智能系统学报
CAAI TRANSACTIONS ON INTELLIGENT SYSTEMS

# 基于时空-动作自适应融合网络的油田作业行为识别

田枫, 卫宁彬, 刘芳, 韩玉祥, 赵玲, 张思睿, 马贵宝

引用本文:
田枫, 卫宁彬, 刘芳, 等. 基于时空-动作自适应融合网络的油田作业行为识别[J]. 智能系统学报, 2024, 19(6): 1407-1418.
TIAN Feng, WEI Ningbin, LIU Fang, et al. Oilfield operation behavior recognition based on spatio-temporal and action adaptive fusion network[J]. *CAAI Transactions on Intelligent Systems*, 2024, 19(6): 1407-1418.

在线阅读 View online: https://dx.doi.org/10.11992/tis.202309021

**您可能感兴趣的其他文章**

双向特征融合与注意力机制结合的目标检测
Target detection based on bidirectional feature fusion and an attention mechanism
智能系统学报. 2021, 16(6): 1098–1105 https://dx.doi.org/10.11992/tis.202012029

基于CNN–BLSTM的化妆品违法违规行为分类模型
Classification model for judging illegal and irregular behavior for cosmetics based on CNN–BLSTM
智能系统学报. 2021, 16(6): 1151–1157 https://dx.doi.org/10.11992/tis.202104001

基于反馈注意力机制和上下文融合的非模式实例分割
Feedback attention mechanism and context fusion based amodal instance segmentation
智能系统学报. 2021, 16(4): 801–810 https://dx.doi.org/10.11992/tis.202007042

可能性匹配知识迁移原型聚类算法
Possibility–matching based knowledge transfer prototype clustering algorithm
智能系统学报. 2020, 15(5): 978–989 https://dx.doi.org/10.11992/tis.201810028

基于注意力融合的图像描述生成方法
An image caption generation method based on attention fusion
智能系统学报. 2020, 15(4): 740–749 https://dx.doi.org/10.11992/tis.201910039

时空域融合的骨架动作识别与交互研究
Research on skeleton–based action recognition with spatiotemporal fusion and humanrobot interaction
智能系统学报. 2020, 15(3): 601–608 https://dx.doi.org/10.11992/tis.202006029

**DOI**: 10.11992/tis.202309021
**网络出版地址**: https://link.cnki.net/urlid/23.1538.tp.20240910.1943.012

# 基于时空−动作自适应融合网络的油田作业行为识别

田枫, 卫宁彬, 刘芳, 韩玉祥, 赵玲, 张思睿, 马贵宝
(东北石油大学 计算机与信息技术学院, 黑龙江 大庆 163318)

**摘 要**: 为解决油田作业现场复杂环境对行为识别算法造成干扰而引起的误检、漏检问题, 提出一种时空−动作自适应融合网络, 用于油田作业现场的人员行为识别。构建的网络首先使用稀疏采样的策略对视频进行处理, 再通过特征提取网络进行特征提取, 其核心模块分别为时空注意力模块、动作强化模块和自适应特征融合模块。时空注意力模块完成特征的时空重要性再分配, 建立不同帧之间的时间关联; 动作强化模块完成背景的弱化、人体动作的强化, 使模型聚焦于人体动作; 特征融合模块在二者并行特征强化后进行自适应特征融合, 最终通过全连接层和 Softmax 层来实现行为的分类。为验证所提网络的效果, 分别在公共数据集和油田自制数据集上将所提模型与经典网络进行对比, UCF101 数据集上的 Top-1 准确率相较于 SlowOnly(SlowFast 模型的 Slow 分支)和 TSM(temporal shift module)分别提升了 3.33% 和 1.61%, HMDB51 数据集上的 Top-1 准确率相较于 SlowOnly 和 TSM 分别提升了 8.56% 和 1.83%, 在油田自制数据集上与 TSN(temporal segment network)、TSM、SlowOnly 进行对比, 结果显示所提模型准确率得到大幅提升, 验证了时空−动作自适应融合网络在油田作业现场环境下的有效性, 更适用于油田作业环境下的行为识别任务。

**关键词**: 行为识别; ResNet50; 注意力机制; 油田作业; 特征融合; 时空注意力; 动作注意力; 复杂场景
**中图分类号**: TP391 **文献标志码**: A **文章编号**: 1673−4785(2024)06−1407−12

**中文引用格式**: 田枫, 卫宁彬, 刘芳, 等. 基于时空−动作自适应融合网络的油田作业行为识别 [J]. 智能系统学报, 2024, 19(6): 1407–1418.
**英文引用格式**: TIAN Feng, WEI Ningbin, LIU Fang, et al. Oilfield operation behavior recognition based on spatio-temporal and action adaptive fusion network[J]. CAAI transactions on intelligent systems, 2024, 19(6): 1407–1418.

# Oilfield operation behavior recognition based on spatio-temporal and action adaptive fusion network

TIAN Feng, WEI Ningbin, LIU Fang, HAN Yuxiang, ZHAO Ling, ZHANG Sirui, MA Guibao
(School of Computer and Information Technology, Northeast Petroleum University, Daqing 163318, China)

**Abstract**: A spatiotemporal and action adaptive fusion network is proposed for personnel behavior recognition in oilfield operation sites to address the problems of false positives and negatives caused by the complex environment of oilfield operations interfering with behavior recognition algorithms. First, the videos are processed on the constructed network using a sparse sampling strategy, and features on the feature extraction network are then extracted. The core modules of the network include spatiotemporal attention, action reinforcement, and adaptive feature fusion modules. The spatiotemporal attention module redistributes the spatiotemporal importance of features, establishing temporal correlations between different frames. The action reinforcement module weakens the background and enhances human body movements, allowing the model to focus on human actions. The feature fusion module adaptively combines the parallel features after reinforcement. Finally, behavior classification is achieved through fully connected layers and a SoftMax layer. The model is compared with classic networks on public and self-built oilfield datasets to verify the effectiveness of the proposed network. The Top-1 accuracy on the UCF101 dataset shows a 3.33% improvement over SlowOnly, the Slow branch of the SlowFast model, and a 1.61% improvement over the temporal shift module (TSM). On the HMDB51 dataset, the Top-1 accuracy improves by 8.56% and 1.83% compared to SlowOnly and TSM, respectively. Additionally, when evaluated on the self-built oilfield dataset, the proposed model shows a notable improvement in accuracy over the temporal segment network, TSM, and SlowOnly. This result validates the effectiveness of the spatiotemporal and action adaptive fusion network in oilfield operations and confirms its suitability for behavior recognition tasks in such environments.

**Keywords**: behavior recognition; ResNet50; attention mechanism; oilfield operation; feature fusion; spatio-temporal attention; action attention; complex scenes

**收稿日期**: 2023−09−11. **网络出版日期**: 2024−09−11.
**基金项目**: 黑龙江省自然科学基金项目(LH2021F004).
**通信作者**: 刘芳. E-mail: lfliufang1983@126.com.

石油产业作为国家的支柱产业之一, 其安全防控也是产业内容中的重中之重, 定期进行设备

©《智能系统学报》编辑部版权所有

检修、安全检查是保证作业安全的常规手段，然而生产过程中作业人员的操作行为作为不可控因素，例如摘安全帽、摘手套等，也是导致发生安全事故的重要原因。采用视频监控的方式进行监督，其效果受人为因素的影响，难以及时准确地发现不规范行为。目前，将算法接入监控视频，通过行为识别算法[1-2]对连续帧的画面进行识别并对危险行为发出警报是油田安全防控的重要手段。

行为识别作为计算机视觉的热门研究方向之一，按照输入的特征划分为两大类别，分别是基于图像的算法和基于人体关键点的算法。其中，基于图像的算法具体又分为单分支网络和多分支网络，Tran 等[3]提出了 C3D（convolutional 3D）模型，在卷积核上引入了一个额外的时间维度，使得模型可以在三维空间对视频序列进行卷积操作，同时获取时间和空间信息，该网络的设计启发了后续的神经网络，如 I3D（inflated 3D）[4]和 R(2+1)D（residual 3D）[5]。Two-Stream CNN[6]是最早引入双流网络结构的模型之一，该模型由 2 个独立的分支组成，分别处理色彩信息和光流信息，2 个分支的特征在融合后使用全连接层进行分类。TSN（temporal segment network）[7]在双流网络的基础上通过分段处理视频序列来利用时间信息；TSM（temporal shift module）[8]则是在双流网络中引入了时间位移模块来对时间信息进行处理，二者均在行为识别领域取得较好的效果。Feichtenhofer 等[9]提出的 SlowFast 模型则是将双流网络与 C3D 结合起来，旨在同时捕获视频中的快速动作与慢速动作，SlowFast 模型在行为识别领域表现出色，其设计思想也启发了其他模型的发展，如 SlowOnly 模型（SlowFast 模型的 Slow 分支）和 FastOnly 模型（SlowFast 模型的 Fast 分支）等，这些模型都在识别任务中取得了突破和应用。Yan 等[10]提出 ST-GCN 模型，该模型将人体关键点当作自然连接的三维拓扑结构，通过图卷积进行特征提取，从而更好地捕捉人体姿态序列的时空特征和动作特征。PoseC3D 是一种基于人体关键点的三维卷积神经网络，通过利用人体姿态的先验信息和三维卷积的特性来有效处理时空信息，更加准确地识别人体行为[11]。

近些年，行为识别算法在作业场景中得到了一定程度的发展和应用。田枫等[12]提出一种基于图卷积的行为识别方法，按照定位、提取关键点、人员追踪和图卷积分类的流程进行作业现场行为识别，该方法在油田现场的不规范阀门操作识别率达到 96.7%。陆昱翔等[13]提出基于 Transformer 时空自注意力的行为识别算法，模型在提取空间特征的基础上增加了时间特征的分析，从时空维度对视频帧进行处理，最终在自制数据集上达到了 98.54% 的准确率。饶天荣等[14]将 C3D 与 ST-GCN 结合，使用 C3D 提取图像特征，使用 ST-GCN 提取人体关键点特征，使用交叉注意力将得到的 2 种特征进行融合并进行分类，在煤矿数据集上得到了有效提升。以上算法一定程度上解决了作业现场进行行为识别的问题，但仍存在缺陷：基于 Transformer 的方法计算量庞大，识别效率不高[15]；基于人体关键点的算法需要提前进行关键点提取，且训练时需要大量关键点数据，扩展性不强；部分基于双流框架的算法也存在效率低下和需要提前获取如光流或人体关键点等先验知识的问题。

考虑到三维卷积神经网络的可扩展性、一次卷积即可实现时间和空间维度的特征提取，同时基于图像特征的行为识别算法可以实现端到端的训练和推理，因此本研究采用三维卷积进行网络设计，用单分支网络来完成行为识别任务。ResNet50 网络凭借其优秀的残差结构及其表达能力，成为许多模型骨干网络的首选[16]，本研究在此基础上构建一种并行双注意力机制同时自适应特征融合的行为识别模型，通过显式地引入时空注意力模块与动作强化模块并进行自适应特征融合来完成油田现场行为识别的任务，时空注意力模块对特征进行时空维度上的重要性再分配，强化网络的时空信息提取能力，动作强化模块通过时间维度上的错位相减，过滤掉不相关的背景信息，增强网络对于动作变化的敏感度，自适应特征融合通过计算时空特征与动作特征来实现时空与动作 2 种特征有侧重、自适应地融合。

## 1　时空−动作自适应融合网络构建

本研究构建的时空−动作自适应融合网络以三维 ResNet50 作为其基础网络，其核心为提出的时空注意力模块、动作强化模块以及自适应特征融合模块。具体地说，先将时空注意力与动作强化 2 个模块并联在残差网络的数据处理层后面，后面紧接着自适应特征融合模块对 2 种特征进行融合，融合方式如图 1 上半部分所示。此外，残差网络的 4 个残差层之后都以同样的方式并联 2 个模块及自适应特征融合模块，详细的组合方式如图 1 下半部分结构所示，每个残差层都由若干个 Bottleneck 块组成，本研究将提出的时空注意力与

动作强化 2 个模块连接在 Bottleneck 块之后, 二者对视频特征分别处理再自适应融合, 起到特征再处理的效果, 图 1 中⊙代表逐元素相乘, ⊕代表逐元素相加; $C$ 为图像通道数, $W$ 为图像宽度, $H$ 为图像高度。

对于输入视频, 采用稀疏采样的策略处理为 $N\times T$ 个分辨率为256×256的视频帧作为网络的输入, 其中, $N$代表采样后的视频片段数目, $T$代表视频片段的视频帧数目, 首先经过数据处理层得到$N\times 64\times T\times 64\times 64$的特征向量, 之后依次经过改进后的 4 个残差层, 每经过一个残差层通道数翻倍、特征宽高减半, 然后经过全连接层(fully connected layers, 图 1 中记为 FC)及 Softmax 层得到形状为$N\times C_{\text{lasses}}$的变量, 其代表$N$个片段的分类得分, 其中$C_{\text{lasses}}$代表类别数量, 最后将各片段的分类得分求平均(average polling, 图 1 中记为 AP)作为整个视频的分类得分。模型在训练时使用交叉熵损失函数计算模型的分类损失:

$$L=\frac{1}{M}\times\sum_{i=1}^{M}L_i=-\frac{1}{M}\times\sum_{i=1}^{M}\sum_{c=1}^{C_{\text{lasses}}}y_{ic}\log p_{ic} \tag{1}$$

式中: $M$代表样本数量; $y_{ic}$代表样本$i$真实类别是否等于$c$, 是的话则$y_{ic}=1$, 否则$y_{ic}=0$; $p_{ic}$代表样本$i$预测类别为$c$的概率。

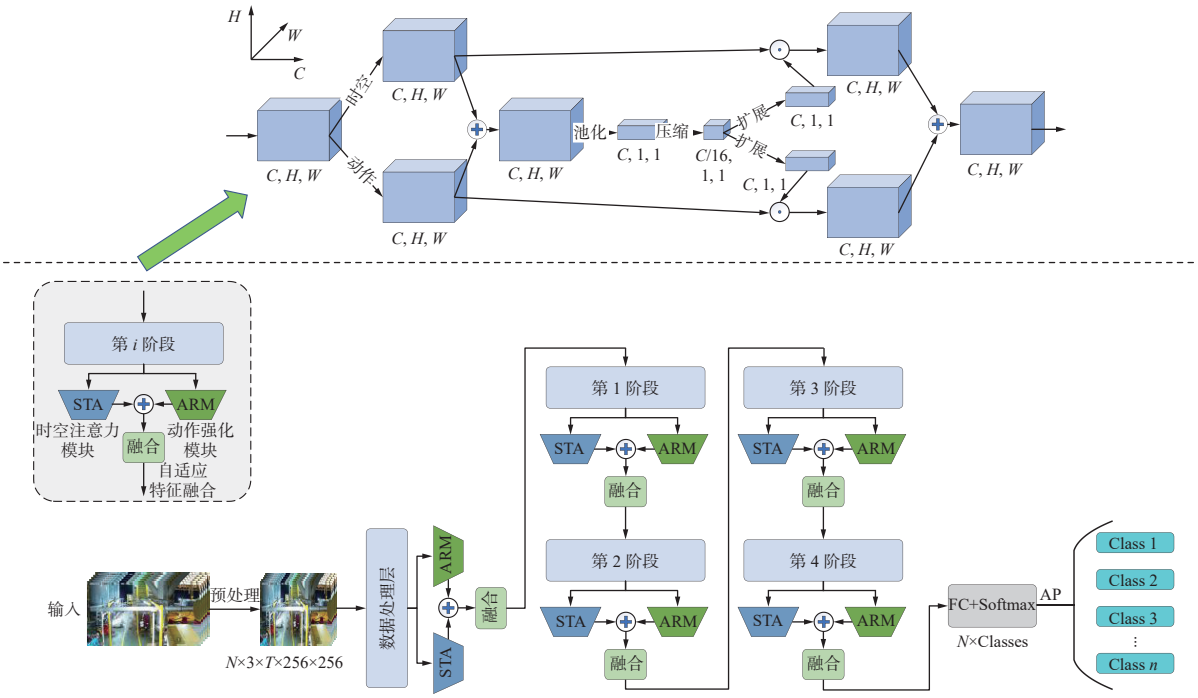



**图 1　时空−动作自适应融合网络**

**Fig. 1　Spatio-temporal and action adaptive fusion network**

### 1.1　数据预处理

对于每个视频片段, 通常帧率为 25~30 帧, 整个视频都进行特征提取会含有大量重复视频帧, 增加计算量, 因此采用均匀采样的方式对视频进行预处理。均匀采样用于从连续的视频序列中提取离散的帧来代表整个视频内容, 以降低数据量并减少处理所需的计算资源, 具体操作为将待检测视频平均分为若干个片段, 每个片段中以相同的时间间隔选择帧, 以确保所选帧在时间上均匀分布。

对于输入视频, 采用均匀采样的策略, 整个视频被平均切分为设定好的$N$个片段, 每个片段中以固定间隔$I$抽取, $T$帧作为一个片段的采样序列, 最终整个视频被处理为$N\times T$个视频帧作为网络的输入。本研究中设定$N$为 8, $I$为 4, 采样完成后对视频帧进行缩放、正则化等处理, 最终得到$N\times T$个分辨率为256×256的视频帧作为网络的输入。

### 1.2　时空注意力模块

时空注意力模块被设计为一个在时间和空间维度上计算特征重要性的模块, 该模块根据输入的特征张量 $\boldsymbol{X}\in\mathbf{R}^{N\times C\times T\times H\times W}$ 计算出一个系数矩阵 $\boldsymbol{M}\in\mathbf{R}^{N\times 1\times T\times H\times W}$, 根据这个矩阵对输入的特征张量进行时空维度上的特征重要性再分配, 如图 2 所示。模块接收一个特征张量 $\boldsymbol{X}\in\mathbf{R}^{N\times C\times T\times H\times W}$, 首先对特征张量分别进行跨通道的平均池化与最大池化, 将二者拼接后以此来获得输入特征的全局空间信息($\boldsymbol{F}$), 公式化表达为

$$\boldsymbol{F} = \mathrm{cat}(\boldsymbol{F}_{\mathrm{Avg}}, \boldsymbol{F}_{\mathrm{Max}}) \tag{2}$$

$$\boldsymbol{F}_{\mathrm{Avg}} = \frac{1}{C} \times \sum_{i=1}^{C} \boldsymbol{X}[:, i, :, :, :] \tag{3}$$

$$\boldsymbol{F}_{\mathrm{Max}} = \max_{i} \boldsymbol{X}[:, i, :, :, :] \tag{4}$$

式中：$\boldsymbol{F} \in \mathbf{R}^{N\times 2\times T\times H\times W}$，$\boldsymbol{F}_{\mathrm{Avg}}$、$\boldsymbol{F}_{\mathrm{Max}} \in \mathbf{R}^{N\times 1\times T\times H\times W}$，cat代表张量拼接操作，$C$代表通道数，$i$为其中一个通道的索引。

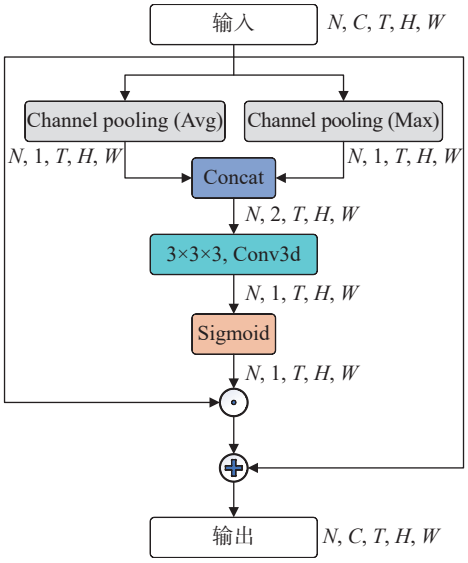



**图 2　时空注意力模块**

**Fig. 2　Spatio-temporal attention module**

接着将提取到的全局空间信息$\boldsymbol{F}$使用3×3×3的三维卷积核$\boldsymbol{K}$进行卷积，通过卷积操作学习特征的时空重要性，使用Sigmoid激活函数将结果映射至0~1便得到了时空重要性矩阵$\boldsymbol{M}$，公式化表达为

$$\boldsymbol{M} = \delta(\boldsymbol{F} * \boldsymbol{K}) \tag{5}$$

式中：$\delta$代表sigmoid激活函数。

最后，将输入的特征张量与矩阵$\boldsymbol{M}$相乘，便完成了输入特征的时空重要性再分配，从输入到输出的完整公式化表达为

$$\boldsymbol{Y} = \boldsymbol{M} \odot \boldsymbol{X} \oplus \boldsymbol{X} \tag{6}$$

### 1.3　动作强化模块

动作强化模块被设计为在通道和时间维度上聚焦关键动作的模块，该模块对输入特征张量$\boldsymbol{X} \in \mathbf{R}^{N\times C\times T\times H\times W}$的相邻帧间的动作信息进行建模，依据相邻帧间的动作信息计算出一个系数矩阵$\boldsymbol{M} \in \mathbf{R}^{N\times C\times T\times 1\times 1}$，以此来强化关键动作帧、弱化静态背景帧，如图3所示。该模块接收一个特征张量$\boldsymbol{X} \in \mathbf{R}^{N\times C\times T\times H\times W}$，与时空注意力模块类似，首先将特征张量分别进行跨通道的平均池化与最大池化，将二者拼接后得到输入特征的全局空间信息$\boldsymbol{F} \in \mathbf{R}^{N\times 2\times T\times H\times W}$。

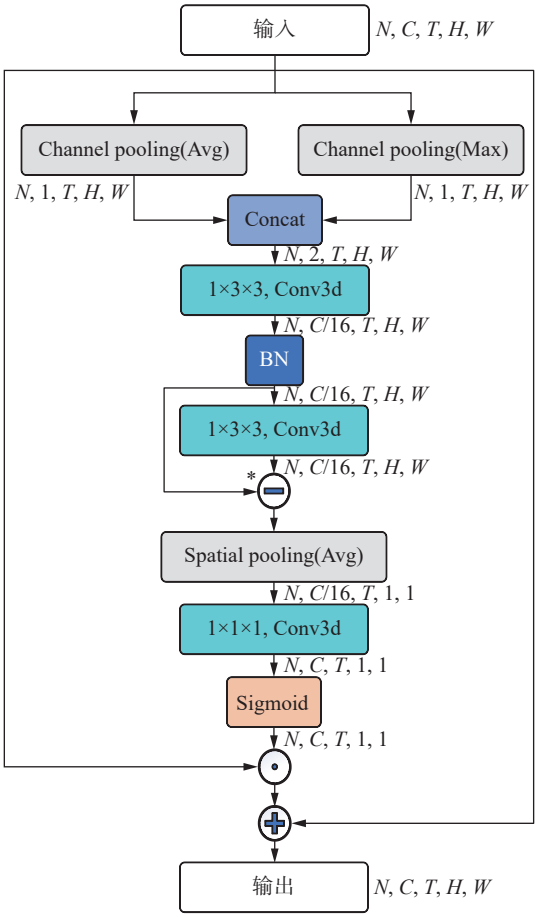



**图 3　动作强化模块**

**Fig. 3　Action reinforcement module**

接着对提取到的全局空间信息$\boldsymbol{F}$，使用一个1×3×3的卷积核$\boldsymbol{K}_1$进行卷积，在进一步提取特征的同时，将特征的通道维度从2扩展为$C$/16，卷积后对结果进行归一化处理就得到了通道压缩后的特征信息$\boldsymbol{F}_{C_1} \in \mathbf{R}^{N\times C/16\times T\times H\times W}$，公式化表达为

$$\boldsymbol{F}_{C_1} = \mathrm{bn}(\boldsymbol{F} * \boldsymbol{K}_1) \tag{7}$$

式中bn代表归一化操作。

将特征通道压缩为1/16后对动作信息进行建模，在保持通道数不变的情况下，使用1×3×3的卷积核$\boldsymbol{K}_2$对$\boldsymbol{F}_{C_1}$进行卷积得到$\boldsymbol{F}_{C_2}$，将$\boldsymbol{F}_{C_2}$与$\boldsymbol{F}_{C_1}$在时间维度上进行错位相减后即可得到$T$−1个时间点上的运动信息，将最后一个时间点$T$的信息进行补0操作来保证数据维度的统一，最终得到输入特征的动作信息$\boldsymbol{F}_{\mathrm{A}}$，公式化表达为

$$\boldsymbol{F}_{\mathrm{A}} = [F_{\mathrm{a}}(1)\ \ F_{\mathrm{a}}(2)\ \ \cdots\ \ F_{\mathrm{a}}(T-1)\ \ 0] \tag{8}$$

$$\boldsymbol{F}_{\mathrm{a}} = \boldsymbol{F}_{C_2}[:, :, 2:T, :, :] - \boldsymbol{F}_{C_1}[:, :, 1:T-1, :, :] \tag{9}$$

$$\boldsymbol{F}_{C_2} = \boldsymbol{F}_{C_1} * \boldsymbol{K}_2 \tag{10}$$

式(8)代表补零操作，式(9)代表错位相减，将错位相减再补零的操作简化为图3中的$^{*}\ominus$符号。

将动作信息$\boldsymbol{F}_{\mathrm{A}}$在空间维度上进行平均池化

以得到通道及时间维度上对应空间信息的权重 $\boldsymbol{F}'_{\mathrm{A}}$, 然后使用1×1×1的卷积核 $\boldsymbol{K}_3$ 进行卷积, 将特征的通道数从 $C/16$ 扩展为 $C$, 完成通道数的复原操作, 最后使用Sigmoid激活函数将结果映射至0~1便得到了动作重要性矩阵 $\boldsymbol{M}$, 公式化表达为

$$\boldsymbol{M}=\delta(\boldsymbol{F}'_{\mathrm{A}}*\boldsymbol{K}_3) \tag{11}$$

$$\boldsymbol{F}'_{\mathrm{A}}=\frac{1}{H\times W}\sum_{i=1}^{H}\sum_{j=1}^{W}\boldsymbol{F}_{\mathrm{A}}[:,:,:,i,j] \tag{12}$$

最后, 将输入的特征张量与矩阵 $\boldsymbol{M}$ 相乘, 便完成了输入特征的运动信息增强, 从输入到输出的完整公式化表达为

$$\boldsymbol{Y}=\boldsymbol{M}\odot\boldsymbol{X}\oplus\boldsymbol{X} \tag{13}$$

### 1.4　自适应特征融合模块

经过时空注意力模块与动作强化模块 2 个模块的特征提取后, 得到了时空特征 $\boldsymbol{F}_{\mathrm{st}}$ 与动作特征 $\boldsymbol{F}_{\mathrm{a}}$, 自适应特征融合机制将会分别计算出 2 种特征的重要性矩阵 $\boldsymbol{W}_{\mathrm{a}}$ 和 $\boldsymbol{W}_{\mathrm{b}}$, 动态地分配给其不同的权重, 实现有侧重地对时空维度和动作模式两类特征进行融合, 如图 4 所示。将时空特征与动作特征先逐元素相加, 再在空间维度上进行平均池化, 将其调整形状以便后续通道维度上的操作, 最终得到跨通道的权重 $\boldsymbol{W}_{\mathrm{gb}}\in\mathbf{R}^{N\times T\times C\times 1\times 1}$, 公式化表达为

$$\boldsymbol{W}_{\mathrm{gb}}=\frac{1}{H\times W}\sum_{i=1}^{H}\sum_{j=1}^{W}(\boldsymbol{F}_{\mathrm{st}}+\boldsymbol{F}_{\mathrm{a}})[:,:,:,i,j] \tag{14}$$

接着进行类似于 SENet[17] 的通道压缩和扩张操作, 使用 2 个全连接层 $\boldsymbol{F}_{C_1}$ 和 $\boldsymbol{F}_{C_2}$ 来实现, $\boldsymbol{F}_{C_1}$ 将通道压缩为原来的 1/16, $\boldsymbol{F}_{C_2}$ 将通道扩张为原来的 2 倍, 最终得到 $\boldsymbol{W}_{\mathrm{ab}}\in\mathbf{R}^{N\times T\times 2\times C\times 1\times 1}$:

$$\boldsymbol{W}_{\mathrm{ab}}=\boldsymbol{F}_{C_2}(\boldsymbol{F}_{C_1}(W_{\mathrm{gb}})) \tag{15}$$

得到权重矩阵 $\boldsymbol{W}_{\mathrm{ab}}$, 经过 Softmax 后将其在通道维度上进行拆分得到 $\boldsymbol{W}_{\mathrm{a}}$ 和 $\boldsymbol{W}_{\mathrm{b}}$ 2 部分权重, 将其形状调整为 $N\times C\times T\times 1\times 1$ 后, 与时空和动作特征相乘再求和即可得到时空、动作融合后的特征:

$$\boldsymbol{Y}=\boldsymbol{F}_{\mathrm{st}}\odot W_{\mathrm{a}}\oplus\boldsymbol{F}_{\mathrm{a}}\odot W_{\mathrm{b}} \tag{16}$$

$$\boldsymbol{W}_{\mathrm{a}}=\mathrm{Softmax}(\boldsymbol{W}_{\mathrm{ab}})[:,1,:,:] \tag{17}$$

$$\boldsymbol{W}_{\mathrm{b}}=\mathrm{Softmax}(\boldsymbol{W}_{\mathrm{ab}})[:,2,:,:] \tag{18}$$

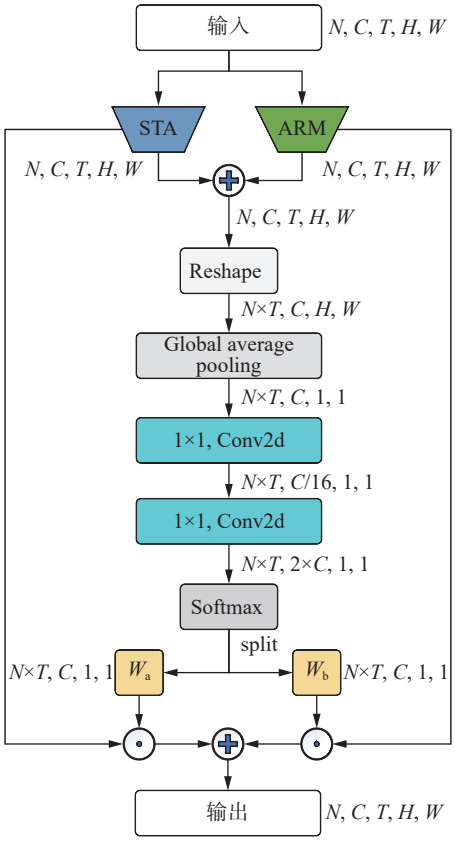



**图 4　自适应特征融合**

**Fig. 4　Adaptive feature fusion**

## 2　实验与结果分析

### 2.1　实验数据集

本研究使用公开数据集 HMDB51(human motion database)[18]、UCF101(University of Central Florida)[19] 和油田自制数据集进行模型的对比评价, 各个数据集的信息见表 1。

**表 1　数据集视频数量、标签数量及行为种类**

**Table 1　Number of videos in the dataset, the number of labels, and the types of behaviors**

| 数据集名称 | 视频数量 | 标签数量 | 行为种类 |
|---|---|---|---|
| HMDB51 | 6 766 | 51 | 5(普通的面部动作、复杂的面部动作、普通的肢体动作、复杂的肢体动作、多人互动肢体动作) |
| UCF101 | 13 320 | 101 | 5(人与物体交互、人与人互动、单纯的肢体动作、演奏乐器、体育运动) |
| 油田自制数据集 | 1 800 | 16 | 3(人与物体交互、单纯的肢体动作、人机协作) |

HMDB51 是一个来自于电影、公共数据库和网络视频等途径的 6 766 个真实视频片段组成的行为识别数据集, 整个数据集包含了 5 种行为: 普通面部动作、复杂面部动作、普通肢体动作、复杂肢体动作、多人互动肢体动作。数据集中的视频片段被划分为 51 个, 如踢球、骑自行车和拥抱等具体的行为类别, 每类行为所包含的视频片段至少为 101 个, 其中, 70% 的片段用来训练, 30% 的

片段用来测试。

UCF101 是由来自于 YouTube 的 13 320 个真实视频片段组成的行为识别数据集，整个数据集包含了五大行为种类：人与物体交互、人与人互动、肢体动作、演奏乐器、体育运动。数据集中的视频片段被划分为 101 个，如俯卧撑、弹钢琴和投篮等具体的行为类别，同时每个类别中的视频按照背景、视角等因素又被划分为 25 组，每一组有 4~7 个视频。按照 70% 与 30% 的比例将数据集划分为训练集和测试集。

油田自制数据集主要由作业现场监控视频、网络素材和模拟拍摄视频组成，作业现场不同监控涉及到不同作业任务，因此以摄像头为单位筛选出有效的视频片段并进行裁剪，同时模拟拍摄了不规范作业视频作为数据集的补充，每个类别不同视角、不同光照进行重复拍摄以保证数据集的多样性。采用与 HMDB51 和 UCF101 同样的数据组织和标注方式，整个数据集由 1 800 个视频组成，包含了三大行为种类：人与物体交互、肢体动作、人机协作，具体划分为 16 个行为类别，分别有蹲下、站起、摔倒、上楼梯、下楼梯、摘手套、戴手套、摘安全帽、戴安全帽、拿起喷枪、放下喷枪、拿起钻头、放下钻头、递喷枪、递钻头和踩踏小盖板，同样按照 70% 与 30% 的比例将数据集划分为训练集和测试集，数据集部分如图 5 所示。

(a) 下楼梯　(b) 拿起喷枪

(c) 摘手套　(d) 摘安全帽

**图 5　油田自制数据集**

**Fig. 5　Presentation of self-built oilfield dataset**

相较于 HMDB51 数据集、UCF101 数据集在背景复杂度、视角变换和照明条件等因素上都更为复杂，动作种类也更多，但 HMDB51 数据集对低级特征（颜色、场景等）的依赖程度更低，更适合用高级特征去做识别任务。油田自制数据集面向油田作业场景，是为解决油田作业场景下施工行为识别问题而建立，在此数据集上进行识别任务能有效对施工现场作业行为进行监管。

### 2.2　实验环境

本研究使用的模型初始化权重为在 kinetics-400[20] 数据集上进行训练的 3D ResNet50，采用的优化策略为随机梯度下降，初始学习率为 0.000 25，动量为 0.9，衰减率为 0.000 1，迭代次数为 40 次，在第 16 次和第 31 次迭代时对模型的学习率进行调整。

实验配置为：操作系统为 Ubuntu 20.04，CPU 型号为 Intel Xeon(R) Silver 4214，GPU 型号为 Tesla P40，使用 Pytorch 11.0.3 作为深度学习框架，CUDA 版本为 11.7，编译器为 Pycharm 2022.3。

### 2.3　评价指标

本研究使用准确率（Accuracy，记为 $A_{\text{ccuracy}}$）[21] 作为模型的性能评价指标，其用来衡量模型在数据集上的分类能力：

$$A_{\text{ccuracy}}=\frac{T_{\text{P}}+T_{\text{N}}}{T_{\text{P}}+F_{\text{P}}+T_{\text{N}}+F_{\text{N}}} \tag{19}$$

式中：$T_{\text{P}}$ 代表预测为正的正样本数目，$T_{\text{N}}$ 代表预测为负的负样本数目，$F_{\text{P}}$ 代表预测为正的负样本数目，$F_{\text{N}}$ 代表预测为负的正样本数目。

本研究使用 Top-1 Acc 和 Top-5 Acc 2 种准确率（记为 $T_{\text{op-}k\text{ Acc}}$，$k$ 取值为 1 或 5）来对模型进行评价，其中，Top-1 Acc 使用得分最高的类别作为预测值，该预测值与真实标签相同则分类正确，否则分类错误；而 Top-5 Acc 则是看得分最高的 5 个类别中是否出现真实标签，如果出现则分类正确，否则分类错误。具体分别为

$$T_{\text{op-}k\text{ Acc}}=\frac{1}{M}\times\sum_{i=1}^{M}p_{\text{red}_i} \tag{20}$$

$$p_{\text{red}_i}=\begin{cases}1, & y_i \text{ in argsort}(p_i,k)\\ 0, & y_i \text{ not in argsort}(p_i,k)\end{cases} \tag{21}$$

式中：$k$=1 代表 Top-1 准确率，$k$=5 代表 Top-5 准确率；$M$ 代表样本数量；$p_{\text{red}i}$ 代表样本 $i$ 是否分类正确，正确则 $p_{\text{red}i}$=1，否则 $p_{\text{red}i}$=0；$y_i$ 代表样本 $i$ 的真实标签；$p_i$ 代表样本 $i$ 的预测概率；argsort($p_i$,$k$) 代表获取样本 $i$ 中预测概率最高的前 $k$ 个标签。

### 2.4　公共数据集实验结果与分析

首先在公共数据集 HMDB51 和 UCF101 上，将本研究模型与主流的模型 Two-Stream、Two-Stream+LSTM[22]、C3D[23]、SlowOnly、TSN、TSM 进行对比，结果见表 2。

**表 2　不同模型在公共数据集上的实验结果**
**Table 2　Experimental results of different models on public datasets**　%

| 模型 | 骨干网络 | UCF101 | | HMDB51 | |
|---|---|---|---|---|---|
| | | Top-1 | Top-5 | Top-1 | Top-5 |
| Two-Stream | | 88.00 | — | 59.40 | — |
| Two-Stream+LSTM | | 88.60 | — | 65.20 | — |
| C3D | ResNext101[24] | 83.27 | 95.90 | 51.60 | |
| SlowOnly | ResNet50 | 92.78 | 99.42 | 65.95 | 91.05 |
| TSN | | 83.03 | 96.78 | 56.08 | 84.31 |
| TSM | | 94.50 | 99.58 | 72.68 | 92.03 |
| 本研究 | | 96.11 | 99.68 | 74.51 | 93.99 |

由表 2 可知, 各模型在 HMDB51 数据集上的精度差别较大, 可见该数据集更能检验模型的高级特征提取能力, 在 UCF101 数据集上各模型的精度差异较小, 可见该数据集对于低级特征的依赖程度较大。其中 SlowOnly 模型作为 SlowFast 模型的 Slow 分支, 其使用较少帧数与较大通道数来学习视频中的空间语义信息, 具有较强的空间特征提取能力; TSM 模型则是在 TSN 模型的基础上加入了时间位移操作, 使模型能够关联前后帧中的特征, 从而有效地利用视频中的时序信息。以上 2 个模型在 UCF101 数据集和 HMDB51 数据集上都有着优秀的性能, 因此本研究将重点关注 SlowOnly 模型与 TSM 模型, 与其进行对照。

在 UCF101 数据集上, SlowOnly 模型仅依靠连续帧的色彩信息在 Top-1 准确率达到了 92.78%, 而 TSM 模型在 TSN 模型的基础上加入了时间位移操作, 使得其 Top-1 准确率从 83.03% 提高到 94.50%, 本研究模型的 Top-1 准确率相较于 SlowOnly 和 TSM 分别提高了 3.33 百分点和 1.61 百分点, 可以看出在空间信息与时间信息的提取上本研究模型有着更好的性能; 在 HMDB51 数据集上, SlowOnly 模型的 Top-1 准确率为 65.95%, TSM 模型的 Top-1 准确率为 72.68%, 其相较于 TSN 模型提升了 16.6 百分点, 可见时序信息的利用在该数据集上作用更大, 本研究模型的 Top-1 准确率相较于 SlowOnly 提高了 8.56 百分点, Top-1 准确率相较于 TSM 的 72.68% 提升了 1.83 百分点, 因此本研究模型的信息提取能力更好。

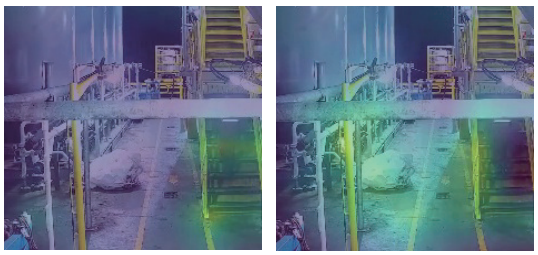
### 2.5　油田自制数据集实验结果与分析

进一步将本研究模型与几个经典模型进行对比, 实验结果见表 3。由实验结果可知, 几个主流模型的准确率相较于公共数据集均有大幅度下降, 这是因为油田作业环境下场景较为复杂, 容易受到光照、视角和背景等因素的干扰, 由此可以看出, 复杂场景尤其是工业场景的行为识别任务具有挑战性。本研究模型在油田自制数据集上的性能相较于几个主流模型具有优势, 例如前面在 UCF101 与 HMDB51 数据集上均有优越性能的 TSM 模型, 其在油田自制数据集上的 Top-1 准确率为 49.52%, Top-5 准确率为 94.17%, 虽然较其他几个主流模型高, 但是与本研究模型的 Top-1 与 Top-5 准确率相比, 分别低了 24.77 百分点和 5.71 百分点。结合公共数据集与油田自制数据集上的比较分析, 说明本研究模型具有较好的信息提取能力, 更能适应复杂场景下的识别任务。

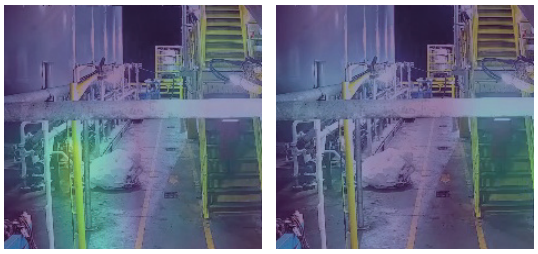
**表 3　不同模型在油田自制数据集上的实验结果**
**Table 3　Experimental results of different models on a self-built oilfield dataset**　%

| 模型 | 油田自制数据集 | |
|---|---|---|
| | Top-1 | Top-5 |
| TSN | 40.36 | 93.81 |
| TSM | 49.52 | 94.17 |
| SlowOnly | 30.24 | 86.43 |
| 本研究 | 74.29 | 99.88 |

通过可视化结果来进一步说明本研究模型的场景适应性, 使用 GradCam[25] 对图片进行热图可视化, 该方法通过计算目标分类对目标卷积层的梯度, 进行相关计算后得到类激活图, 通过类激活图可以了解到模型的决策依据, 了解模型对不同区域的关注程度, 颜色越亮则关注程度越高, 如图 6 所示。由图 6 可以看到, 本研究模型的关注点更聚焦于人体动作本身及其附近环境, 而 SlowOnly 和 TSN 则被图片中的前景物体所干扰, SlowOnly 除了一小部分集中在人体身上, 更多关注点放到了管道上, 而 TSN 则基本上只关注到管道, TSM 虽然也聚焦于人体动作及其附近环境, 但是对比人体上热图的颜色, 显然本研究模型关注程度更高一些。



(a) 本研究模型　(b) SlowOnly



(c) TSN　(d) TSM

**图 6　模型可视化结果对比**

**Fig. 6　Comparison of model visualization results**

表 4 给出的是表 3 中提到模型的复杂度和速度，其中，使用计算量和参数量衡量模型的复杂度，速度指标计算的是模型在训练过程中的每迭代一个批次的数据所花费的时间，将其作为间接反映模型推理速度的指标。由此可以看到 TSN 在速度和复杂度方面都位于第一，但是由表 3 可知，其 Top-1 准确率较低；TSM 在保持较低复杂度的情况下，将 Top-1 准确率提升至 49.52%，而速度方面降低至 0.689 3 s/iter；本研究所提出模型在 ResNet50 的基础上加入 2 个增强模块和融合模块，Top-1 准确率提升至 74.29%，与 SlowOnly 相比，计算量增加 0.36 G，参数量增加 1.61 M，速度降低了 0.190 6 s。综合来看本研究模型在增加有限计算量和参数量的同时，在训练速度可接受的范围内大幅度提升了模型的检测准确率。

**表 4　不同模型的复杂度与速度对比**

**Table 4　Comparison of complexity and speed among different models**

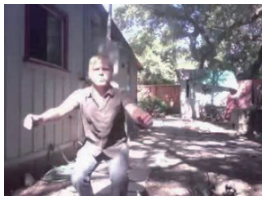
| 模型 | FLOPs/$10^9$ | Params/$10^6$ | Speed/ (s/iter) |
|---|---|---|---|
| TSN | 43.05 | 23.57 | 0.447 7 |
| TSM | 43.05 | 23.57 | 0.689 3 |
| SlowOnly | 54.86 | 31.70 | 0.632 3 |
| 本研究 | 55.22 | 33.31 | 0.822 9 |

## 2.6　可视化结果及分析

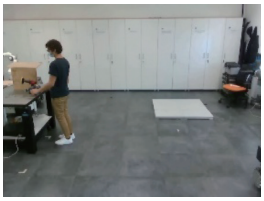
将本研究模型、SlowOnly、TSN 以及 TSM 在不同数据集上的识别结果进行可视化展示，见表 5，时间序列热力图，如图 7。

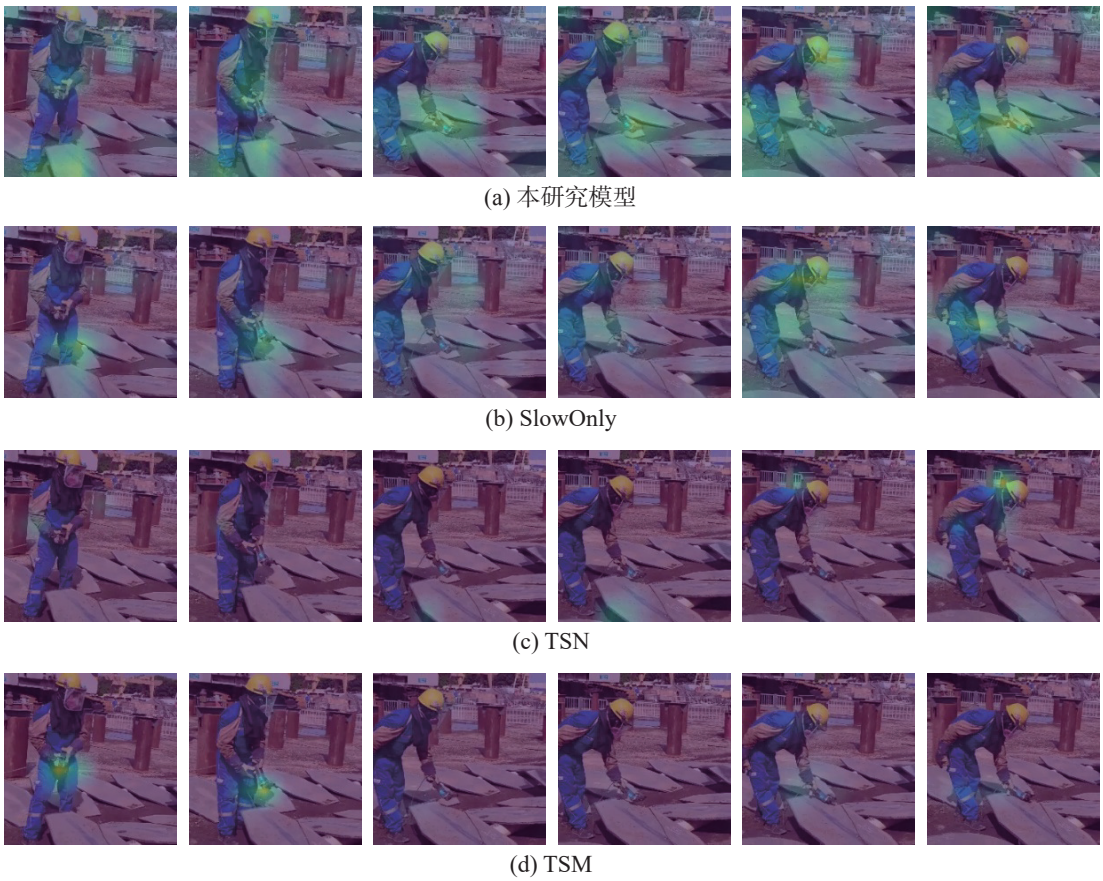
**表 5　不同数据集上各模型预测结果对比**

**Table 5　Comparison of prediction results of models on different datasets**

| 数据集 | 输入视频 | 标签 | 不同模型预测结果 | |
|---|---|---|---|---|
| UCF101 |  | 高低杠（UnevenBars） | 本研究模型 | 高低杆（UnevenBars） |
| | | | SlowOnly | 跳跃（PommelHorse） |
| | | | TSN | 双杠（ParallelBars） |
| | | | TSM | 双杠（ParallelBars） |
| |  | 深蹲（BodyWeightSquats） | 本研究模型 | 深蹲（BodyWeightSquats） |
| | | | SlowOnly | 开合跳（JumpingJack） |
| | | | TSN | 杂耍抛球（JugglingBalls） |
| | | | TSM | 开合跳（JumpingJack） |
| HMDB51 |  | 跳跃（Jump） | 本研究模型 | 跳跃（Jump） |
| | | | SlowOnly | 接住（Catch） |
| | | | TSN | 翻滚（Somersault） |
| | | | TSM | 翻滚（Somersault） |
| |  | 剑术练习（Sword_Exercise） | 本研究模型 | 剑术练习（Sword_Exercise） |
| | | | SlowOnly | 拔剑（Draw_Sword） |
| | | | TSN | 拔剑（Draw_Sword） |
| | | | TSM | 拔剑（Draw_Sword） |

续表 5

| 数据集 | 输入视频 | 标签 | 不同模型预测结果 | |
|---|---|---|---|---|
| 油田自制数据集 |  | 拿起钻头（PickUpDrill） | 本研究模型 | 拿起钻头（PickUpDrill） |
| | | | SlowOnly | 放下钻头（PutDownDrill） |
| | | | TSN | 拿起喷枪（PickUpBurner） |
| | | | TSM | 拿起喷枪（PickUpBurner） |
| |  | 放下喷枪（PutDownBurner） | 本研究模型 | 放下喷枪（PutDownBurner） |
| | | | SlowOnly | 放下钻头（PutDownDrill） |
| | | | TSN | 拿起钻头（PickUpDrill） |
| | | | TSM | 放下钻头（PutDownDrill） |



(a) 本研究模型

(b) SlowOnly

(c) TSN

(d) TSM

**图 7　油田自制数据集上的模型可视化结果对比**

**Fig. 7　Comparison of model visualization results on a self-made dataset in the oilfield**

由表 5 可知, 在所列场景中本研究模型均能正确识别, 而其他模型均出现交互对象以及肢体动作的错误识别。交互对象的识别方面, SlowOnly 将高低杠识别为跳马、将剑术练习识别为拔剑、喷枪与钻头错误识别, TSN 与 TSM 将高低杠识别为双杠、将剑术练习识别为拔剑、钻头与喷枪错误识别; 肢体动作的识别方面, SlowOnly 将深蹲识别为开合跳、将跳跃识别为接住、拿起与放下错误识别, TSN 将深蹲识别为杂耍抛球、将跳跃识别为翻滚、将放下识别为拿起, TSM 将深蹲识别为开合跳、将跳跃识别为翻滚。对比以上结果, SlowOnly、TSN 与 TSM 在交互对象的识别方面均存在不足; 肢体动作的识别方面, TSN 和 SlowOnly 均出现了深蹲、跳跃、拿起、放下的错误识别, TSM 通过加入时间位移操作, 使模型有效利用时序信息, 能正确识别拿起和放下, 其效果优于 TSN 和 SlowOnly, 但对于深蹲和跳跃, 仍然无法正确识别。

图 7 给出的是油田自制数据集上的时间序列热力图可视化。可以观察到本研究模型对于人体

动作及其交互对象的整体关注程度较高，中间两帧更关注交互对象，后 4 帧因为涉及到“放下”而有着更高的关注度；而 SlowOnly 对于行为全局的关注度则不如本研究模型，且缺少对于交互对象的关注；TSN 对于最后两帧较为关注，对于行为前期的关注程度较低；TSM 对于前两帧中的交互对象较为关注，对于行为后期的关注程度较低。

通过以上可视化展示，证明本研究模型行为识别任务的有效性，能更关注人体动作及其交互对象，同时能够在行为过程中更加关注关键动作帧，通过时空特征与动作特征的有效融合完成识别任务。

### 2.7　消融实验

为了验证本研究模型中各模块的有效性以及网络深度对模型效果的影响，在 UCF101 数据集、HMDB51 数据集以及油田自制数据集上进行消融实验，结果见表 6~8，表 6 中×和√分别代表网络中未使用和使用该模块。各模型在油田自制数据集上训练时的准确率曲线和损失曲线如图 8 所示。

**表 6　不同模块的实验结果**

**Table 6　Experimental results of different modules**　　%

| 时空注意力模块 | 动作强化模块 | 自适应特征融合 | SE模块 | UCF101 | | HMDB51 | | 油田自制数据集 | |
|---|---|---|---|---|---|---|---|---|---|
| | | | | Top-1 | Top-5 | Top-1 | Top-5 | Top-1 | Top-5 |
| × | × | × | × | 92.78 | 99.42 | 65.95 | 91.05 | 30.24 | 86.43 |
| √ | × | × | × | 95.51 | 99.58 | 72.48 | 94.38 | 51.55 | 99.17 |
| × | √ | × | × | 95.48 | 99.39 | 71.90 | 93.14 | 43.93 | 97.62 |
| √ | √ | × | × | 95.64 | 99.61 | 72.70 | 93.90 | 67.38 | 99.29 |
| √ | √ | × | √ | 96.04 | 99.65 | 73.59 | 94.12 | 71.19 | 100.00 |
| √ | √ | √ | × | **9**6.11 | 99.68 | 74.51 | 93.99 | 74.29 | 99.88 |

**表 7　不同网络深度的实验结果**

**Table 7　Experimental results with different network depths**　　%

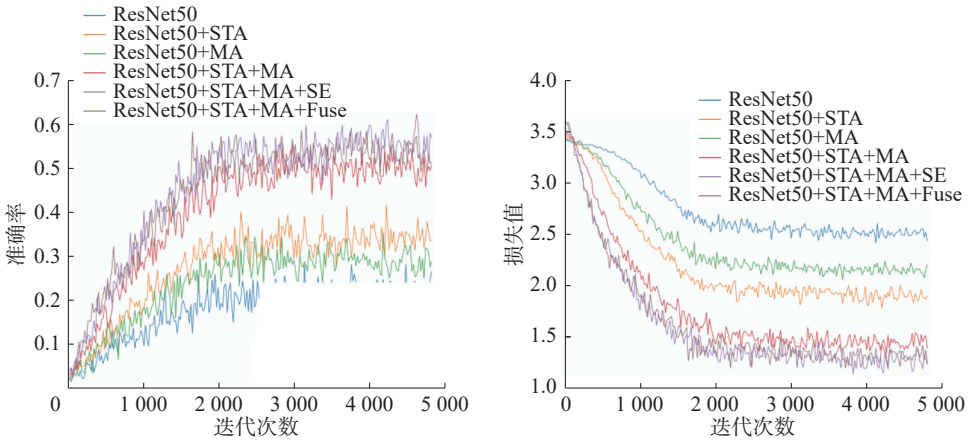
| 网络深度 | UCF101 | | HMDB51 | | 油田自制数据集 | |
|---|---|---|---|---|---|---|
| | Top-1 | Top-5 | Top-1 | Top-5 | Top-1 | Top-5 |
| ResNet34 | 51.39 | 77.72 | 23.14 | 55.16 | 38.10 | 94.76 |
| ResNet50 | 96.11 | 99.68 | 74.51 | 93.99 | 74.29 | 99.88 |
| ResNet101 | 96.54 | 99.84 | 77.84 | 95.16 | 77.57 | 99.76 |

**表 8　不同网络深度的模型复杂度**

**Table 8　Model complexity at different network depths**

| 网络深度 | FLOPs/$10^9$ | Params/$10^6$ |
|---|---|---|
| ResNet34 | 79.51 | 60.95 |
| ResNet50 | 55.22 | 33.31 |
| ResNet101 | 112.43 | 61.21 |

由表 6 可知，相较于未使用任何模块，仅使用时空注意力模块或动作强化模块时准确率均有提升，当 2 个模块采用直接相加进行特征融合时准确率进一步提升，当使用 SE 模块进行特征融合时准确率高于直接相加的融合方式，当采用自适应特征融合时模型在数据集上的准确率达到最高。观察到在 2 个公共数据集上，时空注意力模块与动作强化模块同时使用与单独使用一个模块相比，虽然准确率均有提升，但是当未采用自适应特征融合时准确率提升幅度较小，使用 SE 模块进行特征融合的效果虽然优于直接相加的融合方式，但是当采用自适应特征融合时，模型通过自适应地调整时空特征与动作特征融合方式使得准确率得到有效提升，达到最高水平。而在场景更为复杂的油田自制数据集上，当仅使用时空注意力模块时准确率提升 21.31%，当仅使用动作强化模块时准确率提升 13.69%，该结果表明动作特征较时空特征更难以提取，当同时使用 2 个模块且未采用自适应特征融合时准确率达到 67.38%，当同时使用 2 个模块且采用 SE 模块进行特征融合时准确率为 71.19%，当同时使用 2 个模块且采用自适应特征融合时准确率达到最高水平 74.29%。该结果表明时空特征与动作特征二者缺一不可，即便简单相加进行融合带来的提升也是巨大的，而融合方式也是决定模型性能的关键因素，虽然 SE 模块通过通道注意力给模型带来性能提升，但是自适应特征融合模块将时空特征、动作特征、直接相加融合后的特征以及特征各自的权重系数这几个因素相关联，自发地寻找特征之间的关系以达到自适应特征融合的目的，使模型达到最高水平。油田自制数据集相较于公共数据集，其场景更复杂，干扰因素更多，更能检验模型的场景适应性与信息提取能力，通过在公共数据集与油田自制数据集上的消融实验有效验证了本研究模型中各模块的有效性。




(a) 准确率曲线对比　　(b) 损失曲线对比

**图 8　模型训练曲线对比**

**Fig. 8　Comparison chart of model training curve**

由表 7 和表 8 可知, 使用 ResNet34 作为骨干网络时各个数据集上的准确率均为最低; 而使用 ResNet101 作为骨干网络时各个数据集上的准确率均为最高, 而二者参数量处于同一水平, 分别为 $60.95\times10^6$ 和 $61.21\times10^6$。这是因为 ResNet34 的残差块均为 3×3 卷积, 而 ResNet101 的残差块为 1×1 卷积和 3×3 卷积组成, 使得参数可控且能进一步增加网络深度。由此可见, 网络过浅即便提升参数量对于性能提升也不大, 增加网络深度可以提取深层次特征从而带来性能的提升。使用 ResNet50 作为骨干网络, 在复杂度远低于以 ResNet34 作为骨干网络的同时, 实现了性能的大幅度提升, 在 UCF101 数据集上的准确率为 96.11%, 在 HMDB51 数据集上的准确率为 74.51%, 在油田自制数据集上的准确率为 74.29%, 虽然相较于以 ResNet101 作为骨干网络时准确率分别低了 0.43%、3.33% 和 3.28%, 但是计算量仅为后者的 49.12%, 参数量仅为后者的 54.42%。由此可见, 增加网络深度在加大模型复杂度的同时并不能总是带来可观的性能提升, 选择合适的网络深度可以使得模型在复杂度较低的同时保持良好的性能。通过对比分析模型在各数据集上的性能以及复杂度, 验证了本研究使用 ResNet50 作为骨干网络的合理性。

## 3　结束语

本研究提出了一种基于时空−动作自适应融合的行为识别网络, 通过显式地引入时空注意力模块和动作强化模块同时自适应地融合特征, 用于油田作业现场的行为识别。该网络可以端到端地进行训练和测试, 不需要提前提取光流等先验信息, 同时分别在公共数据集 UCF101 和 HMDB51 上进行对比实验, 验证了本研究算法的有效性和通用性, 在油田自制数据集上进行对比实验, 验证了本研究算法的高效性和适应性。通过一系列实验, 表明本研究算法具有较强的场景适应性、较高的识别准确率, 更加适用于作业现场的复杂场景, 具有实际应用价值。

## 参考文献:

[1] 富倩. 人体行为识别研究[J]. 信息与电脑 (理论版), 2017(24): 146−147.
FU Qian. Analysis of human behavior recognition[J]. China computer & communication(theoretical edition), 2017(24): 146−147.

[2] 梁绪, 李文新, 张航宁. 人体行为识别方法研究综述[J]. 计算机应用研究, 2022, 39(3): 651−660.
LIANG Xu, LI Wenxin, ZHANG Hangning. Review of research on human action recognition methods[J]. Application research of computers, 2022, 39(3): 651−660.

[3] TRAN D, BOURDEV L, FERGUS R, et al. Learning spatiotemporal features with 3D convolutional networks[C]// 2015 IEEE International Conference on Computer Vision. Santiago: IEEE, 2015: 4489-4497.

[4] 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. Honolulu: IEEE, 2017: 4724−4733.

[5] TRAN D, WANG Heng, TORRESANI L, et al. A closer look at spatiotemporal convolutions for action recognition[C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018: 6450−6459.

[6] SIMONYAN K, ZISSERMAN A. Two-stream convolutional networks for action recognition in videos[C]//Proceedings of the 27th International Conference on Neural Information Processing Systems. Montreal: MIT Press, 2014: 568−576.

[7] WANG Limin, XIONG Yuanjun, WANG Zhe, et al. Temporal segment networks: towards good practices for

deep action recognition[M]//Lecture Notes in Computer Science. Cham: Springer International Publishing, 2016: 20−36.

[8] LIN Ji, GAN Chuang, HAN Song. TSM: temporal shift module for efficient video understanding[C]//2019 IEEE/CVF International Conference on Computer Vision. Seoul: IEEE, 2019: 7082−7092.

[9] FEICHTENHOFER C, FAN Haoqi, MALIK J, et al. SlowFast networks for video recognition[C]//2019 IEEE/CVF International Conference on Computer Vision. Seoul: IEEE, 2019: 6201−6210.

[10] YAN Sijie, XIONG Yuanjun, LIN Dahua. Spatial temporal graph convolutional networks for skeleton-based action recognition[C]//Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence. New Orleans: AAAI, 2018: 7444−7452.

[11] DUAN Haodong, ZHAO Yue, CHEN Kai, et al. Revisiting skeleton-based action recognition[C]//2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans: IEEE, 2022: 2959−2968.

[12] 田枫, 孙晓悦, 刘芳, 等. 基于图卷积的作业行为实时检测方法[J]. 计算机工程与设计, 2022, 43(10): 2944−2952.
TIAN Feng, SUN Xiaoyue, LIU Fang, et al. Real time detection method of work behavior based on graph convolution[J]. Computer engineering and design, 2022, 43(10): 2944−2952.

[13] 陆昱翔, 徐冠华, 唐波. 基于视觉 Transformer 时空自注意力的工人行为识别[J]. 浙江大学学报 (工学版), 2023, 57(3): 446−454.
LU Yuxiang, XU Guanhua, TANG Bo. Worker behavior recognition based on temporal and spatial self-attention of vision Transformer[J]. Journal of Zhejiang university (engineering science edition), 2023, 57(3): 446−454.

[14] 饶天荣, 潘涛, 徐会军. 基于交叉注意力机制的煤矿井下不安全行为识别[J]. 工矿自动化, 2022, 48(10): 48−54.
RAO Tianrong, PAN Tao, XU Huijun. Unsafe action recognition in underground coal mine based on cross-attention mechanism[J]. Industry and mine automation, 2022, 48(10): 48−54.

[15] DOSOVITSKIY A, BEYER L, KOLESNIKOV A, et al. An image is worth 16x16 words: transformers for image recognition at scale[EB/OL]. (2020−10−22)[2022−03−24]. http://arxiv.org/abs/2010.11929.

[16] HE Kaiming, ZHANG Xiangyu, REN Shaoqing, et al. Deep residual learning for image recognition[C]//2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas: IEEE, 2016: 770−778.

[17] HU Jie, SHEN Li, SUN Gang. Squeeze-and-excitation networks[C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018: 7132−7141.

[18] 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: IEEE, 2011: 2556−2563.

[19] SOOMRO K, ZAMIR A R, SHAH M. UCF101: a dataset of 101 human actions classes from videos in the wild [EB/OL]. (2012−12−03)[2021−11−05]. http://arxiv.org/abs/1212.0402.

[20] KAY W, CARREIRA J, SIMONYAN K, et al. The kinetics human action video dataset[EB/OL]. (2017−05−19)[2021−11−05]. http://arxiv.org/abs/1705.06950.

[21] FISHER R A. The use of multiple measurements in taxonomic problems[J]. Annals of eugenics, 1936, 7(2): 179−188.

[22] GAMMULLE H, DENMAN S, SRIDHARAN S, et al. Two stream LSTM: a deep fusion framework for human action recognition[C]//2017 IEEE Winter Conference on Applications of Computer Vision. Santa Rosa: IEEE, 2017: 177−186.

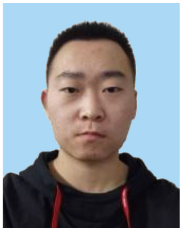
[23] JI Shuiwang, XU Wei, YANG Ming, et al. 3D convolutional neural networks for human action recognition[J]. IEEE transactions on pattern analysis and machine intelligence, 2013, 35(1): 221−231.

[24] XIE Saining, GIRSHICK R, DOLLÁR P, et al. Aggregated residual transformations for deep neural networks[C]//2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu: IEEE, 2017: 5987−5995.

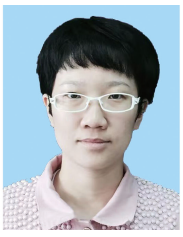
[25] SELVARAJU R R, COGSWELL M, DAS A, et al. Grad-CAM: visual explanations from deep networks via gradient-based localization[J]. International journal of computer vision, 2020, 128(2): 336−359.

## 作者简介：

田枫，教授，博士生导师，博士，计算机与信息技术学院院长，主要研究方向为智能油气地质、计算机视觉、智能数据分析处理。主持和参与国家自然科学基金项目、国家科技重大专项项目 8 项，专利授权 16 项，发表学术论文 30 余篇。E-mail：tianfeng1980@163.com。



卫宁彬，硕士研究生，主要研究方向为计算机视觉、智能数据分析处理。E-mail：1205542631@qq.com。



刘芳，副教授，博士，主要研究方向为智能油气地质、智慧教育、多媒体与现代教育技术、计算机视觉。获黑龙江省科技进步二等奖 1 项、大庆市科技进步二等奖 1 项，主持和参与国家自然科学基金项目、黑龙江省自然科学基金项目 6 项，发表学术论文 20 余篇。E-mail：lfliufang1983@126.com。