Modeling and Simulation 建模与仿真, 2025, 14(2), 449-459
Published Online February 2025 in Hans. https://www.hanspub.org/journal/mos
https://doi.org/10.12677/mos.2025.142166

Hans汉斯

基于双流Mobile Vit与通道剪枝的工业过程故障诊断

马旺兵，田　颖

上海理工大学光电信息与计算机工程学院，上海

收稿日期：2025年1月27日；录用日期：2025年2月20日；发布日期：2025年2月28日

摘　要

工业过程视频监控数据的时间连续性和空间连续性，但是已有的基于视频数据的故障诊断模型有规模庞大而难以部署。为此，本研究提出了一种基于双流Mobile Vit的轻量化视频分类模型用于故障诊断，并且利用权重剪枝技术来降低模型大小。首先提取工业视频的视频帧和稠密光流分别作为工业过程的空间特征和时序特征，再使用两条轻量化主干网络Mobile Vit提取视频的空间特征和时序特征。最终在双流模型的尾部使用卷积注意力融合机制使光流特征和特征充分融合用于最终的诊断。为使模型更加轻量化，模型权重剪枝被用来降低双流Mobile Vit的参数量。通过实验表明，相比于其他故障诊断模型,本研究所提出的模型和所应用的剪枝方法在取得较高的诊断精度的同时，模型大小也远远低于其他模型。

关键词

视频监控，模型剪枝，双流模型，Mobile Vit

Industrial Process Fault Diagnosis Based on Two Stream Mobile Vit and Channel Pruning

Wangbing Ma, Ying Tian

School of Optical-Electrical and Computer Engineering, University of Shanghai for Science and Technology, Shanghai

Received: Jan. 27th, 2025; accepted: Feb. 20th, 2025; published: Feb. 28th, 2025

Abstract

Industrial process video surveillance data has the temporal and spatial continuity，however the existing fault diagnosis models based on video data is too large be deployed in real actual industrial process. Therefore, this study proposes a lightweight video classification model based on Two-stream Mobile Vit for fault diagnosis, and the weights of the Two-Stream Mobile Vit model are

文章引用: 马旺兵, 田颖. 基于双流 Mobile Vit 与通道剪枝的工业过程故障诊断[J]. 建模与仿真, 2025, 14(2): 449-459.
DOI: 10.12677/mos.2025.142166

pruned. Firstly, video frames and dense optical flows of industrial videos are extracted as spatial and temporal features of industrial processes, respectively. Then, the backbones of the proposed model Mobile Vit is used to extract deep spatial features and temporal features. Finally, the Convolutional Attention Fusion Mechanism is used at the tail of the Two-Stream model to fully fuse spatial features and temporal features for final diagnosis. In order to make the proposed model more lightweight, model weight pruning is used to reduce the number of parameters of the whole model. It is shown through experiments that compared with other fault diagnosis models, the model proposed and the pruning method applied in this study achieve higher diagnostic accuracy while the model size is much lower than other models.

Keywords

Video Monitoring, Model Prune, Two Stream Model, Mobile Vit

1. 引言

随着工业生产的不断发展，工业过程的稳定性和效率成为确保生产高产量和高质量的关键因素。在这一背景下，故障诊断的准确性和及时性对于保障工业过程的持续运行至关重要。

传统故障诊断方法都是基于数据模型，现代工业生产过程的集成度和复杂度的不断提升，生产过程通常表现出规模庞大、包含多个单元并且流程复杂的特点使得精确地建立数学模型变得极具挑战性。然而，随着传感器技术的发展，更多的过程变量数据可以被收集到。因此，基于过程数据(传感器数据)的机器学习算法用于工业过程故障诊断，例如：Principle Component Analysis (PCA) [1]、Support Vector Machine (SVM) [2]和 Random Forest (RF) [3]。

更进一步地，随着大数据技术的发展，工业过程视频数据被拍摄收集。因此，基于视频数据的故障诊断成为研究的热点。N Davari [4]线路的视频中提取帧，使用 Faster R-CNN 在每一帧中检测电源设备，然后在整个视频每一帧中对其进行跟踪，然后，使用双流充气 3D 卷积(Inflated 3D ConvNet, I3D)来分别识别每个设备的图像中的电晕放电，确定初始故障严重程度。徐磊[5]等人在 Swin Transformer Block 中加入 3D 卷积形成 Swinc Transformer，并以 Swinc Transformer 为主干网络构建双流模型对化工过程视频进行故障分类，相比较于其他基于视频数据的故障诊断模型取得更高的准确率。

尽管深度学习模型在故障诊断任务表现很好，但其参数量过于庞大难以被应用到实际生产过程中。为了进一步解决上述问题，本文结合工业过程实际特点提出了双流 Mobile Vit 视频分类模型用于故障诊断。首先，视频数据被预处理成 RGB 帧和光流的形式。随后，两条主干网络 Mobile Vit 分别提取 RGB 帧和光流中所包含的空间特征和时序特征。在两条主干网络的末端 Convolution Attention Fusion Mechanism 将被提取出来的空间特征和时序特征充分融合。最终融合后的特征被输入到分类器中实现系统状态诊断。尽管双流 Mobile Vit 中主干网络采用的是 Mobile Vit，其重要组成中有参数量过大卷积层和全连接层。因此，模型需要进一步被剪枝和实现轻量化。该模型主要优势如下：

(1) Mobile Vit 核心模块 Mobile Vit Block 中采用相比于传统卷积更加轻量化的 Transformer 进行全局建模；

(2) 采用 Mobile Vit 作为主干网络，构建双流模型将 Mobile Vit 中的轻量化方法从图像数据迁移到视

频数据；

(3) 剪枝和作为先进的轻量化方法在确保模型的故障诊断进度的前提下可以有效的降低模型的大小。

下文可以被总结为如下几个部分:第二部分介绍了相关工作；第三部分详细介绍双流 Mobile Vit 模型的结构、组成和所用到的轻量化方法；第四部分介绍和分析所用的数据集以及实验结果；第五部分给出本研究的总结和展望。

2. 相关工作

在过去十年，深度学习不仅在图像任务上取得了长足的发展，在视频任务上也在不断突破。在视频任务中，研究人员起初将图像中的方法用到视频上，Karpathy [6]将视频中的每一帧提取出来，再用 2DCNN 方法对视频分类。为了得到视频的动态特征，Simonyan [7]等人提出了 Two-Stream Convolutional Networks，利用 CNNS 同时对视频的对视频每一帧 RGB 图像和光流图进行特征提取用于视频分类。Wang L [8]在发现了从视频中间隔式提取视频帧的数据处理模式对时间维度建模和对分类结果有着很大影响的基础上提出 TSN。为了简化模型和训练效率，Du Tran [9]在 2D 卷积核的基础上增加了时间维度构建 3D 卷积提取时空特征。

由于现有移动诊断设备无法支撑模型的落地使用，因此对基于深度学习的故障诊断模型轻量化操作始终是研究热点。深度学习模型轻量化主要有模型设计和模型压缩两个方面构成。在模型设计方面：Andrew [10]等人利用深度可分离卷积来构建量级卷积神经网络；Zhang [11]等人为提高计算效率，引入通道混洗和逐点分组卷积；Forrest N. Iandola [12]等人在 Squeeze Net 中提出 Fire Module 以压缩模型体量；Alexey DosoVitskiy [13]等人将 NLP 中的 Transformer 思想迁移到图像数据上构建全局学习模型 Vit 替代 CNN; Sachin Mehta [14]等人在 Vit 的基础上引入 unfolding operation 并依赖 CNNs 形成超量化模型 Mobile Vit。在模型压缩方面：Hinton [14]等人将一个复杂的、训练良好的复杂神经网络(教师网络)的知识蒸馏提取出来，然后用一个简单的神经网络(学生网络)来学习这个知识，以达到和复杂网络相似的性能。随后，Zagoruyko [15]等人通过正确定义卷积神经网络的注意力，迫使学生 CNN 网络模仿强大的教师网络的注意力图，从而显著提高学生 CNN 网络的性能。Liu [16]等人提出了通道剪枝策略以降低模型复杂度，通道重要性的评价标准使用的是 Batch Normalization 层中的缩放因子。这两种轻量化方法各有优缺点，设计较少可训练参数的模型可以更深度学习数据特征，但在降低模型体量上不具备优势。模型压缩可以大幅度地降低模型体量，但不能学习更深层数据特征。

在本研究中，我们主要设计了适用于工业视频数据分类模型达到对工业过程实施故障诊断的目的。如图 1 所示，本文提出了基于工业过程视频双流模型用于故障诊断的双流 Mobile Vit，该模型由两条完全相同的主干网络 Mobile Vit 和 Convolution Attention Fusion Mechanism (CAFM) [17]构成。为使模型更加轻量化，权重剪枝技术被用来剪去双流 Mobile Vit 中卷积层和全连接层中根据剪枝标准不重要的权重，以达到进一步降低模型大小的目的。

3. 基本原理

3.1. 双流 Mobile Vit 网络结构

本研究的主要目的在于解决基于过程数据和图像数据的工业过程故障诊断模型诊断精度较低以及模型体量过大无法被部署到实际工业生产过程中。我们设计了轻量化双流并行网络双流 Mobile Vit 提取视频的时序特征和空间特征，如图 1 所示，我们的双流 Mobile Vit 除了视频预处理模块由两条不仅在结构上且在参数上完全相同的主干 2D 轻量化网络 Mobile Vit 构成。一条主干用于提取视频的空间特征，另一条用于提取视频的时间特征，并且在两条主干网络的末端使用轻量化卷积注意力机制 CAFM 按照重要程

度进行充分融合。在整个网络模型，由 Mobile Vit block、MV2 [18]组成的 Mobile Vit 作为模型的主干网络；在融合层中，主要用了分组卷积和激活函数。除以上介绍的模块，如图 7 所示的结构化通道剪枝将用来剪去双流 Mobile Vit 中卷积层和全连接层中相对不重要的权重。

为了让输入符合模型结构，首先提取视频帧并形成光流图。对于一个视频数据 $V \in \mathbb{R}^{T\times C\times H\times W}$，其中 T 代表一个视频中的帧数，C，H，W 分别代表一帧 RGB 图像的通道数，高度和宽度。经过预处理，一个视频转变成 $0.2\times T$ 张形状为 $C\times H\times W$ 的 RGB 图像和 $0.2\times T$ 张形状相同的光流图。

视频特征主要体现在时序特征和空间特征两个维度上，时间特征通过视频中相邻的两帧随着时间变化生成的光流图来表征，空间特征即为视频中每一帧的 RGB 特征。为了确保两个维度上特征的提取同步性,双流并行网络可以提取完整的时空特征。

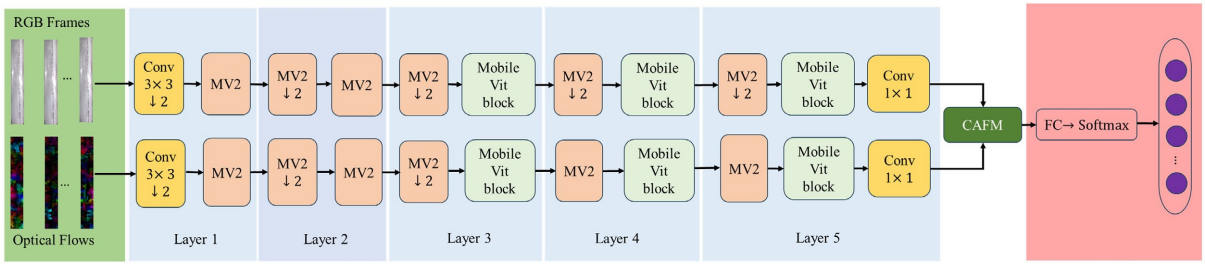


Figure 1. Entire diagram of two-stream Mobile Vit
图 1. 双流 Mobile Vit 的完整结构图

3.1.1. Mobile Vit Block

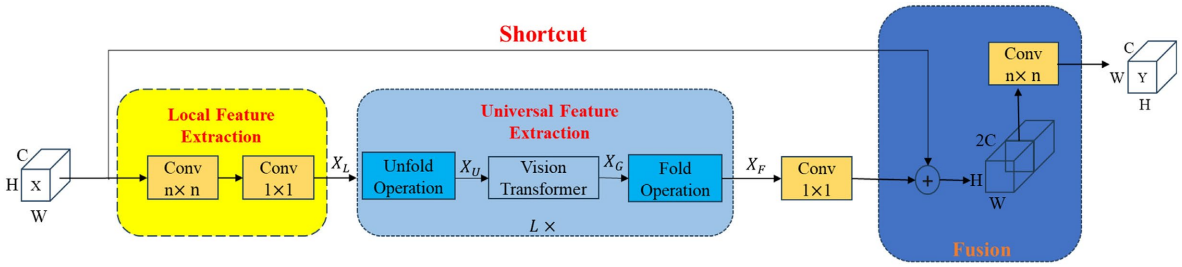


Figure 2. Diagram of Mobile Vit block
图 2. Mobile Vit 块结构图

如图 2 所示，Mobile Vit Block 由 Local Feature Extraction 模块、Universal Features 模块构成和 Fusion 构成, 分别用于提取输入 X 的局部特征和全局特征以及保证工业过程视频上数据的完整性。下面分别对以上每个模块详细介绍。

(1) 在 Local Feature Extraction 中,先用 $n\times n$ 的标准卷积提取输入 $\mathbb{R}^{C\times H\times W}$ 的局部特征，一个逐点卷积随后将局部特征映射到更高维的 d 维空间。

(2) 在 Universal Feature Extraction 中，拥有更广阔视野域的 Vision Transformer 被用来进行全局特征建模。为降低 Transformer 中自注意力的运算成本和利用空间归纳偏差学习图像的全局特征，在全局特征建模之前，首先将局部特征 X_L 经过 Unfolding operation 得到 X_U。Unfolding operation (如图 3 所示)可以被分为两步：第一步将输入图像打成 N 块大小为 $h\times w$ 的 patch；第二步将每个 patch 中像素展平并在通道维度上拼接。如图 4 所示，Vision Transformer 作为特征提取的核心。整个学习过程如下：为降低模型发生过拟合的可能性和提升收敛速度，采用 layer norm 机制对输入 unfolded patches 归一化。随后，如图

4 所示的 Multiple Head Attention 对归一化数据进行多头注意力计算，具体计算过程如式(1)所示，其中 h_n 为注意力机制的头数。

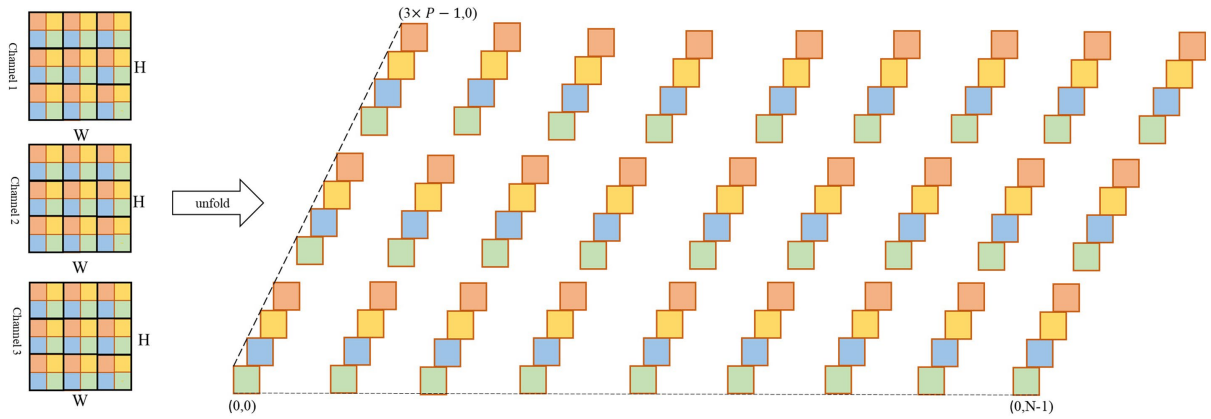


Figure 3. Flowchart of Unfold operation
图 3. Unfold 操作流程图

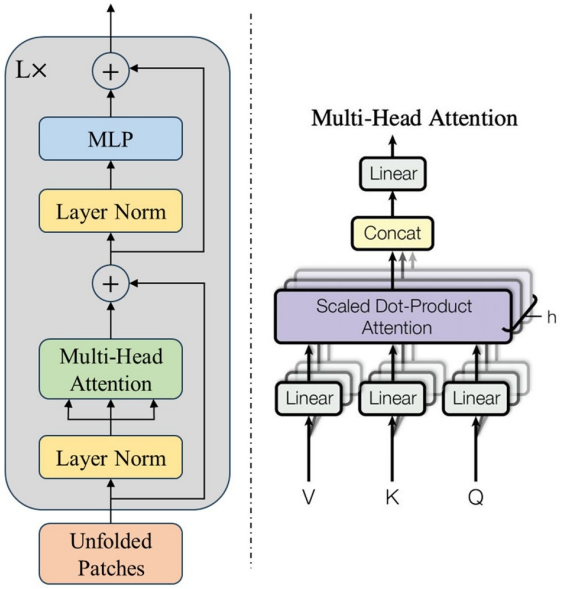


Figure 4. Structure diagram of MHA
图 4. MHA 结构图

$$\begin{cases} q^{ih} = W^{qh} \times W^{q} \times \alpha_i \\ k^{ih} = W^{kh} \times W^{k} \times \alpha_i,\ i = \{1,2,3,\cdots,3\times P-1\},\ h = \{1,2,3,\cdots,h_n\} \\ v^{ih} = W^{vh} \times W^{v} \times \alpha_i \end{cases} \tag{1}$$

经过两次投影后，scaled Dot-Product Attention 利用每个头的 q^{ih}， k^{ih}， v^{ih} 经行注意力计算，计算式如下： $\text{MultiHead} = \text{Concat}\left(\text{head}_1,\cdots,\text{head}_{h_n}\right)W^{O}$, $\left(W^{O} \in \mathbb{R}^{h_n d_k \times d_m}\right)$

$$\text{head}_h = \text{softmax}\left(\frac{q^{ih}\left(k^{ih}\right)^{\mathrm{T}}}{\sqrt{d_k}}\right)v^{ih} \tag{2}$$

为了确保模型在不同位置共同关注来自不同表征子空间特征，MHA 可以将多个头拼接。Multi-Head Attention 利用参数矩阵 W^O 降低每个头数维度以确保总的运算复杂度与单头注意力保持相同[13]，完整计算过程如下式所示：

$$\text{MultiHead} = \text{Concat}\left(\text{head}_1, \cdots, \text{head}_{h_n}\right) W^O, \ \left(W^O \in \mathbb{R}^{h_n d_k \times d_m}\right) \tag{3}$$

3.1.2. MV2 与 MV2

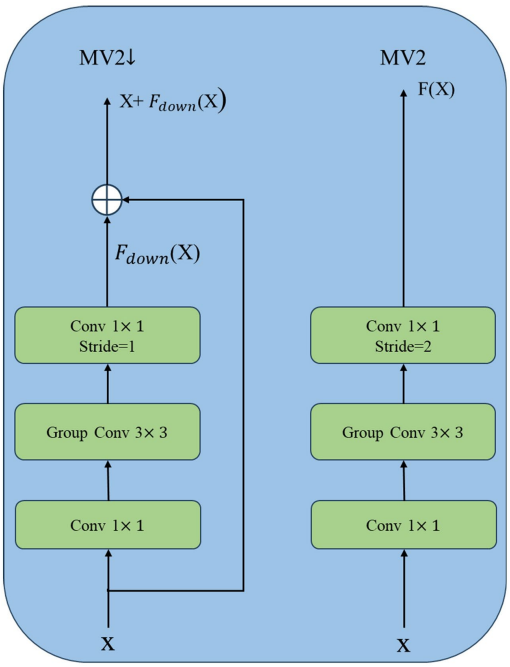


Figure 5. Structure diagram of MV2 and MV2↓
图 5. MV2 和 MV2↓结构图

MV2 指 MobileNetV2，是一种轻量化卷积神的经网络，重要作用是辅助 Mobile Vit Block 特征学习。本网络使用了图 5 所示两种结构，一种是 MV2↓倒残差结构，一种是直流结构 MV2。MV2↓中“↓”表明 MV2↓中使用步长为 1 逐点卷积 Conv 1 × 1 对特征图进行下采样。MV2↓中的倒残差结构不仅可以缓解模型中 CNN 部分梯度消失问题，还可以通过直接将输入 X 直接传输到输出来保护工业过程特征的完整性。

3.1.3. Convolution Attention Fusion Mechanism

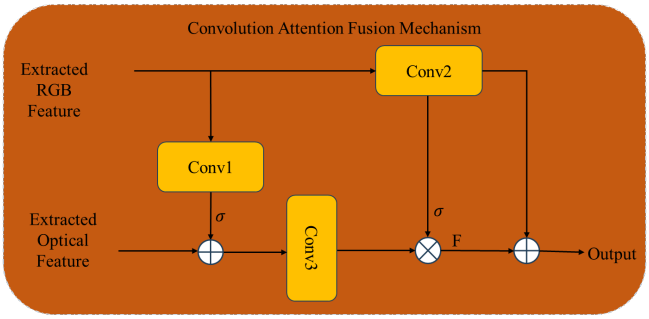


Figure 6. Structure diagram of CAFM
图 6. CAFM 结构图

本研究中提出的模型由两条主干网络构成的工业视频分类网络，分别用于提取视频的空间(RGB 特征)特征和时间特征(Optical 特征)。有效的融合两种不同维度的特征对模型的鲁棒性和提升分类准确率有着重要意义，我们提出如图 6 所示的由卷积和激活函 σ 构成 Convolution Attention Fusion Mechanism。卷积通过自身局部监督和激活函数的归一化功能生成注意力图，使有效特征能充分流入到下一阶段。注意力的引导特性在特征传播过程中具有降低数据冗余的作用。

3.2. 结构化剪枝

结构化剪枝是一种模型压缩技术，目的是在减少模型计算量和存储需求的同时，尽可能保持模型的性能。它通过剪除网络中的整个结构单元，使剪枝后的网络仍具有规则的结构，从而便于高效推理和硬件加速[19]。其中，通道剪枝是结构化剪枝的一种常用方法，专注于剪掉冗余的通道以减少模型参数量和计算复杂度。通道剪枝的主要步骤为依次进行通道重要性评估、剪枝决策、通道裁剪，模型微调。通道重要性评估一般是通过计算每个通道的重要性得分。剪枝决策是通过设定阈值或剪枝率裁决定重要性较低的通道。通道裁剪是剪去重要性较低的通道。微调是将被剪枝后的模型再次训练实现更好的性能。本文中采用的通道剪枝方法是通过 L1 范数判定整个模型中卷积层和全连接层中每个通道权重的重要性得分，再对重要性得分排序并设定剪枝率决定被保留的通道权重。剪枝具体方法如图 7 所示，其中 h_c、λ 分别代表原有的权重通道数和剪枝率，$[\cdot]$ 代表取整函数。

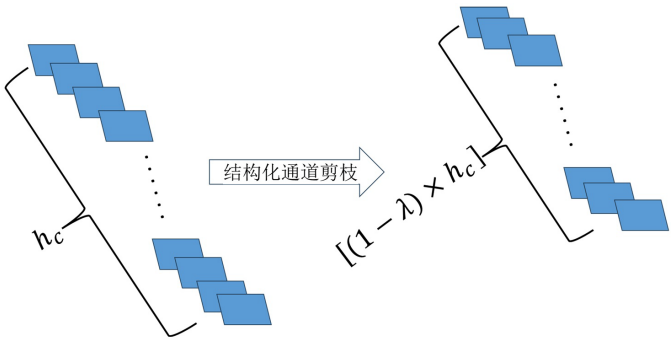


Figure 7. Structural channel pruning
图 7. 结构化通道剪枝

4. 数据集与实验

4.1. PRONTO 视频数据集与数据增强

本研究采用的是 PRONTO [20]基准视频数据集，该视频数据集收集自克兰菲尔德大学过程系统工程实验室的全自动、高压、多相流设备。该设施为研究多相流的输送、测量和控制而设计的，允许对包括水、空气和油在内的多相流进行研究。该设施，描述了不同操作条件下的测试和诱发的故障。

本研究使用的视频数据集是根据 PRONTO 基准数据通过人为仿真三种系统状态并记录得到的，该数据集中包含 3 种系统状态，分别是：正常、空气泄露以及分流，其中空气泄露和分流是通过人为操作引发的故障。本实验将视频按照 7:3 的比例划分训练集和验证集。

考虑到数据来源的单一性，对数据进行增强及其重要。数据增强部分由图像随机裁剪，图像大小随机调整，图像随机垂直翻转和归一化三大部分组成。图像随机裁剪，图像大小随机调整，图像随机垂直翻转能提高数据的多样性以改善模型的泛化能力。对图像数据的归一化不仅可以降低模型过拟合的可能性，还可以降低数据分布范围广而导致溢出问题的可能性。

4.2. 实验

本实验是建立在 Python 3.7 and PyTorch 1.7 环境之上，cuda 版本为 11.1，操作系统为 Ubuntu 22。由两张 NVIDIA GeForce RTX 3090 GPU 驱动代码执行，超参数配置如表 1 所示。

Table 1. Configuration of hyper parameter
表 1. 超参数配置

参数	参数值
批量大小	8
训练轮数	150
学习率	0.0001
损失函数	交叉熵损失函数
优化器	Adam [21]
权重衰减率	0.001
剪枝率	0.5

我们采取准确率，精度，召回率和 F_1Score 四种性能指标来全面分析并评估提出的模型的故障诊断能力，采用准确率表示模型正确分类的样本占总样本数的比例；精确率是在预测为正例的样本中，真正是正例的比例；召回率是在真实正例样本中，被正确预测为正例的比例；精确率和召回率的调和平均值，用于综合衡量分类模型的性能。四种指标的计算公式如式(4)所示。

$$\begin{cases} Accuracy = \dfrac{TP+TN}{TP+TN+FN+FT} \\ \Pr ecision = \dfrac{TP}{TP+FP} \\ \operatorname{Re} call = \dfrac{TP}{TP+FP} \\ F_1 score = \dfrac{2\times \Pr ecision\times \operatorname{Re} call}{\Pr ecision+\operatorname{Re} call} = \dfrac{2TP}{2TP+FP+FN} \end{cases} \tag{4}$$

TP 代表模型正确地将属于某个类别的样本分类为该类别的样本数。*TN* 表示模型正确地将不属于某个类别的样本分类为其他类别的样本数。*FP* 代表模型错误地将不属于某个类别的样本分类为该类别的样本数；*FN* 表示模型错误地将属于某个类别的样本分类为其他类别的样本数。

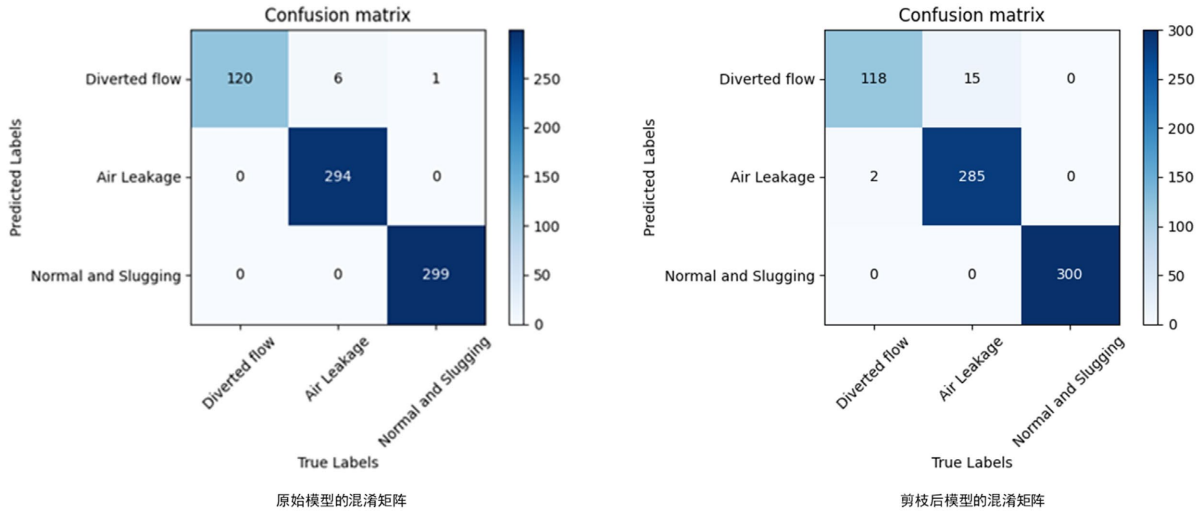
除上述四种性能指标以外，*Paras* 将被采用衡量模型的大小。*Paras* 这一指标表明整个模型中所包含的参数的数量。

4.2.1. 不同故障诊断模型之间的比较

由表 2 可得，双流 Mobile Vit 取得了最好的故障诊断性能。由于双流 Mobile Vit 中使用的是双流并行架构以及 Mobile Vit Block 中采用先局部特征提取后全局特征提取的顺序，为此取得了最高的故障诊断精度。在模型大小上，双流 Mobile Vit 中 Mobile Vit Block 中在对全局特征学习前，首先对特征进行 unfold 操作，用过降低了运算复杂度来减小参数量。为进一步降低所提出来的模型对的大小，结构化通道剪枝这一模型轻量化方法被用来降低模型的参数量。从实验结果来看，虽然被剪枝后的模型的故障诊断准确率比原来模型的准确率低了两个百分点左右，但是依然优于其他三个模型效果。在模型大小上，结构化剪枝方法使模型的参数量大约降低了百分之三十。

Table 2. Diagnosis results of the different model based video data
表 2. 不同基于视频数据的故障诊断模型的诊断结果

模型	*Accuracy*	*Precision*	*Recall*	*F1score*	*Paras*
C3D	73.26	73.15	75.33	74.32	79.92M
Two Stream Swinc Transformer	95.26	95.68	95.06	95.73	32.58M
双流 Mobile Vit	99.02	99.63	99.72	99.82	3.25M
剪枝后的双流 Mobile Vit	97.36	99.56	98.32	99.23	2.13M



原始模型的混淆矩阵　　剪枝后模型的混淆矩阵

Figure 8. Comparison between confusion matrixes of the original proposed model and the pruned proposed model
图 8. 原始模型混淆矩阵与剪枝后模型混淆矩阵对比

4.2.2. 消融实验

本实验是为了探究 Mobile Vit Block 中 shortcut 连接对工业过程特征的完整性是否有保护作用。为验证其作用，本实验采取对比验证方式，一个实验结果是整个被提出来模型的 Mobile Vit block 带有中 shortcut 连接，另一个实验模型中不带有 shortcut 连接。实验结果如表 3 所示，模型在使用带有 shortcut 连接的情况下，故障诊断准确率要比不使用 shortcut 连接的准确率高出 5 个百分点。以上结果表明在提取工业过程视频特征的过程中，shortcut 连接对工业过程特征的完整性有着显著的保护性。

Table 3. The impact of Shortcuts on the integrity of industrial process video data
表 3. Shortcut 对工业过程视频数据完整性的影响

模型中 Mobile Vit block 是否带有 Shortcut 链接	*Accuracy*
Yes	99.02
No	94.44

5. 结论与总结

本文提出了一种基于轻量化 Vision Transformer 模型 Mobile Vit 的双流工业过程视频故障诊断模型 Two Stream Mobile Vit 来解决基于过程数据和图像数据的故障诊断精度较低和模型无法被应用到实际工业生产中的双重问题。首先，利用对视频数据进行预处理，预处理的过程包括提取视频帧和计算稠密光

流形成光流图，视频帧和稠密光流分别作为工业过程的时序特征和空间特征。第二步，利用 Mobile Vit 提取时空特征。第三步利用 CAFM 融合时空特征，最终分类器利用融合后的特征对工业过程视频数据分类实现状态诊断。从实验结果来看，本文中所提出的模型在视频数据上取得的准确率比其他模型都高。另一方面在取得较高准确率的同时，参数极大地降低。为了更大幅度地降低模型大小，结构化剪枝技术降低了原模型百分之三十的参数量，并且剪枝后的模型各项指标也高于其他模型。除此之外，Mobile Vit block 中 Shortcut 连接可以有效保护工业过程的完整性。未来的工作可以将 Mobile Vit 的思想直接从二维数据迁移到三维数据上，这样模型结构会更加简单。

基金项目

本工作受国家自然科学基金(61903251)资助。

参考文献

[1] Garcia-Alvarez, D., Bregon, A., Pulido, B. and Alonso-Gonzalez, C.J. (2023) Integrating PCA and Structural Model Decomposition to Improve Fault Monitoring and Diagnosis with Varying Operation Points. *Engineering Applications of Artificial Intelligence*, **122**, Article 106145. https://doi.org/10.1016/j.engappai.2023.106145

[2] Lian, R., Xu, Z. and Lu, J. (2013) Online Fault Diagnosis for Hydraulic Disc Brake System Using Feature Extracted from Model and an SVM Classifier. 2013 *Chinese Automation Congress*, Changsha, 7-8 November 2013, 228-232. https://doi.org/10.1109/cac.2013.6775733

[3] Wang, Y. and Liu, H. (2019) Centrifugal Pump Fault Diagnosis Based on MEEMD-PE Time-Frequency Information Entropy and Random Forest. 2019 *CAA Symposium on Fault Detection, Supervision and Safety for Technical Processes* (*SAFEPROCESS*), Xiamen, 5-7 July 2019, 932-937. https://doi.org/10.1109/safeprocess45799.2019.9213261

[4] Davari, N., Akbarizadeh, G. and Mashhour, E. (2021) Intelligent Diagnosis of Incipient Fault in Power Distribution Lines Based on Corona Detection in UV-Visible Videos. *IEEE Transactions on Power Delivery*, **36**, 3640-3648. https://doi.org/10.1109/tpwrd.2020.3046161

[5] 徐磊, 田颖. 基于双流 Swinc Transformer 的工业过程故障诊断[J]. 建模与仿真, 2023, 12(2): 777-785.

[6] Karpathy, A., Toderici, G., Shetty, S., Leung, T., Sukthankar, R. and Fei-Fei, L. (2014) Large-Scale Video Classification with Convolutional Neural Networks. 2014 *IEEE Conference on Computer Vision and Pattern Recognition*, Columbus, 23-28 June 2014, 1725-1732. https://doi.org/10.1109/cvpr.2014.223

[7] Simonyan, K. and Zisserman, A. (2014) Two-Stream Convolutional Networks for Action Recognition in Videos. *Advances in Neural Information Processing Systems*, **27**, 568-575.

[8] Wang, L., Xiong, Y., Wang, Z., Qiao, Y., Lin, D., Tang, X., *et al*. (2016) Temporal Segment Networks: Towards Good Practices for Deep Action Recognition. In: *Lecture Notes in Computer Science*, Springer, 20-36. https://doi.org/10.1007/978-3-319-46484-8_2

[9] Tran, D., Bourdev, L., Fergus, R., Torresani, L. and Paluri, M. (2015) Learning Spatiotemporal Features with 3D Convolutional Networks. 2015 *IEEE International Conference on Computer Vision* (*ICCV*), Santiago, 7-13 December 2015, 4489-4497. https://doi.org/10.1109/iccv.2015.510

[10] Howard, A.G., Zhu, M., Chen, B., *et al*. (2017) Mobile-Nets: Efficient Convolutional Neural Networks for Mobile Vision Applications.

[11] Zhang, H., Hao, Y. and Ngo, C. (2021) Token Shift Transformer for Video Classification. *Proceedings of the* 29*th ACM International Conference on Multimedia*, China, 20-24 October 2021, 917-925. https://doi.org/10.1145/3474085.3475272

[12] Iandola, F.N., Han, S., Moskewicz, M.W., *et al*. (2016) SqueezeNet: AlexNet-Level Accuracy with 50x Fewer Parameters and <0.5 MB Model Size.

[13] Dosovitskiy, A., Beyer, L., Kolesnikov, A., *et al*. (2020) An Image Is Worth 16x16 Words: Transformers for Image Recognition at Scale.

[14] Mehta, S. and Rastegari, M. (2021) MobileViT: Light-Weight, General-Purpose, and Mobile-Friendly Vision Transformer.

[15] Hinton, G., Vinyals, O. and Dean, J. (2015) Distilling the Knowledge in a Neural Network.

[16] Zagoruyko, S. and Komodakis, N. (2016) Paying More Attention to Attention: Improving the Performance of Convolutional Neural Networks via Attention Transfer.

[17] 徐磊. 基于多源异构信息的工业过程故障诊断策略研究[D]: [硕士学位论文]. 上海: 上海理工大学, 2023.

[18] Sandler, M., Howard, A., Zhu, M., Zhmoginov, A. and Chen, L. (2018). MobileNetV2: Inverted Residuals and Linear Bottlenecks. 2018 *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Salt Lake City, 18-23 June 2018, 4510-4520. https://doi.org/10.1109/cvpr.2018.00474

[19] Zhu, M.J., Tang, Y.H. and Han, K. (2019) Vision Transfomer Pruning.

[20] Stief, A., Tan, R., Cao, Y., Ottewill, J.R., Thornhill, N.F. and Baranowski, J. (2019) A Heterogeneous Benchmark Dataset for Data Analytics: Multiphase Flow Facility Case Study. *Journal of Process Control*, **79**, 41-55. https://doi.org/10.1016/j.jprocont.2019.04.009

[21] Kingma, D. and Ba, J. (2014) Adam: A Method for Stochastic Optimization. arXiv:1412.6980. https://arxiv.org/abs/1412.6980