1 数据集获取与预处理

1.1 任务背景与数据集构建动机

在计算机视觉领域，**手语识别（Sign Language Recognition, SLR）**与传统的**手势识别（Gesture Recognition）**存在本质区别。手势识别通常关注静态的手部姿态（如“握拳”、“V手势”）或简单的短时指令，往往缺乏语言学属性。而手语是一门完备的视觉语言，具有复杂的语法结构、丰富的词汇语义以及长时序的动态演变特性。一个完整的手语动作不仅依赖于手型，还高度依赖于手部的运动轨迹、空间位置变化以及双手间的协同配合。

尽管目前存在如NVGesture、EgoGesture等手势数据集，但针对**中国手语（Chinese Sign Language, CSL）**的高质量公开数据集相对稀缺。现有的CSL数据集大多采集于受控的实验室环境，背景单一且缺乏多样性的光照和视角变化，难以满足模型在真实场景下的泛化需求。此外，针对特定课程任务的小样本词汇训练集更是空白。

鉴于此，为了构建一个既符合中国手语语言习惯，又具有一定鲁棒性的验证基准，本研究决定自主搭建手语视频采集系统，构建专用的手语数据集。这不仅能够保证数据的针对性，也为后续验证双流CNN-LSTM融合网络在动态时序建模上的有效性提供了数据基础。

1.2 数据集的采集与分布

本项目依据《中国手语》标准词汇，选取了日常交流中频率最高的6个核心词汇作为识别对象，分别为“好”、“他”、“早上”、“谢谢”、“很”和“你”。

数据采集过程邀请了多位志愿者参与，尽可能覆盖不同的打手语习惯（如动作幅度、速度）。视频录制设备采用标准RGB摄像头，分辨率为 1920×1080，帧率保持在 30fps。最终经过人工清洗，剔除模糊严重或动作不完整的片段，共获取有效原始样本 **318** 条。各类别的样本分布如表 1 所示。

表 1: 自建原始数据集样本数量分布

手语词汇	好	他	早上	谢谢	很	你	总计
样本数量	49	88	48	36	49	48	318

1.3 基于 MediaPipe 的动态特征提取

为了从冗余的视频像素中提取紧凑且富有判别力的语义信息，本研究设计了基于骨骼关键点的特征提取流水线。

1.3.1 手部关键点提取与空间拓扑构建

我们利用 **MediaPipe Hands** 框架对视频序列进行逐帧解析。相较于传统的OpenPose，MediaPipe在移动端和轻量级设备上具有更高的推理效率。

1. **检测配置：** 设置 `max_num_hands=2` 以支持双手交互动作的捕捉，并关闭 `static_image_mode`（设为 False），利用上一帧的检测结果指导当前帧的追踪，从而显著减少抖动并提高时序连贯性。

2. **关键点定义：** 系统每帧输出每只手的 **21 个** 骨骼关键点坐标 $\mathbf{p} = (x, y, z)$。其中，x, y 为归一化平面坐标，z 为以腕关节为原点的相对深度坐标。

为了显式地建模手部物理结构，我们基于生物学解剖特征，定义了一个 21×21 的邻接矩阵 $\mathbf{A}$，描述了手腕到指尖的 20 条骨骼连接。这为后续基于图结构的特征分析奠定了拓扑基础。

1.3.2 基于能量阈值的自适应动作分割

原始录制视频通常包含动作前后的预备动作（Pre-stroke）和撤回动作（Post-stroke），以及动作中间的自然停顿。为了精确截取语义动作段，我们提出了一种**基于运动能量且具备停顿容忍度（Pause-Tolerant）**的自动分割算法。

算法核心步骤如下：

1. **瞬时能量计算：** 定义第 t 帧的运动能量 E_t 为双手所有关键点相对于上一帧的位移向量的 L_2 范数之和：

$$E_t = \sum_{h \in \{L,R\}} \sum_{i=1}^{21} \|\mathbf{p}_{h,i}^{(t)} - \mathbf{p}_{h,i}^{(t-1)}\|_2^2 \tag{1}$$

   其中 $\mathbf{p}_{h,i}^{(t)}$ 表示第 t 帧第 h 只手第 i 个关键点的坐标。

2. **能量平滑：** 为消除检测噪声引起的能量抖动，对 E_t 序列进行窗口大小为 3 的移动平均平滑。

3. **自适应边界判定：** 设定自适应阈值 $\tau = 0.5 \cdot \text{mean}(E) + \epsilon$。

   - 当 $E_t > \tau$ 时，判定动作开始。
   - 当 $E_t < \tau$ 时，并不立即判定结束，而是启动计数器 C_{pause}。只有当连续静止帧数 $C_{pause} > \text{max\_pause\_len}$（本实验设为25帧）时，才判定动作真正结束。

这一机制有效解决了复杂手语动作中因短暂亦扬顿挫而被错误切分为多个片段的问题。

1.3.3 时序统计特征向量构建

为了降低计算复杂度并保留动作的统计分布特性，我们将原始的 21×3 坐标点压缩为高层的统计特征。对于每一个动作区间，逐帧提取 12 维特征向量 V_t：

$$V_t = [\mu_L, \sigma_L^2, \mu_R, \sigma_R^2]^\top \in \mathbb{R}^{12} \tag{2}$$

其中 $\mu \in \mathbb{R}^3$ 和 $\sigma^2 \in \mathbb{R}^3$ 分别代表手部关键点群在 x, y, z 轴上的均值和方差。均值特征反映了手部在画面中的**绝对位置**轨迹，而方差特征则隐含了手势的**张开程度和形态变化**信息。

1.4 时序规范化与多策略数据增强

1.4.1 时序归一化

由于不同志愿者打手语的语速存在差异，导致提取出的动作序列长度不一。为满足神经网络固定输入维度的要求，本研究采用**重采样技术**将所有样本的时序长度统一为 $T = 30$ 帧。

- 对于短序列（$L < 30$），采用**末帧填充（Padding）**策略，保持动作语义的完整性。
- 对于长序列（$L > 30$），采用**线性插值采样**，保证关键动作帧不丢失。

1.4.2 时空数据增强机制

深度学习模型的性能高度依赖于数据的规模与多样性。考虑到自建数据集仅有318个样本，极易导致模型过拟合。因此，本研究设计了一套包含四种变换策略的**时空数据增强（Spatio-Temporal Augmentation）**方案，旨在模拟真实应用场景中的各种干扰因素。通过不同随机种子的组合，将数据集扩充至 **2862** 条（扩增倍数 9×）。具体增强方法如下：

1. 时间拉伸与压缩 (Time Warping) 模拟不同用户语速的快慢差异。通过对时序索引 t 进行非线性变换 $t' = \alpha t + \beta$，再通过插值重新生成特征序列。这迫使模型学习动作的内在演变规律，而非简单记忆帧与帧之间的绝对对应关系，从而提升对语速变化的鲁棒性。

2. 坐标高斯噪声扰动 (Gaussian Jittering) 模拟摄像头成像噪声或MediaPipe检测的微小抖动。在每个关键点特征 $v_{t,i}$ 上叠加高斯噪声 $\epsilon \sim \mathcal{N}(0, \sigma^2)$：

$$v'_{t,i} = v_{t,i} + \epsilon \tag{3}$$

这不仅增强了模型对输入噪声的容忍度，也相当于在特征空间中对样本分布进行了平滑，防止模型陷入局部极小值。

3. 时间平移 (Time Shift) 模拟动作检测模块可能存在的边界定位误差（即动作开始/结束时间的判定偏移）。通过将整个特征序列沿时间轴向前或向后平移 Δt 帧，并在空缺处进行补零或边缘填充。这训练了模型在非严格对齐的情况下识别关键动作模式的能力。

4. 空间旋转变换 (Spatial Rotation) 模拟摄像头拍摄角度的偏差或用户手掌朝向的多样性。构建二维旋转矩阵 $R(\theta)$ 对手部 x, y 坐标特征进行变换：

$$\begin{bmatrix} x' \\ y' \end{bmatrix} = \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} \tag{4}$$

其中旋转角度 θ 在一定范围内随机采样（如 $[-15°, 15°]$）。这极大地提升了模型对视角变化的泛化能力。

图 1 展示了在原始手语关键点序列的基础上，经过不同数据增强策略处理后，右手食指指尖在 X 轴方向上的轨迹变化情况。原始轨迹通常存在轻微抖动、不规则波动或由于采集环境造成的噪声；而通过加入 **时序平滑、随机扰动、时间尺度拉伸/压缩、空间仿射变换** 等增强方式，可以有效模拟不同使用者的手势差异、采集设备的轻微偏移，以及动作执行速度的变化。对比图中可以观察到：

1. **原始轨迹**：数值分布较为集中，但局部波动明显。
2. **平滑增强后轨迹**：去除了高频噪声，整体走势更加平稳，有利于模型捕捉动作的主趋势。
3. **随机扰动增强后轨迹**：在主轨迹基础上加入轻微随机偏移，增强模型对个体差异的鲁棒性。
4. **时间拉伸/压缩增强后轨迹**：轨迹走势一致，但在时序维度上被重新采样，模拟不同速度的动作执行方式。
5. **空间变换增强后轨迹**：整体曲线形状保持一致，但位置有小幅平移或缩放，用于模拟摄像头角度与距离变化。

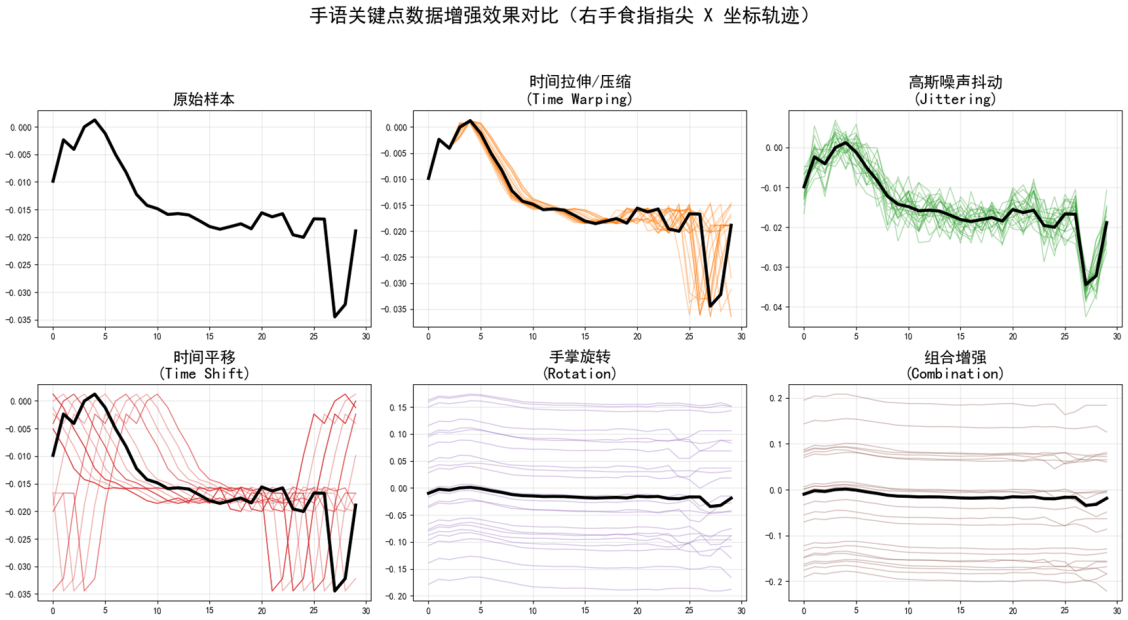

图 1: 手语关键点数据增强效果对比(右手食指指尖X坐标轨迹)

通过这些增强方式，最终得到的轨迹在保持动作语义不变的前提下显著增加了样本的多样性，有助于模型提升对实际复杂场景中手语动作的泛化能力。

综上所述，通过上述严谨的数据采集、精细的特征工程以及全面的数据增强，我们构建了一个高质量、抗干扰能力强的手语数据集，为后续模型的训练奠定了坚实基础。