基于多重信息自注意力的综采工作面目标行为识别

杨　艺[1,2,3]，杨艳磊[1,2]，王　田[4]，王科平[1,2]

(1. 河南理工大学 电气工程与自动化学院，河南 焦作　454003；2. 河南理工大学 河南省煤矿装备智能检测与控制重点实验室，河南 焦作　454003；3. 郑州煤矿机械集团股份有限公司，河南 郑州　450000；4. 北京航空航天大学 人工智能研究院，北京　100191)

摘　要：综采工作面关键设备和人员的行为识别是开采环境信息智能感知的基础和核心。然而，综采工作面光照条件普遍较差，煤尘和水雾等干扰容易引起视频画面模糊，导致识别目标行为的关键特征难以提取，使得设备和人员的行为识别准确度无法达到实际工程应用的标准。为此，基于 ResT 网络架构，建立一种包含空间、时间、通道的多重信息自注意力模型和特征融合机制，扩展了模型特征提取的信息源，将其从单纯的空间信息扩展到空间、时间和通道的多重信息，提升了模型对目标行为的表征能力。其中，空间信息是对目标行为在空间上的深度解析，展现了目标的纹理、位置和形状等一系列深层特征；时间信息是从连续的视频帧中提取目标行为的时序特征，反映了行为发生的顺序以及演变关系；通道信息则是对空间和时间层面上的扩展与深入，从多角度挖掘空间和时间信息，并将原始数据表征在特征通道上，提供了目标行为的全局特征。算法的有效性在综采工作面行为识别数据集上进行了验证和对比试验。结果表明：在真实综采工作面环境下，行为识别的准确度可达到 96.90%。相较于 Swin-Transformer、Timesformer 等主流的行为识别算法，识别准确率分别提升了 11.06% 和 10.62%。算法经过 ONNX 模型转换和 TensorRT 加速后，在 GPU 上实现了推理，具备工程应用价值。据此，研发了综采工作面行为识别系统，并将算法模型以插件的形式嵌入到行为识别系统的 Pipeline 中，实现在 DeepStream 框架下对综采工作面关键设备和人员行为的实时推理和准确识别。

关键词：工作面；行为识别；空间−时间−通道信息；网络模型；工程部署

中图分类号：TD67　　**文献标志码：**A　　**文章编号：**0253−9993(2025)02−1421−18

Target behavior recognition of fully mechanized mining face based on multi-information self-attention

YANG Yi[1,2,3], YANG Yanlei[1,2], WANG Tian[4], WANG Keping[1,2]

(1. *School of Electrical Engineering and Automation, Henan Polytechnic University, Jiaozuo*　454003, *China*; 2. *Henan Key Laboratory of Intelligent Inspection and Control of Coal Mine Equipment, Henan Polytechnic University, Jiaozuo*　454003, *China*; 3. *Zhengzhou Coal Mining Machinery Group, Zhengzhou*　450000, *China*; 4. *Institute of Artificial Intelligence, Beihang University, Beijing*　100191, *China*)

Abstract: Recognizing the behavior of key equipment and personnel in fully mechanized mining faces is the foundation and core of intelligent sensing of mining environment information. However, the lighting conditions in fully mechanized mining faces are generally poor. coal dust and water mist can easily cause blurring of the video image, making it difficult

收稿日期：2024−05−11　　**策划编辑：**郭晓炜　　**责任编辑：**宫在芹　　**DOI：**10.13225/j.cnki.jccs.2024.0056

基金项目：国家自然科学基金资助项目 (61972016)；中国国家铁路集团有限公司资助项目 (N2023X005)；河南省科技攻关资助项目 (232102210040)

作者简介：杨　艺 (1980—)，男，湖北利川人，副教授，博士 (后)，硕士生导师，E-mail：yangyi@hpu.edu.cn

通讯作者：杨艳磊 (2000—)，男，河南周口人，硕士研究生，E-mail：yangyanlei@home.hpu.edu.cn

引用格式：杨艺，杨艳磊，王田，等. 基于多重信息自注意力的综采工作面目标行为识别[J]. 煤炭学报，2025，50(2)：1421−1438.

YANG Yi, YANG Yanlei, WANG Tian, et al. Target behavior recognition of fully mechanized mining face based on multi-information self-attention[J]. Journal of China Coal Society, 2025, 50(2): 1421−1438.

移动阅读

to extract key features for identifying target behaviors. As a result, this affects the accuracy of identifying the behavior of equipment and personnel, failing to meet the requirements for practical engineering purposes. In order to address this problem, a multi-information self-attention model and feature fusion mechanism have been developed based on the ResT network architecture. This model expands the information source for feature extraction from pure spatial information to multi-information, including space, time, and channel. This enhancement improves the model's capability to recognize the target's behavior. Among the aforementioned categories, spatial information is a detailed spatial analysis of the target behavior, showcasing a range of deep features such as texture, location, and shape of the target. Temporal information refers to extracting temporal features of the target behavior from continuous video frames, reflecting the order of occurrence of the behavior as well as the evolutionary relationship. Channel information represents the expansion and depth of spatial and temporal levels by extracting spatial and temporal information from multiple perspectives. It characterizes the raw data on feature channels, which provide global features of the target behavior. The effectiveness of our algorithm has been validated through comparative experiments on the dataset for behavior recognition in fully mechanized mining faces. The experimental results demonstrate that the accuracy of behavior recognition can reach 96.90% in the fully mechanized mining faces environment. In comparison with mainstream behavior recognition algorithms such as Swin-Transformer and Timesformer, the recognition accuracy is enhanced by 11.06% and 10.62% respectively. The algorithm is transformed by an ONNX model and accelerated using TensorRT to enable GPU inference, thereby enhancing its value for engineering applications. Consequently, the fully mechanized mining face behavior recognition system was developed. The algorithm model was embedded into the pipeline of the behavior recognition system as a plug-in unit. This integration enables real-time analysis and accurate recognition of crucial equipment and personnel behaviors on the fully mechanized mining faces within the DeepStream framework.

Key words: working face; behavior recognition; spatial-temporal-channel information; network model; project deployment

0 引 言

煤炭作为我国最重要的一次能源，是确保我国经济社会高速发展的关键因素之一[1-2]。目前，我国绝大多数煤矿还是以井工开采为主，但井下工作环境恶劣，难以吸引优质人力资源。而且，部分矿区瓦斯、水害、高应力等灾害较为明显[3]。虽然经过多年治理，上述灾害引起的重大事故已经大幅减少，但是仍然难以消除普通人对井下安全的担忧，这使得煤矿人力资源现状更加窘迫。因此，推进煤矿智能化开采，实现煤矿少人化、甚至是无人化开采，是解决煤矿人力资源困境的有效方法[4]。2020 年，国家发改委、国家能源局等八部委联合下发《关于加快煤矿智能化发展的指导意见》[5]，其中明确指出智能化是煤炭行业发展的核心。2023 年，煤炭“十四五”规划再次指出要加速实现煤炭智能化改造。

计算机视觉是工作面信息智能感知和决策的关键技术，也是实现煤炭智能化的重要支撑[6]。目前，计算机视觉在煤矿的应用主要集中在视频目标检测与跟踪[7-12]，但这仅是对视觉信息的初步感知。而视频行为识别是在目标检测基础上更高级的视觉信息感知。它通过捕捉帧间信息获取目标的动态特征，从而实现行为识别，进而可以确定关键设备和人员的趋势或意图。这为工作面开采工艺的智能决策、设备智能控制、故障诊断及事故预警等方面提供了强有力的信息支撑。

然而，针对煤矿视频目标动作识别的研究绝大部分集中在巷道、车场等条件较好的环境中，对目标的行为识别可以取得较好效果。而工作面光照条件差，普遍存在光照不均匀、光比过大、饱和度低等问题，这使得同一对象的关键特征在同一工作面不同的光照条件下差异较大，导致表征对象行为的关键特征失效[13]。此外，回采过程产生大量煤尘、水雾等干扰，使得成像结果存在大量的模糊现象，从而影响了对目标行为关键特征的提取，导致行为识别失败。为此，笔者针对工作面视频目标行为识别存在的模糊、光照等引起的识别准确度不高的问题，在前期工作基础上提出一种多重信息提取模型 STC-ResT(Spatial-Temporal-Channel ResT)。多重信息是指从单帧图像、连续多帧以及信息分解 3 个层面对目标动作行为信息的表征。其中，目标在单帧图像上的形体特征和空间位置表征为空间信息；目标在连续多帧中形态和位置的变化过程以及演变规律表征为时间信息；将空间和时间信息分解到多个通道，以不同的角度观察到的目标特性表

征为通道信息。

STC-ResT 模型从空间−时间−通道等多个方面提取工作面目标的行为特征, 降低了模型对空间特征的依赖性。并利用高效多头自注意力机制, 增强视频帧间和不同通道间的关联度, 从而解决因视频模糊导致行为识别失效的问题。最后, 将 STC-ResT 网络模型部署到综采工作面行为识别系统, 引入 TensorRT 对模型的推理实现优化, 在 DeepStream 框架下实现视频目标的行为识别。

1　行为识别相关研究

关于行为识别的研究最早可以追溯到 20 世纪 70 年代, JOHANSSON[14]率先通过实验提出以人体若干关节点的运动来描述行为的产生, 并为之后行为识别研究[15-16]奠定了基础。到 20 世纪 90 年代, 行为识别进入快速发展时期, 产生了大量的行为识别方法, 主要可以分为 3 类: 基于时空体积特征[17-19]、基于 STIP(Space-Time Interest Points) 特征[20-22]和基于轨迹特征[23, 24]。

2012 年, 深度学习点燃人工智能的新一轮研究热潮[25], 卷积神经网络[26]CNN(Convolutional Neural Network) 引领人工智能各个研究领域, 并迅速在行为识别领域得到应用。

2014 年, SIMONYAN[27]首次将卷积运算和光流法融合, 分别提取运动目标的空间特征和时间特征, 从而大幅提高行为识别的准确度。在后续的近 10 a 间, 研究人员尝试在时间维度对视频帧进行叠加从而将特征组合成立方体结构, 然后使用 3D 卷积核对其进行特征提取, 于是 3D 卷积神经网络[28-34]应运而生。

2017 年, 以自注意力为核心的 Transformer[35]网络架构被提出, 并成功应用于自然语言处理领域[36-38]。2018 年, Transformer[39]被首次引入计算机视觉领域。由于其可以直接关注全局信息, 几乎在所有计算机视觉任务中, Transformer 的处理效果都大幅超过 CNN, 并成为大模型的核心和基础, 如 ChatGPT[40]、文心一言[41]、盘古大模型[42]等。同时, 也为行为识别领域[43-48]的研究带来重大进展。

Transformer 作为目前最主流的网络模型, 被广泛应用于行为识别[49-51]、目标检测[52, 53]、图像分类[54, 55]等视觉任务中。但面向煤矿井下行为识别的研究起步相对较晚, 主要工作包括: 2016 年, 胡杨杨[56]针对煤矿井下巷道等环境, 提出一种实现人体运动区域提取的融合算法, 并使用支持向量机对行为进行分类。同年, 杨超宇[57]等基于 Spring MVC(Spring Model View Controller) 框架实现了煤矿行为识别系统。2019 年, 陈庆峰[58]采用 Kinect 传感器提取矿工关节点进行行为识别, 并使用 SVM(Support Vector Machines) 对不安全行为进行建模。同年, 杨赛烽[59]针对罐笼内矿工的不安全行为提出一种基于决策级融合的矿工交互行为识别方法。随着深度神经网络的兴起[60], 煤矿井下行为识别研究取得了较大的突破。2020 年, 党伟超等[61]将时间和空间的特征进行加权融合, 提出基于改进双流法的井下配电室行为识别。同年, 温廷新等[62]提出迁移学习结合深度残差网络的行为识别方法。2021 年, 黄瀚等[63]提出基于 DA-GCN(Domain-aware Attentive Graph Convolution Network) 的煤矿人员行为识别, 通过动态注意力和多层感知图卷积对人体行为进行识别。同年, 刘浩等[64]融合多种行为识别模型实现对井下人员不安全行为的智能识别系统。2022 年, 饶天荣等[65]提出一种基于交叉注意力的多特征融合行为识别模型, 可以有效的融合图像和人体关键点的特征, 从而提高识别准确度。2023 年, 张雷等[66]针对特征提取手段单一的问题, 提出了基于融合网络的井下人员行为识别方法。

但是针对工作面的行为识别研究十分稀少。2016 年, 岳志奇[67]通过行为抽样法实测综采工作面员工不安全行为数据, 得出综采工作面各个主要类型的不安全行为发生率, 并在理论上提出了分析综采工作面不安全行为形成机理、影响因素的模型, 并通过实验加以验证。2019 年, 杨峰等[68]提出了一种基于视觉关系检测的综采工作面不安全行为识别方法, 通过构建不安全行为数据库再由目标识别算法进行识别, 最后与数据库不安全行为进行对比。

课题组自 2020 年起将 Transformer 应用于工作面人员和关键设备的行为识别, 先后建立了 Transformer 和 Swin Transformer 模型[48]来自动提取空间特征, 并在此基础上引入限制对比度自适应直方图均衡算法, 有效解决了工作面光照条件较差的问题。然而, 对煤尘、水雾引起的视频模糊进而导致行为识别准确度不高的问题, 并未得到很好解决。基于空间特征的 Transformer 方法本质上是在单帧图像上提取空间特征。若图像模糊, 则容易引起特征提取失效, 导致行为识别失败。2021 年, ResT(Res-Transformer) 模型[69]在 Transformer 基础上提出高效多头自注意力机制, 一定程度上提升了识别的准确性, 但对解决因模糊引起行为识别失败的问题, 效果并不明显。

以 Transformer 为基础架构的模型是以注意力的方式在空间维度上做特征提取, 并不直接考虑视频帧间的关联度。为此, 本文在课题组前期研究的基础上, 以 ResT 为基础网络, 构建 STC-ResT (Spatial-Tempor-

al-Channel ResT) 多维特征提取模型，将特征信息扩展到空间、时间和通道维度，以提升模型对特征的提取能力，提升目标行为识别的准确度。

2 ResT 空间注意力运行机理

2021 年，ZHANG 等[69]在传统 Transformer 模型的基础上，以 ResNet[70]作为信息处理框架，提出一种具备高效提取的空间注意力网络模型框架 ResT，如图 1 所示。ResT 以空间注意力为基础，在 Transformer 基础上引入深度卷积操作，较好的解决了 Transformer 模型在计算量以及推理开销较大的问题。

2.1 Transformer 的自注意力机制

Transformer 的自注意力是以处理序列为主的机制，可以根据输入序列中不同位置的相关性自动分配注意力权重。将 Transformer 应用于计算机视觉时，首先将图像分块，转化成序列，再计算其自注意力值。而多头自注意力值的计算量主要跟输入序列的长度和隐藏层的维度有关。对于单个注意力计算头来说注意力的计算量 ∂ 如式 (1) 所示。

$$\partial(MSA) = N^2\frac{d}{m} = (h\times w)^2\frac{d}{m} \tag{1}$$

式中：输入图像分块后序列长度 N 为 $h\times w$；h 为图像块的高；w 为图像块的宽；d 为隐藏层的维度；m 为注意力头数。

2.2 ResT 的自注意力机制

ResT 的空间多头自注意力机制 SMSA(Spatial Multi-Head Self-Attention) 是在 Transformer 多头自注意力机制的基础上增加了深度卷积操作，在扩大感受野的同时降低计算量。

ResT 注意力的计算量 F 如式 2 所示。

$$\partial(SMSA) = N'^2\frac{d}{m} = \frac{1}{81}\left((h\times w)^2\frac{d}{m}\right) \tag{2}$$

式中：N' 为经过深度卷积之后的输入序列长度；d 和 m 依旧为隐藏层的维度和注意力头数。

在输入、隐藏层的维度以及注意力头数都不变的情况下，设 ResT 的卷积核形状为 3×3。经过深度卷积后其输入序列长度 N' 变为 $\frac{h\times w}{9}$，而根据式 (1) 中定义的输入图像分块后的序列长度 N 为 $h\times w$。对比式 (1) 和 (2) 可知，ResT 的计算量 ∂ 降低为原计算量的 $\frac{1}{81}$。因此，ResT 采用深度卷积操作可减少输入序列的长度，从而大幅度降低模型的计算量。

此外，ResT 的卷积操作有助于扩大局部感受野，能够提取更广泛的上下文信息。感受野 RF(Receptive Field) 计算式如式 (3) 所示：

$$RF_i = RF_{i-1} + (k-1)S_{i-1} \tag{3}$$

式中：RF_i 为卷积后的感受野，RF_{i-1} 为卷积前的感受野。k 为卷积核的大小。S_{i-1} 为卷积前步长的乘积。通过对式 (3) 分析，使用深度卷积后的感受野等于深度卷积前的感受野加上一个正数。由此可以证明经过深度卷积可以增强模型的感受野。

ResT 在经过深度卷积之后不仅减少了空间自注意力的计算量，而且增加了 ResT 的感受野，使该模型更好的感知上下文之间的关系。因此 ResT 的空间自注意力在添加深度卷积之后可以获得更好的效果。

需要注意的是：ResT 的空间注意力是采用分块级的空间自注意力机制，这样虽然能减少计算开销但却舍弃了全局依赖关系，丧失了对全局信息的掌握。在煤矿井下，摄像头所处的环境通常是光线昏暗，色彩饱和度低，且尘雾干扰严重。在图像信息上表现为对

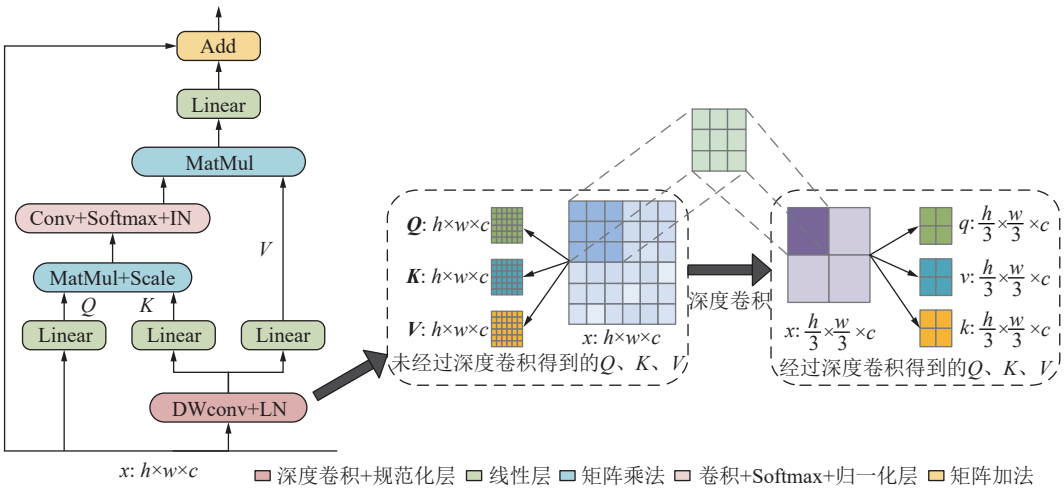


图 1 ResT 空间注意力

Fig.1 ResT spatial attention

比度低、纹理信息简单、视频模糊。ResT 以空间信息为基础提取特征, 缺乏对对比度、纹理信息的深度解析。因此, 直接采用 ResT 处理煤矿井下视频信息, 可能会造成行为识别准确度低, 甚至识别失败。

3　STC-ResT 模型

针对 ResT 的空间注意力分块化处理流程导致模型缺少对全局信息的掌握, 以及 ResT 以空间信息为基础的特征提取方式在光线昏暗、色彩饱和度低和尘雾干扰严重的综采工作面环境下缺乏对目标信息的深度解析等问题。本文提出了基于空间−时间−通道注意力机制的 STC-ResT 网络模型。如图 2 所示。

该模型的输入 (Input) 是经过剪裁、旋转、抽帧等预处理后得到的视频帧。模型主体由底层特征提取模块 BE-Block(Basics Extraction Block)、空间−时间−通道信息融合模块和 Head 模块 3 部分组成。

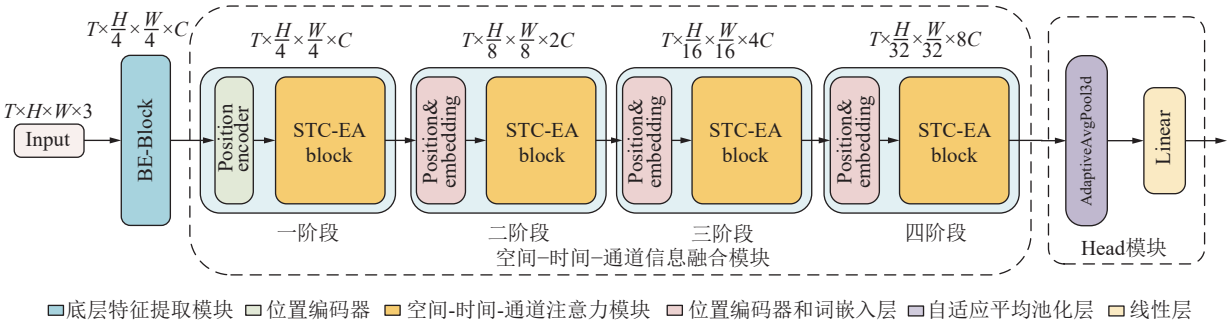


图 2　STC-ResT 模型

Fig.2　STC-ResT model

底层特征提取模块整体由 3DCNN 实现, 主要负责提取目标的轮廓、边缘、颜色等底层特征。这些底层特征虽然语义信息较少, 但是细节特征较多, 可以给 STC-ResT 提供充足的位置信息。

3.1　STC-ResT 运行机制

空间−时间−通道信息融合模块是本框架的核心内容, 包括 4 个阶段: 第 1 阶段由位置编码器 (Position Encoder) 和空间−时间−通道注意力模块 STC-EA Block(Spatial-Temporal-Channel Efficient Attention Block) 组成。后面 3 个阶段则是由位置编码器、词嵌入层 (Patch Embedding) 和 STC-EA Block 三部分组成。其中, 位置编码器是对其位置信息进行编码。Patch Embedding 是通过减小输入空间分辨率的同时增加通道数, 以获取输入的多尺度特征信息。STC-EA Block 是实现对空间、时间、通道维度特征获取的主要模块。该模块通过对这 3 类信息特征进行提取, 不仅可以减少模型对空间维度的依赖性, 使 STC-ResT 在模糊环境中依旧保持精准识别。同时, 还能提取到输入的全局特征。

STC-EA Block 中的空间注意力 SMSA(Spatial Multi-Head Self-Attention) 是以单个空间块作为计算单元, 计算目标特征的空间位置关系, 做到对目标特征的精准提取。时间注意力 TMSA(Temporal Multi-Head Self-Attention) 是以单一时间段作为计算单元, 计算当前视频段内所有位置的时序信息, 从而增强 STC-ResT 的时序建模能力。通道注意力 CMSA (Channel Multi-Head Self-Attention) 是以单个通道作为计算单元, 计算当前通道内所有位置的特征信息, 从而强化 STC-ResT 提取全局特征的能力。

Head 模块是由全局自适应池化层和线性层组成。其中, 全局自适应池化层负责将输入转化为一维序列, 再由线性层将输入映射到分类数上进行输出。

STC-ResT 对输入提取特征信息的具体流程如下。首先, 将视频帧的大小处理为统一的$T\times H\times W\times 3$来作为输入, 其中$T$表示视频帧的帧数, H和W表示视频帧的高和宽。其次, 再用底层特征提取模块 BE-Block 对输入的轮廓、边缘、颜色等底层特征进行提取, 使输入变为$T\times\frac{H}{4}\times\frac{W}{4}\times C$, 其中$C$为通道维度。然后, 空间−时间−通道信息融合模块对输入进行多尺度的特征信息提取。空间−时间−通道信息融合模块由四个阶段组成, 每经历一个阶段, 空间维度缩减为之前的 1/2, 通道维度扩展 2 倍。在经过 4 个阶段处理之后输入变为了$T\times\frac{H}{32}\times\frac{W}{32}\times 8C$。最后, Head 模块中的全局自适应池化层将输入由二维矩阵变为一维向量之后由线性层进行输出。

3.2　空间−时间−通道注意力模块

空间−时间−通道注意力模块 (STC-EA Block) 是将空间注意力模块、时间注意力模块和通道注意力模块通过残差连接进行结合, 从而达到相互补充和增强的作用。其空间−时间−通道注意力模块如图 3 所示。

空间−时间−通道注意力模块中, 空间注意力模块、

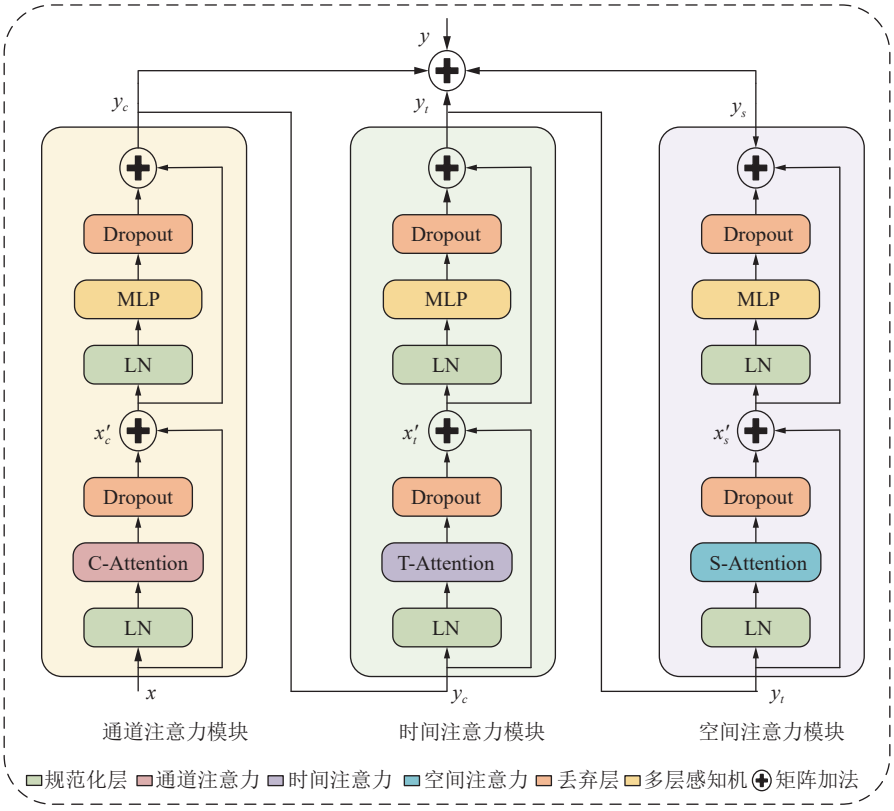


图 3 空间−时间−通道注意力模块

Fig.3 Spatial-Temporal-Channel attention module

时间注意力模块和通道注意力模块采用相同的流程架构。从下往上依次为规范化层 (Layer- Norm), 负责对输入进行规范化处理, 增加输入数据分布范围的稳定性; 自注意力层 (Attention), 负责对输入的空间、时间、通道维度进行自注意力计算以充分提取目标的特征信息; Dropout, 防止训练过程中出现过拟合; 多层感知机 MLP(Multi Layer Perceptron), 增强网络模型的泛化和建模能力。之后通过残差连接能更好的拟合网络模型以获得更高的分类准确度。同时, 还能降低后面部分对前面部分的计算结果造成拉伸、压缩或信息丢失等影响。

在自注意力层中, 空间注意力是基于图像块的局部感受野, 可以根据不同位置信息的重要性自动进行精准特征提取。时间注意力是基于帧间信息的局部感受野, 能够对不同时刻信息进行加权并剔除背景冗余, 减少计算开销。通道注意力是基于通道维度的全局感受野, 用于提取输入的全局特征并传递给后者。STC-ResT 通过空间−时间−通道注意力模块深度解析了空间、时间以及通道之间的结构和信息关系, 全面挖掘了其中蕴含的目标特征信息。空间注意力模块、时间注意力模块和通道注意力模块三者之间有序配合, 大幅增强了 STC-ResT 对图像信息的感知能力, 并降低了视频模糊对识别造成的影响。

空间−时间−通道注意力模块整体输出如式 (4) 所示。

$$y = y_s + y_t + y_c \tag{4}$$

$$y_c = x'_c + \mathrm{MLP}(LN(x'_c)) \tag{5}$$

$$x'_c = y_t + \mathrm{SMSA}(LN(y_t)) \tag{6}$$

$$y_t = x'_t + \mathrm{MLP}(LN(x'_t)) \tag{7}$$

$$x'_t = y_s + \mathrm{TMSA}(LN(y_s)) \tag{8}$$

$$y_s = x'_s + \mathrm{MLP}(LN(x'_s)) \tag{9}$$

$$x'_s = x + \mathrm{CMSA}(LN(x)) \tag{10}$$

式中: x为输入; y为输出; x'_s、x'_t、x'_c分别为经过空间注意力模块、时间注意力模块、通道注意力模块之后的输出; y_s、y_t、y_c分别为每个模块计算后的输出。下面分别详细阐述空间−时间−通道注意力模块。

3.2.1 空间注意力模块

空间注意力模型可以帮助模型更精细地识别目标不同位置的行为特征。在综采工作面行为识别中, 模型依靠空间注意力能够准确的识别出目标的空间关系, 提高识别效率。

传统的 Transformer 是以巨大的计算量换来了超

高的准确率, 并且计算量根据输入的大小呈平方倍增长, 而直接将综采工作面视频数据集输入模型会造成巨大的训练和推理开销。针对这一问题, STC-ResT 在空间注意力模块中添加了卷积操作, 不仅极大的降低了计算量与推理花销还扩宽了感受野。于是本文在空间处理模块沿用了 ResT 的时间注意力模块。其空间注意力计算式如式 (11) 所示。

$$F_{\mathrm{SMSA}}(Q',K',V')=\mathrm{IN}\left\{\mathrm{Softmax}\left[\mathrm{Conv}\left(\frac{Q'K'^{T}}{\sqrt{d_k}}\right)\right]\right\}V' \tag{11}$$

式中: Q'(Query)、K'(Key)、V'(Value) 为查询向量、键向量和值向量; d_k为词嵌入的维度。

空间−时间−通道注意力模块是从输入的空间维度、时间维度和通道维度进行充分特征提取的模块。其中, 在空间和时间注意力模块上增加卷积操作是为了扩大模型的感受野, 提升模型对特征的提取能力; 而通道注意力模块本身以全局感受野提取目标的行为特征, 故不用引入卷积操作。

3.2.2　时间注意力模块

在综采工作面场景下存在不同时间尺度的行为特征, 如工人的行走或站立为长时行为、护帮板的打开、关闭为短时行为。而时间注意力模块可以有效的整合这些多尺度的时序特征信息, 从而提高行为识别的准确性。其次, 视频行为识别主要是通过前后帧之间像素的变化信息对目标行为进行分析。而时序信息中就包含了这些变化信息的演变过程以及上下文关系。所以时间维度的变化信息对综采工作面目标的行为识别尤其重要。

时间维度相较于空间维度和通道维度来说比较小, 所以在使用多头自注意力操作时每个注意力头所分到的维度会更小。而且每个注意力头之间缺少直接联系, 如果直接进行注意力计算, 会导致模型不能发挥全部的性能甚至丧失时间维度上的连贯性。于是, 针对以上问题提出时间注意力模块 TMSA(Temporal Multi-head Self-Attention), 如图 4 所示。

该模块与传统多头自注意力 MSA(Multi-head Self-Attention) 操作类似, 通过获得Q、K、V后进行矩阵运算从而提取到时间维度上的信息。但不同的是 TMSA 通过卷积操作模拟多头之间的相互作用, 增强了多头之间的联动。其注意力函数如式 (12) 所示。

$$F_{\mathrm{TMSA}}(Q,K,V)=\mathrm{IN}\left\{\mathrm{Softmax}\left[\mathrm{Conv}\left(\frac{QK^{T}}{\sqrt{d_k}}\right)\right]\right\}V \tag{12}$$

式中: Q、K、V为查询向量、键向量和值向量, d_k为词嵌入维度的大小。Conv()是一个1×1的卷积操作, 经过卷积可以在不破坏其原本维度的同时增强每个注意力头之间的交互性, 从而让每个注意力函数都依赖于所有的Q、K、V向量。最后, 为了保证模型的多样性, 加入了实例归一化层IN()。

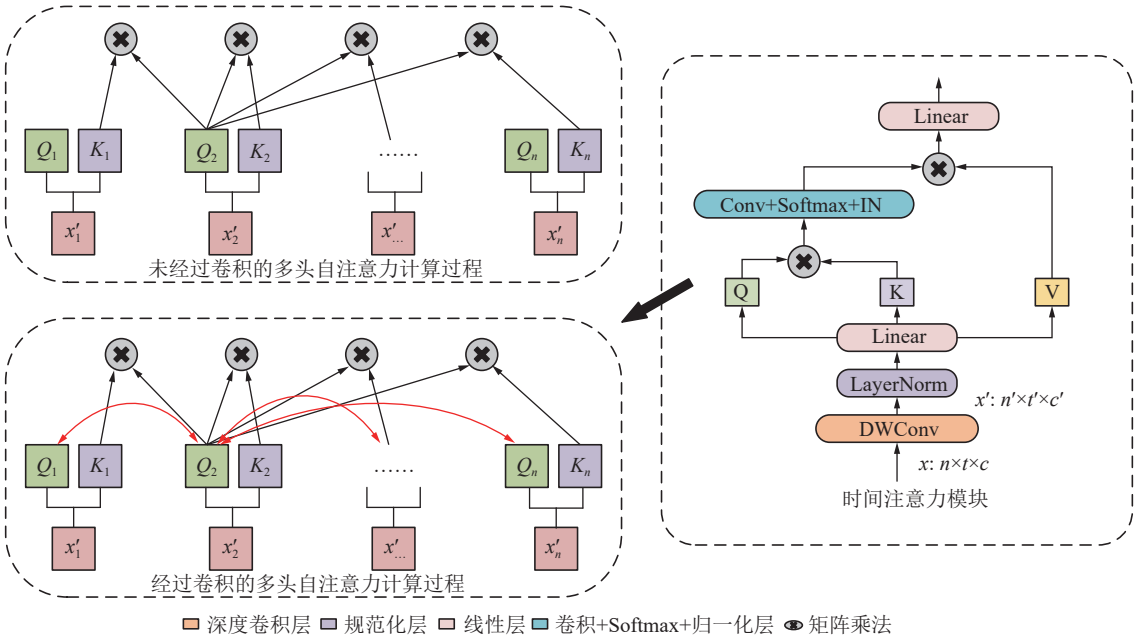


图 4　时间注意力模块

Fig.4　Temporal attention module

3.2.3　通道注意力模块

由于空间−时间−通道注意力模块整合了输入特征的局部信息和全局信息, 这使得在面对综采工作面视频模糊这种情况时, 模型依旧可以在有限的输入特

征中有效提取有用信息,从而确保识别准确度不下降。

因为综采工作面环境灯光昏暗、煤尘水雾等常附着于目标表面,易对空间特征信息造成干扰。如果单独以空间特征为主会使得目标行为特征缺失,严重会导致行为识别失效。为此,本文在 STC-ResT 中加入通道注意力模块,从通道维度出发提取目标的特征信息,减低模型对空间特征的依赖性。

在综采工作面视频数据中,最初视频帧中包含 RGB 三色彩通道。后来,随着多尺度特征信息提取,通道数逐渐增加。因此,通道维度中蕴含的信息也越来越多,通道维度的重要性也逐渐显现。而通道注意力就是以捕获通道维度信息为主,会自动地对每个通道的重要程度进行加权。并且通道注意力是将每个通道单元作为一个输入,如图 5 所示。通道注意力会考虑到单个通道内所有元素的空间位置关系,从而建立全局的信息关联。

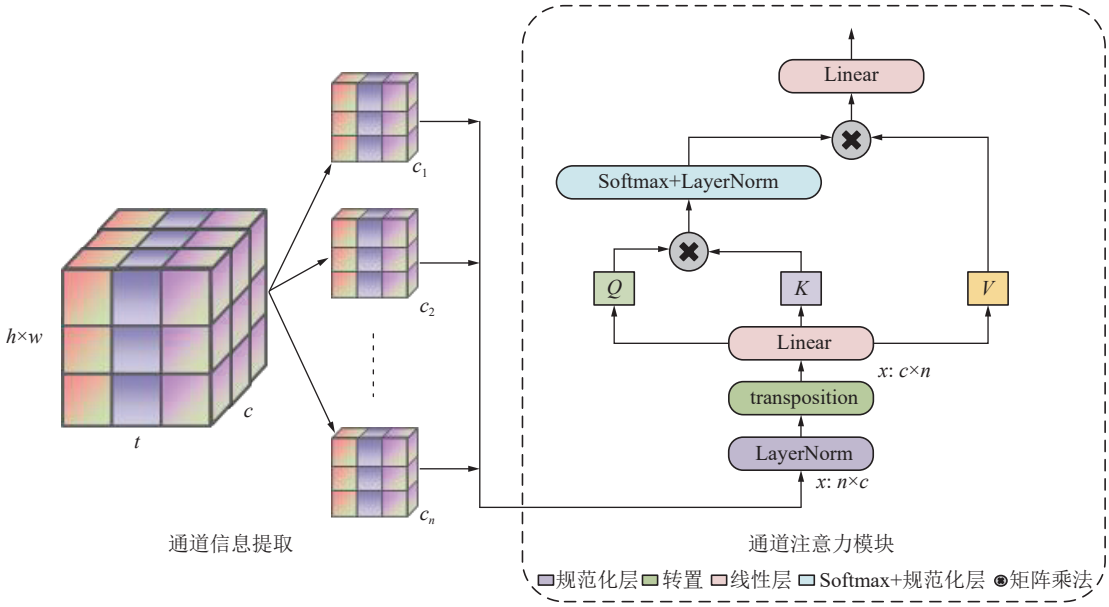


图 5 通道注意力模块

Fig.5 Channels the attention module

通道注意力的计算不同于空间上分块计算和时间上分帧计算的方式,而是以全局注意力的方式进行特征提取。特别是查询向量$\boldsymbol{Q}$、键向量$\boldsymbol{K}$、值向量$\boldsymbol{V}$的获取也是由通道维度进行线性投影所得到。因此,保证了通道维度的权重不受其它因素的影响。其通道注意力函数如式 (13) 所示。

$$F_{\mathrm{CMSA}}=\mathrm{Softmax}\left(\frac{\boldsymbol{Q}^{\mathrm{T}}\boldsymbol{K}}{\sqrt{C}}\right)\boldsymbol{V}^{\mathrm{T}} \tag{13}$$

其中,$\boldsymbol{Q}$、$\boldsymbol{K}$、$\boldsymbol{V}$为查询向量、键向量和值向量,C为通道的维度。特别的是通道注意力通过转置的方式将原始输入的维度由$n\times c$变为$c\times n$。这样既保证了通道注意力的独立性,又增强了通道注意力泛化性,使通道注意力模块可以适应任意大小的输入。

4 实验设置与分析

4.1 工作面行为识别数据集

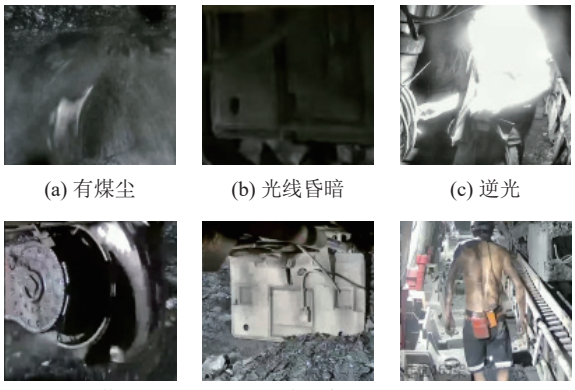
由于不同综采工作面的工作环境、开采设备、现场布置情况等都各不相同,这就导致了数据集的通用性降低。如通常情况下液压支架采用一级护帮板,而对于煤层较软、采高较大的工作面液压支架采用两级护帮,其对护帮板打开和关闭动作特征差异较大。

为了增加模型的适应广度,本文数据集主要来源于 3 个真实综采工作面的监控视频,煤层采高为 2～5 m。其中,液压支架的护帮板包含一级护帮板和无护帮板两种情况。这使得数据集涵盖了部分开采条件相似的工作面,从而使得模型具有一定的泛化能力。

在综采工作面回采过程中,人员动作、护帮板动作以及采煤机动作是安全生产和智能控制的关键信息。其中,人员站立、行走是工作面最常见的行为,同时也是危险区域闯入报警、操作到位检测等后续任务的基础信息;而采煤机开采与停采、护帮板打开与关闭这 4 个动作则在采煤机与支架协同控制中起到关键的耦合作用。为此,本文选择上述 6 种基本动作来研究工作面视频目标行为识别算法。数据集共收集了 4 856 个有效视频段,并将其按照 8∶1∶1 的比例划分为训练集、验证集和测试集。单个输入视频的时长为 5 s,帧率为 25 帧/s,视频的高和宽皆为 224。

综采工作面环境复杂,且回采过程中产生大量煤

尘、水雾等干扰, 会弥漫在空气中或附着于摄像头和物体表面, 如图 6a 所示。相比于图 6d 无煤尘的情况下, 煤尘、水雾等干扰导致成像结果存在大量的模糊现象, 对识别结果造成严重影响。其次, 综采工作面光照条件差, 普遍存在光线昏暗、光照不均等情况, 如图 6b 所示。相较于图 6e 光线明亮的情况下, 光线昏暗导致识别目标与背景融为一体, 难以对目标特征进行区分, 使得识别目标特征出现差异。另外, 由于逆光等造成的过度曝光、光比过大等问题, 如图 6c 所示。相较于图 6f 顺光的情况下, 逆光会导致识别目标关键特征丢失, 进而导致识别失效。



(a) 有煤尘　(b) 光线昏暗　(c) 逆光

(d) 无煤尘　(e) 光线明亮　(f) 顺光

图 6　综采工作面特殊环境数据集

Fig.6　Dataset of special environment of fully mechanized mining face

4.2　实验环境与配置

经过反复实验和优化, 本文选择 SDG 优化器, 并设置了 20 个训练轮数。在每个轮次学习中, 从视频数据集提取 8 帧作为输入, 每帧输入图像的高和宽都是 224。初始学习率设置为 0.01, 在第 11 轮以后降低为 0.001。

图 7 呈现了模型在训练以及测试过程的详细流程图。整个过程由前向传播和反向传播两个过程组成。在前向传播阶段, 模型对输入数据进行特征提取, 生成权重矩阵。利用权重矩阵在验证集进行验证, 并通过计算得到误差。根据实际要求对误差进行评估, 判断其是否满足要求, 从而决定下一步操作。若误差满足要求则在测试集上进行测试得到最终结果。若不满足要求则进行反向传播, 而反向传播是一个关键的反馈调节的过程, 可根据误差重新提取特征对权重矩阵进行调整, 实现对模型性能的优化。这一迭代过程会持续进行, 直至达到训练轮次。

表 1　实验环境

Table 1　Experimental environment

名称	型号
System	Ubuntu20.04.6 LTS
Frame	PyTorch
CPU	Intel Core i7-8700K
GPU	Nvidia GeForce RTX 4 080
Librarys	OpenCV、CUDA、CUDNN
IDE	PyCharm、QT、DeepStream

4.3　对比实验

为了验证 STC-ResT 的性能以及提高实验的说服力, 本文分别从基于卷积和 Transformer 的模型中选出几个主流框架, 并制作统一的数据集和相同实验环境对当下流行的几类行为识别模型进行对比。其中 C3D[28]、R3D[33]、R2+1D[34]和 SlowFast[71]这 4 种模型是以卷积神经网络为基础的主流框架。Timesformer[46]、Vivit[47]和 Swin Transformer[48]则是以 Transformer 为基础的模型框架。

因为视频行为识别的主要任务是准确识别视频中目标的行为, 以及在系统实际部署中的真实效果。所以本文更多侧重在参数量和准确度方面能否满足工作面行为识别的要求。所以选取了以下 5 种评价指标作为评价标准。分别为 Acc(Accuracy)、浮点运算次数 Flops(Floating-point Operations Per second)、模型参数大小、准确度−计算量比和准确度−参数量比这 5 个方面作为评价指标。其中, Acc 是指模型在单独测试集上预测正确的个数与预测总数的比值, 反映了模型在场景中对目标识别结果判定的对与错, 称为准确度。Flops 是指模型进行加乘操作的次数, 反映了模型的计算复杂度。模型参数大小则是反应了模型的参数量以及结构配置等信息, 可以评判网络模型

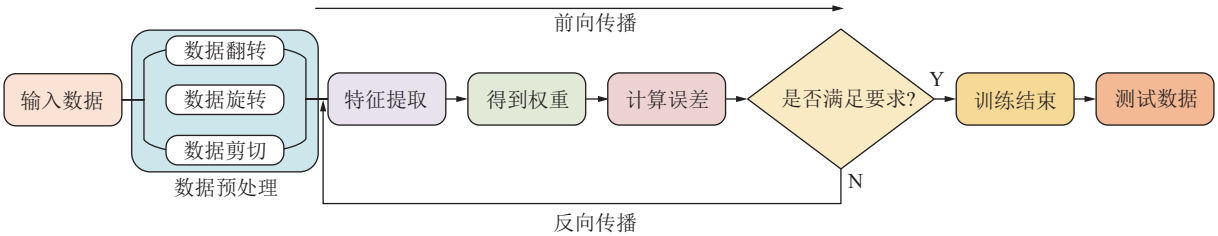


图 7　训练、测试流程

Fig.7　Flow chart of train and test

的大小。准确度−计算量比为准确度和 Flops 的比值，反映了模型在识别准确度和计算复杂度下的识别效率，数值越大代表效率越高。准确度−参数量比则是准确度和参数量的比值，反映了模型对参数利用率，数值越大表示利用率越高。

从表 2 可以看出 STC-ResT 的准确率明显高于其他模型，且模型参数大小远低于其它模型。与 Vivit 相比准确率高出 20.79%，与 R3D 相比准确率虽然只高出了 4.86%，但 Flops 和模型参数大小仅有 R3D 的 1/2。STC-ResT 与 Timesformer 相比 Flops 为它的 1/6，模型参数大小为它的 1/6，而准确率却高出 10.62%。实验结果显示，STC-ResT 拥有最高的准确率和最低的模型参数大小。但是，由于综采工作面环境复杂、特征难以提取，如果用较小计算量的模型难以捕捉到数据中的复杂关系和丰富的特征信息。为了提升行为识别的准确度，STC-ResT 将提取特征分为空间、时间和通道 3 个维度，并通过多次迭代实现对 3 个维度特征信息的提取。这使得模型的计算量高于 Slow-Fast 和 Swin-Transformer 这两种算法，从而导致了 STC-ResT 的精确度−计算量之比并未达到最高。

表 2 对比实验结果

Table 2 Comparative experimental results

模型	Acc/%	FIOPs/G	模型参数大小/M	准确度−计算量比	准确度−参数量比
C3D	86.73	150.48	78.02	0.57	1.11
R3D	92.04	202.99	47.34	0.45	1.94
R2+1D	86.73	163.49	33.18	0.53	2.61
SlowFast	80.09	12.34	33.57	6.49	2.38
Timesformer	86.28	371.20	121.05	0.23	0.71
Vivit	76.11	264.98	170.65	0.28	0.44
Swin-Transformer	85.84	17.27	27.51	4.97	3.12
STC-ResT	**96.90**	**74.63**	**21.15**	**1.29**	**4.58**

综采工作面行为识别结果需要与其他设备联动，所以动作识别准确度直接关系到后续任务的执行。同时，受现阶段技术水平限制，动作识别算法大部分都要调用 GPU 进行运算，这使得对 GPU 显存的要求较高。为此，提升准确度−参数量之比是更加重要的性能指标。由于 STC-ResT 通过迭代方式获取特征的空间、时间和通道三个维度信息，模型的参数量减小较为明显，从而达到了最大的准确度−参数量之比。

4.4 消融实验

本文在 ResT 原有空间注意力的基础上加入了时间注意力模块和通道注意力模块，其中，SMSA(Spatial Multi-Head Self-Attention) 表示空间注意力模块、TMSA(Temporal Multi-Head Self-Attention) 表示时间注意力模块、CMSA(Channel Multi-Head Self-Attention) 表示通道注意力模块。为了验证各个模块的作用，本文通过消融实验进行验证，其结果见表 3。

如表 3 所示，实验 1 为 ResT 在只有空间注意力模块的情况下准确度为 88.94%。实验 2 是在 ResT 中有空间和通道注意力模块下的识别结果，相较于实验 1 准确率可提升 2.21%；实验 3 为在 ResT 中有空间和时间注意力模块下的识别结果，相较于实验 1 准确率提升了 4.87%；实验 4 为 ResT 中有时间和通道注意力模块下的识别结果，识别的准确率相较于实验 1 提升 3.98%；实验 5 为同时有空间、时间、通道注意力模块下的实验结果，其准确率提升 7.96%。通过实验证明，本文所提的空间注意力模块、时间注意力模块和通道注意力模块对于视频行为识别都有较大的提升。

表 3 消融实验结果

Table 3 Results of ablation experiment

实验名称	CMSA	TMSA	SMSA	Acc/%	提升值/%
实验1			√	88.94	0
实验2	√		√	91.15	2.21
实验3	√	√		92.92	4.87
实验4		√	√	93.81	3.98
实验5	√	√	√	**96.90**	**7.96**

为了分析模型的各个模块对目标行为识别的真实效果，以及方便后续的研究。本文绘制了单一模块下对同一测试数据集的识别结果混淆矩阵，如图 8 所示。图中对角线为识别正确类别的准确率，其余的数据为判断错误的类型，数据的横坐标为真实标签，纵坐标为预测标签。从图 8 中可以看出模型在整体识别准确率上都达到了一个不错的结果，对采煤机、护

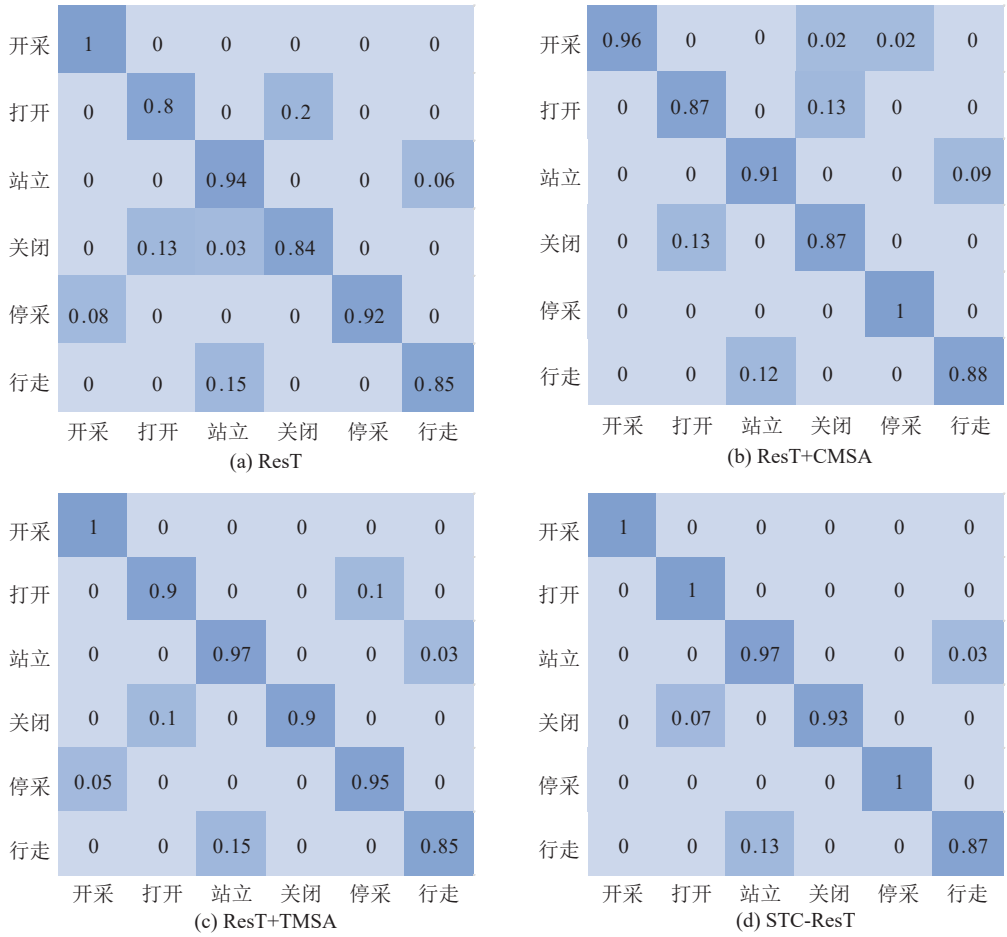


图 8　混淆矩阵对比

Fig.8　Confusion matrix contrast diagram

帮板以及人员等主体的识别都达到了较好的准确率。而大部分的错误识别则是分布在同一主体的不同动作上, 如护帮板的打开和关闭、人员的站立和行走。

通过对混淆矩阵结果进行深入分析后发现, 时间注意力是提取目标在一段时间内的变化信息, 也是对目标动作特征的重要提取手段。在该模块的作用下, 识别准确度出现较大提升, 各个动作的识别准确率都优于基础模型 ResT。通道注意力是对目标全局信息的提取, 相较于 ResT 能更好的提取到目标的整体信息。如在护帮板的打开、关闭和采煤机的停采这一组动作中, 准确率会优于基础的 ResT。但同时在采煤机的开采以及人员站立识别准确率出现下降。这是因为在识别错误的测试视频中发现多目标存在同一场景下, 导致采煤机的开采动作出现错误, 如在采煤机的视频数据集中同时出现了护帮板。而人员的站立则是模型对出现小范围移动更加敏锐, 使得模型对站立、行走这一动作的分辩造成干扰。而将时间注意力和通道注意力结合则是将识别准确率进一步提高, 使得模型能够更加全面的提取目标特征。

本文在空间注意力模块和时间注意力模块中添加了卷积操作, 以减少模型计算开销、增加模型感受野以及增强注意力操作过程中的多头之间的交互性。为测试卷积的性能, 本文通过消融实验进行验证, 其结果见表 4。

表 4　消融实验结果

Table 4　4Results of ablation experiment

模型	准确率/%	浮点运算次数/G
STC-ResT(No CNN)	93.36	71.82
STC-ResT	**96.90**	**74.63**

从表 4 可以看出, 在增加卷积的情况下模型的准确度提升 3.5%。但是卷积操作同样需要进行加乘操作, 所以导致模型 Flops 上升 2.8%。而模型的整体计算量由参数量和加乘次数的乘积组成, 通过前面提出的计算量计算公式 2 可以证明。相较于 Flops 的少量上升, 卷积操作使模型计算量大幅下降, 所以模型整体计算量依旧是下降趋势。图 9 展现了在有无卷积情况下对人员站立这一行为的识别结果。

图中人员在站立情况下虽然总体位置没有发生较大变化, 但是人员上半身并未完全静止, 所以会对

图 9 有无卷积消融实验结果

Fig.9 There are no convolution ablation results

研究结果造成较大干扰。而有卷积操作相较于没卷积操作识别准确度提升了 10%。这是因为在注意力模块中添加卷积操作可以增强模型的感受野,同时增加每个注意力头之间的联动性,使得模型可以关注到更全面的信息,能在大感受野的情况下捕捉到更完整的目标特征。

4.5 识别结果及分析

为验证模型对综采工作面关键设备和人员行为的真实识别效果,本文对测试集识别结果进行展示,如图 10 所示。可以看出 STC-ResT 对目标行为均达到了一个较高的识别准确度,特别是采煤机的“开采”、“停采”和护帮板的“打开”的特征十分明显,所以这 3 类动作在数据集上的识别准确度达到了 100%。但是,实际工程部署中,受样本覆盖率的影响,上述 3 种动作的识别准确率要远低于数据集上的识别准确率。而识别错误的类别都集中在同一物体或人员的不同动作之间。经过对测试集和结果进行分析发现,相较于护帮板打开这一过程,护帮板关闭的过程比较缓慢,且在护帮板关闭视频数据集的前 1～2 s 护帮板都是打开的状态,所以会对识别结果造成一定的干扰。其次,人员的站立并非绝对的静止不动,而是小范围移动,或者是上半身动作较大下半身小范围移动。所以使得人员的行走和站立出现较大的错误识别结果。

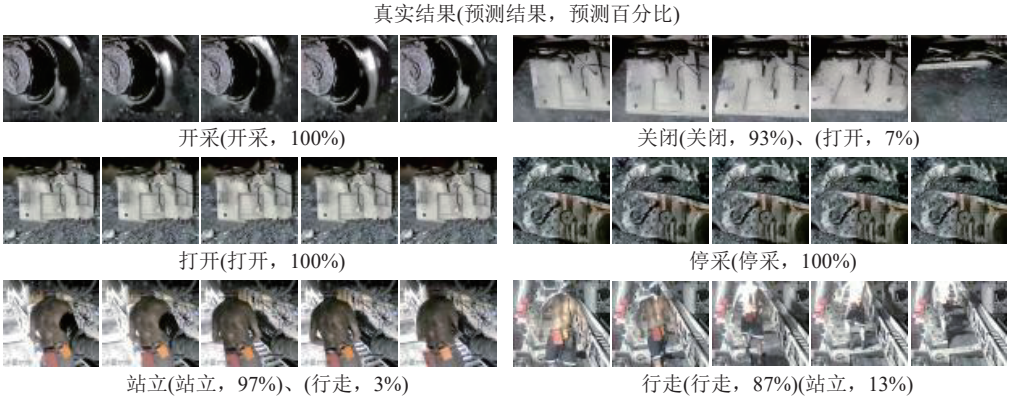


图 10 行为识别结果

Fig.10 Behavior recognition result

5 综采工作面行为识别系统

针对综采工作面视频目标行为识别,本文以 STC-ResT 网络模型为基础搭建了针对综采工作面行为识别系统,如图 11 所示。整套系统由前端、后端和服务器端 3 端组成。端与端之间采用传输速率高、误码率

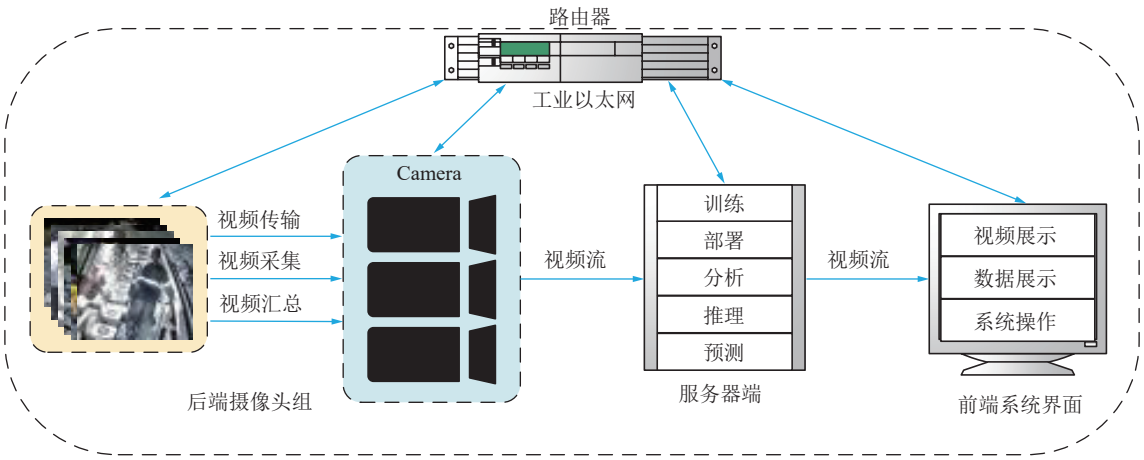


图 11 综采工作面行为检测系统

Fig.11 Behavior detection system of fully mechanized mining face

低且具有专用性质的工业以太网进行通讯, 以保证信息传输的安全。

视频目标行为识别系统的信息在工业以太网 TCP/IP 协议支持下, 完成 RTSP 视频数据流及信息的传输。首先, 工作面摄像头的 RTSP 数据流通过以太网传输至云端服务器; 然后在服务器的 DeepStream 框架下完成模型加载、视频流推理、附加信息和视频流推送; 接着将推理结果存入数据库中; 最后, 系统前端通过 TCP/IP 协议调取数据库信息和 DeepStream 推送的 RTSP 数据流, 完成结果展示和人机交互。

在实际工程应用时, 由于工作面设备配置和环境不同, 可能会导致行为识别的准确率出现一定程度下降。通常情况下可以通过扩充数据集, 增加样本对不同场景的覆盖度来缓解。需要注意的是, 目前采用 DeepStream 还无法实现深度神经网络的在线学习和模型替代。需要在完成模型的离线学习、ONNX 转化、TensorRT 加速后, 再替换原推理模型。替换过程一般在检修班进行, 替换时间一般为几分钟。

系统各部分之间的协同互助, 实现了对视频流和数据流的展示、记录、实时分析等功能。实验证明该系统通过高效的信息传输和处理, 能够在复杂的工作环境中, 提供准确可靠的综采工作面行为识别情况。

5.1　前　端

前端部分用 Qt 制作了一个统筹全局的系统界面, 以便对综采工作面实施远程监控。Qt 是一个跨平台的、支持多操作系统的 C++框架, 主要用于图像界面的开发等。该界面可与服务器端和后端设备相连接, 以便于直接获取现场的视频流和推理分析后得到的数据流, 实现对现场情况实时查看并进行危险预警。

整个前端的操作界面如图 12 所示, 左侧是数据显示区域, 负责对综采工作面识别结果进行实时展示; 中间是视频展示画面, 通过后端的摄像头将视频流传输到系统界面进行播放; 右侧是功能区, 实现对画面调整、摄像头调用、移动检测、入侵报警等一系列功能。当出现异常动作时界面会发布报警信息, 从而反馈给操作人员进行下一步操作。

5.2　后　端

后端部分作为整个系统唯一的 RTSP 视频流获取单元, 具有至关重要的作用。但井下空气潮湿且漂浮着大量煤尘, 极易给摄像头造成腐蚀侵害。于是本系统采用多组井下专用防爆摄像头来采集视频数据。由于其可抵抗复杂环境的影响且具有成像清晰、传输速度快、效率高等优点而被广泛应用于综采工作面等恶劣环境。整个后端通过防爆摄像头实现对综采工作面全方位、多角度的视频流获取、汇聚和管理, 并将得到的 RTSP 视频流分别传入前端和服务器端。

5.3　服务器端网络模型部署

服务器端是整个系统运行的核心, 决定着系统识

图 12　系统操作界面

Fig.12　System operation interface

别的实时性、可行性以及准确度。因此，本系统以 DeepStream 处理框架为基础，搭建综采工作面行为识别系统的推理模型。DeepStream 是一种用于流分析的插件式模块化工具包，可以将视频流作为输入，通过构建 Pipeline 实现编码、解码、分析等功能。其次，搭配 TensorRT 将训练得到的网络模型转化为推理引擎并对后续的视频流进行推理优化。

在综采工作面行为识别系统工作中，重要的是把训练得到的网络模型部署到服务器端，确保视频流可以得到实时处理并达到较好的检测效果。但是，在深度神经网络部署过程中，通常需要将 Pytorch 等算法框架通过 TensorRT 转化为 ONNX 文件后生成 Engine 文件，才能在服务器端实现 GPU 加速。为此，首先需要将网络模型转换为 ONNX 模型文件和 Engine 推理引擎，如图 13 所示。

而模型经过 TensorRT 进行优化和加速之后，能够在 GPU 上实现高性能的推理。这对于行为识别这类推理速度要求较高的应用尤其重要。其次，通过 TensorRT 的优化，模型的推理延迟得到降低，可以适用于对实时性要求高的应用场景。最后，实现 GPU 加速可以充分利用硬件资源，提高模型的资源利用效率，从而更好地满足一些对硬件资源有限制的嵌入式系统或边缘设备的要求。

在部署中，将 ONNX 模型转化为 Engine 模型这一过程，主要是对参数类型和分支进行优化，并生成适配 C++语言调用的模型，面向硬件最大幅度提高计算速度。而在 Engine 模型部署过程中，主要是以 GStreamer 插件管理为基础，搭建适配综采工作面行为识别的 Pipeline 和插件链接的方式，完成行为识别的高效推理。

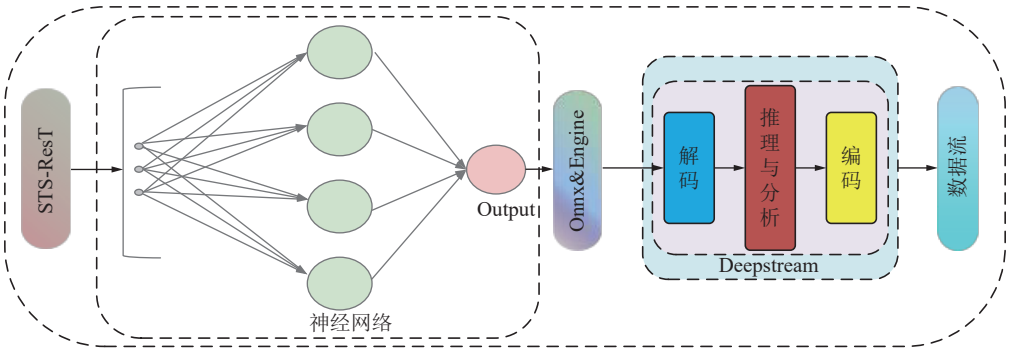


图 13 工程部署过程

Fig.13 Project deployment process

5.4 工程部署与实验

为验证系统的 RTSP 流传输速度、Deepstream 框架推理速度，本文对该系统进行实验验证。实验设备配置及版本信息见表 5。

表 5 系统实验配置

Table 5 System experiment configuration table

名称	版本
CPU	Intel Core i7-12 700
GPU	Nvidia GeForce RTX 4 090
CUDA	12.0
CUDNN	8.8.0
TensorRT	8.6.1
DeepStream	6.0
QT	12.0.1

其中，系统选用分辨率为 1 920×1 080，帧频为 25 帧/s，编码协议 H.264，码率 4 096 Kbps 的摄像头对 RTSP 视频流进行采集。在实验室环境下的以太网组网如图 14 所示。系统采用路由器组成星形网络，其中，路由器为前端、后端和服务器分配 IP 地址，并管理信息交互通道。在工程实践中，以太网组网方式可以采用星形、环形或多级组网方式。

1) 视频 RTSP 推流速度。系统通过以太网读取摄像头的 RTSP 流，并通过 DeepStream 的数据流解码、推理、合并、编码、推流等操作，将目标行为识别结果通过 RTSP 流推送给前端。按照一级路由器组网方式，前端展示的视频与真实场景的延迟大约在 500～600 ms，满足实时性要求。在工程实践中，由于设备过多，以太网可能由多级路由器联网组成。这可能会造成 RTSP 视频流传输时间比实验室环境下偏长。但总体时间延迟基本能控制在 1.5 s 以内，可满足工程需求。

2)DeepStream 推理速度。系统采用 DeepStream 框架对 RTSP 视频流实现推理。过程中采用了 TensorRT 对推理模型进行加速。由于 DeepStream 采用 Pipline 方式对视频流进行插件式管理，因此 DeepStream 可以实现同时对多路视频的处理。其

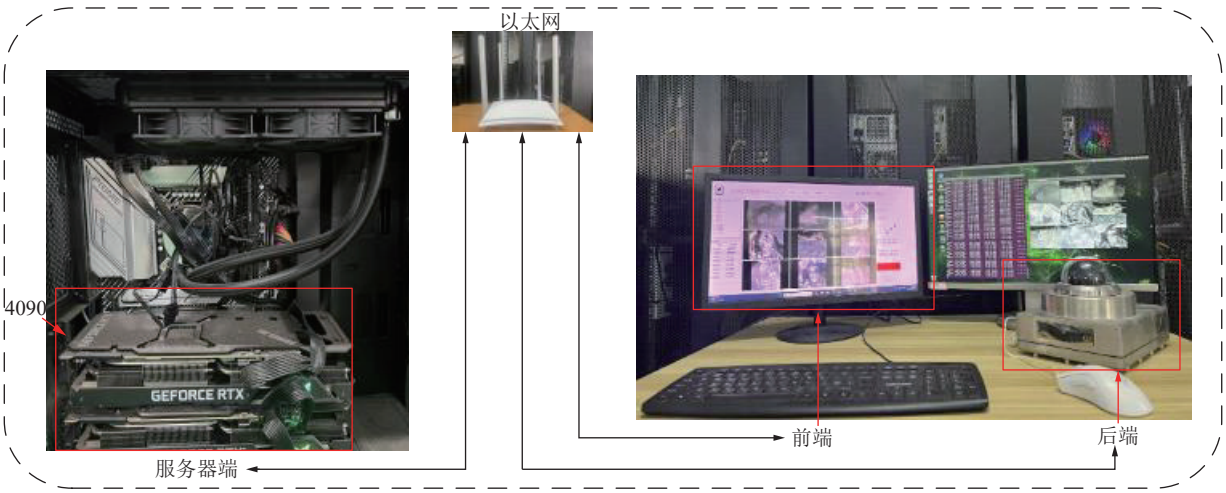


图 14　系统展示

Fig.14　display of the system

推理速度与视频路数有关 (表 6)。

表 6　系统实验结果

Table 6　System experiment results

视频路数	推理时间/ms	帧率/(帧·s^{-1})
1路	16.8	63.5
4路	39.4	25.3
9路	93.9	10.6

当处理 1 路视频时, 推理时间仅为 16.8 ms, 且每秒可处理 63.5 帧图像。随着视频路数增加, 推理速度和帧率同时会降低。当处理 4 路视频时, 可以保持 25.3 帧/s 的识别效果, 大于摄像头的帧频 25 帧/s。这表明本实验的配置可对摄像头的所有视频帧进行推理。当工程应用中的视频路数较多时, 通常采用间隔抽帧的方式降低 DeepStream 的处理负担, 从而提高视频推理的实时性。

6　总结与展望

1) 针对工作面光照普遍条件较差, 且煤尘和水雾等容易引起视频模糊, 导致目标行为的关键特征提取困难, 使得模型对设备和人员的行为识别准确度不高; 以及 ResT 模型以空间信息为主, 缺少对全局信息和对图像的对比度、纹理信息的深度解析等问题构建了 STC-ResT 模型。该模型使用空间−时间−通道信息融合模块对空间维度、时间维度和通道维度的特征进行全方位的提取。其中, 空间注意力则是对目标纹理、位置、形状等深度信息进行表征。时间注意力提供了行为发生的顺序以及演变关系。通道注意力提供了行为的全局特征。在综采工作面数据集中进行验证和实验结果表明, STC-ResT 的准确率为 96.90%, 与主流网络模型相比, 更能够满足综采工作面行为识别准确度的要求。

2) 搭建了综采工作面的远程监控系统, 系统以前端、后端、服务器端进行划分。前端采用 QT 制作了一个远程监控界面对综采工作面的工作情况进行实时监控和记录; 后端以摄像头组构成, 记录了综采工作面全方位、多角度的视频流获取、汇聚和管理; 服务器端则对视频流进行分析和推理。整个系统通过工业以太网的方式进行连接, 提高了信息传输的速率, 并且保证了信息的安全性。

3) 完成了 STC-ResT 网络模型在服务器端的部署。以 DeepStream 流媒体处理框架为基础, 搭建综采工作面行为识别框架。通过 TensorRT 将模型框架转化为 ONNX 文件, 实现服务器中 GPU 对分析与推理过程的加速。之后由 ONNX 文件生成 Engine 文件, 组成综采工作面行为识别的 Pipeline。最终实现对综采工作面行为识别的准确识别和实时推理。

受井下环境的限制, 仍然需要在以下几个方面进一步展开研究:

1) 行为识别的本质是对物体在一段时间内的变化过程的识别。然而, 综采工作面摄像头的视野小, 目标出现的时间短。因此, 如何在更短的时间段内提取目标的时序变化信息, 以提升行为识别的准确性是下一步的主要研究内容。

2) 本文搭建的行为识别系统是以区分目标的动作为目的, 尚未与设备控制形成联动机制。在后续研究中, 将重点开展基于行为识别的设备智能控制方法, 以提升设备智能控制水平。

3) 本文的行为识别数据集是以采煤机开采停采、护帮板的打开关闭和人员的站立行走为主。在后续的研究中会根据实际情况将动作种类扩展到支架的降−移−升、采煤机的进刀方式以及人员的违规操作等

其他动作。

参考文献(References):

[1] 王国法, 刘峰, 庞义辉, 等. 煤矿智能化: 煤炭工业高质量发展的核心技术支撑[J]. 煤炭学报, 2019, 44(2): 349−357.
WANG Guofa, LIU Feng, PANG Yihui, et al. Coal mine intellectualization: The core technology of high quality development[J]. Journal of China Coal Society, 2019, 44(2): 349−357.

[2] 王国法, 刘峰, 孟祥军, 等. 煤矿智能化 (初级阶段) 研究与实践[J]. 煤炭科学技术, 2019, 47(8): 1−36.
WANG Guofa, LIU Feng, MENG Xiangjun, et al. Research and practice on intelligent coal mine construction (primary stage)[J]. Coal Science and Technology, 2019, 47(8): 1−36.

[3] 许献磊, 马正, 陈令洲. 煤矿地质灾害隐患透明化探测技术进展与思考[J]. 绿色矿山, 2023, 1(1): 56−69.
XU Xianlei, MA Zheng, CHEN Lingzhou. Progress and thinking of transparent detection technology for hidden geological hazards in coal mines[J]. Journal of Green Mine, 2023, 1(1): 56−69.

[4] 鲍久圣, 张可琨, 王茂森, 等. 矿山数字孪生 MiDT: 模型架构、关键技术及研究展望[J]. 绿色矿山, 2023, 1(1): 166−177.
BAO Jiusheng, ZHANG Kekun, WANG Maosen, et al. Mine Digital Twin: Model architecture, key technologies and research prospects [J]. Journal of Green Mine, 2023, 1(1): 166−177.

[5] 关于加快煤矿智能化发展的指导意见 [N]. 2020−03−05.
Guiding Opinions on Accelerating the Intelligent Development of Coal Mines [N]. 2020−03−05.

[6] 谭明, 沈政昌, 杨义红. 矿物分选装备技术研究进展[J]. 绿色矿山, 2024, 2(1): 85−93.
TAN Ming, SHEN Zhengchang, YANG Yihong. Research progress of mineral processing equipment technology[J]. Journal of Green Mine, 2024, 2(1): 85−93.

[7] 阮顺领, 李少博, 顾清华, 等. 基于双向特征融合的露天矿区道路障碍检测[J]. 煤炭学报, 2023, 48(3): 1425−1438.
RUAN Shunling, LI Shaobo, GU Qinghua, et al. Road obstacle detection in open-pit mines based on bidirectional feature fusion[J]. Journal of China Coal Society, 2023, 48(3): 1425−1438.

[8] 王国法, 张良, 李首滨, 等. 煤矿无人化智能开采系统理论与技术研发进展[J]. 煤炭学报, 2023, 48(1): 34−53.
WANG Guofa, ZHANG Liang, LI Shoubin, et al. Progresses in theory and technological development of unmanned smart mining system[J]. Journal of China Coal Society, 2023, 48(1): 34−53.

[9] 郝帅, 张旭, 马旭, 等. 基于 CBAM-YOLOv5 的煤矿输送带异物检测[J]. 煤炭学报, 2022, 47(11): 4147−4156.
HAO Shuai, ZHANG Xu, MA Xu, et al. Foreign object detection in coal mine conveyor belt based on CBAM-YOLOv5[J]. Journal of China Coal Society, 2022, 47(11): 4147−4156.

[10] 王卫东, 张康辉, 吕子奇, 等. 基于机器视觉的煤中杂物智能分选系统研究[J]. 选煤技术, 2020, 48(2): 87−91.
WANG Weidong, ZHANG Kanghui, LYU Ziqi, et al. A study of the machine vision-based intelligent separation system for extraction of tramp materials in raw coal[J]. Coal Preparation Technology, 2020, 48(2): 87−91.

[11] 程健. 煤矿巷道机器人管线视觉辅助定位与导航方法研究[J]. 煤炭科学技术, 2020, 48(7): 226−232.
CHENG Jian. Study on pipeline vision-aided positioning and navigation method for coal mine tunnel robot[J]. Coal Science and Technology, 2020, 48(7): 226−232.

[12] 杨建辉, 黄子洋, 汪梅, 等. 机器视觉灰度化金字塔卷积模型的煤流异物识别[J]. 煤炭科学技术, 2022, 50(11): 194−201.
YANG Jianhui, HUANG Ziyang, WANG Mei, et al. Recognition of unwanted objects in coal flow based on gray pyramid convolution model of machine vision[J]. Coal Science and Technology, 2022, 50(11): 194−201.

[13] 王学文, 王孝亭, 谢嘉成, 等. 综采工作面 XR 技术发展综述: 从虚拟 3D 可视化到数字孪生的演化[J]. 绿色矿山, 2024, 2(1): 75−84.
WANG Xuewen, WANG Xiaoting, XIE Jiacheng, et al. Review of XR technology development in fully mechanized mining faces: From 3D visualization to digital twin[J]. Journal of Green Mine, 2024, 2(1): 75−84.

[14] JOHANSSON G. Visual motion perception[J]. Scientific American, 1975, 232(6): 76−88.

[15] O'ROURKE J, BADLER N I. Model-based image analysis of human motion using constraint propagation[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1980(6): 522−536.

[16] NAGEL H H. From image sequences towards conceptual descriptions[J]. Image and Vision Computing, 1988, 6(2): 59−74.

[17] KLAESER A, MARSZALEK M, SCHMID C. A spatio-temporal descriptor based on 3D-gradients[C]//Proceedings ofthe British Machine Vision Conference 2008. British Machine Vision Association, 2008.

[18] ZHANG Z M, HU Y Q, CHAN S, et al. Motion context: A new representation for human action recognition[C]// Computer Vision – ECCV 2008. Berlin, Heidelberg: Springer Berlin Heidelberg, 2008: 817−829.

[19] SOMASUNDARAM G, CHERIAN A, MORELLAS V, et al. Action recognition using global spatio-temporal features derived from sparse representations[J]. Computer Vision and Image Understanding, 2014, 123: 1−13.

[20] DAS DAWN D, SHAIKH S H. A comprehensive survey of human action recognition with spatio-temporal interest point (STIP) detector[J]. The Visual Computer, 2016, 32(3): 289−306.

[21] NGUYEN T V, SONG Z, YAN S C. STAP: Spatial-temporal attention-aware pooling for action recognition[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2015, 25(1): 77−86.

[22] PENG X J, WANG L M, WANG X X, et al. Bag of visual words and fusion methods for action recognition: Comprehensive study and good practice[J]. Computer Vision and Image Understanding, 2016, 150: 109−125.

[23] GAIDON A, HARCHAOUI Z, SCHMID C. Activity representation with motion hierarchies[J]. International Journal of Computer Vision, 2014, 107(3): 219−238.

[24] WANG H, ONEATA D, VERBEEK J, et al. A robust and efficient video representation for action recognition[J]. International Journal of Computer Vision, 2016, 119(3): 219−238.

[25] TECHNICOLOR T S, RELATED S O R, TECHNICOLOR T S, et

al. ImageNet Classification with Deep Convolutional Neural Networks [50] [J].

[26] FUKUSHIMA K. Neocognitron: A self organizing neural network model for a mechanism of pattern recognition unaffected by shift in position[J]. Biological Cybernetics, 1980, 36(4): 193−202.

[27] SIMONYAN K, ZISSERMAN A. Two-stream convolutional networks for action recognition in videos[J]. ArXiv e-Prints, 2014: 1406.2199.

[28] TRAN D, BOURDEV L, FERGUS R, et al. Learning spatiotemporal features with 3D convolutional networks[C]//2015 IEEE International Conference on Computer Vision (ICCV). Piscataway, NJ: IEEE, 2015: 4489−4497.

[29] 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). Piscataway, NJ: IEEE, 2017: 4724−4733.

[30] DIBA A L, FAYYAZ M, SHARMA V, et al. Temporal 3D ConvNets: New architecture and transfer learning for video classification[EB/OL]. 2017: 1711.08200. https://arxiv.org/abs/1711.08200v1.

[31] XIE S N, SUN C, HUANG J, et al. Rethinking spatiotemporal feature learning: Speed-accuracy trade-offs in video classification[C]// Computer Vision – ECCV 2018. Cham: Springer International Publishing, 2018: 318−335.

[32] QIU Z F, YAO T, MEI T. Learning spatio-temporal representation with pseudo-3D residual networks[C]//2017 IEEE International Conference on Computer Vision (ICCV). Piscataway, NJ: IEEE, 2017: 5534−5542.

[33] HARA K, KATAOKA H, SATOH Y. Can spatiotemporal 3D CNNs retrace the history of 2D CNNs and ImageNet? [C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2018: 6546−6555.

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

[35] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[J]. Advances in neural information processing systems, 2017, 30.

[36] DEVLIN J, CHANG M W, LEE K, et al. BERT: Pre-training of deep bidirectional transformers for language understanding[EB/OL]. 2018: 1810.04805. https://arxiv.org/abs/1810.04805v2.

[37] YANG Z L, DAI Z H, YANG Y M, et al. XLNet: Generalized autoregressive pretraining for language understanding[J]. 2019: 5753−5763.

[38] RAFFEL C, SHAZEER N, ROBERTS A, et al. Exploring the limits of transfer learning with a unified text-to-text transformer[J]. Journal of Machine Learning Research, 2020, 21(1): 5485−5551.

[39] PARMAR N, VASWANI A, USZKOREIT J, et al. Image transformer[EB/OL]. 2018: 1802.05751. https://arxiv.org/abs/1802.05751v3.

[40] RADFORD A, WU J, CHILD R, et al. Language models are unsupervised multitask learners[J]. OpenAI blog, 2019, 1(8): 9.

[41] SUN Y, WANG S H, FENG S K, et al. ERNIE 3.0: Large-scale knowledge enhanced pre-training for language understanding and generation[J]. ArXiv e-Prints, 2021: arXiv: 2107.02137.

[42] BI K F, XIE L X, ZHANG H H, et al. Accurate medium-range global weather forecasting with 3D neural networks[J]. Nature, 2023, 619(7970): 533−538.

[43] GIRDHAR R, JOÃO CARREIRA J, DOERSCH C, et al. Video action transformer network[C]//2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway, NJ: IEEE, 2019: 244−253.

[44] KONSTANTINIDIS D, PAPASTRATIS I, DIMITROPOULOS K, et al. Multi-manifold attention for vision transformers[J]. IEEE Access, 2023, 11: 123433−123444.

[45] LI Y, WU C, FAN H, et al. Improved multiscale vision transformers for classification and detection. arXiv 2021 [J]. arXiv preprint arXiv: 211201526.

[46] BERTASIUS G, WANG H, TORRESANI L. Is space-time attention all you need for video understanding? [EB/OL]. 2021: 2102.05095. https://arxiv.org/abs/2102.05095v4.

[47] ARNAB A, DEHGHANI M, HEIGOLD G, et al. ViViT: A video vision transformer[C]//2021 IEEE/CVF International Conference on Computer Vision (ICCV). Piscataway, NJ: IEEE, 2021: 6816−6826

[48] LIU Z, LIN Y T, CAO Y, et al. Swin transformer: Hierarchical vision transformer using shifted windows[C]//2021 IEEE/CVF International Conference on Computer Vision (ICCV). Piscataway, NJ: IEEE, 2021: 9992−10002.

[49] TONG Z, SONG Y, WANG J, et al. VideoMAE: Masked Autoencoders are Data-Efficient Learners for Self-Supervised Video Pre-Training [J]. 2022.

[50] HE K M, CHEN X L, XIE S N, et al. Masked autoencoders are scalable vision learners[C]//2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway, NJ: IEEE, 2022: 15979−15988.

[51] FAN H Q, XIONG B, MANGALAM K, et al. Multiscale vision transformers[C]//2021 IEEE/CVF International Conference on Computer Vision (ICCV). Piscataway, NJ: IEEE, 2021: 6804−6815

[52] KIM D, ANGELOVA A, KUO W C. Region-centric image-language pretraining for open-vocabulary detection[M]//Computer vision–ECCV 2024. Cham: Springer Nature Switzerland, 2024: 162−179.

[53] REKAVANDI A M, RASHIDI S, BOUSSAID F, et al. Transformers in small object detection: A benchmark and survey of state-of-the-art[J]. arxiv preprint arxiv: 2309.04902, 2023.

[54] DOSOVITSKIY A, BEYER L, KOLESNIKOV A, et al. Transformers for image recognition at scale [J]. arXiv preprint arXiv: 201011929, 2020.

[55] TOLSTIKHIN I O, HOULSBY N, KOLESNIKOV A, et al. Mlp-mixer: An all-mlp architecture for vision[J]. Advances in neural information processing systems, 2021, 34: 24261−24272.

[56] 胡杨杨. 矿井视频中人体行为检测方法研究[D]. 武汉: 武汉理工大学, 2016.
HUYANG Yang. Research on human behavior detection method in mine video [D]. Wuhan: Wuhan University of Technology, 2016.

[57] 杨超宇, 李策, 苏剑臣, 等. 基于视频的煤矿安全监控行为识别系

统研究[J]. 煤炭工程, 2016, 48(4): 111−113, 117.
YANG Chaoyu, LI Ce, SU Jianchen, et al. Research on video-based system of activity recognition for coal mine safety surveillance[J]. Coal Engineering, 2016, 48(4): 111−113,117.

[58] 陈庆峰. 矿井皮带区域矿工不安全行为识别方法的研究[D]. 徐州: 中国矿业大学, 2019.
CHEN Qingfeng. Research on the identification method of miners' unsafe behaviors in the belt area of a mine[D]. Xuzhou: China University of Mining and Technology, 2019.

[59] 杨赛烽. 基于 Kinect 的罐笼内矿工不安全行为识别方法研究[D]. 徐州: 中国矿业大学, 2019.
YANG Saifeng. Research on the identification method of miners' unsafe behaviors in the belt area of a mine[D]. Xuzhou: China University of Mining and Technology, 2019.

[60] LECUN Y, BENGIO Y, HINTON G. Deep learning[J]. Nature, 2015, 521(7553): 436−444.

[61] 党伟超, 张泽杰, 白尚旺, 等. 基于改进双流法的井下配电室巡检行为识别[J]. 工矿自动化, 2020, 46(4): 75−80.
DANG Weichao, ZHANG Zejie, BAI Shangwang, et al. Inspection behavior recognition of underground power distribution room based on improved two-stream CNN method[J]. Industry and Mine Automation, 2020, 46(4): 75−80.

[62] 温廷新, 王贵通, 孔祥博, 等. 基于迁移学习与残差网络的矿工不安全行为识别[J]. 中国安全科学学报, 2020, 30(3): 41−46.
WEN Tingxin, WANG Guitong, KONG Xiangbo, et al. Identification of miners' unsafe behaviors based on transfer learning and residual network[J]. China Safety Science Journal, 2020, 30(3): 41−46.

[63] 黄瀚, 程小舟, 云霄, 等. 基于 DA-GCN 的煤矿人员行为识别方法[J]. 工矿自动化, 2021, 47(4): 62−66.
HUANG Han, CHENG Xiaozhou, YUN Xiao, et al. DA-GCN-based coal mine personnel action recognition method[J]. Industry and Mine Automation, 2021, 47(4): 62−66.

[64] 刘浩, 刘海滨, 孙宇, 等. 煤矿井下员工不安全行为智能识别系统[J]. 煤炭学报, 2021, 46(S2): 1159−1169.
LIU Hao, LIU Haibin, SUN Yu, et al. Intelligent recognition system of unsafe behaviors of underground coal mine employees J]. Journal of China Coal Society, 2021, 46(S2): 1159−1169.

[65] 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.

[66] 张雷, 冉凌鏄, 代婉婉, 等. 基于融合网络的井下人员行为识别方法[J]. 工矿自动化, 2023, 49(3): 45−52.
ZHANG Lei, RAN Lingbo, DAI Wanwan, et al. Behavior recognition method for underground personnel based on fusion network[J]. Industry and Mine Automation, 2023, 49(3): 45−52.

[67] 岳志奇. 煤矿综采工作面复杂条件下人的安全行为模式研究[D]. 焦作: 河南理工大学, 2016.
YUE Zhiqi. Research on human safety behavior model under complex conditions of coal mine working face[D]. Jiaozuo: Henan Polytechnic University, 2016.

[68] 杨峰, 徐友庆, 孟祥峰, 等. 基于视觉关系检测的煤矿综采工作面不安全行为识别方法: CN201910360181.1[P]. CN110119701A[2025−02−27].
YANG Feng, XU Youqing, MENG Xiangfeng, et al. Unsafe Behavior Recognition Method of fully mechanized coal mine Working Face Based on visual relationship Detection: CN201910360181.1[P]. CN110119701A[2025−02−27].

[69] ZHANG Q L, YANG Y B. ResT: An efficient transformer for visual recognition[EB/OL]. 2021: 2105.13677. https://arxiv.org/abs/2105.13677v5.

[70] HE K M, ZHANG X Y, REN S Q, et al. Deep residual learning for image recognition[C]//2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway, NJ: IEEE, 2016: 770−778.

[71] FEICHTENHOFER C, FAN H Q, MALIK J, et al. SlowFast networks for video recognition[C]//2019 IEEE/CVF International Conference on Computer Vision (ICCV). IEEE, 2019.