# 基于时空注意力模块的视频行为识别

摘要：为了使识别网络聚焦于目标及其局部运动模式下的时空信息，提出一种时空注意力模块下的行为识别方法．时空注意力模块由通道-时间分支与空间-时间分支并联组成．通道-时间分支利用多尺度池化提取各通道空间信息数列，经卷积得到各通道的注意权重，接着对权重进行矩阵操作与softmax得到时间维上的权重值，最后利用元素乘法实现特征重标定．空间-时间分支采用最大池化与平均池化将所有通道的空间信息压缩到2个特征图上，然后采用矩阵操作与softmax得到包含时间信息的特征图并映射到原始特征图．两分支输出相加后作为时空注意力模块的最终输出．将时空注意力模块插入基础网络并在常见的人体行为识别数据集UCF101和HDBM51上进行实验分析，实现了准确率的提升．
关键词：行为识别，通道时间注意力，空间时间注意力

# Video Action Recognition Based on Spatio-temporal Attention Module

Abstract：In order to make the network focus on targets and their local motion patterns, spatial position and time information, an action recognition method based on spatio-temporal attention module is proposed. Temporal and spatial attention module is composed of channel-temporal branch and spatial-temporal branch in parallel. Channel-temporal branch uses multi-scale pooling to extract the spatial information sequence of each channel, and then gets the attention weight of each channel by convolution. Then, it performs matrix operation and softmax to get the weight value in time dimension, and finally realizes feature recalibration by element multiplication. Spatial-temporal branch uses maximum pooling and average pooling to compress the spatial information of all channels into two feature maps, and then uses matrix operation and softmax to get the feature map containing time information and map it to the original feature map. The outputs of the two branches are added as the final output of the spatio-temporal attention module. The spatio-temporal attention module is inserted into the basic network, and experimental analysis is carried out on the common human behavior recognition data sets UCF101 and HDBM51, which improves the accuracy.
Keywords：Action recognition, Channel-temporal Attention, Spatial-temporal Attention

## 1 引言

在计算机视觉领域，对人类行为识别的研究既能发展相关理论基础又能扩大其工程应用范围．对于理论基础，行为识别领域融合了图像处理、计算机视觉、人工智能、人体运动学和生物科学等多个学科的知识，对人类行为识别的研究可以促进这些学科的共同进步．对于工程应用，视频中的人类行为识别系统有着丰富的应用领域和巨大的市场价值．其应用领域包括自动驾驶、人机交互、智能安防监控等．

早期的行为识别方法主要依赖较优异的人工设计特征，如密集轨迹特征[1]、视觉增强单词包法[2]等．得益于神经网络的发展，目前基于深度学习的行为识别方法已经领先于传统的手工设计特征的方法．Karpathy等[3]率先将神经网络运用于行为识别，其将单张RGB图作为网络的输入，这只考虑了视频的空间表观特征而忽略了时域上的运动信息．对此，Simonyan等[4]提出了双流网络．该方法使用基于RGB图片的空间流卷积神经网络(Spatial stream ConvNet)和基于光流图的时间神经网络网络(Temporal stream ConvNet)分别提取人类行为的静态特征和动态特征最后将双流信息融合进行识别．Wang等[5]提出了TSN结构来处

理持续时间较长的视频，其将一个输入视频分成 K 段(segment)，然后每个段中随机采样得到一个片段(snippet)．不同片段的类别得分采用段共识函数进行融合来产生段共识，最后对所有模型的预测融合产生最终的预测结果．为了解决背景信息干扰，Zhou 等[6]结合目标检测使神经网络有侧重地学习人体的动作信息．Liu 等[7]提出融合卷积网络与 LSTM 的行为识别方法来指导网络学习更有效的特征．Jie 等[8]提出了通道注意力模块(SE-block)来显式地建模特征通道间的相互依赖关系，从而提升网络性能．在 SENet 的基础上，WOO 等[9]提出了包含通道与特征空间信息的注意力模块(CBAM)．考虑到视频的时序特性，WANG 等[10]提出了 Non-local 模块同时建模空间与时间信息的重要性．

上述方法中，普通的卷积网络没有侧重地学习重要特征信息，从而导致整体结果不佳；SENet 与 CBAM 引入了注意力模块，但它们只关注空间信息，忽略了视频中的时序信息；Non-local 虽然同时考虑空间与时间信息，但其计算量太过庞大．基于此，本文提出全新的时空注意力模块，通过通道-时间分支提取输入特征的通道依赖关系及时序相关性，空间-时间分支则建模输入特征的空间与时序信息，然后将两分支信息进行相加融合．该模块是一种即插即用模块，能嵌入主流卷积网络中，提升网络特征表示能力，同时只引入较少的额外计算量．

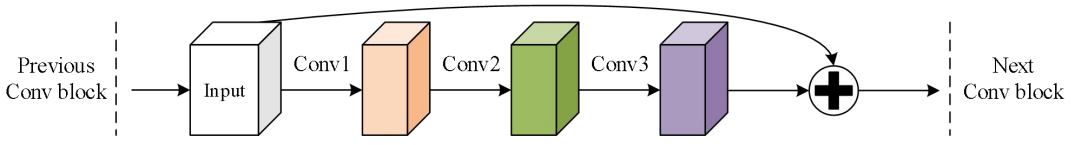
## 2 时空注意力网络的整体架构

本节首先介绍 ResNet50 的构成模块—Bottleneck；接着介绍将时空注意力模块引入 Bottleneck 后的网络构成模块；最后给出本文网络的整体架构．时空注意力模块将在第 3 节详细介绍．

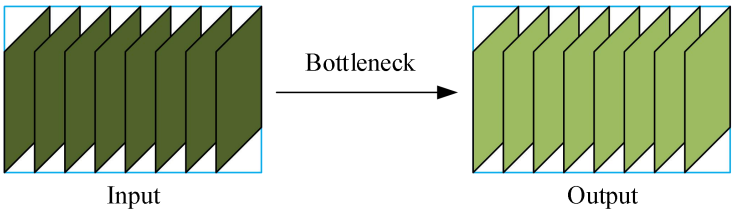
### 2.1 ResNet50 基础模块问题分析

ResNet50 由 4 个网络层(stage1-4)构成，每层构成模块是 Bottleneck，其结构过程如图 1(a)所示．该模块首先对输入特征图使用 1×1 卷积(Conv1)进行卷积操作，主要目的是为了减少参数的数量，从而减少计算量，且在降维之后可以更加有效、直观地进行数据的训练和特征提取．接着使用感受野较大的 3×3 卷积(Conv2)来提取更细化特征．最后使用 1×1 卷积(Conv3)进行升维操作，使输出能与原始输入维度相匹配，从而进行特征相加．

从信号角度来分析，每个特征图是输入信号的一个分信息，每个分信息对整体的重要性不一样，因此，让网络学习重要的分信息是很重要的．图 1(b)为特征经过 Bottleneck 前后的表示对比示意图，从图中可以看出，输入特征只是简单地经过几次卷积操作，并没有突出重要的特征信息．




(a) Bottleneck 示意图



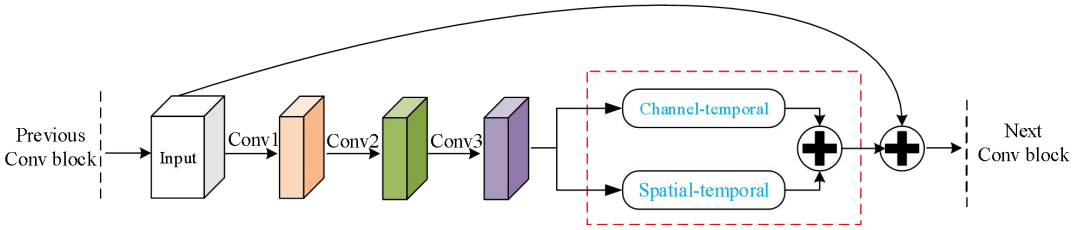

(b) 输入与输出对比

图 1 ResNet50 构成模块和输出示意图

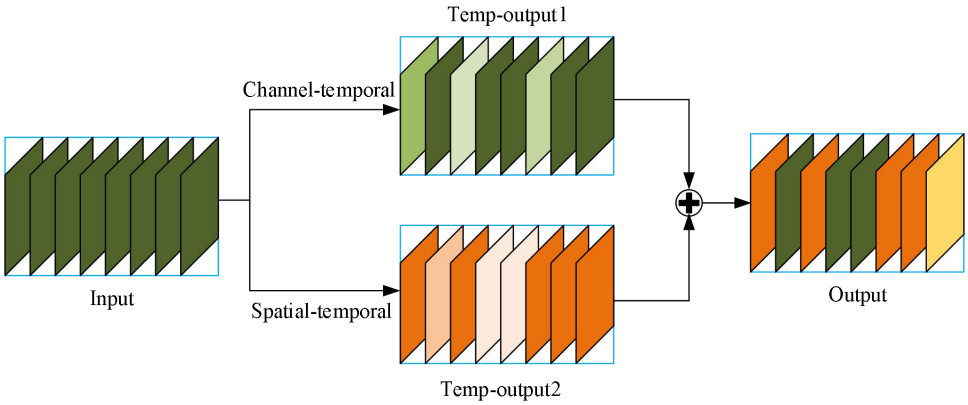
为了解决 Bottleneck 存在的问题，本文提出了时空注意力模块并将其插入 Bottleneck 构建 Spatio-temporal-attention-block(STAB)，该模块流程如图 2(a)所示．输入特征首先经过 3 次卷积操作，然后将输出送入 Channel-temporal 与 Spatial-temporal 分支提取重要性较高的特

征信息，接着将两部分输出相加融合，最后与原始输入进行特征相加.

图 2(b)显示了经过 STAB 后输入与输出的对比结果. 输入特征在经过 attention 分支后得到 Temp-output，相比于原输入，Temp-output 对特征进行了选择，然后将两个 Temp-output 进行相加得到最终的输出. 和 Bottleneck 的输出相比，本模块的输出对输入特征按重要性进行了选择，得到了对识别任务更有用的特征信息.




(a) Spatio-temporal-attention-block




(b) 输入与输出对比

图 2 改进后的 ResNet50 构成模块和输出示意图

## 2.2 行为识别网络整体架构

本文的网络整体结构如图 3 所示. 图中，黑色虚线框表示网络层(stage)，紫红色立方体表示输入，黄色矩形表示卷积核为 7×7 的卷积操作，蓝色立方体表示 ResNet50 原本的 Bottleneck 模块，红色立方体表示 STAB，绿色框表示全连接层(Fc)，橙色椭圆表示交叉熵损失函数，紫色圆圈则是每类结果的预测得分. 输入首先经过 7×7 卷积操作进行特征图尺寸缩减，然后将输出送入 Bottleneck 与 STAB 进行特征提取，最后的高层特征经过全连接层拉平后得到每类结果的预测得分，完成行为识别任务. 网络结构中 STAB 与 Bottleneck 交替使用，原因分析详见 4.3.2 小节.

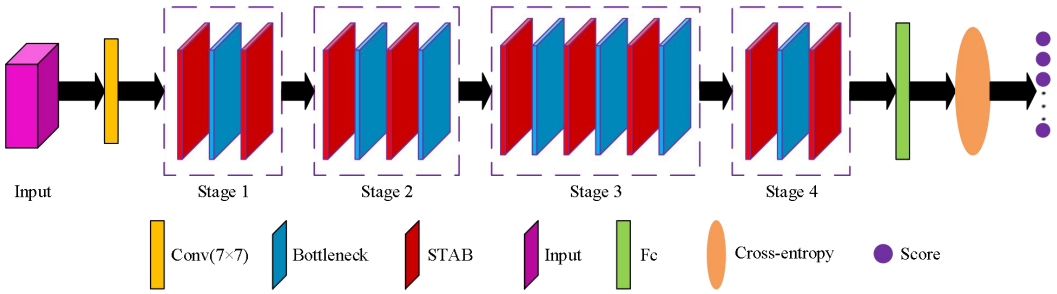



图 3 行为识别网络整体架构

# 3 时空注意力模块

本小节首先介绍 channel-attention、spatial-attention 和 temporal-attention 三个子模块，然后介绍 channel-temporal 和 spatial-temporal 分支.

## 3.1 通道注意力模块

特征图的每个通道都被视为特征检测器，因此建模特征的通道间关系来产生通道关注图能帮助网络关注更有意义的信息．本文的通道注意力模块如图 4 所示．从图中可以看出，该模块由金字塔池化、一维卷积和激活映射组成.

首先，利用自适应均值池化操作来聚合特征图的空间信息，生成 4 个不同尺寸的空间上下文描述符 $F_1$-$F_4$(见图 5)，然后将所有描述符拉伸展平为一维向量，接着采用级联操作将 4 个一维向量拼接成长度为 110 的一维向量.

通过金字塔池化[11]，C 个输入特征图就变成了 C 个 110 维的特征向量，接着使用一维卷积操作为每个特征向量学习一个权重，该权重用来显示地建模特征通道间的相关性．然后利用 Sigmoid 函数将权重归一化到[0,1]，最后通过乘法逐通道加权到先前的特征上，完成在通道维度上的对原始特征的重标定.

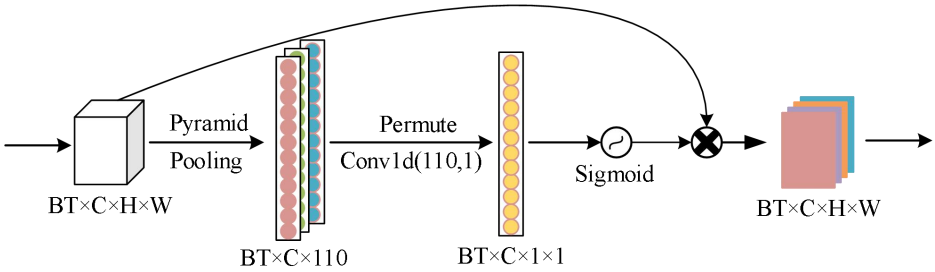



图 4 channel-attention 示意图

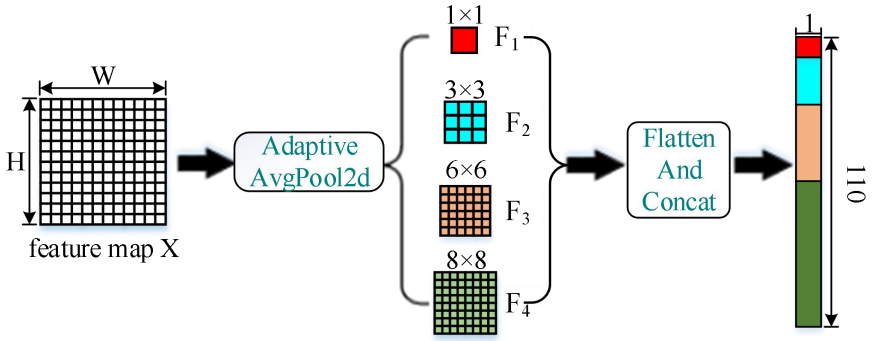



图 5 金字塔池化

## 3.2 空间注意力模块

通道注意力帮助网络聚焦于”what”信息，空间注意力则侧重于”where”信息．为了计算空间注意力，本文使用了图 6 所示的空间注意力模块.

首先，沿着通道方向对输入特征图依次应用平均池化和最大池化操作来聚集特征映射的通道信息．然后将得到的 2 个新的特征图连接起来生成有效的特征描述符．接着对特征描述符进行卷积操作得到空间注意图．最后对空间注意图使用 Sigmoid 函数进行归一化并将其与原始输入相乘，实现空间维度上的对原始特征的重标定.

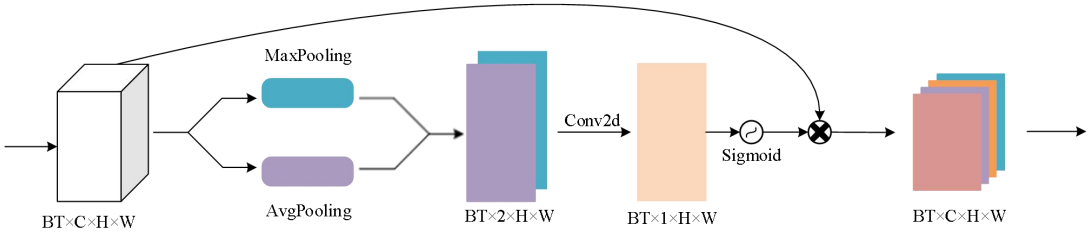



图 6 spatial-attention 示意图

## 3.3 时间注意力模块

出于简化角度，省略输入特征的 batch_size 维度，此时输入特征尺寸为[C, T, W, H]，T 表示输入帧数，C 表示通道数，H、W 则是特征图尺寸大小．时间注意力流程如图 7 所示．

首先，利用维度转换操作将输入转换成[CHW, T]（蓝色矩形）和[T, CHW]（紫色矩形），接着将两者进行矩阵相乘．红色虚线框给出了具体操作，框中 X 的第 $i$ 行表示第 $i$ 帧图像的所有信息，同理，Y 中的第 $j$ 列表示第 $j$ 帧图像的所有信息，两者相乘得到 Z．此时，Z 中第 $(i, j)$ 位置的元素为输入序列中第 $i$ 帧对第 $j$ 帧的影响，从而实现任意两帧之间的依赖关系．接着，对 Z 使用 Softmax 进行行归一化得到 U，归一化后每一行之和为 1．对于 U 中的 $(i, j)$ 位置，可理解为第 $i$ 帧对第 $j$ 帧的权重，所有的 $i$ 对 $j$ 的权重之和为 1，此时的 U 便是时间维度上的注意力图．

再次利用矩阵乘法将注意力图 U 作用到原特征上，接着通过维度转换操作将输出转变成原始尺寸并与原特征相加实现时间维度上的特征重标定．

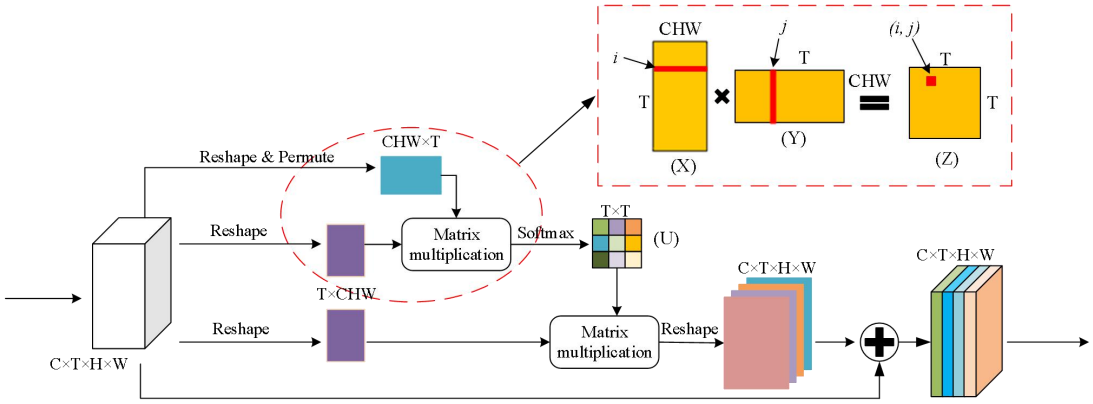
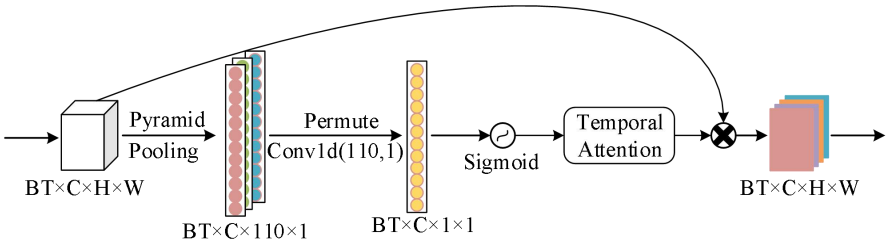



图 7 temporal-attention 示意图

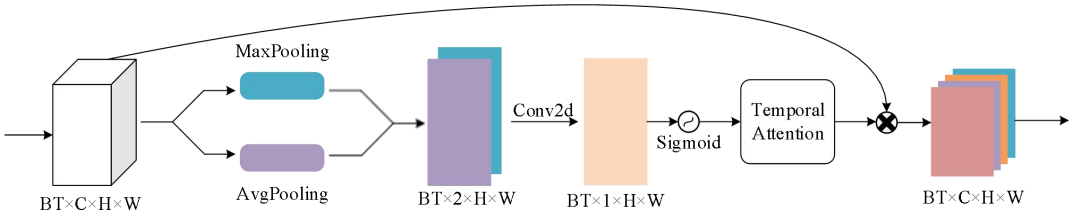
## 3.4 时空注意力模块

3.1、3.2 与 3.3 小节分别介绍了 channel-attention、spatial-attention 和 temporal-attention 模块，若是直接将他们运用到网络中，可能会破坏视频的时空特性，因此本文将 channel-attention 和 temporal-attention 结合在一起，同时也把 spatial-attention 和 temporal-attention 关联起来，构建了两个时空注意力分支：channel-temporal-attention 分支和 spatial-temporal-attention 分支．

图 8(a)表示 channel-temporal-attention 分支，图 8(b)表示 spatial-temporal-attention 分支．




(a) channel-temporal-attention 分支示意图




(b) spatial-temporal-attention 分支示意图

图 8 时空注意力模块

# 4 实验与分析

本小节首先介绍实验数据集，然后阐述 Baseline 设置与实验设置，最后对具体实验结果进行分析.

## 4.1 数据集介绍

本文在最常见的行为识别数据集 UCF101 和 HMDB51 上对本文网络结构进行评估实验，以便将其性能与目前主流的方法进行比较.

UCF101 数据集是从 YouTube 收集的具有 101 个动作类别的逼真动作视频的动作识别数据集. 101 个动作类别中的视频分为 25 组，每组可包含 4-7 个动作视频. 来自同一组的视频可能共享一些共同的功能，例如类似的背景，类似的观点等.

HMDB51 数据集内容主要来自电影，一小部分来自公共数据库，如 YouTube 视频. 该数据集包含 6849 个剪辑，分为 51 个动作类别，每个动作类别至少包含 101 个剪辑.

## 4.2 实验设置

### 4.2.1 Baseline 设置

本文 Baseline 采用 ResNet50 作为主干网络. 针对每个视频输入，首先将其分为 8 个片段，然后在每个片段随机采样 1 帧，共计 8 帧作为输入. 网络对每帧作出预测，然后将 8 个预测值取平均作为最终预测值.

### 4.2.2 训练设置

本文实验中，卷积神经网络基于 PyTorch 平台设计实现. 网络采用 ResNet50 作为主干网络，训练采用小批量随机梯度下降法，动量为 0.9，权值在第 15、35、55 个 epoch 时衰减一次，衰减率为 0.1，总训练数设置为 70. 初始学习率设为 0.001. Dropout 设置为 0.8. 实验采用 2 张 TITAN 1080TI GPU 进行，batchsize 设置为 24.

## 4.3 实验结果与分析

### 4.3.1 可视化结果分析

本实验采用 ResNet50 作为主干网络，然后将 SE_Block、CBAM_Block、Non-local 以及本文的 STAB 分别嵌入到网络中，经过精心训练后，对输入进行可视化输出. 可视化结果详见图 9.

图 9 可视化结果

在图 9 中，每一行各代表一种注意力模块的可视化结果．对于第一列结果和第二列结果，四种注意力模块都成功关注到了重要部分，但 STAB 关注的无关区域相对较少；对于第三列结果，SE 和 CBAM 都没能关注到人体，Non-local 也只关注到部分人体，STAB 则关注到了完整的人体区域；对于第四列结果，SE 和 CBAM 只关注到了人的下半身，忽略了上半身动作区域，Non-local 和 STAB 都关注到了完整的人体区域.

综上所述，STAB 在四种注意力模块中表现最好.

4.3.2 模块嵌入位置分析

图 3 给出了本文网络的整体框架，本小节将通过实验说明这么搭建的原因. 实验结果如表 1 所示. 表 1 中, all 表示将原始的 bottleneck 全部替换成 STAB, part1 表示将 stage1-4 的奇数残差块替换成 STAB, part2 表示将 stage2-3 的奇数残差块替换成 STAB.

从结果来看，随着 STAB 个数的增多，网络提取关键信息的能力提升了．但是 all 却表现很差，造成这一现象的原因是过多的 STAB 导致信息冗余从而造成不必要的信息增强.

综上所述，本文采用 part2 为最终网络结构，即图 3．后续的所有实验均在此结构上进行.

表 1 模块嵌入位置

| 网络结构 | UCF101(%) |
|---|---|
| ResNet50 | 88.9 |
| ResNet50-part1 | 90.2 |
| ResNet50-part2 | 90.6 |
| ResNet50-all | 89.7 |

4.3.3 网络计算量分析

一个优秀的注意力模块不仅能给网络带来结果正确率的提升，同时引入的额外计算量也应该很少．基于此，本实验采用 ResNet50 作为主干网络，定量分析四种注意力模块对网络计算量的影响，采用每秒浮点运算次数(FLOPs)作为评价指标，该指标值越大则意味着网络需要更多的计算资源．结果详见表 2.

表 2 网络计算量与 UCF101 正确率

| 网络 | FLOPs(G) | UCF101(%) |
|---|---|---|
| Baseline | 32.89 | 88.9 |
| Baseline+SE | 32.93 | 89.1 |
| Baseline+CBAM | 32.94 | 89.0 |
| Baseline+Non-local | 49.38 | 90.2 |
| Baseline+STAB | 33.03 | 90.6 |

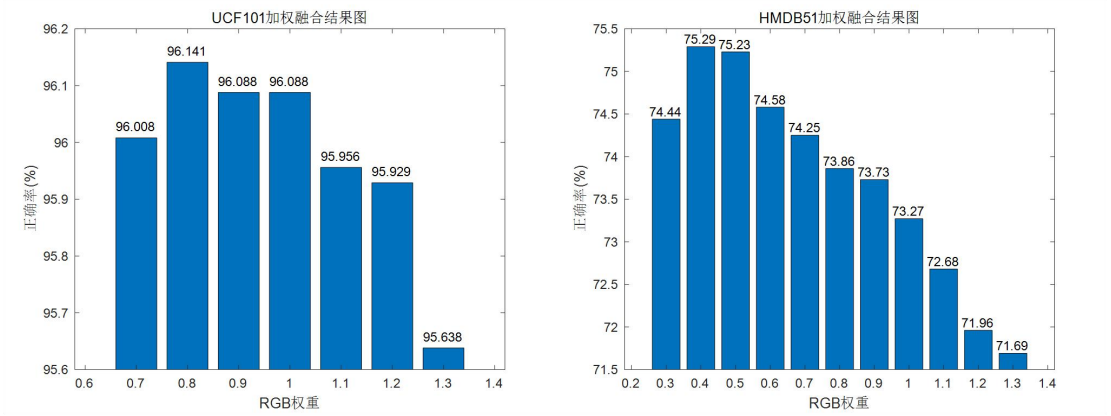
从表 2 结果来看，本文的 STAB 相比于 baseline 只增加了 0.14G 额外计算量，但正确率却提高了 1.7．虽然 SE 和 CBAM 增加的额外计算量相对最少，但正确率几乎没有提升．Non-local 正确率接近 STAB，但其额外计算量比 STAB 多了 49.5%．综上所述，STAB 是一个高效的注意力模块.

4.3.3 双流结果融合

本文采用 RGB 和 Flow 两种输入来训练网络，并对双流结果进行决策层融合作为最终的识别结果．在进行融合时，采用控制变量的思想，固定 Flow 权重为 1，改变 RGB 权重来寻找最高的准确率．图 10 给出了 UCF101 和 HMDB51 两个数据集上 RGB 权重对识别结果的影响.

在 UCF101 上，Flow 表现比 RGB 略好，因此在图 10(a)中，RGB 权重最低值选为 0.7．在 HMDB51 上，Flow 表现远远好于 RGB，因此在图 10(b)中，RGB 权重最低值选为 0.3．由图 10 可以看出，对于 UCF101，当 RGB 权重为 0.8 时，识别结果最高；对于 HMDB51，当

RGB 权重为 0.4 时，识别结果最高.




(a) UCF101 结果　　(b) HMDB51 结果

图 10 RGB 权重对识别结果的影响

4.3.4 与主流方法比较

在本小节中， 将通过具体实验进一步展示所提出的时空注意力网络与最先进的动作识别方法的比较结果. 关于 UCF101 和 HMDB51 的相关结果详见表 3. 表 3 中，S 代表 Sports-1M 数据集，大小约为 128G；K 代表 Kinetics 数据集，大小约为 132G；I 代表 ImageNet 数据集，大小约为 155G；mK 代表 mini−Kinetics 数据集，大小约为 70G.

表 3 与主流方法的比较

| 方法 | 输入 | Pre-train | UCF101(%) | HMDB51(%) |
|---|---|---|---|---|
| C3D+IDT[16] | RGB+Flow | S | 90.4 | - |
| R(2+1)D[17] | RGB+Flow | K+S | 95.0 | 72.7 |
| LTC[18] | RGB+Flow | I | 91.7 | 64.8 |
| TS+LSTM[19] | RGB+Flow | I+S | 88.6 | - |
| TLE[12] | RGB+Flow | I+S | 95.4 | 71.1 |
| TSN[5] | RGB+Flow+RGB_Diff | I | 94.2 | 69.4 |
| MiCT[15] | RGB+Flow | K | 94.7 | 70.5 |
| TSM[20] | RGB | K | 94.5 | 70.7 |
| P3D[14] | RGB+Flow | I | 93.7 | - |
| I3D[13] | RGB+Flow | mK | 95.7 | 74.3 |
| StNet[21] | RGB | K | 93.5 | - |
| Hidden-two-stream[22] | RGB+Flow | I | 93.2 | 66.8 |
| STH [23] | RGB | I+K | 96.0 | 74.8 |
| Ours(STAB) | RGB+Flow | mk | **96.1** | **75.3** |

在当下主流方法中，有的采用了先进的时空融合方法来获得高效的网络特征，如 TLE[12]；有的则利用 CNN 和 LSTM 网络的结合体来获得输入帧之间的序列信息以此来获得比单纯 RGB 表观信息更丰富的时空信息；I3D[13]直接将最先进的 2D CNN 架构膨胀成 3D CNN 网络，以利用训练好的 2D 模型；为了减少参数量，P3D[14]通过将 3D 卷积分解为沿空间维度的 2D 卷积和沿时间维度的 1D 卷积来建模时空信息，从而学习非常深的时空特征；MiCT[15]则提出混合 2D/3D 卷积模块，利用 2D 卷积提取 RGB 图像的表观信息，利用 3D 卷积提取序列间的相关信息.

从表 3 结果来看，在 UCF101 数据集上，本文的 STAB 在使用最小预训练集的条件下，凭借 96.1%的正确率排在所有方法中第一位；在 HMDB51 数据集上，STAB 的正确率同样

排在第一名，结果为 75.3%.

综上所述，本文的 STAB 确实能给现有基础网络带来较大的效果提升.

## 5 结论

本文提出了时空注意力模块下的人体行为识别方法．通过分析现有基础网络的局限性，提出了时空注意力模块．为了验证模块的有效性，分别从可视化结果、额外计算量、精度提升等方面进行实验验证．最后在通用数据集上与其他主流方法进行比较，实验结果再次证明了时空注意力模块的高效性.

## References


[1] IKIZLER-CINBIS N and SCLAROFF S, Object, sence and actions: Combining multiple features for human action recognition[C]. In: *European Conference on Computer Vision*, Heraklion, Crete, Greece, 2010, 6311: 494-507.

[2] 张良，鲁梦梦，姜华．局部分布信息增强的视觉单词描述与动作识别[J]. *电子与信息学报*, 2016, 38(3): 549-556. doi: 10.11999/JEIT150410.

[3] KARPATHY A, TODERICI G, Shetty S, et al. Large-scale video classi-fication with convolutional neural networks[C]. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Colombus, USA, 2014: 1725-1732.

[4] SIMONYAN K, ZISSERMAN A. Two-stream convolutional networks for action recognition in videos[C]. In: *Proceedings of Advances in Neural Information Processing Systems*, Montreal, Canada, 2014: 568-576.

[5] WANG L, XIONG Y, WANG Z, et al. Temporal segment networks: Towards good practices for deep action recogni-tion[C]. In: *Proceedings of European Conference on Computer Vision*, Springer, Cham, 2016: 20-36.

[6] 张冰冰，葛疏雨，王旗龙，李培华．基于多阶信息融合的行为识别方法研究[J/OL]．*自动化学报*. https://doi.org/10.16383/j.aas.c180265.

[7] 刘天亮，谯庆伟，万俊伟，戴修斌，罗杰波．融合空间-时间双网络流和视觉注意的人体行为识别[J]. *电子与信息学报*, 2018, 40(10):2395-2401.

[8] Hu J, Shen L, Sun G. Squeeze-and-excitation networks[C]. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Salt Lake City, Utah, 2018: 7132-7141.

[9] Woo S, Park J, Lee J Y, et al. Cbam: Convolutional block attention module[C]. In: *Proceedings of the European Conference on Computer Vision*, Munich, Germany, 2018: 3-19.

[10] Wang X, Girshick R, Gupta A, et al. Non-local neural networks[C]. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Salt Lake City, Utah, 2018: 7794-7803.

[11] He K, Zhang X, Ren S, et al. Spatial pyramid pooling in deep convolutional networks for visual recognition[J]. *IEEE transactions on pattern analysis and machine intelligence*, 2015, 37(9): 1904-1916.

[12] Diba A, Sharma V, Van Gool L. Deep temporal linear encoding networks[C]. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Hawaii, 2017: 2329-2338.

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

[14] Qiu Z, Yao T, Mei T. Learning spatio-temporal representation with pseudo-3d residual networks[C]. In: *Proceedings of the IEEE International Conference on Computer Vision*, Venice, Italy, 2017: 5533-5541.

[15] Zhou Y, Sun X, Zha Z J, et al. Mict: Mixed 3d/2d convolutional tube for human action recognition[C]. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Salt Lake City, Utah, 2018: 449-458.

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

[17] Tran D, Wang H, Torresani L, et al. A closer look at spatiotemporal convolutions for action recognition[C]. In: *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, Salt Lake City, Utah, 2018: 6450-6459.

[18] G. Varol, I. Laptev, and C. Schmid. Long-term temporal convolutions for action recognition[J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017, 40(6): 1510-1517.

[19] Yue-Hei Ng J, Hausknecht M, Vijayanarasimhan S, et al. Beyond short snippets: Deep networks for video classification[C]. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Boston, USA, 2015: 4694-4702.

[20] Lin J, Gan C, Han S. Tsm: Temporal shift module for efficient video understanding[C]. In: *Proceedings of the IEEE International Conference on Computer Vision*, Seoul, south Korea, 2019: 7083-7093.

[21] He D, Zhou Z, Gan C, et al. Stnet: Local and global spatial-temporal modeling for action recognition[C]. In : *Proceedings of the AAAI Conference on Artificial Intelligence*, Hawaii, 2019, 33: 8401-8408.

[22] Zhu Y, Lan Z, Newsam S, et al. Hidden two-stream convolutional networks for action recognition[C]. In: *Asian Conference on Computer Vision*, Springer, Cham, 2018: 363-378.

[23] Li X, Wang J, Ma L, et al. STH: Spatio-Temporal Hybrid Convolution for Efficient Action Recognition[J]. *arXiv preprint arXiv:2003.08042*, 2020.

创新性说明：

1） 通道注意力模块

2） 时间注意力模块

3） 注意力子模块间的结合方式及输出结果的结合方式