计算机科学
COMPUTER SCIENCE

多维特征激励网络用于视频行为识别

罗会兰，于亚威，王婵娟

引用本文

罗会兰，于亚威，王婵娟. 多维特征激励网络用于视频行为识别[J]. 计算机科学，2023，50(11A): 230300115-8.

LUO Huilan, YU Yawei, WANG Chanjuan. Multi-dimensional Feature Excitation Network for Video Action Recognition [J]. Computer Science, 2023, 50(11A): 230300115-8.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.230300115

多维特征激励网络用于视频行为识别

罗会兰　于亚威　王婵娟
江西理工大学信息工程学院　江西 赣州 341000

摘　要　在动作识别任务中,由于视频数据存在内容多样和背景复杂的特性,因此提取有效的时空特征是研究的主要难点。为了利用深度网络学习时空特征,研究者们通常采用双流网络和3D卷积网络。但是,双流网络中光流信息缺乏捕获长距离时间关系的能力,且光流提取需占用很大的内存和时间;而3D卷积与2D卷积相比,增加了一个数量级的计算成本,容易导致过拟合和收敛缓慢。为解决以上问题,提出了一种基于注意力的多维度特征激励融合网络MFARs(Multi-dimensional Feature Activation Residual networks)用于视频行为识别。MFARs采用2D卷积网络解决时序特征表达学习问题,利用运动补足激励模块建模时序特征,激发时间通道运动信息;同时利用联合特征激励模块,通过时序特征激励通道和空间信息,以学习到更好的时空特征表达。MFARs在行为识别数据集UCF101和HMDB51上的准确度分别达到了96.5%和73.6%。与当前的主流行为识别模型相比,提出的多维特征激励方法能够有效地表达时空特征,更好地平衡复杂度和分类准确率。

关键词:行为识别;深度学习;2D卷积网络;注意力机制;视频特征表达
中图法分类号　TP391

Multi-dimensional Feature Excitation Network for Video Action Recognition

LUO Huilan,YU Yawei and WANG Chanjuan
College of Information Engineering,Jiangxi University of Technology,Ganzhou,Jiangxi 341000,China

Abstract　Due to the diversity of video content and the complexity of video background,how to effectively extract spatio-temporal features is the main challenge of the video action recognition. In order to use deep networks to learn spatio-temporal features,researchers usually use two-stream networks and 3D convolution networks. Two-stream networks use the optical flow as its input to learn temporal features,but optical flow cannot express long-distance temporal relationships and the calculation of optical flow requires a lot of memory and time. On the other hand,3D convolution networks increase the computational cost by an order of magnitude compared with 2D convolution networks,which easily leads to over-fitting and slow convergence. To solve these problems,an attention-based multi-dimensional feature activation residual networks(MFARs) is proposed for video action recognition. A motion supplement excitation module is proposed to model temporal information and stimulate motion information. A united information excitation module is proposed to use temporal features to stimulate channels and spatial information in order to learn a better spatio-temporal features. Combing these two modules,MFARs is constructed for video action recognition. The proposed method obtainsan accuracy of 96.5% and 73.6% respectively on datasets UCF101 and HMDB51. Compared with the current mainstream action recognition models,the proposed multi-dimensional feature excitation method can effectively express spatial and temporal characteristics,and achieve a better balance of computation complexity and classification accuracy.

Keywords　Action recognition,Deep learning,2D convolution network,Attention mechanism,Video feature representation

视频行为识别是指利用计算机视觉自动分析视频中人的行为,对视频中人的行为和动作进行判断、分析和理解。视频行为识别广泛应用于智能生产和安全生产监控中,对异常行为和不规范行为进行智能预警,从而提高生产效率,降低安全事故。

视频行为识别方法主要分为两类:基于传统的手工提取特征方法[1-5]和基于深度学习的方法[6-17]。传统的手工提取特征方法通过人工设计算法来提取特征和编码,然后输入到分类器中进行分类。由于手动提取特征的方法将特征提取和分类器训练分为两个独立的过程,导致提取到的特征不能很好地服务于分类,所以其对于复杂背景下的视频行为识别的效果并不如人意。相比之下,基于深度学习的方法则是一个端到端的学习过程,通过设计网络利用迭代学习,自动从视频中提取适合于分类任务的特征,然后采用反向传播的方式对模型进行训练,效果有了很大的提升。

双流网络是比较经典的用于视频行为识别或动作识别的

基金项目:国家自然科学基金(61862031);江西省主要学科技术带头人领军人才计划资助项目(20213BCJ22004);江西省学位与研究生教育教学改革研究重点项目(JXYJG-2020-120)
This work was supported by the National Natural Science Foundation of China(61862031),Project Supported by the Leading Talents Plan for the Technical Leaders of Major Disciplines in Jiangxi Province(20213BCJ22004) and Jiangxi Province Degree and Postgraduate Education and Teaching Reform Research Key Project(JXYJG-2020-120).
通信作者:罗会兰(luohuilan@sina.com)

深度模型之一，Simonyan 等[6]根据人类视觉皮层的构造提出双流网络，分别使用堆叠的 RGB 图像帧和光流图作为输入，构建空间流和时间流神经网络，再将这两个网络分别训练，最后将两个网络的平均分类得分作为最后的输出结果。在此基础上，时间分段网络(Temporal Segment Networks，TSN)[7]，采用稀疏采样的方法，首先将视频均匀划分成片段，然后从每个视频片段中随机采样一帧，替代长视频进行时空特征学习，在保持准确率的同时，减少了计算负担。TRN(Temporal Relation Network，TRN)[8]在 TSN 的基础上，提出时序关系推理，通过学习不同尺度之间视频帧的时序关系来更好地学习运动信息。Diba 等[9]针对双流网络融合方式的单一化，提出了时序线性编码层，对视频分段后经过双流网络得到的特征图进行融合编码，以更好地学习时空特征。虽然双流网络解决了时序特征表达的部分问题，但由于空间流和光流分支缺少充分交互，很难学习到时空信息的关联性。

为了更好地学习时空特征，Ji 等[10]提出三维卷积(3-Dimensional Convolution Neural Network，3DCNN)，将时间维度加入卷积计算中，运用 3D 卷积同时从空间和时间维度来提取特征，更好地学习空间与时序维度间的关系。在此基础上，Tran 等[11]使用 3D 卷积和 3D 池化构造 3D 卷积网络 C3D (Convolutional 3D Network)，并且通过实验确定最合适的 3D 卷积核大小为 3×3×3。为了能够利用已训练好的图像分类模型，Carreira 等[12]提出了 I3D(Inflated 3D ConvNet，I3D)网络，将用于图像分类的 2D 网络的卷积核通过“膨胀”得到 3D 网络卷积核，准确度全面超越了 C3D 以及双流 2D 卷积网络，自此 3D 卷积网络逐渐成为主流。但是，由于 3D 卷积添加了时间维度，它的参数量和计算量比 2D 卷积大得多，因此需要更多的训练数据和计算资源。为了提高 3D 卷积模型的效率，Xie 等[13]提出了 S3D(Separable 3D Convolution Neural Network)模型，将 3D 卷积核分解为时间维度的 1D 卷积和空间维度的 2D 卷积，提高了视频识别的效率。同时，Tran 等[14]提出了 R(2+1)D((2+1)Dimensional ResNets)网络，将 3D-ResNet 网络中的 3D 卷积核分解为时间维度的 1D 卷积和空间维度的 2D 卷积，并且将空间信息和时间信息的优化过程也分解开来，使得 2+1 维的卷积更容易优化。Huang 等[13]在双流架构的时间流子网络和空间流子网络采用了改进的 R(2+1)D 卷积，以实现性能和效率的提升。

双流卷积网络将视频分解为空间和时间成分，分别独立进行训练，只在最后一层进行融合，缺少充分的时空交互；而且在时间流部分，光流信息的提取需要额外的时间和内存消耗。另一方面，3D 卷积网络虽然不用提取光流信息，但是由于视频本身存在大量冗余信息，直接将 2D 卷积扩展到时间维度，不仅会使得参数量和计算量成倍增加，导致网络推理速度变慢，而且无法利用视频冗余性的特点。为了解决上述问题，本文在 2D 卷积基础上，提出了运动补足激励模块(Motion Supplement Excitation，MSE)来替代光流信息，建模短距离时序依赖关系，学习视频中的运动信息；同时，提出联合特征激励模块(United Information Excitation，UIE)，联合通道、时间以及空间特征对全局上下文信息进行建模，将学习到的运动信息注入到时间维度和空间维度，以实现不同维度的信息交互，有效学习时空特征。

本文的主要贡献如下：

1)提出运动补足激励模块 MSE，建模局部上下文运动信息，实现时序特征的有效表达。

2)提出联合特征激励模块 UIE，联合时间、通道、及空间特征建模全局上下文信息，有效学习时空特征。

利用 MSE 和 UIE 两个模块构建了多维特征激励残差网络(MFARs)，实验证明它能更加有效地提取视频时空特征，在 UCF101 和 HMDB51 上获得了更高的行为识别准确率。

1 相关工作

1.1 时空特征的 2D 卷积提取方法

通过计算光流图获得时序运动信息，计算成本较高，且时空信息较难融合；而 3D 卷积虽然适用于视频时空特征的提取，但参数量大，增加了学习难度。因此，许多研究者基于 2D 卷积网络，设计出不使用光流信息，能有效学习时空特征的网络模型。Lin 等[18]提出 TSM(Temporal Shift Module，TSM)模块，该模块通过在时序维度上对部分通道进行平移来建模时序特征，使得每一帧都跟附近的帧交换信息，达到在时序上增大感受野的效果，然后将模块插入到现有的分类网络中，进行复杂的时空特征提取。Sudhakaran 等[19]在 TSM 的基础上，构建完整的时空特征提取模块 GSM(Gate-Shift Module，GSM)，该模块分为并联的两个部分，分别是空间 2D 卷积和时序 Shift 操作，然后将两部分的输出进行融合，达到时空建模的目的。STM[20] (SpatioTemporal and Motion Encoding)采用(2+1)D 的方式来建模时空特征，以串联的方式将 3D 卷积分解成 1D 时序卷积和 2D 空间卷积；然后添加额外的运动信息模块(Channel-wise Motion Module，CMM)来增强运动表示。进一步，Liu 等[21]提出新的学习时序特征的范式，首先使用运动增强模块增强运动信息表示，然后加上时序交互模块，以通道的方式补充时序上下文信息，最后以串联的方式联合两个模块建模时空信息。TEA[22] (Temporal Excitation and Aggregation Network)网络采用逐渐扩张感受野的方式来建模长距离的时空关系，在通道维度上进行拆分，分别输入采用不同分辨率的(2+1)D 子网络，对时空信息进行建模，最后再逐步整合。TDN[23] (Temporal Difference Networks)将输入视频分段，在各视频段内用短距离时序差分模块学习局部时序信息，在各视频段间利用长距离时序差分模块学习全局时序信息，实现多尺度的时空信息表达。

1.2 注意力机制

注意力机制可以使模型聚焦关键区域，自动实现重要特征的增强，从而提高识别性能。SENet[24]采用压缩激励模块(Squeeze and Excitation，SE)，显式地建模通道之间的相互依赖关系，增强特定通道特征。Woo 等[25]在通道激励模块的基础上，增加空间注意力，构建注意力模块(Convolutional Block Attention Module，CBAM)，对输入特征图进行自适应特征细化。Fu 等[26]提出双注意力网络，基于自注意力机制来分别捕获空间维度和通道维度中的特征依赖关系。Qiu 等[27]提出立方体信息嵌入注意力(Cubic Information-Embedding Attention，CIEA)，建模空间和通道的相互关系。

上述 SENet，CBAM 和 CIEA 等方法都是图像识别研究领域设计的注意力模块。在视频识别中，Wang 等[28]提出混

合注意力机制，设计时空注意力、通道注意力和运动注意力，将它们联合起来用于动作识别。TimeSformer[29]基于自注意力架构 Transformer[30]，将视频分块，学习各视频块间的相互关系，获得时空特征表达，进而提高分类精确度。类似地，Aranb 等[31]采用自注意力机制解决视频问题，提出了 4 种不同的时空自注意力模块，实现时空特征的有效学习。但是基于 Transformer 的方法需要计算每个视频块之间的关系，会消耗大量的计算成本，对设备的要求很高。

与上述视频特征注意力机制方法不同，本文基于通道注意力和空间注意力机制设计联合特征激励模块，同时建模时间、通道以及空间的语义关系，在捕获 3 个特征维度关系的同时，对 3 个维度进行注意力激励，达到时空特征增强的效果。

2 提出的方法

2.1 网络总体结构

本文提出的用于视频行为识别的多维特征激励残差网络（MFARs）如图 1 所示，网络的主干是由多维特征激励残差块（Multi-dimensional Feature Activation Residual block，MFAR）堆叠而成。MFAR 由提出的运动补足激励模块（MSE）和联合特征激励模块（UIE）并联起来构成，它能够充分学习长短时运动信息。网络首先采用稀疏时间采样策略对输入视频进行采样，得到 T 帧图片；然后使用一个 7×7 的卷积层来对图片进行初步的特征提取；接着输入主干网络提取时空特征；最后输入分类网络，得到最终的分类结果。

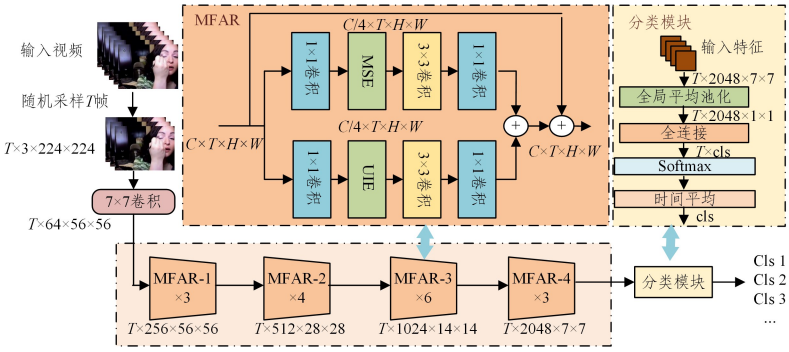

图 1 视频行为识别网络 MFARs 的整体结构图

Fig. 1 Overall structure of MFARs

2.2 多维特征激励残差块 MFAR

如图 1 所示，MFARs 的主干网络是由多组不同分辨率的 MFAR 模块组成。MFAR 模块由 MSE 和 UIE 两个模块并联而成，其中 MSE 学习短时运动特征，UIE 学习长短时运动信息激励的跨通道和空间维度的综合时空特征。输入特征首先经过两个平行的 1×1 卷积减少通道数，然后分别输入 MSE 和 UIE 模块，达到建模运动信息和联合特征信息的目的；各自再通过 1×1 卷积恢复通道数后，对两个模块输出的特征图进行相加得到最终的时空特征。

2.3 运动补足激励模块 MSE

视频中的时序特征表达对于视频动作识别来说非常重要，为了在 2D 卷积网络中更好地提取时序信息，本文提出了运动补足激励模块 MSE。如图 2 所示，与 STM[20]中的 CMM 模块和 TEA[20]中的 ME 模块相比，本文提出的 MSE 模块旨在获得多级运动信息激励和增强的特征。MSE 模块在运动信息获取中能够捕获多级的运动特征从而使得运动信息的表征更加完善。

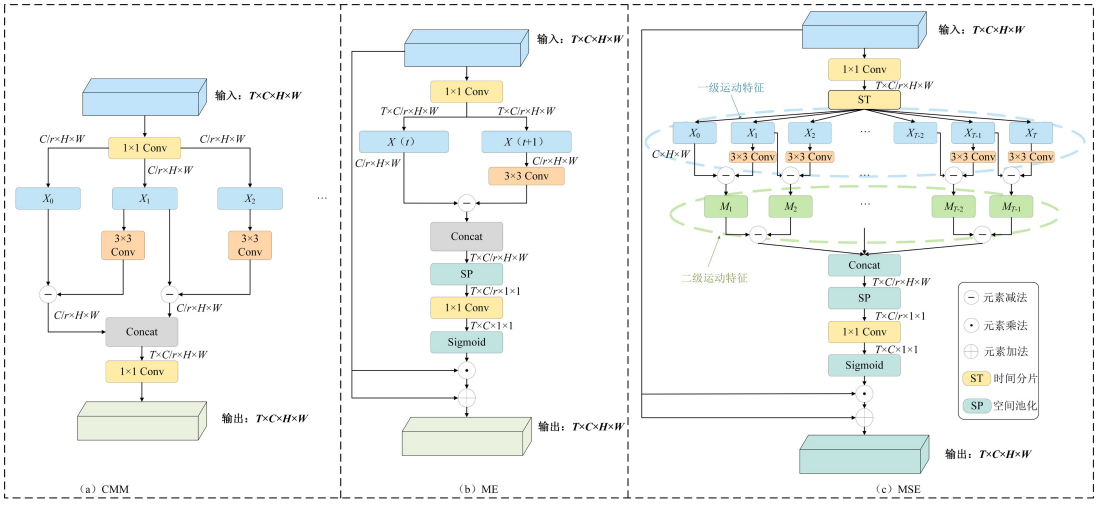

图 2 运动激励模块比较

Fig. 2 Motion excitation module comparison

MSE 的结构如图 2(c)所示，输入特征图首先经过一个 2D 的 1×1 卷积来降低通道的数量，然后将输入特征图沿时间维度进行相邻两帧特征图相减。考虑到动作主体会随着时间发生偏移，在相减前，后一帧使用一个 2D 的 3×3 卷积来学习动作偏移偏差，此过程如式(1)所示：

$$\boldsymbol{M}_t = conv_{3\times3}(\boldsymbol{X}_t) - \boldsymbol{X}_{t-1},\ 1 \leqslant t < T \tag{1}$$

其中，$\boldsymbol{M}_t$ 表示第 t 帧特征图，$conv_{3\times3}$ 表示 3×3 卷积。为了保证相减后时间维度上的帧数不变，对输入特征图在时间维度上进行了填充，即添加一个全 0 的特征图作为第 0 帧 $\boldsymbol{X}_0$。

在获得移动量级别的运动信息 $\boldsymbol{M}_t$ 后，为了去除背景运动噪声的影响，学习到更加可靠的运动特征，对获得的运动特征再次进行相邻帧相减，以学习到更加可靠和更加充分的运动信息。此过程如式(2)所示，将得到的特征图 $\boldsymbol{M}_{t+1}$ 与 $\boldsymbol{M}_t$ 相减得到 $\boldsymbol{V}_t$，来进一步学习更高级别的运动信息。

$$\boldsymbol{V}_t=\boldsymbol{M}_{t+1}-\boldsymbol{M}_t,1\leqslant t<T-1 \tag{2}$$

同样地，为了保持时间维度不变，使用了一个值全为 0 的特征图作为第 $T+1$ 帧进行填充。将得到的 T 帧运动信息特征图 $\boldsymbol{V}_t$ 在时间维度上进行拼接得到 V^c。

接下来，通过一个空间全局平均池化，对每帧上的运动信息进行全局量化，以消除空间细节对运动信息学习的影响。接着，使用一个 2D 的 1×1 卷积来恢复通道数，然后通过 sigmoid 函数来得到运动注意力激励信息。另外，为了增强运动激励信息，将 sigmoid 函数的输出区间从[0,1]扩展到[−1,1]，这个过程如式(3)所示：

$$A=2\delta(conv_{1\times1}(\boldsymbol{V}^c))-1 \tag{3}$$

其中，δ 代表 sigmoid 函数，$conv_{1\times1}$ 代表 1×1 卷积。

最后，将运动注意力激励信息与原输入特征图相乘，完成运动特征的学习和增强。同时增加了残差连接，以减少梯度消失问题。此过程如式(4)所示：

$$\boldsymbol{X}_{out}=\boldsymbol{X}+\boldsymbol{X}\odot A \tag{4}$$

其中，$\odot$ 表示通道级的元素乘法。

2.4 联合特征激励模块 UIE

运动补足激励模块 MSE 旨在将短距离运动信息补充到空间信息中，以学习到时空特征。而联合特征激励模块 UIE 一方面旨在将通道信息补充到运动信息流中，即通道注意力激励时序维；另一方面旨在将空间信息补充到运动信息流中，使用空间注意力激励时序维。通过建模时序与通道和空间的相互依赖关系，学习到长距离时空特征。UIE 的结构如图 3 所示，它包含两个部分：长短时运动信息激励的通道注意力(long-short term Motion Activated Channel Attention，MACA)，和长短时运动信息激励的空间注意力(long-short term Motion Activated Spatial Attention，MASA)。

如图 3 左侧所示，MASA 结构有两个分支，其中左边的分支只使用了一个卷积核大小为 3 的 1D 时序卷积来学习短时运动信息；而右边的分支用于获取长时运动信息。长时运动信息分支首先将输入特征经过卷积核大小为 3 的 1D 时序卷积学习短时运动信息，然后使用一个卷积核大小为 T 的 1D 时序卷积实现时序维度的池化，学习全局时序信息。包含全局时序信息的特征在进行通道维的 softmax 后，与短时运动信息进行矩阵相乘，获得通道信息池化后的特征图。传统的方法将通道进行简单池化，会造成通道信息的较大损失，MASA 采用矩阵相乘的方法，将通道信息融合到空间和时间维度中，能在不损失通道特征的前提下学习到更好的空间注意力。将通道信息池化后的特征图通过一个 3×3 卷积进一步学习空间信息后，经过 sigmoid 函数激活得到空间注意力图，操作过程如式(5)所示：

$$\boldsymbol{MASA}(\boldsymbol{X})=\delta(conv_{3\times3}(conv_{1D}(\boldsymbol{X})\otimes\mathrm{softmax}(RS(TP(RS(conv_{1D}(\boldsymbol{X})))))) \tag{5}$$

其中，$conv_{1D}$ 代表卷积核为 3 的时序 1D 卷积，RS 代表 Reshape 操作，TP 代表 Temporal Pooling 操作，$\otimes$ 代表矩阵乘法。

如图 3 右侧所示，MACA 同样有两个分支：左分支使用卷积核大小为 3 的 1D 时序卷积和 Reshape 操作建模短距离时空特征；右分支在卷积核大小为 3 的 1D 时序卷积之后，再使用卷积核大小为 T 的 1D 时序卷积对时间维度进行池化，reshape 之后获得全局时空特征；然后将全局时空特征通过空间维的 softmax 之后与短距离时空特征进行矩阵相乘，使得空间信息融入到通道和时间维度中，在不损失空间特征的情况下学习更好的通道注意力。操作过程如式(6)所示：

$$\boldsymbol{MACA}(\boldsymbol{X})=\delta(RS(RS(conv_{1D}(\boldsymbol{X}))\otimes\mathrm{softmax}(TP(RS(conv_{1D}(\boldsymbol{X})))))) \tag{6}$$

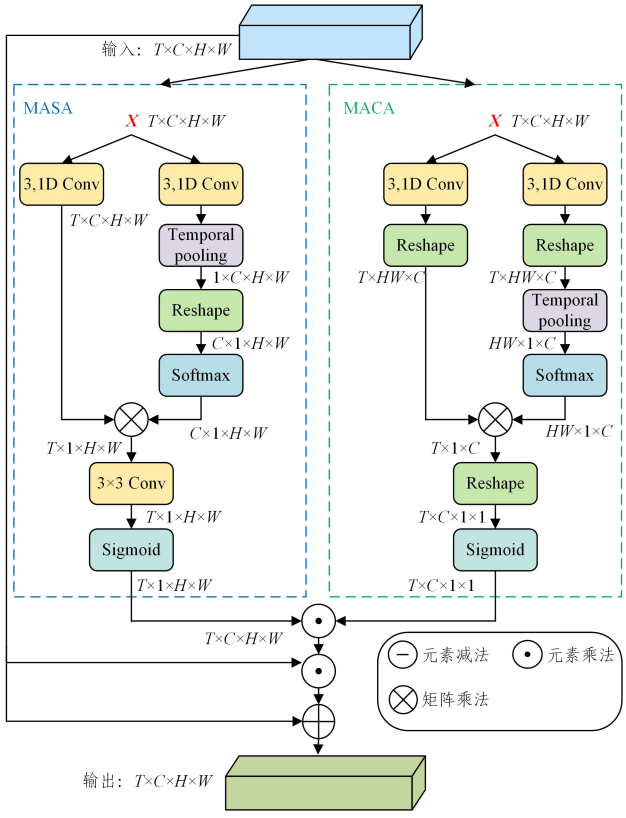

图 3 联合特征激励模块 UIE

Fig. 3 United information excitation module

在获得长短时运动信息激励的空间注意力和长短时运动信息激励的通道注意力后，UIE 模块将它们相乘，以获得基于注意力的多维特征激励信息。UIE 模块的完整操作过程如式(7)所示：

$$\boldsymbol{Y}=\boldsymbol{X}\odot(\boldsymbol{MASA}(\boldsymbol{X})\odot\boldsymbol{MACA}(\boldsymbol{X}))+\boldsymbol{X} \tag{7}$$

其中，$\boldsymbol{X}$ 表示输入的特征图，$\odot$ 代表元素级的乘法，$\boldsymbol{MASA}()$ 和 $\boldsymbol{MACA}()$ 分别表示长短时运动信息激励的空间注意力和长短时运动信息激励的通道注意力功能块。

3 实验结果与分析

3.1 数据集

本小节在两个经典的动作识别数据集 UCF101[32] 和 HMDB51[33] 上评估了 MFARs 的性能。其中 UCF101 数据

集有 13 320 个视频片段，包含五大类：人与物体互动、人体动作、人与人互动、乐器演奏和体育运动，共 101 类人体行为。HMDB51 数据集共有 6 849 个视频片段，包含五大类：面部动作、面部操作、身体动作、交互动作和人体动作，共 51 类人体行为。HMDB51 数据集中的视频片段主要来自于 YouTube 视频和电影，场景复杂，光照条件变化较大，所以比 UCF101 更具有挑战性。

3.2 实验设置

本文采用分类准确率来评估模型的识别能力，采用 GFLOPs(Giga Floating-point Operations)，即浮点运算次数来衡量模型效率。实验用的软件平台为 Pytorch，硬件采用了型号为 GeForce RTX 2080Ti 的显卡。和其他文献一样，将 UCF101 和 HMDB51 的训练数据划分为 spilt1，spilt2 和 spilt3，每个 spilt 中的视频数据按 7∶3 的比例划分为训练集和测试集。实验中，除了与其他先进方法对比时，MFARs 采用了在 Kinetics 上预训练过的 ResNet50 作为基础骨干网络，其他实验均采用在 ImageNet 上预训练过的 ResNet50 作为基础骨干网络，在其上增加的运动补足激励模块(MSE)和联合特征激励模块(UIE)都没有经过预训练。

输入视频采用 TSN 提出的稀疏时间采样策略[7]进行采样，设置 $T=8$。在训练中，将视频帧裁剪为多个 224×224 的图像，以扩充数据量。优化方法采用动量为 0.9 的随机梯度下降方法，训练的权重衰减系数设置为 0.000 5。初始学习率为 0.001，每经过 20 个 epoch 将学习率缩减为之前的 1/10，一共训练 60 个 epoch。

3.3 行为识别可视化分析(MFAR 模块可视化分析)

为了验证提出的 MFAR 模块能够有效提取时空特征，我们采用 grad-CAM[34]对网络进行了可视化。如图 4 所示，第一行为输入视频，其中第一列和第二列表示 HMDB51 数据集中的“run”类别中的两帧图像，第三列和第四列表示 HMDB51 数据集中“jump”类别中的两帧图像，而第五列和第六列代表了实拍视频中两帧图像，表示的是“跳”这个类别。同样的，第七和第八列则表示实拍视频“跑”这个类别的两帧图像。第二行显示的是使用 ResNet50 作为骨干时，conv4_x 的最后一层提取到的特征的热力图。第三行显示的是 MFAR-4 的最后一层提取到的特征的热力图。红色部分代表图片被关注的部分，蓝色部分则代表不被网络关注的部分。从图 4 可以看出，不论是在数据集还是实拍视频中，ResNet50 无法准确捕捉到具体的动作信息，关注区域比较广泛；而 MFARs 可以准确关注到运动的区域。说明 MFAR 模块可以帮助网络更加关注运动区域，减少背景信息的影响，学习到更具表达性的时空特征。

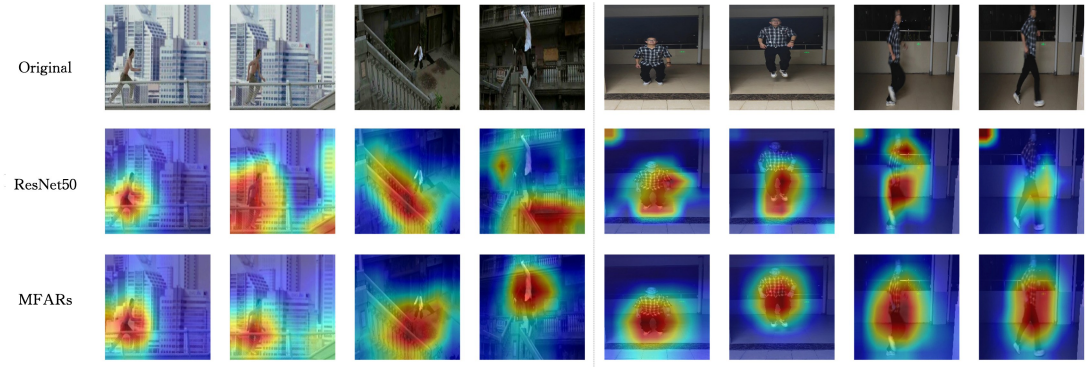

图 4 MFARs 和 ResNet50 的特征可视化对比

Fig. 4 Visualization of features extracted by MFARs and ResNet50

图 4 比较了 ResNet50 与 MFARs 网络在个体行为上可视化的区别，图 5 则进一步比较了人与人之间相互交互的行为在 ResNet50 与 MFARs 网络上不同层级上的特征热力图。如图 5 所示，第一列显示的是 HMDB51 数据集中的“hug”类中的一帧关键帧。第二列到第五列分别显示了 ResNet50 网络和 MFARs 网络在 4 个层级提取到的特征热力图。从图 5 可以看出，ResNet50 网络和 MFARs 网络在浅层中对动作捕捉都不敏感，而 MFARs 在最后两层中与 ResNet50 网络相比能更准确地关注到关键运动区域。证明 MFAR 模块在网络的深层更能发挥作用。

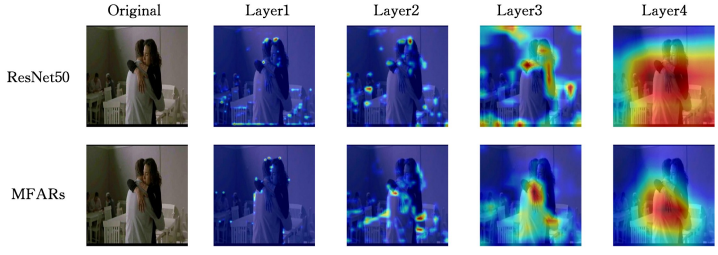

图 5 HMDB51 数据集中“hug”类的特征热力图

Fig. 5 Feature visualization of the “ hug ” class in HMDB51 datas

3.4 与现有先进方法的对比实验

表 1 列出了本文方法与其他先进的动作识别算法在 UCF101 和 HMDB51 数据集上的识别准确率对比结果。本文方法在 UCF101 和 HMDB51 数据集上分别获得了 96.5%和 73.6%的识别准确率。实验结果总结如下：

1)与双流卷积网络 TSN[7]、压缩激励残差网络[35]、STCA-Net[36]和 Improved Two-stream[37]方法相比，本文方法在 UCF101 和 HMDB51 上的准确度均有提升。相比于 STCA-Net，本文方法在 UCF101 和 HMDB51 上分别提升了 2.4%和 4.1%。由于 HMDB51 数据集对动作信息更加

敏感，因此本文方法在 HMDB51 数据集上方法性能提升更大。与 FSTFN[38] 相比，本文方法分别在 UCF101 和 HMDB51 数据集上提升了 3.8%和 7.7%。与改进的双流网络方法 Improved Two-stream 相比，本文方法的参数量只有该方法的 1/3，但是在 UCF101 数据集上的准确度提升了 5.25%。这充分验证了本文提出的运行信息学习模块能有效捕获运动特征，从而超越使用了光流的双流网络。

2）与 3D 卷积模型 I3D[12]，ECO[39]，3D-TDC[40] 和 MRTP[41] 相比，本文方法相较于 3D-TDC 在 UCF101 上提升了 2.68%，在 HMDB51 上提升了 6.77%；相较于 ECO，在 UCF101 上提升了 1.7%，在 HMDB51 上提升了 1.2%；相较于 MRTP，本文方法在 UCF101 上提升了 0.9%。与 I3D 方法相比，在 UCF101 上提升了 0.9%，但是在 HMDB51 上准确率不及 I3D 网络。在模型运算量方面，I3D 的浮点运算量为 108GFLOPs，ECO 的浮点运算量为 267GFLOPs，3D-TDC 的浮点运算量为 84GFLOPs，而本文方法只有 34GFLOPs。这证明本文方法在性能与效率上取得了较好的平衡。

3）与同是采用 2D 卷积方法来进行视频动作识别的方法对比，相较于 TSM[18] 和 sTnet[42]，本文方法在 UCF101 数据集上的识别准确率分别高出 2.0%和 3.0%，在 HMDB51 上高出 TSM 方法 2.9%。与 STM[20] 方法对比，本文方法在 UCF101 数据集上识别准确率高出 0.3%，在 HMDB51 上高出 1.4%，而参数量减少了 11.08×10^6。与 TEA[22] 和 TEInet[21] 方法相比，虽然本文方法在 UCF101 数据集上的识别准确率略低（分别低 0.4%和 0.2%），但是在 HMDB51 数据集上的准确率比 TEA 高 0.3%，比 TEInet 高 1.5%。HMDB51 数据集相比于 UCF101 数据集，动作更复杂，动作识别更加依赖于运动信息的提取，本文方法在 HMDB51 数据集上的优秀表现证明了本文提出的方法能够学习到更加丰富的运动特征。与 Sep-3D RAN[43] 和 S3D RANs[44] 相比，本文方法在 UCF101 数据集上的准确率分别高出 4.8%和 3.2%，在 HMDB51 数据集上比 S3D RANs 方法高 2.4%。这些实验结果充分验证了本文提出的运动补足激励模块（MSE）和联合特征激励模块（UIE）的有效性。

表 1 与现有先进方法在 UCF101 和 HBDB51 上的对比结果

Table 1 Comparison results with existing advanced methods on UCF101 and HBDB51

	方法	年份	预训练数据集	FLOPs/G	UCF101/%	HMDB51/%
双流卷积网络	Two-Stream[6]	2014	—	—	88.00	59.40
	TSN[7]	2016	ImageNet	80.00	91.10	63.20
	压缩激励残差网络[35]	2019	ImageNet	—	92.04	69.30
	Improved Two-stream[37]	2021	—	—	91.25	—
	STCA-Net[36]	2022	ImageNet	68.49	94.1	69.50
	FSTFN[38]	2022	—	—	92.70	65.90
3D 卷积网络	I3DRGB[12]	2017	ImageNet+Kinetics	108.00	95.60	74.80
	ECO[39]	2018	Kinetics	267.00	94.80	72.40
	3D-TDC[40]	2021	Kinetics	84.00	93.82	66.83
	MRTP[41]	2022	Kinetics	—	95.60	—
2D 卷积网络	TSM[18]	2019	ImageNet+Kinetics	33.00	94.50	70.70
	sTNet[42]	2019	ImageNet+Kinetics	53.00	93.50	—
	STM[20]	2019	ImageNet+Kinetics	67.00	96.20	72.20
	TEA[22]	2020	ImageNet+Kinetics	35.00	**96.90**	73.30
	TEInet[21]	2020	ImageNet+Kinetics	33.00	96.70	72.10
	Sep-3D RAN[43]	2022	—	—	91.70	—
	S3D RANs[44]	2022	Kinetics	161.50	93.30	71.20
	Ours		Kinetics*	34.00	96.50	**73.60**

注：Kinetics* 表示只有部分网络结构在 Kinetics 上预训练过。

3.5 消融实验

本小节的消融实验分别在数据集 UCF101-spilt1 和 HMDB51-spilt1 上分析了 MSE 模块、UIE 模块、MSE 模块与 UIE 模块的融合方式对模型性能的影响。

1）MSE 模块的效果分析

表 2 比较了 3 种不同的网络模型分别在 UCF101-spilt1 和 HMDB51-spilt1 上的准确率，分别是：ResNet50；在 ResNet50 基础上添加 CMM 模块，也就是 STM 网络[20]；在 ResNet50 基础上添加 ME 模块，也就是 TEA 网络[22]；在 ResNet50 中添加 MSE 模块。

表 2 MSE 模块的性能影响

Table 2 Effect of MSE module

方法	参数量	UCF101 准确率/%	HMDB51 准确率/%
ResNet50	25.50×10^6	81.30	49.39
ResNet50+CMM	25.70×10^6	82.10	54.18
ResNet50+ME	25.72×10^6	85.35	54.35
ResNet50+MSE	25.72×10^6	**85.96**	**55.01**

由表 2 中的结果可以看出，在 UCF101-spilt1 上，添加 MSE 模块使得 ResNet50 准确度提升了 4%，比添加 ME 模块高出 0.61%。比添加 CMM 模块高了 4.66%；在 HMDB51-spilt1 上，添加 MSE 模块使得 ResNet50 提升了 5%的准确度，比添加 CMM 模块高 0.83%，比添加 ME 模块高 0.66%。实验结果证明 MSE 模块能够提取更加精细的运动特征，从而提高模型识别准确率。

2）UIE 模块的效果分析

为了分析 UIE 模块的作用，表 3 在 UCF101-spilt1 和 HMDB51-spilt1 上比较了 4 种结构的识别准确率：ResNet50、在 ResNet50 基础上添加 UIE 模块、在 ResNet50 加 MSE 模块基础上再添加 SE 模块、在 ResNet50 加 MSE 模块基础上再添加 UIE 模块。

由表 3 中的结果可以看出，在添加 UIE 模块后，ResNet50 在 UCF101-spilt1 和 HMDB51-spilt1 上的识别准确率分别提升了 4.5%和 7.38%，证明 UIE 模块可以有效学习到全局的时空特征，从而获得很大的性能提升。与添加 SE 模块[24] 相比，添加 UIE 模块的识别准确率分别提高了 2.4%和

3.41%，这是因为 UIE 模块在分别进行通道和空间激励的时候，保留了更加完善的特征信息，使得时空信息表达效果更好。

表 3　UIE 模块的性能影响

Table 3　Effect of UIE module

方法	参数量	UCF101 准确率/%	HMDB51 准确率/%
ResNet50	25.50×10^6	81.30	49.39
ResNet50＋UIE	29.50×10^6	85.80	56.77
ResNet50＋MSE＋UIE	29.72×10^6	**87.31**	**58.67**
ResNet50＋MSE＋SE	28.50×10^6	84.91	55.26

3）MSE 模块与 UIE 模块的融合方式对性能的影响

表 4 比较了 MSE 与 UIE 模块的两种不同融合方式对模型性能的影响。如表 4 所列，MSE 与 UIE 串联融合结构在 UCF101-spilt1 和 HMDB51-spilt1 上的识别准确率比 ResNet50 分别高出 6.01%和 9.28%。而 MSE 与 UIE 并联融合结构（即本文方法）在 UCF101-spilt1 和 HMDB51-spilt1 上的识别准确率比 ResNet50 分别高出 13.45%和 22.31%。MSE 与 UIE 模块并联比 MSE 与 UIE 模块串联效果更好，这是因为两个模块都学习了时序信息，串联会造成时序特征的紊乱，而并联能更好地相互补充，进而提升特征的表达能力。同时，比较表 2、表 3 和表 4 的结果，可以看出融合 MSE 与 UIE 模块的效果比单独使用一个模块更好，这证明两个模块分别学习了不同的时空特征，融合后实现了互补提升。

表 4　MSE 模块与 UIE 模块的融合方式对性能的影响

Table 4　Influence of the fusion mode of MSE module with UIE module on performance

（单位：%）

方法	UCF101 准确率	HMDB51 准确率
ResNet50	81.30	49.39
串联	87.31	58.67
并联	94.75	71.70

结束语　本文提出了一种基于注意力机制的多维联合特征激励网络 MFARs 用于视频行为识别。MFARs 网络由多组不同分辨率的 MFAR 模块构成，它包含用于建模局部时空信息的 MSE 模块，和使用多维注意力方式学习时空特征、建模全局时空信息的 UIE 模块。两个模块通过并联的方式相互完善和补充时空信息，使网络学习到更加全面和丰富的时空特征。通过与双流卷积网络和 3D 卷积网络的实验对比，证明了 MFARs 能够有效表达视频时空信息，既保留了双流卷积网络参数量低的优势，又有与 3D 卷积网络一样有效的时空特征建模能力，获得了性能与效率的平衡。下一步将研究多级运动信息表达，用于学习更细微的时空特征，使模型具有细粒度视频动作分类能力，进一步提升行为识别的效率和准确率。

参 考 文 献

[1] WANG H，KLASER A，SCHMID C，et al. Action recognition by dense trajectories[C]// Computer Vision and Pattern Recognition. IEEE，2011：3169-3176.

[2] WANG H，SCHMID C. Action Recognition with Improved Trajectories[C]// IEEE International Conference on Computer Vision. IEEE，2013：3551-3558.

[3] YILMAZ A，MUBARAK S. Actions Sketch：A Novel Action Representation[C]// Computer Vision and Pattern Recognition. IEEE，2005：984-989.

[4] BOBICK A，DAVIS J. An appearance-based representation of action[C]// International Conference on Pattern Recognition. IEEE，1996：307-312.

[5] WANG H，ULLAH M M，KLASER A，et al. Evaluation of local spatio-temporal features for action recognition[C]// Procedings of the British Machine Vision Conference. London：British Machine Vision Association，2009：124.1-124.11.

[6] SIMONYAN K，ZISSERMAN A. Two-Stream Convolutional Networks for Action Recognition in Videos[C]// Neural Information Processing Systems. Curran Associates，Inc.. 2014：568-576.

[7] WANG L，XIONG Y，WANG Z，et al. Temporal Segment Networks：Towards Good Practices for Deep Action Recognition [C]// European Conference on Computer Vision. ECCV，2016：20-36.

[8] ZHOU B，ANDONIAN A，OLIVA A，et al. Temporal Relational Reasoning in Videos[C]// European Conference on Computer Vision. ECCV，2018：831-846.

[9] DIBA A，SHARMA V，VAN GOOL L. Deep Temporal Linear Encoding Networks[C]// 2017 IEEE Conference on Computer Vision and Pattern Recognition. IEEE，2017：1541-1550.

[10] 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.

[11] TRAN D，BOURDEV L，FERGUS R，et al. Learning Spatiotemporal Features with 3D Convolutional Networks[C]// 2015 IEEE International Conference on Computer Vision (ICCV). Santiago，Chile：IEEE，2015：4489-4497.

[12] CARREIRA J，ZISSERMAN A. Quo Vadis，Action Recognition? A New Model and the Kinetics Dataset[C]// 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Honolulu，HI：IEEE，2017：4724-4733.

[13] XIE S，SUN C，HUANG J，et al. Rethinking Spatiotemporal Feature Learning：Speed-Accuracy Trade-offs in Video Classification[C]// European Conference on Computer Vision. ECCV，2018：318-335.

[14] TRAN D，WANG H，TORRESANI L，et al. A Closer Look at Spatiotemporal Convolutions for Action Recognition[C]// IEEE Conference on Computer Vision and Pattern Recognition，2018：6450-6459.

[15] HUANG M，SHANG R X，QIAN H M. Composite Deep Neural Network for Human Activities Recognition in Video[J]. Pattern Recognition and Artificial Intelligence，2022，35(6)：562-570.

[16] ZHANG H B，FU D M，ZHOU K. Video-Based Temporal Enhanced Action Recognition[J]. Pattern Recognition and Artificial Intelligence，2020，33(10)：951-958.

[17] ONG A Y，TANG C，WANG W J. Human Action Recognition Fusing Two-Stream Networks and SVM[J]. Pattern Recognition and Artificial Intelligence，2021，34(9)：863-870.

[18] LIN J，GAN C，HAN S. TSM：Temporal Shift Module for Efficient Video Understanding[C]// 2019 IEEE/CVF International Conference on Computer Vision. Seoul，Korea (South)：IEEE，2019：7082-7092.

[19] SUDHAKARAN S，ESCALERA S，LANZ O. Gate-Shift Networks for Video Action Recognition[C]// 2020 IEEE/CVF

Conference on Computer Vision and Pattern Recognition. Seattle, WA, USA: IEEE, 2020: 1099-1108.
[20] JIANG B, WANG M, GAN W, et al. STM: SpatioTemporal and Motion Encoding for Action Recognition[C]// 2019 IEEE/CVF International Conference on Computer Vision. Seoul, Korea (South): IEEE, 2019: 2000-2009.
[21] LIU Z, LUO D, WANG Y, et al. TEINet: Towards an Efficient Architecture for Video Recognition[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2020, 34(7): 11669-11676.
[22] LI Y, JI B, SHI X, et al. TEA: Temporal Excitation and Aggregation for Action Recognition[C]// 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, WA, USA: IEEE, 2020: 906-915.
[23] WANG L, TONG Z, JI B, et al. TDN: Temporal Difference Networks for Efficient Action Recognition[C]// IEEE Conference on Computer Vision and Pattern Recognition. Computer Vision Foundation. IEEE, 2021: 1895-1904.
[24] HU J, SHEN L, ALBANIE S, et al. Squeeze-and-Excitation Networks[C]// IEEE/CVF Conference on Computer Vision and Pattern Recognition. IEEE, 2018: 7132-7141.
[25] WOO S, PARK J, LEE J Y, et al. CBAM: Convolutional Block Attention Module[C]// Proceedings of the European Conference on Computer Vision (ECCV). 2018: 3-19.
[26] FU J, LIU J, TIAN H, et al. Dual Attention Network for Scene Segmentation[C]// IEEE/CVF Conference on Computer Vision and Pattern Recognition. IEEE, 2019: 3146-3154.
[27] QIU Y, LIU Y, CHEN Y, et al. A2SPPNet: Attentive Atrous Spatial Pyramid Pooling Network for Salient Object Detection [J]. IEEE Transactions on Multimedia, 2022, 25: 1991-2006.
[28] WANG Z, SHE Q, SMOLIC A. ACTION-Net: Multipath Excitation for Action Recognition[C]// IEEE/CVF Conference on Computer Vision and Pattern Recognition. IEEE, 2021: 13209-13218.
[29] BERTASIUS G, WANG H, TORRESANI L. Is Space-Time Attention All You Need for Video Understanding? [C]// International Conference on Machine Learning. PMLR, 2021: 813-824.
[30] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is All you Need[J]. arXiv: 1706.03762, 2017.
[31] ARNAB A, DEHGHANI M, HEIGOLD G, et al. ViViT: A Video Vision Transformer[C]// IEEE/CVF International Conference on Computer Vision. IEEE, 2021: 6816-6826.
[32] SOOMRO K, ZAMIR A R, SHAH M. UCF101: A Dataset of 101 Human Actions Classes From Videos in The Wild[J]. Computer Science, 2012, 3(12): 1-9.
[33] KUEHNE H, JHUANG H, GARROTE E, et al. HMDB: A large video database for human motion recognition[C]// 2011 International Conference on Computer Vision. Barcelona, Spain: IEEE, 2011: 2556-2563.
[34] ZHOU B, KHOSLA A, LAPEDRIZA A, et al. Learning Deep Features for Discriminative Localization[C]// IEEE Conference on Computer Vision and Pattern Recognition. IEEE, 2016: 2921-2929.
[35] LUO H L, TONG K, YUAN P. Spatiotemporal squeeze-and-excitation residual multiplier network for video action recognition [J]. Journal on Communications, 2019, 40(10): 189-198.
[36] LUO H L, CHEN H. Spatial-Temporal Convolution Attention Network for Action Recognition[J]. Computer Engineering and Applications, 2023(9): 150-158.
[37] WANG Y, LIU W, XING W. Improved Two-stream Network for Action Recognition in Complex Scenes[C]// 2021 International Conference on Artificial Intelligence and Electromechanical Automation (AIEA). IEEE, 2021: 361-365.
[38] YANG G, ZOU W. Deep learning network model based on fusion of spatiotemporal features for action recognition[J]. Multimedia Tools and Applications, 2022, 81(7): 9875-9896.
[39] ZOLFAGHARI M, SINGH K, BROX T. ECO: Efficient Convolutional Network for Online Video Understanding[C]// European Conference on Computer Vision (ECCV). 2018: 695-712.
[40] MING Y, FENG F, LI C, et al. 3D-TDC: A 3D temporal dilation convolution framework for video action recognition[J]. Neurocomputing, 2021, 450: 362-371.
[41] ZHANG K, YANG J, ZHANG D, et al. MRTP: Multi-Temporal Resolution Real-Time Action Recognitionpproach by Time-Action Perception[J]. Joural of Xi'an Jiaotong University, 2022, 56(3): 22-32.
[42] HE D, ZHOU Z, GAN C, et al. StNet: Local and Global Spatial-Temporal Modeling for Action Recognition[C]// Proceedings of the AAAI Conference on Artificial Intelligence. 2019: 8401-8408.
[43] ZHANG Z, PENG Y, GAN C, et al. Separable 3D residual attention network for human action recognition[J]. Multimedia Tools and Applications, 2022, 82(4): 5435-5453.
[44] CHEN B, TANG H, ZHANG Z, et al. Video-based action recognition using spurious-3D residual attention networks[J]. IET Image Processing, 2022, 16(11): 3097-3111.

LUO Huilan, born in 1974, Ph.D, professor, Ph.D supervisor. Her main research interests include computer vision and machine learning.