基于轻量化 Transformer 的手语识别系统研究

邓宇航 李佳慧 陈辉金 王海涛 *
泉州信息工程学院 机械与电气工程学院，中国 · 福建 泉州 362000

摘要：为应对手语识别系统在边缘设备部署中面临的计算资源受限问题，本文提出一种基于轻量化 Transformer 的实时手语识别方法。该方法利用 MediaPipe 提取手部 21 个关键点的三维坐标作为特征输入，构建包含 3 层轻量化 Transformer 编码器的时序建模网络，将模型参数量控制在 61.2 万以内。在自建的 12 类中国手语数据集上，通过与 CNN1D、LSTM、GRU 等基线模型进行对比实验，验证了所提方法的有效性。实验结果表明，轻量化 Transformer 模型在验证集上的识别准确率达到 98.63%，F1 分数为 0.9864，其参数量仅为 LSTM 模型的 24.5%。通过系统的消融实验，进一步揭示了模型维度、注意力头数、网络层数以及序列长度等关键超参数对模型性能的影响规律，为手语识别技术在可穿戴设备等边缘场景中的实际部署提供了可行的轻量化解决方案。

关键词：手语识别；Transformer；轻量化；关键点检测；MediaPipe

Research on sign language recognition system based on lightweight Transformer

Deng Yuhang, Li Jiahui, Chen Huijin, Wang Haitao*
Quanzhou University of Information Engineering, School of Mechanical and Electrical Engineering, China Fujian Quanzhou 362000

Abstract: To address the computational resource constraints faced by sign language recognition systems when deployed on edge devices, this paper proposes a real-time sign language recognition method based on a lightweight Transformer. The method extracts 3D coordinates of 21 hand key points using MediaPipe as feature input, and designs a temporal modeling network with a 3-layer lightweight Transformer encoder, keeping the number of model parameters under 612,000. On a self-built Chinese sign language dataset containing 12 categories, comparative experiments with baseline models such as CNN1D, LSTM, and GRU demonstrate the effectiveness of the proposed approach. Experimental results show that the lightweight Transformer model achieves a recognition accuracy of 98.63% and an F1-score of 0.9864 on the validation set, with its parameter count being only 24.5% of that of the LSTM model. Systematic ablation studies further reveal the influence patterns of key hyperparameters—including model dimension, number of attention heads, network depth, and sequence length—on model performance. This work provides a feasible lightweight solution for deploying sign language recognition technology on wearable devices and other edge computing scenarios.

Keywords: Sign language recognition; Transformer; Lightweight; Keypoint detection; MediaPipe

0 引言

手语作为全球听障人士的主要沟通方式，其自动识别与转换技术对于促进信息无障碍沟通具有重要意义。据统计，全球约有 4.66 亿人患有听力障碍，其中中国听障人士数量达 2780 万，约占全国总人口的 2%。智能手语识别技术能够实现手语与文字 / 语音之间的自动转换，从而为听障群体提供高效的沟通辅助工具，具有显著的社会应用价值。

近年来，随着深度学习技术的快速发展，基于卷积神经网络（CNN）、循环神经网络（RNN）以及 Transformer 等架构的手语识别方法取得了显著进展。Camgoz 等人[1] 提出的 Sign Language Transformers 首次将 Transformer 架构应用于连续手语识别任务，并在 PHONENIX-2014T 数据集上取得了突破性成果。Rastogi 等人[2] 提出 YOLOv10-ST 模型，通过将 Swin-Transformer 集成到 YOLOv10 框架中，在印度手语数据集上实现了 97.50% 的识别准确率与 48.7 FPS 的实时推理速度。在国内研究中，宋浩翔等人[3] 基于 MediaPipe 提取的 21 个手部关键点，结

合前馈神经网络实现了低延迟的实时识别；路飞等人[4]提出了融合轻量3D CNNs与Transformer的手语识别方法，进一步提升了模型效率。

早期手语识别研究多依赖传统计算机视觉方法，在复杂光照和背景条件下泛化能力有限[5]。刘星辰等[6]利用OpenCV和MediaPipe框架，通过手部关键点归一化特征构建了中国通用手语识别系统，实现了交互式识别。马旭冉等[7]提出基于改进YOLOv5算法的手语翻译方法，在HaGRID数据集上取得了优于同类算法的准确率和F1分数。王雅锐[8]提出SA-YOLOv8n模型，融合Sobel边缘检测与CBAM注意力机制，检测精度较原版YOLOv8n提升1.5%，并在嵌入式平台上实现了实时推理部署。

在骨架关键点特征表示方面，岳伟[9]以三维骨架信息为基础，结合引入注意力机制的时空图卷积网络（ST-GCN）和Transformer进行连续手语识别，在PHOENIX-2014-T和CSL数据集上验证了方法的有效性。边辉等[10]提出基于图卷积网络（GCN）和CTC/Attention编解码器的连续手语识别框架，融合空间注意力机制赋予骨骼关键点差异化权重，在自建SSLD数据集上平均字准确率达94.41%。杨逸峰[11]构建了覆盖120种短语的传感器手语语料库，提出融合手势姿态特征的识别模型，其中基于TinyML的轻量化子系统识别准确率达98.89%，具备边缘实时部署能力。

Vaswani等[12]提出的Transformer架构以多头自注意力机制革新了序列建模范式，并被广泛引入手语识别与翻译领域。王祉元[13]系统研究了基于Transformer的手语孤立词识别方法；格梦茹[14]提出TDD-SLT翻译模型，通过融合方向距离位置编码与“马卡龙”卷积结构，BLEU-4较基线提升0.58个点；孙意翔[15]提出双路稀疏注意力（DSA）模块，有效降低了注意力计算复杂度；刘凯[16]构建多模态传感器手语数据集，所提轻量化Transformer融合模型识别准确率达97.33%。Alsharif等[17]结合YOLOv11与MediaPipe手部关键点跟踪，系统mAP@0.5达98.2%，展示了关键点与检测框架协同设计的潜力。

Transformer的序列建模能力还被拓展至其他感知与工程领域。屈乐乐等[18]针对毫米波雷达小样本手势识别问题，提出基于Transformer元学习网络的识别方法，有效提升了少样本场景下的识别准确率。李沼洁等[19]提出融合Transformer与CNN的双分支伪装目标检测方法，为Transformer-CNN混合结构设计提供了参考借鉴。范忠岗等[20]将Transformer与强化学习结合应用于工程优化设计，进一步验证了Transformer在复杂序列决策任务中的广泛适用性。

然而，当前研究仍存在以下局限性：（1）模型计算复杂度高，多数高精度模型参数量在百万级以上，难以在资源受限的边缘设备中实现实时部署；（2）特征表示存在冗余，传统方法通常提取全身姿态特征，引入了大量与手语语义无关的噪声信息；（3）场景适应能力有限，在受控实验室环境下取得的高性能往往难以迁移至光照、背景多变的真实应用场景。

针对上述问题，本文提出一种基于轻量化Transformer的手语识别方法，其主要贡献包括：（1）构建了一个基于胸口第一视角采集的12类中国手语数据集，并采用仅包含手部关键点的特征表示，使特征计算量降低50%以上；（2）设计了一个参数量小于100万的轻量化Transformer编码器结构；（3）通过全面的对比实验和消融实验，验证了所提方法的有效性及其在轻量化部署方面的优势。

1 手语识别系统设计

1.1 系统架构

本文采用“特征提取—时序分类”的两阶段识别架构。第一阶段基于MediaPipe Hands框架从输入视频帧中实时检测并提取双手共42个关键点的三维坐标，如图1、图2所示；第二阶段将关键点序列输入轻量化Transformer编码器进行时序建模与分类。系统整体流程架构可概括为：视频输入→关键点提取→序列标准化→Transformer编码→分类输出，如图3所示。

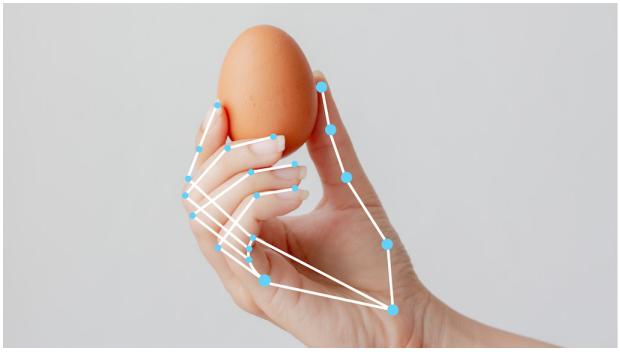
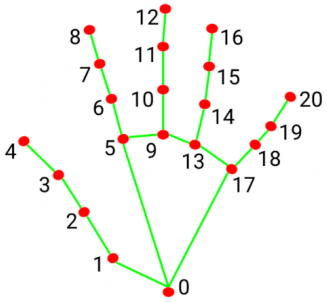


图1 手部21个关键点示意图



0. WRIST
1. THUMB_CMC
2. THUMB_MCP
3. THUMB_IP
4. THUMB_TIP
5. INDEX_FINGER_MCP
6. INDEX_FINGER_PIP
7. INDEX_FINGER_DIP
8. INDEX_FINGER_TIP
9. MIDDLE_FINGER_MCP
10. MIDDLE_FINGER_PIP
11. MIDDLE_FINGER_DIP
12. MIDDLE_FINGER_TIP
13. RING_FINGER_MCP
14. RING_FINGER_PIP
15. RING_FINGER_DIP
16. RING_FINGER_TIP
17. PINKY_MCP
18. PINKY_PIP
19. PINKY_DIP
20. PINKY_TIP

图 2 MediaPipe手部21个关键点示意图

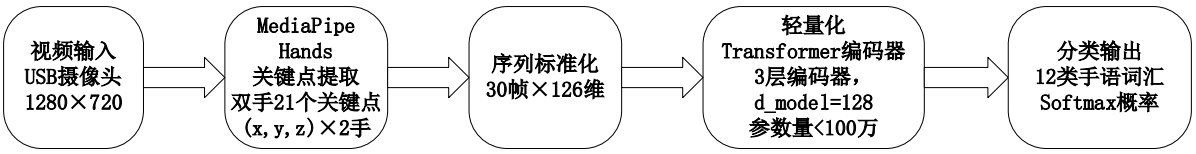


图 3 系统整体流程框架示意图

1.2 数据集构建

为更好地模拟可穿戴设备的实际使用场景，本研究构建了一个基于胸口第一视角的中国手语数据集。该数据集包含 12 类常用手语词汇，如数字 0~5 以及“OK”“名字”“什么”“你好”“谢谢”“再见”等。每类词汇采集约 30−50 个原始视频样本。采用 USB 摄像头（1280×720 分辨率）进行采集，视角固定于胸口位置以模拟可穿戴设备佩戴位置的典型视角。为提高鲁棒性，采用了数据增强策略，包括亮度调整（0.7−1.3 倍）、对比度调整（1.1−1.4 倍）、小角度旋转（±10°）等，为避免手语语义反转，排除了水平翻转操作。经数据增强后，总样本数扩充至约 1000 个。

1.3 特征提取

本系统采用 Google MediaPipe Hands 框架进行手部关键点的检测。每只手提取 21 个关键点，其中手腕拥有 1 个关键点，各手指分别拥有 4 个关键点。每个关键点包含三维坐标 (x,y,z)，因此双手联合特征维度的计算如式（1）所示。

$$d_{in} = N_k \times N_d \times N_h \quad (1)$$

式中：N_k 为单手关键点数；N_d 为每个关键点的坐标维度数（即 x,y,z 三维）；N_h 为手的数量。本文中 N_k=21，N_d=3，N_h=2，d_{in}=126。

为消除手部绝对位置对特征识别的影响，以手腕坐标为基准点，对每一帧的关键点坐标 (x_i, y_i, z_i) 进行归一化处理，如式（2）所示。

$$x_i^{'} = x_i - x_{wrist}, \quad y_i^{'} = y_i - y_{wrist}, \quad z_i^{'} = z_i - z_{wrist} \quad (2)$$

设定序列长度 L_{seq} 统一为 30 帧，对不足帧进行补零，过长序列进行截断。该特征提取策略将原始视频像素数据（以分辨率 1280×720×3 为例，约 277 万维）压缩至 126 维，维度压缩比约为 22000，计算过程如式（3）所示。

$$R_c = \frac{(W \times H \times C)}{d_{in}} \quad (3)$$

式中：R_c 为压缩比；W 为图像宽度；H 为图像高度；C 为颜色通道数；d_{in} 为关键点特征维度。本文中 W=1280，H=720，C=3，d_{in}=126，$R_c \approx 22000$。

这一显著的特征降维降低了计算复杂度，为后续轻量化模型设计奠定了基础。

1.4 Transformer 模型设计

本文设计的轻量化 Transformer 编码器结构及其核心参数如表 1 所示。输入关键点序列首先通过线性层映射到 d_{model} 维空间，并加入正弦位置编码以注入序列顺序信息。随后，数据经过多层 Transformer 编码器进行特征建模，最终通过全局平均池化与线性分类层输出类别概率。

表 1 Transformer模型参数配置

参数	数值	说明
d_{model}	128	模型特征维度
n_{head}	8	注意力头数量
n_{layers}	3	编码器层数
$d_{feedforward}$	512	前馈网络维度
$p_{dropout}$	0.1	随机失活率
总参数量	612,620	约61.2万

输入关键点序列 $X_{input} \in R^{n\times 126}$ 首先通过线性投影层映射到模型维度空间，如式（4）所示。

$$X_{embed}=X_{input}\cdot W_{embed}+b_{embed} \quad (4)$$

式中：X_{input} 为输入关键点序列，$n=30$ 为序列帧数，126 为双手关键点特征维度（$21\times 3\times 2$）；$W_{embed} \in R^{126\times d_{model}}$ 为嵌入权重矩阵；$b_{embed} \in R^{d_{model}}$ 为偏置向量；$d_{model}=128$ 为模型维度。

由于自注意力机制本身不包含位置信息，需通过位置编码为序列中的每个位置注入时序信息。本文采用正弦－余弦位置编码，如式（5）和式（6）所示。

$$PE_{(pos,2i)}=\sin\left(\frac{pos}{10000^{\frac{2i}{d_{model}}}}\right) \quad (5)$$

$$PE_{(pos,2i+1)}=\cos\left(\frac{pos}{10000^{\frac{2i}{d_{model}}}}\right) \quad (6)$$

式中：pos 为序列中的位置索引；i 为维度索引；$d_{model}=128$ 为模型维度。位置编码矩阵 PE 与嵌入向量相加，即 $X=X_{embed}+PE$，作为编码器的输入。

自注意力机制是 Transformer 编码器的核心模块，其目的是计算序列中各位置之间的关联强度，从而实现全局信息的聚合。为此，输入序列 X 通过三组独立的线性变换，分别生成查询矩阵 Q、键矩阵 K 和值矩阵 V，如式（7）所示。

$$Q=XW_Q,\ K=XW_K,\ V=XW_V \quad (7)$$

式中：$X \in R^{n\times d_{model}}$ 为编码器的输入序列；W_Q、W_K、$W_V \in R^{d_{model}\times d_k}$ 为可学习的投影参数矩阵 $d_k=\frac{d_{model}}{h}$；为每个注意力头的维度，h 为注意力头数。

在得到 Q、K、V 后，通过缩放点积运算计算注意力输出，如式（8）所示。

$$\text{Attention}(Q,K,V)=\text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (8)$$

式中：$QK^T \in R^{n\times n}$ 为注意力分数矩阵，衡量各位置之间的相似度；$\sqrt{d_k}$ 为缩放因子，用于防止点积结果过大导致 *softmax* 函数进入梯度饱和区；*softmax* 对每一行进行归一化，使注意力权重之和为 1。

单个注意力头仅能捕获一种特征关系，为增强模型的表达能力，Transformer 采用多头注意力机制，通过 h 个注意力头并行计算，使模型能够同时关注不同子空间的特征信息。每个注意力头独立进行式（7）~（8）的计算，如式（9）所示；各头的输出沿特征维度拼接后，经线性投影得到最终输出，如式（10）所示。

$$\text{head}_i=\text{Attention}(XW_{Q_i},XW_{K_i},XW_{V_i}) \quad (9)$$

$$\text{MultiHead}(X)=\text{Concat}(\text{head}_1,\cdots,\text{head}_h)\cdot W_O \quad (10)$$

式中：W_{Q_i}、W_{K_i}、$W_{V_i} \in R^{d_{model}\times d_k}$ 为第 i 个注意力头的投影矩阵（$i=1,2,...,h$）；$W_O \in R^{d_{model}\times d_{model}}$ 为输出投影矩阵；表示沿特征维度拼接操作。本文采用 个注意力头，每个头的维度 $d_k=\frac{d_{model}}{h}=\frac{128}{8}=16$。

在多头注意力之后，每个编码器层包含一个前馈神经网络（FFN），对各位置的特征独立地进行非线性变换，如式（11）所示。

$$\text{FFN}(x)=\text{ReLU}(xW_1+b_1)W_2+b_2 \quad (11)$$

式中：$x \in R^{d_{model}}$ 为某位置的特征向量；$W_1 \in R^{d_{model}\times d_{ff}}$、$W_2 \in R^{d_{ff}\times d_{model}}$ 为全连接层权重矩阵；b_1、b_2 为偏置向量；$d_{ff}=512$ 为前馈网络中间层维度；*ReLU* 为激活函数，引入非线性表达能力。

为加速训练收敛并防止梯度消失，每个子层（多头注意力或前馈网络）均采用残差连接与层归一化，如式（12）和式（13）所示。

$$Z_{attn}=\text{LayerNorm}(X+\text{MultiHead}(X)) \quad (12)$$

$$Z_{out}=\text{LayerNorm}(Z_{attn}+\text{FFN}(Z_{attn})) \quad (13)$$

式中：X 为子层的输入；$Z_{out} \in R^{n\times d_{model}}$ 为单个编码器层的输出。本文堆叠 3 层编码器，逐层传递特征。

编码器输出经全局平均池化聚合为固定长度的特征向量，再通过线性分类层输出各类别概率，如式（14）所示。

$$\hat{y}=\text{softmax}\left(\text{GAP}\left(Z^{(3)}\right)\cdot W_{cls}+b_{cls}\right) \quad (14)$$

式中：$Z^{(3)} \in R^{n\times d_{model}}$ 为第 3 层编码器的输出；GAP 为全局平均池化，沿序列维度取均值，将序列特征压缩为 d_{model} 维向量；$W_{cls} \in R^{d_{model}\times C}$ 为分类权重矩阵；$b_{cls} \in R^C$ 为分类偏置向量；$C=12$ 为手语类别数。

1.5 训练策略

模型训练采用交叉熵损失函数，如式（15）所示。

$$L=-\frac{1}{N}\sum_{i=1}^{N}\sum_{c=1}^{C}y_c^{(i)}\log\left(\hat{y}_c^{(i)}\right) \quad (15)$$

其中 y_c 为真实类别的 one-hot 编码，$\hat{y}_c$ 为模型预测的概率分布。使用 Adam 优化器，初始学习率设为 0.001，批大小为 32，训练轮数为 30。数据集按 8∶2 的比例随机划

分为训练集与验证集，并采用分层抽样以保证各类别分布均衡。

2 实验结果与分析

2.1 实验设置

实验硬件与软件配置如表 2 所示。

表 2 实验环境配置

项目	配置
操作系统	Windows 11
CPU	Intel Core i7
GPU	NVIDIA RTX 4060
内存	16 GB
深度学习框架	PyTorch 2.0
Python版本	3.9
CUDA版本	11.7

2.2 对比实验结果

为验证所提轻量化 Transformer 模型的有效性，本文与一维卷积网络（CNN1D）、长短时记忆网络（LSTM）以及门控循环单元（GRU）三种经典时序模型进行了对比实验，结果如表 3 所示。

表 3 不同模型性能对比

模型	准确率(%)	精确率	召回率	F1分数	参数量	训练时间(s)
CNN1D	98.63	0.9878	0.9863	0.9862	1,413,132	6.3
LSTM	65.75	0.5270	0.4932	0.4998	2,497,804	14.0
GRU	84.93	0.8431	0.7397	0.7498	1,906,956	13.2
Transformer	98.63	0.9766	0.9726	0.9720	612,620	11.7

从表 3 可以看出：（1）CNN1D 与 Transformer 模型均取得了 98.63% 的超高准确率，显著优于 LSTM（65.75%）与 GRU（84.93%）；（2）本文提出的 Transformer 模型在保持高性能的同时，参数量仅为 61.2 万，分别是 CNN1D、LSTM 和 GRU 的 43.4%、24.5% 和 32.1%，体现了其显著的轻量化优势；（3）LSTM 在该任务上表现较差，推测是由于其难以有效捕捉手语序列的长距离依赖关系。

Transformer 模型在验证集上的混淆矩阵如图 4 所示。可见大部分类别均能被准确识别，仅“你好”与“谢谢”两类之间存在少量混淆，可能与两类手势在运动轨迹上具有一定相似性有关。

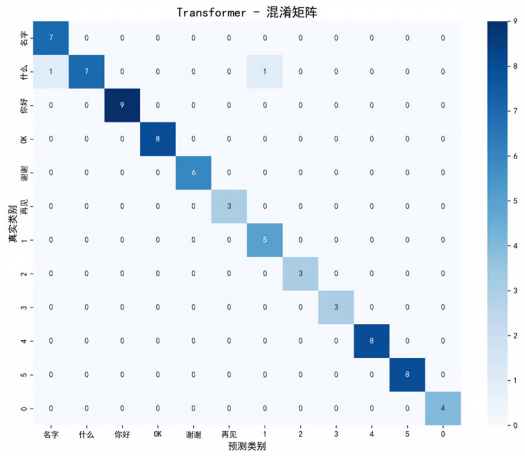


图4 Transformer模型混淆矩阵

各模型在训练过程中的曲线对比如图 5、图 6、图 7 所示。可以看出，Transformer 与 CNN1D 模型收敛较快且训练过程稳定，而 LSTM 模型收敛缓慢且波动较大，进一步说明其在该时序建模任务上的局限性。

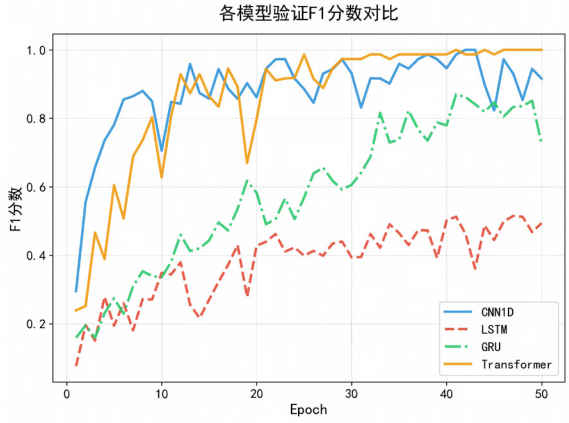


图5 各模型验证F1分数对比曲线

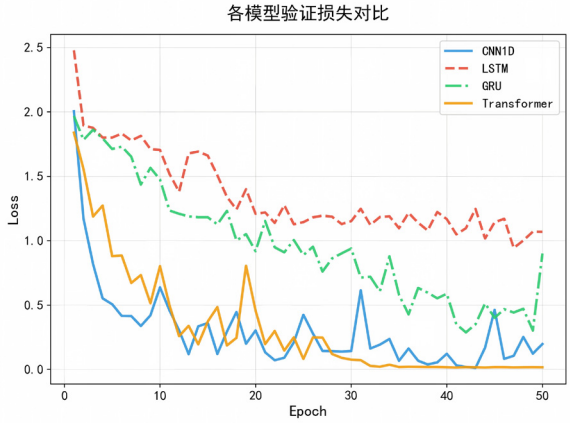


图6 各模型验证损失对比曲线

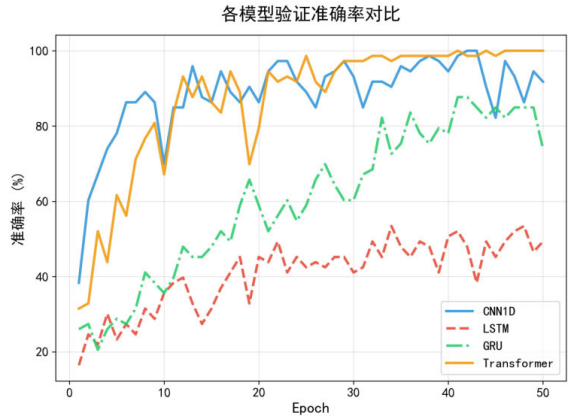


图7 各模型验证准确率对比曲线

2.3 消融实验分析

为探究各最优超参数的配置对模型性能的影响，本文对系统进行了消融实验，重点关注模型维度（d_{model}）、注意力头数（n_{head}）、Transformer 层数（n_{layers}）以及序列长度（L_{seq}）四个关键因素，结果汇总于表 4 所示。

表 4 消融实验结果

实验项	配置	准确率(%)	F1分数	参数量
d_{model}	64	98.63	0.9468	158,860
	128	98.63	0.9727	612,620
	256	94.52	0.7746	2,404,876
n_{head}	2	97.26	0.9732	–
	4	98.63	0.9443	–
	8	98.63	0.9864	–
n_{layers}	1	95.89	0.9591	216,076
	2	100.00	0.9201	414,348
	3	98.63	0.9590	612,620
L_{seq}	20	98.63	0.9592	–
	30	97.26	0.8969	–
	40	100.00	0.9864	–

消融实验结果表明：（1）模型维度为 128 时，在性能与效率之间达到最佳平衡，继续增大到 256 会导致过拟合，性能下降；（2）注意力头数为 8 时，模型获得最高分数（0.9864），说明适度的多注意力机制有助于捕捉手语的多维度特征；（3）编码器层数为 2 时，虽然准确率达到 100%，但 F1 分数相对较低，可能存在过拟合风险，3 层结构在准确率与 F1 分数上表现更为稳健；（4）序列长度为 40 帧时，性能最优，但考虑到实时性要求，20 帧在精度与计算开销之间提供了较好的折衷。

如图 8、图 9、图 10、图 11 所示，综合展示了各参数变化对模型准确率与 F1 分数的影响趋势。

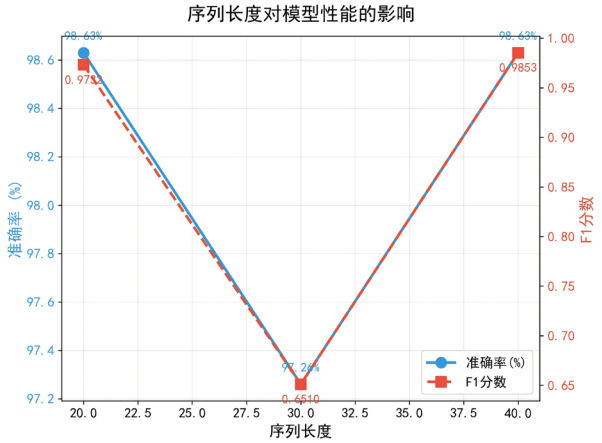


图 8 序列长度对模型性能的影响

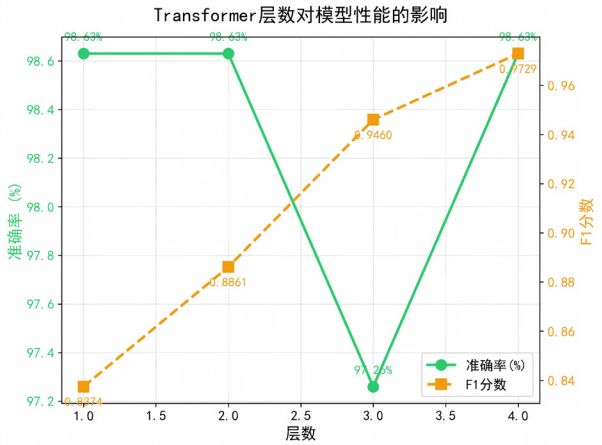


图 9 Transformer层数对模型性能的影响

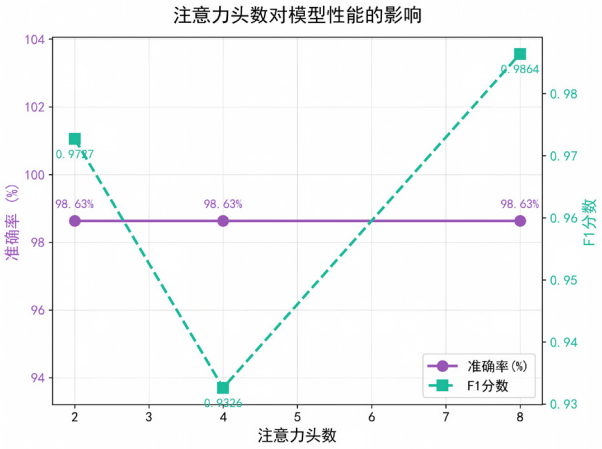


图 10 注意力头数对模型性能的影响

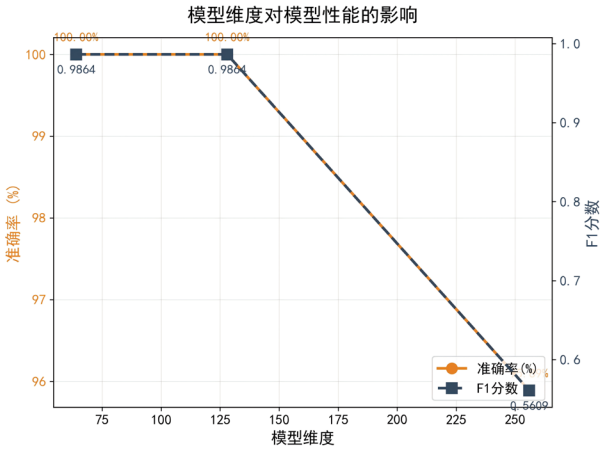


图 11 模型维度对模型性能的影响

2.4 结语

实验分析可得到以下结论：（1）模型性能并非随参数规模单调增长，过大的模型维度或过深的网络层数易导致过拟合，说明“更大”不一定“更好”；（2）仅使用手部关键点特征即可实现高精度识别，验证了剔除身体其余部分冗余信息的有效性，为极致轻量化提供了方向；（3）多头注意力机制能够并行捕捉手语序列中不同子空间的特征关联，是提升模型判别能力的关键；（4）在轻量化设计中需综合考虑准确率、F1 分数、参数量及推理速度等多方面指标，单一指标的优化可能掩盖模型在泛化能力或效率上的缺陷。

3 结语

本文面向边缘设备部署需求，提出了一种基于轻量化 Transformer 的实时手语识别方法。通过采用 MediaPipe 提取手部关键点作为紧凑特征表示，并设计参数量仅为 61.2 万的 3 层 Transformer 编码器，在自建的中国手语数据集上实现了 98.63% 的识别准确率与 0.9864 的 F1 分数，其参数量不足 LSTM 模型的四分之一。对比实验与消融实验充分验证了所提方法在精度与轻量化方面的综合优势。未来工作将围绕以下方面展开：

（1）扩展数据集的规模与类别覆盖，增加连续手语及句子级样本；

（2）研究从孤立词识别向连续手语识别及手语翻译任务的拓展；

（3）探索更加极致的模型压缩与加速技术，推动该系统在智能眼镜、移动终端等实际可穿戴设备上的落地应用。

参考文献：

[1] CAMGOZ N C, KOLLER O, HADFIELD S, et al. Sign language transformers: joint end-to-end sign language recognition and translation[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. IEEE, 2020: 10023-10033.

[2] RASTOGI U, MAHAPATRA R P, KUMAR S. YOLOv10-ST: A hybrid model integrating Swin Transformer for improved sign language recognition[J]. Scientific Reports, 2024, 14(1): 18496.

[3] 宋浩翔 , 张旭 , 沈宏晔等. 基于 MediaPipe 的手语识别系统[J].物联网技术 ,2025,15(10):4-6.

[4] 路飞 , 韩祥祖 , 程显鹏等. 基于轻量 3D CNNs 和 Transformer 的手语识别[J].华中科技大学学报 (自然科学版),2023,51(5):13-18.

[5] 王琪. 基于深度学习的手语识别系统设计与实现[D].西安电子科技大学 ,2021.

[6] 刘星辰 , 杨瑞 , 刘林鑫等. 基于深度学习的中国通用手语识别系统[J].电脑与电信 ,2024,(11):43-47.DOI:10.15966/j.cnki.dnydx.2024.11.006.

[7] 马旭冉 , 陈圣贤 , 李熙文等. 基于改进 YOLOv5 算法的智能手语翻译方法研究[J].电脑知识与技术 ,2025,21(24):36-39.

[8] 王雅锐. 基于卷积神经网络的手势识别[D].安徽理工大学 , 2025.

[9] 岳伟.基于三维骨架信息的连续手语识别研究[D].长安大学 , 2024.

[10] 边辉 , 孟畅乾 , 李子涵等. 基于图卷积网络和 CTC/Attention 的连续手语识别[J].计算机科学 , 2025,52(S1):562-570.

[11] 杨逸峰. 融合手势姿态特征的传感器手语识别模型[D]. 江西师范大学 , 2024.

[12] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[C]//Advances in Neural Information Processing Systems. 2017: 5998-6008.

[13] 王祉元. 基于 Transformer 的手语孤立词识别方法研究[D]. 哈尔滨工程大学 , 2023.

[14] 格梦茹. 基于 Transformer 的手语翻译研究[D]. 天津理工大学 , 2024.

[15] 孙意翔. 基于异构注意力机制的手语翻译方法研究[D]. 中南大学 , 2023.

[16] 刘凯. 基于多模态数据融合的手语识别研究[D]. 江西师范大学 , 2025.

[17] ALSHARIF B, ALALWANY E, IBRAHIM A, et al. Real-time American sign language interpretation using deep learning and keypoint tracking[J]. Sensors, 2025, 25(7): 2138.

[18] 屈乐乐 , 洪雨云.基于 Transformer 元学习网络的毫米波雷达手势识别方法[J].雷达科学与技术 ,2025,23(4):407-416.

[19] 李沼洁 , 郭家毅 , 杨英东等. 融合 Transformer 和 CNN 的伪装目标检测[J].福建理工大学学报 , 2025, 23(6):557-563.

[20] 范忠岗 , 吴跃腾 , 巴顿等. 基于 Transformer 与强化学习的叶片与机匣处理一体化优化设计[J]. 工程热物理学报 , 2025,46(12):3995-4004.

基金项目：大学生创新创业训练计划项目，项目的名称是《一种背带式智能穿戴设备》，项目编号是 202613766002。

作者简介：邓宇航（2005-），男，本科在读，研究方向：机器视觉。

* 通讯作者：王海涛（1997-），男，助教，工学硕士，研究方向：机器视觉，运动轨迹优化。