让我们学习到用于视频识别的时空特征.虽然他提出了一些被后来的文章证明了非常好的模型.但是在UCF101上迁移学习的性能降低了20%.不仅如此，他还发现以单帧图像作为输入的模型的表现和以视频(堆叠起来的帧)作为输入的模型的性能是差不多的.这就意味着这个模型学到的特征无法捕捉到人的运动

With the rise of deep learning [107], researchers started to adapt CNNs for video problems. The seminal work DeepVideo [99] proposed to use a single 2D CNN model on each video frame independently and investigated several temporal connectivity patterns to learn spatio-temporal features for video action recognition, such as late fusion, early fusion and slow fusion. Though this model made early progress with ideas that would prove to be useful later such as a multi-resolution network, its transfer learning performance on UCF101 [190] was 20% less than hand-crafted IDT features (65.4% vs 87.9%). Furthermore, DeepVideo [99] found that a network fed by individual video frames, performs equally well when the input is changed to a stack of frames. This observation might indicate that the learnt spatio-temporal features did not capture the motion well. It also encouraged people to think about why CNN models did not outperform traditional hand-crafted features in the video domain unlike in other computer vision tasks [107, 171].

3.2. Two-stream networks

Since video understanding intuitively needs motion information, finding an appropriate way to describe the temporal relationship between frames is essential to improving the performance of CNN-based video action recognition.

光流可以准确的描述物体的运动.使用光流的优点就在于相比于RGB图像，光流提供了正交的信息

Optical flow [79] is an effective motion representation to describe object/scene movement. To be precise, it is the pattern of apparent motion of objects, surfaces, and edges in a visual scene caused by the relative motion between an observer and a scene. We show several visualizations of optical flow in Figure 5. As we can see, optical flow is able to describe the motion pattern of each action accurately. The advantage of using optical flow is it provides orthogonal information compared to the the RGB image. For example, the two images on the bottom of Figure 5 have cluttered backgrounds. Optical flow can effectively remove the nonmoving background and result in a simpler learning problem compared to using the original RGB images as input.

例如在图5中，光流可以有效地去除非运动的背景，因此相比于RGB图像来说是一种更加简单并且易于学习的问题

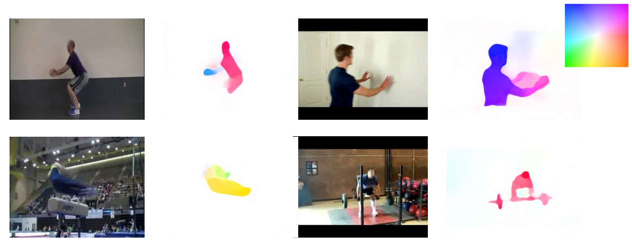


Figure 5. Visualizations of optical flow. We show four image-flow pairs, left is original RGB image and right is the estimated optical flow by FlowNet2 [85]. Color of optical flow indicates the directions of motion, and we follow the color coding scheme of FlowNet2 [85] as shown in top right.

光流的可视化图，左边是原图.右边是可视化的光流，颜色表示运动的方向，按照右上角的颜色方向对应图

除了光流能够提供额外的信息以外，光流也已经被证明在视频处理任务上可以工作的很好.传统的手工设计的特征例如IDT中也包含了类似于光流的特征，例如HOF和MBH

In addition, optical flow has been shown to work well on video problems. Traditional hand-crafted features such as IDT [210] also contain optical-flow-like features, such as Histogram of Optical Flow (HOF) and Motion Boundary Histogram (MBH).

Hence, Simonyan *et al.* [187] proposed two-stream networks, which included a spatial stream and a temporal stream as shown in Figure 6. This method is related to the two-streams hypothesis [65], according to which the human visual cortex contains two pathways: the ventral stream (which performs object recognition) and the dorsal stream (which recognizes motion). The spatial stream takes raw video frame(s) as input to capture visual appearance information. The temporal stream takes a stack of optical flow images as input to capture motion information between video frames. To be specific, [187] linearly rescaled the horizontal and vertical components of the estimated flow (i.e., motion in the x-direction and y-direction) to a [0, 255] range and compressed using JPEG. The output corresponds to the two optical flow images shown in Figure 6. The compressed optical flow images will then be concatenated as the input to the temporal stream with a dimension of $H \times W \times 2L$, where H, W and L indicates the height, width and the length of the video frames. In the end, the final prediction is obtained by averaging the prediction scores from both streams.

因此，Simonyan等人提出了双流网络.双流网络是基于人类是视觉皮层包含两条通路的假设:ventral流负责完成物体的识别，而dorsal流负责识别物体的运动.因此，他们提出的双流网络中的空间流以输入的帧做数如入来捕获视觉信息.而事件流以堆叠的光流作为输入来捕获运动信息

By adding the extra temporal stream, for the first time, a CNN-based approach achieved performance similar to the previous best hand-crafted feature IDT on UCF101 (88.0% vs 87.9%) and on HMDB51 [109] (59.4% vs 61.1%). [187] makes two important observations. First, motion information is important for video action recognition. Second, it is still challenging for CNNs to learn temporal information directly from raw video frames. Pre-computing optical flow as the motion representation is an effective way for deep learning to reveal its power. Since [187] managed to close the gap between deep learning approaches and traditional hand-crafted features, many follow-up papers on two-stream networks emerged and greatly advanced the devel-

通过添加额外的时间流，基于CNN的方法第一次得到了和传统方法相近的性能.

187提出的双流网络提供了两个重要的理解:1.对于视频动作识别来说，运动信息是非常重要的 2.对于CNN来说，直接从视频帧中学习得到时间信息是很困难的. 从提前计算得到的光流中进行学习对于深度学习来说是有效地

双流网络中的空间流用于识别物体，其以原始视频帧作为输入，而时间流用于识别物体的运动，其输入是输入视频的光流.具体来说是提取了原始视频x和y方向上的光流，然后把值压缩到0-255之后编码成JPEG格式，和原始输入格式对齐.接下来各自截取一半，融合之后作为输入.最终的结果是时间流和空间流的预测分数平均之后进行分类

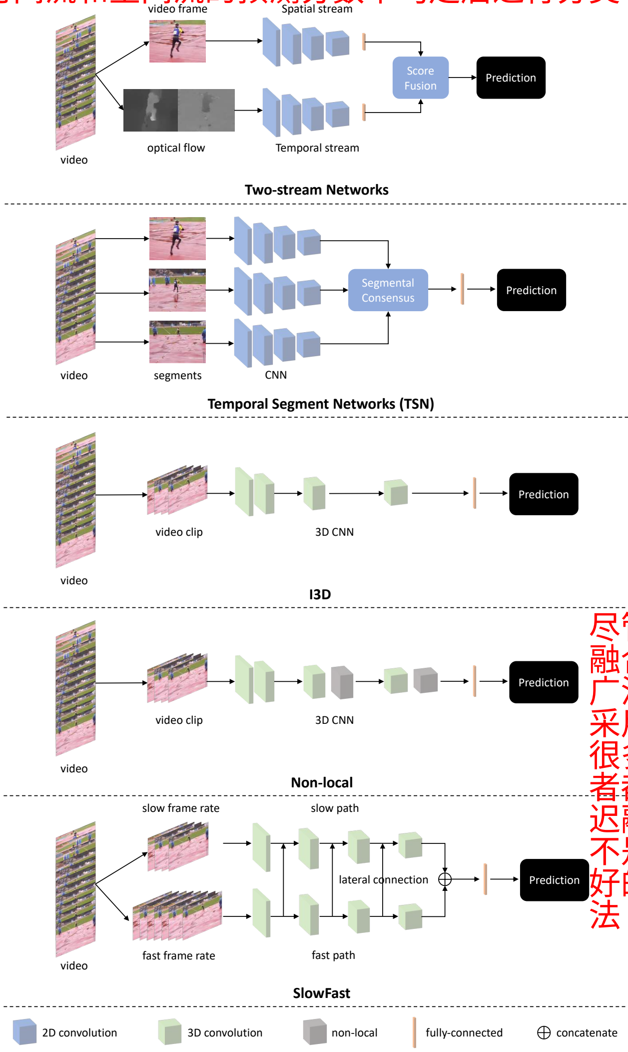


Figure 6. **Workflow of five important papers**: two-stream networks [187], temporal segment networks [218], I3D [14], Non-local [219] and SlowFast [45]. Best viewed in color.

因为187中提出的双流网络第一次弥补了深度学习方法和传统方法之间的差距，因此再这之后出现了很多双流系列的网络来改进其性能.这些方法极大地推动了视频动作识别的发展.因此我们下面将会分解介绍这些方法

opment of video action recognition. Here, we divide them into several categories and review them individually.

3.2.1 Using deeper network architectures

因为187中提出的双流网络使用的是比较浅的网络，因此一个非常自然的扩展就是使用深层网络

Two-stream networks [187] used a relatively shallow network architecture [107]. Thus a natural extension to the two-stream networks involves using deeper networks. However, Wang *et al.* [215] finds that simply using deeper networks does not yield better results, possibly due to overfitting on the small-sized video datasets [190, 109]. Recall from section 2.1, UCF101 and HMDB51 datasets only have thousands of training videos. Hence, Wang *et al.* [217] introduce a series of good practices, including cross-modality initialization, synchronized batch normalization, corner cropping and multi-scale cropping data augmentation, large dropout ratio, etc. to prevent deeper networks from overfitting. With these good practices, [217] was able to train a two-stream network with the VGG16 model [188] that outperforms [187] by a large margin on UCF101. These good practices have been widely adopted and are still being used. Later, Temporal Segment Networks (TSN) [218] performed a thorough investigation of network architectures, such as VGG16, ResNet [76], Inception [198], and demonstrated that deeper networks usually achieve higher recognition accuracy for video action recognition. We will describe more details about TSN in section 3.2.4.

但是Wang等人却发现简单的使用更深层的网络并不会带来性能上的提升.这很有可能是由于overfit了小数据集.因此，Wang等人在217中提出了一系列非常好用的训练技巧，包括跨模态初始化，同步batch normalization，角落切割，多尺度切割，大dropout率等等方式.最终成功使用VGG16作为backbone，并且在UCF101数据集上极大地超越了原始的双流网络

直到现在，Wang等人提出的技巧仍然在被广泛的使用

接下来TSN中对不同的网络结构进行了详细的比较，例如VGG16，最后发现更深的网络往往会得到更好的性能

因为双流网络中存在两条通路，因此就一定会存在两条通路信息融合的阶段.这个阶段通常称为时空融合阶段.

3.2.2 Two-stream fusion

Since there are two streams in a two-stream network, there will be a stage that needs to merge the results from both networks to obtain the final prediction. This stage is usually referred to as the spatial-temporal fusion step.

实现时空融合最简单的一步就是迟融合，即对两个通路的网络最后的结果进行加权平均

尽管迟融合被广泛的采用，很多学者都说迟融合不是最好的方法

The easiest and most straightforward way is late fusion, which performs a weighted average of predictions from both streams. Despite late fusion being widely adopted [187, 217], many researchers claim that this may not be the optimal way to fuse the information between the spatial appearance stream and temporal motion stream. They believe that earlier interactions between the two networks could benefit both streams during model learning and this is termed as early fusion.

诸多学者相信在特征提取的初期进行特征融合是更加有益的

Fusion [50] is one of the first of several papers investigating the early fusion paradigm, including how to perform spatial fusion (e.g., using operators such as sum, max, bilinear, convolution and concatenation), where to fuse the network (e.g., the network layer where early interactions happen), and how to perform temporal fusion (e.g., using 2D or 3D convolutional fusion in later stages of the network). [50] shows that early fusion is beneficial for both streams to learn richer features and leads to improved performance over late fusion. Following this line of research, Feichtenhofer *et al.* [46] generalizes ResNet [76] to the spatiotemporal domain by introducing residual connections between the two streams. Based on [46], Feichtenhofer *et al.* [47] further propose a multiplicative gating function for residual networks to learn better spatio-temporal features. Concurrently, [225] adopts a spatio-temporal pyramid to perform hierarchical early fusion between the two streams.

Fusion是最早研究早融合范式的文章，包括如何进行空间融合(即探究不同的运算方法，sum，max，卷积或者拼接)，在那里进行融合以及如何进行时间融合.50中的结果最终表明早融合对于两条路来说都是有益的，并且相比于晚融合会提升性能

Feichtenhofer等人把残差网络泛化到了两条路的时空领域，并且在47中更仅有一步提出了更好的函数

3.2.3 Recurrent neural networks

Since a video is essentially a temporal sequence, researchers have explored Recurrent Neural Networks (RNNs) for temporal modeling inside a video, particularly the usage of Long Short-Term Memory (LSTM) [78].

LRCN [37] and Beyond-Short-Snippets [253] are the first of several papers that use LSTM for video action recog-

说到底，视频还是一个时序数据，因此用一些研究者尝试如何把RNN(尤其是LSTM)运用到视频动作识别领域.LRCN和BSS是最早的几篇把LSTM作为双流网络中一路从而把LSTM运用到视频动作识别任务上的文章

LRCN和BSS把CNN得到的feature map作为LSTM的输入，然后通过聚合帧级别的CNN特征来得到最后的视频级别的预测.需要注意的是,他们都是在两条路中分别使用的LSTM,然后进行的迟融合.但是,使用LSTM相比于原始的双流网络来说并没有非常多的提升.

跟随着CNN-LSTM的范式，有很多变体被提出来了，包括双向LSTM,CNN-LSTM融合,分层LSTM融合

nition under the two-stream networks setting. They take the feature maps from CNNs as an input to a deep LSTM network, and aggregate frame-level CNN features into video-level predictions. Note that they use LSTM on two streams separately, and the final results are still obtained by late fusion. However, there is no clear empirical improvement from LSTM models [253] over the two-stream baseline [187]. Following the CNN-LSTM framework, several variants are proposed, such as bi-directional LSTM [205], CNN-LSTM fusion [56] and hierarchical multi-granularity LSTM network [118]. [125] described VideoLSTM which includes a correlation-based spatial attention mechanism and a lightweight motion-based attention mechanism. VideoLSTM not only show improved results on action recognition, but also demonstrate how the learned attention can be used for action localization by relying on just the action class label. Lattice-LSTM [196] extends LSTM by learning independent hidden state transitions of memory cells for individual spatial locations, so that it can accurately model long-term and complex motions. ShuttleNet [183] is a concurrent work that considers both feedforward and feedback connections in a RNN to learn long-term dependencies. FASTER [272] designed a FAST-GRU to aggregate clip-level features from an expensive backbone and a cheap backbone. This strategy reduces the processing cost of redundant clips and hence accelerates the inference speed.

125中介绍的VideoLSTM使用了基于相关性的空间注意力机制，以及一个轻量的基于动作的注意力机制.V-LSTM的结果不仅提高了视频动作识别的精度上限,并且说明了学习到的注意力可以帮助进行动作定位

上面说到的方法使用不同的LTSM,因此他们之间的区别不好比较

However, the work mentioned above [37, 253, 125, 196, 183] use different two-stream networks/backbones. The differences between various methods using RNNs are thus unclear. Ma *et al.* [135] build a strong baseline for fair comparison and thoroughly study the effect of learning spatio-temporal features by using RNNs. They find that it requires proper care to achieve improved performance, e.g., LSTMs require pre-segmented data to fully exploit the temporal information. RNNs are also intensively studied in video action localization [189] and video question answering [274], but these are beyond the scope of this survey.

Ma等人首先建立了一个强大的baseline作为比较的基准,然后使用相同的LSTM作为Backbone来研究不同的LSTM的性能.最后他们发现LSTM需要预分段的数据来完全获取时间信息

3.2.4 Segment-based methods

通过光流进行学习使得模型可以学习到短时的时空信息,但是却无法学习到长时信息.因此Wang等人提出了TSN来进行视频级别的动作识别

Thanks to optical flow, two-stream networks are able to reason about short-term motion information between frames. However, they still cannot capture long-range temporal information. Motivated by this weakness of two-stream networks , Wang *et al.* [218] proposed a Temporal Segment Network (TSN) to perform video-level action recognition. Though initially proposed to be used with 2D CNNs, it is simple and generic. Thus recent work using either 2D or 3D CNNs, are still built upon this framework.

To be specific, as shown in Figure 6, TSN first divides a whole video into several segments, where the segments distribute uniformly along the temporal dimension. Then TSN randomly selects a single video frame within each segment and forwards them through the network. Here, the network shares weights for input frames from all the segments. In the end, a segmental consensus is performed to aggregate information from the sampled video frames. The segmental consensus could be operators like average pooling, max pooling, bilinear encoding, etc. In this sense, TSN is capable of modeling long-range temporal structure because the model sees the content from the entire video. In addition, this sparse sampling strategy lowers the training cost over long video sequences but preserves relevant information.

虽然TSN最早被提出来是用于2D的卷积,但是其思想却非常简单并且易于泛化.因此最近的许多文章2D或者3D的卷积网络都是建立在TSN的框架的基础上的

TSN首先把整个视频分成几个在时间上均匀分布视频段,然后在每一段中抽出来单独的一帧,让网络进行推理.最后会进行段间共识来聚合不同采样的帧之间的信息.段间共识的计算方法可以是平均池化,最大池化,线性插值等等.TSN可以建模长期的时间结构,因为模型看完了视频的所有内容.此外,稀疏采样策略也帮助减少了长时间视频序列的训练代价并且能够保存相关的信息

Given TSN's good performance and simplicity, most two-stream methods afterwards become segment-based two-stream networks. Since the segmental consensus is simply doing a max or average pooling operation, a feature encoding step might generate a global video feature and lead to improved performance as suggested in traditional approaches [179, 97, 157]. Deep Local Video Feature (DVOF) [114] proposed to treat the deep networks that trained on local inputs as feature extractors and train another encoding function to map the global features into global labels. Temporal Linear Encoding (TLE) network [36] appeared concurrently with DVOF, but the encoding layer was embedded in the network so that the whole pipeline could be trained end-to-end. VLAD3 and ActionVLAD [123, 63] also appeared concurrently. They extended the NetVLAD layer [4] to the video domain to perform video-level encoding, instead of using compact bilinear encoding as in [36]. To improve the temporal reasoning ability of TSN, Temporal Relation Network (TRN) [269] was proposed to learn and reason about temporal dependencies between video frames at multiple time scales. The recent state-of-the-art efficient model TSM [128] is also segment-based. We will discuss it in more detail in section 3.4.2.

因为TSN的简单和高效,接下来许多双流模型都使用了基于段的双流网络.因为段间共识就是进行简单的操作,因此再进行一个特征编码就可能获得视频的全局特征,并且会带来性能的提升.因此,DVOF提出了把接受局部视频的网络作为特征提取器,然后在训练一个编码函数来把全局特征对应到类上

类似的,VLAD3和ActionVLAD把NetVLAD扩展到了视频领域来获得视频级别的编码.此外为了提高TSN的时间推理能力,研究者提出了能够学习到不同帧之间的时间相关性的TRN

3.2.5 Multi-stream networks

双流网络之所以可以成功,就是因为出现和运动是视频最重要的两个特征.但是除了这两个特征以外,还有一些重要的因素可以帮助视频动作识别,例如姿态,物体,音频等等

Two-stream networks are successful because appearance and motion information are two of the most important properties of a video. However, there are other factors that can help video action recognition as well, such as pose, object, audio and depth, etc.

Pose information is closely related to human action. We can recognize most actions by just looking at a pose (skeleton) image without scene context. Although there is previous work on using pose for action recognition [150, 246], P-CNN [23] was one of the first deep learning methods that successfully used pose to improve video action recognition. P-CNN proposed to aggregates motion and appearance information along tracks of human body parts, in a similar spirit to trajectory pooling [214]. [282] extended this pipeline to a chained multi-stream framework, that computed and integrated appearance, motion and pose. They

姿态信息和人的动作有紧密的关系.对于我们人来说,我们能够通过姿态(骨架)就可以识别出来绝大多数人的动作.尽管先前有一些工作使用姿态来识别动作,但是P-CNN是第一个成功使用深度学习的方法来提升视频动作识别的准确率.P-CNN中提出了通过人体部分的追踪来聚合运动和外观信息.282中的多流网络来获取出现,运动和姿态信息

282中引入了马尔科夫链来把出现,运动和姿态连续的加入进行推测,并且最终在动作分类和动作识别上获取到不错的效果.25中提出的PoTion是P-CNN的后续作品,但是引入了更加强大的特征表示,即人体语义关键点的运动.PoTion中首先通过一个人体姿态估计器来抽取到人体关节点的热力图.然后在时空上不断聚合获取到的关节点的概率图最终得到分类的结果

introduced a Markov chain model that added these cues successively and obtained promising results on both action recognition and action localization. PoTion [25] was a follow-up work to P-CNN, but introduced a more powerful feature representation that encoded the movement of human semantic keypoints. They first ran a decent human pose estimator and extracted heatmaps for the human joints in each frame. They then obtained the PoTion representation by temporally aggregating these probability maps. PoTion is lightweight and outperforms previous pose representations [23, 282]. In addition, it was shown to be complementary to standard appearance and motion streams, e.g. combining PoTion with I3D [14] achieved state-of-the-art result on UCF101 (98.2%).

PoTion相比于先前的使用姿态表示模型来说是一个轻量的模型,并且超越了他们的表现.此外PoTion和标准的出现和运动流对比是补充性质的,因此结合了I3D和PoTion就得到了SOTA的表现

除了姿势以外,物体信息也是非常重要的信息

Object information is another important cue because most human actions involve human-object interaction. Wu [232] proposed to leverage both object features and scene features to help video action recognition. The object and scene features were extracted from state-of-the-art pretrained object and scene detectors. Wang *et al*. [252] took a step further to make the network end-to-end trainable. They introduced a two-stream semantic region based method, by replacing a standard spatial stream with a Faster RCNN network [171], to extract semantic information about the object, person and scene.

Wu等人提出了利用物体特征和场景信息来帮助进行视频动作识别

音频信号也存在于视频中,而且和视觉信息是互补的

Audio signals usually come with video, and are complementary to the visual information. Wu *et al*. [233] introduced a multi-stream framework that integrates spatial, short-term motion, long-term temporal and audio in videos to digest complementary clues. Recently, Xiao *et al*. [237] introduced AudioSlowFast following [45], by adding another audio pathway to model vision and sound in an unified representation.

在RGBD动作识别领域使用深度信息是标准的做法,但是对于单目相机拍摄到的图片中其实是没有深度信息的,280中尝试使用深度估计器来估计深度信息

In RGB-D video action recognition field, using depth information is standard practice [59]. However, for vision-based video action recognition (e.g., only given monocular videos), we do not have access to ground truth depth information as in the RGB-D domain. An early attempt Depth2Action [280] uses off-the-shelf depth estimators to extract depth information from videos and use it for action recognition.

Essentially, multi-stream networks is a way of multi-modality learning, using different cues as input signals to help video action recognition. We will discuss more on multi-modality learning in section 5.12.

3.3. The rise of 3D CNNs

预局算的光流网络对计算量的需求非常大,并且需要大量的存储空间.对于大规模训练和部署来说是非常不友好的.因此在概念上非常简单的方式就是把一个视频视为3D张量,有两个空间维度和一个时间维度.因此使用3D卷积作为时空信息建模的单元就很自然

Pre-computing optical flow is computationally intensive and storage demanding, which is not friendly for large-scale training or real-time deployment. A conceptually easy way to understand a video is as a 3D tensor with two spatial and one time dimension. Hence, this leads to the usage of 3D CNNs as a processing unit to model the temporal information in a video.

91是最早使用3D卷积网络的文章,虽然非常有启发意义,但是它们的模型并不是很深,因此没有展现出来非常强大的性能.因此,Tran等人在202中将91中的模型扩展为了更深的3D卷积网络,称为C3D.C3D的模型设计遵从了188中的模范结构.188中的结构可以说是VGG网络的翻版.虽然C3D在标准的基准测试集上的表现并不是非常好,但是却展现出了强大的泛化能力,因此可以用于多种视频任务通用的特征抽取器

The seminal work for using 3D CNNs for action recognition is [91]. While inspiring, the network was not deep enough to show its potential. Tran *et al*. [202] extended [91] to a deeper 3D network, termed C3D. C3D follows the modular design of [188], which could be thought of as a 3D version of VGG16 network. Its performance on standard benchmarks is not satisfactory, but shows strong generalization capability and can be used as a generic feature extractor for various video tasks [250].

然而,3D卷积核是很难进行优化的,为了训练3D卷积核,人们需要大型的数据集,能够包含多种不同的说过内容和动作类别

However, 3D networks are hard to optimize. In order to train a 3D convolutional filter well, people need a large-scale dataset with diverse video content and action categories. Fortunately, there exists a dataset, Sports1M [99] which is large enough to support the training of a deep 3D network. However, the training of C3D takes weeks to converge. Despite the popularity of C3D, most users just adopt it as a feature extractor for different use cases instead of modifying/fine-tuning the network. This is partially the reason why two-stream networks based on 2D CNNs dominated the video action recognition domain from year 2014 to 2017.

尽管有大型的数据集支持C3D的训练,C3D却需要几个周的时间收敛

因此绝大多数研究者就是把C3D当做了特征提取器.而高昂的训练代价可能就是14-17年这几年间双流网络占据主要地位的原因.这一现象在2017年Carreira提出了I3D之后得到了解决.

The situation changed when Carreira *et al*. [14] proposed I3D in year 2017. As shown in Figure 6, I3D takes a video clip as input, and forwards it through stacked 3D convolutional layers. A video clip is a sequence of video frames, usually 16 or 32 frames are used. The major contributions of I3D are: 1) it adapts mature image classification architectures to use for 3D CNN; 2) For model weights, it adopts a method developed for initializing optical flow networks in [217] to inflate the ImageNet pre-trained 2D model weights to their counterparts in the 3D model. Hence, I3D bypasses the dilemma that 3D CNNs have to be trained from scratch. With pre-training on a new large-scale dataset Kinetics400 [100], I3D achieved a 95.6% on UCF101 and 74.8% on HMDB51. I3D ended the era where different methods reported numbers on small-sized datasets such as UCF101 and HMDB51[2]. Publications following I3D needed to report their performance on Kinetics400, or other large-scale benchmark datasets, which pushed video action recognition to the next level. In the next few years, 3D CNNs advanced quickly and became top performers on almost every benchmark dataset. We will review the 3D CNNs based literature in several categories below.

I3D中使用堆叠的视频片段作为输入,每一个视频片段都是16或者32帧的片段.I3D的主要贡献有三个:1.能够把成熟的图像分类网络运用到视频分类上来.2.模型参数初始化可以通过膨胀技术来把在ImageNet上预训练的权重迁移到3D模型上来.此外,在Kinect400上预训练的得到权重在UCF101和HMDB51上的准确度达到了

We want to point out that 3D CNNs are not replacing two-stream networks, and they are not mutually exclusive. They just use different ways to model the temporal relationship in a video. Furthermore, the two-stream approach is a generic framework for video understanding, instead of a specific method. As long as there are two networks, one for spatial appearance modeling using RGB frames, the other for temporal motion modeling using optical flow, the

[2]As we can see in Table 2.

需要说明的是,3D卷积网络并没有取代双流网络,因为两者并不是互斥的,3D卷积网络只不过是另外一种建模视频中的时空信息的方式.其实双流网络是指一类网络,只要模型中有两个网络,一个用于建模空间信息,一个用于建模时间信息就可以被归类为双流网络

在I3D的原始文章中,其实也构建了一条时间通路,I3D结合时间通路得到的性能甚至更高.因此最终的I3D的模型是3D卷积网络和双流网络的结合.但是这页从某个角度说明I3D本身并不是依赖于时间流,即与双流网络是不同的方法

method may be categorized into the family of two-stream networks. In [14], they also build a temporal stream with I3D architecture and achieved even higher performance, 98.0% on UCF101 and 80.9% on HMDB51. Hence, the final I3D model is a combination of 3D CNNs and two-stream networks. However, the contribution of I3D does not lie in the usage of optical flow.

2D卷积网络从大规模的图像数据集(ImageNet/Places205)中受益,而且图像领域的数据远远不是视频数据集可比的.因此就有很多的研究者致力于把2D图像的领先运用到3D领域

3.3.1 Mapping from 2D to 3D CNNs

2D CNNs enjoy the benefit of pre-training brought by the large-scale of image datasets such as ImageNet [30] and Places205 [270], which cannot be matched even with the largest video datasets available today. On these datasets numerous efforts have been devoted to the search for 2D CNN architectures that are more accurate and generalize better. Below we describe the efforts to capitalize on these advances for 3D CNNs.

ResNet直接使用2D卷积网络的结构,并且把卷积核更换为3D卷积核受到了ResNext的启发,Chen等人提出了一个多纤维网络.通过把一个复杂的网络分割为多个轻量级的网络(fiber)的集合

ResNet3D [74] directly took 2D ResNet [76] and replaced all the 2D convolutional filters with 3D kernels. They believed that by using deep 3D CNNs together with large-scale datasets one can exploit the success of 2D CNNs on ImageNet. Motivated by ResNeXt [238], Chen *et al.* [20] presented a multi-fiber architecture that slices a complex neural network into an ensemble of lightweight networks (fibers) that facilitate information flow between fibers, reduces the computational cost at the same time. Inspired by SENet [81], STCNet [33] propose to integrate channel-wise information inside a 3D block to capture both spatial-channels and temporal-channels correlation information throughout the network.

3.3.2 Unifying 2D and 3D CNNs

为了减小3D卷积计算的复杂度,P3D和R2+1D探索了3D卷积核分解.具体来说,一个3D卷积核可以分解为两个独立的卷积操作

To reduce the complexity of 3D network training, P3D [169] and R2+1D [204] explore the idea of 3D factorization. To be specific, a 3D kernel (e.g., $3\times3\times3$) can be factorized to two separate operations, a 2D spatial convolution (e.g., $1\times3\times3$) and a 1D temporal convolution (e.g., $3\times1\times1$). The differences between P3D and R2+1D are how they arrange the two factorized operations and how they formulate each residual block. Trajectory convolution [268] follows this idea but uses deformable convolution for the temporal component to better cope with motion.

另外一种简化3D卷积的方式是在一个网络中混合2D和3D卷积.MiCTNet中通过混合2D和3D卷积实现了更深的网络,并且得到的特征更加具有意义

Another way of simplifying 3D CNNs is to mix 2D and 3D convolutions in a single network. MiCTNet [271] integrates 2D and 3D CNNs to generate deeper and more informative feature maps, while reducing training complexity in each round of spatio-temporal fusion. ARTNet [213] introduces an appearance-and-relation network by using a new building block. The building block consists of a spatial branch using 2D CNNs and a relation branch using 3D CNNs. S3D [239] combines the merits from approaches mentioned above. It first replaces the 3D convolutions at

ARTNet中通过他们提出的模块构建了出现-关系网络.这个木块包含使用2D卷积的空间分支和使用3D卷积的事件分支

S3D中首先发现把网络底端的3D卷积核更换为2D卷积核能够提高篇网络的准确率,因此S3D中就把所有剩余的卷积核(类似P3D和R2+1D一样的)进行了分解.同时期的工作ECO也采用了类似的头重脚轻的网络来实现在线级别的视频理解

the bottom of the network with 2D kernels, and find that this kind of top-heavy network has higher recognition accuracy. Then S3D factorizes the remaining 3D kernels as P3D and R2+1D do, to further reduce the model size and training complexity. A concurrent work named ECO [283] also adopts such a top-heavy network to achieve online video understanding.

在3D卷积网络中,长期时间练习可以通过堆叠多个短时卷积网络.但是有用时间信息可能在网络的后半部分丢失,尤其是在很后面的阶段

3.3.3 Long-range temporal modeling

In 3D CNNs, long-range temporal connection may be achieved by stacking multiple short temporal convolutions, e.g., $3\times3\times3$ filters. However, useful temporal information may be lost in the later stages of a deep network, especially for frames far apart.

为了能够实现长时时间建模,LTC引入并且评价了在多个视频帧之间的长期卷积.然而由于GPU显存的限制,LTC的作者牺牲了输入的分辨率.在这之后T3D使用了密集连接结构来保证元时输入信息在深层网络也能够得到很好的保留

In order to perform long-range temporal modeling, LTC [206] introduces and evaluates long-term temporal convolutions over a large number of video frames. However, limited by GPU memory, they have to sacrifice input resolution to use more frames. After that, T3D [32] adopted a densely connected structure [83] to keep the original temporal information as complete as possible to make the final prediction. Later, Wang *et al.* [219] introduced a new building block, termed non-local. Non-local is a generic operation similar to self-attention [207], which can be used for many computer vision tasks in a plug-and-play manner. As shown in Figure 6, they used a spacetime non-local module after later residual blocks to capture the long-range dependence in both space and temporal domain, and achieved improved performance over baselines without bells and whistles. Wu *et al.* [229] proposed a feature bank representation, which embeds information of the entire video into a memory cell, to make context-aware prediction. Recently, V4D [264] proposed video-level 4D CNNs, to model the evolution of long-range spatio-temporal representation with 4D convolutions.

为了进一步提升3D卷及网络的运算效率(即每秒进行的浮点计算次数,模型参数,计算延迟),很多3D卷积网络的变体被提出来了

3.3.4 Enhancing 3D efficiency

In order to further improve the efficiency of 3D CNNs (i.e., in terms of GFLOPs, model parameters and latency), many variants of 3D CNNs begin to emerge.

收到了2D卷积网络高效性的启发,不少研究者开始研究如何在视频领域使用逐深度的分组卷积

Motivated by the development in efficient 2D networks, researchers started to adopt channel-wise separable convolution and extend it for video classification [111, 203]. CSN [203] reveals that it is a good practice to factorize 3D convolutions by separating channel interactions and spatio-temporal interactions, and is able to obtain state-of-the-art performance while being 2 to 3 times faster than the previous best approaches. These methods are also related to multi-fiber networks [20] as they are all inspired by group convolution.

Recently, Feichtenhofer *et al.* [45] proposed SlowFast, an efficient network with a slow pathway and a fast pathway. The network design is partially inspired by the biological Parvo- and Magnocellular cells in the primate visual systems. As shown in Figure 6, the slow pathway operates at low frame rates to capture detailed semantic information, while the fast pathway operates at high temporal resolution to capture rapidly changing motion. In order to incorporate motion information such as in two-stream networks, SlowFast adopts a lateral connection to fuse the representation learned by each pathway. Since the fast pathway can be made very lightweight by reducing its channel capacity, the overall efficiency of SlowFast is largely improved. Although SlowFast has two pathways, it is different from the two-stream networks [187], because the two pathways are designed to model different temporal speeds, not spatial and temporal modeling. There are several concurrent papers using multiple pathways to balance the accuracy and efficiency [43].

Following this line, Feichtenhofer [44] introduced X3D that progressively expand a 2D image classification architecture along multiple network axes, such as temporal duration, frame rate, spatial resolution, width, bottleneck width, and depth. X3D pushes the 3D model modification/factorization to an extreme, and is a family of efficient video networks to meet different requirements of target complexity. With similar spirit, A3D [276] also leverages multiple network configurations. However, A3D trains these configurations jointly and during inference deploys only one model. This makes the model at the end more efficient. In the next section, we will continue to talk about efficient video modeling, but not based on 3D convolutions.

3.4. Efficient Video Modeling

With the increase of dataset size and the need for deployment, efficiency becomes an important concern.

If we use methods based on two-stream networks, we need to pre-compute optical flow and store them on local disk. Taking Kinetics400 dataset as an illustrative example, storing all the optical flow images requires 4.5TB disk space. Such a huge amount of data would make I/O become the tightest bottleneck during training, leading to a waste of GPU resources and longer experiment cycle. In addition, pre-computing optical flow is not cheap, which means all the two-stream networks methods are not real-time.

If we use methods based on 3D CNNs, people still find that 3D CNNs are hard to train and challenging to deploy. In terms of training, a standard SlowFast network trained on Kinetics400 dataset using a high-end 8-GPU machine takes 10 days to complete. Such a long experimental cycle and huge computing cost makes video understanding research only accessible to big companies/labs with abundant computing resources. There are several recent attempts to speed up the training of deep video models [230], but these are still expensive compared to most image-based computer vision tasks. In terms of deployment, 3D convolution is not as well supported as 2D convolution for different platforms. Furthermore, 3D CNNs require more video frames as input which adds additional IO cost.

Hence, starting from year 2018, researchers started to investigate other alternatives to see how they could improve accuracy and efficiency at the same time for video action recognition. We will review recent efficient video modeling methods in several categories below.

3.4.1 Flow-mimic approaches

One of the major drawback of two-stream networks is its need for optical flow. Pre-computing optical flow is computationally expensive, storage demanding, and not end-to-end trainable for video action recognition. It is appealing if we can find a way to encode motion information without using optical flow, at least during inference time.

[146] and [35] are early attempts for learning to estimate optical flow inside a network for video action recognition. Although these two approaches do not need optical flow during inference, they require optical flow during training in order to train the flow estimation network. Hidden two-stream networks [278] proposed MotionNet to replace the traditional optical flow computation. MotionNet is a lightweight network to learn motion information in an unsupervised manner, and when concatenated with the temporal stream, is end-to-end trainable. Thus, hidden two-stream CNNs [278] only take raw video frames as input and directly predict action classes without explicitly computing optical flow, regardless of whether its the training or inference stage. PAN [257] mimics the optical flow features by computing the difference between consecutive feature maps. Following this direction, [197, 42, 116, 164] continue to investigate end-to-end trainable CNNs to learn optical-flow-like features from data. They derive such features directly from the definition of optical flow [255]. MARS [26] and D3D [191] used knowledge distillation to combine two-stream networks into a single stream, e.g., by tuning the spatial stream to predict the outputs of the temporal stream. Recently, Kwon *et al*. [110] introduced MotionSqueeze module to estimate the motion features. The proposed module is end-to-end trainable and can be plugged into any network, similar to [278].

3.4.2 Temporal modeling without 3D convolution

A simple and natural choice to model temporal relationship between frames is using 3D convolution. However, there are many alternatives to achieve this goal. Here, we will review some recent work that performs temporal modeling without 3D convolution.

Lin *et al*. [128] introduce a new method, termed temporal shift module (TSM). TSM extends the shift operation

[228] to video understanding. It shifts part of the channels along the temporal dimension, thus facilitating information exchanged among neighboring frames. In order to keep spatial feature learning capacity, they put temporal shift module inside the residual branch in a residual block. Thus all the information in the original activation is still accessible after temporal shift through identity mapping. The biggest advantage of TSM is that it can be inserted into a 2D CNN to achieve temporal modeling at zero computation and zero parameters. Similar to TSM, TIN [182] introduces a temporal interlacing module to model the temporal convolution.

There are several recent 2D CNNs approaches using attention to perform long-term temporal modeling [92, 122, 132, 133]. STM [92] proposes a channel-wise spatio-temporal module to present the spatio-temporal features and a channel-wise motion module to efficiently encode motion features. TEA [122] is similar to STM, but inspired by SENet [81], TEA uses motion features to recalibrate the spatio-temporal features to enhance the motion pattern. Specifically, TEA has two components: motion excitation and multiple temporal aggregation, while the first one handles short-range motion modeling and the second one efficiently enlarge the temporal receptive field for long-range temporal modeling. They are complementary and both light-weight, thus TEA is able to achieve competitive results with previous best approaches while keeping FLOPs as low as many 2D CNNs. Recently, TEINet [132] also adopts attention to enhance temporal modeling. Note that, the above attention-based methods are different from non-local [219], because they use channel attention while non-local uses spatial attention.

3.5. Miscellaneous

In this section, we are going to show several other directions that are popular for video action recognition in the last decade.

3.5.1 Trajectory-based methods

While CNN-based approaches have demonstrated their superiority and gradually replaced the traditional hand-crafted methods, the traditional local feature pipeline still has its merits which should not be ignored, such as the usage of trajectory.

Inspired by the good performance of trajectory-based methods [210], Wang *et al*. [214] proposed to conduct trajectory-constrained pooling to aggregate deep convolutional features into effective descriptors, which they term as TDD. Here, a trajectory is defined as a path tracking down pixels in the temporal dimension. This new video representation shares the merits of both hand-crafted features and deep-learned features, and became one of the top performers on both UCF101 and HMDB51 datasets in the year 2015. Concurrently, Lan *et al*. [113] incorporated both Independent Subspace Analysis (ISA) and dense trajectories into the standard two-stream networks, and show the complementarity between data-independent and data-driven approaches. Instead of treating CNNs as a fixed feature extractor, Zhao *et al*. [268] proposed trajectory convolution to learn features along the temporal dimension with the help of trajectories.

3.5.2 Rank pooling

There is another way to model temporal information inside a video, termed rank pooling (a.k.a learning-to-rank). The seminal work in this line starts from VideoDarwin [53], that uses a ranking machine to learn the evolution of the appearance over time and returns a ranking function. The ranking function should be able to order the frames of a video temporally, thus they use the parameters of this ranking function as a new video representation. VideoDarwin [53] is not a deep learning based method, but achieves comparable performance and efficiency.

To adapt rank pooling to deep learning, Fernando [54] introduces a differentiable rank pooling layer to achieve end-to-end feature learning. Following this direction, Bilen *et al*. [9] apply rank pooling on the raw image pixels of a video producing a single RGB image per video, termed dynamic images. Another concurrent work by Fernando [51] extends rank pooling to hierarchical rank pooling by stacking multiple levels of temporal encoding. Finally, [22] propose a generalization of the original ranking formulation [53] using subspace representations and show that it leads to significantly better representation of the dynamic evolution of actions, while being computationally cheap.

3.5.3 Compressed video action recognition

Most video action recognition approaches use raw videos (or decoded video frames) as input. However, there are several drawbacks of using raw videos, such as the huge amount of data and high temporal redundancy. Video compression methods usually store one frame by reusing contents from another frame (i.e., I-frame) and only store the difference (i.e., P-frames and B-frames) due to the fact that adjacent frames are similar. Here, the I-frame is the original RGB video frame, and P-frames and B-frames include the motion vector and residual, which are used to store the difference. Motivated by the developments in the video compression domain, researchers started to adopt compressed video representations as input to train effective video models.

Since the motion vector has coarse structure and may contain inaccurate movements, Zhang *et al*. [256] adopted knowledge distillation to help the motion-vector-based temporal stream mimic the optical-flow-based temporal stream.

双流网络的结果表明:1.没有动作/时间建模,纯卷积(DeepVideo)的表现非常烂.2.传统方法中得到的知识(TDD的Trajectory Pooling,TLE的Global Feature Encoding)对于深度网络来说非常有助

3D卷积的结果表明:1.优化3D卷积核其实很难(C3D在大型的视频数据集上进行了训练,但是性能上却表现得比同时期的一些工作(主要是双流网络)要差), 因此最近有一些工作,例如I3D,P3D,R2+1D和S3D把3D卷积核分解为2D空间卷积核和1D的时间卷积核来简化训练

	Method	Pre-train	Flow	Backbone	Venue	UCF101	HMDB51	Kinetics400
双流网络	DeepVideo [99]	I	-	AlexNet	CVPR 2014	65.4	-	-
	Two-stream [187]	I	✓	CNN-M	NeurIPS 2014	88.0	59.4	-
	LRCN [37]	I	✓	CNN-M	CVPR 2015	82.3	-	-
	TDD [214]	I	✓	CNN-M	CVPR 2015	90.3	63.2	-
	Fusion [50]	I	✓	VGG16	CVPR 2016	92.5	65.4	-
	TSN [218]	I	✓	BN-Inception	ECCV 2016	94.0	68.5	**73.9**
	TLE [36]	I	✓	BN-Inception	CVPR 2017	**95.6**	**71.1**	-
3D卷积(或者把3D卷积分解为2D空间卷积和1D时间卷积).	C3D [202]	S	-	VGG16-like	ICCV 2015	82.3	56.8	59.5
	I3D [14]	I,K	-	BN-Inception-like	CVPR 2017	95.6	74.8	71.1
	P3D [169]	S	-	ResNet50-like	ICCV 2017	88.6	-	71.6
	ResNet3D [74]	K	-	ResNeXt101-like	CVPR 2018	94.5	70.2	65.1
	R2+1D [204]	K	-	ResNet34-like	CVPR 2018	96.8	74.5	72.0
	NL I3D [219]	I	-	ResNet101-like	CVPR 2018	-	-	77.7
	S3D [239]	I,K	-	BN-Inception-like	ECCV 2018	96.8	**75.9**	74.7
	SlowFast [45]	-	-	ResNet101-NL-like	ICCV 2019	-	-	79.8
	X3D-XXL [44]	-	-	ResNet-like	CVPR 2020	-	-	**80.4**
	TPN [248]	-	-	ResNet101-like	CVPR 2020	-	-	78.9
	CIDC [121]	-	-	ResNet50-like	ECCV 2020	**97.9**	75.2	75.5
关注与训练效率的方法	Hidden TSN [278]	I	-	BN-Inception	ACCV 2018	93.2	66.8	72.8
	OFF [197]	I	-	BN-Inception	CVPR 2018	96.0	74.2	-
	TSM [128]	I	-	ResNet50	ICCV 2019	95.9	73.5	74.1
	STM [92]	I,K	-	ResNet50-like	ICCV 2019	96.2	72.2	73.7
	TEINet [132]	I,K	-	ResNet50-like	AAAI 2020	96.7	72.1	76.2
	TEA [122]	I,K	-	ResNet50-like	CVPR 2020	**96.9**	73.3	76.1
	MSNet [110]	I,K	-	ResNet50-like	ECCV 2020	-	**77.4**	**76.4**

Table 2. **Results of widely adopted methods on three scene-focused datasets**. Pre-train indicates which dataset the model is pre-trained on. I: ImageNet, S: Sports1M and K: Kinetics400. NL represents non local.

However, their approach required extracting and processing each frame. They obtained comparable recognition accuracy with standard two-stream networks, but were 27 times faster. Wu *et al*. [231] used a heavyweight CNN for the I frame and lightweight CNN's for the P frames. This required that the motion vectors and residuals for each P frame be referred back to the I frame by accumulation. DMC-Net [185] is a follow-up work to [231] using adversarial loss. It adopts a lightweight generator network to help the motion vector capturing fine motion details, instead of knowledge distillation as in [256]. A recent paper SCSampler [106], also adopts compressed video representation for sampling salient clips and we will discuss it in the next section 3.5.4. As yet none of the compressed approaches can deal with B-frames due to the added complexity.

3.5.4 Frame/Clip sampling

Most of the aforementioned deep learning methods treat every video frame/clip equally for the final prediction. However, discriminative actions only happen in a few moments, and most of the other video content is irrelevant or weakly related to the labeled action category. There are several drawbacks of this paradigm. First, training with a large proportion of irrelevant video frames may hurt performance. Second, such uniform sampling is not efficient during inference.

Partially inspired by how human understand a video using just a few glimpses over the entire video [251], many methods were proposed to sample the most informative video frames/clips for both improving the performance and making the model more efficient during inference.

KVM [277] is one of the first attempts to propose an end-to-end framework to simultaneously identify key volumes and do action classification. Later, [98] introduce AdaScan that predicts the importance score of each video frame in an online fashion, which they term as adaptive temporal pooling. Both of these methods achieve improved performance, but they still adopt the standard evaluation scheme which does not show efficiency during inference. Recent approaches focus more on the efficiency [41, 234, 8, 106]. AdaFrame [234] follows [251, 98] but uses a reinforcement learning based approach to search more informative video clips. Concurrently, [8] uses a teacher-student framework, i.e., a see-it-all teacher can be used to train a compute efficient see-very-little student. They demonstrate that the efficient student network can reduce the inference time by 30% and the number of FLOPs by approximately 90% with negligible performance drop. Recently, SCSampler [106] trains a lightweight network to sample the most salient video clips based on compressed video representations, and achieve state-of-the-art performance on both Kinetics400 and Sports1M dataset. They also empirically show that such saliency-based sampling is not only efficient, but also enjoys higher accuracy than using all the video frames.

3.5.5 Visual tempo

Visual tempo is a concept to describe how fast an action goes. Many action classes have different visual tempos. In most cases, the key to distinguish them is their visual tempos, as they might share high similarities in visual appearance, such as walking, jogging and running [248]. There are several papers exploring different temporal rates (tempos) for improved temporal modeling [273, 147, 82, 281, 45, 248]. Initial attempts usually capture the video tempo through sampling raw videos at multiple rates and constructing an input-level frame pyramid [273, 147, 281]. Recently, SlowFast [45], as we discussed in section 3.3.4, utilizes the characteristics of visual tempo to design a two-pathway network for better accuracy and efficiency tradeoff. CIDC [121] proposed directional temporal modeling along with a local backbone for video temporal modeling. TPN [248] extends the tempo modeling to the feature-level and shows consistent improvement over previous approaches.

We would like to point out that visual tempo is also widely used in self-supervised video representation learning [6, 247, 16] since it can naturally provide supervision signals to train a deep network. We will discuss more details on self-supervised video representation learning in section 5.13.

本节我们在几个基准数据集上比较了流行的方法.更明确的说,

4. Evaluation and Benchmarking

我们首先介绍了几个标准的评价方法.然后我们把现有的数据集分成了三种:1.场景数据集 2.动作数据集 3.多标签数据集,我们再这三类数据及上对现有的了公平的方法的效率和精度进行比较

In this section, we compare popular approaches on benchmark datasets. To be specific, we first introduce standard evaluation schemes in section 4.1. Then we divide common benchmarks into three categories, scene-focused (UCF101, HMDB51 and Kinetics400 in section 4.2), motion-focused (Sth-Sth V1 and V2 in section 4.3) and multi-label (Charades in section 4.4). In the end, we present a fair comparison among popular methods in terms of both recognition accuracy and efficiency in section 4.5.

4.1. Evaluation scheme

During model training, we usually randomly pick a video frame/clip to form mini-batch samples. However, for evaluation, we need a standardized pipeline in order to perform fair comparisons.

在模型训练阶段,我们通常会随机选取视频帧或者视频的片段来形成mini-batch的example.但是在测试阶段,我们需要标准的pipeline来得到公平的比较

对于卷积网络而言,一个广泛采用的评价方案就是从每个视频中均匀抽取出来25帧形成测试样本.对于抽出来的每一帧,都会进行ten-crop数据增强,具体来说就是把一张图像的中心和四个角切出来,然后再进行水平反转.这样一个测试样本就由250帧图像组成.

对于3D卷积网络,一种广泛采用的评价方法就是从每个视频里面抽取得到10帧图像,然后短边放缩到256个像素,然后用三段来包括整个空间信息

For 2D CNNs, a widely adopted evaluation scheme is to evenly sample 25 frames from each video following [187, 217]. For each frame, we perform ten-crop data augmentation by cropping the 4 corners and 1 center, flipping them horizontally and averaging the prediction scores (before softmax operation) over all crops of the samples, i.e., this means we use 250 frames per video for inference.

For 3D CNNs, a widely adopted evaluation scheme termed 30-view strategy is to evenly sample 10 clips from each video following [219]. For each video clip, we perform three-crop data augmentation. To be specific, we scale the shorter spatial side to 256 pixels and take three crops of 256×256 to cover the spatial dimensions and average the prediction scores.

However, the evaluation schemes are not fixed. They are evolving and adapting to new network architectures and different datasets. For example, TSM [128] only uses two clips per video for small-sized datasets [190, 109], and perform three-crop data augmentation for each clip despite its being a 2D CNN. We will mention any deviations from the standard evaluation pipeline.

然而需要注意的是,评价方法并不是固定的,他们随着时间是在一直演化的,例如新的网络/数据集有新的评价方法,因此对于这些变体我们在下面会有额外的说明

In terms of evaluation metric, we report *accuracy* for single-label action recognition, and *mAP (mean average precision)* for multi-label action recognition.

对于评价的指标,对于单标签动作识别使用accuracy,多标签动作识别使用mAP作为指标

4.2. Scene-focused datasets

这部分我们在关注于场景的三个数据集:UCF101,HMDB51和Kinecti400上比较了最近流行的方法的性能.之所以称他们是场景数据集,是因为这些数据集中的视频是非常短的.并且通过周围的背景和语义信息就可以完成判断

Here, we compare recent state-of-the-art approaches on scene-focused datasets: UCF101, HMDB51 and Kinetics400. The reason we call them scene-focused is because most action videos in these datasets are short, and can be recognized by static scene appearance alone as shown in Figure 4.

Following the chronology, we first present results for early attempts of using deep learning and the two-stream networks at the top of Table 2. We make several observations. First, without motion/temporal modeling, the performance of DeepVideo [99] is inferior to all other approaches. Second, it is helpful to transfer knowledge from traditional methods (non-CNN-based) to deep learning. For example, TDD [214] uses trajectory pooling to extract motion-aware CNN features. TLE [36] embeds global feature encoding, which is an important step in traditional video action recognition pipeline, into a deep network.

按照编年表顺序,我们先给出了早期模型和双流网络的性能.结果表明:1.在没有动作/时间建模的话,DeepVideo的表现非常烂.2.传统方法中得到的知识对于深度网络来说非常有助

We then compare 3D CNNs based approaches in the middle of Table 2. Despite training on a large corpus of videos, C3D [202] performs inferior to concurrent work [187, 214, 217], possibly due to difficulties in optimization of 3D kernels. Motivated by this, several papers - I3D [14], P3D [169], R2+1D [204] and S3D [239] factorize 3D convolution filters to 2D spatial kernels and 1D temporal kernels to ease the training. In addition, I3D introduces an inflation strategy to avoid training from scratch by bootstrap-

我们接下来比较了3D卷积的方法.虽然C3D在大型的视频数据集上进行了训练,但是性能上却表现得比同时期的一些工作(主要是双流网络)要差,这就表明了优化3D卷积核其实很难.因此最近有一些工作,例如I3D,P3D,R2+1D和S3D把3D卷积核分解为2D空间卷积核和1D的时间卷积核来简化训练

Method	Pre-train	Backbone	Frames×Views	Venue	V1 Top1	V2 Top1
TSN [218]	I	BN-Inception	8×1	ECCV 2016	19.7	-
I3D [14]	I,K	ResNet50-like	32×6	CVPR 2017	41.6	-
NL I3D [219]	I,K	ResNet50-like	32×6	CVPR 2018	44.4	-
NL I3D + GCN [220]	I,K	ResNet50-like	32×6	ECCV 2018	46.1	-
ECO [283]	K	BNIncep+ResNet18	16×1	ECCV 2018	41.4	-
TRN [269]	I	BN-Inception	8×1	ECCV 2018	42.0	48.8
STM [92]	I	ResNet50-like	8×30	ICCV 2019	49.2	-
STM [92]	I	ResNet50-like	16×30	ICCV 2019	50.7	-
TSM [128]	K	ResNet50	8×1	ICCV 2019	45.6	59.1
TSM [128]	K	ResNet50	16×1	ICCV 2019	47.2	63.4
bLVNet-TAM [43]	I	BLNet-like	8×2	NeurIPS 2019	46.4	59.1
bLVNet-TAM [43]	I	BLNet-like	16×2	NeurIPS 2019	48.4	61.7
TEA [122]	I	ResNet50-like	8×1	CVPR 2020	48.9	-
TEA [122]	I	ResNet50-like	16×1	CVPR 2020	51.9	-
TSM + TPN [248]	K	ResNet50-like	8×1	CVPR 2020	49.0	62.0
MSNet [110]	I	ResNet50-like	8×1	ECCV 2020	50.9	63.0
MSNet [110]	I	ResNet50-like	16×1	ECCV 2020	**52.1**	**64.7**
TIN [182]	K	ResNet50-like	16×1	AAAI 2020	47.0	60.1
TEINet [132]	I	ResNet50-like	8×1	AAAI 2020	47.4	61.3
TEINet [132]	I	ResNet50-like	16×1	AAAI 2020	49.9	62.1

Table 3. **Results of widely adopted methods on Something-Something V1 and V2 datasets**. We only report numbers without using optical flow. Pre-train indicates which dataset the model is pre-trained on. I: ImageNet and K: Kinetics400. View means number of temporal clip multiples spatial crop, e.g., 30 means 10 temporal clips with 3 spatial crops each clip.

通过这些技巧，这些3D网络在没有使用光流网的前提下超越了双流网络的性能.此外，通过使用更多的训练样本，额外的网络路径和架构搜索.

此外，I3D中使用了膨胀策略以使用2D网络中训练好的参数来简化3D模型的训练，避免了从头训练一个网络

ping the 3D model weights from well-trained 2D networks. By using these techniques, they achieve comparable performance to the best two-stream network methods [36] without the need for optical flow. Furthermore, recent 3D models obtain even higher accuracy, by using more training samples [203], additional pathways [45], or architecture search [44].

最后，我们展示了最近的模型的性能.我们可以看到这些方法得到了比双流网络更高的精度，可以和3D卷积的精度相媲美

Finally, we show recent efficient models in the bottom of Table 2. We can see that these methods are able to achieve higher recognition accuracy than two-stream networks (top), and comparable performance to 3D CNNs (middle). Since they are 2D CNNs and do not use optical flow, these methods are efficient in terms of both training and inference. Most of them are real-time approaches, and some can do online video action recognition [128]. We believe 2D CNN plus temporal modeling is a promising direction due to the need of efficiency. Here, temporal modeling could be attention based, flow based or 3D kernel based.

我们认为，2D网络结合时间建模是非常有前景的方向，因为它非常高效.此外，时间建模的方法可以是注意力机制光流的方法或者是3D卷积的方法

4.3. Motion-focused datasets

In this section, we compare the recent state-of-the-art approaches on the 20BN-Something-Something (Sth-Sth) dataset. We report top1 accuracy on both V1 and V2. Sth-Sth datasets focus on humans performing basic actions with daily objects. Different from scene-focused datasets, background scene in Sth-Sth datasets contributes little to the final action class prediction. In addition, there are classes such as "Pushing something from left to right" and "Pushing something from right to left", and which require strong motion reasoning.

在本节，我们比较了最近的SOTA在SSV1和SSV2上的表现.我们报告了V1和V2上的top1准确率.和场景数据集不一样，动作数据集中的背景对动作识别的贡献非常的小.此外，数据集中还存在把东西从左移到右和从右移到左这两类动作，因此动作数据集要求模型有很强的动作理解能力

通过比较先前的工作在SSV2上的表现，我们发现：1.相同的网络使用更长的输入得到的性能更好. 2.使用时间卷积的模型比简单的3D模型更好

By comparing the previous work in Table 3, we observe that using longer input (e.g., 16 frames) is generally better. Moreover, methods that focus on temporal modeling [128, 122, 92] work better than stacked 3D kernels [14]. For example, TSM [128], TEA [122] and MSNet [110] insert an explicit temporal reasoning module into 2D ResNet backbones and achieves state-of-the-art results. This implies that the Sth-Sth dataset needs strong temporal motion reasoning as well as spatial semantics information.

4.4. Multi-label datasets

In this section, we first compare the recent state-of-the-art approaches on Charades dataset [186] and then we list some recent work that use assemble model or additional object information on Charades.

Comparing the previous work in Table 4, we make the following observations. First, 3D models [229, 45] generally perform better than 2D models [186, 231] and 2D models with optical flow input. This indicates that the spatio-temporal reasoning is critical for long-term complex concurrent action understanding. Second, longer input helps with the recognition [229] probably because some actions require long-term feature to recognize. Third, models with strong backbones that are pre-trained on larger datasets gen-

通过对先前文章的比较，我们得到如下的观察：1.3D模型通常表现的比2D模型要好，即便2D模型以光流作为输入. 2.使用更长的输入更有利于动作的辨识，可能是因为有一些动作本身就需要长的输入. 3.以在大型数据及上预训练的强大的模型作为backbone的方法通常表现的要好

Method	Extra-information	Backbone	Pre-train	Venue	mAP
2D CNN [186]	-	AlexNet	I	ECCV 2016	11.2
Two-stream [186]	flow	VGG16	I	ECCV 2016	22.4
ActionVLAD [63]	-	VGG16	I	CVPR 2017	21.0
CoViAR [231]	-	ResNet50-like	-	CVPR 2018	21.9
MultiScale TRN [269]	-	BN-Inception-like	I	ECCV 2018	**25.2**
I3D [14]	-	BN-Inception-like	K400	CVPR 2017	32.9
STRG [220]	-	ResNet101-NL-like	K400	ECCV 2018	39.7
LFB [229]	-	ResNet101-NL-like	K400	CVPR 2019	42.5
TC [84]		ResNet101-NL-like	K400	ICCV 2019	41.1
HAF [212]	IDT + flow	BN-Inception-like	K400	ICCV 2019	43.1
SlowFast [45]	-	ResNet-like	K400	ICCV 2019	42.5
SlowFast [45]	-	ResNet-like	K600	ICCV 2019	**45.2**
Action-Genome [90]	person + object	ResNet-like	-	CVPR 2020	**60.1**
AssembleNet++ [177]	flow + object	ResNet-like	-	ECCV 2020	59.9

Table 4. **Charades evaluation using mAP**, calculated using the officially provided script. NL: non-local network. Pre-train indicates which dataset the model is pre-trained on. I: ImageNet, K400: Kinetics400 and K600: Kinetics600.

此外，一些研究者最近研究了不同的方法来处理并发的动作识别，例如集成模型或者使用额外的人物交互信息

这些方法极大地超过了原有的方法，这表明了探索人物交互中的时空信息有助于进行并发动作理解

erally have better performance [45]. This is because Charades is a medium-scaled dataset which doesn't contain enough diversity to train a deep model.

Recently, researchers explored the alternative direction for complex concurrent action recognition by assembling models [177] or providing additional human-object interaction information [90]. These papers significantly outperformed previous literature that only finetune a single model on Charades. It demonstrates that exploring spatio-temporal human-object interactions and finding a way to avoid overfitting are the keys for concurrent action understanding.

4.5. Speed comparison

在真实世界中部署模型就要求模型的速度能够达到可接受的最小程度

因此在这一部分我们从四个角度进行了比较：1.参数量 2.FLOPS 3.latency 4.FPS

To deploy a model in real-life applications, we usually need to know whether it meets the speed requirement before we can proceed. In this section, we evaluate the approaches mentioned above to perform a thorough comparison in terms of (1) number of parameters, (2) FLOPS, (3) latency and (4) frame per second.

We present the results in Table 5. Here, we use the models in the GluonCV video action recognition model zoo[3] since all these models are trained using the same data, same data augmentation strategy and under same 30-view evaluation scheme, thus fair comparison. All the timings are done on a single Tesla V100 GPU with 105 repeated runs, while the first 5 runs are ignored for warming up. We always use a batch size of 1. In terms of model input, we use the suggested settings in the original paper.

As we can see in Table 5, if we compare latency, 2D models are much faster than all other 3D variants. This is

[3]To reproduce the numbers in Table 5, please visit `https://github.com/dmlc/gluon-cv/blob/master/scripts/action-recognition/README.md`

从延迟的角度来说，2D模型相比于3D模型可以更快的获得输出

其次，FLOPS与真实的处理延迟之间并不是强相关的

probably why most real-world video applications still adopt frame-wise methods. Secondly, as mentioned in [170, 259], FLOPS is not strongly correlated with the actual inference time (i.e., latency). Third, if comparing performance, most 3D models give similar accuracy around 75%, but pre-training with a larger dataset can significantly boost the performance[4]. This indicates the importance of training data and partially suggests that self-supervised pre-training might be a promising way to further improve existing methods.

大多数3D模型的性能都在75%左右，但是在大型的数据集上预训练可以极大地提升准确度

5. Discussion and Future Work

We have surveyed more than 200 deep learning based methods for video action recognition since year 2014. Despite the performance on benchmark datasets plateauing, there are many active and promising directions in this task worth exploring.

5.1. Analysis and insights

More and more methods haven been developed to improve video action recognition, at the same time, there are some papers summarizing these methods and providing analysis and insights. Huang *et al.* [82] perform an explicit analysis of the effect of temporal information for video understanding. They try to answer the question "how important is the motion in the video for recognizing the action". Feichtenhofer *et al.* [48, 49] provide an amazing visualization of what two-stream models have learned in order to understand how these deep representations work and what they are capturing. Li *et al.* [124] introduce a concept, representation bias of a dataset, and find that current datasets

[4]Note that, R2+1D-ResNet152∗ and CSN-ResNet152∗ in Table 5 are pretrained on a large-scale IG65M dataset [60].

Latency这里是延迟的意思

Model	Input	FLOPS (G)	# of params (M)	FPS	Latency (s)	Acc (%)
TSN-ResNet18 [218]	$3 \times 224 \times 224$	3.671	21.49	151.96	0.0066	69.85
TSN-ResNet34 [218]	$3 \times 224 \times 224$	1.819	11.382	264.01	0.0038	66.73
TSN-ResNet50 [218]	$3 \times 224 \times 224$	4.110	24.328	114.05	0.0088	70.88
TSN-ResNet101 [218]	$3 \times 224 \times 224$	7.833	43.320	59.56	0.0167	72.25
TSN-ResNet152 [218]	$3 \times 224 \times 224$	11.558	58.963	36.93	0.0271	72.45
I3D-ResNet50 [14]	$3 \times 32 \times 224 \times 224$	33.275	28.863	1719.50	0.0372	74.87
I3D-ResNet101 [14]	$3 \times 32 \times 224 \times 224$	51.864	52.574	1137.74	0.0563	75.10
I3D-ResNet50-NL [219]	$3 \times 32 \times 224 \times 224$	47.737	38.069	1403.16	0.0456	75.17
I3D-ResNet101-NL [219]	$3 \times 32 \times 224 \times 224$	66.326	61.780	999.94	0.0640	75.81
R2+1D-ResNet18 [204]	$3 \times 16 \times 112 \times 112$	40.645	31.505	804.31	0.0398	71.72
R2+1D-ResNet34 [204]	$3 \times 16 \times 112 \times 112$	75.400	61.832	503.17	0.0636	72.63
R2+1D-ResNet50 [204]	$3 \times 16 \times 112 \times 112$	65.543	53.950	667.06	0.0480	74.92
R2+1D-ResNet152∗ [204]	$3 \times 32 \times 112 \times 112$	252.900	118.227	546.19	0.1172	81.34
CSN-ResNet152∗ [203]	$3 \times 32 \times 224 \times 224$	74.758	29.704	435.77	0.1469	83.18
I3D-Slow-ResNet50 [45]	$3 \times 8 \times 224 \times 224$	41.919	32.454	1702.60	0.0376	74.41
I3D-Slow-ResNet50 [45]	$3 \times 16 \times 224 \times 224$	83.838	32.454	1406.00	0.0455	76.36
I3D-Slow-ResNet50 [45]	$3 \times 32 \times 224 \times 224$	167.675	32.454	860.74	0.0744	77.89
I3D-Slow-ResNet101 [45]	$3 \times 8 \times 224 \times 224$	85.675	60.359	1114.22	0.0574	76.15
I3D-Slow-ResNet101 [45]	$3 \times 16 \times 224 \times 224$	171.348	60.359	876.20	0.0730	77.11
I3D-Slow-ResNet101 [45]	$3 \times 32 \times 224 \times 224$	342.696	60.359	541.16	0.1183	78.57
SlowFast-ResNet50-4x16 [45]	$3 \times 32 \times 224 \times 224$	27.820	34.480	1396.45	0.0458	75.25
SlowFast-ResNet50-8x8 [45]	$3 \times 32 \times 224 \times 224$	50.583	34.566	1297.24	0.0493	76.66
SlowFast-ResNet101-8x8 [45]	$3 \times 32 \times 224 \times 224$	96.794	62.827	889.62	0.0719	76.95
TPN-ResNet50 [248]	$3 \times 8 \times 224 \times 224$	50.457	71.800	1350.39	0.0474	77.04
TPN-ResNet50 [248]	$3 \times 16 \times 224 \times 224$	99.929	71.800	1128.39	0.0567	77.33
TPN-ResNet50 [248]	$3 \times 32 \times 224 \times 224$	198.874	71.800	716.89	0.0893	78.90
TPN-ResNet101 [248]	$3 \times 8 \times 224 \times 224$	94.366	99.705	942.61	0.0679	78.10
TPN-ResNet101[248]	$3 \times 16 \times 224 \times 224$	187.594	99.705	754.00	0.0849	79.39
TPN-ResNet101[248]	$3 \times 32 \times 224 \times 224$	374.048	99.705	479.77	0.1334	79.70

带*的是在大型的IG65M上预训练的模型

Table 5. **Comparison on both efficiency and accuracy**. Top: 2D models and bottom: 3D models. FLOPS means floating point operations per second. FPS indicates how many video frames can the model process per second. Latency is the actual running time to complete one network forward given the input. Acc is the top-1 accuracy on Kinetics400 dataset. TSN, I3D, I3D-slow families are pretrained on ImageNet. R2+1D, SlowFast and TPN families are trained from scratch.

are biased towards static representations. Experiments on such biased datasets may lead to erroneous conclusions, which is indeed a big problem that limits the development of video action recognition. Recently, Piergiovanni *et al.* introduced the AViD [165] dataset to cope with data bias by collecting data from diverse groups of people. These papers provide great insights to help fellow researchers to understand the challenges, open problems and where the next breakthrough might reside.

5.2. Data augmentation

Numerous data augmentation methods have been proposed in image recognition domain, such as mixup [258], cutout [31], CutMix [254], AutoAugment [27], FastAutoAug [126], etc. However, video action recognition still adopts basic data augmentation techniques introduced before year 2015 [217, 188], including random resizing, random cropping and random horizontal flipping. Recently, SimCLR [17] and other papers have shown that color jittering and random rotation greatly help representation learning. Hence, an investigation of using different data augmentation techniques for video action recognition is particularly useful. This may change the data pre-processing pipeline for all existing methods.

5.3. Video domain adaptation

Domain adaptation (DA) has been studied extensively in recent years to address the domain shift problem. Despite the accuracy on standard datasets getting higher and higher, the generalization capability of current video models across datasets or domains is less explored.

There is early work on video domain adaptation [193, 241, 89, 159]. However, these literature focus on small-scale video DA with only a few overlapping categories, which may not reflect the actual domain discrepancy and may lead to biased conclusions. Chen *et al.* [15] introduce two larger-scale datasets to investigate video DA and find that aligning temporal dynamics is particularly useful. Pan *et al*. [152] adopts co-attention to solve the temporal misalignment problem. Very recently, Munro *et al*. [145] explore a multi-modal self-supervision method for fine-grained video action recognition and show the effectiveness of multi-modality learning in video DA. Shuffle and Attend [95] argues that aligning features of all sampled clips results in a sub-optimal solution due to the fact that all clips do not include relevant semantics. Therefore, they propose to use an attention mechanism to focus more on informative clips and discard the non-informative ones. In conclusion, video DA is a promising direction, especially for researchers with less computing resources.

5.4. Neural architecture search

Neural architecture search (NAS) has attracted great interest in recent years and is a promising research direction. However, given its greedy need for computing resources, only a few papers have been published in this area [156, 163, 161, 178]. The TVN family [161], which jointly optimize parameters and runtime, can achieve competitive accuracy with human-designed contemporary models, and run much faster (within 37 to 100 ms on a CPU and 10 ms on a GPU per 1 second video clip). AssembleNet [178] and AssembleNet++ [177] provide a generic approach to learn the connectivity among feature representations across input modalities, and show surprisingly good performance on Charades and other benchmarks. AttentionNAS [222] proposed a solution for spatio-temporal attention cell search. The found cell can be plugged into any network to improve the spatio-temporal features. All previous papers do show their potential for video understanding.

Recently, some efficient ways of searching architectures have been proposed in the image recognition domain, such as DARTS [130], Proxyless NAS [11], ENAS [160], one-shot NAS [7], etc. It would be interesting to combine efficient 2D CNNs and efficient searching algorithms to perform video NAS for a reasonable cost.

5.5. Efficient model development

Despite their accuracy, it is difficult to deploy deep learning based methods for video understanding problems in terms of real-world applications. There are several major challenges: (1) most methods are developed in offline settings, which means the input is a short video clip, not a video stream in an online setting; (2) most methods do not meet the real-time requirement; (3) incompatibility of 3D convolutions or other non-standard operators on non-GPU devices (e.g., edge devices).

Hence, the development of efficient network architecture based on 2D convolutions is a promising direction. The approaches proposed in the image classification domain can be easily adapted to video action recognition, e.g. model compression, model quantization, model pruning, distributed training [68, 127], mobile networks [80, 265], mixed precision training, etc. However, more effort is needed for the online setting since the input to most action recognition applications is a video stream, such as surveillance monitoring. We may need a new and more comprehensive dataset for benchmarking online video action recognition methods. Lastly, using compressed videos might be desirable because most videos are already compressed, and we have free access to motion information.

5.6. New datasets

Data is more or at least as important as model development for machine learning. For video action recognition, most datasets are biased towards spatial representations [124], i.e., most actions can be recognized by a single frame inside the video without considering the temporal movement. Hence, a new dataset in terms of long-term temporal modeling is required to advance video understanding. Furthermore, most current datasets are collected from YouTube. Due to copyright/privacy issues, the dataset organizer often only releases the YouTube id or video link for users to download and not the actual video. The first problem is that downloading the large-scale datasets might be slow for some regions. In particular, YouTube recently started to block massive downloading from a single IP. Thus, many researchers may not even get the dataset to start working in this field. The second problem is, due to region limitation and privacy issues, some videos are not accessible anymore. For example, the original Kinetcis400 dataset has over 300K videos, but at this moment, we can only crawl about 280K videos. On average, we lose 5% of the videos every year. It is impossible to do fair comparisons between methods when they are trained and evaluated on different data.

5.7. Video adversarial attack

Adversarial examples have been well studied on image models. [199] first shows that an adversarial sample, computed by inserting a small amount of noise on the original image, may lead to a wrong prediction. However, limited work has been done on attacking video models.

This task usually considers two settings, a white-box attack [86, 119, 66, 21] where the adversary can always get the full access to the model including exact gradients of a given input, or a black-box one [93, 245, 226], in which the structure and parameters of the model are blocked so that the attacker can only access the (input, output) pair through queries. Recent work ME-Sampler [260] leverages the motion information directly in generating adversarial videos, and is shown to successfully attack a number of video classification models using many fewer queries. In summary, this direction is useful since many companies provide APIs for services such as video classification, anomaly detection, shot detection, face detection, etc. In addition, this topic is also related to detecting DeepFake videos. Hence, investigating both attacking and defending methods is crucial to keeping these video services safe.

5.8. Zero-shot action recognition

Zero-shot learning (ZSL) has been trending in the image understanding domain, and has recently been adapted to video action recognition. Its goal is to transfer the learned knowledge to classify previously unseen categories. Due to (1) the expensive data sourcing and annotation and (2) the set of possible human actions is huge, zero-shot action recognition is a very useful task for real-world applications.

There are many early attempts [242, 88, 243, 137, 168, 57] in this direction. Most of them follow a standard framework, which is to first extract visual features from videos using a pretrained network, and then train a joint model that maps the visual embedding to a semantic embedding space. In this manner, the model can be applied to new classes by finding the test class whose embedding is the nearest-neighbor of the model's output. A recent work URL [279] proposes to learn a universal representation that generalizes across datasets. Following URL [279], [10] present the first end-to-end ZSL action recognition model. They also establish a new ZSL training and evaluation protocol, and provide an in-depth analysis to further advance this field. Inspired by the success of pre-training and then zero-shot in NLP domain, we believe ZSL action recognition is a promising research topic.

5.9. Weakly-supervised video action recognition

Building a high-quality video action recognition dataset [190, 100] usually requires multiple laborious steps: 1) first sourcing a large amount of raw videos, typically from the internet; 2) removing videos irrelevant to the categories in the dataset; 3) manually trimming the video segments that have actions of interest; 4) refining the categorical labels. Weakly-supervised action recognition explores how to reduce the cost for curating training data.

The first direction of research [19, 60, 58] aims to reduce the cost of sourcing videos and accurate categorical labeling. They design training methods that use training data such as action-related images or partially annotated videos, gathered from publicly available sources such as Internet. Thus this paradigm is also referred to as webly-supervised learning [19, 58]. Recent work on omni-supervised learning [60, 64, 38] also follows this paradigm but features bootstrapping on unlabelled videos by distilling the models' own inference results.

The second direction aims at removing trimming, the most time consuming part in annotation. UntrimmedNet [216] proposed a method to learn action recognition model on untrimmed videos with only categorical labels [149, 172]. This task is also related to weakly-supervised temporal action localization which aims to automatically generate the temporal span of the actions. Several papers propose to simultaneously [155] or iteratively [184] learn models for these two tasks.

5.10. Fine-grained video action recognition

Popular action recognition datasets, such as UCF101 [190] or Kinetics400 [100], mostly comprise actions happening in various scenes. However, models learned on these datasets could overfit to contextual information irrelevant to the actions [224, 227, 24]. Several datasets have been proposed to study the problem of fine-grained action recognition, which could examine the models' capacities in modeling action specific information. These datasets comprise fine-grained actions in human activities such as cooking [28, 108, 174], working [103] and sports [181, 124]. For example, FineGym [181] is a recent large dataset annotated with different moves and sub-actions in gymnastic videos.

5.11. Egocentric action recognition

Recently, large-scale egocentric action recognition [29, 28] has attracted increasing interest with the emerging of wearable cameras devices. Egocentric action recognition requires a fine understanding of hand motion and the interacting objects in the complex environment. A few papers leverage object detection features to offer fine object context to improve egocentric video recognition [136, 223, 229, 180]. Others incorporate spatio-temporal attention [192] or gaze annotations [131] to localize the interacting object to facilitate action recognition. Similar to third-person action recognition, multi-modal inputs (e.g., optical flow and audio) have been demonstrated to be effective in egocentric action recognition [101].

5.12. Multi-modality

Multi-modal video understanding has attracted increasing attention in recent years [55, 3, 129, 167, 154, 2, 105]. There are two main categories for multi-modal video understanding. The first group of approaches use multimodalities such as scene, object, motion, and audio to enrich the video representations. In the second group, the goal is to design a model which utilizes modality information as a supervision signal for pre-training models [195, 138, 249, 62, 2].

Multi-modality for comprehensive video understanding Learning a robust and comprehensive representation of video is extremely challenging due to the complexity

Method	Dataset	Video	Audio	Text	Size	Backbone	Venue	UCF101 Linear	UCF101 FT	HMDB51 Linear	HMDB51 FT
AVTS [105]	K400	✓	✓	–	224	R(2+1)D-18	NeurIPS 2018	–	86.2	–	52.3
AVTS [105]	AS	✓	✓	–	224	R(2+1)D-18	NeurIPS 2018	–	89.1	–	58.1
CBT [194]	K600+	✓	–	✓	112	S3D	arXiv 2019	54.0	79.5	29.5	44.6
MIL-NCE [138]	HTM	✓	–	✓	224	S3D	CVPR 2020	82.7	91.3	53.1	61.0
ELO [162]	YT8M	✓	✓	–	224	R(2+1)D-50	CVPR 2020	–	93.8	64.5	67.4
XDC [3]	K400	✓	✓	–	224	R(2+1)D-18	NeurIPS 2020	–	86.8	–	52.6
XDC [3]	AS	✓	✓	–	224	R(2+1)D-18	NeurIPS 2020	–	93.0	–	63.7
XDC [3]	IG65M	✓	✓	–	224	R(2+1)D-18	NeurIPS 2020	–	94.6	–	66.5
XDC [3]	IG-K	✓	✓	–	224	R(2+1)D-18	NeurIPS 2020	–	**95.5**	–	68.9
AVID [144]	AS	✓	✓	–	224	R(2+1)D-50	arXiv 2020	–	91.5	–	64.7
GDT [154]	K400	✓	✓	–	112	R(2+1)D-18	arXiv 2020	–	89.3	–	60.0
GDT [154]	AS	✓	✓	–	112	R(2+1)D-18	arXiv 2020	–	92.5	–	66.1
GDT [154]	IG65M	✓	✓	–	112	R(2+1)D-18	arXiv 2020	–	95.2	–	72.8
MMV [2]	AS+ HTM	✓	✓	✓	200	S3D	NeurIPS 2020	89.6	92.5	62.6	69.6
MMV [2]	AS+ HTM	✓	✓	✓	200	TSM-50x2	NeurIPS 2020	**91.8**	95.2	**67.1**	**75.0**
OPN [115]	UCF101	✓	–	–	227	VGG	ICCV 2017	–	59.6	–	23.8
3D-RotNet [94]	K400	✓	–	–	112	R3D	arXiv 2018	–	62.9	–	33.7
ST-Puzzle [102]	K400	✓	–	–	224	R3D	AAAI 2019	–	63.9	–	33.7
VCOP [240]	UCF101	✓	–	–	112	R(2+1)D	CVPR 2019	–	72.4	–	30.9
DPC [71]	K400	✓	–	–	128	R-2D3D	ICCVW 2019	–	75.7	–	35.7
SpeedNet [6]	K400	✓	–	–	224	S3D-G	CVPR 2020	–	81.1	–	48.8
MemDPC [72]	K400	✓	–	–	224	R-2D3D	ECCV 2020	54.1	86.1	30.5	54.5
CoCLR [73]	K400	✓	–	–	128	S3D	NeurIPS 2020	**74.5**	87.9	**46.1**	54.6
CVRL [167]	K400	✓	–	–	224	R3D-50	arXiv 2020	–	92.2	–	66.7
CVRL [167]	K600	✓	–	–	224	R3D-50	arXiv 2020	–	**93.4**	–	**68.0**

Table 6. **Comparison of self-supervised video representation learning methods.** Top section shows the multi-modal video representation learning approaches and bottom section shows the video-only representation learning methods. From left to right, we show the self-supervised training setting, e.g. dataset, modalities, resolution, and architecture. Two last right columns show the action recognition results on two datasets UCF101 and HMDB51 to measure the quality of self-supervised pre-trained model. HTM: HowTo100M, YT8M: YouTube8M, AS: AudioSet, IG-K: IG-Kinetics, K400: Kinetics400 and K600: Kinetics600.

of semantics in videos. Video data often includes variations in different forms including appearance, motion, audio, text or scene [55, 129, 166]. Therefore, utilizing these multi-modal representations is a critical step in understanding video content more efficiently. The multi-modal representations of video can be approximated by gathering representations of various modalities such as scene, object, audio, motion, appearance and text. Ngiam *et al*. [148] was an early attempt to suggest using multiple modalities to obtain better features. They utilized videos of lips and their corresponding speech for multi-modal representation learning. Miech *et al*. [139] proposed a mixture-of embedding-experts model to combine multiple modalities including motion, appearance, audio, and face features and learn the shared embedding space between these modalities and text. Roig *et al*. [175] combines multiple modalities such as action, scene, object and acoustic event features in a pyramidal structure for action recognition. They show that adding each modality improves the final action recognition accuracy. Both CE [129] and MMT [55], follow a similar research line to [139] where the goal is to combine multiple-modalities to obtain a comprehensive representation of video for joint video-text representation learning. Piergiovanni *et al*. [166] utilized textual data together with video data to learn a joint embedding space. Using this learned joint embedding space, the method is capable of doing zero-shot action recognition. This line of research is promising due to the availability of strong semantic extraction models and also success of transformers on both vision and language tasks.

Multi-modality for self-supervised video representation learning Most videos contain multiple modalities such as audio or text/caption. These modalities are great source of supervision for learning video representations [3, 144, 154, 2, 162]. Korbar *et al*. [105] incorporated the natural synchronization between audio and video as a supervision signal in their contrastive learning objective for self-supervised representation learning. In multi-modal self-supervised representation learning, the dataset plays an important role. VideoBERT [195] collected 310K cooking videos from YouTube. However, this dataset is not publicly available. Similar to BERT, VideoBERT used a "masked language model" training objective and also quantized the visual representations into "visual words". Miech *et al.* [140] introduced HowTo100M dataset in 2019. This dataset includes 136M clips from 1.22M videos with their corresponding text. They collected the dataset from YouTube with the aim of obtaining instructional videos (how to perform an activity). In total, it covers 23.6K instructional tasks. MIL-NCE [138] used this dataset for self-supervised cross-modal representation learning. They tackled the problem of visually misaligned narrations, by considering multiple positive pairs in the contrastive learning objective. ActBERT [275], utilized HowTo100M dataset for pre-training of the model in a self-supervised way. They incorporated visual, action, text and object features for cross modal representation learning. Recently AVLnet [176] and MMV [2] considered three modalities visual, audio and language for self-supervised representation learning. This research direction is also increasingly getting more attention due to the success of contrastive learning on many vision and language tasks and the access to the abundance of unlabeled multimodal video data on platforms such as YouTube, Instagram or Flickr. The top section of Table 6 compares multi-modal self-supervised representation learning methods. We will discuss more work on video-only representation learning in the next section.

5.13. Self-supervised video representation learning

Self-supervised learning has attracted more attention recently as it is able to leverage a large amount of unlabeled data by designing a pretext task to obtain free supervisory signals from data itself. It first emerged in image representation learning. On images, the first stream of papers aimed at designing pretext tasks for completing missing information, such as image coloring [262] and image reordering [153, 61, 263]. The second stream of papers uses instance discrimination [235] as the pretext task and contrastive losses [235, 151] for supervision. They learn visual representation by modeling visual similarity of object instances without class labels [235, 75, 201, 18, 17].

Self-supervised learning is also viable for videos. Compared with images, videos has another axis, temporal dimension, which we can use to craft pretext tasks. Information completion tasks for this purpose include predicting the correct order of shuffled frames [141, 52] and video clips [240]. Jing *et al.* [94] focus on the spatial dimension only by predicting the rotation angles of rotated video clips. Combining temporal and spatial information, several tasks have been introduced to solve a space-time cubic puzzle, anticipate future frames [208], forecast long-term motions [134] and predict motion and appearance statistics [211]. RSPNet [16] and visual tempo [247] exploit the relative speed between video clips as a supervision signal.

The added temporal axis can also provide flexibility in designing instance discrimination pretexts [67, 167]. Inspired by the decoupling of 3D convolution to spatial and temporal separable convolutions [239], Zhang *et al.* [266] proposed to decouple the video representation learning into two sub-tasks: spatial contrast and temporal contrast. Recently, Han *et al.* [72] proposed memory augmented dense predictive coding for self-supervised video representation learning. They split each video into several blocks and the embedding of future block is predicted by the combination of condensed representations in memory.

The temporal continuity in videos inspires researchers to design other pretext tasks around correspondence. Wang *et al.* [221] proposed to learn representation by performing cycle-consistency tracking. Specifically, they track the same object backward and then forward in the consecutive video frames, and use the inconsistency between the start and end points as the loss function. TCC [39] is a concurrent paper. Instead of tracking local objects, [39] used cycle-consistency to perform frame-wise temporal alignment as a supervision signal. [120] was a follow-up work of [221], and utilized both object-level and pixel-level correspondence across video frames. Recently, long-range temporal correspondence is modelled as a random walk graph to help learning video representation in [87].

We compare video self-supervised representation learning methods at the bottom section of Table 6. A clear trend can be observed that recent papers have achieved much better linear evaluation accuracy and fine-tuning accuracy comparable to supervised pre-training. This shows that self-supervised learning could be a promising direction towards learning better video representations.

6. Conclusion

In this survey, we present a comprehensive review of 200+ deep learning based recent approaches to video action recognition. Although this is not an exhaustive list, we hope the survey serves as an easy-to-follow tutorial for those seeking to enter the field, and an inspiring discussion for those seeking to find new research directions.

本文，我们对200多篇视频动作识别的文章进行了综述．虽然文章很多，但是我们希望这篇文章读起来不是很难懂，并且希望这个综述可以作为一个很容易上手的新手教程（对于刚入门的人），或者是一个非常有启发的讨论（对于寻找新的研究方向的这个领域的研究者）

Acknowledgement

We would like to thank Peter Gehler, Linchao Zhu and Thomas Brady for constructive feedback and fruitful discussions.

References

[1] Sami Abu-El-Haija, Nisarg Kothari, Joonseok Lee, Paul Natsev, George Toderici, Balakrishnan Varadarajan, and

Sudheendra Vijayanarasimhan. YouTube-8M: A Large-Scale Video Classification Benchmark. *arXiv preprint arXiv:1609.08675*, 2016.

[2] Jean-Baptiste Alayrac, Adrià Recasens, Rosalia Schneider, Relja Arandjelović, Jason Ramapuram, Jeffrey De Fauw, Lucas Smaira, Sander Dieleman, and Andrew Zisserman. Self-Supervised MultiModal Versatile Networks, 2020.

[3] Humam Alwassel, Dhruv Mahajan, Bruno Korbar, Lorenzo Torresani, Bernard Ghanem, and Du Tran. Self-Supervised Learning by Cross-Modal Audio-Video Clustering. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

[4] Relja Arandjelović, Petr Gronat, Akihiko Torii, Tomas Pajdla, and Josef Sivic. NetVLAD: CNN Architecture for Weakly Supervised Place Recognition. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[5] Moez Baccouche, Franck Mamalet, Christian Wolf, Christophe Garcia, and Atilla Baskurt. Sequential Deep Learning for Human Action Recognition. In *the Second International Conference on Human Behavior Understanding*, 2011.

[6] Sagie Benaim, Ariel Ephrat, Oran Lang, Inbar Mosseri, William T. Freeman, Michael Rubinstein, Michal Irani, and Tali Dekel. SpeedNet: Learning the Speediness in Videos. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[7] Gabriel Bender, Pieter-Jan Kindermans, Barret Zoph, Vijay Vasudevan, and Quoc Le. Understanding and Simplifying One-Shot Architecture Search. In *The International Conference on Machine Learning (ICML)*, 2018.

[8] Shweta Bhardwaj, Mukundhan Srinivasan, and Mitesh M. Khapra. Efficient Video Classification Using Fewer Frames. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[9] Hakan Bilen, Basura Fernando, Efstratios Gavves, Andrea Vedaldi, and Stephen Gould. Dynamic Image Networks for Action Recognition. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[10] Biagio Brattoli, Joseph Tighe, Fedor Zhdanov, Pietro Perona, and Krzysztof Chalupka. Rethinking Zero-Shot Video Classification: End-to-End Training for Realistic Applications. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[11] Han Cai, Ligeng Zhu, and Song Han. ProxylessNAS: Direct Neural Architecture Search on Target Task and Hardware. In *The International Conference on Learning Representations (ICLR)*, 2019.

[12] Joao Carreira, Eric Noland, Andras Banki-Horvath, Chloe Hillier, and Andrew Zisserman. A short note about kinetics-600. *arXiv preprint arXiv:1808.01340*, 2018.

[13] Joao Carreira, Eric Noland, Chloe Hillier, and Andrew Zisserman. A short note on the kinetics-700 human action dataset. *arXiv preprint arXiv:1907.06987*, 2019.

[14] Joao Carreira and Andrew Zisserman. Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[15] Min-Hung Chen, Zsolt Kira, Ghassan AlRegib, Jaekwon Yoo, Ruxin Chen, and Jian Zheng. Temporal Attentive Alignment for Large-Scale Video Domain Adaptation. In *The IEEE International Conference on Computer Vision (ICCV)*, 2019.

[16] Peihao Chen, Deng Huang, Dongliang He, Xiang Long, Runhao Zeng, Shilei Wen, Mingkui Tan, and Chuang Gan. RSPNet: Relative Speed Perception for Unsupervised Video Representation Learning, 2020.

[17] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A Simple Framework for Contrastive Learning of Visual Representations. *arXiv preprint arXiv:2002.05709*, 2020.

[18] Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. Improved Baselines with Momentum Contrastive Learning. *arXiv preprint arXiv:2003.04297*, 2020.

[19] Xinlei Chen and Abhinav Gupta. Webly Supervised Learning of Convolutional Networks. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 1431–1439, 2015.

[20] Yunpeng Chen, Yannis Kalantidis, Jianshu Li, Shuicheng Yan, and Jiashi Feng. Multi-Fiber Networks for Video Recognition. In *The European Conference on Computer Vision (ECCV)*, 2018.

[21] Zhikai Chen, Lingxi Xie, Shanmin Pang, Yong He, and Qi Tian. Appending Adversarial Frames for Universal Video Attack. *arXiv preprint arXiv:1912.04538*, 2019.

[22] Anoop Cherian, Basura Fernando, Mehrtash Harandi, and Stephen Gould. Generalized Rank Pooling for Activity Recognition. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[23] Guilhem Cheron, Ivan Laptev, and Cordelia Schmid. P-CNN: Pose-based CNN Features for Action Recognition. In *The IEEE International Conference on Computer Vision (ICCV)*, 2015.

[24] Jinwoo Choi, Chen Gao, Joseph CE Messou, and Jia-Bin Huang. Why Can't I Dance in the Mall? Learning to Mitigate Scene Bias in Action Recognition. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 853–865, 2019.

[25] Vasileios Choutas, Philippe Weinzaepfel, Jérôme Revaud, and Cordelia Schmid. PoTion: Pose MoTion Representation for Action Recognition. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[26] Nieves Crasto, Philippe Weinzaepfel, Karteek Alahari, and Cordelia Schmid. MARS: Motion-Augmented RGB Stream for Action Recognition. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[27] Ekin D. Cubuk, Barret Zoph, Dandelion Mane, Vijay Vasudevan, and Quoc V. Le. AutoAugment: Learning Augmentation Strategies From Data. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[28] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Sanja Fidler, Antonino Furnari, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, and Michael Wray. The EPIC-KITCHENS Dataset: Col-

lection, Challenges and Baselines. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2020.

[29] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Antonino Furnari, Evangelos Kazakos, Jian Ma, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, et al. Rescaling Egocentric Vision. *arXiv preprint arXiv:2006.13256*, 2020.

[30] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR*, 2009.

[31] Terrance DeVries and Graham W Taylor. Improved Regularization of Convolutional Neural Networks with Cutout. *arXiv preprint arXiv:1708.04552*, 2017.

[32] Ali Diba, Mohsen Fayyaz, Vivek Sharma, Amir Hossein Karami, Mohammad Mahdi Arzani, Rahman Yousefzadeh, and Luc Van Gool. Temporal 3D ConvNets: New Architecture and Transfer Learning for Video Classification. *arXiv preprint arXiv:1711.08200*, 2017.

[33] Ali Diba, Mohsen Fayyaz, Vivek Sharma, M. Mahdi Arzani, Rahman Yousefzadeh, Juergen Gall, and Luc Van Gool. Spatio-Temporal Channel Correlation Networks for Action Classification. In *The European Conference on Computer Vision (ECCV)*, 2018.

[34] Ali Diba, Mohsen Fayyaz, Vivek Sharma, Manohar Paluri, Jürgen Gall, Rainer Stiefelhagen, and Luc Van Gool. Large Scale Holistic Video Understanding. In *European Conference on Computer Vision*, pages 593–610. Springer, 2020.

[35] Ali Diba, Ali Mohammad Pazandeh, and Luc Van Gool. Efficient Two-Stream Motion and Appearance 3D CNNs for Video Classification. *arXiv preprint arXiv:1608.08851*, 2016.

[36] Ali Diba, Vivek Sharma, and Luc Van Gool. Deep Temporal Linear Encoding Networks. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[37] Jeff Donahue, Lisa Anne Hendricks, Marcus Rohrbach, Subhashini Venugopalan, Sergio Guadarrama, Kate Saenko, and Trevor Darrell. Long-term Recurrent Convolutional Networks for Visual Recognition and Description. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[38] Hao Dong Duan, Yue Zhao, Yuanjun Xiong, Wentao Liu, and Dahu Lin. Omni-sourced Webly-supervised Learning for Video Recognition. In *European Conference on Computer Vision*, 2020.

[39] Debidatta Dwibedi, Yusuf Aytar, Jonathan Tompson, Pierre Sermanet, and Andrew Zisserman. Temporal Cycle-Consistency Learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1801–1810, 2019.

[40] Bernard Ghanem Fabian Caba Heilbron, Victor Escorcia and Juan Carlos Niebles. ActivityNet: A Large-Scale Video Benchmark for Human Activity Understanding. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[41] Hehe Fan, Zhongwen Xu, Linchao Zhu, Chenggang Yan, Jianjun Ge, and Yi Yang. Watching a small portion could be as good as watching all: Towards efficient video classification. In *IJCAI International Joint Conference on Artificial Intelligence*, 2018.

[42] Lijie Fan, Wenbing Huang, Chuang Gan, Stefano Ermon, Boqing Gong, and Junzhou Huang. End-to-End Learning of Motion Representation for Video Understanding. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[43] Quanfu Fan, Chun-Fu (Richard) Chen, Hilde Kuehne, Marco Pistoia, and David Cox. More Is Less: Learning Efficient Video Representations by Big-Little Network and Depthwise Temporal Aggregation. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

[44] Christoph Feichtenhofer. X3D: Expanding Architectures for Efficient Video Recognition. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[45] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. SlowFast Networks for Video Recognition. In *The IEEE International Conference on Computer Vision (ICCV)*, 2019.

[46] Christoph Feichtenhofer, Axel Pinz, and Richard P. Wildes. Spatiotemporal Residual Networks for Video Action Recognition. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2016.

[47] Christoph Feichtenhofer, Axel Pinz, and Richard P Wildes. Spatiotemporal Multiplier Networks for Video Action Recognition. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[48] Christoph Feichtenhofer, Axel Pinz, Richard P. Wildes, and Andrew Zisserman. What Have We Learned From Deep Representations for Action Recognition? In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[49] Christoph Feichtenhofer, Axel Pinz, Richard P. Wildes, and Andrew Zisserman. Deep insights into convolutional networks for video recognition. *International Journal of Computer Vision (IJCV)*, 2019.

[50] Christoph Feichtenhofer, Axel Pinz, and Andrew Zisserman. Convolutional Two-Stream Network Fusion for Video Action Recognition. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[51] Basura Fernando, Peter Anderson, Marcus Hutter, and Stephen Gould. Discriminative Hierarchical Rank Pooling for Activity Recognition. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[52] Basura Fernando, Hakan Bilen, Efstratios Gavves, and Stephen Gould. Self-supervised video representation learning with odd-one-out networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3636–3645, 2017.

[53] Basura Fernando, Efstratios Gavves, Jose Oramas M., Amir Ghodrati, and Tinne Tuytelaars. Modeling Video Evolution For Action Recognition. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[54] Basura Fernando and Stephen Gould. Learning End-to-end Video Classification with Rank-Pooling. In *The International Conference on Machine Learning (ICML)*, 2016.

[55] Gabeur et al. Multi-modal Transformer for Video Retrieval. *arxiv:2007.10639*, 2020.

[56] Harshala Gammulle, Simon Denman, Sridha Sridharan, and Clinton Fookes. Two Stream LSTM: A Deep Fusion Framework for Human Action Recognition. In *The IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2017.

[57] Chuang Gan, Ming Lin, Yi Yang, Yueting Zhuang, and Alexander G.Hauptmann. Exploring Semantic Inter-Class Relationships (SIR) for Zero-Shot Action Recognition. In *AAAI*, 2015.

[58] Chuang Gan, Chen Sun, Lixin Duan, and Boqing Gong. Webly-supervised video recognition by mutually voting for relevant web images and web video frames. In *European Conference on Computer Vision*, pages 849–866. Springer, 2016.

[59] Nuno C. Garcia, Pietro Morerio, and Vittorio Murino. Modality Distillation with Multiple Stream Networks for Action Recognition. In *The European Conference on Computer Vision (ECCV)*, 2018.

[60] Deepti Ghadiyaram, Matt Feiszli, Du Tran, Xueting Yan, Heng Wang, and D. Mahajan. Large-Scale Weakly-Supervised Pre-Training for Video Action Recognition. *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[61] Spyros Gidaris, Praveer Singh, and Nikos Komodakis. Unsupervised representation learning by predicting image rotations. *arXiv preprint arXiv:1803.07728*, 2018.

[62] Simon Ging, Mohammadreza Zolfaghari, Hamed Pirsiavash, and Thomas Brox. Coot: Cooperative hierarchical transformer for video-text representation learning. In *Advances in Neural Information Processing Systems*, 2020.

[63] Rohit Girdhar, Deva Ramanan, Abhinav Gupta, Josef Sivic, and Bryan Russell. ActionVLAD: Learning Spatio-Temporal Aggregation for Action Classification. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[64] Rohit Girdhar, Du Tran, Lorenzo Torresani, and Deva Ramanan. Distinit: Learning video representations without a single labeled video. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 852–861, 2019.

[65] M. A. Goodale and A. D. Milner. Separate Visual Pathways for Perception and Action. *Trends in Neurosciences*, 1992.

[66] Ian Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In *International Conference on Learning Representations*, 2015.

[67] Daniel Gordon, Kiana Ehsani, Dieter Fox, and Ali Farhadi. Watching the World Go By: Representation Learning from Unlabeled Videos. *arXiv preprint arXiv:2003.07990*, 2020.

[68] Priya Goyal, Piotr Dollár, Ross Girshick, Pieter Noordhuis, Lukasz Wesolowski, Aapo Kyrola, Andrew Tulloch, Yangqing Jia, and Kaiming He. Accurate, Large Minibatch SGD: Training ImageNet in 1 Hour. *arXiv preprint arXiv:1706.02677*, 2017.

[69] Raghav Goyal, Samira Ebrahimi Kahou, Vincent Michalski, Joanna Materzynska, Susanne Westphal, Heuna Kim, Valentin Haenel, Ingo Fruend, Peter Yianilos, Moritz Mueller-Freitag, et al. The" Something Something" Video Database for Learning and Evaluating Visual Common Sense. In *The IEEE International Conference on Computer Vision (ICCV)*, 2017.

[70] Chunhui Gu, Chen Sun, David A. Ross, Carl Vondrick, Caroline Pantofaru, Yeqing Li, Sudheendra Vijayanarasimhan, George Toderici, Susanna Ricco, Rahul Sukthankar, Cordelia Schmid, and Jitendra Malik. AVA: A Video Dataset of Spatio-Temporally Localized Atomic Visual Actions. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[71] Tengda Han, Weidi Xie, and Andrew Zisserman. Video representation learning by dense predictive coding. In *Workshop on Large Scale Holistic Video Understanding, ICCV*, 2019.

[72] Tengda Han, Weidi Xie, and Andrew Zisserman. Memory-augmented dense predictive coding for video representation learning. In *European Conference on Computer Vision*, 2020.

[73] Tengda Han, Weidi Xie, and Andrew Zisserman. Self-supervised co-training for video representation learning. In *Neurips*, 2020.

[74] Kensho Hara, Hirokatsu Kataoka, and Yutaka Satoh. Can Spatiotemporal 3D CNNs Retrace the History of 2D CNNs and ImageNet? In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[75] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. *arXiv preprint arXiv:1911.05722*, 2019.

[76] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[77] Samitha Herath, Mehrtash Harandi, and Fatih Porikli. Going Deeper into Action Recognition: A Survey. *arXiv preprint arXiv:1605.04988*, 2016.

[78] Sepp Hochreiter and Jürgen Schmidhuber. Long Short-Term Memory. *Neural Computation*, 1997.

[79] Berthold K.P. Horn and Brian G. Rhunck. Determining Optical Flow. *Artificial Intelligence*, 1981.

[80] Andrew Howard, Mark Sandler, Grace Chu, Liang-Chieh Chen, Bo Chen, Mingxing Tan, Weijun Wang, Yukun Zhu, Ruoming Pang, Vijay Vasudevan, Quoc V. Le, and Hartwig Adam. Searching for MobileNetV3. In *The IEEE International Conference on Computer Vision (ICCV)*, 2019.

[81] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-Excitation Networks. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[82] De-An Huang, Vignesh Ramanathan, Dhruv Mahajan, Lorenzo Torresani, Manohar Paluri, Li Fei-Fei, and Juan Carlos Niebles. What Makes a Video a Video: Analyzing Temporal Information in Video Understanding Models and Datasets. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[83] Gao Huang, Zhuang Liu, Laurens van der Maaten, and Kilian Q. Weinberger. Densely Connected Convolutional Networks. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[84] Noureldien Hussein, Efstratios Gavves, and Arnold WM Smeulders. Timeception for complex action recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 254–263, 2019.

[85] E. Ilg, N. Mayer, T. Saikia, M. Keuper, A. Dosovitskiy, and T. Brox. FlowNet 2.0: Evolution of Optical Flow Estimation with Deep Networks. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[86] Nathan Inkawhich, Matthew Inkawhich, Yiran Chen, and Hai Li. Adversarial attacks for optical flow-based action recognition classifiers. *arXiv preprint arXiv:1811.11875*, 2018.

[87] Allan Jabri, Andrew Owens, and Alexei A Efros. Space-time correspondence as a contrastive random walk. *Advances in Neural Information Processing Systems*, 2020.

[88] Mihir Jain, Jan C van Gemert, Thomas Mensink, and Cees GM Snoek. Objects2action: Classifying and Localizing Actions without Any Video Example. In *ICCV*, 2015.

[89] Arshad Jamal, Vinay P Namboodiri, Dipti Deodhare, and KS Venkatesh. Deep Domain Adaptation in Action Space. In *The British Machine Vision Conference (BMVC)*, 2018.

[90] Jingwei Ji, Ranjay Krishna, Li Fei-Fei, and Juan Carlos Niebles. Action genome: Actions as compositions of spatio-temporal scene graphs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10236–10247, 2020.

[91] Shuiwang Ji, Wei Xu, Ming Yang, and Kai Yu. 3D Convolutional Neural Networks for Human Action Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 2012.

[92] Boyuan Jiang, MengMeng Wang, Weihao Gan, Wei Wu, and Junjie Yan. STM: SpatioTemporal and Motion Encoding for Action Recognition. In *The IEEE International Conference on Computer Vision (ICCV)*, 2019.

[93] Linxi Jiang, Xingjun Ma, Shaoxiang Chen, James Bailey, and Yu-Gang Jiang. Black-box adversarial attacks on video recognition models. In *Proceedings of the 27th ACM International Conference on Multimedia*, pages 864–872, 2019.

[94] Longlong Jing, Xiaodong Yang, Jingen Liu, and Yingli Tian. Self-supervised spatiotemporal feature learning via video rotation prediction. *arXiv preprint arXiv:1811.11387*, 2018.

[95] Samuel Schulter Jinwoo Choi, Gaurav Sharma and Jia-Bin Huang. Shuffle and Attend: Video Domain Adaptation. In *The European Conference on Computer Vision (ECCV)*, 2020.

[96] Valter Luís Estevam Junior, Helio Pedrini, and David Menotti. Zero-Shot Action Recognition in Videos: A Survey. *arXiv preprint arXiv:1909.06423*, 2019.

[97] Hervé Jégou, Matthijs Douze, Cordelia Schmid, and Patrick Pérez. Aggregating Local Descriptors into a Compact Image Representation. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2010.

[98] Amlan Kar, Nishant Rai, Karan Sikka, and Gaurav Sharma. AdaScan: Adaptive Scan Pooling in Deep Convolutional Neural Networks for Human Action Recognition in Videos. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[99] Andrej Karpathy, George Toderici, Sanketh Shetty, Thomas Leung, Rahul Sukthankar, and Li Fei-Fei. Large-Scale Video Classification with Convolutional Neural Networks. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014.

[100] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, Mustafa Suleyman, and Andrew Zisserman. The Kinetics Human Action Video Dataset. *arXiv preprint arXiv:1705.06950*, 2017.

[101] Evangelos Kazakos, Arsha Nagrani, Andrew Zisserman, and Dima Damen. EPIC-Fusion: Audio-Visual Temporal Binding for Egocentric Action Recognition. In *ICCV*, 2019.

[102] Dahun Kim, Donghyeon Cho, and In So Kweon. Self-Supervised Video Representation Learning with Space-Time Cubic Puzzles. In *AAAI*, 2019.

[103] Takuya Kobayashi, Yoshimitsu Aoki, Shogo Shimizu, Katsuhiro Kusano, and Seiji Okumura. Fine-grained action recognition in assembly work scenes by drawing attention to the hands. In *2019 15th International Conference on Signal-Image Technology & Internet-Based Systems (SITIS)*, pages 440–446. IEEE, 2019.

[104] Yu Kong and Yun Fu. Human Action Recognition and Prediction: A Survey. *arXiv preprint arXiv:1806.11230*, 2018.

[105] Bruno Korbar, Du Tran, and Lorenzo Torresani. Cooperative learning of audio and video models from self-supervised synchronization, 2018.

[106] Bruno Korbar, Du Tran, and Lorenzo Torresani. SCSampler: Sampling Salient Clips From Video for Efficient Action Recognition. In *The IEEE International Conference on Computer Vision (ICCV)*, 2019.

[107] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. ImageNet Classification with Deep Convolutional Neural Networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2012.

[108] Hilde Kuehne, Ali Arslan, and Thomas Serre. The language of actions: Recovering the syntax and semantics of goal-directed human activities. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 780–787, 2014.

[109] Hildegard Kuehne, Hueihan Jhuang, Estíbaliz Garrote, Tomaso Poggio, and Thomas Serre. HMDB: A Large Video Database for Human Motion Recognition. In *The IEEE International Conference on Computer Vision (ICCV)*, 2011.

[110] Heeseung Kwon, Manjin Kim, Suha Kwak, and Minsu Cho. Motionsqueeze: Neural motion feature learning for video understanding. In *ECCV*, 2020.

[111] Okan Köpüklü, Neslihan Kose, Ahmet Gunduz, and Gerhard Rigoll. Resource Efficient 3D Convolutional Neural Networks. In *The IEEE International Conference on Computer Vision (ICCV) Workshop*, 2019.

[112] Zhenzhong Lan, Ming Lin, Xuanchong Li, Alexander G. Hauptmann, and Bhiksha Raj. Beyond Gaussian Pyramid:

Multi-skip Feature Stacking for Action Recognition. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[113] Zhenzhong Lan, Dezhong Yao, Ming Lin, Shoou-I Yu, and Alexander Hauptmann. The Best of Both Worlds: Combining Data-independent and Data-driven Approaches for Action Recognition. *arXiv preprint arXiv:1505.04427*, 2015.

[114] Zhenzhong Lan, Yi Zhu, Alexander G. Hauptmann, and Shawn Newsam. Deep Local Video Feature for Action Recognition. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2017.

[115] Hsin-Ying Lee, Jia-Bin Huang, Maneesh Kumar Singh, and Ming-Hsuan Yang. Unsupervised Representation Learning by Sorting Sequence, 2017.

[116] Myunggi Lee, Seungeui Lee, Sungjoon Son, Gyutae Park, and Nojun Kwak. Motion Feature Network: Fixed Motion Filter for Action Recognition. In *The European Conference on Computer Vision (ECCV)*, 2018.

[117] Ang Li, Meghana Thotakuri, David A Ross, João Carreira, Alexander Vostrikov, and Andrew Zisserman. The ava-kinetics localized human actions video dataset. *arXiv preprint arXiv:2005.00214*, 2020.

[118] Qing Li, Zhaofan Qiu, Ting Yao, Tao Mei, Yong Rui, and Jiebo Luo. Action Recognition by Learning Deep Multi-Granular Spatio-Temporal Video Representation. In *The ACM International Conference on Multimedia Retrieval (ICMR)*, 2016.

[119] Shasha Li, Ajaya Neupane, Sujoy Paul, Chengyu Song, Srikanth V Krishnamurthy, Amit K Roy Chowdhury, and Ananthram Swami. Adversarial perturbations against real-time video classification systems. *arXiv preprint arXiv:1807.00458*, 2018.

[120] Xueting Li, Sifei Liu, Shalini De Mello, Xiaolong Wang, Jan Kautz, and Ming-Hsuan Yang. Joint-task self-supervised learning for temporal correspondence. In *Advances in Neural Information Processing Systems*, pages 317–327, 2019.

[121] Xinyu Li, Bing Shuai, and Joseph Tighe. Directional temporal modeling for action recognition. In *European Conference on Computer Vision*, pages 275–291. Springer, 2020.

[122] Yan Li, Bin Ji, Xintian Shi, Jianguo Zhang, Bin Kang, and Limin Wang. TEA: Temporal Excitation and Aggregation for Action Recognition. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[123] Yingwei Li, Weixin Li, Vijay Mahadevan, and Nuno Vasconcelos. VLAD3: Encoding Dynamics of Deep Features for Action Recognition. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[124] Yingwei Li, Yi Li, and Nuno Vasconcelos. RESOUND: Towards Action Recognition without Representation Bias. In *The European Conference on Computer Vision (ECCV)*, 2018.

[125] Zhenyang Li, Kirill Gavrilyuk, Efstratios Gavves, Mihir Jain, and Cees GM Snoek. VideoLSTM Convolves, Attends and Flows for Action Recognition. *Computer Vision and Image Understanding (CVIU)*, 2018.

[126] Sungbin Lim, Ildoo Kim, Taesup Kim, Chiheon Kim, and Sungwoong Kim. Fast AutoAugment. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

[127] Ji Lin, Chuang Gan, and Song Han. Training Kinetics in 15 Minutes: Large-scale Distributed Training on Videos. In *Advances in Neural Information Processing Systems (NeurIPS) Workshop*, 2019.

[128] Ji Lin, Chuang Gan, and Song Han. TSM: Temporal Shift Module for Efficient Video Understanding. In *The IEEE International Conference on Computer Vision (ICCV)*, 2019.

[129] Liu et al. Use What You Have: Video Retrieval using Representations from Collaborative Experts. *arxiv:1907.13487*, 2019.

[130] Hanxiao Liu, Karen Simonyan, and Yiming Yang. DARTS: Differentiable Architecture Search. In *The International Conference on Learning Representations (ICLR)*, 2019.

[131] Miao Liu, Siyu Tang, Yin Li, and James Rehg. Forecasting human object interaction: Joint prediction of motor attention and egocentric activity. In *ECCV*, 2020.

[132] Zhaoyang Liu, Donghao Luo, Yabiao Wang, Limin Wang, Ying Tai, Chengjie Wang, Jilin Li, Feiyue Huang, and Tong Lu. TEINet: Towards an Efficient Architecture for Video Recognition. In *The Conference on Artificial Intelligence (AAAI)*, 2020.

[133] Zhaoyang Liu, Limin Wang, Wayne Wu, Chen Qian, and Tong Lu. TAM: Temporal Adaptive Module for Video Recognition. *arXiv preprint arXiv:2005.06803*, 2020.

[134] Zelun Luo, Boya Peng, De-An Huang, Alexandre Alahi, and Li Fei-Fei. Unsupervised learning of long-term motion dynamics for videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2203–2212, 2017.

[135] Chih-Yao Ma, Min-Hung Chen, Zsolt Kira, and Ghassan AlRegib. TS-LSTM and Temporal-Inception: Exploiting Spatiotemporal Dynamics for Activity Recognition. *Signal Processing: Image Communication*, 2019.

[136] Minghuang Ma, Haoqi Fan, and Kris M Kitani. Going deeper into first-person activity recognition. In *CVPR*, 2016.

[137] Pascal Mettes and Cees G. M. Snoek. Spatial-Aware Object Embeddings for Zero-Shot Localization and Classification of Actions. In *ICCV*, 2017.

[138] Antoine Miech, Jean-Baptiste Alayrac, Lucas Smaira, Ivan Laptev, Josef Sivic, and Andrew Zisserman. End-to-End Learning of Visual Representations from Uncurated Instructional Videos. In *CVPR*, 2020.

[139] Antoine Miech, Ivan Laptev, and Josef Sivic. Learning a text-video embedding from incomplete and heterogeneous data. *arXiv preprint arXiv:1804.02516*, 2018.

[140] Antoine Miech, Dimitri Zhukov, Jean-Baptiste Alayrac, Makarand Tapaswi, Ivan Laptev, and Josef Sivic. HowTo100M: Learning a Text-Video Embedding by Watching Hundred Million Narrated Video Clips. In *ICCV*, 2019.

[141] Ishan Misra, C Lawrence Zitnick, and Martial Hebert. Shuffle and learn: unsupervised learning using temporal order verification. In *European Conference on Computer Vision*, pages 527–544. Springer, 2016.

[142] Mathew Monfort, Alex Andonian, Bolei Zhou, Kandan Ramakrishnan, Sarah Adel Bargal, Tom Yan, Lisa Brown, Quanfu Fan, Dan Gutfruend, Carl Vondrick, et al. Moments in Time Dataset: One Million Videos for Event Understanding. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 2019.

[143] Mathew Monfort, Kandan Ramakrishnan, Alex Andonian, Barry A McNamara, Alex Lascelles, Bowen Pan, Quanfu Fan, Dan Gutfreund, Rogerio Feris, and Aude Oliva. Multi-moments in time: Learning and interpreting models for multi-action video understanding. *arXiv preprint arXiv:1911.00232*, 2019.

[144] Pedro Morgado, Nuno Vasconcelos, and Ishan Misra. Audio-visual instance discrimination with cross-modal agreement. 2020.

[145] Jonathan Munro and Dima Damen. Multi-Modal Domain Adaptation for Fine-Grained Action Recognition. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[146] Joe Yue-Hei Ng, Jonghyun Choi, Jan Neumann, and Larry S. Davis. ActionFlowNet: Learning Motion Representation for Action Recognition. In *The IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2018.

[147] Joe Yue-Hei Ng and Larry S. Davis. Temporal Difference Networks for Video Action Recognition. In *The IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2018.

[148] Jiquan Ngiam, Aditya Khosla, Mingyu Kim, Juhan Nam, Honglak Lee, and Andrew Y Ng. Multimodal deep learning. In *ICML*, 2011.

[149] Phuc Nguyen, Ting Liu, Gautam Prasad, and Bohyung Han. Weakly supervised action localization by sparse temporal pooling network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6752–6761, 2018.

[150] Bruce Xiaohan Nie, Caiming Xiong, and Song-Chun Zhu. Joint Action Recognition and Pose Estimation From Video. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[151] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

[152] Boxiao Pan, Zhangjie Cao, Ehsan Adeli, and Juan Carlos Niebles. Adversarial Cross-Domain Action Recognition with Co-Attention. In *The Conference on Artificial Intelligence (AAAI)*, 2020.

[153] Deepak Pathak, Philipp Krähenbühl, Jeff Donahue, Trevor Darrell, and Alexei A. Efros. Context encoders: Feature learning by inpainting. In *CVPR*, 2016.

[154] Mandela Patrick, Yuki M. Asano, Ruth Fong, João F. Henriques, Geoffrey Zweig, and Andrea Vedaldi. Multimodal self-supervision from generalized data transformations. *arXiv preprint arXiv:2003.04298*, 2020.

[155] Sujoy Paul, Sourya Roy, and Amit K Roy-Chowdhury. W-talc: Weakly-supervised temporal activity localization and classification. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 563–579, 2018.

[156] Wei Peng, Xiaopeng Hong, and Guoying Zhao. Video Action Recognition Via Neural Architecture Searching. In *The IEEE International Conference on Image Processing (ICIP)*, 2019.

[157] Xiaojiang Peng, Limin Wang, Xingxing Wang, and Yu Qiao. Bag of Visual Words and Fusion Methods for Action Recognition: Comprehensive Study and Good Practice. *arXiv preprint arXiv:1405.4506*, 2014.

[158] Xiaojiang Peng, Changqing Zou, Yu Qiao, and Qiang Peng. Action Recognition with Stacked Fisher Vectors. In *The European Conference on Computer Vision (ECCV)*, 2014.

[159] Toby Perrett and Dima Damen. DDLSTM: Dual-Domain LSTM for Cross-Dataset Action Recognition. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[160] Hieu Pham, Melody Y. Guan, Barret Zoph, Quoc V. Le, and Jeff Dean. Efficient Neural Architecture Search via Parameter Sharing. In *The International Conference on Machine Learning (ICML)*, 2018.

[161] AJ Piergiovanni, Anelia Angelova, and Michael S. Ryoo. Tiny Video Networks. *arXiv preprint arXiv:1910.06961*, 2019.

[162] AJ Piergiovanni, Anelia Angelova, and Michael S Ryoo. Evolving Losses for Unsupervised Video Representation Learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 133–142, 2020.

[163] AJ Piergiovanni, Anelia Angelova, Alexander Toshev, and Michael S. Ryoo. Evolving Space-Time Neural Architectures for Videos. In *The IEEE International Conference on Computer Vision (ICCV)*, 2019.

[164] AJ Piergiovanni and Michael S. Ryoo. Representation Flow for Action Recognition. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[165] AJ Piergiovanni and Michael S. Ryoo. Avid dataset: Anonymized videos from diverse countries, 2020.

[166] AJ Piergiovanni and Michael S. Ryoo. Learning multimodal representations for unseen activities, 2020.

[167] Rui Qian, Tianjian Meng, Boqing Gong, Ming-Hsuan Yang, Huisheng Wang, Serge Belongie, and Yin Cui. Spatiotemporal contrastive video representation learning. *arXiv preprint arXiv:2008.03800*, 2020.

[168] Jie Qin, Li Liu, Ling Shao, Fumin Shen, Bingbing Ni, Jiaxin Chen, and Yunhong Wang. Zero-Shot Action Recognition with Error-Correcting Output Codes. In *CVPR*, 2017.

[169] Zhaofan Qiu, Ting Yao, and Tao Mei. Learning Spatio-Temporal Representation with Pseudo-3D Residual Networks. In *The IEEE International Conference on Computer Vision (ICCV)*, 2017.

[170] Ilija Radosavovic, Raj Prateek Kosaraju, Ross Girshick, Kaiming He, and Piotr Dollár. Designing Network Design Spaces. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[171] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2015.

[172] Alexander Richard, Hilde Kuehne, and Juergen Gall. Weakly supervised action learning with rnn based fine-to-coarse modeling. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 754–763, 2017.

[173] Itsaso Rodríguez-Moreno, José María Martínez-Otzeta, Basilio Sierra, Igor Rodriguez, and Ekaitz Jauregi. Video Activity Recognition: State-of-the-Art. *Sensors*, 2019.

[174] Marcus Rohrbach, Anna Rohrbach, Michaela Regneri, Sikandar Amin, Mykhaylo Andriluka, Manfred Pinkal, and Bernt Schiele. Recognizing fine-grained and composite activities using hand-centric features and script data. *International Journal of Computer Vision*, pages 1–28, 2015.

[175] C. Roig, M. Sarmiento, D. Varas, I. Masuda, J. C. Riveiro, and E. Bou-Balust. Multi-modal pyramid feature combination for human action recognition. In *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*, pages 3742–3746, 2019.

[176] Andrew Rouditchenko, Angie Boggust, David Harwath, Dhiraj Joshi, Samuel Thomas, Kartik Audhkhasi, Rogerio Feris, Brian Kingsbury, Michael Picheny, Antonio Torralba, and James Glass. Avlnet: Learning audio-visual language representations from instructional videos, 2020.

[177] Michael S Ryoo, AJ Piergiovanni, Juhana Kangaspunta, and Anelia Angelova. Assemblenet++: Assembling modality representations via attention connections. *arXiv preprint arXiv:2008.08072*, 2020.

[178] Michael S. Ryoo, AJ Piergiovanni, Mingxing Tan, and Anelia Angelova. AssembleNet: Searching for Multi-Stream Neural Connectivity in Video Architectures. In *The International Conference on Learning Representations (ICLR)*, 2020.

[179] Jorge Sanchez, Florent Perronnin, Thomas Mensink, and Jakob Verbeek. Image Classification with the Fisher Vector: Theory and Practice. *International Journal of Computer Vision (IJCV)*, 2013.

[180] Fadime Sener, Dipika Singhania, and Angela Yao. Temporal aggregate representations for long-range video understanding. In *European Conference on Computer Vision*, pages 154–171. Springer, 2020.

[181] Dian Shao, Yue Zhao, Bo Dai, and Dahua Lin. FineGym: A Hierarchical Video Dataset for Fine-grained Action Understanding. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[182] Hao Shao, Shengju Qian, and Yu Liu. Temporal interlacing network, 2020.

[183] Yemin Shi, Yonghong Tian, Yaowei Wang, Wei Zeng, and Tiejun Huang. Learning Long-Term Dependencies for Action Recognition With a Biologically-Inspired Deep Network. In *The IEEE International Conference on Computer Vision (ICCV)*, 2017.

[184] Zheng Shou, Hang Gao, Lei Zhang, Kazuyuki Miyazawa, and Shih-Fu Chang. Autoloc: Weakly-supervised temporal action localization in untrimmed videos. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 154–171, 2018.

[185] Zheng Shou, Xudong Lin, Yannis Kalantidis, Laura Sevilla-Lara, Marcus Rohrbach, Shih-Fu Chang, and Zhicheng Yan. DMC-Net: Generating Discriminative Motion Cues for Fast Compressed Video Action Recognition. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[186] Gunnar A. Sigurdsson, Gül Varol, Xiaolong Wang, Ali Farhadi, Ivan Laptev, and Abhinav Gupta. Hollywood in Homes: Crowdsourcing Data Collection for Activity Understanding. In *The European Conference on Computer Vision (ECCV)*, 2016.

[187] Karen Simonyan and Andrew Zisserman. Two-Stream Convolutional Networks for Action Recognition in Videos. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2014.

[188] Karen Simonyan and Andrew Zisserman. Very Deep Convolutional Networks for Large-Scale Image Recognition. In *The International Conference on Learning Representations (ICLR)*, 2015.

[189] Bharat Singh, Tim K. Marks, Michael Jones, Oncel Tuzel, and Ming Shao. A Multi-Stream Bi-Directional Recurrent Neural Network for Fine-Grained Action Detection. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[190] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. UCF101: A Dataset of 101 Human Actions Classes From Videos in The Wild. *arXiv preprint arXiv:1212.0402*, 2012.

[191] Jonathan C. Stroud, David A. Ross, Chen Sun, Jia Deng, and Rahul Sukthankar. D3D: Distilled 3D Networks for Video Action Recognition. In *The IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2020.

[192] Swathikiran Sudhakaran, Sergio Escalera, and Oswald Lanz. Lsta: Long short-term attention for egocentric action recognition. In *CVPR*, 2019.

[193] Waqas Sultani and Imran Saleemi. Human Action Recognition across Datasets by Foreground-Weighted Histogram Decomposition. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014.

[194] Chen Sun, Fabien Baradel, Kevin Murphy, and Cordelia Schmid. Learning video representations using contrastive bidirectional transformer, 2019.

[195] Chen Sun, Austin Myers, Carl Vondrick, Kevin Murphy, and Cordelia Schmid. VideoBERT: A Joint Model for Video and Language Representation Learning. In *The IEEE International Conference on Computer Vision (ICCV)*, 2019.

[196] Lin Sun, Kui Jia, Kevin Chen, Dit-Yan Yeung, Bertram E. Shi, and Silvio Savarese. Lattice Long Short-Term Memory for Human Action Recognition. In *The IEEE International Conference on Computer Vision (ICCV)*, 2017.

[197] Shuyang Sun, Zhanghui Kuang, Lu Sheng, Wanli Ouyang, and Wei Zhang. Optical Flow Guided Feature: A Fast and Robust Motion Representation for Video Action Recognition. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[198] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going Deeper with Convolutions. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[199] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing Properties of Neural Networks. *arXiv preprint arXiv:1312.6199*, 2013.

[200] Graham W. Taylor, Rob Fergus, Yann LeCun, and Christoph Bregler. Convolutional Learning of Spatio-temporal Features. In *The European Conference on Computer Vision (ECCV)*, 2010.

[201] Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive multiview coding. *arXiv preprint arXiv:1906.05849*, 2019.

[202] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning Spatiotemporal Features with 3D Convolutional Networks. In *The IEEE International Conference on Computer Vision (ICCV)*, 2015.

[203] Du Tran, Heng Wang, Lorenzo Torresani, and Matt Feiszli. Video Classification With Channel-Separated Convolutional Networks. In *The IEEE International Conference on Computer Vision (ICCV)*, 2019.

[204] Du Tran, Heng Wang, Lorenzo Torresani, Jamie Ray, Yann LeCun, and Manohar Paluri. A Closer Look at Spatiotemporal Convolutions for Action Recognition. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[205] A. Ullah, J. Ahmad, K. Muhammad, M. Sajjad, and S. W. Baik. Action Recognition in Video Sequences using Deep Bi-Directional LSTM With CNN Features. *IEEE Access*, 2017.

[206] Gül Varol, Ivan Laptev, and Cordelia Schmid. Long-term Temporal Convolutions for Action Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 2018.

[207] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is All You Need. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

[208] Carl Vondrick, Hamed Pirsiavash, and Antonio Torralba. Anticipating visual representations from unlabeled video. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 98–106, 2016.

[209] Heng Wang, Alexander Kläser, Cordelia Schmid, and Liu Cheng-Lin. Action Recognition by Dense Trajectories. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2011.

[210] Heng Wang and Cordelia Schmid. Action Recognition with Improved Trajectories. In *The IEEE International Conference on Computer Vision (ICCV)*, 2013.

[211] Jiangliu Wang, Jianbo Jiao, Linchao Bao, Shengfeng He, Yunhui Liu, and Wei Liu. Self-supervised spatio-temporal representation learning for videos by predicting motion and appearance statistics. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4006–4015, 2019.

[212] Lei Wang, Piotr Koniusz, and Du Huynh. Hallucinating idt descriptors and i3d optical flow features for action recognition with cnns. In *Proceedings of the 2019 International Conference on Computer Vision*. IEEE, Institute of Electrical and Electronics Engineers, 2019.

[213] Limin Wang, Wei Li, Wen Li, and Luc Van Gool. Appearance-and-Relation Networks for Video Classification. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[214] Limin Wang, Yu Qiao, and Xiaoou Tang. Action Recognition With Trajectory-Pooled Deep-Convolutional Descriptors. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[215] Limin Wang, Zhe Wang, Yuanjun Xiong, and Yu Qiao. CUHK and SIAT Submission for THUMOS15 Action Recognition Challenge. *THUMOS'15 Action Recognition Challenge*, 2015.

[216] Limin Wang, Yuanjun Xiong, Dahua Lin, and Luc Van Gool. Untrimmednets for weakly supervised action recognition and detection. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 4325–4334, 2017.

[217] Limin Wang, Yuanjun Xiong, Zhe Wang, and Yu Qiao. Towards Good Practices for Very Deep Two-Stream ConvNets. *arXiv preprint arXiv:1507.02159*, 2015.

[218] Limin Wang, Yuanjun Xiong, Zhe Wang, Yu Qiao, Dahua Lin, Xiaoou Tang, and Luc Van Gool. Temporal Segment Networks: Towards Good Practices for Deep Action Recognition. In *The European Conference on Computer Vision (ECCV)*, 2016.

[219] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-Local Neural Networks. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[220] Xiaolong Wang and Abhinav Gupta. Videos as Space-Time Region Graphs. In *The European Conference on Computer Vision (ECCV)*, 2018.

[221] Xiaolong Wang, Allan Jabri, and Alexei A. Efros. Learning Correspondence from the Cycle-Consistency of Time. In *CVPR*, 2019.

[222] Xiaofang Wang, Xuehan Xiong, Maxim Neumann, AJ Piergiovanni, Michael S. Ryoo, Anelia Angelova, Kris M. Kitani, and Wei Hua. Attentionnas: Spatiotemporal attention cell search for video classification, 2020.

[223] Xiaohan Wang, Linchao Zhu, Yu Wu, and Yi Yang. Symbiotic attention for egocentric action recognition with object-centric alignment. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020.

[224] Yang Wang and Minh Hoai. Pulling actions out of context: Explicit separation for effective combination. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7044–7053, 2018.

[225] Yunbo Wang, Mingsheng Long, Jianmin Wang, and Philip S. Yu. Spatiotemporal Pyramid Network for Video Action Recognition. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[226] Zhipeng Wei, Jingjing Chen, Xingxing Wei, Linxi Jiang, Tat-Seng Chua, Fengfeng Zhou, and Yu-Gang Jiang. Heuristic black-box adversarial attacks on video recognition models. *arXiv preprint arXiv:1911.09449*, 2019.

[227] Philippe Weinzaepfel and Grégory Rogez. Mimetics: Towards understanding human actions out of context. *arXiv preprint arXiv:1912.07249*, 2019.

[228] Bichen Wu, Alvin Wan, Xiangyu Yue, Peter Jin, Sicheng Zhao, Noah Golmant, Amir Gholaminejad, Joseph Gonzalez, and Kurt Keutzer. Shift: A Zero FLOP, Zero Parameter Alternative to Spatial Convolutions. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[229] Chao-Yuan Wu, Christoph Feichtenhofer, Haoqi Fan, Kaiming He, Philipp Krahenbuhl, and Ross Girshick. Long-Term Feature Banks for Detailed Video Understanding. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[230] Chao-Yuan Wu, Ross Girshick, Kaiming He, Christoph Feichtenhofer, and Philipp Krähenbühl. A Multigrid Method for Efficiently Training Video Models. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[231] Chao-Yuan Wu, Manzil Zaheer, Hexiang Hu, R. Manmatha, Alexander J. Smola, and Philipp Krähenbühl. Compressed Video Action Recognition. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[232] Zuxuan Wu, Yanwei Fu, Yu-Gang Jiang, and Leonid Sigal. Harnessing Object and Scene Semantics for Large-Scale Video Understanding. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[233] Zuxuan Wu, Yu-Gang Jiang, Xi Wang, Hao Ye, and Xiangyang Xue. Multi-Stream Multi-Class Fusion of Deep Networks for Video Classification. In *The ACM Conference on Multimedia (MM)*, 2016.

[234] Zuxuan Wu, Caiming Xiong, Chih-Yao Ma, Richard Socher, and Larry S. Davis. AdaFrame: Adaptive Frame Selection for Fast Video Recognition. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[235] Zhirong Wu, Yuanjun Xiong, Stella X Yu, and Dahua Lin. Unsupervised feature learning via non-parametric instance discrimination. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3733–3742, 2018.

[236] Zuxuan Wu, Ting Yao, Yanwei Fu, and Yu-Gang Jiang. Deep Learning for Video Classification and Captioning. *arXiv preprint arXiv:1609.06782*, 2016.

[237] Fanyi Xiao, Yong Jae Lee, Kristen Grauman, Jitendra Malik, and Christoph Feichtenhofer. Audiovisual SlowFast Networks for Video Recognition. *arXiv preprint arXiv:2001.08740*, 2020.

[238] Saining Xie, Ross Girshick, Piotr Dollar, Zhuowen Tu, and Kaiming He. Aggregated Residual Transformations for Deep Neural Networks. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[239] Saining Xie, Chen Sun, Jonathan Huang, Zhuowen Tu, and Kevin Murphy. Rethinking Spatiotemporal Feature Learning: Speed-Accuracy Trade-offs in Video Classification. In *The European Conference on Computer Vision (ECCV)*, 2018.

[240] Dejing Xu, Jun Xiao, Zhou Zhao, Jian Shao, Di Xie, and Yueting Zhuang. Self-supervised spatiotemporal learning via video clip order prediction. 2019.

[241] Tiantian Xu, Fan Zhu, Edward K. Wong, and Yi Fang. Dual Many-to-One-Encoder-based Transfer Learning for Cross-Dataset Human Action Recognition. *Image and Vision Computing*, 2016.

[242] Xun Xu, Timothy Hospedales, and Shaogang Gong. Multi-Task Zero-Shot Action Recognition with Prioritised Data Augmentation. In *ECCV*, 2016.

[243] Xun Xu, Timothy Hospedales, and Shaogang Gong. Transductive Zero-Shot Action Recognition by Word-Vector Embedding. *IJCV*, 2017.

[244] Zhongwen Xu, Yi Yang, and Alexander G. Hauptmann. A Discriminative CNN Video Representation for Event Detection. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[245] Huanqian Yan, Xingxing Wei, and Bo Li. Sparse black-box video attack with reinforcement learning. *arXiv preprint arXiv:2001.03754*, 2020.

[246] Sijie Yan, Yuanjun Xiong, and Dahua Lin. Spatial temporal graph convolutional networks for skeleton-based action recognition. In *AAAI*, 2018.

[247] Ceyuan Yang, Yinghao Xu, Bo Dai, and Bolei Zhou. Video Representation Learning with Visual Tempo Consistency. In *arXiv preprint arXiv:2006.15489*, 2020.

[248] Ceyuan Yang, Yinghao Xu, Jianping Shi, Bo Dai, and Bolei Zhou. Temporal Pyramid Network for Action Recognition. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[249] Xitong Yang, Xiaodong Yang, Sifei Liu, Deqing Sun, Larry Davis, and Jan Kautz. Hierarchical contrastive motion learning for video action recognition, 2020.

[250] Li Yao, Atousa Torabi, Kyunghyun Cho, Nicolas Ballas, Christopher Pal, Hugo Larochelle, and Aaron Courville. Describing Videos by Exploiting Temporal Structure. In *The IEEE International Conference on Computer Vision (ICCV)*, 2015.

[251] Serena Yeung, Olga Russakovsky, Greg Mori, and Li Fei-Fei. End-to-end learning of action detection from frame glimpses in videos. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[252] Wang Yifan, Jie Song, Limin Wang, Luc Van Gool, and Otmar Hilliges. Two-Stream SR-CNNs for Action Recognition in Videos. In *The British Machine Vision Conference (BMVC)*, 2016.

[253] Joe Yue-Hei Ng, Matthew Hausknecht, Sudheendra Vijayanarasimhan, Oriol Vinyals, Rajat Monga, and George Toderici. Beyond Short Snippets: Deep Networks for Video Classification. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[254] Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo. CutMix: Regularization Strategy to Train Strong Classifiers With Localizable Features. In *The IEEE International Conference on Computer Vision (ICCV)*, 2019.

[255] Christopher Zach, Thomas Pock, and Horst Bischof. A Duality based Approach for Realtime TV-L1 Optical Flow. *Joint Pattern Recognition Symposium*, 2007.

[256] Bowen Zhang, Limin Wang, Zhe Wang, Yu Qiao, and Hanli Wang. Real-time Action Recognition with Enhanced Motion Vector CNNs. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[257] Can Zhang, Yuexian Zou, Guang Chen, and Lei Gan. Pan: Towards fast action recognition via learning persistence of appearance, 2020.

[258] Hongyi Zhang, Moustapha Cisse, Yann N. Dauphin, and David Lopez-Paz. Mixup: Beyond Empirical Risk Minimization. In *The International Conference on Learning Representations (ICLR)*, 2018.

[259] Hang Zhang, Chongruo Wu, Zhongyue Zhang, Yi Zhu, Zhi Zhang, Haibin Lin, Yue Sun, Tong He, Jonas Muller, R. Manmatha, Mu Li, and Alexander Smola. ResNeSt: Split-Attention Networks. *arXiv preprint arXiv:2004.08955*, 2020.

[260] Hu Zhang, Linchao Zhu, Yi Zhu, and Yi Yang. Motion-Excited Sampler: Video Adversarial Attack with Sparked Prior. In *The European Conference on Computer Vision (ECCV)*, 2020.

[261] Hong-Bo Zhang, Yi-Xiang Zhang, Bineng Zhong, Qing Lei, Lijie Yang, Ji-Xiang Du, and Duan-Sheng Chen. A Comprehensive Survey of Vision-Based Human Action Recognition Methods. *Sensors*, 2018.

[262] Richard Zhang, Phillip Isola, and Alexei A Efros. Colorful image colorization. In *ECCV*, 2016.

[263] Richard Zhang, Phillip Isola, and Alexei A. Efros. Split-brain autoencoders: Unsupervised learning by cross-channel prediction. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[264] Shiwen Zhang, Sheng Guo, Weilin Huang, Matthew R. Scott, and Limin Wang. V4D:4D Convolutional Neural Networks for Video-level Representation Learning. In *The International Conference on Learning Representations (ICLR)*, 2020.

[265] Xiangyu Zhang, Xinyu Zhou, Mengxiao Lin, and Jian Sun. ShuffleNet: An Extremely Efficient Convolutional Neural Network for Mobile Devices. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[266] Zehua Zhang and David Crandall. Hierarchically decoupled spatial-temporal contrast for self-supervised video representation learning, 2020.

[267] Hang Zhao, Zhicheng Yan, Lorenzo Torresani, and Antonio Torralba. HACS: Human Action Clips and Segments Dataset for Recognition and Temporal Localization. *arXiv preprint arXiv:1712.09374*, 2019.

[268] Yue Zhao, Yuanjun Xiong, and Dahua Lin. Trajectory Convolution for Action Recognition. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018.

[269] Bolei Zhou, Alex Andonian, Aude Oliva, and Antonio Torralba. Temporal Relational Reasoning in Videos. In *The European Conference on Computer Vision (ECCV)*, 2018.

[270] Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 million Image Database for Scene Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 2017.

[271] Yizhou Zhou, Xiaoyan Sun, Zheng-Jun Zha, and Wenjun Zeng. MiCT: Mixed 3D/2D Convolutional Tube for Human Action Recognition. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[272] Linchao Zhu, Du Tran, Laura Sevilla-Lara, Yi Yang, Matt Feiszli, and Heng Wang. Faster recurrent networks for efficient video classification. In *AAAI*.

[273] Linchao Zhu, Zhongwen Xu, and Yi Yang. Bidirectional Multirate Reconstruction for Temporal Modeling in Videos. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[274] Linchao Zhu, Zhongwen Xu, Yi Yang, and Alex G. Hauptmann. Uncovering Temporal Context for Video Question Answering. *International Journal of Computer Vision (IJCV)*, 2017.

[275] Linchao Zhu and Yi Yang. ActBERT: Learning Global-Local Video-Text Representations. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[276] Sijie Zhu, Taojiannan Yang, Matias Mendieta, and Chen Chen. A3d: Adaptive 3d networks for video action recognition, 2020.

[277] Wangjiang Zhu, Jie Hu, Gang Sun, Xudong Cao, and Yu Qiao. A Key Volume Mining Deep Framework for Action Recognition. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[278] Yi Zhu, Zhenzhong Lan, Shawn Newsam, and Alexander G. Hauptmann. Hidden Two-Stream Convolutional Networks for Action Recognition. In *The Asian Conference on Computer Vision (ACCV)*, 2018.

[279] Yi Zhu, Yang Long, Yu Guan, Shawn Newsam, and Ling Shao. Towards Universal Representation for Unseen Action Recognition. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[280] Yi Zhu and Shawn Newsam. Depth2Action: Exploring Embedded Depth for Large-Scale Action Recognition. In *The European Conference on Computer Vision (ECCV) Workshop*, 2016.

[281] Yi Zhu and Shawn Newsam. Random Temporal Skipping for Multirate Video Analysis. In *The Asian Conference on Computer Vision (ACCV)*, 2018.

[282] Mohammadreza Zolfaghari, Gabriel L. Oliveira, Nima Sedaghat, and Thomas Brox. Chained Multi-Stream Networks Exploiting Pose, Motion, and Appearance for Action Classification and Detection. In *The IEEE International Conference on Computer Vision (ICCV)*, 2017.

[283] Mohammadreza Zolfaghari, Kamaljeet Singh, and Thomas Brox. ECO: Efficient Convolutional Network for Online Video Understanding. In *The European Conference on Computer Vision (ECCV)*, 2018.