文献综述专题——视频动态特征提取的深度学习方法

姓名 曲宇勋

班级 1603

学号 201928014628016

摘要

近年来，随着静态图像的几大基本问题，如图像分类，目标检测，语义分割等都取得重大进展，越来越多研究者开始关注视频序列这种动态图像的理解。区别于静态图像，视频序列中存在一种能够描述运动信息的特征，被称为视频动态特征。视频动态特征的提取是视频理解最基层的任务，一个优秀的视频特征描述将有利于下游任务的性能提升。

一个优秀的视频描述子应该具有强大的时序表述能力，对于不同长度视频的适应性，以及高效与实时性等等。本文依照这些特性，在视频理解与生成的相关领域，即降水预测与行为识别领域找到了用于提取视频动态特征的多种模型，在简单介绍两个任务的背景与基准知乎，依照模型架构，将模型分为 RNN 结构、光流结构、3D 结构与 2D 近似及新架构四个分支，并分别分析了其利弊。在文章的最后，总结了几个模型的发展过程与优缺点，并提出了未来的展望。

目录

一、引言

随着计算机视觉领域的发展，静态图片的分类、检测、分割等问题都已经得到比较好的解答。此时更多的研究者将目光投向图片的另一个模态——视频中来。相比于静态的图片，视频更加贴近人观察事物的模式，也蕴含着更多的潜在信息。

然而，相较于静态的 2D 图片，视频中拥有的增量信息主要是动态信息，比如其主要描述存在于多帧之间动态联系，所以视频输入不能仅仅当作多个通道的二维图像处理，而是要找到不同帧之间共有的联系，如人开关门的动作，跑步的姿态等，都非一两帧能够完整表述的。能够描述这种动态信息的特征被称为视频动态特征。

视频动态特征的相关研究广泛存在于视频内容理解的各个子领域中，尤其是行为识别，基于 RGB 图像的行为识别可以认为是视频理解的基础，是视频中的分类问题，其地位就像图像分类一样。成熟的图像分类模型可以作为 backbone 嵌入下游任务中，成熟的行为识别模型也可以作为视频特征提取器插入视频理解的下游任务中去。所以视频动态特征本就是行为识别研究的核心问题。

1.1 什么是一个好的视频描述特征

特征的描述通常难以直接用定量指标来衡量，实验时也更多以降维和分类的方法间接地表述特征的好坏。所以这里进入正文之前，首先定性地分析一下一个好的视频描述特征应该具有什么样的特点。

首先，视频描述子必须具有时序的描述性，如果我们将视频的每一帧送入一个深度网络提取特征，然后将每一帧的特征利用平均池化合并起来得到整个视频的特征，那么这一特征并不具有时序性，这里可以举一个例子，我们将一个开门的视频倒放就可以得到一个关门的视频，然而采用上述的特征提取手段，二者的特征会是相同的，静态的网络并不能区分动作完全不同的两个类别。所以，要描述一段视频的特征，必须要捕获视频中的前后关系、因果关系，理解多帧之间的关联性。

其次，视频描述子需要有不同时序长度的适应性，不因动作短而遗漏，也不因动作长而耗费太多资源。比如原始的 Two-stream 框架需要每一帧的光流图，就不适用于长的视频序列，而 TSN 这种框架由于采用分段采样融合的方式，可以适用于长视频序列。再比如 3D 类方法捕获长时间跨度的关系能力较差，而 Non-Local 的方法捕获长距能力较强。

最后，视频描述子需要有实时性，这一点在研究的中后期的重要程度甚至能比得上前两者，光流和 3D 的方法的性能已经足够优越，然而其计算复杂，并不能满足实时性要求，才会有 2D 近似和新框架的出现。

1.2 本文内容概述

另外一个本文中会提及的领域是降水预测[1][2]，降水预测是视频/时序预测问题的一个子问题，由于任务的特殊性，使得在行为识别领域失去一席之地的 RNN 类型方法在这个领域发展起来。为了确保内容的完整性，在大篇幅描述行为识别中的 3D/光流类方法之前，会首先花一些篇幅介绍降水预测问题以及其中的 RNN 方法，同时也会阐释其不适合行为识别领域的原因。

本文中将会提到用于视频动态特征提取的几种结构，包括在降水预测中的 RNN 类型结构，在行为识别的 3D/光流/Attention/2D 结构。

RNN 模型毋庸置疑是建立时序问题的首选模型，但因为并行性差，训练困难等原因逐渐被行为识别主流所淘汰。但其却因为不可忽视的灵活性在降水预测这种实时性要求不高的任务中大放异彩。从 ConvLSTM[1]与 TrajGRU[2]中，我们可以看到 RNN 结构如何被设计来提取视频中的时空域特征。

光流类方法起源远远早于深度学习，早期行为识别中的 SoTA iDT 就有利用到光流轨迹。而在 NIPS2014 中出现了利用光流图和静态帧图的 Two-Stream 框架，自此深度学习开始进入行为识别领域[3]。CVPR2016 中的 TSN 则提出了一整套包括了框架、正则化方法、训练方式、增广方式的方案，使得双流法可被用于长视频的行为分类[4]，该框架也被后来其他模型广泛借鉴。TSN 开始，光流法逐渐成熟，然而光流法依赖于光流特征的提取，其并不能被集成入深度框架内，且需要耗费较大的计算资源，使得实时性不能被实现，所以后期直至现在，开始涌现出一批光流近似的方法，OFF 就是这样一个例子，其参考光流原理，设计了一套用于近似光流效果的端到端模型，取得不错成效[5]。

3D 类型方法最早可能可以追溯回 13 年[6]，但广为人知的模型 C3D 还是 15 年的工作[7]。3D 类方法添加了时间的卷积维度，理解起来十分自然，很快就得到广泛关注。其在 17 年的 I3D 中发展到顶峰，I3D 中明确了 2D 到 3D 架构的迁移方式，使得 3D 的研究可以建立在 2D 的研究之上，另外其将光流与 3D 结合起来取得极高性能[8]。I3D 以超高的计算耗费取得了超高的性能，这也引起了研究者的反思，更多研究者开始思考更加高效的形式。这在 CVPR18 的(2+1)D 得到了体现，其开始思考是否 3D 是否具有冗余的参数[9]。后期的研究中，3D 的声音逐渐弱了下去。

随之崛起的是新框架与新模型，2D 模型开始成为主流，2D 模型提倡利用简单的 2D 结构代替 3D 和双流模型，2+1D[9]，OFF[5]以及 ICCV2019 的 STM[11]均属于这一类。另外，还有另辟蹊径，采用 attention 方法捕获时域信息的，如 18 年的 Non-Local 模型[10]等。

图中详细表述了十篇文献的相互联系，而正文中会详细阐释模型的细节以及

方法的思想以及发展情况。

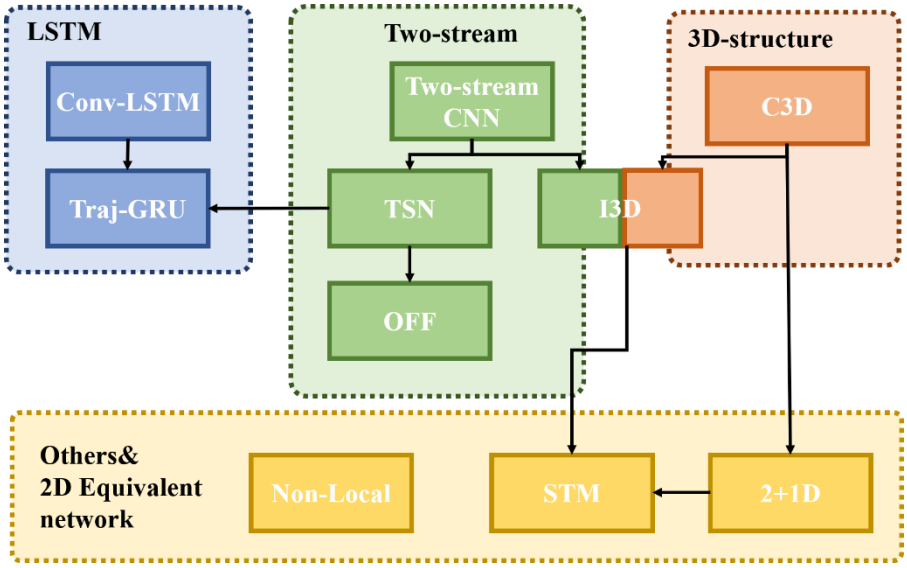


图 1 文中十篇文献的联系

二、降水预测/视频预测中的动态特征提取

2.1 降水预测/视频预测任务简介

2.1.1 视频预测问题

时空序列预测是指输入与要预测的目标均为时空序列的预测问题。若假设可观测序列以张量$\{\mathcal{X}_t\}$表示，被观测采样的序列可由$\{\mathcal{X}_t\}$表示，时空序列预测问题就是用之前的 J 观测对象预测未来 K 个观测对象[1]。

$$\mathcal{X}_{t+1},...,\mathcal{X}_{t+K}=\underset{\mathcal{X}_{t+1},...,\mathcal{X}_{t+K}}{\operatorname{argmax}} p\left(\mathcal{X}_{t+1},...,\mathcal{X}_{t+K}\mid\mathcal{X}_{t-J+1},\mathcal{X}_{t-J+2}...,\mathcal{X}_t\right)$$

当$\{\mathcal{X}_t\}$为一段图像序列的时候，这个问题变为采用视频中的前几帧预测未出现的后几帧图像，也即视频预测问题。从学术上来讲，视频预测问题可以利用大量无监督的数据来学习视频的动态特征，对于视频内容理解有着重要意义，从应用上来说，在安防领域、自动驾驶等方面有着重要应用。视频预测属于视频内容理解中的生成任务，高质量的视频特征提取与逐帧图像生成都是必不可少的。近期的研究主要在于如何利用一些经典的视频特征，如光流特征等驱动 GANs 这中成熟的生成器以生成新的图片。

本次报告中引用的方法却不在这一框架内，本次报告的重点将放在视频特征的提取，视频特征本质上还是时空序列特征，基于 RNN 的方法是不可不提的。但是由于实时性等原因，RNN 类方法在行为识别这种实时性要求高的领域内没

有发展出一定体系，而在视频预测领域较为前期的一系列工作中，我们发现由于降水预测任务的特殊性，RNN 类方法在这类问题中展现出它的优势。这也是本文选取降水预测问题作为视频特征提取的一个子问题的原因。

2.1.2 降水预测问题

降水预测问题是视频预测问题的一个特例，其利用过去一段时间的雷达回波序列去预测未来的雷达图，实际应用中，常常会利用每 6～10min 采样一次的天气雷达的雷达图来预测未来 6～10min 的雷达回波情况。在降水预测任务中，$\mathcal{X} \in \mathbf{R}^{P \times M \times N}$，其代表具有 M 行 N 列的空间范围，P 个预测通道的雷达序列。需要利用过去 J 个雷达图组成的序列 $\mathcal{X}$ 预测将来 K 个雷达图序列的情况。

相比于一些简单的时空序列建模，降水预测问题困难的关键点在于变量空间的庞大，对于一个长度为 K 的序列而言，其变量就多达 $O(M^K N^K P^K)$，所以在实践中，挖掘问题的空间与时间结构，降低问题的维度是解决问题的关键[1]。

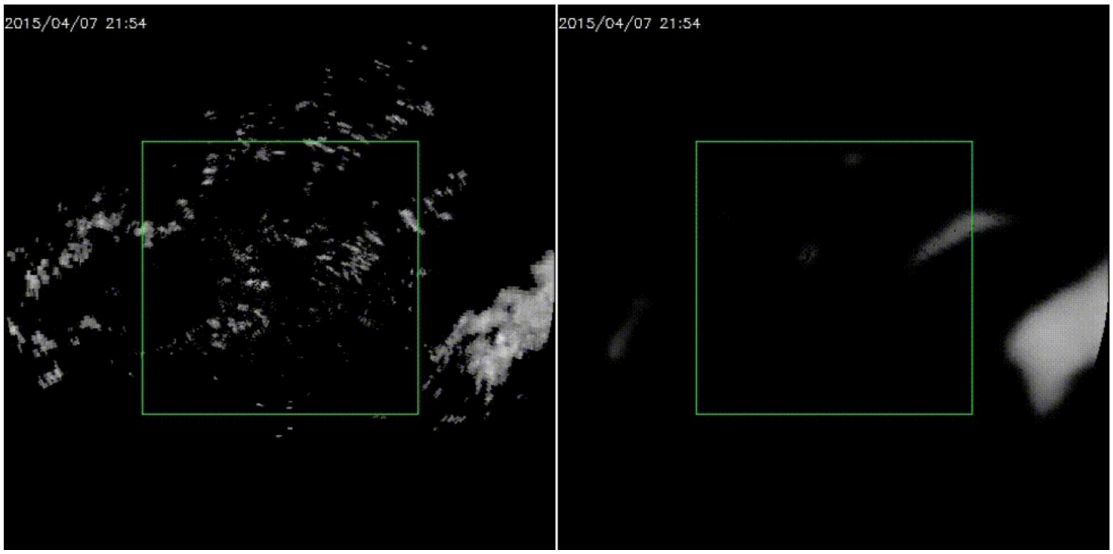


图 2 降水预测任务，左图为采集的雷达 ground truth，右图为预测的结果

2.2 视频/降水预测 BechMark 与指标

2.2.1 视频预测的 toy 数据集和指标

Moving-MNIST 是视频预测领域的 MNIST，属于视频预测领域前期被广泛提及的 Toy 数据集。在 Moving-MNIST 中，每一帧均是一个或多个来自于 MNIST 的字母以随机的运动速度和方向前进，每一帧图片的大小为 64*64，帧数为 20。共有 10000 个训练序列，2000 个验证序列和 3000 个测试序列。[1]

Moving MINIST++是 Moving MINIST 的改进数据集。相比与 Moving MINIST，

添加了随机旋转，随机缩放以及随机形变等运动元素。[2]

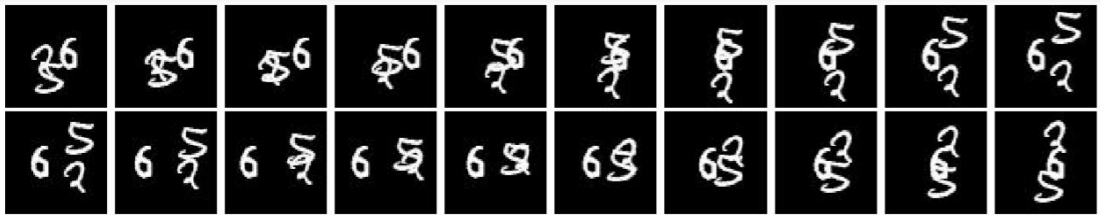


图 3 Moving-MNIST 数据集[1]

对于这个数据集而言，每个像素点是二值的，所以指标上采用的是二值的交叉熵，即 BCELoss。

2.2.2 降水预测的数据集和指标

在研究早期，雷达预测问题相对比较小众，开源的基准也有限，如 ConvLSTM 中采用的雷达图就来自于香港气象台 2011 到 2013 年采集的数据，这一部分并未开源。在雷达数据集中，采用以下气象学描述指标:

$$CSI = \frac{hits}{hits + misses + falsealarms}$$

$$FAR = \frac{falsealarms}{hits + falsealarms}$$

$$POD = \frac{hits}{hits + misses}$$

$$correlation = \frac{\sum_{i,j} P_{ij} T_{ij}}{\sqrt{\sum_{ij} P_{ij}^2 \sum_{ij} T_{ij}^2}}$$

其中，hits 代表预测为 1，真实标签也为 1 情况，misses 表示预测为 0，真实为 1 的情况，falsealarms 表示预测为 1，真实为 0 的情况，P 表示预测的值而 T 表示真实值（均为 0 或 1）。[1]

HKO-7 Dataset 是 TrajGRU 中提出的数据集，也是第一份开源的降水预测数据集。其包含了 2009 到 2015 年 HKO 的雷达数据。有 812 天的训练数据，50 天的验证数据与 121 天的测试数据。雷达反射率已采用 Z-R 方程转换为降雨量。

其提供了两种测试手段，离线测试是模型始终接受 5 帧的输入，并依此预测未来 20 帧。在线测试是连续接收 5 帧信号并且预测将来 20 帧。

测试指标主要是 CSI 和 HSS，另外还有对目标分段加权的 B-MSE 和 B-MAE，CSI 和 HSS 都可以设立不同的阈值 r=0.5,2,5,10,30,依据这个阈值可以将问题变为二分类问题。指标定义如下。[2]

$$CSI = \frac{TP}{TP + FN + FP}$$

$$HSS = \frac{TP \times TN - FN \times FP}{(TP + FN)(FN + TN) + (TP + FP)(FP + TN)}$$

2.3 降水预测 Baseline

降水预测的传统方法主要分为两个方面，一个是 NWP，即采用气象学方程推导预测出之后的降水情况，另一个是雷达图外推的方法，主要是基于光流特征的 ROVER。ROVER 是领域内的 the state of the art，效果比基于方程推算的方法更快更准，但基于光流的方法仍然存在光流估计和外推过程相互隔离并非端到端，模型参数调试困难等问题。

另外，降水预测是一种特殊的时空序列建模问题。在时空序列建模中，RNN 和 LSTM 是不可忽略的 Baseline，RNN 是循环神经网络，通过 LSTM 在 RNN 的基本架构中，引入记忆单元通路，能够有效缓解梯度消失问题，本文的 baseline 是 FC-LSTM，一种 LSTM 的多元拓展。FC-LSTM 的一个 cell 计算公式如下。

$$\begin{aligned}
i_t &= \sigma\left(W_{xi}x_t + W_{hi}h_{t-1} + W_{ci} \circ c_{t-1} + b_i\right) \\
f_t &= \sigma\left(W_{xf}x_t + W_{hf}h_{t-1} + W_{cf} \circ c_{t-1} + b_f\right) \\
c_t &= f_t \circ c_{t-1} + i_t \circ \tanh\left(W_{xc}x_t + W_{hc}h_{t-1} + b_c\right) \\
o_t &= \sigma\left(W_{xo}x_t + W_{ho}h_{t-1} + W_{co} \circ c_t + b_o\right) \\
h_t &= o_t \circ \tanh(c_t)
\end{aligned}$$

可以看出，这个模型对时序信息的建模已经相当精巧，然而，其对空间信息却缺乏设计，ConvLSTM 的方法就是在 FC-LSTM 的基础上挖掘其空间信息，使其能够用于降水预测的任务。[1]

而从 17 年的 TrajGRU 研究开始，就正式将降水预测纳入视频预测的领域。那时的视频预测领域的深度方法分为三大类：RNN 结构，2D-CNN 结构与 3D-CNN 结构。RNN 结构利用 CNN 提取空域特征，利用 RNN 提取运动特征并融合；2D-CNN 结构将时间轴上的图片堆叠进张量的不同通道中中，再采用 CNN 的方法识别；3D-CNN 则将时间轴作为深度，除了空间的邻域以外引入时间的邻域设计结构。[2]

2.4 RNN 类方法

2.4.1 ConvLSTM[1]

2.4.1.1 ConvLSTM 单元结构

ConvLSTM 是 FC-LSTM 在考虑空间结构的一种改进，这种结构中，$\mathcal{C}_t$ 的更新仅仅与 $\mathcal{X}_t, \mathcal{H}_t$ 中对应的格点的邻域相关。所以 input-to-state 和 state-to-state 的处理可以用卷积代替全连接，也即下式。

$$
\begin{aligned}
i_t &= \sigma\left(W_{xi} * \mathcal{X}_t + W_{hi} * \mathcal{H}_{t-1} + W_{ci} \circ \mathcal{C}_{t-1} + b_i\right) \\
f_t &= \sigma\left(W_{xf} * \mathcal{X}_t + W_{hf} * \mathcal{H}_{t-1} + W_{cf} \circ \mathcal{C}_{t-1} + b_f\right) \\
c_t &= f_t \circ \mathcal{C}_{t-1} + i_t \circ tanh\left(W_{xc} * \mathcal{X}_t + W_{hc} * \mathcal{H}_{t-1} + b_c\right) \\
o_t &= \sigma\left(W_{xo} * \mathcal{X}_t + W_{ho} * \mathcal{H}_{t-1} + W_{co} \circ \mathcal{C}_t + b_o\right) \\
h_t &= o_t \circ tanh(\mathcal{C}_t)
\end{aligned}
$$

另外，ConvLSTM 模型中，每个时间步都需要上一个时刻的状态作为输入，其中，初始状态 $\mathcal{H}_0$ 可被设定为全零张量。另外为了保持每个单元内张量大小一致性，卷积的 padding 方式设定为 0-padding。

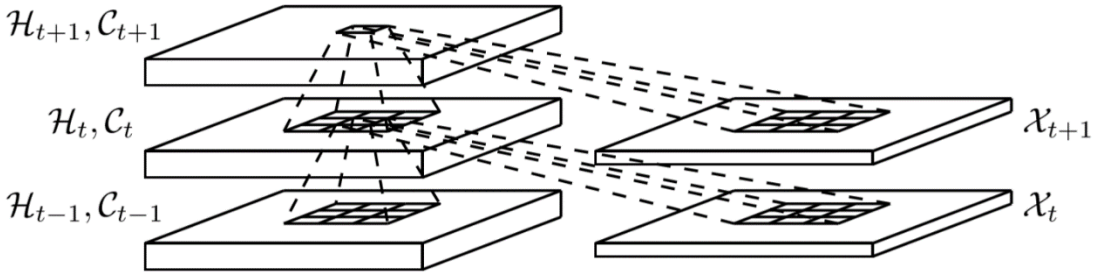


图 4 ConvLSTM 结构[1]

2.4.1.2 基于 RNN 类网络的降水预测框架

该论文的另一个亮点就是其提出了一种适用于降水预测，乃至于视频预测的 RNN 架构。该架构由 Encoder 和 Forecaster 组成，Encoder 部分由多层 Conv-RNN 组成，底层 RNN 的输出作为上层 RNN 的输入。Conv 负责压缩提取空间特征，RNN 部分压缩提取时空特征后，将其融合为状态 $\mathcal{H}$ 输出给 Forecaster。不同层级的 Forecaster 部分以底层的特征作为输入，以 Encodeing 部分的状态特征 $\mathcal{H}$ 作为初始状态继续运行，并将各个不同层的特征融合处理为最终的输出。其最终的结构如下图所示。

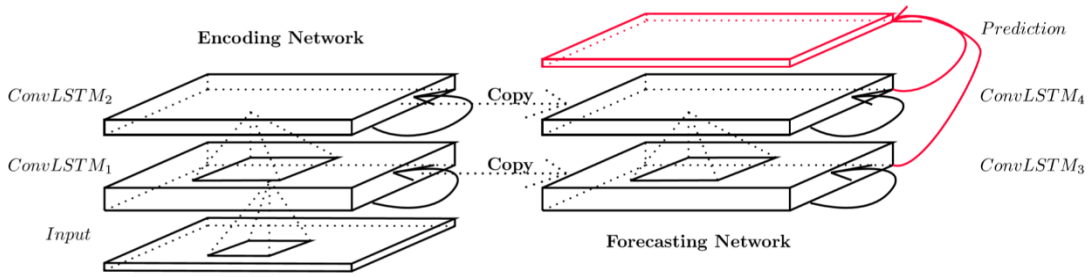


图 5 降水预测框架[1]

2.4.1.3 ConvLSTM 优缺点

ConvLSTM 的优点在于（1）充分考虑了空间的结构，利用卷积的结构以较少的参数量建模复杂的空间信息，利用 LSTM 架构建模时序信息。（2）提出一种通用的 Encoding-Forecasting 框架，这种框架中通过多个层次的 RNN 连接，融合时序的同时也能融合具有不同粒度的信息。

其缺点在于（1）对于物体的变形和旋转等操作效果不佳。其根本原因在于该模型仅仅能融合卷积规定的邻域内的特征，若对于变形和旋转运动，其理论的邻域不再是卷积规定的邻域，而是一个扭曲场。（2）对于这个 RNN 架构而言，由于特征图大小不变，所以其感受野完全由卷积层提供，事实上，感受野并不大，对于大图像中一些快速运动的物体可能存在检测的困难。

2.4.2 TrajGRU[2]

2.4.2.1 结构化 RNN

结构化的 RNN 架构就是在 RNN 架构中添加结构信息，如 ConvLSTM 就添加了空域的结构信息，即空域上的卷积邻域，相类似的，还有 SocialLSTM 和 S-RNN，其分别基于不同个体之间的距离与时空图构建结构化信息。

然而这些模型中的结构都是相对固定的，如 ConvLSTM 模型中，结构就是卷积决定的邻域。而在具有旋转和变形的系统中，由于各点的速度方向的差异，不同点的邻域空间也会有不少差异，也就是对于不同的点，其邻域结构会产生变化。所以，作者提出了 trajGRU，该模型中，邻域结构不再确定，而是可以通过学习地方法自动地调整。

2.4.2.2 新的 RNN 预测框架 Encoding-Forecasting

TrajGRU 针对 ConvLSTM 中提出的视频预测架构做了一定改进。

首先是 Forecaster 的数据流动方向，在上一篇文章中，Forecaster 的数据与 Encoder 一样，是从底层向高层流动的，但本文认为由获取更多全局信息的高层来指导更多细节信息的底层更加合理，而且最终结果中细节信息的影响应该更加

重要。所以本文中 Forecaster 的流向是由底层向高层流动，这一思想同样也体现于 FCN，UNET，FPN 等相关网络结构。

其次是层级的上下采样，由于 Forecaster 不再采用密集加和的形式，不同层的 feature map 就可以拥有不同的大小了。在不同层级中间夹杂入上下采样模块，一方面可以减少计算量，一方面也可以扩大感受野。上下采样可以由 stride=2 的卷积与反卷积构成。

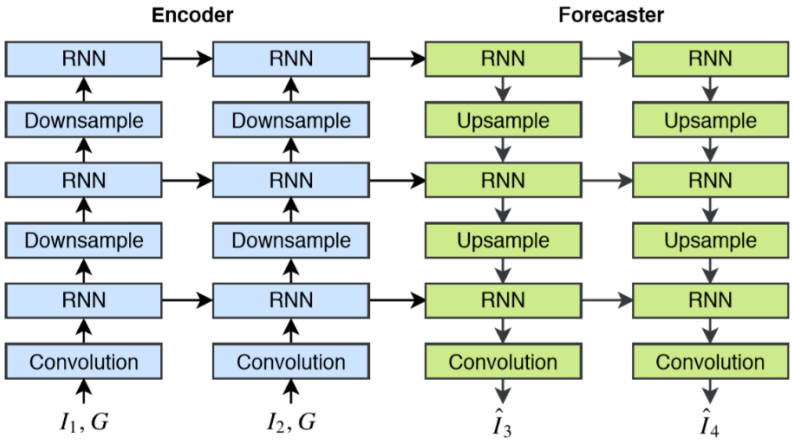


图 6 Encoding-Forecasting 框架[2]

2.4.2.3 ConvGRU 和 TrajGRU

GRU 是 LSTM 的一种改进，GRU 将 LSTM 的三个门控简化为两个门控:重置门与更新门，重置门用于调控前一个状态的对输出影响，更新门则用于控制前一个状态对后一个状态的影响。经过广泛的试验，其能以较少的参数实现与 LSTM 相类似的效果。同理，ConvGRU 也是 ConvLSTM 的一种改进方法，其将卷积的基本结构迁移到 GRU 上，使得 GRU 也能捕获视频流中的空间结构。

$$
\begin{aligned}
\mathcal{Z}_t &= \sigma\left(\mathcal{W}_{xz} * \mathcal{X}_t + \mathcal{W}_{hz} * \mathcal{H}_{t-1}\right) \\
\mathcal{R}_t &= \sigma\left(\mathcal{W}_{xr} * \mathcal{X}_t + \mathcal{W}_{hr} * \mathcal{H}_{t-1}\right) \\
\mathcal{H}'_t &= f\left(\mathcal{W}_{xh} * \mathcal{X}_t + \mathcal{R}_t \circ \left(\mathcal{W}_{hh} * \mathcal{H}_{t-1}\right)\right) \\
\mathcal{H}_t &= (1-\mathcal{Z}_t) \circ \mathcal{H}'_t + \mathcal{Z}_t \circ \mathcal{H}_{t-1}
\end{aligned}
$$

在 ConvGRU 中采用的仍然是一般的卷积过程，在前向传递的过程中，每个卷积的感受野仍然没有变化。为了引入运动过程中邻域的变化，可以利用光流去扭曲空间场。可以举一个例子，假设我们在预测一个极速收缩的球体，那么在一个时刻的状态与下一个时刻的信息进行融合的时候，其应该拥有更大的感受野。如图中左图所示，蓝色框是一般既定的感受野，光流方向用箭头表示，此时由于物体正在收缩，所以其转移到下一个状态的时候，应该具有比既定更大的感受野，如图中红色框所示。

然而，可变感受野的卷积并不容易实现，即使实现了也很难描述不规则的感受野。为了快速便捷地实现这一点，可以反其道而行之，不去改变卷积，而是利

用光流将原图进行改变，再在改变后的原图上进行一般的卷积，如图中右图所示。利用光流将球体先压缩之后再做卷积，就可以实现等效感受野的增大。

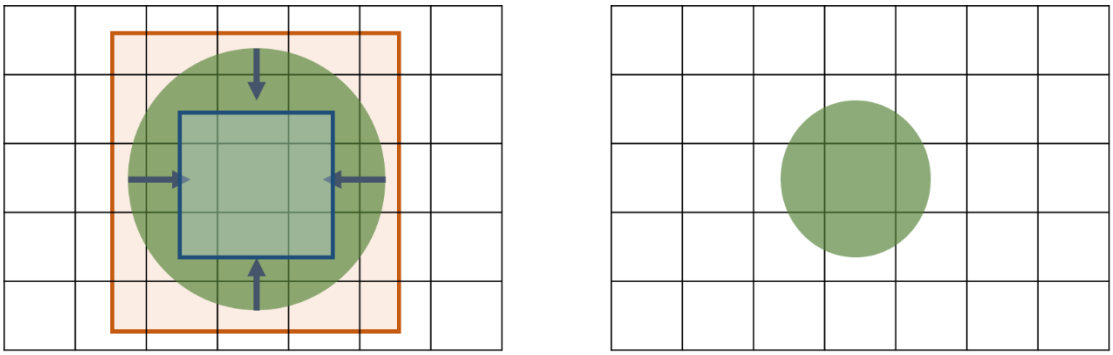
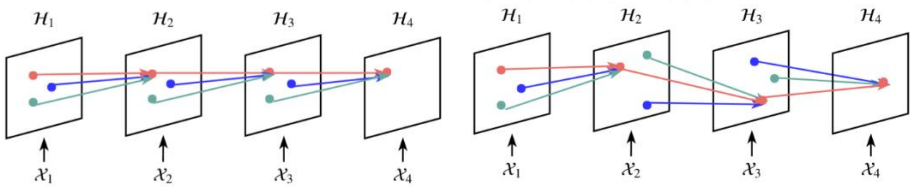


图 7 感受野的变化与光流变化等效

下面考虑用数学语言描述一下这一过程。实际上，如果不考虑 Reset gate 和输入的影响，那么后一个状态与前一个状态的关系如下公式所示。其中 $\mathcal{N}_{i,j}^{h}$ 代表在 (i, j) 处的邻域，其邻域大小由相应的卷积核决定，这也是 ConvRNN 中实际引入的空间结构。这一空间结构的邻域是由卷积核的感受野引入的，卷积核的大小不会随着时间和位置变化。然而对于不同模式的运动而言，这一假设并不成立。例如旋转与尺度变化，这两种运动模式在不同的位置的速度大小方向都不相同。

$$\mathcal{H}'_{t,:,i,j} = f(\mathbf{W}_{hh} concat(< \mathcal{H}_{t-1,:,p,q} \mid (p,q) \in \mathcal{N}_{i,j}^{h} >)) = f(\sum_{l=1}^{|\mathcal{N}_{i,j}^{h}|} \mathbf{W}_{hh}^{l} \mathcal{H}_{t-1,:,p_{l,i,j},q_{l,i,j}})$$



(a) For convolutional RNN, the recurrent connections are fixed over time.

(b) For trajectory RNN, the recurrent connections are dynamically determined.

图 8 TrajRNN 的可变感受野[2]

所以，trajGRU 就利用当前的输入与前时刻的状态来产生一个动态的邻域，但是若邻域的索引是不可微分的，只能用连续光流的插值来代替目标索引。具体操作如下。

$$
\begin{aligned}
\mathcal{U}_t, \mathcal{V}_t = &\quad \gamma(\mathcal{X}_t, \mathcal{H}_{t-1}) \\
\mathcal{Z}_t = &\quad \sigma\left(\mathcal{W}_{xz} * \mathcal{X}_t + \sum_{l=1}^{L} \mathcal{W}_{hz}^{l} * warp(\mathcal{H}_{t-1}, \mathcal{U}_{t,l}, \mathcal{V}_{t,l}) \right) \\
\mathcal{R}_t = &\quad \sigma\left(\mathcal{W}_{xr} * \mathcal{X}_t + \sum_{l=1}^{L} \mathcal{W}_{hr}^{l} * warp(\mathcal{H}_{t-1}, \mathcal{U}_{t,l}, \mathcal{V}_{t,l}) \right) \\
\mathcal{H}'_t = &\quad f\left(\mathcal{W}_{xh} * \mathcal{X}_t + \mathcal{R}_t \circ \left(\sum_{l=1}^{L} \mathcal{W}_{hr}^{l} * warp(\mathcal{H}_{t-1}, \mathcal{U}_{t,l}, \mathcal{V}_{t,l}) \right) \right) \\
\mathcal{H}_t = &\quad (1 - \mathcal{Z}_t) \circ \mathcal{H}'_t + \mathcal{Z}_t \circ \mathcal{H}_{t-1}
\end{aligned}
$$

其中，$\mathcal{U}_t, \mathcal{V}_t \in \mathbf{R}^{L \times H \times W}$ 是光流网络产生的光流域，储存了当前运动形态下的邻域信息。而 $\mathcal{M}_{c,i,j} = warp(\mathcal{I}, \mathbf{U}, \mathbf{V})$ 则是通过双线性插值的方式，将光流信息用于领域的矫正与修整。

$$
\mathcal{M}_{c,i,j} = \quad warp(\mathcal{I}, \mathbf{U}, \mathbf{V}) = \sum_{m=1}^{H} \sum_{n=1}^{W} \mathcal{I}_{c,m,n} \max(0, 1 - |i + \mathbf{V}_{i,j} - m|) \max(0, 1 - |j + \mathbf{U}_{i,j} - n|)
$$

2.4.2.4 TrajGRU 的优缺点

TrajGRU 具有如下优点：（1）改进了 ConvLSTM 中提出的视频预测架构，扩大感受野，减少参数量，提高模型效率。（2）提出一种利用轨迹特征动态调整领域的 TrajGRU 结构，能够动态地捕获运动的邻域结构。

其缺点在于训练并不简单，收敛性能欠佳。由于中间出现了插值模块，模型的可微性能下降。可能需要二阶段的训练更加稳当。

2.5 总结

视频预测领域中的降水预测由于雷达每 5min 接受一次数据，比起用于实时监控的行为识别，用于自动驾驶的视频预测等视频理解与生成任务，其实时性要求并不高，这就使得 RNN 架构能够在这种任务的土壤上发展起来。

ConvLSTM 模型是 RNN 架构的第一个具有强大影响力的视频预测模型，其通过将卷积引入 LSTM cell 中，在 LSTM 的时序模型中首次引入了空间结构。另外参考 LSTM 的基本结构很容易可以加深 LSTM 的层数，以实现复杂度的提升。[1]

而 TrajGRU 模型则是思考更符合视频语境的结构化 RNN，其利用估计的光流来插值重构状态，并在重构的状态下进行卷积并与下一时刻结合，通过这种方式来实现根据运动方式调整的感受野，用于适应形变，缩放等运动类型。这一设计十分符合 RNN 逐时间点运算的架构。另外参考了 UNet 设计了一个多尺度的

RNN 视频预测结构。[2]

可以看出来，RNN 结构由于逐时刻运算的特征，其在设计上具有很大的灵活性，可以依据先验去设计符合视频特点的结构化 RNN。同时也由于设计的特异性，其能取得很优异的效果。但是，正如之前提到的，RNN 结构无法进行并行性运算，递归性计算会使得其训练和推理效率低下，这使得其在大部分实时性视频任务中都退出了主流地位。

三、行为识别中的动态特征提取

3.1 行为识别任务简介

行为识别是视频内容分析的一个分支，即识别出视频中的运动类别，一般一个视频中仅包含一类的人类行为。在视频内容理解中，行为识别是最为基础的一个问题，其地位可以参考静态图像中的图像分类问题。

就像图像分类中的模型可以作为 backbone 应用在目标检测与图像分割中，成熟的行为识别模型也可以作为一个预训练的视频特征提取器加入到视频提名、视频检测、视频问答等模型中去。所以我们可以认为，行为识别问题的核心，其实就是视频特征的提取。这也是为何我们会选取行为识别来做为视频特征专题中的一个部分。

3.2 行为识别任务的 bechmark

在研究前期，行为识别的数据集主要包括 UCF-101 和 HMDB-51。其中 UCF-101 包括了 13K 个视频，平均每个视频 180 帧，共有 101 个动作类别；HMDB-51 包括了 6.8K 个视频，共有 51 个动作类别。[3]

然而，这两个数据集规模都比较小，DeepMind 的研究者在 17 年的 I3D 中提出，ImageNet 的数据集由于具有 1000 类的图像，其训练效果具有很好的泛化性能，甚至用其预训练的 backbone 可以用于其他任务且能获得不少提升，而在视频理解中的分类问题——行为识别问题中，却缺少这样的一个数据集，所以 Kinetics 人类动作行为识别数据集被提出，该数据集的数据量远远大于 UCF-101 与 HMDB-51。这个数据集后来也被证明是十分有效的，在其上训练的模型迁移到其他数据集上能有十分亮眼的表现。

Kinetics 数据集，包括了单人动作、双人交互动作、人与物体互动的动作等。共包含 400 个类别的视频，每个类别超过 400 个视频片段，共有 240k 的视频片段，每个片段 10s，而在测试集中，每个类别共有 100 个视频片段。[8]

在研究的后期，随着新的数据集不断增加，人们又发现行为识别的数据集根

据数据特点的不同可以进行分类。STM 文中提出，数据集可以被分成两类，以动作为主的时序数据集和以单帧场景理解为主的场景数据集。举个例子，打开电脑这个动作通过单帧图像难以辨别出来的，所以这种数据集为时序数据集，另一种动作比如骑马，仅仅通过单帧图像的场景就能推理出动作，这种数据集为场景数据集。

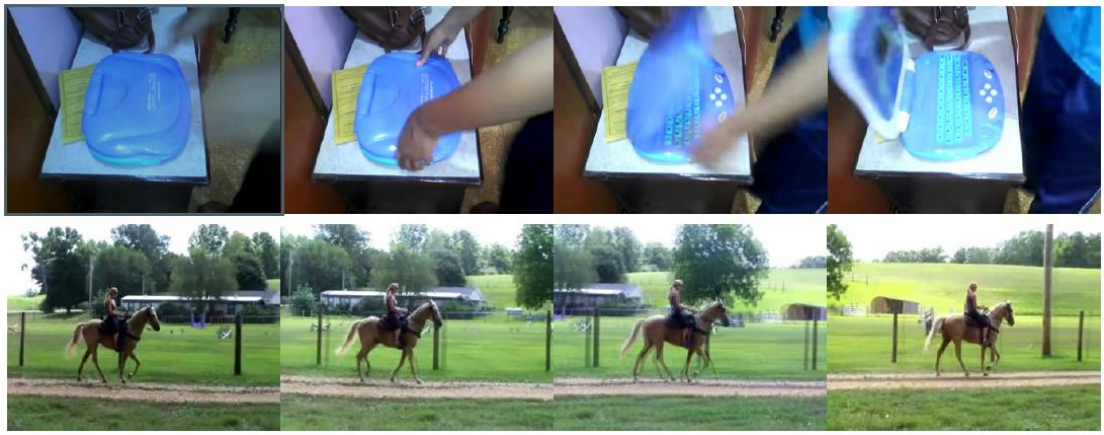


图 9 时序数据集和场景数据集[11]

常见的时序数据集又 Something-Something v1，v2 , Jester，常见的场景数据集包括 Kinetics400, UCF-101, HMDB-51 等。[11]

3.3 光流与光流近似方法

光流特征是指采用光流法解算出的图像的运动特征，分为稀疏光流与密集光流。在像素点亮度不变与小运动的前提下，可以得到如下约束。

$$I(x,y,t)=I(x+\Delta x,y+\Delta y,t+\Delta t)$$

在位移小，时间间隔短的前提下，可以对上式子进行泰勒展开，从而得到下面的推论。

$$\begin{aligned} I(x,y,t)= \quad & I(x,y,t)+\frac{\partial I}{\partial x}\Delta x+\frac{\partial I}{\partial y}\Delta y+\frac{\partial I}{\partial t}\Delta t+\varepsilon \\ 0= \quad & \frac{\partial I}{\partial x}\Delta x+\frac{\partial I}{\partial y}\Delta y+\frac{\partial I}{\partial t}\Delta t+\varepsilon \end{aligned}$$

两边同时除以 Δt，在计算过程中，可以忽略高阶无穷小 ϵ，则可以得到下面的结论。

$$
\begin{aligned}
0 = & \frac{\partial I}{\partial x}\Delta x + \frac{\partial I}{\partial y}\Delta y + \frac{\partial I}{\partial t}\Delta t + \varepsilon \\
0 = & \frac{\partial I}{\partial x}\frac{\Delta x}{\Delta t} + \frac{\partial I}{\partial y}\frac{\Delta y}{\Delta t} + \frac{\partial I}{\partial t} \\
0 = & \frac{\partial I}{\partial x}v_x + \frac{\partial I}{\partial y}v_y + \frac{\partial I}{\partial t}
\end{aligned}
$$

令 $I_x = \frac{\partial I}{\partial x}, I_y = \frac{\partial I}{\partial y}, I_t = \frac{\partial I}{\partial t}$,那么原式变为

$$
0 = (I_x, I_y, I_t)(v_x, v_y, 1)
$$

这就是光流法的约束方程，其中 I_x, I_y, I_t 可分别用 Sobel 算子和帧间差分实现近似，利用该方程我们可以得知 v_x, v_y 的关系。另外，我们可以引入一些频域或者空域上的约束来进一步求解 v_x, v_y，从而解得图中关键点或是每一个像素点的速度大小与方向，这就是光流图。

光流法是一种十分古老的方法，其可以追溯回行为识别的传统方法 iDT，iDT 利用了光流轨迹和传统特征提取子提取视频特征，其在传统方法中一直占据着绝对的地位。

深度学习的潮流到来之后，研究者开始研究将光流特征融入深度结构中去。Two-stream 框架应运而生，在当时 Two-stream 框架取得了一骑绝尘的性能，但 two-stream 框架由于密集的光流计算导致耗费计算量巨大，难以应用于实时性要求高的任务中去。[3]TSN 框架改密集计算为稀疏计算，就很好解决了这一问题。光流类方法的性能十分强悍，现在许多数据集中的 SoTA 性能，仍然由带有光流特征的 TSN 以及其改进方法所取得。[4]但其仍然存在一些问题，比如光流计算复杂，难以实现实时性计算，在后续的研究中，也有更多的研究者将精力放在光流的等效与近似上。

在本节中，将着重介绍基本框架 Two-stream、TSN 与近似光流特征 OFF。

3.3.1 Two stream[3]

Two-stream 法是深度学习在人类行为识别上的开山之作，在此之前深度学习的效果都难以超越传统特征的方法。视频分类任务中最为重要的一点是时序关系的提取，文章提取时序的方式也十分简单，就是采用光流特征，在空间流之外独立地添加一路时间流，最后将特征结合起来用于分类。

3.3.1.1 two-stream 框架

区别于一般图像分类，行为识别的网络框架分为两个部分，分别用于提取空

域和时域的信息。两个部分的网络的结构是类似的，不同点在于输入特征的组织。空域部分的输入为单帧的图像，而时域的部分输入则为多帧的光流信息，也即光流栈/轨迹栈。下图为双流 CNN 的框架。

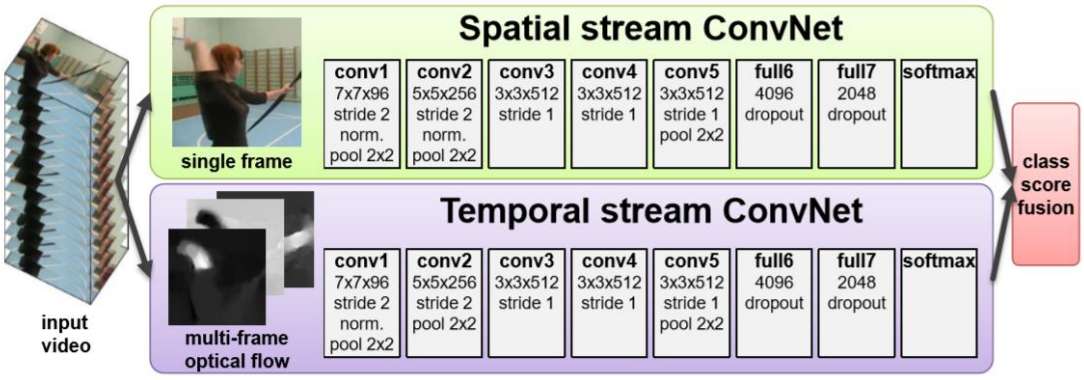


图 10 双流法 Two-stream 框架[3]

3.3.1.2 光流栈特征

光流法是一种传统的提取视频运动信息的方法，如下图所示，中间的图展示了局部的光流场。在光流场中，每个像素都被赋予一个向量，这个向量的方向代表该像素点的运动方向，而其大小代表像素点的速度大小。光流法可以分为稀疏与密集光流，其中密集光流指的是逐像素地计算光流场，最终得出与原图相同分辨力的光流方法。

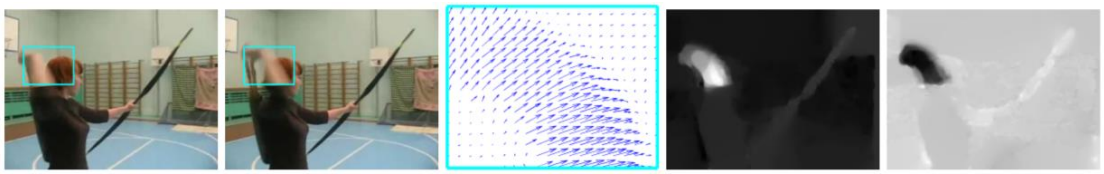


图 11 光流与轨迹示意[3]

具体应用在输入特征中，可以分为光流堆叠与轨迹堆叠，光流堆叠是直接将每一帧的光流特征堆叠起来，而轨迹堆叠则是追踪某一点在多帧光流中的位置变化。另外还有双向的光流网络等推广的光流特征。

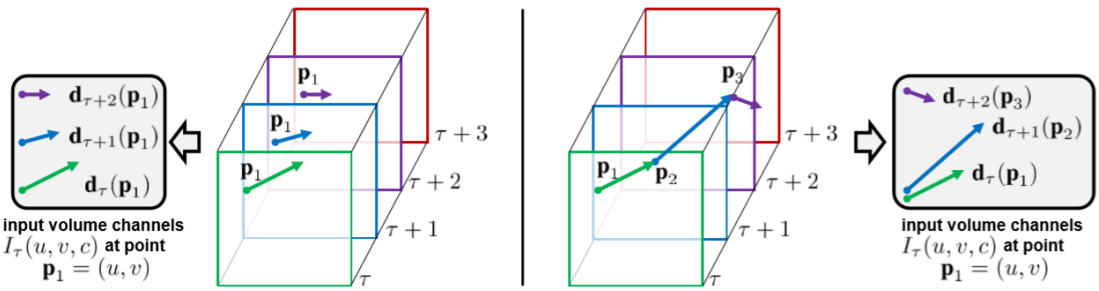


图 12 光流栈示意[3]

经过实验证明，双流框架中，光流栈的特征通常比轨迹栈有效，双向的光流更加有效。空域模型采用预训练模型通常也能取得更好的效果。

在当时数据集数据缺乏的前提下，本文在两个数据集上同时训练一个网络，用不同的 softmax 层训练两个数据集的任务。这种做法可以有效地减少过拟合程度，但对现在的足量的数据集并不具有很大的参照意义。

3.3.1.3 Two-stream 方法的优缺点

Two-stream 方法的优点在于其是行为识别领域上的第一个深度模型，提出了双流的框架，第一次提出引入光流为代表的时序特征。至今为止，SoTA 的性能依然依赖于光流特征的提取。

但其作为开山之作，仍然存在不少缺陷，（1）对于每一帧都求光流特征，会消耗大量的计算资源，这使得其对于长视频片段与实时性任务的处理变得不显式。（2）时序网络和光流网络的设计以及特征的融合都过于粗糙。

3.3.2 TSN[4]

TSN 提出的年代，行为识别主要的挑战是对于长时间的时序信息难以捕获，Two-stream 主要将静态的表观特征以及短时间内的动态特征作为重点，而提取长时间时序特征的能力较弱。针对该问题，TSN 设计了一个基于采样与分段处理可用于长时间视频的高效行为识别双流结构。TSN 的框架自提出以来就被广泛应用，截至目前，TSN 仍然是行为识别领域不得不提的 baseline 之一。

3.3.2.1 TSN 框架

对于一个视频V,将其分为K个持续时间相同的部分$\{S_1, S_2, ...S_K\}$。TSN 架构可以表述为以下公式。

$$TSN(T_1, T_2, ..., T_K) = \mathcal{H}(\mathcal{G}(\mathcal{F}(T_1, \mathbf{W}), \mathcal{F}(T_2, \mathbf{W}), ..., \mathcal{F}(T_K, \mathbf{W})))$$

其中，$T_1, T_2, ...T_K$.分别是从$\{S_1, S_2, ...S_K\}$中采样得到的时间点，$\mathcal{F}(T_k, \mathbf{W})$是第$k$段视频的采样帧图像用双流网络提取出的特征。$\mathcal{G}(.)$表示将各段特征聚合的函数。而$\mathcal{H}(.)$表示 softmax 函数。

简单来说，其仍然使用了双流 CNN 的基本结构，但与之不同的是，不再采用所有帧的图片和光流，而是将视频分段，每段中仅仅用单帧图片与多帧光流及其他模态的特征来描述段内的特征，再将多段的特征联合起来。

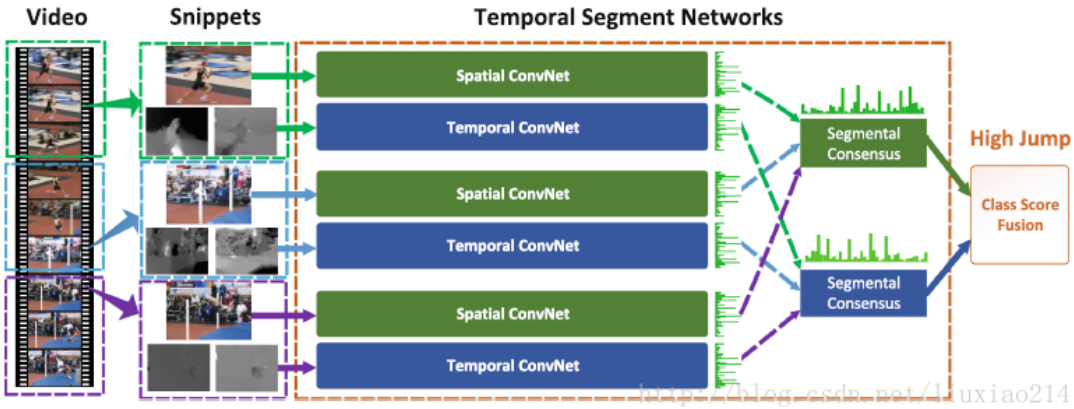


图 13 TSN 结构[4]

3.3.2.2 TSN 的一般配置

网络结构上，原 two-stream 方法中采用的 backbone 较浅，为了充分利用现代网络的潜力，本文用 BN-Inception 的框架来代替原 two-stream 的网络。

网络输入上，除了原先的光流网络，本文还提出另外两种输入，分别是 RGB 差值和 wrapped 光流场。二者都是为描述物体的运动补充的时序特征，其中 RGB 差值是将图像与上一帧图像相减的结果作为特征。而 wrapped 光流场则是受到了 iDT 方法的启发，由于光流特征中耦合了相机本身的运动，就采用 wrapped 光流场特征来抑制背景的运动。这两种特征在对比实验中都独立作为一个分支被训练。

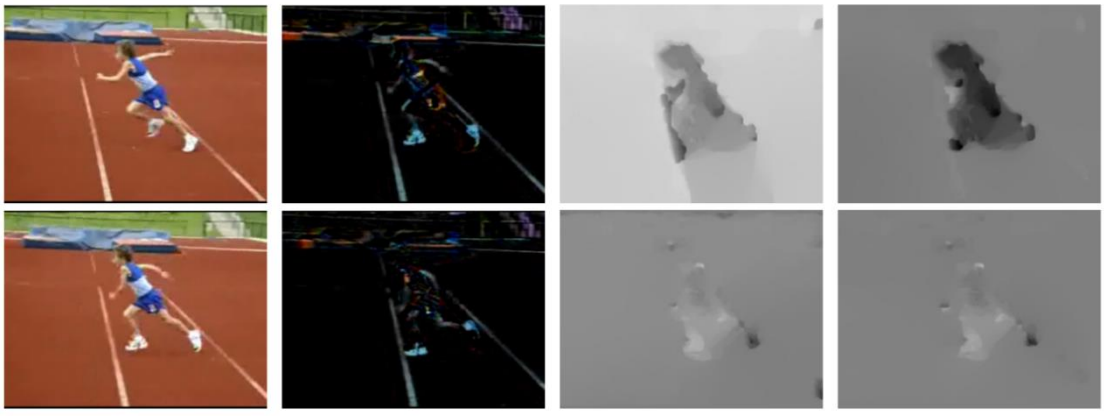


图 14 不同模态的动态特征[4]

网络训练上，RGB 图像分支可以直接采用预训练模型，之后再 fine-tuning，但是时间流不行。所以，这里提出将时间流网络一样用 RGB 网络的预训练参数初始化，并将原网络的第一层修改为适用于光流的特征通道数的卷积。另外第一层的参数在输入 channel 上将原卷积核取平均，再将其按照光流特征的通道数复制多份即可。通过这种方式即可迁移参数。

正则化上，为了避免过拟合，冻结 BN 的参数，另外在之后添加 dropout 层。

数据增强上，新提出角裁剪（从图片边角或中心提取，避免仅仅关注图像中

心）和尺度抖动（即改变裁剪的长宽比与大小）。

实验证明，最为有效的特征组合是光流+wrapped 光流+RGB 图像，不同段落的特征聚合方式中最为有效的是平均。

3.3.2.2 TSN 的优缺点

TSN 的优点在于（1）提出了一种段落采样的框架，使得长时间距离的关系能够被捕获。（2）提出了几种可以替代或者是辅助光流的特征，拓宽了视频序列中时序特征描述的路子。

其缺点在于（1）仍然使用光流特征，计算仍然较慢。（2）聚合不同段落的特征等环节设计较为粗糙，还有改进空间。

3.3.3 OFF[5]

双流框架中光流图的计算会限制双流架构的运行效率。有一些研究尝试仅在训练阶段引入光流特征，或是提出一种简化的光流向量来代替光流特征，其都不能完全替代光流的效果。OFF 的方法在时序上参考了光流法的原理，建立基于光流的时序特征来辅助行为的分类。

3.3.3.1 光流等效特征 OFF

光流法基本假设公式规定了若物体在短时内有较小的位移，且其不同时刻的对应点灰度值不改变，其满足

$$\left[\frac{\partial f(I,\omega)(p)}{\partial x},\frac{\partial f(I,\omega)(p)}{\partial y},\frac{\partial f(I,\omega)(p)}{\partial t}\right][v_x,v_y,1] \quad =0$$

可以做以下符号假设

$$v=(v_x,v_y,1),\vec{F}(I,\omega)(p)=\left[\frac{\partial f(I,\omega)(p)}{\partial x},\frac{\partial f(I,\omega)(p)}{\partial y},\frac{\partial f(I,\omega)(p)}{\partial t}\right]$$

可以看出，在之前的光流法中，是引入约束去计算光流矢量 v 的近似。OFF 方法认为，可以直接利用 v 与 $\vec{F}(I,\omega)(p)$ 呈正交关系这一点，将 $\vec{F}(I,\omega)(p)$ 直接作为光流特征的一种替代，称为光流指导特征(OFF)。

OFF 特征中，x,y 方向的偏导数都可以用对应方向的 Sobel 算子替代，而时间上偏导数则可以用帧间的差分替代。这种 OFF 特征可以存在于各个层次中，可以捕获不同尺度的时序特征。

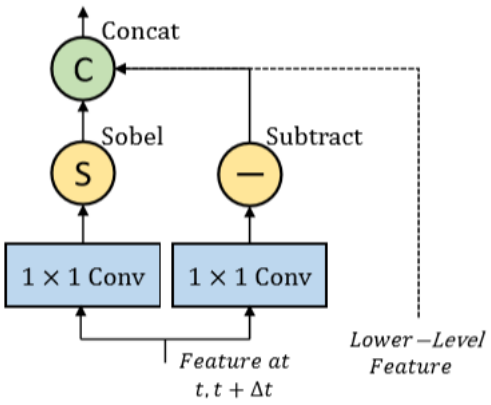


图 15 OFF 模块实现[5]

3.3.3.2 OFF 整体网络

网络的整体结构如下，网络的 backbone 是 BN-Inception，相邻帧与 OFF 层分别计算出 score 之后求平均得到最后的 score。网络整体上采用 TSN 的框架与配置，即将训练集的视频切分为多段，对于每一段仅采样两帧用于下图的网络，并将最终 score 平均下来得到最后 score。训练方式上采用的是双阶段训练，第一阶段用 TSN 作为 pretrain 模型来训练特征提取网络，第二阶段冻结特征提取层的参数并加入 OFF 层进行训练。另外采用了中间监督以辅助训练，采用训练集采样小于测试集来加速训练。

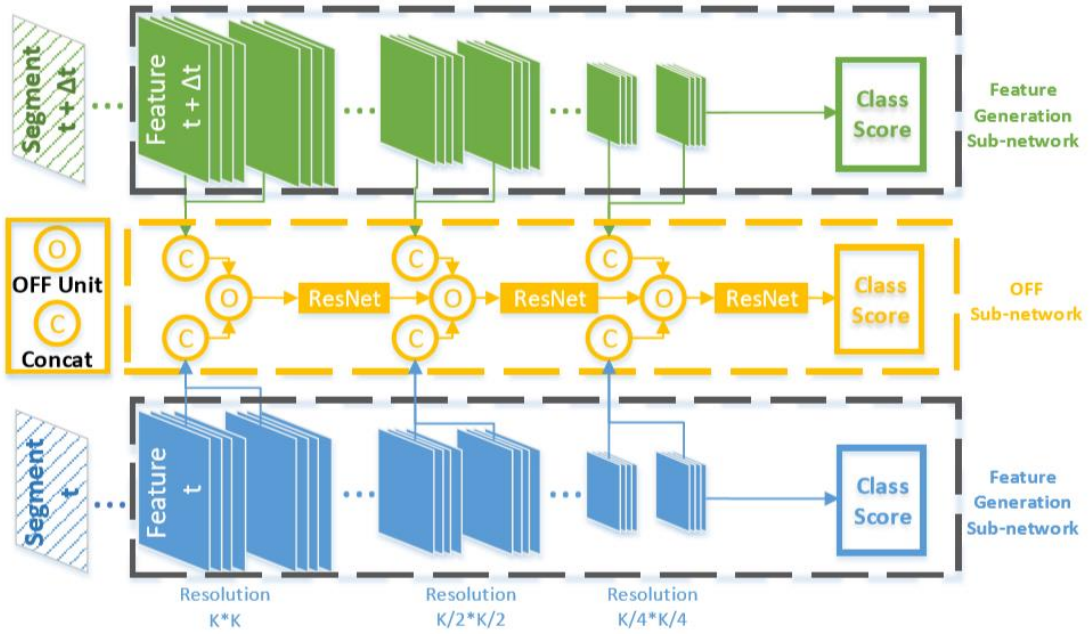


图 16 OFF 整体网络结构[5]

实验证明 OFF 框架中利用 RGB 以及 RGB-diff 特征作为基本特征，可以在提速超过 10 倍的情况下，达到与 TSN(RGB+Flow)相近的效果。另外，添加入 RGB-diff 特征作为基本特征能够在不过多降低速度的前提下显著提升性能。

3.3.3.3 OFF 方法的优缺点

OFF 方法的优点在于（1）提出了一种 OFF 模块，可以作为光流法的一种有效替代。在与 RGB+Flow 的 TSN 达到类似的性能前提下，速度甚至可能超过 200 帧/s。（2）OFF 模块作为一个模块，可以很容易地嵌入现有的架构中，如 TSN 等。

其缺点在于虽然其效率高，但是其难以端到端训练，结构相对固定，且性能仍然与光流特征有一定差距，其仍具有改进空间。

3.3.4 光流类方法的总结

光流框架作为继承了传统 iDT 的一类方法，利用光流特征来捕获动态特征，并将其与静态的特征相融合。由于光流法仅能建模相邻帧之间的动态信息，早期的光流法采用堆叠的光流图来表征整段视频的动态特征，对于长视频而言，长的连续光流栈的高额计算成本使得其基本不能被实现。

所以 TSN 框架的提出具有里程碑的意义，其提出了一种分段采样后融合的思路，使得长视频的处理变得可能，另外其在光流特征的基础上引入了其他类光流特征以补充动态特征信息。TSN 框架一出世就引起极大关注，其不仅仅提供了一种模型，更是提供了包括训练方式、增广、正则化等等一系列内容的一整套模式。后续有影响力的模型，如 TSM 等基本都是建立在 TSN 的框架之下的，许多数据集的 SoTA 也仍然由带有光流的 TSN 所保持。

TSN 已经证明光流类方法在视频特征提取上的一夫当关的地位，然而其高额的计算成本却使得其难以落地到一些现实中实时性要求高的任务中去。所以在后期，光流类方法的研究趋势开始转向快速的光流近似方法，其希望能以较低的运算成本来代替行为识别框架中光流的地位。

OFF 就是这样一个例子，其追溯回光流原理，利用光流原理中与光流向量正交的另一个向量来间接建模光流，也取得了不错的成效。另外还有不少研究利用网络去近似光流的计算。

关于光流的作用其实也一直存在争议，也有研究指出，光流类方法并不是真正地提取出动态特征，而仅仅是隐藏了不重要的表观特征。比如向前运动的人，光流的作用仅仅是隐藏人的细节体征，包括衣服、五官等，从而使得模型能够关注到运动本身。

相关的观点也很多，光流的争议也未曾中断过，不过实验也向我们证实了，作为视频描述特征子，光流类方法具有其他方法所不能比拟的强大的性能。

3.4 3D-CNN 以及 3D 近似方法

3D-CNN 方法是行为识别领域关于时序不同于光流的另一条探索。3D-CNN 在时间维度引入卷积来建模时序信息，将 2D 的卷积上升至 3D。其发展从 C3D 起始，到 I3D 达到高潮。3D 类方法是一种十分自然的想法，比起光流类方法，其对计算资源的依赖确实更低，然而其并称不上是一个轻量级的方法，新增的维度会引入成倍增长的参数，这导致训练难以为继，另一方面，在巨量的参数下性能的提升却十分受限。所以在研究的后期，3D 类方法逐渐走向瓶颈，越来越多研究者开始怀疑 3D 类方法是否是一个足够高效直接的方法，越来越多 3D 的近似方法开始被提出。

3.4.1 C3D[7]

C3D 是 3D-CNN 在行为识别领域的开山之作，其在 2D-CNN 之上增加了时间维度，提出了一种适用于行为识别的 3D-CNN 的结构。

3.4.1.1 3D 卷积结构

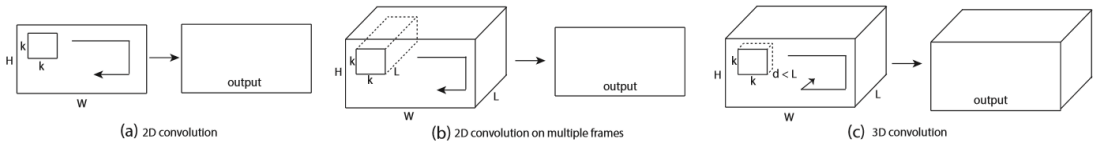


图 17 3D 卷积结构[7]

3D-CNN 和 2D-CNN 的区别主要在于对于时间轴的处理上，图中示意了 2D-CNN 和 3D-CNN 的处理，2D-CNN 对于时间的处理一般有两种，一种是将时间合并入样本维度，相当于对每一帧图像分开处理；另一种方式将时间合并入通道维度，时间上采用全连接，相当于处理一个通道数为 3*T 的图像。而 3D-CNN 在时间上也采用的是 CNN 的邻域思想和参数共用思想，也就是说，区别于 2D-CNN 仅对空间的邻域做卷积，3D 的卷积层卷积的对象是一点与其空间与时间的邻域。池化也是一个道理，同时考虑空域和时域的邻域做的平均/最大操作。、

3.4.1.2 C3D 整体架构

与 2D-CNN 同理，3D-CNN 也需要堆叠才能产生效果，然而 3D-CNN 时间维度的 kernel size 设置又成了一个可以讨论的问题。C3D 做了一系列的实验，包括了 kernel size 为 1，3，5，7，以及逐层递增与逐层递减的情况，事实证明，kernel size 取 3 的时候能够以较少的参数量与计算量取得较优的性能。

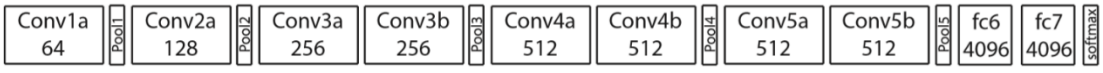


图 18 C3D 结构[7]

3D-CNN 的基本结构如图所示，其中 3D 卷积核采用的大小是 3*3*3，池化层第一层采用的是 1*2*2，之后采用的都是 2*2*2。第一层池化时间维度的 kernel size 为 1 的原因是为了在浅层更好地捕获单帧的特征。

C3D 论文在 Sport-1M 上，做了一个简单的可视化，提取出了 conv-5b 中激活值最高的部分在对应帧上的投影，可以看出 C3D 在视频的初期特征主要集中在整张图片的特征，也即主要集中于空域特征，而随着时间推移，后期的特征主要集中于运动部分的特征。

图 19 C3D 可视化结果[7]

3.4.1.3 C3D 方法的优缺点

C3D 方法的优点在于其是利用 3D-CNN 在行为识别任务上应用的开山之作，在视频序列中十分自然地引入了 3D 卷积操作。

其缺点在于，在当时的设计中，由于本身与 2D 网络结构有较大差别，所以难以将在 ImageNet 上预训练的现代深度网络直接迁移入网络中。

3.4.2 I3D [8]

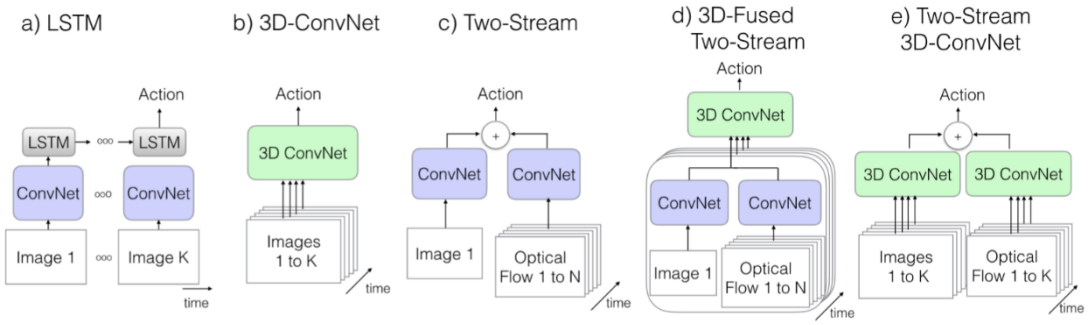


图 20 I3D 之前的主流结构，图 e 为 I3D 结构[8]

相比于 C3D，I3D 是一个更加全面的工作，其总结了 17 年以前行为识别领域的主流方法，找到了一种将 2D-CNN 的结构与参数迁移至 3D-CNN 的方式，并且将 3D 的方法融入双流架构中。笔者认为，将 I3D 称为集大成者毫不为过。

3.4.2.1 2D 结构到 3D 的迁移

本文采用了一个最为简便的方式做迁移，就是不改变原网络的形态，仅仅为其中的卷积和池化层添加一个时间轴，具体来说，就是将 N*N 的卷积和池化修改为 N*N*N。参数迁移上，文中假定了 ImageNet 的图片自身堆叠形成了视频格式，文中称为"boring vedio"，希望对于该视频每层提取出的特征能与 2D-CNN 对图片提取出的特征对应，那么显然其对应位置上采用其原来的参数，并在时间维度上削减幅度即可提取特征。具体而言，N*N*M 的卷积核的参数就是 N*N 的卷积核的参数复制 M 倍并且除以 M。

另一个需要注意的问题空间上宽度和长度的感受野应该保持一致，但是时间上不一定是这样，时间上的感受野与视频的帧率息息相关。对于视频帧率为 25 帧每秒的视频而言，I3D 在网络浅层减少了时间维度卷积的 stride。

3.4.2.2 I3D 的结构

I3D 迁移了 Inception 的结构，其具体结构如图所示。另外，虽然 3D-CNN 已经可以一定程度上描述时序信息，但文中认为其对相邻帧信息的捕获仍然不够全面，所以引入一个光流分支做为其补充，两个分支的预测结果平均后作为输出，两个分支均采用 Inception-V1 结构。I3D 的整体架构如上图(e)所示。

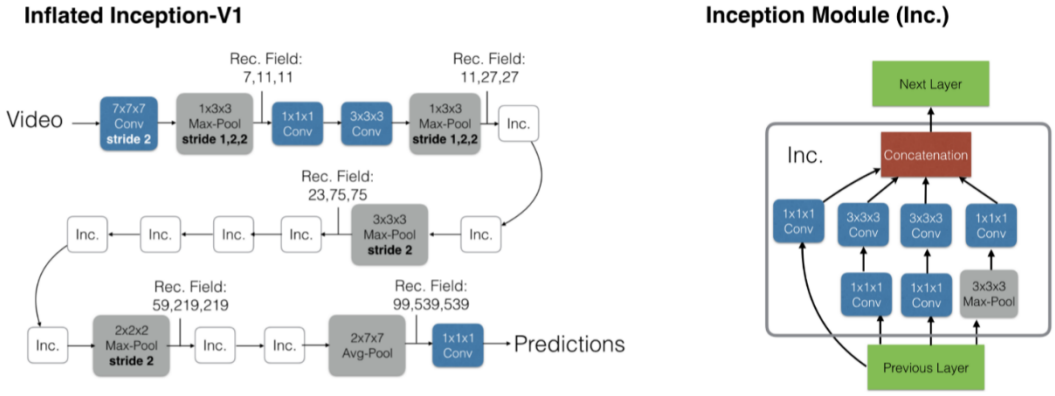


图 21 I3D 的具体网络结构，图中激活函数，softmax 和 BN 层均略去，左图中 Rec Field 代表对应层级的时空感受野大小，Inc 具体结构见右图[8]

3.4.2.3 I3D 的优缺点

I3D 的优点在于（1）其提供了一种将现成的现代 2D 网络结构和参数迁移为 3D 网络的方法，避免了在 3D 网络中重复探索结构。（2）将当下最主流的 3D 结构和双流结构结合，得出具有超高性能的结构。

I3D 同样是一个具有很明确的缺点的模型，它本质上并没有解决光流和 3D 的计算复杂的问题，二者的结合反而使得网络更加难以训练与推理。

需要说明的是，就我个人观点看来 I3D 是一个分水岭，自 I3D 之前，研究者主要关心如何提升模型的性能，但在 I3D 巨大的运算量下，更多关于模型轻量化的研究才开始展开，这一点在背后会详细说明。

3.4.3 2+1D 与 3D 变体 [9]

2+1D 提出的时期，有学界声音认为动态特征实际上是无用的，静态的特征已经可以实现视频行为识别的高精度。这篇文章反驳了这种观点，采用了一系列简化的 3D-CNN 结构进行举证与实验。在这些实验中，最为突出的一个是 2+1D 模型，其类似于将 3D 模型进行了时间维度上的张量拆解，将 3D 拆解为 2D+1D，使得模型降低冗余参数，变得更加高效。

3.4.3.1 2+1D 网络

2+1D 的网络是对 3D 网络的一种拆解，是将 3D 网络的空间维度（2D）和时间维度（1D）拆开来分析。对于一个 $N_i \times N_{i+1} \times t \times d \times d$ 的卷积核而言，将其拆解为 $N_i \times M_i \times 1 \times d \times d$ 的 2D 卷积核串接上一个 $M_i \times N_{i+1} \times t \times 1 \times 1$ 的一个 1D 卷积核，其中 M_i 表示时空串接维度，其计算公式如下。

$$M_i = floor(\frac{td^2 N_i N_{i+1}}{d^2 N_i + t N_{i+1}})$$

其设计的基本原则是和 3D 的参数量保持相同。2+1D 的方法在保持类似的参数量的前提下，增强了整个网络的非线性。具体体现在原来一个非线性激活函数经过拆分之后变为两个非线性层。这个思想类似于 VGG，另外 2D+1D 的结构也更加易于优化。

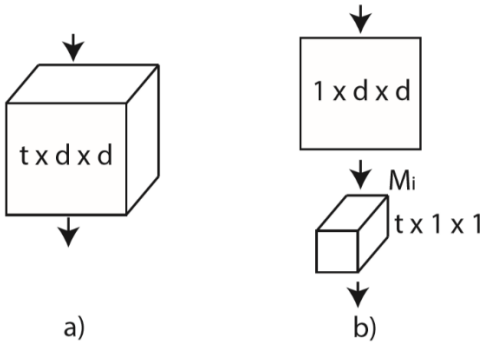


图 22 几种 2+1D 结构[9]

3.4.3.2 其他 3D 的变体以及 3D 有效性实验

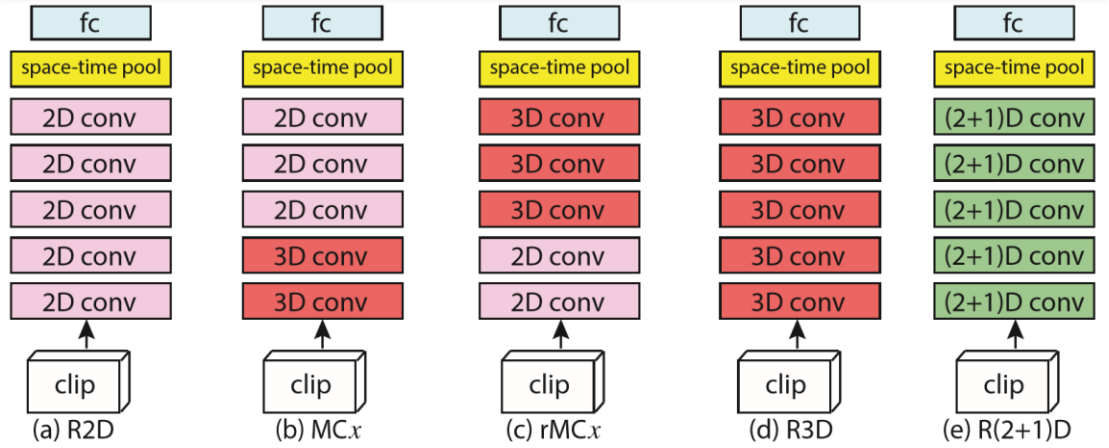


图 23 几种 2D 与 3D 结构[9]

为了证明时域特征的有效性，文章设计了三种模型进行对比试验。

第一种是经典 2D 网络来解决视频问题，R2D 是将视频按照时间轴直接合并起来再放入 2D-CNN 中，而 f-R2D 是独立处理每一帧，将每一帧放入 2D-CNN 中再将多帧的特征融合分类。

第二种是 3D-CNN，包括 R3D、MCx 和 rMCx，R3D 是 ResNet 改造成的 3D-CNN，MCx 和 rMCx 则是基于两种不同的假设设计的 CNN。MCx 系列基于的假设是在视频理解任务中，动作信息在底层更加重要，而到了高层由于高度语义化，动作信息不再必要。所以这种结构在底层使用了 3D-CNN 结构，而在高层还原回 2D 卷积。而 rMCx 的假设正好与之相反，rMCx 结构认为动作信息在高层更加重要，所以在高层采用了 3D 卷积的结构。

第三种就是 2+1D 网络，网络详细内容已在上一节中阐述。

其通过大量的实验给出了以下几个结论（1）所有 3D 网络效果基本都比 2D 网络效果好，说明时域的特征是有效的。（2）2+1D 的在具有相同的参数量的情况下，比 3D 的网络更加有效，说明 3D 网络的设计确实是有冗余的。（3）MC 和 rMC 的网络结构总体上差别不大，也就是说运动特征在底层或者高层有效这一点并不十分严谨，然而其效果都优于一般的 R3D 网络。

3.4.3.3 2+1D 的优缺点

2+1D 的优点在于训练更加容易，且具有更强的非线性。将 3D 拆解之后也有效降低了冗余。其缺点在于 2+1D 时域上仍然采用卷积的思路，对于一些长时间跨度的动作仍然缺少把控。

3.4.4 3D 类方法总结

相比于光流，3D 类方法具有更加自然的思路，其通过卷积学习时域关系，也比光流更加灵活。从 C3D 在 2D 中引入时间维度开始，就有大量研究集中在如

何更好地利用 3D-CNN 结构上，I3D 给出了一个很好的解决方案，即将 2D 的一些先进的网络结构迁移到 3D 的架构上去。到 I3D 为止，3D 方法达到高峰，并且和光流方法相互结合，得到了一个“巨无霸”模型 I3D，这个模型以极大的计算代价换来了极高的性能，至今在许多数据集上仍有一席之地。

个人理解中，I3D 其实是 3D 方法乃至行为识别领域的分水岭，在此之前，研究者关心的是如何提升性能，但从 I3D 屠榜之后，研究者开始反思，以超大的参数量与运算代价换取的性能是否值当。在这个研究背景下，2+1D 横空出世，其在肯定了 3D 的作用的前提下，利用低维卷积来实现高维卷积的等价。2+1D 的成功昭示着 3D 类的方法可能确实存在参数的冗余。

2+1D 的问世带给研究者更多的思考，3D 结构是否是一个高效，乃至于将来可以应用于工业界的结构呢？3D 的方法确实具有很高的灵活性，但随之而来的可能是参数的过多、训练的低效与过拟合。近两年来，在行为识别领域有关于 3D 的研究越来越少，同时，以用较低运算量替代 3D 和光流结构为主要目标的新 2D 乃至于 attention 的架构开始崛起，这也是后面一节将要主要介绍的内容。

3.5 2D 近似结构与新架构

在 2017 年 I3D 之后，行为识别迎来了新的时期，这一时期内，人们开始关注 3D 和光流的一些 2D 近似，以实现视频的实时性识别，上文中的 OFF 和 2+1D 都可以归入这一类中来。另外，3D 和光流方法都是基于邻域的方法，3D 方法虽然随着卷积的堆叠可以实现较大的感受野，然而由于卷积关注的重点仍然在中央，其对远距远时的关系捕获仍然不到位，所以一些可捕获远距关系的新结构被提出。

3.5.1 Non-Local[10]

3.5.1.2 非局部的长距相关性

卷积和循环操作都是捕获邻域关系的特例，而实际上，有一些距离较远（可能是空间距离也可能是时间距离）的格点具有很强的相关性，这种关系可以称为长距依赖性(long-range dependencies)。在过去的方法中，长距依赖性只能由邻域关系反复组合得来，这样的效率是很低的。

在视频类的任务中，就有这样一个案例。捕获与融合视频中的信息的最好方式，是捕获到人的身体部位和球在各帧中的对应关系，第一帧球的邻域信息与第二帧相同位置的信息关系其实不大，而是与第二帧中球的邻域的信息关系更大。如果能够找到一种方法可以更加高效地捕获这种信息，就可以更加高效地处理视频类的任务。

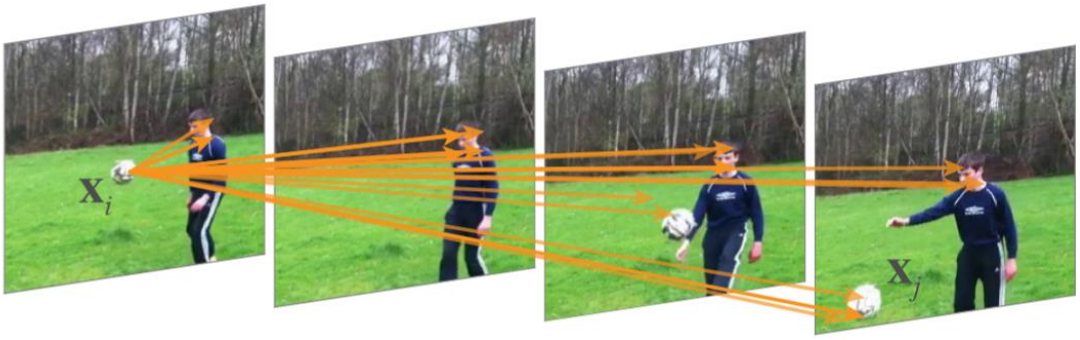


图 24 视频行为识别中的远距关系[10]

光流的方法建模的是相邻帧的动态信息，理论上相隔数帧的信息就难以获取，3D 方法则是通过堆叠卷积的方法来扩大感受野，但也有研究表明随着尺度的增大，对于较远距离的像素点权重都较小，捕获长距离非局部关联性的能力较差。而本文要通过 self-attention 的手段来端到端地学习出远距的关联性。

3.5.1.2 Non-Local block

non-local 的基本核心操作公式如下，$f(\mathbf{x}_i,\mathbf{x}_j)$ 为 $\mathbf{x}_i,\mathbf{x}_j$ 像素点的相似度。这种操作不同于卷积和循环操作的地方在于 $\mathbf{y}_i$ 的值不仅与 $\mathbf{x}_i$ 的时间或者空间的邻域中的像素值有关，而且与其他位置的像素点的值有关。由于不再保持局部性，所以被称为 non-local。

$$\mathbf{y}_i = \frac{1}{\mathcal{C}(\mathbf{x})}\sum_{\forall j} f(\mathbf{x}_i,\mathbf{x}_j)g(\mathbf{x}_j)$$

该操作实现的关键在于函数 f 和 g 的设置，为了简化操作，g 的设置可以用简单的线性操作替代，具体体现在空域上为1×1卷积，体现在时域上则为1×1×1卷积。对于相似度 f 而言，文章中提出以下四种相似度计算方法。

- Gaussian $f(\mathbf{x}_i,\mathbf{x}_j) = \mathrm{e}^{\mathbf{x}_i^T\mathbf{x}_j}$
- Embedded Gaussian $f(\mathbf{x}_i,\mathbf{x}_j) = \mathrm{e}^{\theta(\mathbf{x}_i)^T\phi(\mathbf{x}_j)}$
- Dot product $f(\mathbf{x}_i,\mathbf{x}_j) = \theta(\mathbf{x}_i)^T\phi(\mathbf{x}_j)$
- Concatenation $f(\mathbf{x}_i,\mathbf{x}_j) = ReLU(\mathbf{w}_f^T[\theta(\mathbf{x}_i)^T\phi(\mathbf{x}_j)])$

其中，Embedded Gaussian 的方法等价于 $f(\mathbf{x}_i,\mathbf{x}_j) = softmax(\mathbf{x}^T\mathbf{W}_\theta\mathbf{W}_\phi\mathbf{x})$，这种形式也与 self-attention 操作类似。

经过实验证明，相关性方法的影响并不大，不同相关性函数造成的性能波动

远不如插入 non-local 本身造成的性能提升更大。

上式中定义的 non-local 操作可以被嵌入到其他经典结构中去，一个例子就是 non-local block。其将 non-local 结构嵌入残差连接结构中去。具体的设计如下图所示。

$$\mathbf{z}_i = W_z \mathbf{y}_i + \mathbf{x}_i$$

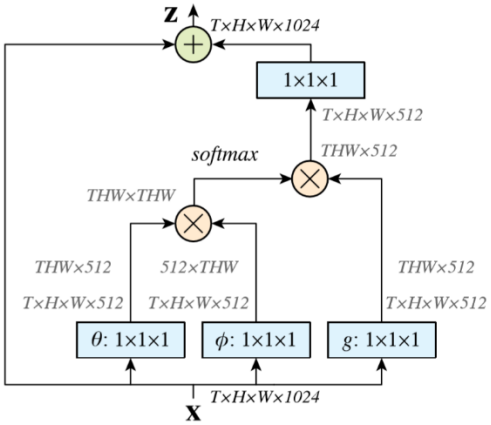


图 25 Non-Local Block[10]

3.5.1.3 其他设计细节

相关实验证明，non-local block 插入的位置也会影响性能。太过于底层会造成计算量过大，计算相关性也不准确，而太过于顶层效果也不佳，原因在于通过多层的卷积，底层特征已经较为充分融合了远距离的信息，此时引入 non-local 效果也不佳。所以最佳的方案是在中层引入。在 ResNet50 的对比实验中，共 5 个 block 的 model 中在第三层或是第四层中引入效果较佳。

另外，non-local 的堆叠对性能的提升也是有效的，实验表明，分别在 ResNet 中添加入 1-non-local block,5-non-local block,10-non-local block,会发现堆叠越多效果越好，一定程度上，这种堆叠比卷积更加有效率。其中 5-block 的 ResNet50 的性能效果已经超越了 baseline 的 ResNet101，即使其只有后者 70\%的参数量与 80\%的 FLOPs。

图中可视化了 non-local block 在 Kinetics 上的测试效果，箭头表示对应像素点相关性最大的几个点，可以看到其与动作的基本趋势相符。

图 26 视频中的远距相关性[10]

3.5.1.4 Non-Local 方法的优缺点

Non-Local 方法优点在于（1）是一种端到端的网络模块，简单且独立性强，可以很容易集成入其他架构中去。（2）相比于传统的 local 算法，non-local 更容易于捕获空域或时域长距离的信息。放置于合适的位置可以减少参数量和运算量的同时，增大感受野提升性能。

其缺点在于该模块的计算量 FLOPs 和特征图的大小有很强的相关性，当特征图偏大的时候，其会引入很大的计算量，通用性没有卷积模块强。

3.5.2 STM[11]

STM 来自于 ICCV2019，其认为时空特征与运动特征的提取是行为识别中的核心的问题。在 I3D 中，利用 3D 卷积来提取时空特征，利用光流分支来提取运动特征。然而前面也提及，这两者所需耗费的计算资源都过大，若要实现模型的实行性，需要找到二者的 2D 替代。所以 STM 模型就是以一种 2D 的 CNN 网络高效地替代 3D 卷积和光流特征。这种方法能够在以时序关联和场景关联的多个数据集上都取得 SoTA 效果。

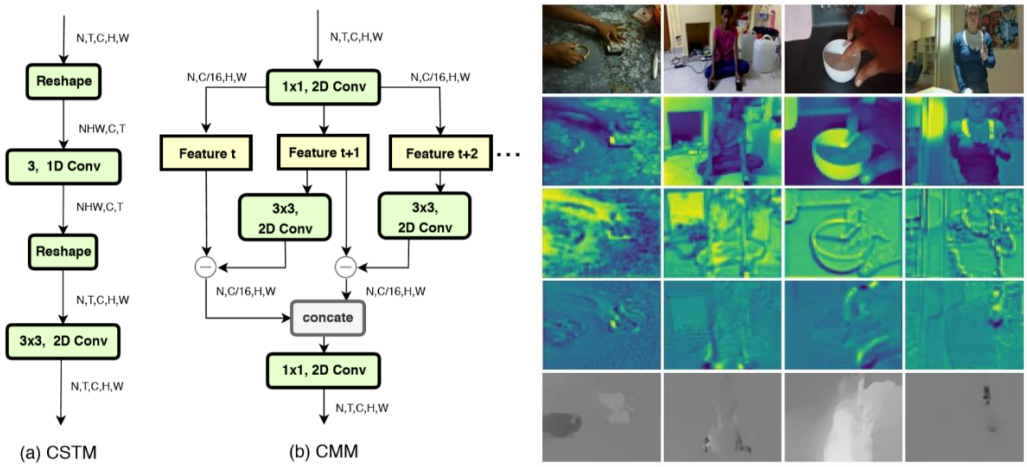


图 27 CSTM 和 CMM 的结构(left), STM block 的特征可视化(right), 第一行为输入帧. 第二行是 Conv2 1 block 的特征图. 第三行是 CSTM 的特征图. 第四行的 CMM 的特征图. 最后一行是光流特征图.[11]

3.5.2.1 STM 中的时空特征模块 CSTM

时空特征代表能够表征视频中的主体的特征。Channel-wise 的时空模型(CSTM)可以认为是一种 3D-CNN 的轻量化改进(和 2+1D-CNN 做法基本一致)，其模型图见上图左(a)，将一个 3D-CNN 分解为时域上的 channel-wise 的 1D-CNN 与空域上的 2D-CNN 的串联。Channel-wise 1D-CNN 等同于全通道的分组卷积，其不仅有利于不同的语义特征的相互隔离，而且能够有效减小计算量。从上图右可以看出第二行的特征图中运动的主体部分被激活。

3.5.2.2 STM 中的动作特征模块 CMM

相比于时空特征对于动作主体表观信息的提取，运动特征更多地是对于帧间边缘运动特征的描述。Channel-wise 的运动模型(CMM)对标的是光流特征，其主要结构如上图左(b)所示。利用 1*1 卷积缩减维度之后，对于相邻帧做带有 channel-wise 的 3*3 卷积的帧间差分，之后再合并在一起扩增维度。由于帧间差分需要保持维度的对应，所以 3*3 卷积必须是分组卷积。从上图右中可以观察到，第三行的特征关注局部边缘的运动，与光流特征类似。

3.5.2.3 STM 整体架构

CSTM 和 CMM 可以结合 residual block 的结构合并入一个 STM block 中。由于 STM block 适用性强，所以其可以嵌套入任意 ResNet 结构中，本文采用的结构是 ResNet50，用 STM block 代替所有 residual block。视频端采用类似 TSN 的段落采样技术，将视频分为 T 段，在每段中采样 1 帧并合并，最后接入 STM 网络进行分类。

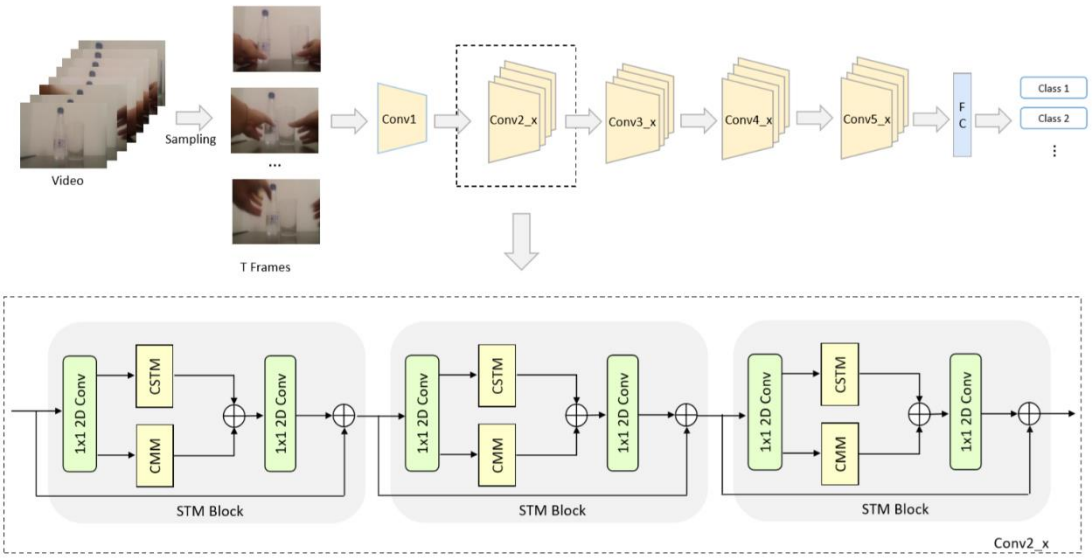


图 28 STM 整体架构.[11]

STM 在各种数据集中都取得了 SoTA 的效果。可以证明其强大的通用性能。另外，这是一个 2D 网络，保持最高性能的同时，也具有与 TSN 相当的 FLOPs。比起 I3D 和 ECO 来说，仅仅耗用 1/10 和 1/2 的 FLOPs 就能实现 6\%左右准确率的提升。

3.5.2.4 STM 方法的优缺点

STM 的优点（1）首先，其是 2D 网络近似 3D 网络的一种成功尝试，本文在 2+1D 模型上做了一定的改动，使得模型更加轻量化。同时，也是 2D 网络对于光流特征的一种有效替代。（2）STM-Block 可以作为一种基本结构嵌入其他结构中，除了文中提到的 ResNet，对于 DenseNet，SENet 等结构也适用。

STM 的缺点，作为 2019 年领域的 SoTA，STM 模型确实十分完备，但在动作特征即光流特征的近似上，仅仅采用卷积后帧查的方法建模确实还欠缺些许考虑，这也可能是为何其性能在有些数据集上与一些加光流的老模型指标相似甚至不如那些老模型的原因。

3.5.3 新结构的总结

跳出 3D 和双流法的框架，我们发现行为识别中还有很多有趣的问题等待我们发现。比如利用 Attention 来直接捕获相隔多帧的区域关联性，而非通过 3D 卷积一点一点地扩大感受野。另外，3D 和光流的冗余性也被深挖，2+1D 就捕获到了 3D 的时间冗余，OFF 从原理上找到一种光流的快速等效，STM 则是精心设计了近似二者的 2D 网络。事实证明，尽管 3D 和光流都是优秀的视频特征，其仍然存在改进与轻量化的空间，这在最近一段时间内也将会是研究的热点。

而 2D 网络和新架构的兴起同时也预示着更多的可能，比如利用网络搜索 NAS 来得到一个高效的结构，再比如结合跨模态信息，利用 video-bert 等新模型来弱监督地学习到一些视频特征。这些都可能会是将来的发展方向。

四、总结

本文主要总结了用于视频时序特征建模的四类模型。

LSTM 以 RNN/LSTM/GRU 为主要结构的特征提取器，主要用于早期的视频预测任务，最早始于雷达降水预测任务 ConvLSTM，后来融入光流/轨迹特征，发展为 TrajGRU 模型。这种模型在早期确实有不错的性能，但是由于 LSTM/GRU 等时序结构较为固定，扩展性能差，同时时序结构这种递推形式也不利于并行计算，会极大拖延运行速度，可能仅适合于雷达降水这种对于实时性要求不高的任务。所以这种方法没有得到大幅推广。

Flow/Two stream 以双流特征为主的视频特征提取器主要出现在行为识别领域，其中 TSN 仍然是行为识别领域经久不衰的 baseline。其主要思路是利用相邻帧的光流来补充视频的动作信息。虽然其能取得不错的性能，但光流法并不是能够融入网络的端到端算法，其本身就需要较大的计算量，这会极大地影响计算的实时性。所以后期中有 OFF，STM 这种借用 2D 网络来模拟实现光流效果的方法被不断提出与改进。

3D-Net 3D-Net 是与双流法齐名的方法，借用 3D-CNN 来提取视频特征，与光流不同，经过堆叠，其提取的视频特征具有更大的时间感受野，可以用于捕获一些长时跨度的视频特征。其一定程度上是与光流特征互补的，I3D 就是基于这一点将二者融合。然而，3D-Net 也会引入更大的计算成本。为了节约这一成本，更多的 2D 等效网络被提出，如 2+1D-CNN，STM 等。其旨在用低成本的运算代替 3D-CNN 结构。

另外还有一些结构不完全在以上分类的网络，比如 Non-Local 网络，利用 Attention 机制捕获时间与空间上的长距离关系。STM 与 2+1D 网络均是对于双流框架和 3D 框架的一种 2D 轻量级改良。

总而言之，视频特征的提取在前期已经积累了较高的性能的前提下，开始不断向着轻量化前进，在未来，还可能出现更多基于 attention、transformer 模型的视频提取子，又或者是结合以上特点的网络结构搜索出的视频特征提取模块。

参考文献

[1] Xingjian SHI, Chen Z, Wang H, et al. Convolutional LSTM network: A machine

learning approach for precipitation nowcasting[C]//Advances in neural information processing systems. 2015: 802-810.

[2] Shi X, Gao Z, Lausen L, et al. Deep learning for precipitation nowcasting: A benchmark and a new model[C]//Advances in neural information processing systems. 2017: 5617-5627.

[3] Simonyan K, Zisserman A. Two-stream convolutional networks for action recognition in videos[C]//Advances in neural information processing systems. 2014: 568-576.

[4] Wang L, Xiong Y, Wang Z, et al. Temporal segment networks: Towards good practices for deep action recognition[C]//European conference on computer vision. Springer, Cham, 2016: 20-36.

[5] Sun S, Kuang Z, Sheng L, et al. Optical flow guided feature: A fast and robust motion representation for video action recognition[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2018: 1390-1399.

[6] Ji S, Xu W, Yang M, et al. 3D convolutional neural networks for human action recognition[J]. IEEE transactions on pattern analysis and machine intelligence, 2013, 35(1): 221-231.

[7] Tran D, Bourdev L, Fergus R, et al. Learning spatiotemporal features with 3d convolutional networks[C]//Proceedings of the IEEE International Conference on Computer Vision. 2015: 4489-4497.

[8] Carreira J, Zisserman A. Quo vadis, action recognition? a new model and the kinetics dataset[C]//proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017: 6299-6308.

[9] Tran D , Bourdev L , Fergus R , et al. A Closer Look at Spatiotemporal Convolutions for Action Recognition [J]. 2018.

[10] Wang X, Girshick R, Gupta A, et al. Non-local neural networks[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2018: 7794-7803.

[11] Jiang B, Wang M M, Gan W, et al. Stm: Spatiotemporal and motion encoding for action recognition[C]//Proceedings of the IEEE International Conference on Computer Vision. 2019: 2000-2009.