HCN:分层式共现网络

Co-occurrence Feature Learning from Skeleton Data for Action Recognition and Detection with Hierarchical Aggregation

分享人：程文卓

前提

熟悉卷积神经网络相关知识！！！

目录

1、概述及背景

1、概述及背景

问题一：论文主要做了什么事情?

海康威视这篇论文主要做的是事情是从人体关键点序列（骨架数据）进行行为检测和识别

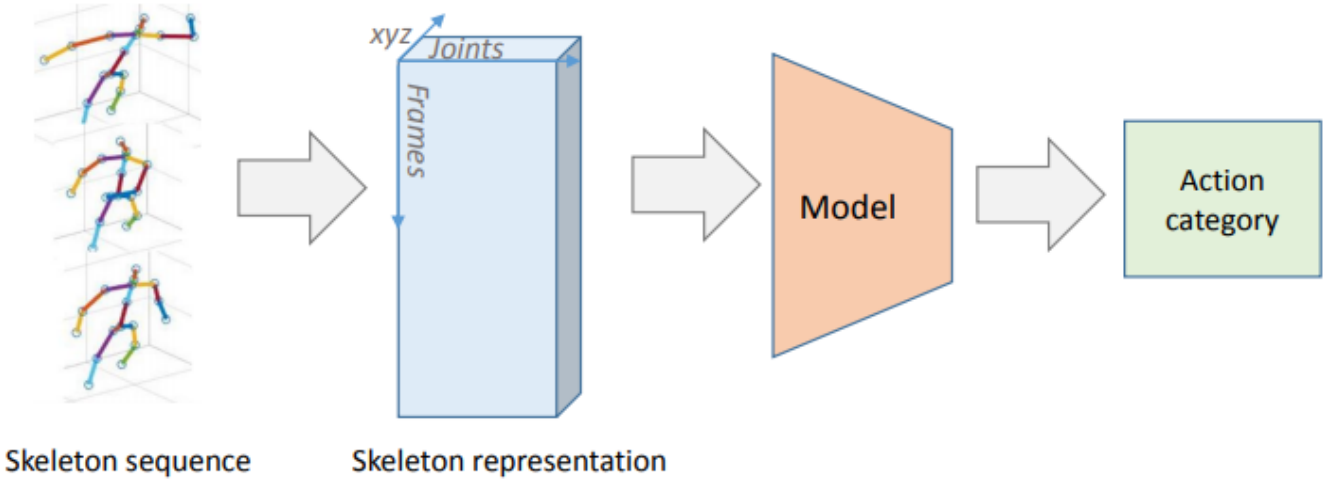


Figure 1: Workflow for skeleton-based human action recognition.

1、概述及背景

问题二：为什么使用骨架数据而不使用RGB图片数据?

一方面，骨架数据在背景噪声中具有固有的稳健性，
并且能提供人体动作的抽象信息和高层面特征。

另一方面，与 RGB 数据相比，骨架数据的规模非常小，这
让我们可以设计出轻量级且硬件友好的模型。

1、概述及背景

问题三：早期研究是如何进行的？

第一类：CNN模型，此类模型优势是提取数据高层面信息方面能力出色

第二类：RNN模型，此类模型的优势是善于处理长期时间依赖关系

缺点：目前这两类已有的研究都没有关注完整骨架所有关节点的全局共现特征

1、概述及背景

问题四：本论文提出了什么方法?

第一：主要提出了全局共现特征的一种提取方法,即通过转置关键点序列,将被卷积的通道维度由坐标维度变为关键点维度,从而提取关键点的全局共现特征

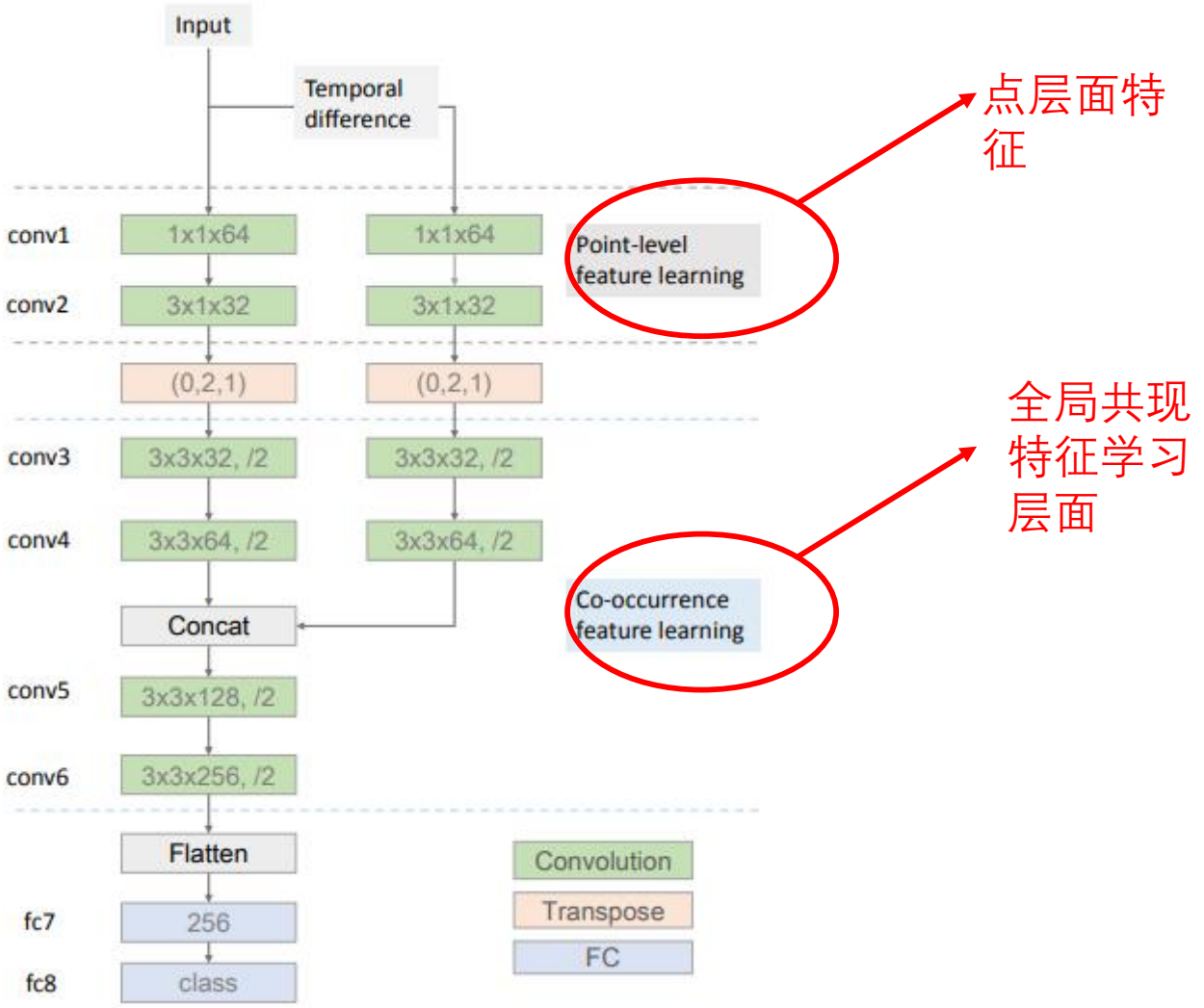
第二类：提出关于多人交互动作问题的解决方法

第三类：提出了一种时域动作检测方法（借鉴Faster RCNN的proposal思想）

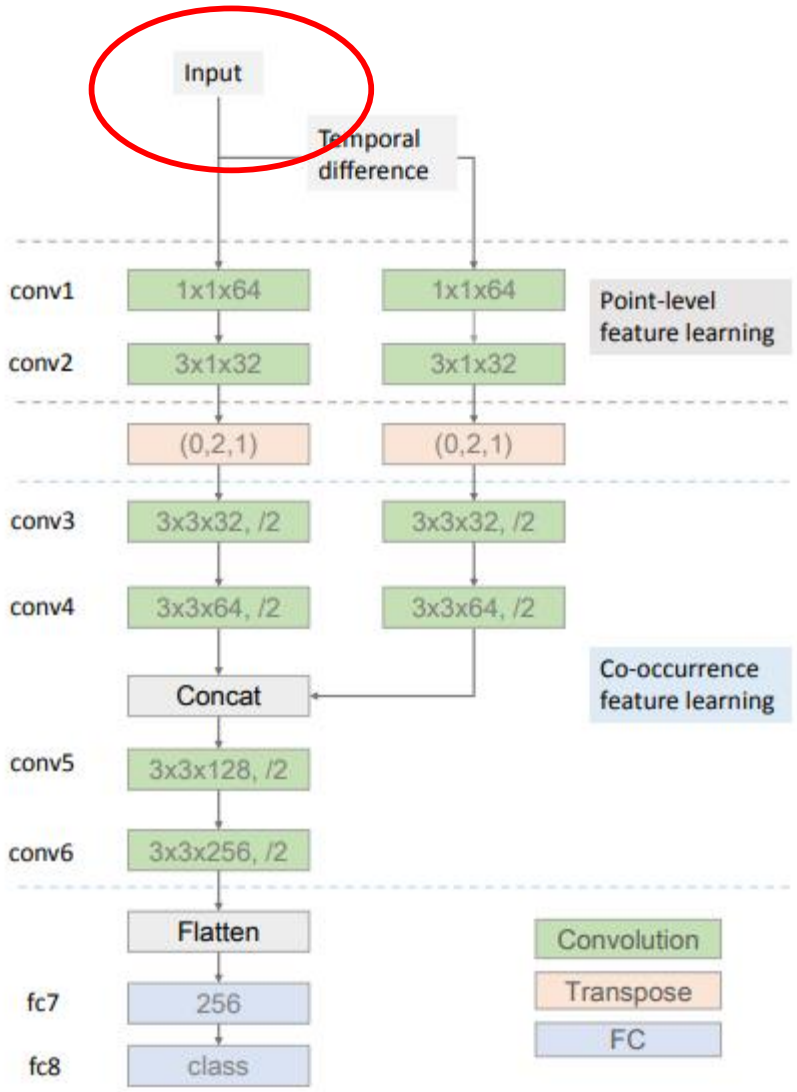
2、论文方法

1、单人动作识别

论文提出的端到端的分层式共现网络（HCN）的概况。绿色模块是卷积层，其中最后一维表示输出通道的数量。后面的「/2」表示卷积之后附带的最大池化层，步幅为 2。转置层是根据顺序参数重新排列输入张量的维度。conv1、conv5、conv6 和 fc7 之后附加了 ReLU 激活函数以引入非线性。

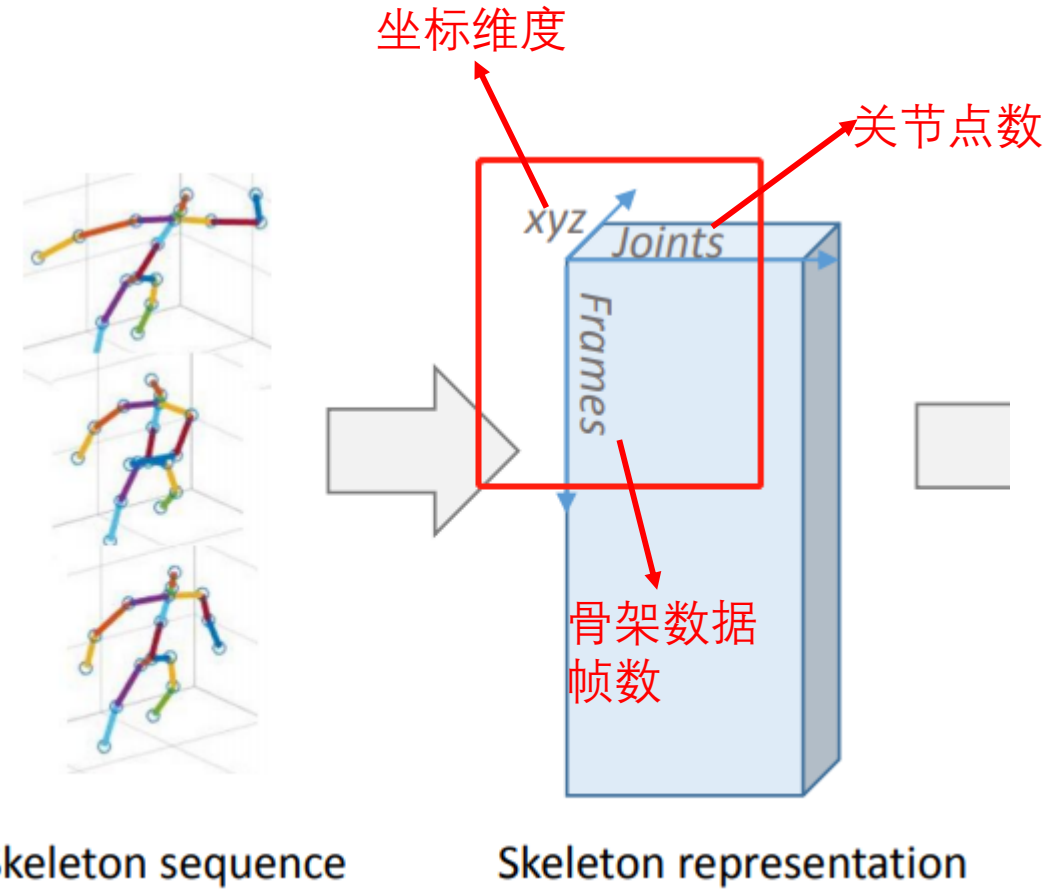


1、单人动作识别

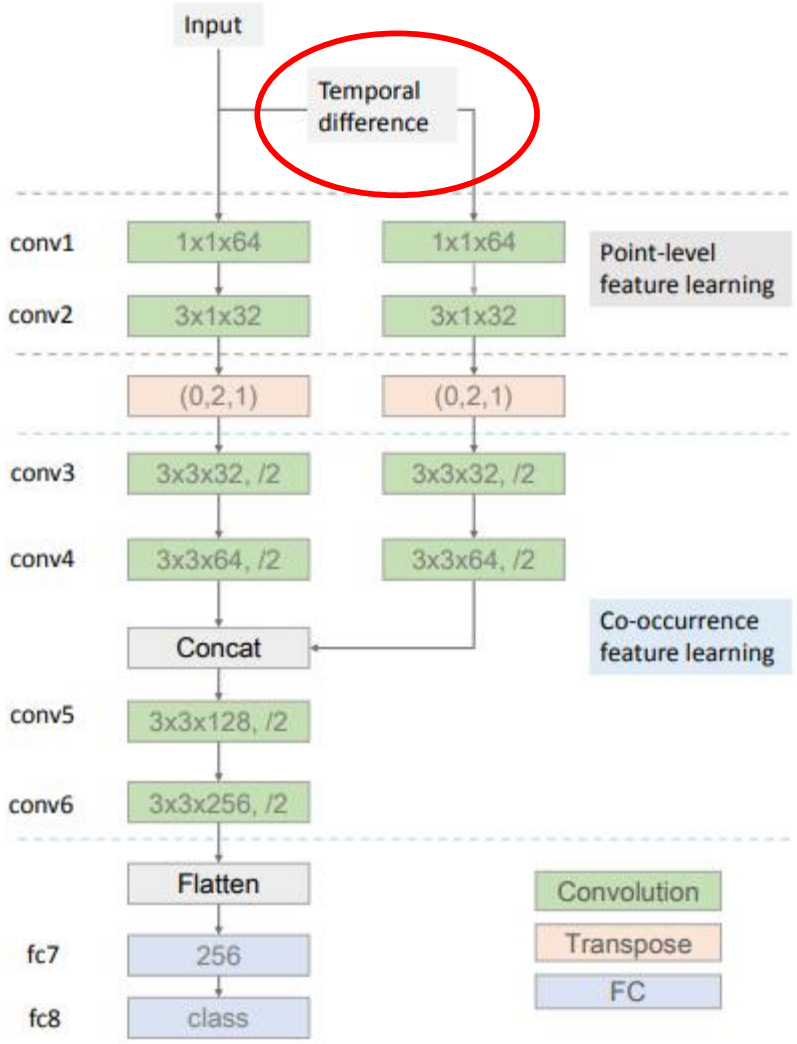

Input
Temporal difference
conv1
1x1x64
1x1x64
Point-level feature learning
conv2
3x1x32
3x1x32
(0,2,1)
(0,2,1)
conv3
3x3x32, /2
3x3x32, /2
conv4
3x3x64, /2
3x3x64, /2
Concat
Co-occurrence feature learning
conv5
3x3x128, /2
conv6
3x3x256, /2
Flatten
fc7
256
fc8
class
Convolution
Transpose
FC

1、单人动作识别

HCN的输入部分：基于 CNN 的方法可以通过将时间动态和骨架关节分别编码成行和列而将骨架序列表示成一张图像，然后就像图像分类一样将图像输入 CNN 来识别其中含有的动作。具体说，将骨架序列表示成了一个形状帧×关节×3（最后一维作为通道）的张量



1、单人动作识别


Input
Temporal difference
conv1
1x1x64
1x1x64
Point-level feature learning
conv2
3x1x32
3x1x32
(0,2,1)
(0,2,1)
conv3
3x3x32, /2
3x3x32, /2
conv4
3x3x64, /2
3x3x64, /2
Concat
Co-occurrence feature learning
conv5
3x3x128, /2
conv6
3x3x256, /2
Flatten
fc7
256
fc8
class
Convolution
Transpose
FC

1、单人动作识别

显式的骨架运动：关节的时间运动是识别潜在行为的关键线索。虽然时间演化模式可以通过CNN隐式学习，但我们认为显性建模更可取。因此，我们引入了骨架运动的表示并将其明确地馈送到网络中。

J=(x,y,z)是3D关节坐标

N是关节数

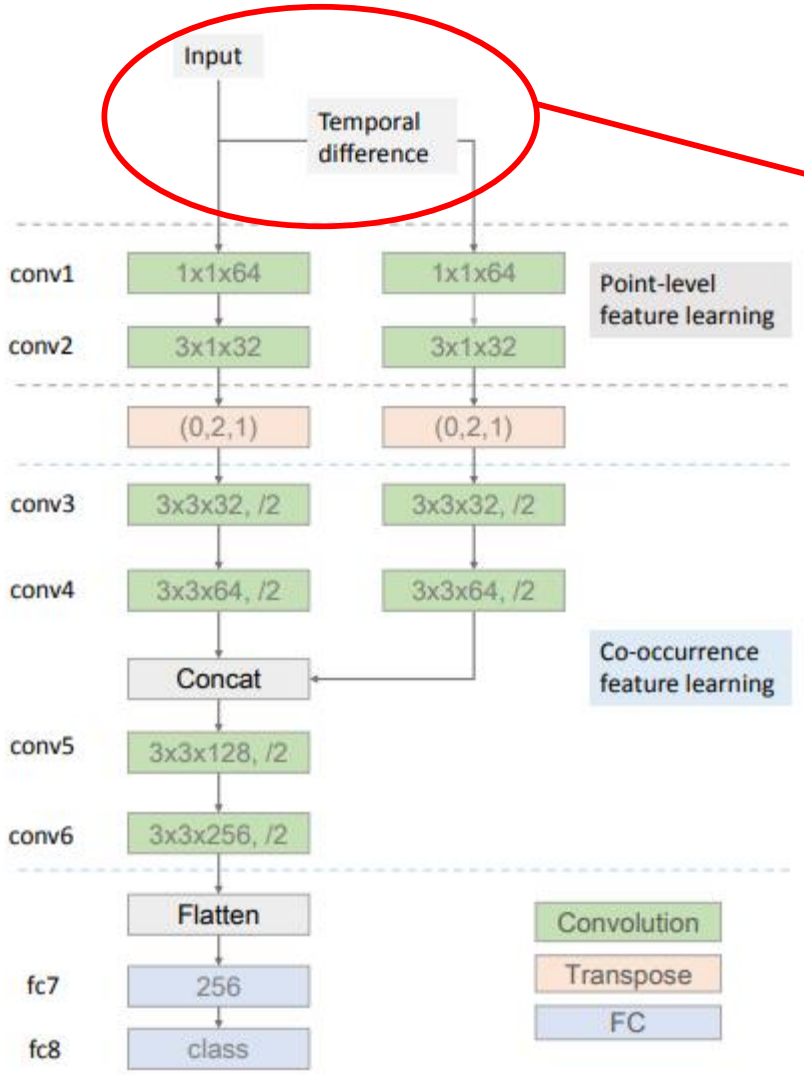
第t帧中人的骨架表示：

$$\boldsymbol{S}^t = \{\boldsymbol{J}_1^t, \boldsymbol{J}_2^t, \ldots, \boldsymbol{J}_N^t\}$$

骨架运动：为两个连续帧之间每个关节的时间差异

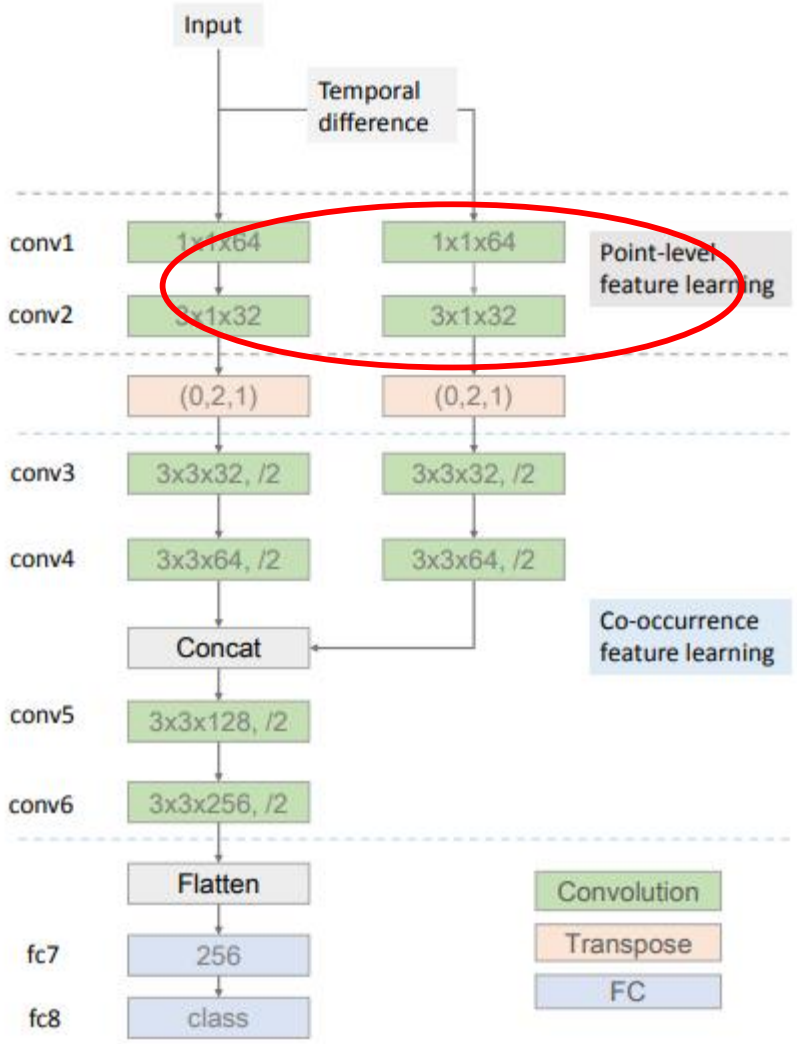
$$\begin{aligned}\boldsymbol{M}^t &= \boldsymbol{S}^{t+1} - \boldsymbol{S}^t \\ &= \{\boldsymbol{J}_1^{t+1} - \boldsymbol{J}_1^t, \boldsymbol{J}_2^{t+1} - \boldsymbol{J}_2^t, \ldots, \boldsymbol{J}_N^{t+1} - \boldsymbol{J}_N^t\}.\end{aligned}$$

1、单人动作识别



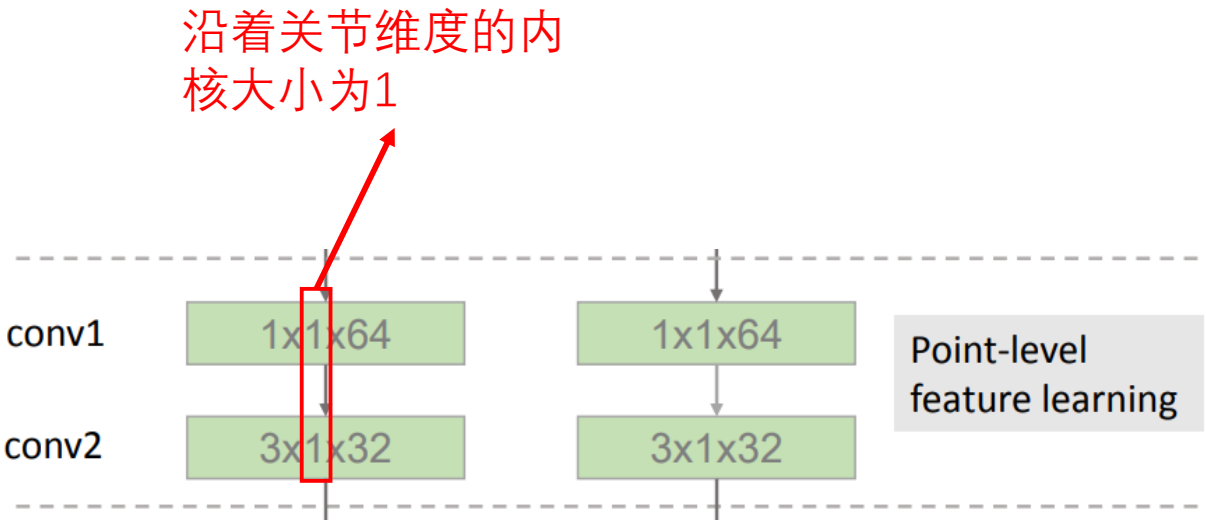
原始骨架坐标S和骨架运动M通过双流范式独立地馈送到网络中。后续层中跨通道连接它们。

1、单人动作识别

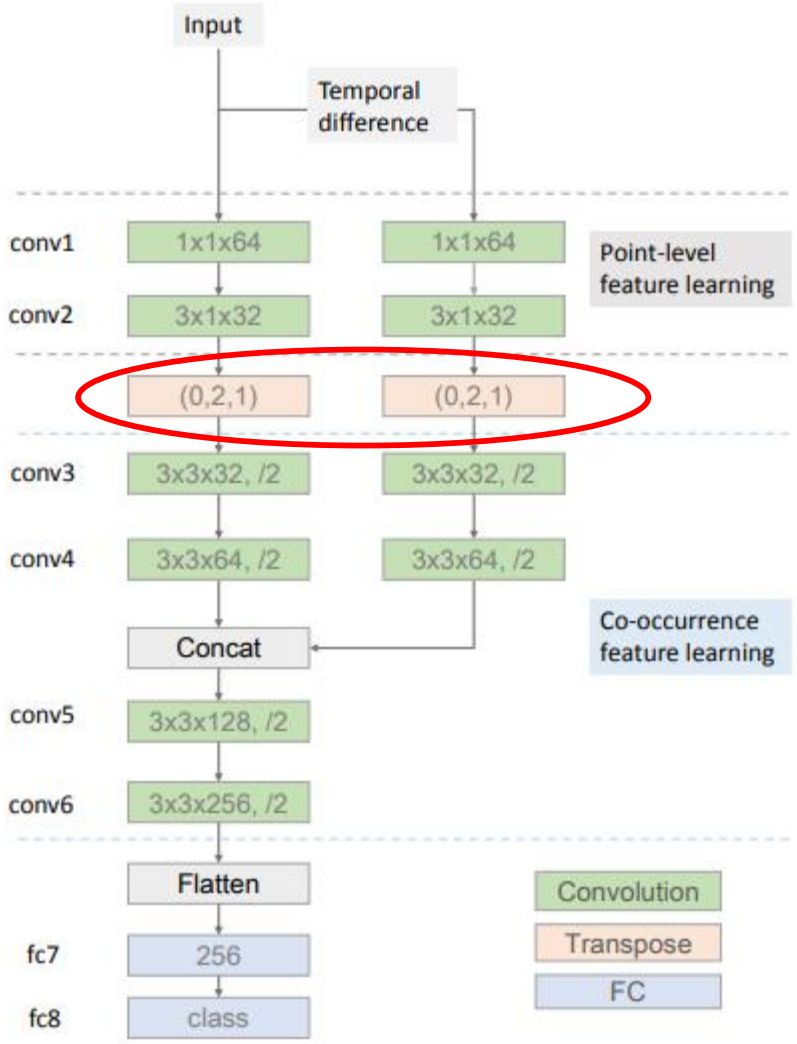

Input
Temporal difference
conv1
1x1x64
1x1x64
Point-level feature learning
conv2
3x1x32
3x1x32
(0,2,1)
(0,2,1)
conv3
3x3x32, /2
3x3x32, /2
conv4
3x3x64, /2
3x3x64, /2
Co-occurrence feature learning
Concat
conv5
3x3x128, /2
conv6
3x3x256, /2
Flatten
fc7
256
fc8
class
Convolution
Transpose
FC

1、单人动作识别

在第一层面中，点层面特征用1×1（conv1）和n×1（conv2）卷积层编码。由于沿着关节维度的内核大小保持为1，因此他们被迫独立地从每个关节的3D坐标学习点层面表示。

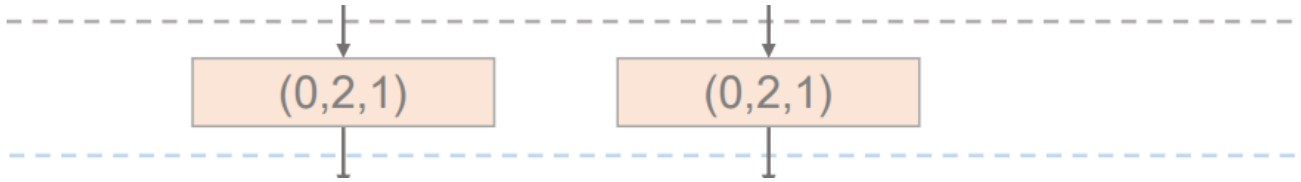


1、单人动作识别


Input
Temporal difference
conv1
1x1x64
1x1x64
Point-level feature learning
conv2
3x1x32
3x1x32
(0,2,1)
(0,2,1)
conv3
3x3x32, /2
3x3x32, /2
conv4
3x3x64, /2
3x3x64, /2
Concat
Co-occurrence feature learning
conv5
3x3x128, /2
conv6
3x3x256, /2
Flatten
fc7
256
fc8
class
Convolution
Transpose
FC

1、单人动作识别

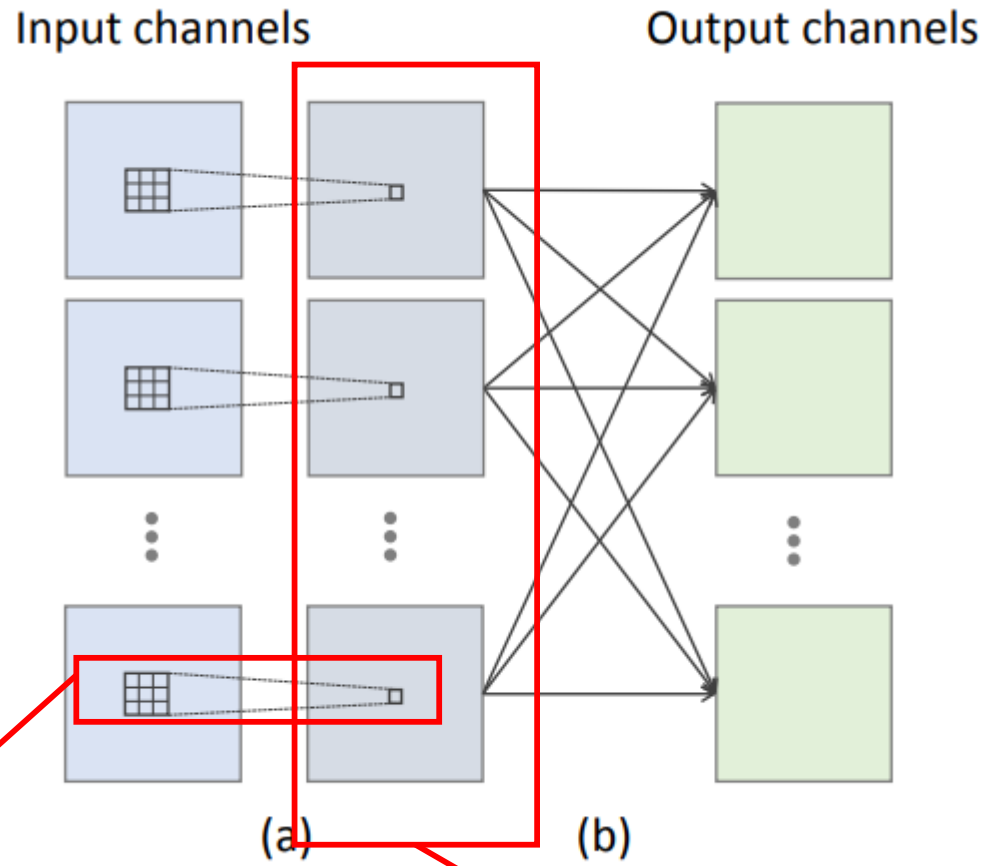
使用参数（0,2,1）转置点层面卷积输出，以便将关节维度移动到张量的通道。那么此时输入张量变为：帧数×坐标维度×关节数（关节数作为通道）



1、单人动作识别

问题五：为什么要转置将关节数的维度作为通道，通道有什么特别之处?

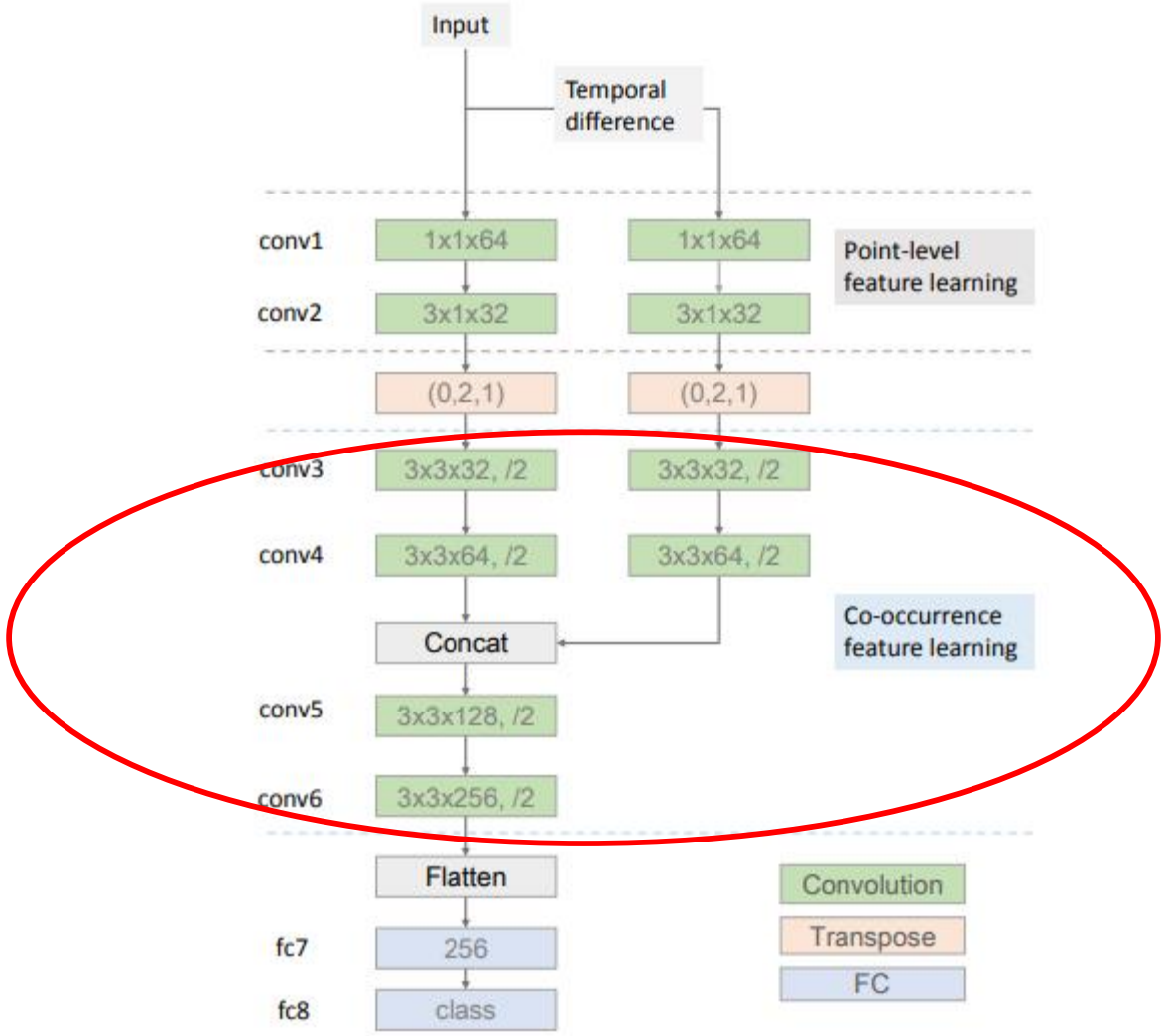
因为当关节数的维度不是通道时，共现特征只能在局部聚合，这可能无法捕获像穿鞋这样的动作中涉及的远程联合交互。为此，我们认为在全局范围内聚合共现特征非常重要，可以提高行动识别性能。通过将关节维度放入CNN输入的通道中可以很容易地实现。



（a）跨空间域的局部特征聚合（高度和宽度）

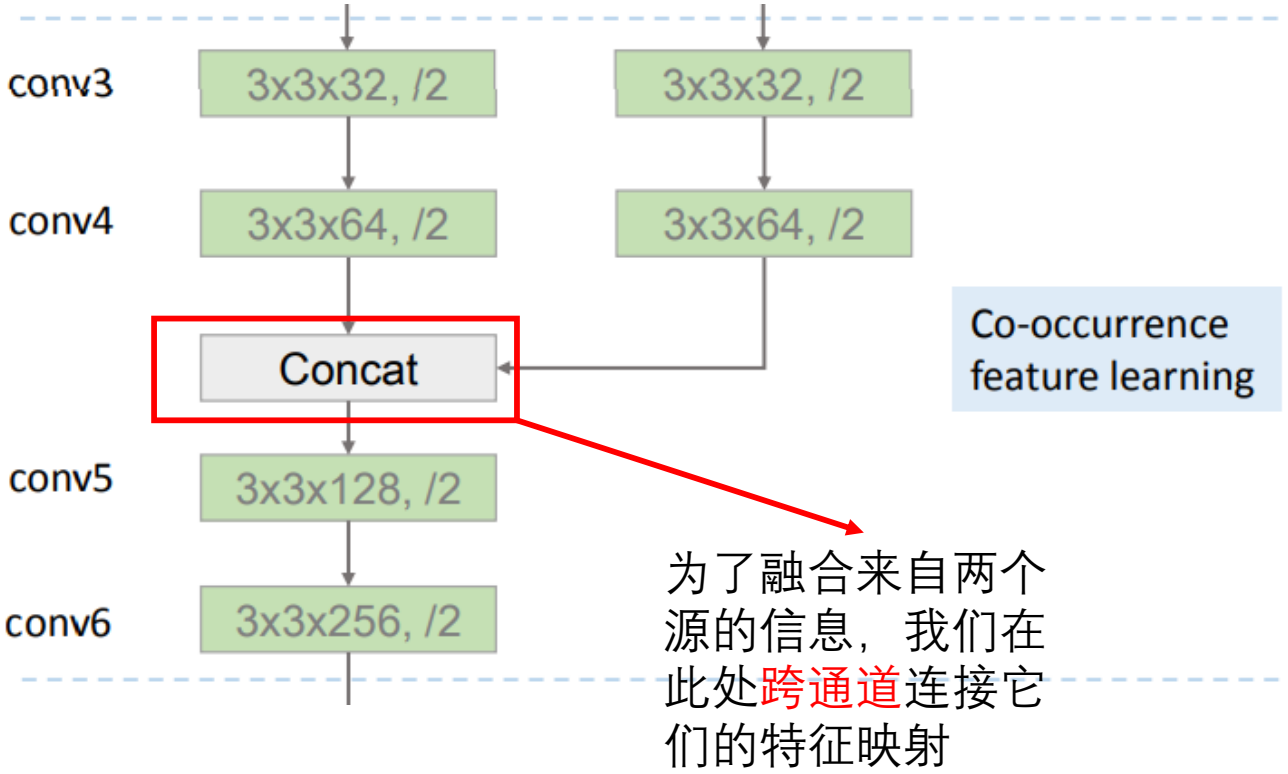
（b）跨通道的全局特征聚合

1、单人动作识别

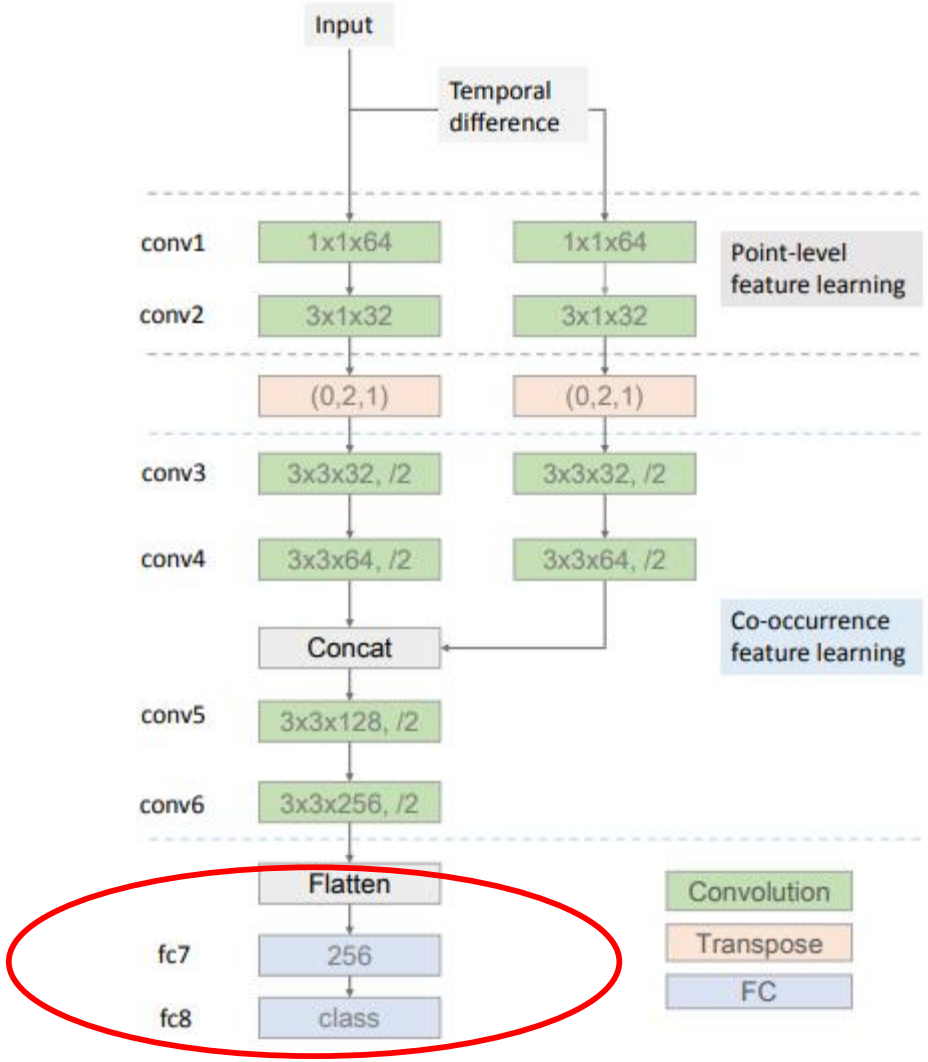

Input
Temporal difference
conv1
1x1x64
1x1x64
Point-level feature learning
conv2
3x1x32
3x1x32
(0,2,1)
(0,2,1)
conv3
3x3x32, /2
3x3x32, /2
conv4
3x3x64, /2
3x3x64, /2
Concat
Co-occurrence feature learning
conv5
3x3x128, /2
conv6
3x3x256, /2
Flatten
fc7
256
fc8
class
Convolution
Transpose
FC

1、单人动作识别

在第二层面中，后续的层分层地聚合来自所有关节地全局特征

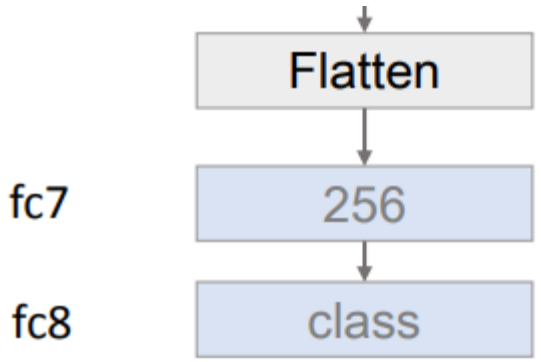


1、单人动作识别


Input
Temporal difference
conv1
1x1x64
1x1x64
Point-level feature learning
conv2
3x1x32
3x1x32
(0,2,1)
(0,2,1)
conv3
3x3x32, /2
3x3x32, /2
conv4
3x3x64, /2
3x3x64, /2
Concat
Co-occurrence feature learning
conv5
3x3x128, /2
conv6
3x3x256, /2
Flatten
fc7
256
fc8
class
Convolution
Transpose
FC

1、单人动作识别

最后，将第二层面地输出展平为矢量，并通过两个完全连接的层进行最终分类。



2、多人交互动作识别

多人交互动作识别：在拥抱和握手等活动中，涉及多个人。为了使我们的框架可以扩展到多人场景，我们对不同的特征融合策略进行了全面的评估。

第一种策略是早期融合：来自多个人的所有关节被堆叠作为网络的输入。对于可变数量的人，如果人数小于预定义的最大数量，则应用零填充。

第二种策略是晚期融合：多个人的输入经过相同的子网，并且他们的conv6特征映射与沿通道的串联/元素最大/平均操作合并。

2、多人交互动作识别

第二种后期融合可以很好地推广到可变数量的人，而第一种早期融合需要预定义的最大数量。此外，与单人相比，没有引入额外的参数。

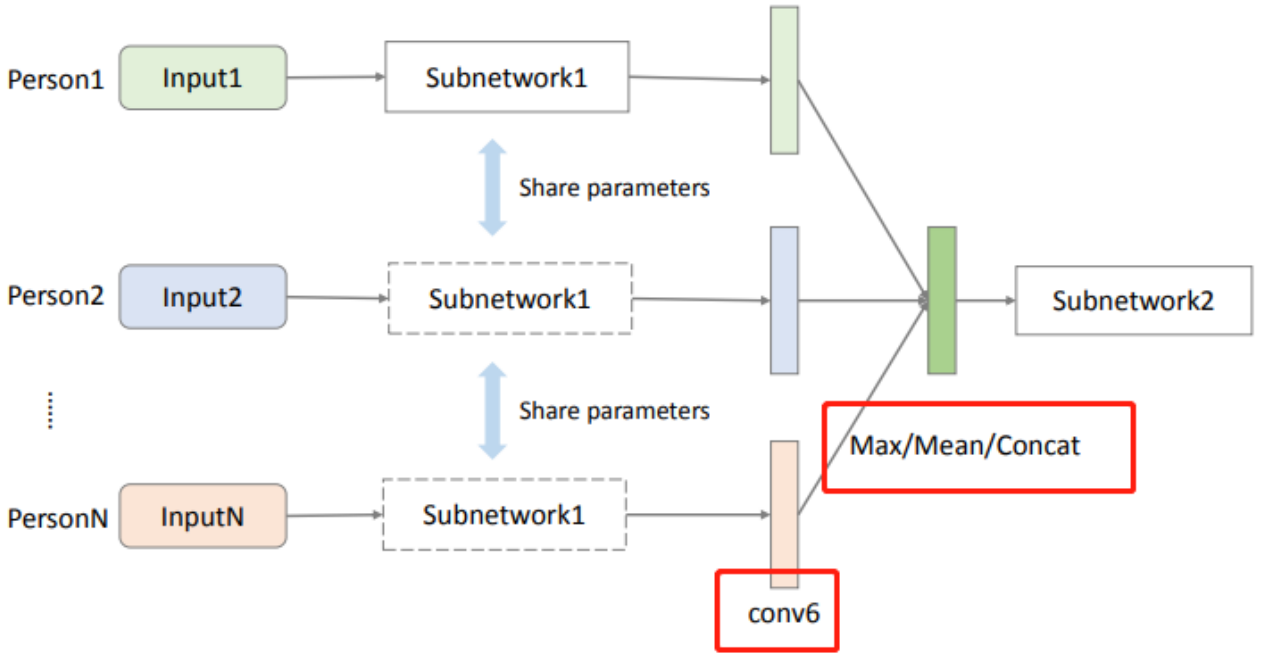


Figure 4: Late fusion diagram for multi-person feature fusion. Maximum, mean and concatenation operations are evaluated in terms of performance and generalization.

3、时域动作检测

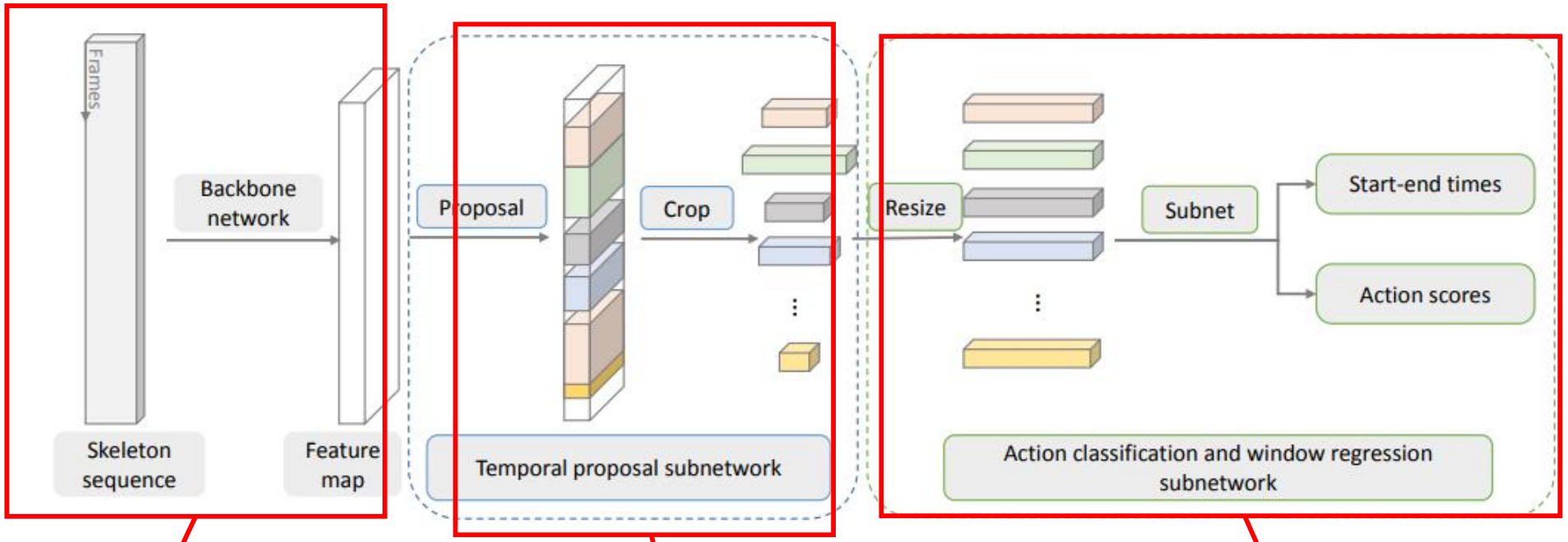


Figure 5: The temporal action detection framework. The backbone network is described in Figure 3. Two subnetworks are designed for temporal proposal segmentation and action classification respectively.

基于主干特征学习网络，在conv5之后附加两个子网

时间提议子网预测可能包含动作的可变长度时间段。使用裁剪和调整大小操作来提取每个提议的相应特征映射。

动作分类子网预测其动作类别。

3、论文实验结果

3、论文实验结果

NTU RGB+D Action Recognition Dataset：这个数据集是新加坡南洋理工大学博云搜索实验室建立的。NTU RGB+D动作识别数据集包含56,880个示例动作，内容有 RGB 视频, 深度图序列， 3D 骨骼数据, 对于每个示例还有红外成像视频。

SBU Kinect Interaction Dataset：这个数据集的数据主要是两个人交互，数据通过微软的Kinect传感器收集(3.3 gb)，含有八种两人互动动作:靠近,离开,推动,踢,拳打, 拥抱 ,握手,交换物品。

PKU-MMD Dataset：这个数据集是一段视频包含多段动作，主要是用来做detection的。

Method		Accuracy (%)
Early fusion		85.2
Late fusion	Mean	85.8
	Concat	85.9
	Max	**86.5**

Table 1: Performance of different fusion methods for multi-person feature on the NTU RGB+D dataset in the cross-subject setting.

Methods	Accuracy (%) CS	CV
Deep LSTM [Shahroudy *et al.*, 2016]	60.7	67.3
Part-aware LSTM [Shahroudy *et al.*, 2016]	62.9	70.3
ST-LSTM+Trust Gate [Liu *et al.*, 2016]	69.2	77.7
STA-LSTM [Song *et al.*, 2017]	73.4	81.2
Clips + CNN + MTLN [Ke *et al.*, 2017]	79.6	84.8
VA-LSTM [Zhang *et al.*, 2017]	79.2	87.7
Two-stream CNN [Li *et al.*, 2017b]	83.2	89.3
Proposed **HCN**	**86.5**	**91.1**

Table 2: Action classification performance on the NTU RGB+D dataset. CS and CV mean the cross-subject and cross-view settings respectively.

Methods	Accuracy (%)
Raw skeleton [Ji *et al.*, 2014]	79.4
Joint feature [Ji *et al.*, 2014]	86.9
ST-LSTM [Liu *et al.*, 2016]	88.6
Co-occurrence RNN [Zhu *et al.*, 2016]	90.4
STA-LSTM [Song *et al.*, 2017]	91.5
ST-LSTM+Trust Gate [Liu *et al.*, 2016]	93.3
VA-LSTM [Zhang *et al.*, 2017]	97.6
Proposed **HCN**	**98.6**

Table 3: Action classification performance on the SBU dataset.

Methods	mAP (%)	
	CS	CV
STA-LSTM [Song *et al.*, 2017]	44.4	13.1
JCRRNN [Li *et al.*, 2016]	32.5	53.3
Skeleton boxes [Li *et al.*, 2017a]	54.8	94.2
Li et al. [Li *et al.*, 2017b]	90.4	93.7
Proposed **HCN**	**92.6**	**94.2**

Table 4: Action detection performance on the PKU-MMD dataset. mAP is measured at an IoU threshold of 0.5.

Importance of Global Co-occurrence Feature Learning

Methods	NTU RGB+D		SBU	PKU-MMD	
	CS	CV	-	CS	CV
HCN-local	83.9	89.7	96.8	91.1	93.9
HCN	86.5	91.1	98.6	92.6	94.2

Table 5: Comparison of HCN-local and HCN in terms of classification accuracy on the NTU RGB+D and SBU datasets and detection mAP on the PKU-MMD dataset.

3、论文实验结果

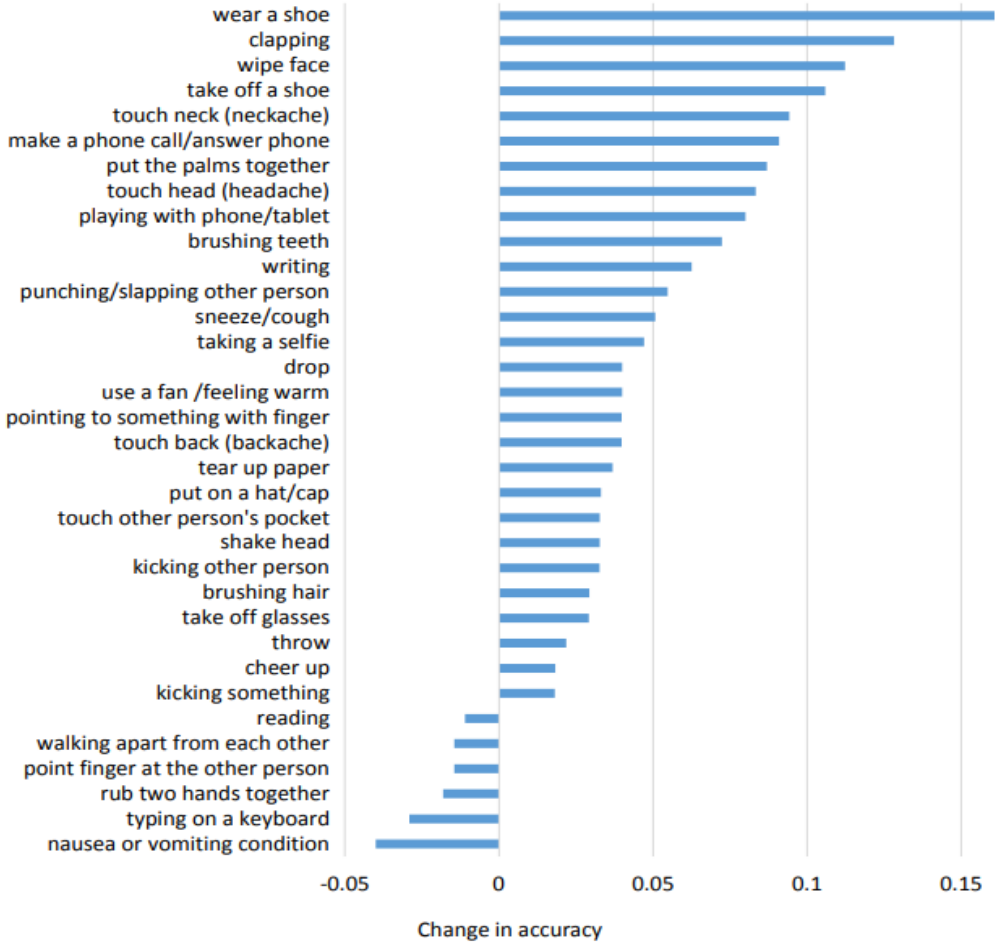


Figure 6: Per-category change in accuracy of HCN over HCN-local on the NTU RGB+D dataset in the cross-subject setting. For clarity only categories with change greater than 1% are shown.

4、答疑

谢谢