

# 基于热成像关键点检测的热轧卷取定尾识别研究

徐靖驰<sup>1</sup> 郭开<sup>1</sup> 张国华<sup>1</sup> 黄雪宾<sup>1</sup> 王智健<sup>2</sup>

1. 宁波钢铁有限公司, 中国·浙江 宁波 315807

2. 中冶赛迪信息(重庆)有限公司, 中国·重庆 401122

**摘要:** 本文提出一种基于热成像技术的关键点检测模型, Thermal-ViTpose。旨在克服带钢尾部自动识别出现的准确性差的技术难题。由于热成像图像一般纹理稀疏、对比度低等固有的特性, 在设计整体框架时, Thermal-ViTpose 又添加了热域显著特征提取模块(TSEM)、热域归一化处理单元(TDN)和多尺度特征整合层(MSFF)这三个重要的部分。引入热域显著特征提取模块(TSEM)来加强目标区域的核心结构信息, 抑制背景噪声的影响, 加入热域归一化处理单元(TDN), 自适应校正温度场分布变化给复杂工况带来的影响, 融合多尺度特征整合层(MSFF), 全局和局部视角的信息协同分析, 准确重构尾部轮廓。实验结果表明, 在自建的热轧生产线尾部热成像数据集中, 该算法的定位精度达到了 93.2%, 优于卷积神经网络以及可见光领域 Transformer 基线方案, 为工业智能化监控系统开发提供技术支持和理论基础。

**关键词:** 卷取定尾; 关键点检测; ViT; 热成像

## Research on Hot-Rolled Coil Tail Recognition Based on Thermal Imaging Keypoint Detection

Xu Jingchi<sup>1</sup>, Guo Kai<sup>1</sup>, Zhang Guohua<sup>1</sup>, Huang Xuebin<sup>1</sup>, Wang Zhijian<sup>2</sup>

1. Ningbo Iron and Steel Co., Ltd., China Zhejiang Ningbo 315807

2. CISDI Information Technology CO., Ltd., China Chongqing 401122

**Abstract:** A thermal imaging-based keypoint detection model, Thermal-ViTpose, is proposed for precise tail-end recognition. Built upon ViTPose, the model introduces three novel modules: a Thermal Saliency Enhancement Module (TSEM) to highlight structural cues and suppress noise, a Thermal Domain Normalization Module (TDN) for adaptive normalization under varying thermal conditions, and a Multi-Scale Feature Fusion Module (MSFF) for integrating global and local features. Experimental results on a self-constructed thermal dataset achieve an mAP of 93.2%, outperforming CNN-based and visible-light Transformer models. The proposed method demonstrates strong accuracy and robustness, offering a practical solution for intelligent hot-rolling coiling inspection.

**Keywords:** Hot rolling coiling; Keypoint detection; ViT; Thermal imaging

## 0 引言

近些年来, Transformer 架构在计算机视觉领域被大量使用, 给工业场景中复杂的视觉任务带来新的解决办法。它用自注意力机制(Self-Attention)大大提高了长程依赖关系的建模能力, 也改善了全局特征的表现效果。相关的研究发现 ViTPose 等研究在人体姿态估计任务上取得较好的效果, 为钢铁带尾部几何形状准确建模提供有力的理论支持。目前大部分的传统方法都是根据可见光成像原理来设计的, 不能很好地适应热成像数据特有的统计特性以及成像机理的要求。

为了克服已有的技术缺陷, 本文提出了一种专门用于热成像技术的关键点检测模型 Thermal-ViTpose, 主要目的就是热轧带钢尾部缺陷进行高精度的自动识别。本模

型以 ViT-Pose 框架为根基, 依照热成像数据的特点加以改良, 包含三部分主要的创新模块:

(1) 热域显著性增强模块(Thermal Saliency Enhancement Module, TSEM)主要是改善尾部结构的关键部位在热图像上可视性的效果, 用噪声抑制的方法提高目标的几何形状的识别精度和整体表现。

(2) 热域规范化模块(Thermal Domain Normalization, TDN)采用动态调节的方式来改善参数, 在应对各种变温场复杂的环境的时候起到作用。大大提高了在强辐射环境下以及有背景噪音的环境下模型的鲁棒性以及抗干扰能力。

(3) 多尺度特征融合模块(Multi-Scale Feature Fusion Module, MSFF)融合不同层级 Transformer 特征, 实现全局轮廓与局部细节的联合建模, 进一步提升模型对尾部边界

的定位精度。

由实验结果可知, Thermal-ViT Pose 模型对本研究所建立的卷取定尾热成像数据集有很强的检测能力, 平均精确召回率 (mAP) 达到了 93.2%。该指标明显大于利用卷积神经网络的方法, 并且大大超过把可见光域预训练的 Transformer 结构直接应用于红外场景时所表现出来的表现水平。

## 1 相关研究

### 1.1 热轧卷取自动化控制研究

卷取系统是钢铁热轧工艺中的一项重要工序, 其自动化程度的好坏直接关系到产品的质量以及生产的效率。传统的卷取过程主要是依靠张力调节、速度同步、末端精准定位等几个要素的共同作用, 用卷径 - 张力 - 速度三个变量之间的非线性数学模型来达到动态优化的目的。近些年来, 在现代控制理论的带动之下, MPC 和模糊逻辑控制这些新技术被用来改善该领域的系统稳定状况以及运行可靠程度。虽然这些创新技术已经有了不少的成果, 但是由于它们过度地依靠精确的系统参数, 而且由于理想化的操作条件的限制, 使得它们的实际应用中也有着很多的挑战。高温、高湿或者强光干扰等环境因素下系统抗干扰能力还要加强才能达到工业生产高标准要求。

提高尾部识别的准确性、稳定性, 把视觉检测技术应用到卷取系统末端定位的研究当中。Kong 等人用高帧率工业相机和边缘检测算法完成了初步的识别功能, 但是其方案对于蒸汽遮挡或者强光照的情况表现不佳。

### 1.2 工业视觉中的深度学习应用

由于深度学习技术不断革新, 视觉检测技术在工业上的应用领域也越来越大, 给解决钢卷尾部缺陷问题提供新的思路和借鉴。

近几年来, 深度学习技术被广泛应用到工业视觉领域, 在解决问题的尾部检测上给出了新的研究途径。卷积神经网络 (CNN) 因为有端到端训练的优势, 在图像深层语义特征的提取上有着很好的表现, 很好地克服了传统手工设计的缺点。在钢铁制造行业里, 利用 CNN 技术来开展的表面缺陷辨识, 焊缝检测以及尺寸测定等工作都取得了比较不错的效果, 且具备很强的鲁棒性以及较高的准确度特点。相关学者 Leng 等用优化过的 Faster R-CNN 实现了高效的精确的钢板瑕疵定位, Tian 团队用注意力机制改进特征提取的效果, 有效提高了对于复杂背景干扰的处理能力。

传统卷积神经网络 (CNN) 由于本身固有的局部感受

野特性, 不能很好地处理长程依赖关系以及复杂的几何结构建模任务。热轧尾部检测场景中, 因为目标区域所占比例小, 受到高反射率或者低分辨率的影响, 容易造成识别准确率降低和漏检现象的发生。因此研究者开始使用 Transformer 模型来加入到深度学习模型中, 使模型更能够抓住全局语义信息。

### 1.3 Transformer 在视觉检测中的应用

作为自然语言处理领域开创性的研究结果, 利用自注意力机制可以克服传统序列模型对于长距离依赖的忽视。计算机视觉领域里, Vision Transformer (ViT) 把图像数据分割成一个个的 Patch, 在此基础上用全局自注意力机制和位置编码来改进网络对于各种规模语义信息的认识程度。此法已经被用来做多种类型的视觉任务, 比如分类、目标检测、语义分割等, 表现出了非常高的准确性。

DETR 架构在目标检测方面依靠 Transformer 框架构建起端到端的检测体系, 把传统的锚框机制给舍弃掉。它的衍生版 Deformable DETR 用更好的训练效率和弱信号识别能力提高模型的性能。而 KeyPoint - TR 就是利用 Transformer 算法对关键点进行定位的系统, 在建立全局语义关联的时候取得了明显进步, 具备很强的抗干扰能力, 在高温高反射复杂环境下热轧尾部的特征提取也提出新的思路。

### 1.4 关键点检测在视觉检测中的应用

关键点检测技术起源于人体姿态估计, 其目的是用节点提取的方式来得到三维几何结构, 从而达到空间定位的准确性。随着研究领域越来越广, 该技术已经被应用到机器人操作、物体抓取、工业检测等各个方面, 而且已经成为解决高精度空间定位问题的主要手段。和传统的边界框标注或者语义分割方式比较起来, 关键点检测可以给出更加细致、准确的空间坐标数据。

一些研究还把关键点检测技术和焊缝路径识别结合起来, 给工业机器人路径规划以及精确控制赋予了可靠的科技支撑。而杨氏研究组也对螺栓的轻量级的关键点检测进行了深入的研究, 在复杂的环境当中可以快速准确地定位出螺栓的位置。目前该技术已经在工业生产中得到了广泛的应用, 但是热轧尾部缺陷检测领域还处在初级的研究阶段, 其应用的价值还有待进一步的挖掘和证明。

## 2 方法设计

为了解决由于高温背景干扰、红外信号波动、目标结构边缘不清等造成的热轧带钢尾部检测的技术问题, 本文提出了一个专为复杂热成像场景的热成像关键点检测模型

叫做 Thermal-ViTpose，并且给出了该模型的具体功能。

本研究基于 ViTPose 架构构建检测模型，创新性地整合了热域显著特征增强模块（Thermal Saliency Enhancement Module, TSEM）、热域归一化模块（Thermal Domain Normalization, TDN）以及多尺度特征融合模块（Multi-Scale Feature Fusion Module, MSFF），致力于提升带钢尾部关键区域的精确定位精度与属性解析水平。第一部分用骨干网络对输入热成像数据进行多层语义特征提取，第二部分用 TSEM 突出目标区域的重要性，第三部分用 TDN 对局部温度分布进行规范化校正，第四部分用 MSFF 将不同分辨率的特征映射融合在一起得到高精度关键点热图来完成检测任务。

## 2.1 热域显著特征增强模块

热成像场景下，带钢尾部和背景之间存在的辐射强度差别很小，在低对比度区域中，传统的 ViTPose 模型就会遇到很大的困难。因此本文提出一种 TSEM 模块，用显著性引导机制增强带钢尾部的特征，抑制热背景噪声来提高目标识别准确率。

TSEM 模块用建立的显著性权重矩阵  $S$  来度量各个区域的重要性。显著性计算方式把通道注意力和空洞注意力机制结合在一起，具体的数学表达如下式子。

$$S = \sigma(\text{Conv}_{1 \times 1}([A_c(F), A_s(F)])) \quad (1)$$

其中  $F$  为输入特征图， $A_c(F) = \text{GAP}(F)$  表示全局平均池化得到的通道注意力， $A_s = \text{Conv}_{3 \times 3}(F)$  表示空间维度的响应， $\sigma$  为 Sigmoid 激活函数。

根据公式（2），模块化特征增强的操作流程就是用显著性权重来完成的一种运算过程。

$$F' = F \otimes (1 + \alpha S) \quad (2)$$

其中  $\otimes$  表示逐元素乘法， $\alpha$  为可学习权重系数，用于控制显著性强化强度。

利用热图把和尾部边界的热量集中起来，明显提高了检测模型对特征提取方面识别效率。

## 2.2 热域归一化模块

热轧时由于时间和温度的变化以及辐射作用等各方面因素的共同影响，在热成像图的空间强度分布上存在着明显的不均匀现象。传统层归一化（LayerNor）方法由于本身具有的统计特性，在这样复杂的环境中很容易出现特征映射偏差、归一化过程失效等问题。本文提出一种基于自适应的热域归一化模块（TDN），它可以实时地去捕捉、校正热场参数的变化，在各种工况下都能达到特征表示一致、稳定的目的。

对于输入特征  $F' \in RC \times H \times WF' \in RC \times H \times W$ ，传统归一化形式为：

$$\hat{F}_i = \frac{F'_i - \mu}{\sqrt{\sigma^2 + \epsilon}} \quad (3)$$

其中  $\mu$  与  $\sigma^2$  分别为特征的均值与方差。

在 TDN 中，这两个统计量通过一个轻量动态网络根据热强度分布自适应生成：

$$\mu_T = f_\mu(\text{GAP}(F')), \quad \sigma_T = f_\sigma(\text{GAP}(F')) \quad (4)$$

其中  $f_\mu$  与  $f_\sigma$  为由全连接层构成的热分布调节函数。最终输出归一化结果为：

$$\tilde{F} = \gamma \cdot \frac{F' - \mu_T}{\sqrt{\sigma_T^2 + \epsilon}} + \beta \quad (5)$$

其中  $\gamma$ 、 $\beta$  为可学习仿射变换参数。

其实质就是一种动态调控特征分布的方式，其目的就是减少高温迁移给系统性能带来的负面影响，从而大大提高模型在复杂环境中稳定性和适应性的表现。

## 2.3 基于 ViT 的关键点检测网络

关键点检测模块以 ViTPose 为基线，ViTPose 的核心是将图像划分为 patch 序列并通过 Transformer 编码器进行全局建模。设卷积特征  $F_e$  划分为  $N$  个 patch，每个 patch 大小为  $P \times P$ ，则得到 patch 序列：

$$x_i = \text{Flatten}(F_e[:, h_i: h_i + P, w_i: w_i + P]), \quad i = 1, 2, \dots, N \quad (6)$$

序列输入到 Transformer 编码器中进行全局特征建模：

$$z_i^{(l)} = \text{MSA}(\text{LN}(z_i^{(l-1)})) + z_i^{(l-1)} \quad (7)$$

$$z_i^{(l)} = \text{MLP}(\text{LN}(z_i^{(l)})) + z_i^{(l)} \quad (8)$$

其中  $z_i^{(l)} = x_i$ ，MSA 表示多头自注意力机制，LN 表示层归一化，MLP 为前馈网络。Transformer 编码器通过自注意力机制捕捉尾部关键点之间的全局依赖关系，使局部细节与整体结构信息得以联合建模。

## 2.4 多尺度特征融合

尾部关键点通常位于图像小区域，边界模糊。为同时保留局部细节和全局信息，本文引入多尺度特征融合模块（MSFF）。设 Transformer 编码器输出多尺度特征为  $\{F_s\}_{s=1}^S$ ，其中  $S$  表示尺度层数，每一尺度特征通过线性投影映射到统一通道维度：

$$\hat{F}_s = W_s F_s + b_s \quad (9)$$

多尺度特征融合采用加权相加形式：

$$F_{fused} = \sum_{s=1}^S \alpha_s \hat{F}_s, \quad \sum_{s=1}^S \alpha_s = 1 \quad (10)$$

权重  $\alpha_s$  通过可学习参数自适应调整, 实现局部与全局特征的平衡, 从而提升尾部关键点的检测精度。

## 2.5 关键点预测与损失函数

融合后的特征图  $F_{fused}$  输入关键点预测头, 生成尾部关键点热力图  $H$ , 其中  $K$  为关键点数量。每个关键点  $k$  的热力图  $H$  使用二维高斯核生成:

$$H_k(u, v) = \exp\left(-\frac{(u-u_k^*)^2 + (v-v_k^*)^2}{2\sigma^2}\right) \quad (11)$$

其中  $(u_k^*, v_k^*)$  为关键点真实坐标,  $\sigma$  高斯核标准差。预测热力图与真实热力图之间的均方误差定义为热力图损失:

$$\mathcal{L}_{heatmap} = \frac{1}{K} \sum_{k=1}^K \|\widehat{H}_k - H_k\|_2^2 \quad (12)$$

为了保证关键点之间的几何一致性, 引入空间约束损失:

$$\mathcal{L}_{geo} = \frac{1}{|\mathcal{E}|} \sum_{(i,j) \in \mathcal{E}} \|\widehat{p}_i - \widehat{p}_j\|_2 - \|p_i^* - p_j^*\|_2\|_2^2 \quad (13)$$

其中  $\mathcal{E}$  为关键点之间的边集合,  $\widehat{p}_i$  和  $p_i^*$  分别为预测与真实坐标。

最终总损失函数为:

$$\mathcal{L} = \lambda_1 \mathcal{L}_{heatmap} + \lambda_2 \mathcal{L}_{geo} \quad (14)$$

其中  $\lambda_1$  和  $\lambda_2$  为权重超参数, 通过交叉验证确定。

## 3 实验与结果分析

### 3.1 数据集构建与划分

为了验证本文所提出的基于关键点检测的热轧卷取自动定尾模型的有效性, 本文自行构建了一个热成像钢卷尾部数据集。该数据集来源于钢铁企业热轧生产线现场采集的红外图像, 覆盖了多种工况下的卷取过程, 图像分辨率统一为  $640 \times 640$ , 总计收集 4000 张热成像图像。针对每张图像, 人工标注了带钢尾部的关键点位置。

根据初始的样本集来创建训练集 8 份 2 和验证集 4 份 1 的样本, 得到总的 3200 份训练样本和 800 份验证样本。为了克服由于不同的采集设备产生的光照非均匀性, 在标注完全部的图像之后对所有的图像进行热值线性化校正、Gamma 曲线调整、边缘增强滤波等一系列标准化的预处理。在实验环节当中加入随机旋转、水平翻转和高斯噪声干扰等诸多的数据增强手段, 目的在于提高模型的泛化能力以及使其更适合复杂的环境。

### 3.2 实验设置

实验在 NVIDIA RTX 4090 GPU 上进行, 深度学习框架采用 PyTorch 2.2。训练过程中, 批量大小 (batch size) 设为 32, 优化器为 AdamW, 初始学习率  $1 \times 10^{-4}$ 。

本文创建起来的评价体系由四个关键指标构成, 分别为 PCK (关键点精准度), 它显示关键点坐标偏差是否处在预定界限之内所占的比例状况, NME (归一化均方根误差), 体现姿态识别期间各个关键点的预测精度情况, mAP (平均精确率 - 召回率曲线下的面积), 体现模型对于姿态检测任务综合性的效能水准, FPS (每秒帧处理速率), 主要用来评判算法的实际运作速度以及计算能力。

### 3.3 对比实验

为了验证本文方法的性能优势, 将其与当前主流关键点检测模型进行对比, 所有模型均在相同数据集与训练配置下重新训练, 表 1 展示了在验证集上的关键性能指标。

表 1 不同模型在热成像钢卷尾部数据集上的性能对比

| 模型             | Backbone  | PCK@0.05 ↑ | NME ↓ | mAP ↑ | FPS ↑ |
|----------------|-----------|------------|-------|-------|-------|
| SimpleBaseline | ResNet50  | 88.7       | 5.3   | 86.1  | 58    |
| HRNet-W48      | HRNet     | 91.2       | 4.8   | 89.6  | 42    |
| DARKPose       | HRNet     | 91.9       | 4.5   | 90.2  | 41    |
| SCNet          | CSPNet    | 89.8       | 5.0   | 87.9  | 73    |
| DEKR           | HRNet     | 90.1       | 4.7   | 88.5  | 40    |
| UDP-Pose       | HRNet     | 92.0       | 4.3   | 90.9  | 43    |
| ViTPose-S      | ViT       | 92.5       | 4.2   | 90.4  | 36    |
| HRFormer-S     | HRFormer  | 93.1       | 4.1   | 91.0  | 39    |
| RTMPose        | CSPResNet | 91.5       | 4.6   | 89.8  | 80    |
| 本文方法           | ViT       | 95.4       | 3.7   | 93.2  | 44    |

从表中可以看出, 本文提出的改进模型在各项指标上均取得最佳性能: PCK@0.05 达到 95.4%, 较原始 ViTPose 提升 2.9%, 较传统 HRNet 提升 4.2%; NME 降至 3.7, 说明关键点定位误差最小; mAP 提升至 93.2, 验证了检测结果的整体一致性; FPS 保持在 44 帧 / 秒, 虽略低于 RTMPose, 但在精度大幅领先的情况下仍具备工程可用性。此外, 从模型类型对比可发现: 卷积型模型 (SimpleBaseline, HRNet, DEKR) 在热成像低纹理场景中特征表达能力受限; 混合型模型 (HRFormer, UDP-Pose) 在特征融合方面表现较好; Transformer 型模型 (ViTPose, Improved ViTPose) 凭借全局注意力机制, 在边缘模糊和低对比度下表现更为稳定。

### 3.4 消融实验

为了验证各改进模块的有效性, 本文在 ViTPose 基线模型上逐步加入所提出的改进模块, 进行消融实验, 分别验证热域显著特征增强模块 (Thermal Saliency Enhancement Module, TSEM)、热域归一化模块 (Thermal Domain Normalization, TDN) 以及多尺度特征融合模块 (Multi-Scale Feature Fusion Module, MSFF) 的效果, 实验结果如表 2 所示。

从表中可以看出: 加入 TSEM 后模型在热成像场景下的边缘识别显著改善; TDN 模块进一步提升了关键点的空

间一致性与多尺度检测能力；MSFF 的引入有效缓解了单窗口自注意力的局部限制，使模型能够同时捕获尾部整体轮廓与细节特征。

表2 模块消融实验结果

| 模型配置                      | TSEM | TDN | MSFF | PCK@0.05 ↑ | NME ↓ |
|---------------------------|------|-----|------|------------|-------|
| Baseline (ViT)            | ×    | ×   | ×    | 92.5       | 4.2   |
| + TSEM                    | ✓    | ×   | ×    | 93.7       | 4.0   |
| + TSEM<br>+ TDN           | ✓    | ✓   | ×    | 94.6       | 3.8   |
| + TSEM<br>+ TDN<br>+ MSFF | ✓    | ✓   | ✓    | 95.4       | 3.7   |

4 结语

本文主要针对热轧带钢尾部高精度自动检测核心技术问题展开研究，并提出了一种新的热成像关键点检测方法叫做 Thermal-ViT Pose 模型。在算法规划的时候采取了热域明显特征加强，温度区间细分严格，多种尺度属性混杂的方式改善模型品质。热域显著特征增强模块（TSEM），使用通道卷积 + 空洞卷积的方式增强热域特征，用双层注意力去除背景干扰，即使在对比度低或者有复杂物体的环境下也能够很好地找到尾部的形状特征，温度区间标准化处理模块（TDN），利用实时采集的数据进行统计分析得到归一化因子来解决极端高温、环境噪音大、局部过热等问题，提高系统的实际应用性，多尺度特征融合模块（MSFF），将多层次 Transformer 编码器和全局轮廓信息、局部细节特征相结合，提升边界定位的精确度以及关键点标注的一致性。这三个子系统共同起作用，提高

了 Thermal-ViT Pose 的综合性能，在实验评价中比现有的主流卷积神经网络以及传统的 Transformer 框架有更好的表现，平均精度指数 (mAP) 大于 93%。该成果可以促使钢铁生产的智能化进程加快，给高质量产品的全流程监测赋予技术支持和理论支撑。

虽然 Thermal-ViT Pose 实验结果较好，但是它实际应用的价值仍然受到许多因素的限制。此模型对于极低对比度的热图、复杂遮挡的情况有明显的不足，并且硬件需求较高，造成能耗问题。为了解决上述问题，后续的研究可以集中于以下几个方面，即加入自监督学习机制来改进预训练的过程，利用跨模态特征融合算法将可见光和红外成像的数据融合起来来增强鲁棒性，并且使用轻量化的网络结构来设计出更加紧凑的运算环境，以此来达到更好的通用性、抗干扰性和节能的目的。

参考文献：

[1] 陈志军. 浅谈热轧卷取机的选择控制[J]. 冶金动力, 2024(01):80-81+110.

[2] 张君威. 基于机器视觉的热轧钢卷卷径测量及卷取张力控制[D]. 燕山大学, 2023.

[3] 沈益钊. 湛钢 2250 热轧卷取机自动定尾控制优化[J]. 新疆钢铁, 2024(04):183-185.

作者简介：徐靖驰（1976.02-），男，汉族，辽宁辽阳人，本科，高级工程师，研究方向：从事电气自动化研究。