文章编号　1004-924X(2025)16-2502-14

基于多尺度注意力机制 TransUNet的双目视觉定位与测量方法

杨　玉，许四祥*，张梦权，吴端正
（安徽工业大学 机械工程学院，安徽 马鞍山 243032）

摘要：针对传统双目视觉特征检测算法检测效率低以及大多网络模型对于全局重要特征的关注度不足、参数量过大等问题，提出一种基于多尺度注意力机制 TransUNet网络的双目视觉的连铸坯定位与测量方法。首先使用标定后的平行双目相机采集连铸坯左右图像数据，构建数据集。接着，以 TransUNet网络作为基础框架，引入改进的 Transformer层实现全局上下文信息的提取；提出一种全局坐标分组注意力模块 GSGA（Global Spatial Group Attention）添加在每个解码器末尾，利用分组和多尺度注意力机制提高对全局重要特征的关注度；在编码器和解码器跳跃连接及双线性插值后加入 CBAM（Convolutional Block Attention Module）模块，结合空间和通道注意力机制提升关键点识别能力。最后，结合双目视觉原理对网络输出的关键点坐标实现三维坐标重建和测距。实验结果显示：相较于 Transformer模型，均方根误差和归一化误差分别下降了 40.96% 和 45.83%，参数量和浮点运算量分别减少了 10.58% 和 8.21%，单批次推理时间缩短了 30.52%。在三维测距上，测量相对误差达到 0.137%，显著优于传统特征检测算法，满足双目视觉定位与测量要求。

关　键　词：双目视觉；TransUNet；关键点检测；注意力机制
中图分类号：TP391.41　　**文献标识码：**A
doi：10.37188/OPE.20253316.2502　　**CSTR：**32169.14.OPE.20253316.2502

Binocular vision localization and measurement method based on TransUNet with multi-scale attention mechanism

YANG Yu, XU Sixiang*, ZHANG Mengquan, WU Duanzheng

(*College of Mechanical Engineering, Anhui University of Technology, Ma'anshan* 243032, *China*)
* *Corresponding author, E-mail: xsxhust@ahut.edu.cn*

Abstract: Aiming at the problems of low detection efficiency of traditional binocular-vision feature-detection algorithms, as well as the insufficient attention to globally important features and the excessive parameter count of most network models, a method of continuous-casting-billet localization and measurement based on TransUNet binocular vision with a multiscale attention mechanism was proposed. Firstly, left and right images of continuous-casting billets were collected with a calibrated parallel binocular camera to build a dataset. Subsequently, taking TransUNet as the backbone, an improved Transformer layer was introduced to extract global context information; a Global Spatial Group Attention (GSGA) module was appended after every decoder block to enhance focus on globally salient features through a

收稿日期：2025-03-20；**修订日期：**2025-05-25.
基金项目：国家自然科学基金资助项目（No. 51374007）；安徽高校自然科学研究重点项目（No. KJ2020A0259）

grouped multiscale attention mechanism; and a Convolutional Block Attention Module (CBAM) was inserted after each encoder-decoder skip connection and bilinear interpolation to boost key-point recognition by combining spatial and channel attention. Finally, 3-D coordinate reconstruction and distance measurement were performed on the network's key-point coordinates by leveraging binocular-vision principles. The experimental results show that compared with the Transformer model, the root-mean-square error and normalized error are reduced by 33.8% and 36.83%, the number of parameters and floating-point operations are reduced by 10.58% and 8.21%, and the single-batch inference time is shortened by 32.30%. In 3D ranging, the relative error of measurement reaches 0.137%, which is significantly better than the traditional feature detection algorithm and meets the binocular vision localization and measurement requirements.

Key words: binocular vision; TransUNet; keypoints detection; attention mechanism

1　引　言

现代钢铁冶金行业连铸坯经火焰切割，切割面上的熔化钢液在切割口的下边缘形成不规则且硬度大的毛刺，对钢材的表面质量和运输辊道造成不利影响。为此，许四祥等人[1]设计了一种用于去除这些毛刺的等离子装置，但这种装置本身并不具备自动定位毛刺区域的能力，因此需要借助定位系统的支持。基于双目视觉技术的非接触式测量特性，将立体匹配与三角测量原理的结合对连铸坯进行定位与测量。

传统的双目视觉特征检测方法有 SIFT (Scale Invariant Feature Transform)[2]，ORB(Oriented FAST and Rotated BRIEF)[3] 和 SURF (Speededup Robust Features)[4]等。杨宇等人[5]提出了一种改进的ORB算法，通过构建高斯核函数加权和字符串描述子响应模型强化特征稳定性，提升定位和测量的准确性；XU 等[6]优化 BRISK (Binary Robust Invariant Scalable Keypoints)算法，提出特征向量动态编码方法，通过构建梯度幅值筛选机制过滤低响应区域，提高了匹配精度；宋超群等[7]改进 FAST(Features from Accelerated Segment Test)算法，像素点邻域灰度均值生成描述子，提高测量精度。龚坤等[8]提出融合点线特征的双目视觉 SLAM 方法，提升弱纹理场景的鲁棒性；张浩等[9]通过一维概率 Hough 变换与局部 Zernike 矩优化双目测量精度。尽管传统的双目视觉检测算法在精度和效率上有了显著提升，但仍然存在关键点冗余或漏提取、匹配过程复杂、计算复杂度高、参数调优困难、缺乏全局上下文信息等问题。

Transforme[10]模型通过多组注意力头实现特征交互的动态权重分配，建立全局上下文依赖，提升了模型对于全局特征的识别精度。Chen 等人[11]首次提出结合 Transformer 和 UNet 的 TransUNet 架构，并将其应用于医疗图像识别，该模型利用 Transformer 多头交叉注意力机制实现跨模态特征融合，提高对医学图像的检测精度。然而，传统 TransUNet 模型并没有充分发挥出 Transformer 的优势[12]。Cao 等人提出 Swin-Unet 模型[13] 设计了层次化窗口注意力机制，通过动态分块自注意力计算实现多尺度特征建模。基于自注意力的视觉 Transformer 架构突破了传统卷积神经网络的感受野限制，但对于局部特征检测仍然具有一定的局限性[14]。Huang 等人提出 MISSFormer[15] 设计精确化 Transformer 机制，提高了局部特征与全局特征的关联性。Heidari 等人提出 Hiformer 模型[16]，设计双级融合模块(Double-Level Fusion Block)来对全局和局部上下文信息进行整合。现有融合架构虽在跨尺度特征整合方面取得进展，但是对于重要特征通道和空间位置信息的特征表示仍然不够准确。

引入注意力机制(Attention Mechanism)可以显著提升模型性能。Bahdanau 等人[17]最早引入了注意力机制，用于神经机器翻译中，通过动态关注源文本的不同部分来提高翻译效果；随后，Luong 等人[18]进一步探索了这一机制的应用，提出了多种不同的注意力模型并验证了其有效性。Vinyals 等人[19]开发了指针网络，利用注意力

机制选择输入序列中的特定元素作为输出，特别适用于组合优化问题。Velickovic等人[20]将注意力机制扩展到图结构数据，提出了图注意力网络(Graph Attention Network，GAT)，通过自注意力机制为每个节点分配权重，提升了特征聚合的效果。Fu等人[21]设计了双注意力网络(Dual Attention Network，DANet)，结合位置和通道注意力模块，以更好地利用全局上下文信息，从而提高了场景分割任务的准确性。尽管这些方法通过引入注意力机制模块提升了模型精度，但在突出重要信息识别能力和提取局部上下文信息方面仍有改进空间，特别是在强调关键区域和细化局部特征表达上表现不足。

由此，提出了基于多尺度注意力机制TransUNet的双目视觉的连铸坯定位与测量方法。首先，提出一种全局坐标分组注意力模块(Global Spatial Group Attention，GSGA)，并在每个解码器末尾融入GSGA模块，通过空间维度解耦(高度/宽度独立处理)和结构感知分组机制，突破传统多尺度方法(如PSA/ASPP)在均匀纹理场景的局限，显著增强对连铸坯关键点的检测能力；接着，在跳跃连接和解码器末端引入卷积块注意力模块(Convolutional Block Attention Module，CBAM)，构建双阶段注意力协同机制，形成GSGA(全局结构建模)和CBAM(局部特征细化)的级联优化，利用空间和通道注意力互补提升关键点识别精度；最后，在Transformer层中加入可学习的位置编码加强模型对于不同位置特征的理解；将嵌入层和MLP层中标准卷积替换成深度可分离卷积，降低模型参数量和计算量；引入高效多头自注意力机制EMSA(Efficient Multi-head Self-Attention)，增强模型对于细节特征的捕捉。最后根据三角测量原理获得连铸坯的三维坐标完成测距任务。

2　本文方法

基于多尺度注意力机制TransUNet网络的双目视觉的连铸坯测定总流程如图1所示，主要包括三个部分，分别是数据集预处理、TransUNet网络模型训练、连铸坯定位与测量。

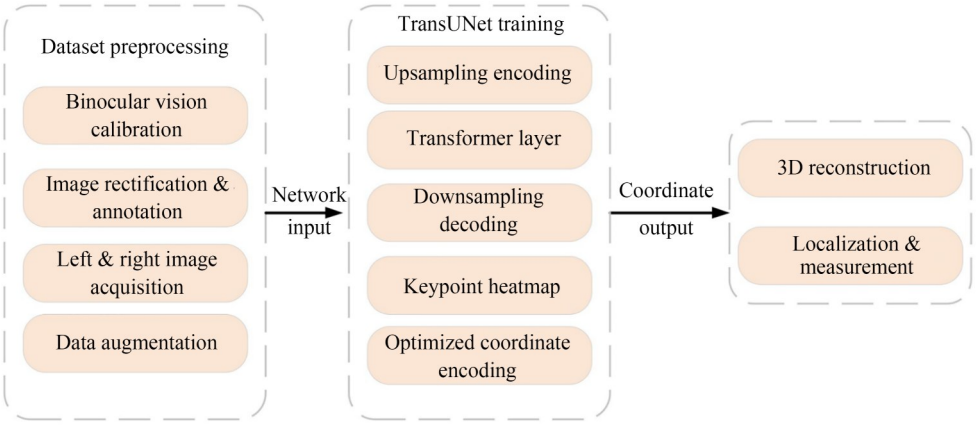

图1　连铸坯测定总体流程

Fig. 1　Overall process of continuous casting billet measurement

2.1　数据集采集与处理

工业生产连铸坯时切割面熔化钢液在切割口下边缘形成不规则毛刺，如图2所示，使用标定后的双目相机采集连铸坯左右视图中四个关键点的位置完整连铸坯测定任务。

对四型连铸坯进行图像采集如图3所示，利用Zhang[22]标定方法对相机进行标定。多角度采集四类连铸坯图像(1 600 pixels×1 200 pixels)，经畸变校正和极线校准处理。将图像统一调整为1 664 pixels×1 152 pixels，最终共收集了1 250张图像，划分为训练集1 180张，测试集70张。经过数据增强处理后，整个数据集的规模达到了

2 425 张图片，其中 2 183 张被分配给了训练集，而剩余的 242 张则构成了测试集。部分经过扩充处理后的样本图像如图 3 所示。

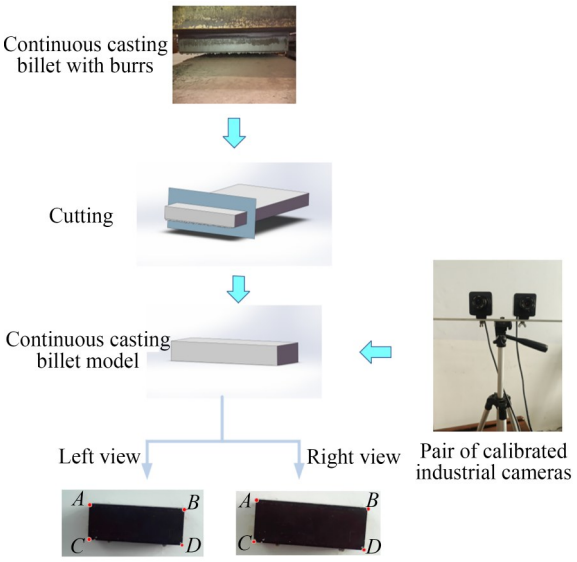

图 2　连铸坯测定流程

Fig. 2　Continuous casting billet measurement process

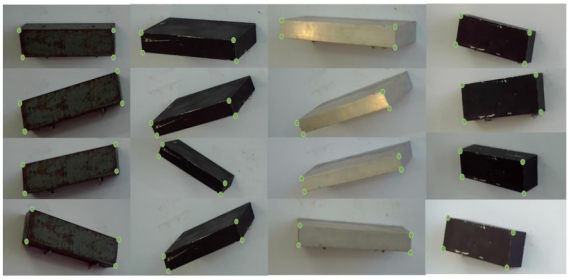

图 3　部分四型连铸坯数据集

Fig. 3　Partial four type continuous casting billet dataset

2.2　TransUNet 模型

本文在 Transformer[23] 模型的基础上改进得到的 TransUNet 网络整体结构如图 4 所示。其中，编码器部分由 CNN 骨干网络和 Transformer 层组成。在 CNN 部分，首先使用预训练的 ResNetV2-50 网络逐步提取图像的局部特征形成多尺度特征图。接着将 CNN 输出的 2D 特征图转换为适合 Transformer 处理的 1D 序列并添加可学习的位置编码。在 Transformer 部分中，

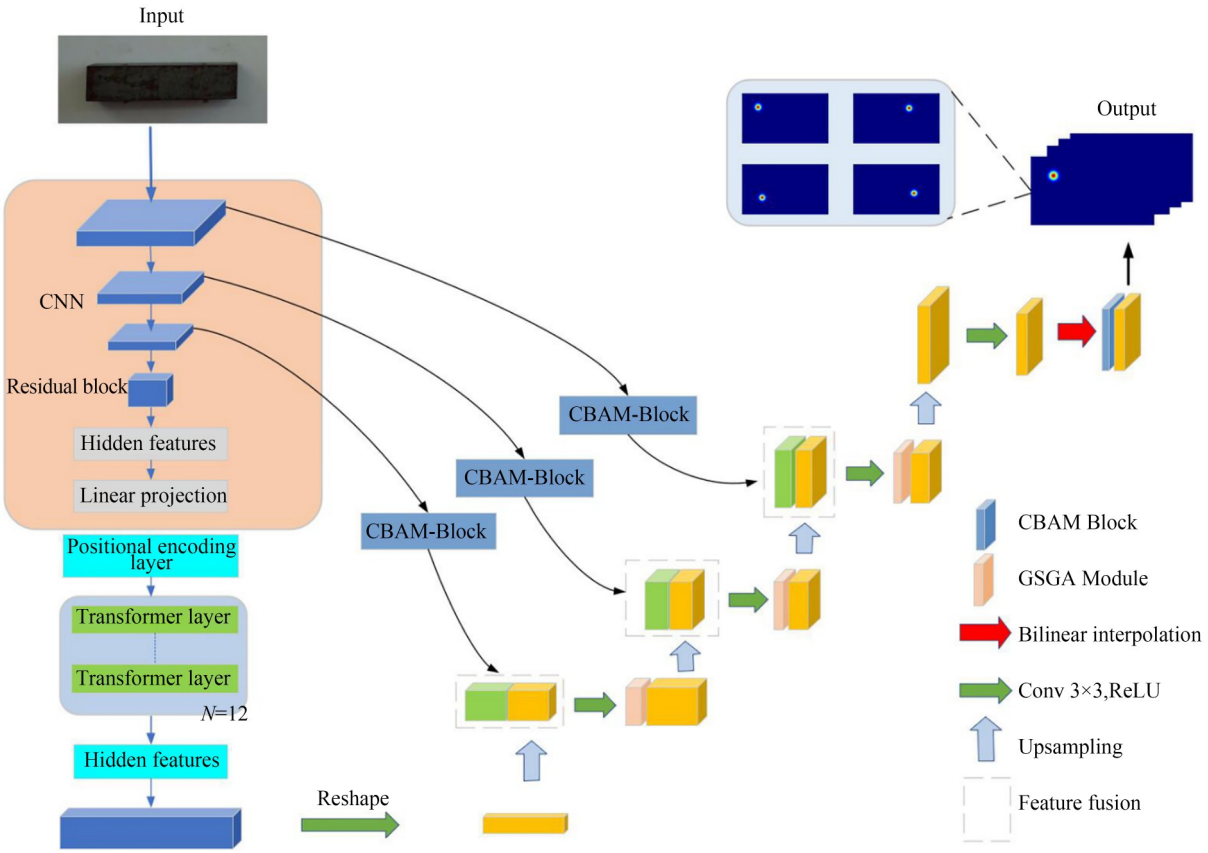

图 4　优化后 TransUNet 网络架构

Fig. 4　Optimized TransUNet network architecture

将序列和保存好的位置编码输入到12层的Transformer模块中进行全局特征信息的提取。引入优化后的多头注意力机制(EMSA)加强模型对于细节特征的捕捉。将嵌入层和MLP层中标准卷积替换成深度可分离卷积,降低模型参数量和计算量。

在解码器部分,输入的特征图首先经历三次上采样,逐步恢复其空间分辨率。与此同时,在每次上采样后进行卷积操作,以融合多尺度信息并进一步提取高级特征。为了弥补上采样过程中部分重要特征的损失,在三层跳跃连接中,融入设计好的CBAM-Block对中低级细节信息进行还原。为了有效融合多尺度信息并精准恢复关键的空间结构细节,特别是针对连铸坯切割面的方向性线性特征,在每个解码器末尾创新性地融入了全局坐标分组注意力(GSGA)模块,接着使用双线性插值确保输出与原始输入具有相同的分辨率并使用CBAM模块提高模型对图像细节的捕捉能力,从而提高模型的性能。

2.2.1　GSGA注意力机制

为了增强模型对空间维度全局信息的感知能力,本文提出一种全局坐标分组注意力模块(GSGA)。

传统多尺度方法(如PSA/ASPP)在均匀纹理场景中易引入噪声,且难以精准捕捉关键点的空间分布模式。为此,本文提出全局坐标分组注意力模块(GSGA),其核心创新为:首先,通过空间维度解耦策略,将输入特征图沿高度与宽度方向独立处理,分别采用全局平均池化与最大池化操作,精准捕获连铸坯切割面纵横方向的空间关联特征,强化关键点位置的全局感知能力;其次,引入结构感知分组策略,在保留空间拓扑结构的前提下进行通道分组,规避多尺度卷积导致的特征离散问题,确保关键点空间上下文的完整性。如图5所示,GSGA通过空间解耦与分组策略的协同作用,显著提升了模型对全局空间信息的建模能力,尤其适用于连铸坯关键点分布的位置依赖性特征检测。

首先将输入特征图分为高度和宽度特征,分别进行全局平均池化和最大池化,捕捉多维度的全局信息,提升特征提取的全面性。接着,将输入特征图分成多个组,每个组内独立地进行通道和空间方向的注意力计算。减少计算复杂度,同时保留更多的局部特征信息。

输入特征图 $\boldsymbol{X}\in\boldsymbol{R}^{B\times C\times H\times W}$ 被分成 N 个组,每组包含通道。对于每个组,分别计算其在高度方向和宽度方向上的全局平均池化(Global Average Pooling,GAP)和全局最大池化(Global Max Pooling,GMP),得到四个张量。接着,通过共享卷积层对这四个张量进行处理,降维和恢复维度,以生成通道和空间的注意力权重 $\boldsymbol{Y}^{N}_{h,\mathrm{avg}}$,$\boldsymbol{Y}^{N}_{h,\max}$,$\boldsymbol{Y}^{N}_{w,\mathrm{avg}}$,$\boldsymbol{Y}^{N}_{w,\max}$。将上述结果相加,并通过Sigmoid函数激活,产生最终的高度方向注意力权重 $\boldsymbol{A}_h$ 和宽度方向注意力权重 $\boldsymbol{A}_w$,其中Sigmoid

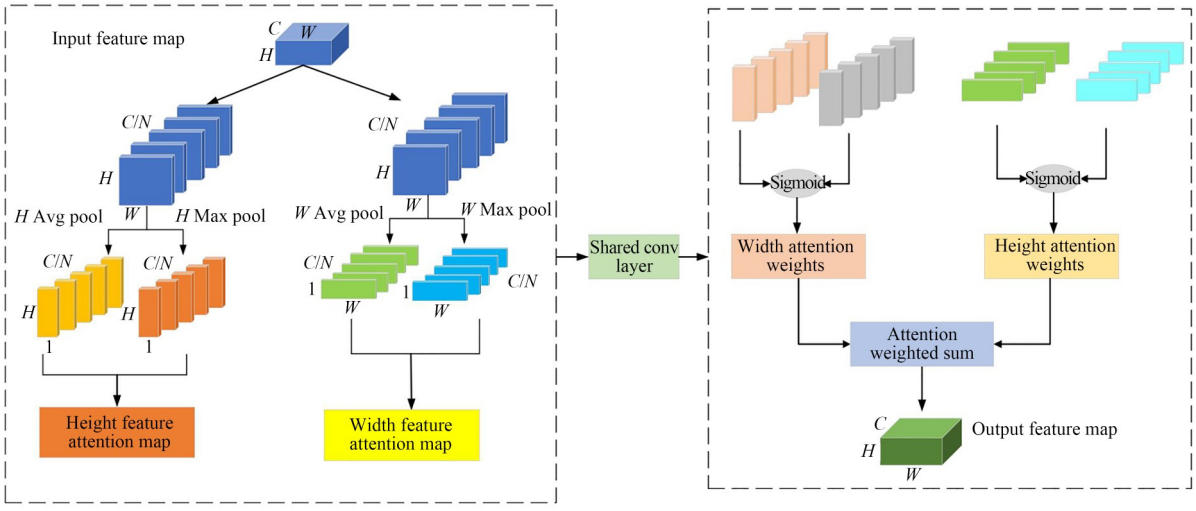

图5　GSGA模块结构图

Fig. 5　GSGA module structure diagram

函数用 σ 表示，公式如式(1)：

$$\begin{cases} A_h^N = \sigma(Y_{h,\text{avg}}^N + Y_{h,\max}^N) \\ A_w^N = \sigma(Y_{w,\text{avg}}^N + Y_{w,\max}^N) \\ Y_N = X_N \bullet A_h^N \bullet A_w^N \\ Y = concat(Y_1, Y_2, \cdots, Y_N) \end{cases} . \tag{1}$$

最后将 A_h^N 和 A_w^N 应用到原始特征图 X 中，通过逐元素乘法的方式调整每个位置的重要性，输出经过加权后的特征图 Y，最终将输出特征图所有组结果相互拼接。

其中，共享卷积层的结构如图 6 所示，该层用于对每个分组特征图进行特征处理并且由两个 1×1 卷积层、批量归一化层和 ReLU 激活函数组成。其主要目的是降低通道维度(降维)，然后恢复到原来的通道数(升维)，同时减少参数并加快计算。

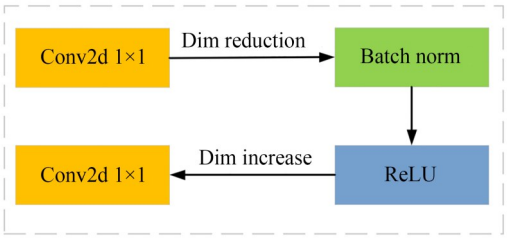

图 6　共享卷积层结构

Fig. 6　Shared Convolutional Layer Structure

在降维阶段通过一个 1×1 卷积将通道数从 C/N 降到 $C/(C \bullet N)$，升维阶段通过另一个 1×1 卷积再将通道数恢复回 C/N，其中 r 是缩减比例 X_* 表示上述四种张量中的任意一种，$Conv_{\text{down}}$ 表示降维的 1×1 卷积，$Conv_{\text{up}}$ 表示升维的 1×1 卷积，BN 表示批量归一化。具体公式如式(2)：

$$\begin{aligned} Z_{*,\text{down}} &= \text{Re}LU(BN(Conv_{\text{down}}(X_*))) \\ Y_* &= \text{Re}LU(BN(Conv_{\text{up}}(Z_{*,\text{down}}))) \end{aligned} . \tag{2}$$

综上所述，完整的共享卷积层操作可以写为：

$$\begin{aligned} f_{\text{conv}}(x) = \text{Re}LU(BN(Conv_{\text{up}} \\ (\text{Re}LU(BN(Conv_{\text{down}}(X_*)))))) \end{aligned} . \tag{3}$$

2.2.2　Transformer 注意力机制

本文网络在编码器部分引入 Transformer 实现全局上下文信息的提取，如图 7 所示。

在 Transformer 模型中，输入首先被转换为特征图，并映射到一个指定的维度。随后，将特征图展平成一系列向量，每个向量 $t_i \in R^h$ 表示序列中的一个元素，其中 $i = 1, 2, \cdots, u$，$h = 768$ 为隐藏层特征图通道数，$u = 468$ 则是展平后得到的序列长度。为了保留各特征向量在原始特征图中的空间位置信息，引入可学习的位置编码矩阵 $E_{\text{pos}} \in R^{h \times u}$ 加强模型对于不同位置特征的理解，帮助捕捉细节信息，输出可以得到二维序列 G_0：

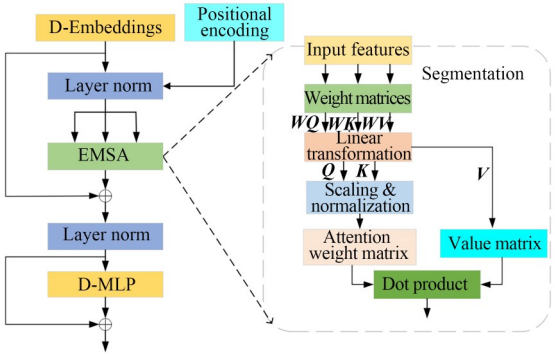

图 7　Transformer 结构图

Fig. 7　Transformer structure diagram

$$G_0 = [t_1, t_2, \cdots, t_u] + E_{\text{pos}}. \tag{4}$$

将得到的二维序列 G_0 输入 EMSA 中得到序列位置信息，EMSA 结构如图 7 中右侧所示。首先，输入特征被分割成多个头，并分别生成查询(Query)、键(Key)和值(Value)。接着，查询与键进行矩阵乘法运算并进行缩放和归一化处理生成注意力权重矩阵。最后，将注意力权重矩阵与值矩阵相乘得到最终的注意力输出 H，确保上下文高度相关的特征信息获得更强的表示，而相关性不强的特征信息则相对减弱。EMSA 层的原理为：

$$H = \sum_{n=1}^{M} soft\max\left(\frac{Q_m \bullet K_n^{\text{T}}}{\sqrt{d_k}}\right) \bullet V_n, \tag{5}$$

其中：Q_m 为第 m 个序列的查询向量，K_n，V_n 为第 n 个序列的键向量和值向量；d_k 为多头注意力的头部维度。

最后通过一个维度调整模块(D-MLP)进行前馈网络处理，进一步增强特征交互和非线性表达能力。如图 8 所示，为了降低 Transformer 层操作计算复杂度和参数数量，将嵌入层和 MLP 层中的标准卷积替换成深度可分离卷积(Depth-wise Separable Convolution)。

深度可分离卷积主要通过深度卷积和逐点卷积处理特征图。首先对每个单独的输入通道

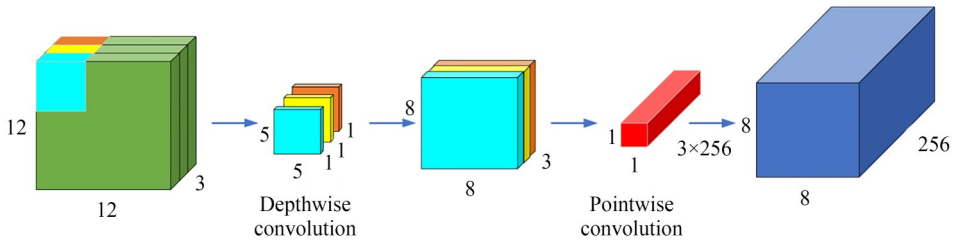

图 8　深度可分离卷积原理图

Fig. 8　Depth separable convolution schematic

应用深度卷积，以产生一组与输入通道数目相等的特征图。随后，采用 1×1 卷积核执行的逐点卷积对深度卷积生成的特征图进行通道混合，生成最终的输出特征图。前者减小模型参数提取空间特征，后者加深模型通道用于提取通道特征。

2.2.3　CBAM-Block 注意力机制

为了减少 TransUNet 模型对于图像全局特征和局部特征识别的误差，在三层跳跃连接中以及解码器末尾引入 CBAM 注意力模块。CBAM 模块内部组成如图 9 所示，包含通道注意力模块和空间注意力模块。

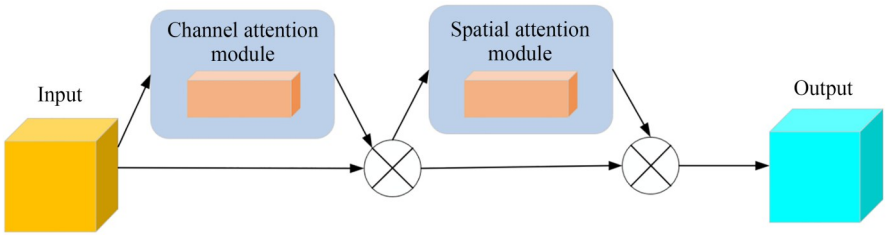

图 9　CBAM 模块内部组成

Fig. 9　Internal composition of CBAM module

如图 10 所示为 CBAM 的通道注意力模块。先通过最大池化层和平均池化层来提取特征图的全局信息，从而生成两个一维向量。随后进行中间特征学习，通过两个共享的全连接层（FC layers）处理这两个向量，生成中间特征。接着使用 ReLU 激活函数对中间特征进行非线性变换。通过 Sigmoid 函数生成最终的通道权重，表示每个通道的相关性。得到的通道注意力 $M_c(F)$ 的计算公式如式（6）：

$$M_C(F)=\sigma(MLP(AvgPoolF+MLP(MaxPool(F)))=\sigma(W_1(W_0(F^C_{avg}))+W_1(W_0(F^C_{max}))) \quad (6)$$

其中：$M_c(F)$ 表示通道注意力模块；F 为最初输入的特征图；F^C_{avg} 为平均池化；F^C_{max} 为最大池化；MLP 的权重由 W_1 和 W_0 共享。

如图 11 所示为 CBAM 的空间注意力模块。

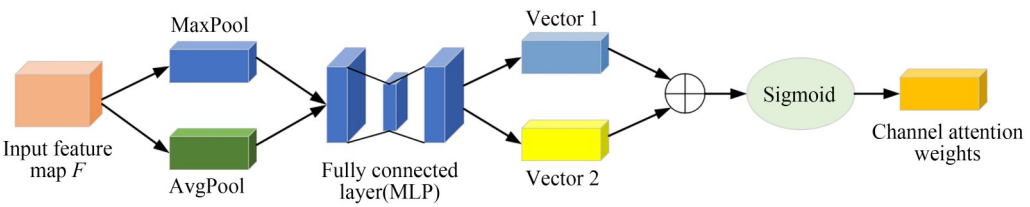

图 10　通道注意力结构图

Fig. 10　Channel attention structure diagram

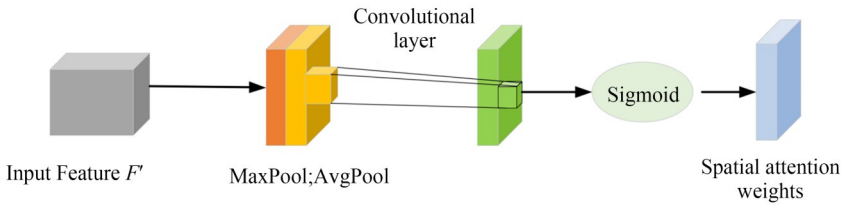

图 11　空间注意力结构图

Fig. 11　Spatial attention structure diagram

对输入特征 F' 先进行最大池化和平均池化再拼接特征图经核为 7 的卷积操作，生成二维向量。最后使用 Sigmoid 函数生成空间注意力权重 $M_s(F')$，公式如式(7)所示：

$$M_S(F')=\sigma(f^{7\times7}([AvgPoolF;\ MLP(MaxPool(F))])=\sigma(f^{7\times7}([F^S_{avg};F^S_{max}])) \tag{7}$$

其中：$M_s(F')$ 表示空间注意力模块；$f^{7\times7}$ 表示 7×7 大小的卷积核；F' 为经过通道注意力模块处理后的特征图；F^S_{avg} 表示平均池化；F^S_{max} 表示最大池化。

3　实验及实验结构分析

3.1　实验环境与训练过程

实验环境采用处理器 13th Gen Intel Core i9-13900HX 2.20 GHz，显卡为 RTX3060，框架是 PyTorch2.3.1。输入图片尺寸为 416 pixels×288 pixels。模型训练采用 SGD 优化器，初始学习率设为 1×10^{-2}，正则化 mask_ratio 初始为 0.05，衰减系数为 0.6，batch_size 大小为 4，迭代训练 15 轮训练过程中，每 2 轮更新学习率，学习率衰减为原来 1/2 倍。使用 MSE 损失函数和 Dice 损失函数的组合形成加权的总损失函数进行训练，组合损失函数公式如式(8)：

$$L_{loss}=\lambda_1\times L_{MSELoss}+\lambda_2\times L_{DiceLoss}, \tag{8}$$

其中：λ_1 和 λ_2 分别表示 MSE 损失函数和 Dice 损失函数的权重，经过实验最优取值为 $\lambda_1=0.3$，$\lambda_2=0.7$。

3.2　网络性能评价指标

在评估关键点检测网络的性能时，本实验采用了多种评价指标，包括均方根误差 E_{rmse}、归一化平均误差 E_{nme}、网络的参数量 N_P、模型的计算量 N_F。均方根误差(Root Mean Square Error，RMSE)用于衡量预测与真实关键点的误差，其表达式如式(9)：

$$E_{rmse}=\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}, \tag{9}$$

其中：n 为关键点数量，y_i，$\hat{y}_i$ 分别是真实值和预测值，均方根误差范围为(0，+∞)。

归一化误差 E_{nme} 表达式如式(10)：

$$E_{nme}=\frac{\sum_{i=1}^{n}\left\|y_i-\hat{y}_i\right\|_2}{n\times d}\times100, \tag{10}$$

其中，d 为图像中连铸坯长边的欧氏距离。

3.3　对比实验

3.3.1　10 折交叉验证下的模型性能验证

为了验证本文改进 TranasUNet 模型的预测性能和稳定性，本文选择 10 折交叉验证(10-Fold Cross Validation)作为评估模型性能的验证方法。基于全部 2 425 张增强后连铸坯数据集进行划分，将数据集划分为 10 个大小相同的子集，其中随机选择 9 个为训练集、1 个为测试集，每个测试子集约含 242 个样本。并进行 10 次训练和验证循环。计算 10 次验证的平均值，获得对于模型整体性能的准确评估结果，实验结果如表 1 所示。

由表 1 所示的 10 折交叉验证实验结果可知，改进后的 TransUNet 模型在连续坯数据集上展

表 1　10 折交叉验证下模型性能对比

Tab. 1　Performance comparison of models under 10-fold cross validation

交叉验证	E_{rmse}	E_{nme}/%
1	1.281 6	0.298 6
2	1.282 7	0.2975
3	1.279 8	0.299 5
4	1.284 3	0.298 3
5	1.287 3	0.299 2

现出优异的预测性能与稳定性，其均方根误差与归一化平均误差（$E_{\mathrm{rmse}}=1.282\pm0.003$，$E_{\mathrm{nme}}=0.299\pm0.002$）均值显著低于工业检测任务的经验阈值，且两项指标的标准差均低于0.003，表明模型在不同数据子集上的预测偏差波动极小。通过十折交叉验证的严格评估框架，模型在所有验证子集上的误差分布呈现高度一致性，验证结果表明了其具有较高的预测性能以及鲁棒性。

3.3.2　不同模型检测性能指标对比

为评估改进的TransUNet模型的有效性，在连铸坯数据集上对比了不同网络的训练结果。为确保对比实验的公正性，所有对比模型（U-Net，HRNet，TransUNet，Transformer等）均采用相同的数据增强策略即统一应用随机水平翻转（概率0.5）、随机旋转（范围±15°）、高斯噪声（$\sigma=0.01$）和亮度调整（±20%）等，所有模型均训练至收敛（验证损失连续5轮无明显下降），最大训练轮数为50，实际平均终止轮数为15～20轮，损失函数均采用MSE和Dice损失函数的组合形成加权的总损失函数进行训练。表2所示为本文算法与原模型HRNet，U-Net，TransUNet和Transformer[23]性能对比。

由表2可知，与U-Net网络对比均方根误差减少73.77%和归一化误差减少79.70%，相较于TransUNet网络均方根误差减少63.68%和归一化平均误差减少65.15%并且参数量降低10.58%和浮点运算量减少8.21%，较大程度地提升了精度的同时又节省了计算量和参数量。与Transformer[21]对比均方根误差减少40.96%和归一化平均误差减少45.83%，一定程度地提升了精度并且单批次模型推理时间降低30.52%。与其他主流方法对比，本文方法在关键点检测精度上有较大提升且推理速度也是最快的。

表2　相同数据集各算法性能对比

Tab. 2　Comparison of algorithm performance on the same dataset

算法	N_p/M	N_F/G	E_{rmse}	E_{nme}/%	T/ms
U-Net	5.70	25.73	4.902	1.473	88.35
HRNet	22.65	18.34	5.357	1.245	118.36
TransUNet	93.20	59.01	3.541	0.858	102.21
Transformer[23]	93.20	60.61	2.178	0.552	108.23
本文算法	83.34	55.63	1.282	0.299	75.20

四种连铸坯不同算法可视化结果如图12所示，其中图片左侧为各型连铸坯图片占总数据集比例，以红色点作为预测点，以绿色点为真实点，两者重合程度可以从侧面反映关键点检测结果（彩图见期刊电子版）。

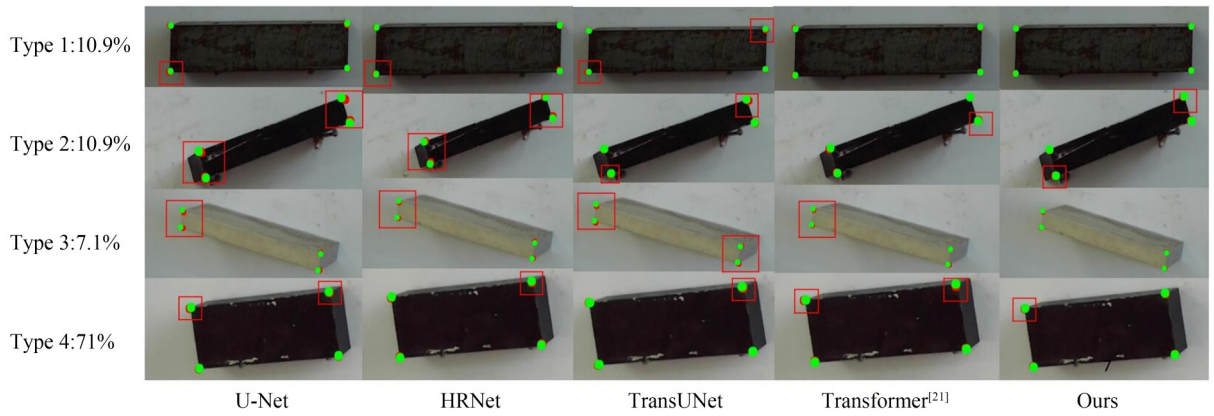

图12　不同模型在数据集上结果对比可视化

Fig. 12　Compare and visualize the results of different models on datasets

3.3.3　本文算法测量结果对比

改进TransUNet网络输出连铸坯左右图关键点的二维坐标，然后利用三角测量原理实现从二维坐标到三维坐标的转换并完成对于连铸坯的测量。如表3所示为对于四种不同类型连铸坯长和宽边的测量结果，其中以测量尺寸和误差的均值以及距离真实值最近的最优值反映测量精度差异。

表 3　本文算法四种连铸坯测量结果对比

Tab. 3　Comparison of measurement results of four types of continuous casting billets using algorithms in this article

类型编号	边	真实尺寸/mm	测量尺寸/mm		绝对误差/mm		相对误差/%	
			均值	最优值	均值	最优值	均值	最优值
1	长	300.34	301.120	300.212	0.780	0.128	0.260	−0.043
	宽	68.30	66.718	68.534	1.582	0.234	−2.316	0.343
2	长	301.90	299.581	302.035	2.319	0.135	−0.768	0.045
	宽	31.20	32.456	31.345	1.256	0.145	4.026	0.465
3	长	204.12	205.165	204.345	1.045	0.225	0.512	0.110
	宽	30.58	32.417	30.312	1.837	0.268	6.008	−0.876
4	长	262.68	261.949	262.805	0.731	0.125	−0.278	0.048
	宽	100.96	100.795	101.015	0.165	0.055	−0.164	0.055

由表 3 可知，对 4 种连铸坯长边测量绝对误差在 3 mm 之内，相对误差均小于 1%，符合测量精度要求。对于 4 型连铸坯的长和宽检测的相对误差绝对值较小，分别为 0.048% 和 0.055%，由图 12 中数据集各类型连铸坯的占比可知，样本数目少的连铸坯类型检测精度明显比样本数多的连铸坯检测精度差。对于 2 型和 3 型的连铸坯宽边检测的相对误差较大，分别达到了 4.026% 和 6.008%。可能的原因如下：长宽边像素密度的差异，长宽边关键点特征稀疏性不同以及相机标定的参数存在误差。

以测量的 1 型连铸坯的长边和 4 型连铸坯的长边为例，如表 4 所示为本文算法与其他算法 SURF，ORB[3]，Proposed ORB[5]，Proposed FAST[7]，Transformer[23] 以及 HRNet[24] 深度学习算法在三维尺寸测量的结果。SURF 采用 Hessian 阈值＝3 000，特征点数＝90，匹配方法为 FLANN＋RANSAC；ORB 设置特征点数＝250，尺度因子＝1.2，金字塔层数＝8，匹配方法为汉明距离＋最近邻比；改进 ORB/FAST 算法参数与文献[5][7]一致，特征筛选阈值为梯度幅值前 20%。深度学习对比模型的训练条件与本文算法一致（包括数据集划分、损失函数、优化器、学习率策略等），所有模型均从零开始训练，未使用预训练权重，评价指标计算方式与本文算法完全相同。

表 4　常见算法测量结果对比

Tab. 4　Comparison of common algorithm measurement results

算法	1 型连铸坯长边			4 型连铸坯长边		
	测量尺寸/mm	真实尺寸/mm	相对误差/%	测量尺寸/mm	真实尺寸/mm	相对误差/%
SURF	306.170		1.941	258.311		−1.663
ORB[3]	295.340		−1.663	267.785		1.944
Proposed ORB[5]	298.773		−0.520	—		—
Proposed FAST[7]	—	300.34	—	261.800	262.28	−0.330
Transformer[23]	299.934		−0.135	261.734		−0.208
HRNet[24]	301.023		0.227	262.609		0.125
本文算法	300.212		−0.043	262..805		0.048

由表 4 可知，本文算法对于 1 型连铸坯长边和 4 型连铸坯长边的测量精度明显高于传统检测算法。1 号长边的检测相对误差达到 0.043%，相较于深度学习算法 Transformer[23]，相对误差减

少 0.092%；相较于深度学习算法 HRNet[24]，相对误差减少 0.184%。由此可见，本文测量长边的精度均高于其他检测算法，有效地提升了测量精度。

3.4 消融实验

3.4.1 验证GSGA模块的有效性

本文通过消融实验的方式来验证全局坐标分组注意力模块 GSGA 的有效性。在保证对于本文网络其他改进不变的情况下，使用相同连铸坯数据集进行训练，分别设定了不加注意力模块(None)、添加坐标注意力模块(CA)、添加注意力模块(SE)、多尺度注意力模块(PSA)和(ASPP)以及添加本文提出的GSGA模块等四种情况，采用 E_{rmse} 和 E_{nme}/% 两项指标反映添加不同多尺度注意力模块对于模型性能的影响。

表 5 GSGA 模块有效性验证消融实验

Tab. 5 GSGA module effectiveness verification ablation experiment

引入模块	E_{rmse}	E_{nme}/%
None	2.025 3	0.546 1
CA	1.647 7	0.385 0
SE	1.557 8	0.365 3
PSA	2.602 8	0.844 1
ASPP	2.266 0	0.657 3
GSGA	1.281 6	0.298 6

由表5可知，引入注意力机制显著优于无注意力模块的基线(None)，其中轻量级的 SE 和 CA 模块分别取得 E_{rmse}=1.557 8 和 1.647 7，证实位置和通道关系在该任务中的重要性。多尺度模块 PSA 和 ASPP 表现欠佳(E_{rmse}=2.602 8 和 2.266 0)，表明传统多尺度方法可能引入冗余噪声。相比之下，本文提出的GSGA模块以 E_{rmse}=1.281 6 和 E_{nme}=0.298 6% 的优异性能显著领先，相比最佳基准(SE)提升 17.7%，验证了其通过全局坐标编码和分组注意力机制有效捕捉连铸坯线性缺陷特征的优势。

3.4.2 验证CBAM-Block模块添加位置的有效性

本文采用实验来评估CBAM-Block在网络不同位置对于模型检测效果的影响。具体来说，保持其他对于模型的改进不变，将CBAM-Block引入到编码器到解码器之间的三层跳跃连接的位置，结果显示 E_{rmse} 从 2.025 3 下降到 1.578 2，且 E_{nme} 也略有下降，这表明在跳跃连接中加入CBAM-Block可以帮助模型增强特征图中的重要特征并抑制不重要的特征，从而提高最终输出的细节和准确性。此外，在解码器末尾加入CBAM-Block，E_{rmse} 从 2.025 3 下降到 1.521 3，E_{nme} 从 0.546 1% 下降到 0.385 4%，可以看出在这个位置插入CBAM-Block可以更好地利用高分辨率的特征图，从而提高分割精度。由实验4可知，同时在跳跃连接和解码器末尾加入CBAM-Block模块，结果显示 E_{rmse} 从 2.025 3 下降到 1.281 6，且 E_{nme} 从 0.546 1% 下降到 0.298 6% 效果最好，可以看出添加CBAM-Block模块可以有效提高模型对于图像全局特征和局部特征识别能力。实验结果如表6所示，其中“√”表示添加。

表 6 CBAM 模块引入位置消融实验结果

Tab. 6 Experimental results of position ablation introduced by CBAM module

实验	跳跃连接	解码器末尾	E_{rmse}	E_{nme}/%
1			2.025 3	0.546 1
2	√		1.578 2	0.378 0
3		√	1.521 3	0.385 4
4	√	√	1.281 6	0.298 6

3.4.3 各模块对模型性能影响

为验证各模块对于网络检测精度的影响，以原始的 TransUNet 网络为基准模型，对优化的 Transformer 注意力机制、CBAM 模块以及 GSGA 模块进行了消融实验。实验结果如表7所示，TF 代表引入位置编码、深度可分离卷积DSC和 EMSA 改进的 Transformer 注意力机制，CBAM 指的是在编码器和解码器三层跳跃位置和解码器末尾同时引入 CBAM-Block 模块，GSGA 指的是在每个解码器末尾引入 GSGA 模块，表格中的“①”和“②”指的是模块的引入顺序，“√”指的是是否引入该模块。

由表7可知，实验1为 TransUNet 基准模型，实验 2 在 TransUNet 基准模型中引入改进的 Transformer 注意力机制使得 E_{rmse} 下降 15.65%，E_{nme} 下降 18.02%，这表明改进后的注意力机制

表 7　引入不同模块消融实验结果

Tab. 7　Introduce different module ablation experimental results

实验	TransUNet	TF	CBAM	GSGA	E_{rmse}	E_{nme}/%
1	√				3.541	0.858
2	√	√			2.987 0	0.703 4
3	√	√	√		2.321 7	0.589 1
4	√	√		√	2.025 3	0.546 1
5	√	√	①	②	1.281 6	0.298 6
6	√	√	②	①	1.351 9	0.328 4

有助于提升模型性能。实验 3 在此基础上添加 CBAM 模块，E_{rmse} 下降 34.43%，E_{nme} 下降 31.34%，显示了 CBAM-Block 在提供更精细特征表示方面的作用。实验 4 在实验 2 的基础上单独加入 GSGA 模块，E_{rmse} 下降 42.82%，E_{nme} 下降 36.35%，证明了 GSGA 模块的有效性，增加了模型对于高度和宽度两个空间维度上的全局信息的捕捉能力。实验 5 和实验 6 在实验 3 基础上同时添加 GSGA 模块和 CBAM 模块，并改变 CBAM 和 GSGA 模块在解码器末尾的引入顺序，实验结果表明，当 GSGA 排列在 CBAM 之前的第六组实验中，尽管仍表现出良好的性能，但 CBAM 模块先于 GSGA 模块应用的情形下，模型达到了最优表现，E_{rmse} 下降 61.82%，E_{nme} 下降 61.72%，验证了上述各个模块的有效性。

4　结　论

本文提出了一种融合多尺度注意力机制 TransUNet 与双目视觉的连铸坯定位与测量方法。通过优化 Transformer 层（引入可学习位置编码、深度可分离卷积）、设计全局坐标分组注意力（GSGA）模块以及优化 CBAM 模块的部署策略，有效提升了模型在复杂背景下对连铸坯关键点的检测精度与效率。

本研究基于双目视觉系统的精确参数标定构建实验数据集，进行实验结果显示，相较于次优方法 Transformer 模型，均方根误差和归一化误差分别下降了 40.96% 和 45.83%。在三维测距任务中，长边测量优势显著：如表 3 所示，4 类连铸坯长边的相对误差均＜1%（均值误差范围：−0.768%～0.512%，最优值 0.043%～0.110%），绝对误差≤2.319 mm，满足工业定位精度要求（≤5 mm）；短边测量存在局限，2 型与 3 型短边相对误差较高（均值 4.026%～6.008%），但最优值显著提升（−0.876%～0.465%），主要源于短边像素密度低（平均 12.3 pixel/mm，仅为长边的 41%）、特征稀疏性及相机标定误差传递效应；综合性能达标：全样本加权平均相对误差为 0.123%，仍显著优于传统算法（SURF/ORB＞1.6%）。

作者贡献声明：

杨玉：测量方法的提出，论文构思和撰写；

许四祥：测量实验的设计及数据整理和分析；

张梦权：论文审核与编辑写作；

吴端正：测量实验数据分析。

参考文献：

[1] 许四祥，陈富强，高培青，等. 一种去除板坯毛刺的系统：中国，102935547B[P]. 2014-10-15.
XU S X, CHEN F Q, GAO P Q, *et al*. System for removing slab burrs: *China*, 102935547B [P]. 2014-10-15.

[2] LOWE D G. Distinctive image features from scale-invariant keypoints [J]. *International Journal of Computer Vision*, 2004, 60(2): 91-110.

[3] RUBLEE E, RABAUD V, KONOLIGE K, *et al*. ORB: An Efficient Alternative to SIFT or SURF [C]. 2011 *International Conference on Computer Vision. November* 6-13, 2011, *Barcelona*, *Spain*. IEEE, 2011: 2564-2571.

[4] BAY H, ESS A, TUYTELAARS T, *et al*. Speeded-up robust features (SURF)[J]. *Computer Vision and Image Understanding*, 2008, 110(3): 346-359.

[5] 杨宇，许四祥，方建中，等. 基于改进ORB算法的双目视觉定位测量方法[J]. 传感技术学报，2019，32(11)：1694-1699.
YANG Y, XU S X, FANG J Z, *et al*. Location and measurement method of binocular vision based on improved ORB algorithm[J]. *Chinese Journal of*

Sensors and Actuators, 2019, 32(11): 1694-1699. (in Chinese)

[6] XU S X, DONG C C, ZHOU S H, *et al*. Binocular measurement method for the continuous casting slab model based on the improved BRISK algorithm [J]. *Applied Optics*, 2022, 61(11): 3019-3025.

[7] 宋超群, 许四祥, 杨宇, 等. 基于改进FAST和BRIEF的双目视觉测量方法[J]. 激光与光电子学进展, 2022, 59(8): 0810013.
SONG C Q, XU S X, YANG Y, *et al*. Binocular vision measurement method using improved FAST and BRIEF[J]. *Laser & Optoelectronics Progress*, 2022, 59(8): 0810013. (in Chinese)

[8] 龚坤, 徐鑫, 陈小庆, 等. 弱纹理环境下融合点线特征的双目视觉同步定位与建图[J]. 光学 精密工程, 2024, 32(5): 752-763.
GONG K, XU X, CHEN X Q, *et al*. Binocular vision SLAM with fused point and line features in weak texture environment [J]. *Opt. Precision Eng.*, 2024, 32(5): 752-763. (in Chinese)

[9] 张浩, 许四祥, 董晨晨, 等. 融入一维概率Hough变换与局部Zernike矩的双目视觉测量[J]. 光学 精密工程, 2023, 31(12): 1793-1803.
ZHANG H, XU S X, DONG C C, *et al*. Binocular vision measurement method incorporating one-dimensional probabilistic Hough transform and local Zernike moment[J]. *Opt. Precision Eng.*, 2023, 31(12): 1793-1803. (in Chinese)

[10] VASWANI A, SHAZEER N, PARMAR N, *et al*. Attention is all you need[C]. 31*st Conference on Neural Information Processing Systems*, *Long Beach*, *CA*, *USA*, 2017.

[11] CHEN J, LU Y, YU Q, *et al*. TransUNet: Transformers Make Strong Encoders for Medical Image Segmentation [J]. *arXiv preprint* arXiv: 2102.04306, 2021.

[12] 蒋婷, 李晓宁. 采用多尺度视觉注意力分割腹部CT和心脏MR图像[J]. 中国图象图形学报, 2024, 29(1): 268-279.
JIANG T, LI X N. Segmentation of abdominal CT and cardiac MR images with multi scale visual attention [J]. *Journal of Image and Graphics*, 2024, 29(1): 268-279. (in Chinese)

[13] CAO H, WANG Y Y, CHEN J, *et al*. *Swin-Unet: Unet-Like Pure Transformer for Medical Image Segmentation*[M]. Computer Vision-ECCV 2022 Workshops. Cham: Springer Nature Switzerland, 2023: 205-218.

[14] AZAD R, ARIMOND R, AGHDAM E K, *et al*. *DAE-former: Dual Attention-Guided Efficient Transformer for Medical Image Segmentation* [M]. Predictive Intelligence in Medicine. Cham: Springer Nature Switzerland, 2023: 83-95.

[15] HUANG X H, DENG Z F, LI D D, *et al*. MISSFormer: an effective transformer for 2D medical image segmentation [J]. *IEEE Transactions on Medical Imaging*, 2023, 42(5): 1484-1494.

[16] HEIDARI M, KAZEROUNI A, SOLTANY M, *et al*. HiFormer: hierarchical multi-scale representations using transformers for medical image segmentation[C]. 2023 *IEEE/CVF Winter Conference on Applications of Computer Vision* (*WACV*). *January* 2-7, 2023, *Waikoloa*, *HI*, *USA*. IEEE, 2023: 6191-6201.

[17] BAHDANAU, D, CHO, K, BENGIO, Y. Neural machine translation by jointly learning to align and translate [C]. *International Conference on Learning Representations* (*ICLR*), *San Diego*, *CA*, *USA*, 2015.

[18] LUONG T, PHAM H, MANNING C D. Effective approaches to attention-based neural machine translation[C]. *Proceedings of the* 2015 *Conference on Empirical Methods in Natural Language Processing. Lisbon*, *Portugal. Stroudsburg*, *PA*, USAACL, 2015.

[19] VINYALS, O, FORTUNATO, M, JAITLY, N. Pointer networks[C]. *Advances in Neural Information Processing Systems* 28, *Montréal*, *QC*, Canada, 2015: 2674-2682.

[20] VELIČKOVIĆ P, CUCURULL G, CASANOVA A, *et al*. Graph attention networks[C]. *International Conference on Learning Representations. Vancouver*, *BC*, Canada, 2018.

[21] FU J, LIU J, TIAN H J, *et al*. Dual attention network for scene segmentation [C]. 2019 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*). *June* 15-20, 2019. *Long Beach*, *CA*, USA. IEEE, 2019.

[22] ZHANG Z. A flexible new technique for camera calibration [J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2000, 22(11): 1330-1334.

[23] 李同谱, 许四祥, 施宇翔, 等. 基于双目视觉与Transformer的连铸坯模型定位与测量[J]. 中南

大学学报(自然科学版)，2024，55(4)：1312-1322.
LI T P，XU S X，SHI Y X，*et al*. Continuous casting slab model positioning and measurement based on binocular vision and Transformer[J]. *Journal of Central South University* (*Science and Technology*)，2024，55(4)：1312-1322. (in Chinese)
[24] 任加琪，许四祥，董宾卉，等. 基于轻量化 HRNet 的双目视觉定位与测量[J/OL]. 中国机械工程，1-9 [2024-12-24]. http://link. cnki. net/urlid/42. 1294. TH. 20241211. 1933. 008.
REN J Q，XU S X，DONG B H，*et al*. Binocular vision positioning and measurement based on lightweight HRNet [J/OL]. *China Mechanical Engineering*，1-9 [2024-12-24]. http://link. cnki. net/urlid/42. 1294. TH. 20241211. 1933. 008.

作者简介：

杨　玉(2001－)，男，安徽安庆人，硕士研究生。研究方向为机器人与机器视觉。E-mail：1308889562@qq. com

通讯作者：

许四祥(1974－)，男，湖北汉川人，博士，教授，2007年于华中科技大学获得博士学位，主要从事机器人与机器视觉方向的研究。E-mail：xsxhust@ahut. edu. cn