DOI: 10.13957/j.cnki.tcxb.2024.01.019

引文格式：

朱永红，戴晨雨，李蔓华，等. 基于深度学习与信息融合技术的陶瓷辊道窑温度智能检测方法[J]. 陶瓷学报，2024, 45(1): 180–190.
ZHU Yonghong, DAI Chenyu, LI Manhua, et al. Intelligent temperature detection method of ceramic roller kiln based on deep learning and information fusion technology [J]. Journal of Ceramics, 2024, 45(1): 180–190.

基于深度学习与信息融合技术的陶瓷辊道窑温度智能检测方法

朱永红，戴晨雨，李蔓华

(景德镇陶瓷大学 机械电子工程学院，江西 景德镇 333403)

摘 要：目前，陶瓷辊道窑烧制温度主要通过热电偶检测。由于热电偶容易老化，造成温度检测精度逐渐变低，以致影响陶瓷制品烧成质量。针对此问题，提出将基于深度学习火焰图像特征识别与热电偶点检测数据融合代替热电偶的辊道窑温度智能检测方法。该方法是针对陶瓷辊道窑火焰图像采用一种基于移位窗口视觉自注意力机制的多尺度特征提取网络，利用卷积神经网络和 Transformer 分支的局部和远程特征来保留更多图像信息，获得较准确的火焰图像特征，并与热电偶点检测数据融合，从而实现对陶瓷辊道窑温度的精确检测。该多尺度特征提取网络模型首先是采用一种基于多层 Transformer 的自编码器网络提取出其中的浅层和多尺度深层特征，其次将多个特征融合到 Transformer 和卷积神经网络中使其获得足够的能力去捕捉特征信息，最后将热电偶点检测获得的数据输入前面的网络中，利用特征级信息融合来是实现火焰图像特征与关键点检测温度数据的融合。实验结果表明，本文提出的融合网路模型比基于卷积神经网络融合方法特征识别平均准确率提高了 1.75%，平均误差产生的减少了 2.67%，其大多数指标上优于卷积神经网络分支或 Transformer 分支图像融合，因而本文的方法有效可行。

关键词：陶瓷辊道窑；深度学习；信息融合；智能温度检测

中图分类号：TQ174.6^{+}53　　**文献标志码：**A　　**文章编号：**1000-2278(2024)01-0180-11

Intelligent Temperature Detection Method of Ceramic Roller Kiln Based on Deep Learning and Information Fusion Technology

ZHU Yonghong, DAI Chenyu, LI Manhua

(School of Mechanical and Electronic Engineering, Jingdezhen Ceramic University, Jingdezhen 333403, Jiangxi, China)

Abstract: At present, the firing temperature of ceramic roller kiln is mainly detected by thermocouples. Due to the easy aging of thermocouples, the temperature detection accuracy of thermocouples gradually becomes low so as to affect the firing quality of ceramic products. For this problem, an intelligent temperature detection method of fusing flame image feature recognition based on deep learning with thermocouple point detection data instead of thermocouple. This method is that a multi-scale feature extraction network based on the shift window visual self-attention mechanism is adopted for the flame image of ceramic roller kiln, convolutional neural network and local and remote features of transformer branch is used for retaining more image information to obtain more accurate flame image features which are fused with thermocouple point detection data, thus, the temperature of ceramic roller kiln can be accurately detected. In the multi-scale feature extraction network model, firstly, an auto-encoder network based on multi-layer transformer is used for extracting shallow and

收稿日期：2023−10−10。　　**修订日期：**2023−11−17。
基金项目：国家自然科学基金(62063010)；江西省重大科技研发专项(20214ABC28W003)。
通信联系人：朱永红(1965–)，男，博士，教授。

Received date: 2023−10−10.　　**Revised date:** 2023−11−17.
Correspondent author: ZHU Yonhong (1965–), Male, Ph.D., Professor.
E-mail: zyh_patrick@163.com

multi-scale deep features, and then multiple features are fused into transformer and convolutional neural network to make it be able to capture feature information, finally, the data obtained from the thermocouple point detection is input into the front network to achieve the fusion of the flame image features and the key point detection temperature data by the feature level information fusion. Experimental results show that the fusion network model proposed in this paper is 1.75% higher in average feature recognition accuracy and 2.67% lower in average error generation than the convolutional neural network fusion method, which is superior to the convolutional neural network branch or transformer branch image fusion in most indicators. Hence, the method proposed in the paper is effective and feasible.

Key words: ceramic roller kiln; deep learning; information fusion; intelligent temperature detection

0　引　言

辊道窑是以转动的辊子作为坯体运载工具的连续烧成隧道窑，按制品在窑内热处理不同也可将辊道窑分为三带：预热带、烧成带、冷却带[1]。温度是影响陶瓷辊道窑陶瓷制品烧成质量的重要因素。如果温度控制不当容易产生过烧或欠烧。过烧是指烧制温度长时间过高造成晶粒粗大，导致陶瓷制品开裂、变形等缺陷；欠烧就是烧制温度低于烧制要求的温度，会导致表面色泽往往不合要求，成瓷状况不好，致密度不足。这些问题会带来烧制过程的稳定性差和经济成本的增加。因此，如何提高辊道窑烧成带温度测量的准确性和可靠性一直是陶瓷产品优质生产的重要研究课题[2]。

目前，陶瓷辊道窑多采用热电偶对陶瓷制品烧制工况的温度进行检测，存在以下不足。首先，热电偶的直接输出电压经信号电路转换成可用的温度读数，若采样处理不当，会引入误差，导致测量精度下降；其次，热电偶由两种不同的金属构成，依赖于冷端温度的测量精度，当测量环境周围存在杂散或磁场时，可能会引起信号变化，抗噪性能差，其长时间使用易造成腐蚀，从而降低测量精度；最后，热电偶一般固定在窑壁处，只能检测窑壁附近的温度，属于温度点检测，对窑内温度信息掌握不全面，不能准确反映出产品烧成环境的温度。

目前，基于火焰图像的窑炉工况的监测和识别方法已经获得一些研究成果[3–4]。Zhou 等[3]设计了一份名为 BASIS 图像监视器的卷积神经网络，以从火焰形状中提取特征，并提前确定火焰的热声状态。该模型获得 99%的总体监测精度可以有效地监测双稳态区域外的燃烧不稳定性 CAM 的分布表明，可以准确地监测 BASIS 的燃烧状态。Zhang 等[4]提出了一种基于 Otsu Kmeans 火焰图像分割和支持向量机的水泥回转窑燃烧状态识别方法，利用视觉检测技术获得回转窑熟料燃烧核心区域的火焰图像，采用 Otsu Kmeans 图像分割方法实现火焰图像目标区域的有效分割。然而，上述文献采用基于火焰图像特征的研究方法，这些方法需要对图像的区域进行划分，但由于工业环境较为恶劣，多种因素会对采集的图像造成不小的影响，使得区域划分不准确，特征提取误差较大，影响窑炉工况识别精度。

为了解决图像采集时产生的误差问题，有些学者采用图像融合的方法来获得精确的数据。Wang 等[5]提出了一种多源信息融合特征来训练多标签深度强化学习(ML-DRL)模型，该方法结合 ML-DRL 理论和自我注意机制理论，提出了一种用于监测多数据源复合故障的 MSIF-DSARL 框架。Liu 等[6]提出了一种用于像素级图像融合的显著性引导深度学习框架，称为 SGFusion，它是一种端到端融合网络，可以通过训练一个模型来应用于各种融合任务，SGFusion 包含三个部分：双引导编码、图像重建解码和显著性检测解码。Chen 等[7]提出可学习图像卷积网络和特征融合(LGCN-FF)的端到端神经网络框架，其由两个模块组成：特征融合网络和可学习图卷积网络。该框架通过可学习的 GCN 和特征融合网络解决了多视图学习问题。

近几年，自注意力模型(Transformer)在自然语言处理领域主流地位的确立，越来越多的工作开始尝试将 Transformer 应用到计算机视觉领域中。Wang 等[8]使用了适用于像素级图像任务的 Transformer 模型，将深度金字塔卷积神经网络(Deep Pyramid CNN)与 Transformer 相结合，通过将初始分辨率调高，并逐层降低分辨率的方式，捕捉更细粒度信息。Carion 等[9]提出一种基于 Trasformer 的端到端物体检测模型，即 DERT 模型，一种基于 Transformer 模型和用于直接集预

测的二分匹配损失的目标检测系统的新设计，该方法在具有挑战性的 COCO 数据集上实现了与优化的 Faster R-CNN 相当的成果，而 DERT 易于实现并且具有灵活的架构，可轻松扩展到全景分割。

基于此，为了更好地精确地检测陶瓷辊道窑温度，本文研究一种基于移位窗口视觉自注意力(Swin Transformer)机制的多尺度特征提取网络，利用卷积神经网络和 Transformer 分支的局部和远程特征来保留更多图像信息，实现陶瓷辊道窑火焰图像特征准确提取，并与热电偶点检测数据融合，为实现陶瓷辊道窑温度精确检测提供技术支持。

1 基于 Swin Transformer 机制的多尺度特征提取网络

将基于 Swin Transformer 机制的多尺度特征提取网络应用于辊道窑火焰图像特征提取，以获得更精确的火焰特征，然后将火焰图像特征与热电偶点检测的信息融合以获得陶瓷辊道窑温度。辊道窑火焰图像是通过安装在窑壁上的工业相机经过冷却隔离透过耐高温玻璃监测的。

1.1 陶瓷辊道窑结构

辊道窑是新型快烧式连续式工业窑炉，其模型结构如图 1 所示。模型主要分为三个区域：预热带，其温度区间为 20 ℃ ~ 900 ℃，利用烧成带的热风将陶瓷砖坯进一步干燥，蒸发掉坯体内部的水分，长度约 24 节，在负压状态下进行操作。烧成带，其温度区间为 900 ℃ ~ 1230 ℃，通过电阻丝或燃烧喷嘴液化气加热，陶瓷坯的微观结构开始转变，长度取 23 节，每节长度 2.12 m，在微正压状态下进行。冷却带，其温度区间为 1230 ℃ ~ 70 ℃，正压，长度取 29 节。

1.2 移位窗口视觉自注意力模块

移位窗口视觉自注意力模块(Swin Transformer Block)建立与正常的 Transformer [10]一样需要先将图片进行预处理，其处理步骤与视觉自注意力模型(VIT)处理是一样的。如图 2 所示，在移动窗口视觉自注意力模块中块的大小为 P=4，因此得到的 $x_p \in N \times 48$，$N=HW/16=H/4 \times W/4$。它首先经过图像块(patch)拆分模块将输入 $H \times W \times 3$ 的 RGB 图像拆分为非重叠等尺寸的 $N \times (P^2 \times 3)$图像块，每个 $p^2 \times 3$ 图像块都被视为一个图像块序列，共拆分出 N 个。

一张 $x \in H \times W \times 3$ 的图片变成 $H/4 \times W/4 \times 48$ 的张量，成为了 $H/4 \times W/4$ 个图像块，每一个块是一个 48 维的词符(token)。接着要进行线性变换(Linear Embedding)对于每一个向量都做一个线性变换，变换后的维度为 C，得到的维度为 $H/4 \times W/4 \times C$。这些图像块词符(patch tokens)被导入若干具有改进自注意力的移位窗口视觉自注意力模块(Swin Transformer blocks)，第一个移动窗口视觉自注意力模块可以保持输入输出词符数值恒定为 $H/4 \times W/4$ 不变，且与线性嵌入层组成一个部分。词符通过图像块合并层来减少数量，产生层次化表示。首个图像块合并层联合了两组 2×2 相邻的图像块，原本的图像块词符变为原先的 1/4，即 $H/8 \times W/8$。而图像块词符的维度则扩大 4 倍，即 4C，接着对 4C 维的图像块拼接特征使用一个线性层，将输出维度降到 2C。使用移位窗口视觉自注意力模块进行特征转换，使其分辨率保持在 $H/8 \times W/8$ 不变。第一个图像块合并层与特征转换移动窗口视觉自注意力模块被视为第二部分。接着重复两次相同的过程，形成第三四两个部分，其输出分辨率为 $H/16 \times W/16$ 和 $H/32 \times W/32$。每一个阶段改变不同张量的维度，形成一种层次化的表征。

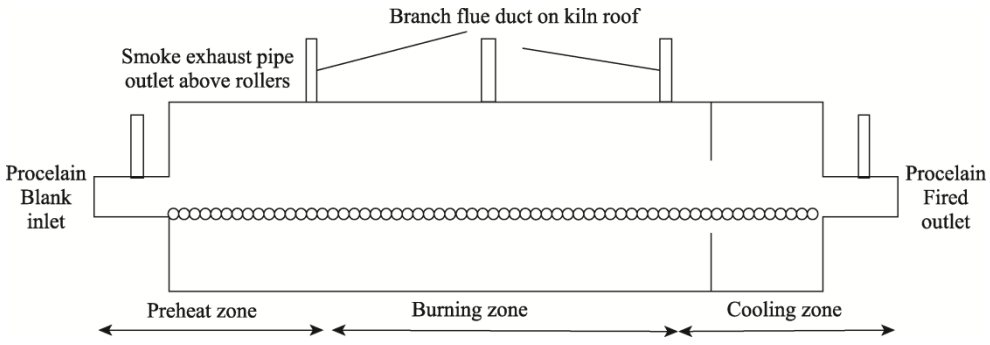

图 1 辊道窑结构图

Fig. 1 Structural diagram of roller kiln

1.3　基于 Swin Transformer 陶瓷辊道窑的火焰图像特征提取

为了能够准确测量火焰图像所反映的瓷坯附近的温度，选用 PTCR 系列的测温环来测量温度。测温环放置陶瓷制品附近的辊道上，经过一定时间的加热，通过工业摄像机拍摄火焰图像，然后将测温环取出。待测温环冷却后测量外径，对照温度换算表得到火焰图像对应的温度值。最后选取了 100 ℃ ~ 1230 ℃ 下的 20 多组图像数据和温度数据，共计 1000 余张火焰图像。不同温度下的火焰图像样本如图 3 所示。

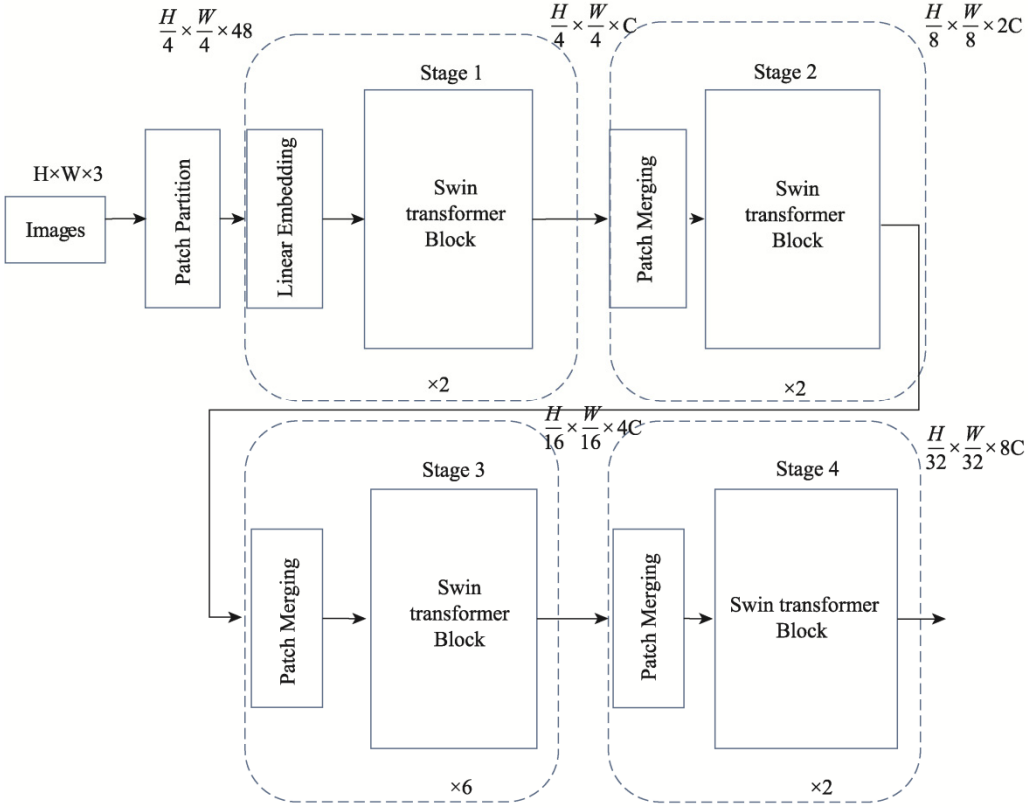

图 2　Swin Transformer 结构图

Fig. 2 Swin Transformer Structure Diagram

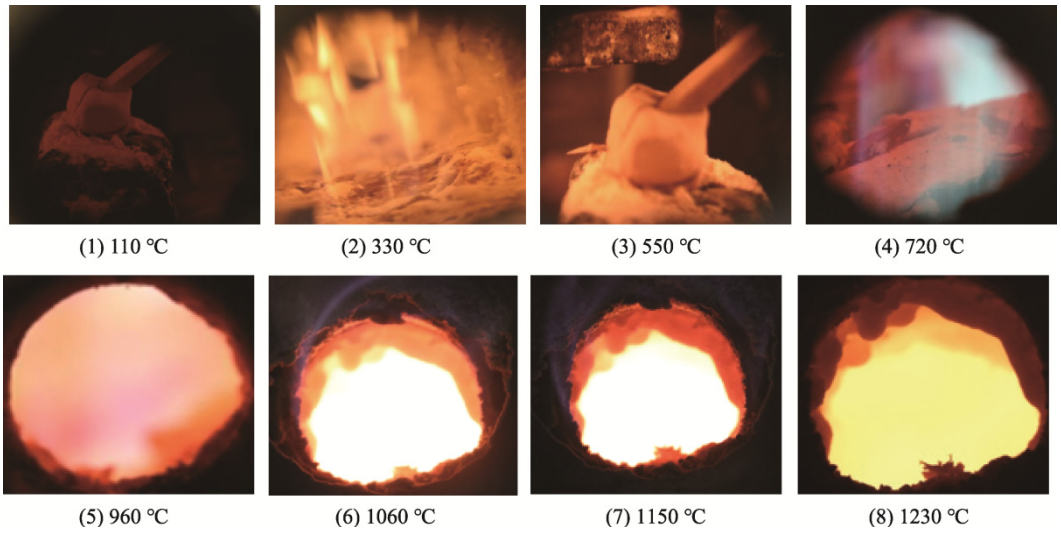

图 3　辊道窑不同温度下的火焰图像

Fig. 3 Flame images of roller kiln at different temperatures

1.4 火焰图像预处理采样过程

本段主要目标是在图像块集合 Patch Merging 层进行下采样，该模块的作用是做降采样，用于缩小分辨率，调整通道数进而形成层次化的设计，同时也能节省一定运算量，具体过程如图 4 所示。而在 CNN 中，则是在每个阶段开始前用的是步幅为 2 的卷积池化层来降低分辨率。Patch Merging 是一个类似于池化的操作，但是比池化操作复杂一些。池化会损失信息而 Patch Merging 不会。每次降采样是两倍，因此，在行方向和列方向上，按位置步幅为 2 选取元素，拼成新的图像块，再把所有图像块都连接起来作为一整个张量，最后展开。此时，通道维度会变成原先的 4 倍(因为 H, W 各缩小 2 倍)，再通过一个全连接层再调整通道维度为原来的两倍。

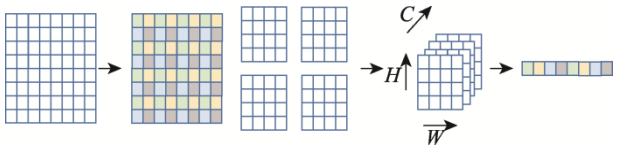

图 4 Patch Merging **降采样过程**
Fig. 4 Patch Merging downsampling process

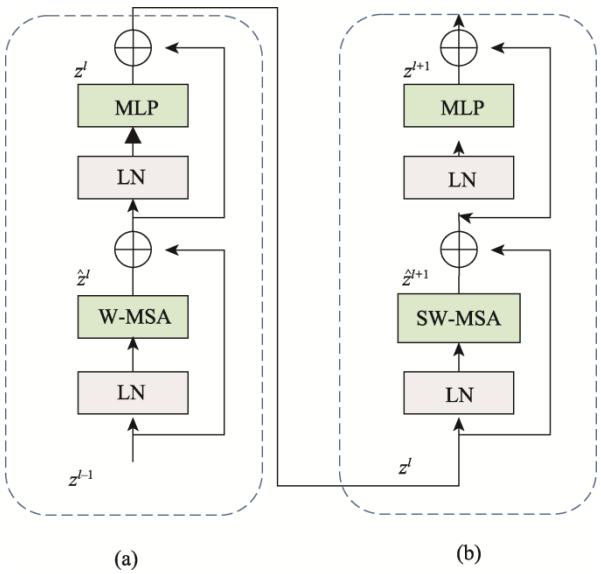

图 5 **两个连续的** Swin Transformer **模块：**
(a) **规则窗口** MSA；(b) **移位窗口** MSA
Fig. 5 Two consecutive swin transformer modules:
(a) Rule Window MSA, (b) Shift Window MSA

1.5 移位窗口视觉自注意力模块(Swin Transformer Block)

两个连续的移位窗口视觉自注意力模块(Swin Transformer Block)如图 5 所示，其中一个移位窗口视觉自注意力模块是由一个带两层多层感知机(MLP) [11]的基于移位窗口的多头自注意力模块(Shifted Window-based MSA)组成，另一个移位窗口视觉自注意力模块是由一个带两层多层感知机的基于窗口的多头自注意力模块(Window-based MSA)组成，在每个多头自注意力模块和每个多层感知机之前使用 LN 层，在多头自注意力模块和多层感知机之后使用残差连接。移位窗口视觉自注意力模块和 ViT 模块[12]的区别就在于将 ViT 的多头注意力机制 MSA 替换为了基于移位窗口的多头自注意力模块和基于窗口的多头自注意力模块其他部分都保持不变。

(1) 基于窗口的多头自注意力模块(Swin Transformer Block: Window-based MSA)

标准 VIT 的多头注意力机制 MSA 采用的是全局自注意力机制，即，计算每个词符和所有其他词符的注意力地图(Attention Map)。全局自注意力机制的计算复杂度是 $O(N^2d)$，该公式来源于 $Attention(Q,K,V)=\mathrm{Softmax}(QK^T/\sqrt{d}+B)V$，其中，$N$ 为词符的数量，d 为嵌入维度，全局自注意力机制的计算量复杂度与序列长度成平方关系。

W-MSA 与 MSA 的不同之处是在于，它是在一个个窗口中去计算自注意力，设定每个窗口中包含 $M\times M$ 个图像块，而 W-MSA 与 MSA 的计算公式为：

$$\begin{aligned}\Omega(\mathrm{MSA}) &= 4hwC^2+2(hw)^2C,\\ \Omega(\text{W-MSA}) &= 4hwC^2+2M^2hwC\end{aligned} \tag{1}$$

式中：h 为特征图高度，w 为特征图宽度，C 为特征图深度，M 为每个窗口的大小。MSA 具有 hw 个图像块，而每个图像块在全局计算 hw 次。W-MSA 则当 M 固定时(默认设为 7)具有线性复杂度(共 hw 个图像块，每个图像块在各自的局部窗口内计算 M^2 次)。hw 对于全局自注意力计算而言是难以承受的，而基于窗口的自注意力则具有良好的扩展性。由于窗口中的 patch 数量 M 远小于图片 patch 的数量 hw，基于窗口的多头自注意力模块的计算量与序列的长度 N=hw 成线性关系。

(2) 基于移位窗口的多头自注意力模块(Shifted Window-based MSA)

基于窗口的多头自注意力模块在节约了计算量的基础上减去了窗口之间关系的建模，这会导致缺乏信息交流影响了模型的表征。基于移位窗口的多头自注意力模块可以解决该问题，在两个连续的移位窗口视觉自注意力模块(Swin Transformer Block)中交替使用 W-MSA 和 SW-MSA。将 8×8 特征图均匀划分为 2×2 个大小

4×4 的窗口，将下面一层的基于移位窗口的多头自注意力模块的窗口位置进行移动，以此得到3×3 个不重合的图像块。

图 6 中两个连续块，其中，第 1 个块具有 4 个窗口，每个窗口是由 M×M=4×4 的图像块组成，第 2 个块是一样的大小但其窗口位置发生偏移。在新的窗口里面做自注意力操作，可以包含窗口的边界来实现建模。

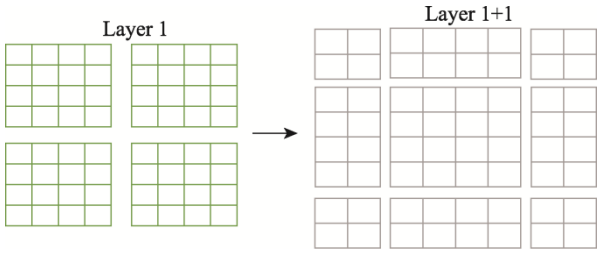

图 6　自注意移位窗口示意图
Fig. 6 Schematic diagram of self attention shift window

两个连续的移位窗口视觉自注意力模块表达式为：

$$\hat{z}^{l} = \text{W-MSA}(\text{LN}(z^{l-1})) + z^{l-1} \tag{2}$$

$$z^{l} = \text{MLP}(\text{LN}(\hat{z}^{l})) + \hat{z}^{l} \tag{3}$$

$$\hat{z}^{l+1} = \text{SW-MSA}(\text{LN}(\hat{z}^{l})) + \hat{z}^{l} \tag{4}$$

$$z^{l+1} = \text{MLP}(\text{LN}(\hat{z}^{l+1})) + \hat{z}^{l+1} \tag{5}$$

式中：$\hat{z}^{l}$ 和 z^{l} 分别表示第 1 个块的(S)W-MSA 模块输出特征和多层感知机模块输出特征。

(3) 移位偏置以及相对位置偏置

移位窗口的接入会导致窗口数量发生变化，需要将窗口做到大小相同。但这会导致窗口数量增加使得计算量增加，即从$[h/M] \times [w/M]$增加到$([h/M]+1) \times ([w/M]+1)$。窗口由特征图中不相邻的子窗口组成，屏蔽机制可以将自注意力机制限制在每个子窗口内。通过移位循环的方法使得处理窗口的数仍与规则内的窗口数相同，经过循环移位后，一个窗口内就可能包含了不同窗口内的内容，需要采用 masked MSA 机制将自注意力限制在子窗口内，如图 7 所示。

在计算自注意力时，要在计算相似度的过程中在每个 head 中加入相对偏置位：

$$B \in R^{M^2 \times M^2}$$

$$Attention(Q,K,V) = Soft\max(QK^{T}/\sqrt{d} + B)V \tag{6}$$

式中：$Q,K,V \in R^{M^2,d}$ 分别为 Q,K,V 矩阵；d 为 Q/V 维度；M^2 为窗口的 patches 数，沿各轴方向的相对位置处于$[-M+1, M-1]$范围内。这个时候需要参数化为一个更小尺寸的偏置矩阵 $\hat{B} \in R^{(2M-1)\times(2M-1)}$，且 B 中的值均取自 $\hat{B}$。

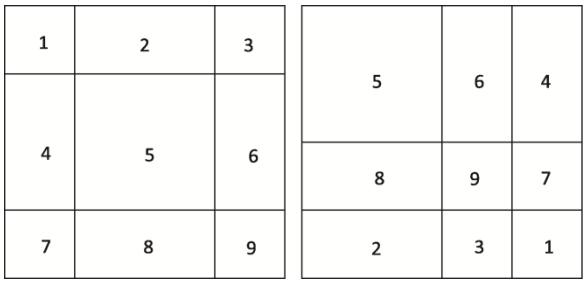

图 7　自注意力移位循环示意图
Fig. 7 Schematic diagram of self attention shift cycle

对于经过编码器获得的浅层特征和多尺度深层特征，分别使用多个图像融合模块(IFB)来进行融合，图像融合模块的构建所示，其中，使用瓶颈层和 3 个具有 3×3 核的卷积层来详细捕获局部特征，如图 8 所示。

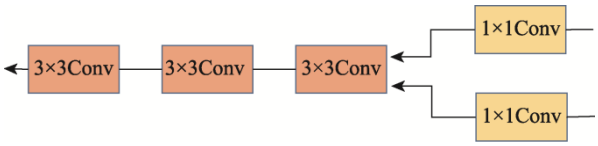

图 8　图像融合模块
Fig. 8 Image fusion module

它由 CNN 和 Transformer 两个分支组成。对于 Transformer 分支，模型采取 Swin Transformer，通过对其机制的远程依赖进行建模来学习全局上下文特性。对于 CNN 分支，用 3×3 卷积将两个特征结合起来，然后通过 3 个残差密集块学习局部残差特征，最后将残差密集块的输出连接起来进行特征融合。

2　陶瓷辊道窑火焰图像特征和点检测数据融合方法

2.1　陶瓷辊道窑火焰图像特征和点检测数据融合温度检测模型

陶瓷辊道窑的温度检测模型的整体框架，如图 9 所示。式(6)显示了每一个注意力模块对信息处理的过程，本文的特征级信息融合[13]就是将每一个注意力模块的输出结果进行融合的过程。在模型训练过程中，对多个并行的注意力层进行计算并获得了多个不同的 Z 矩阵，模型将通过多头自注意力函数进行拼接，并将拼接结果

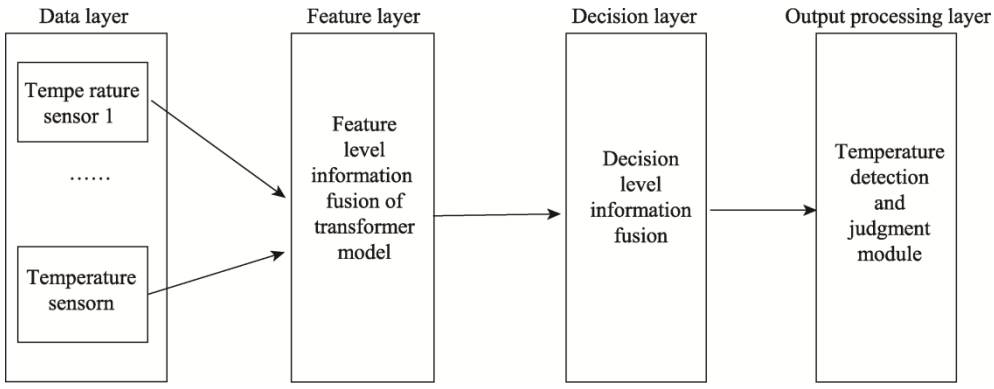

图 9　陶瓷辊道窑的温度检测模型的整体框架

Fig. 9 The overall framework of temperature detection model for ceramic roller kiln

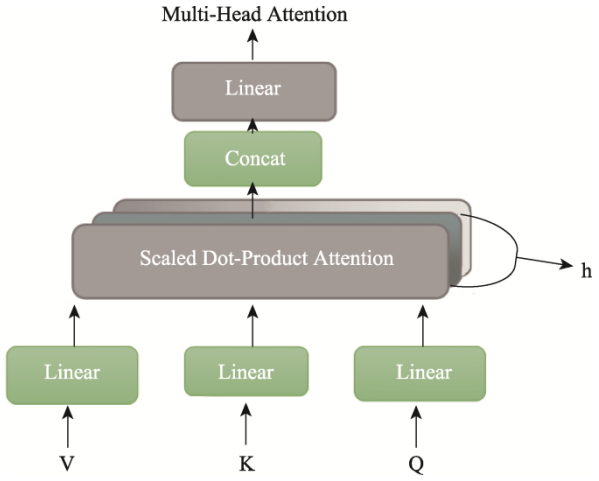

图 10　多头自注意力机制

Fig. 10 Multi head self attention mechanism

输入到线性模型之中，这样就完成了特征级的信息融合，即输入为高维传感器数据，输出为经过线性变换之后的矩阵，其包含的多头自注意力机制如图 10 所示。

在 D-S 理论中，决策级信息融合[14]是定义系统中所有可能出现的状态，并将每一个可能的假设根据其特征质量赋予一个概率。本文的决策级信息融合属于对多源语义数据进行融合，信息融合本质上是一种对传感器数据的融合机制，具体表现为每个节点只处理合成的语义信息。多源决策级语义信息融合通常包含两部分。首先，将多维传感器数据转换为语义信息；之后，通过训练为语义信息提供一个解释。本文的模型将温度传感器信息进行融合，而决策级信息融合则将经过特征及信息融合之后的结果进行解码，进而实现决策级的信息融合。在本模型中，将输入的高维数据经过归一化、线性变换、注意力变换等步骤进行处理和融合，最终输出一个布尔常量，用来标记当前输入数据是否为异常数据。该法具备一个明显优势，具体为不局限于多源异构传感器的数量和数据维度。

2.2　陶瓷辊道窑火焰图像特征和点检测关键数据信息融合算法

温度传感器传输的数据属于多变量时序数据，即输入数据为一个带有时间戳、大小为 N 的序列为：

$$N=\{x_1,x_2,\cdots,x_n\} \tag{7}$$

需要将多源数据传感器[15]进行归一化为：

$$x'_n=\frac{x_t-\min(N)}{\max(N)-\min(N)} \tag{8}$$

式中：max(N)和 min(N)为序列中的最大数据和最小数据，这样可以得到一种归一化后的数据序列 $X'=[x'_1,\ x'_2,\cdots,\ x'_n]$，将特征级信息融合以特征 Q、K、V 作为输入，通过多头自注意力机制进行融合。

本文提出的编码器部分可以提取输入图像的浅层特征和分层深层特征。对于输入图像先提取其特征 F_0，将其映射到高维空间，$H_{sf}()$ 为 3×3 的卷积层，用于浅层特征提取 $F_0\in R^{H\times W\times C}$，其表达式为：

$$F_0\in H_{sf}(I_{input}) \tag{9}$$

与一般 Transformer(如 VIT)的早期处理相比，使用卷积层可以实现更稳定的优化和更容易维度映射。提取多尺度特征 $F_i\in R^{H_i\times W_i\times C_i}$ 来源于 $F_0\in R^{H\times W\times C}$

$$F_i=H_{PM}(H_{STL_i}(F_{i-1})),i=1,2,3,4 \tag{10}$$

$$H_i,W_i,C_i=\frac{H}{2^i},\frac{W}{2^i},C\times 2^i \tag{11}$$

其中，$H_{STL_i}()$ 表示为第 i 层的 Swin Transformer，

H_{PM} 表示为 patch merge，它可以分层表示，通过 patch 池化层减少标签数量。它可以多次 Swin Transformer 和 Patch Merge，将 $H_{sf}()$ 得到的特征分解为多个不同深度的深度特征。

Swin Transformer 层是个多头自注意力 Transformer 层，其是基于局部关注和移位窗口机制发展起来的，对于输出向量其大小为(H×W×C)，其通过 M×M 本地窗口(合计共有 HWM^{-2} 个并且不重叠)转化为 $HWM^{-2}\times W^2\times C$ 特征。对于每个窗口进行相应的定期自我关注，而对于局部窗口 $F_{LW}\in R^{W^2\times C}$，通过局部窗口的自注意力机制计算数量为：

$$Q=F_{LW}P_Q, K=F_{LW}P_K, V=F_{LW}P_V \tag{12}$$

结合式(6)可得，P_Q，P_K，P_V 为不同窗口共享的投影数组，B 是具有学习能力的相对位置编码，d 是从 $Q,K,V\in R^{M^2,d}$ 的维数得到。通过并行执行注意函数 h，并将结果连接为多头自注意并使用具有两个完全聚合层(具有 GELU 非线性)的多层感知器(MLP)进一步转换特征。在使用剩余连接的 MSA 和 MLP 之前，需要进行 LayerNorm (LN)操作。整个操作过程可以描述为：

$$F_{LW}=\mathrm{MSA}(\mathrm{LN}(F_{LW}))+F_{LW} \tag{13}$$

$$F_{LW}=\mathrm{MLP}(\mathrm{LN}(F_{LW}))+F_{LW} \tag{14}$$

3　实验结果分析

3.1　初始训练

为了验证以上提出的基于 transformer 模型与信息融合的火焰图像识别效果，在实验中，所用硬件设备为 R5-5600x 处理器、RTX3080Ti 显卡、32G 内存，开发环境为 python3.9 和 pytorch0.11.0+ cuda113。

本实验模型训练平台采用基于 Python 编程语言的 PyTorch 深度学习框架。在实验中所有图像转换为灰度，并且缩放为 256×256。Bath size、epoch 和学习率设置为 4、4 和 1×10^{-4}。对于网格训练，从基础数据集中选取 400 对图像，通过裁剪、旋转、随机水平翻转、随机旋转等操作，将其转换为 4000 对灰度图像作为训练数据集。在第一阶段 Bath size、epoch 和学习率设置为 4、4 和 1×10^{-4}。

3.2　训练模型过程

编码器的特征提取能力以及解码器利用这些特征恢复图像的能力对图像融合效果非常重要。同时，由于使用了 Transformer 注意力机制，在模型的训练过程中需要更多的计算资源。因此采用两阶段训练方案来训练网络，以保证网络的性能。首先，将编码器和解码器组合成一个自编码器网络，用于重建输入图像。然后，在固定的编码器和解码器下，训练图像融合模块的融合浅层和多尺度的深层特征。

3.2.1　编码器以及解码器设置

首先，将编码器和解码器组合成一个自编码器；然后，对其进行训练。编码器被训练去拾取 4 个尺度特征(通过 Patch merge 进行降采样操作)，并且对解码器网络进行训练以重建具有多尺度深度特征的图像，使用以下损失函数训练自编码器网络 L_{first}：

$$L_{first}=L_{pix}+\alpha L_{SSIM} \tag{15}$$

式中：L_{pix} 和 L_{SSIM} 代表为像素损失和结构相似性损失，α 是参数 L_{pix} 和 L_{SSIM} 的影响因素，L_{pix} 主要作用是在重建图像时使用像素级的输入限制条件；L_{SSIM} 作用是限制输入输出之间的结构相似性。

$$L_{pix}=\left\|I_{output}-I_{input}\right\|_F^2 \tag{16}$$

$$L_{SSIM}=1-SSIM(I_{output}-I_{input}) \tag{17}$$

式中：I_{output} 和 I_{input} 分别代表为输入图像和输出图像了，$\left\|\bullet\right\|_F$ 代表弗罗贝尼乌斯范数，SSIM 代表不同图像的结构相似性。

3.2.2　融合模块设置

该融合方法可以做到保留结构的同时也有不错的前景和背景细节。融合网络的总体训练目标表示为 L_{fuse}：

$$L_{fuse}=\beta L_{feature}+L_{CC} \tag{18}$$

β 为平衡参数，而 $L_{feature}$ 是特征相似度损失其算法如下

$$L_{feature}=\sum_{m=1}^{4}\omega\left\|\phi_f^m-(\omega_1\phi_{I1}^m+\omega_2\phi_{I2}^m)\right\|_F^2 \tag{19}$$

式中：f、I_1、I_2 为融合后的图像，输入图像以及输出图像；ω、ω_1 和 ω_2 为平衡损失大小的三个权衡函数；$\Phi^m{}_f$ 是比例尺的融合特征图；而 Φ_{I1}^m 和 Φ_{I2}^m 分别是输入图像 1 和 2 的编码 m 尺度的特征映射。

互相关[16] (Cross-correlation)是一种常用的图像相似度度量方法，在 A 中首次被用作图像配准的损失函数，且通过 8×8 窗口来计算图像的相关性。使 $\overline{X}(p)$ 和 $\overline{F}(p)$ 分别表示源图像和融合图像的局部平均强度。在实验中，一般设为 n=8 来计算 n^2 区域中的平均值。

$$r_{xf}=\sum_{p\in\Omega}\frac{\sum_{P_i}(X(P_i)-\overline{X}(P))(F(P_i)-\overline{F}(P))}{\sqrt{(\sum_{P_i}(X(P_i)-\overline{X}(P))^2)(\sum_{P_i}(F(P_i)-\overline{F}(P))^2)}} \quad (20)$$

$$L_{CC}=1-\frac{1}{2}(r_{I_1f}+r_{I_2f}) \quad (21)$$

而 P_i 在 P 周围的 n^2 的区域内迭代。

3.3 训练结果

为了验证融合模型的效果，将本文方法与自适应稀疏表示(ASR)、者卷积稀疏表示(CSR)等常见图像融合方法的图像融合效果进行对比，并采用五个客观评价指标来比较融合结果。训练结果如下：

(1) 通过用于图像融合和去噪的 ASR 或 CSR 融合方法生成的融合图像具有较低的对比度以及其他影响因素。先使用拉普拉斯金字塔(The Laplacian pyramid)进行分解，再使用选择最多的规则来进行融合，得到融合的图像具有更多的图像信息，但也会导致其他图像丢失细节和形态信息。

(2) 与拉普拉斯金字塔融合算法一样，NSCT-SR(非下采样轮廓波)和 NSST-PAPCNN(非下采样剪切波)使用 MST 对源图像进行分解，但它们融合高频特征和基特征的策略不同。NSCT-SR 利用 NSCT 对源图像进行分解，并对低通带部分进行稀疏编码融合，但融合后的图像丢失了部分信息，获得了一定的平滑效果。NSST-PAPCNN 使用 NSST 对源图像进行分解，其中，PAPCNN 融合了细节纹理。利用该方法进行图像融合，可以得到对比度更强的多模态图像，但其周边细节模糊。

(3) 使用红外和可见图像的融合方法(DenseFuse)采用编码器-解码器网络结构，但融合后图像的精细纹理被平滑处理。

(4) 采用基于卷积神经网络的通用图像融合框架(IFCNN)选择元素最大化作为融合策略，利用该网络得到的多模态融合图像对比度较低且边缘的细节不够清晰。

(5) 通过用于医学图像融合的多尺度残差金字塔注意力网络(MSRPAN)采用三个 MSRPAN 块提取多尺度特征，并在融合过程中采用特征融合过程的特征能量比策略，但有一定的细节纹理损失。

(6) 采用无监督的增强医学图像融合网络(EMFusion)通过表层和深层约束来增强信息保存，而表层约束则基于显著性和丰度度量。在深度约束中，根据预训练编码器的唯一信道客观地定义唯一信息。它更好地保留了纹理细节和结构，但丢失了一些亮度信息。

所有融合图像的主观评价指标均值如表 1 所示，其中加粗代表最佳值。如表 1 所示，与其他几种融合方法相比，本文提出的融合方法获得了两个最佳值(CC, MI)、一个次优值(SSIM)和一个第三优值(PC)以及差异相关性的总和(SCD)。最大的 CC 和 MI 意味着该融合方法得到的图像与源图像的线性关系更强，并且能够保留源图像的更多信息。

相对较好的 PC 和 SSIM 表明，该方法可以更好地保留图像的显著信息和图像结构特征。该融合网路模型比基于卷积神经网络融合方法特征识别平均准确率提高了 1.75%，平均误差产生的减少了 2.67%，其大多数指标上优于卷积神经网络分支或 Transformer 分支图像融合，这说明网络在获取补充信息、保留源图像的更多信息和保留显著信息方面具有一定的优势。因此，该方法具有较好的融合性能，融合后的图像保留了更多的源图像信息。

3.4 消融实验

将本文提出的网络的编码器部分替换为 CNN 网络的编码器部分，其中，STL 被替换为两个 3×3 卷积层，Patch 被替换为 Max-Pooling。为了排除融合网络的影响，在两个网络中使用 SCA 作为融合策略。如表 2 所示，文中提出的编码器的 CC、PC、MI 和 SSIM 优于 CNN 编码器。这几个客观评价指标的良好表现表明，变压器编码器网络融合方法得到的融合图像可以保留更多的源图像信息。将此归因于 Transformer 自关注在全局特征提取方面优于卷积网络的能力。与使用 CNN 网络相比，虽然该方法需要更多的计算时间，但也可以获得更好的融合效果。总的来说，消融实验表明，本文提出的编码器网络结构在提取图像特征方面具有一定的优势，可以提高融合质量。

为了验证融合模块在网络中的重要性，需要选择了四种常见的融合规则或策略，并仅在 Transformer 和 CNN 融合分支上进行了实验。训练了用于提取多尺度特征并利用融合特征重构融合图像的编码器-解码器网络。通过 5 个质量指标对 CT 和 MRI 图像融合得到的多模态图像进行评价，而得到的索引值如表 3 所示提出的融合网络得到了 4 个最优值，这表明该策略能够更好地保留原始图像的细节和结构信息可以捕获局部和远程特征以理解整体表示可以实现更好的图像融合。

如表 3 所示，本文提出的融合方法通过捕获局部特征和远程特征，在主要指标上优于 CNN 分支或 Transformer 分支的图像融合方法。这证明本文方法可以将远程特征与局部特征相结合来改善图像融合效果。

表 1　几种方法在融合图像方面 5 种评价指标的平均值

Tab. 1 The average value of six evaluation indicators for several methods in fused images

	CC	PC	MI	SSIM	SCD
ASR	0.7846	0.3215	2.1562	1.4165	1.3015
CSR	0.7465	**0.3487**	2.3546	1.3824	1.2458
LP	0.7854	0.3315	1.8642	1.4264	1.4102
NSCT-SR	0.7881	0.3356	1.9245	1.3921	1.4395
NSST-PAPCNN	0.8146	0.2936	1.8223	1.3324	1.5124
DenseFuse	0.8334	0.3479	2.2976	1.0125	1.4231
IFCNN	0.8213	0.3045	1.9841	**1.4691**	**1.4929**
MSRPAN	0.8093	0.3321	2.1023	1.3956	1.4758
EMFusion	0.8348	0.3283	2.3652	1.4309	1.4539
Swin	**0.8594**	0.3420	**2.4158**	1.4672	1.4802

表 2　所提出的编码器的评估指标对比

Tab. 2 Comparison of evaluation indicators for the proposed encoder

Encoder	CC	PC	MI	SSIM	SCD	Time
CNN	0.8096	0.3189	2.2851	1.4621	1.4759	9.2154
Ours	0.8943	0.3516	2.4354	1.4834	1.4572	9.7359

表 3　评估数据集中不同的融合策略评估度量值的比较

Tab. 3 Comparison of evaluation metrics for different fusion strategies in the evaluation dataset

Specimen	CC	PC	MI	SSIM	SCD	Time
Add	0.8186	0.3125	1.9868	1.4465	1.5013	7.0621
Max	0.8266	0.3154	2.0123	1.4532	1.4516	7.0359
11-nom	0.8342	0.3678	2.1548	1.4580	1.4913	7.1546
SCA	0.8531	0.3493	2.4113	1.4571	1.4685	12.2513
Transformer-branch	0.8634	0.3368	2.4048	1.4586	1.4626	9.5821
CNN-branch	0.8512	0.3438	2.3658	1.4413	1.4713	9.3156
SWIN	0.8687	0.3416	2.4387	1.4680	1.4786	9.7695

4　结　论

本文针对目前陶瓷辊道窑温度检测热电偶容易老化造成检测精度低的问题，提出了基于计算机视觉火焰图像特征识别和热电偶点检测数据融合来检测温度方法。该方法是基于自注意力机制的 Swin Transformer 模型与信息融合技术相结合来完成对陶瓷辊道窑火焰图像特征数据与热电偶点检测数据融合，基于编码器—解码器的 Transformer 模型快速完成大规模高维传感器数据的训练和异常内容检测。实验结果表明，在融合多尺度特征的情况下，最终通过基于嵌套连接的网络重构成多模态图像，在特征提取阶段使用 Transformer 与特征级信息融合可以提取更多的图像特征信息，并且在融合过程中通过提取局部和远程信息来提高融合质量，从而实现对温度的较精确检测。总之，本文的方法为陶瓷辊道窑温度智能检测提供了一种新方法和新途径，对陶瓷辊道窑的改进和发展具有实际意义。

参考文献：

[1] 李杰, 朱永红, 罗添, 等. 辊道窑烧成带含坯体流场和温度场的数值模拟[J]. 中国陶瓷, 2022, 58(3): 45–50.
LI J, ZHU Y H, LUO T, et al. Chinese Ceramics, 2022, 58(3): 45–50.

[2] LI W T, WANG D H, CHAI T Y. Burning state recognition of rotary kiln using ELMs with heterogeneous features [J]. Neurocomputing, 2013, 102: 144–153.

[3] ZHOU Y C, ZHNAG C, HAN X, et al. Monitoring combustion instabilities of stratified swirl flames by feature extractions of time-averaged flame images using deep learning method [J]. Aerospace Science and Technology, 2021, 109: 106443.

[4] ZHANG R F, LU S Z, YU H L, et al. Recognition method of cement rotary kiln burning state based on Otsu-Kmeans flame image segmentation and SVM [J]. Optik, 2021, 243: 167418.

[5] WANG Z S, XUAN J P, SHI T L. Multi-source information fusion deep self-attention reinforcement learning framework for multi-label compound fault recognition [J]. Mechanism and Machine Theory, 2023, 179: 105090.

[6] LIU J Y, DIAN R W, LI S T, et al. SGFusion: A saliency guided deep-learning framework for pixel-level image fusion [J]. Information Fusion, 2023, 91: 205–214.

[7] CHEN Z L, FU L L, YAO J, et al. Learnable graph convolutional network and feature fusion for multi-view learning [J]. Information Fusion, 2023, 95: 109–119.

[8] WANG W H, XIE E Z, LI X, et al. Pyramid vision transformer: A versatile backbone for dense prediction without convolutions [C]//2021 IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, 2021: 568–578.

[9] CARION N, MASSA F, SYNNAEYE G, et al. End-to-end object detection with transformers [C]//ECCV 2020: European Conference on Computer Vision, Glasgow, 2020: 213–229.

[10] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need [J/OL]. arXiv.1706.03762, 2017-06-12/2023-09-29.

[11] TOLSTIKHIN I, HOULSBY N, KOLESNIKOW A, et al. MLP-mixer: An all-MLP architecture for vision [J]. Advances in Neural Information Processing Systems, 2021, 34: 24261–24272.

[12] XU Y F, ZHANG J, ZHANG Q M, et al. Vitpose++: Vision transformer foundation model for generic body pose estimation [J]. Journal of Latex Class Files, 2015, 14(8): 1–7.

[13] WANG T S, LIU Z H, OU W H, et al. Domain adaptation based on feature fusion and multi-attention mechanism [J]. Computers and Electrical Engineering, 2023, 108: 108726.

[14] TANG Z Y, XU T Y, LI H, et al. Exploring fusion strategies for accurate RGBT visual object tracking [J]. Information Fusion, 2023, 99: 101881.

[15] YAN X A, YAN W J, XU Y D, et al. Machinery multi-sensor fault diagnosis based on adaptive multivariate feature mode decomposition and multi-attention fusion residual convolutional neural network [J]. Mechanical Systems and Signal Processing, 2023, 202: 110664.

[16] YOO J C, HAN T H. Fast normalized cross-correlation [J]. Circuits, Systems Signal Processing, 2009, 28(6): 819–843.

(编辑　梁华银)