动作识别

讨论two-stream 的fusion机制

动作识别

Spatial Feature

空间特征。边缘、纹理、形状等

Temporal Feature

时序特征。运动轨迹、变化速率等

动作识别

传统方法

利用光流场提取运动轨迹，再沿着轨迹提取HOF，HOG等特征，利用FV等方法对特征进行编码，再使用机器学习进行分类。

Two-stream

利用视频帧和光流图分别输入到两个分支的卷积网络，在fusion该两者特征，进行分类

3D CNN

直接输入原视频，使用3D网络可以提取其中的时间和空间特征。

RNN

RNN通过循环神经网络，能处理序列信息，提取视频中的运动特征

Two-stream

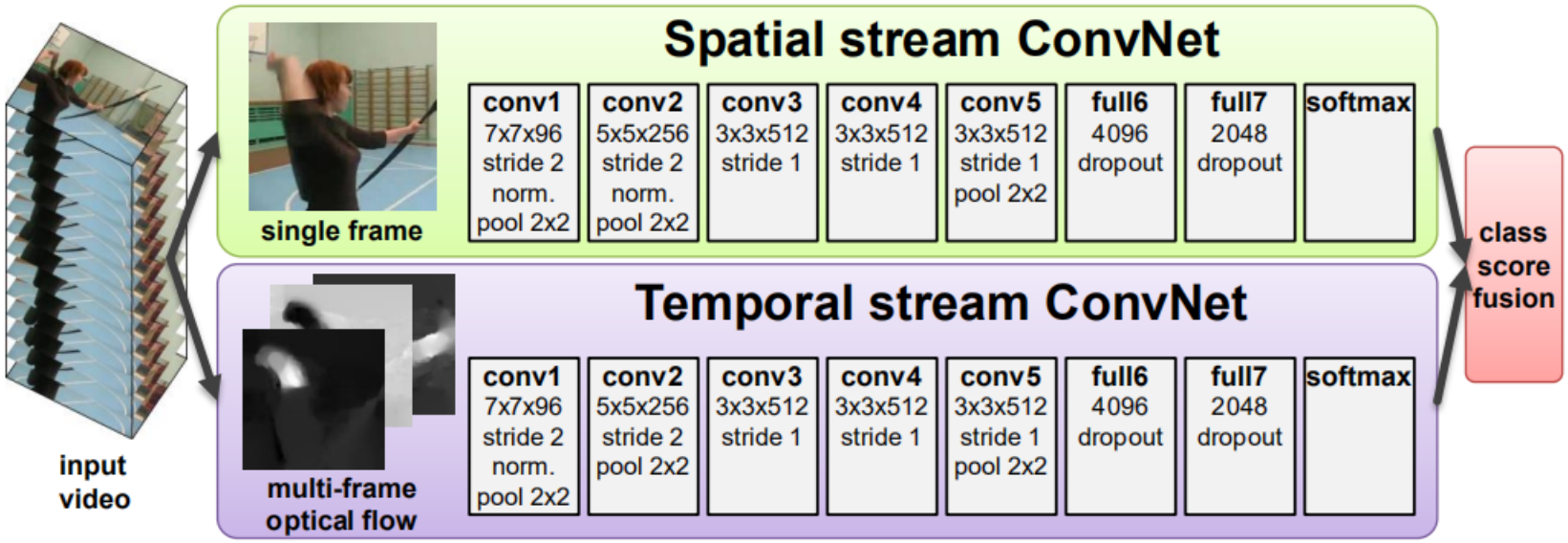


Figure 1: **Two-stream architecture for video classification.**

Simonyan K, Zisserman A. Two-stream convolutional networks for action recognition in videos. 2014

Spatial fusion

- Sum fusion
- Max fusion
- Concatenation fusion
- Conv fusion
- Bilinear fusion

Feichtenhofer C, Pinz A, Zisserman A. Convolutional two-stream network fusion for video action recognition.2016

Spatial fusion

Sum fusion $$y^{\text{sum}}_{i,j,d} = x^{a}_{i,j,d} + x^{b}_{i,j,d},$$

Max fusion $$y^{\max}_{i,j,d} = \max\{x^{a}_{i,j,d}, x^{b}_{i,j,d}\},$$

Feichtenhofer C, Pinz A, Zisserman A. Convolutional two-stream network fusion for video action recognition.2016

Spatial fusion

Concatenation fus $y^{\text{cat}}_{i,j,2d} = x^{a}_{i,j,d} \qquad y^{\text{cat}}_{i,j,2d-1} = x^{b}_{i,j,d},$

Conv fusion $\mathbf{y}^{\text{conv}} = \mathbf{y}^{\text{cat}} * \mathbf{f} + b, \ \mathbf{f} \in \mathbb{R}^{1\times1\times2D\times D}$

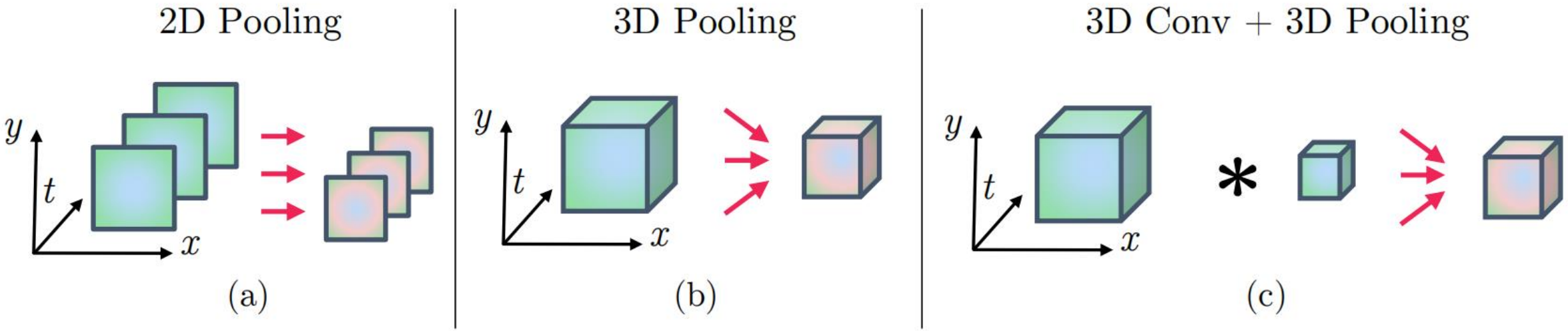
Feichtenhofer C, Pinz A, Zisserman A. Convolutional two-stream network fusion for video action recognition.2016

Spatial fusion

Bilinear fusion $$\mathbf{y}^{\text{bil}} = \sum_{i=1}^{H}\sum_{j=1}^{W} \mathbf{x}_{i,j}^{a\top} \otimes \mathbf{x}_{i,j}^{b}.$$

Feichtenhofer C, Pinz A, Zisserman A. Convolutional two-stream network fusion for video action recognition.2016

Temporal fusion



where fusion