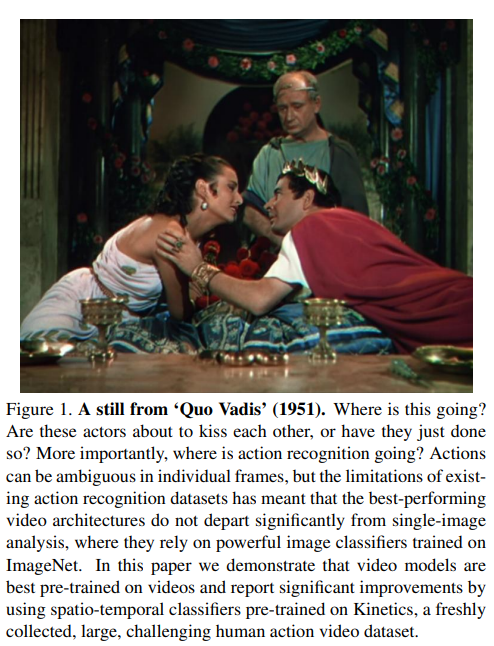
# Title：

**Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset**

New Model：I3D，把一个2D模型扩展到3D（可以使用已经构建好的2D模型了）。

New Dataset：Kinetics Dataset



如果只通过一张图片，不能区分很多动作，比如左图光靠这一张图不能区分这两个演员是否已经亲过。更重要的问题是，接下来如何发展呢？视频领域缺少这一个大的数据集，导致现在学界无法研究视频理解的框架，本文证实了视频的模型最好要在视频上去训练。具体来说，本文提出了一个又大又好的数据集，然后在这个数据集上训练了一个新的模型，叫I3D模型。

# Abstract：

因为目前的数据集（UCF-101，HMDB-51分别有101类和51类）都太小了，没有办法发挥深度神经网络的能力，大家在这两个数据集上的结果都相似，但是这些方法肯定有强有弱，只是数据集无法识别好的网络架构了。所以本文提出了一个新的数据集 Kinetics，它有400类，每一类都有超过400个vedio clip（10s左右）。他们在这个数据集上测试了之前工作的结果，然后利用这个数据集预训练了一个模型在UCF和HMDB上取得了很好的结果。

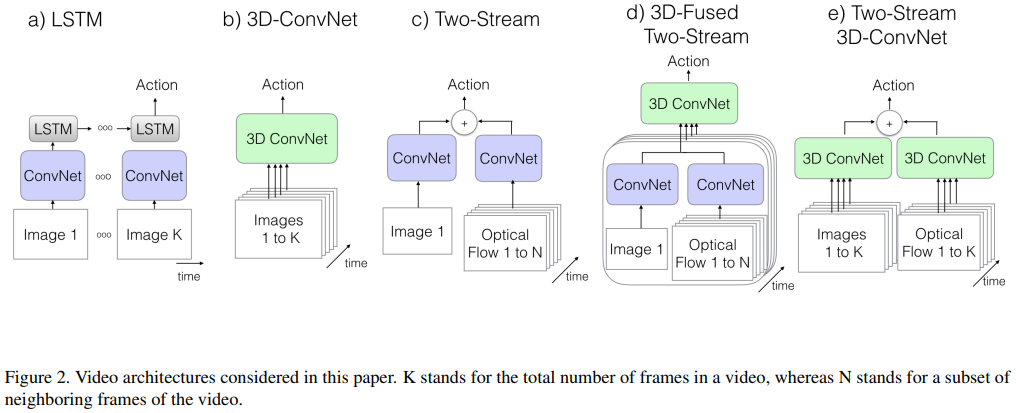
他们提出了一个双流扩充3D卷积神经网络（Two-Stream Inflated 3D ConvNet（I3D））：把一个现成的网络拿过来直接扩充卷积核和池化层，将3\*3扩充成3\*3\*3。这样的好处就是可以利用起来在2D网络架构设计上的经验。最后I3D达到80.9% on HMDB-51 and 98.0% on UCF-101。

# Introduction：

自从用ImageNet训练网络开始，深度学习慢慢把其他任务控制了，因为在其数据集上训练的网络可以很有效地进行抽特征。在视频领域这种范式还未被证明有效，因为本领域数据集量太少，所以作者提出了新的数据集Kinetics（每一个视频都来自于独一无二的YouTube视频）。

作者使用卷积＋LSTM，3D卷积神经网络，双流网络这三种网络在这个数据集上比较，这三种方法也是代表在视频领域大家是如何利用时序信息的。最后发现效果参差不齐，所以他提出了I3D。

# Action Classification Architectures

在2D领域已经有主导的神经网络了。作者通过对比卷积＋LSTM，3D卷积神经网络，双流网络这三种网络，做了大量的实验，最后提出了基于扩充Inception+双流的I3D。

a的方法把视频更多的看做一个图像分类的问题，因为对于Image1~K，是一张张通过卷积神经网络的，然后抽完特征后进入LSTM进行时序建模，最后输出最后时间戳上的结果加上全连接进行分类。这种方式在之前的数据集上表现不好，最后留下3D卷积神经网络和双流网络。

b是3D卷积神经网络，它非常简单粗暴，把视频劈成K个图片组成一个volume作为输入，所以这个网络可以进行时空学习，导致参数量变得很大，在小的数据集上不好训练。

c是双流网络，使用光流进行时间建模，然后用一个2D的卷积处理图片，最后把这两个网络的结果做一个汇聚和late fusion，进入全连接层分类。

d的方法是把b和c有机的结合在了一起，前面还是双流网络，但是汇聚late-fusion用一个小的3D卷积神经网络代替，这种处理是early fusion（在还没出结果的时候把他们融合在一起）

e是作者提出的方法，在有足够数据量的情况下，还是3D在前比较好，但是有些东西学习不到，需要光流帮助它。

因为两个分支都是3D的，所以没有必要再加fusion处理了，后面直接加权平均输出结果

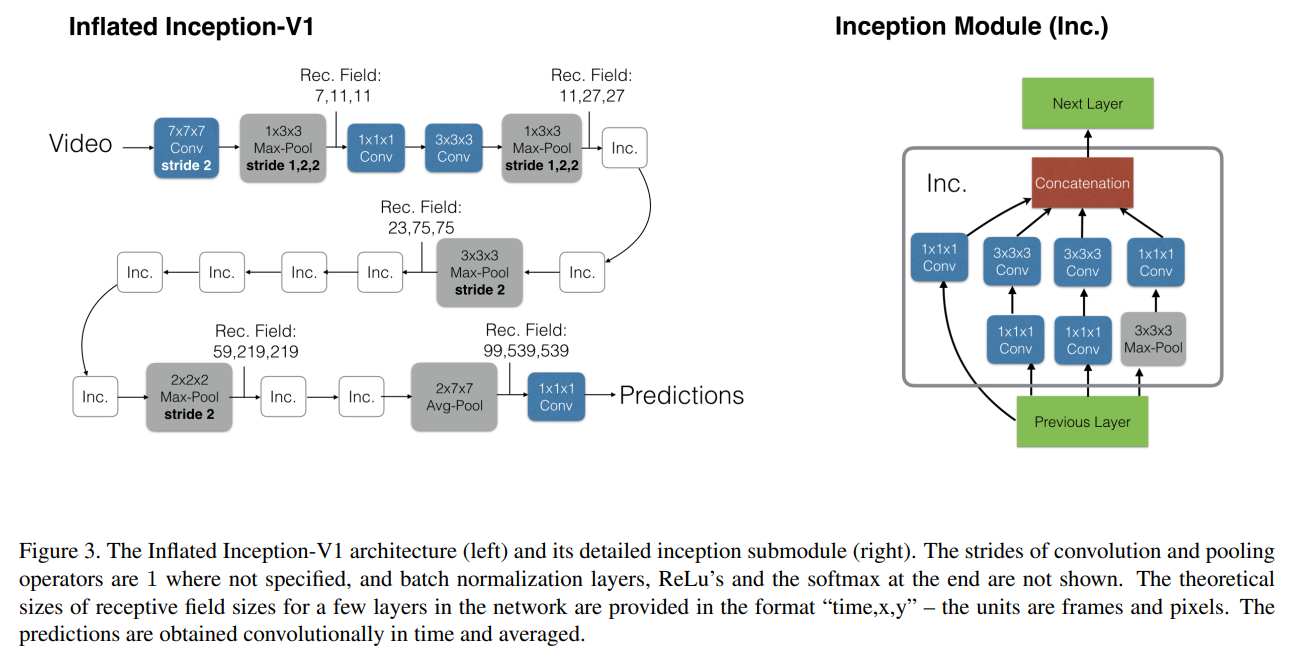
## The New: Two-Stream Inflated 3D ConvNets

**Inflating 2D ConvNets into 3D.**

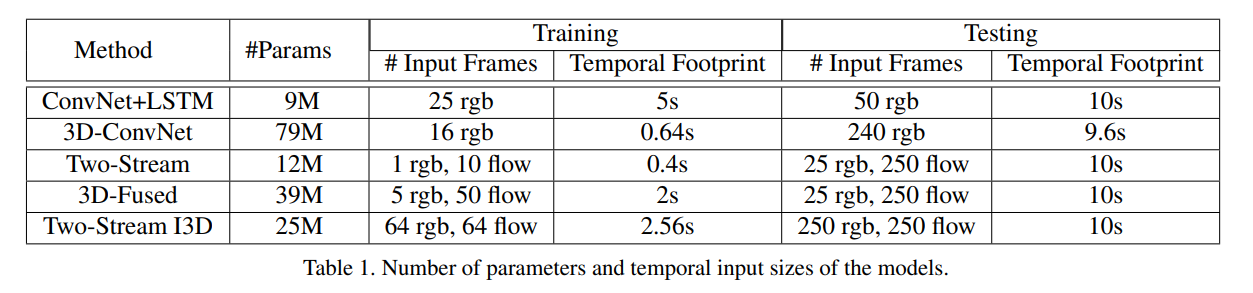
通过从二维架构开始，将所有的过滤器和池化核膨胀，赋予它们额外的时间维度，整体架构保持不变过滤器通常是方形的，只需使它们成为立方体N × N的过滤器变成N × N × N。

**Bootstrapping 3D filters from 2D Filters.**

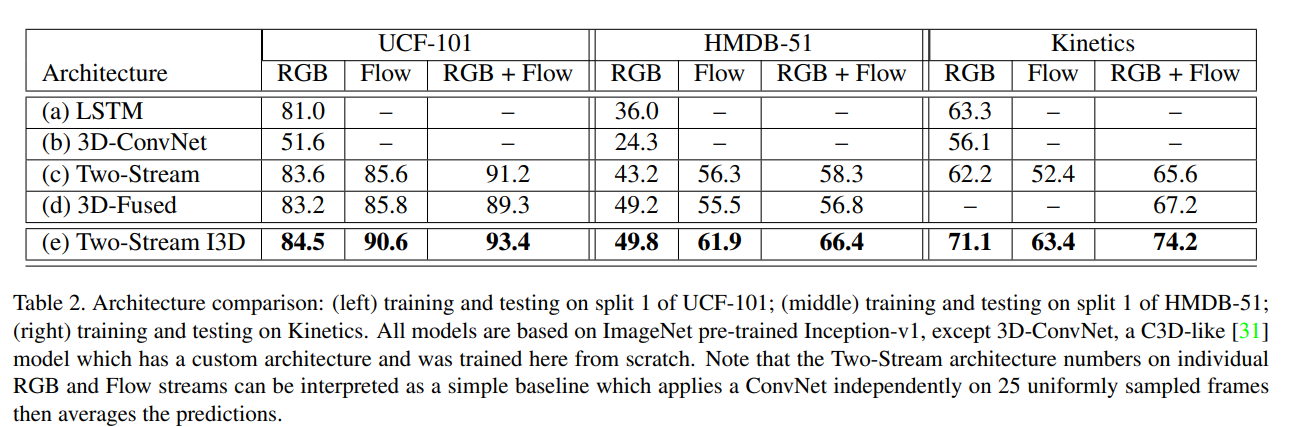
如何从ImageNet上已经训练好的模型出发，让它去初始化一个3D模型，然后在初始化好的3D模型上面继续训练。作者在2D的filter在时间的维度上复制粘贴n次，将输入图片复制粘贴n次，变成“非静止画面”，最后得到n份的W\*H，最后输出除以n，就得到了W\*X，这样就得到了3D的输出和2D的输出是一样的，后面的模块也能够接的住输入了。

注意时间维度不要做下采样，所以变成1\*3\*3而不是3\*3\*3

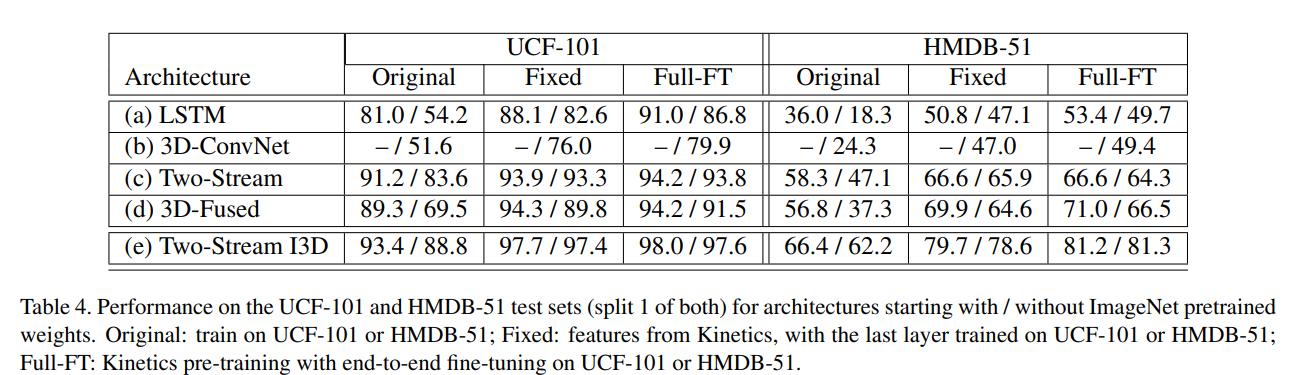
模型大小：



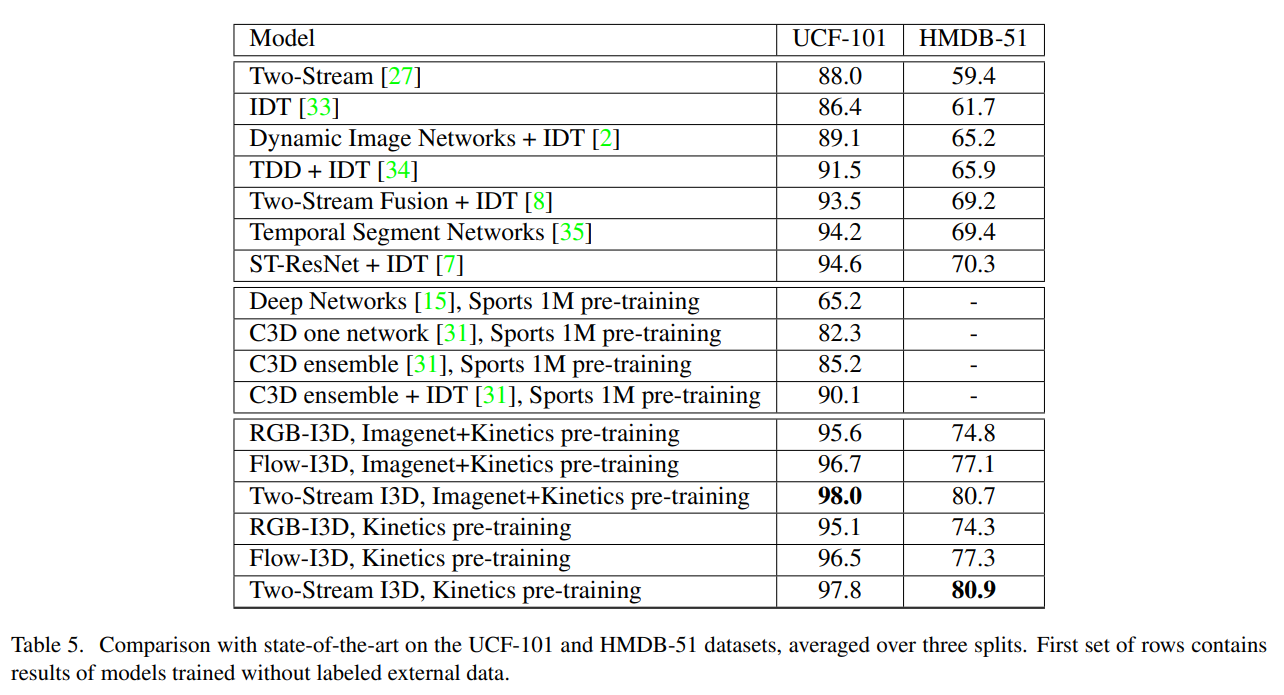
# Experimental Comparison of Architectures

在三个数据集上的表现。

# Comparison with the State-of-the-Art

迁移学习的表现：  
在Full-fine-tune的情况下结果最好。

跟其他最好方法的比较：



# Discussion

在视频领域迁移学习有没有用呢？以本篇文章得到的结论来说，效果是非常明显的。但是其他的一些下游任务，作者还没去做。还有关于模型的探索也没有做的全面。