Computer Science and Application 计算机科学与应用, 2026, 16(11), 156-168
Published Online November 2026 in Hans. https://www.hanspub.org/journal/csa
https://doi.org/10.12677/csa.2026.1611333

Hans汉斯

基于时序卷积密度回归的跳绳视频计数方法

方睿恒[1*]，王少波[1]，智　宇[1]，陈　昂[1,2#]

[1]温州大学计算机与人工智能学院，元宇宙与人工智能研究中心，浙江　温州
[2]温州大学元宇宙与人工智能研究院，浙江　温州

收稿日期：2026年10月7日；录用日期：2026年11月6日；发布日期：2026年11月13日

摘　要

针对人工记录跳绳次数易受速度变化、短暂停顿和观察疲劳影响的问题，本文提出基于姿态特征与一维时序卷积的跳绳视频计数方法。在重叠窗口内预测非负逐帧密度，经同帧平均与积分得到完整视频次数。在RopeAnno-134 (134段视频、19,448次跳跃)固定划分上，5个随机种子运行的平均绝对误差(MAE)为1.1925 ± 0.0867次，一跳以内准确率(OBO)为0.5143 ± 0.0762。补充69名被试的留一被试交叉验证后，视频微平均MAE为5.4229 ± 0.3143次，多数被试维持较小误差，少数较大偏差提示跨被试适应性仍有提升空间。计数网络含149,697个参数，含姿态提取的离线视频处理吞吐率为87.06 ± 0.71帧/s。该方法为姿态驱动的跳绳视频计数提供了结构简洁、结果可追踪的基线。

关键词

跳绳视频计数，重复动作计数，姿态特征，计数密度回归，时序卷积

Jump-Rope Video Counting via Temporal Convolutional Density Regression

Ruiheng Fang[1*], Shaobo Wang[1], Yu Zhi[1], Ang Chen[1,2#]

[1]Research Center for Metaverse and Artificial Intelligence, College of Computing Science and Artificial Intelligence, Wenzhou University, Wenzhou Zhejiang
[2]Institute of Metaverse and Artificial Intelligence, Wenzhou University, Wenzhou Zhejiang

Received: October 7, 2026; accepted: November 6, 2026; published: November 13, 2026

Abstract

Manual jump-rope counting is susceptible to speed changes, brief pauses, and observer fatigue. We

*第一作者。
#通讯作者。

文章引用: 方睿恒, 王少波, 智宇, 陈昂. 基于时序卷积密度回归的跳绳视频计数方法[J]. 计算机科学与应用, 2026, 16(11): 156-168. DOI: 10.12677/csa.2026.1611333

propose a method using pose features and one-dimensional temporal convolution. Non-negative frame-wise densities predicted in overlapping windows are averaged at shared frames and integrated to obtain full-video counts. On the fixed split of RopeAnno-134 (134 videos and 19,448 jumps), five random-seed runs achieve a mean absolute error (MAE) of 1.1925 ± 0.0867 jumps and an off-by-one accuracy (OBO) of 0.5143 ± 0.0762. Supplementary leave-one-subject-out cross-validation over 69 subjects yields a video-micro MAE of 5.4229 ± 0.3143 jumps. Most subjects retain relatively small errors, while larger deviations in a minority indicate room to improve cross-subject adaptability. The counting network has 149,697 parameters, and offline video processing including pose extraction reaches 87.06 ± 0.71 frames/s. The method provides a simple and traceable baseline for pose-driven jump-rope video counting.

Keywords

Jump-Rope Video Counting, Repetition Counting, Pose Features, Count-Density Regression, Temporal Convolution

Copyright © 2026 by author(s) and Hans Publishers Inc.
This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).
http://creativecommons.org/licenses/by/4.0/
Open Access

1. 引言

跳绳次数是训练量记录、课程考核和运动反馈中的基础指标。人工计数虽然直观，却会随动作加速、短暂停顿、遮挡及观察者注意力变化而产生偏差。已有研究分别利用穿戴式传感器或骨架序列开展运动重复计数[1]-[3]；前者能够直接测量周期运动，但增加了设备依赖和使用约束，后者则可由深度相机或普通视频中的人体姿态估计获得。普通视频具有获取方便、可回看和非接触等优点，因此，从单段视频中稳定估计完整跳跃次数具有明确的应用价值。

跳绳计数并不是简单的逐帧分类。一次跳跃持续时间短，视频内节奏可能改变，摆臂、屈膝或身体晃动又可能形成与有效跳跃相似的局部变化。若只检测瞬时峰值，阈值和最小间隔容易随速度变化失效；若直接把整段视频压缩为一个数值，模型又难以保留局部动作证据，并需要处理任意时长输入。重复动作研究已经从模板匹配、周期分析发展到深度时序建模和密度回归[1]-[8]，但面向规模有限、以视频级计数为最终目标的数据，仍需要兼顾结构复杂度、监督目标和长序列累计误差。

本文采用逐帧姿态与运动特征，使用一维时序卷积预测可加的非负计数密度。已有重复动作计数研究表明，密度回归能够利用细粒度周期标注生成连续响应，并通过积分读出动作次数[8]。在此基础上，本文使每次跳跃在时间轴上贡献单位质量，局部窗口预测经重叠融合后在完整视频上积分，因而无需峰值检测或视频特定阈值。与以帧级密度形状为核心的训练方式不同，本文主模型直接优化窗口内预测总量与真实跳跃质量之间的差异，使训练目标与视频计数评价更一致。

本文的主要贡献如下：

1) 构建从姿态特征、单位质量标签、重叠窗口预测到完整视频积分的跳绳计数流程，使局部时序证据与最终计数保持可追踪的加性关系。

2) 设计由四个无空洞、无残差卷积块组成的轻量时序网络，并以窗口级计数损失直接约束密度总量。

3) 在统一数据划分和训练协议下，通过 A0~A6 单因素消融分别检验帧级监督、空洞率、残差相加、窗口重叠和通道宽度，给出与当前测试结果相匹配的结论边界。

2. 相关工作

重复动作计数可按输入表征与时间建模方式归纳为传感器或骨架序列建模、直接视频时序建模以及密度回归三条技术路线。传感器方法利用加速度、角速度等低维周期信号，适合可穿戴训练记录；骨架方法把人体表示为关节点及其时空关系，可降低背景和服装变化的干扰[1]-[3]。二维姿态估计网络为从普通视频获得稳定关节点轨迹提供了基础[9]-[11]，循环网络和时空图卷积则进一步描述关节结构及其时间依赖[12]-[14]。这类方法的共同优势是弱化外观变化，但计数性能仍受传感器部署条件或上游姿态质量制约。

直接视频方法从外观与运动信息中学习重复结构，其中双流网络、时序分段网络和三维卷积网络分别形成了运动建模、稀疏采样与时空卷积的代表性范式[15]-[17]。在重复动作计数任务中，相关方法进一步采用卷积特征累计、时频分析、上下文建模和视频内部自相似关系，以适应未知类别或变速运动[4]-[7]；密度回归则把重复动作表示为可积分的连续响应，使局部预测能够组合为长序列计数[8]。相较于 RGB 端到端模型，姿态输入降低了外观维度但依赖上游关键点质量；相较于全局注意力结构，时序卷积具有并行计算、局部归纳偏置明确和参数规模易控制等特点[18] [19]。

本文关注的不是构造更复杂的通用重复计数器，而是在固定跳绳数据和姿态输入下检验“什么监督与结构足以支持准确积分”。因此，方法保留密度表示的可加性，使用短程串联卷积形成局部动作表征，并通过严格单因素实验区分训练目标、时间范围和融合方式的作用。

3. 方法

3.1. 总体框架

设视频 v 包含 T_v 帧，逐帧输入为 $X^{(v)} \in \mathbb{R}^{(T_v \times 149)}$，真实跳跃次数为 $c^{(v)}$。整体流程如图 1 所示：视频首先转换为姿态与运动特征序列，再按 160 帧窗口和 64 帧步长切分；时序卷积网络为每个窗口输出非负密度；重叠位置的预测映射回原始时间轴并取平均；最终计数为融合密度在所有有效帧上的和。该流程在训练和推理阶段使用一致的窗口定义，避免因分段规则变化引入额外偏差。

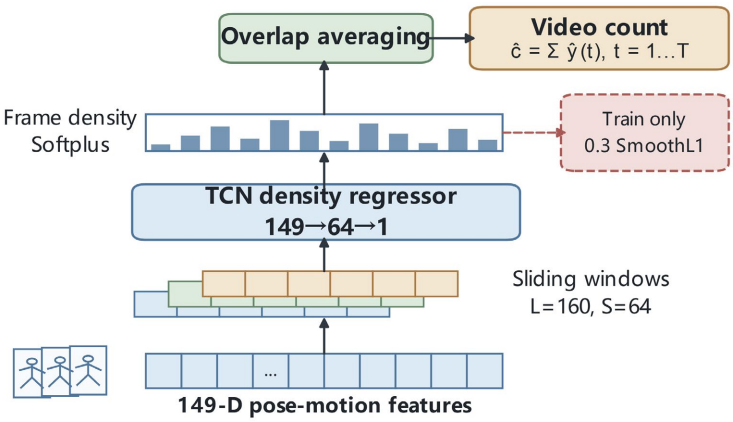


Figure 1. Overall framework of the jump-rope video counting method. The 149-dimensional pose-motion features are divided into overlapping windows of length $L = 160$ and stride $S = 64$. Non-negative frame-wise densities are regressed by a temporal convolutional network (TCN); predictions at overlapping frames are averaged and integrated over time to obtain the video count. The dashed branch is used only during training, and 0.3 Smooth L1 denotes the window-level count loss with a weight of 0.3

图 1. 跳绳视频计数方法总体框架。149 维姿态－运动特征按长度 $L = 160$、步长 $S = 64$ 划分为重叠窗口，经时序卷积网络(TCN)回归逐帧非负密度；重叠帧预测取平均后沿时间积分得到视频计数。虚线分支仅在训练阶段使用，0.3 Smooth L1 表示权重为 0.3 的窗口级计数损失

3.2. 输入特征与计数密度标签

姿态提取采用 Ultralytics YOLOv8n-pose 模型[20]，关键点置信度阈值设为 0.3；当画面中检测到多个人体时，保留平均关键点置信度最高的实例。每帧 149 维特征由 34 维归一化关键点坐标、34 维速度、34 维加速度、5 维关节角度、7 维运动汇总量、17 维关键点置信度、17 维有效性掩码和 1 维有效关键点比例组成。为减小不同量纲对优化的影响，连续特征按训练集统计量标准化：

$$\tilde{\sigma}_j = \sigma_j \left(\sigma_j \geq 10^{-8}\right), \text{otherwise 1};\ \tilde{x}(t,j) = \left[x(t,j) - \mu_j\right] / \tilde{\sigma}_j \tag{1}$$

标准化统计量仅由训练集估计，并原样用于验证集与测试集。缺失或低可信关键点由既定特征管线处理，网络接收固定维度序列，不在测试视频上重新估计尺度参数。

对第 i 次跳跃，依据相位标注确定其中心 τ_i，并在中心附近构造截断高斯核 $g_i(t)$。每个核在有效支持域 Ω_i 内归一化为单位质量：

$$g_i(t) = \exp\left[-(t-\tau_i)^2 / \left(2\sigma^2\right)\right] / \sum_{u\in\Omega_i} \exp\left[-(u-\tau_i)^2 / \left(2\sigma^2\right)\right] \tag{2}$$

视频级目标密度为各事件核之和 $y_t = \sum_i g_i(t)$，从而满足 $\sum_t y_t = c$。该守恒关系使帧级表示可以通过求和直接还原真实次数。核半径设为 4 帧，标准差为 2.0 帧；其作用是把离散事件中心转化为平滑训练目标，而不是把密度解释为逐帧发生概率。

3.3. 无空洞时序卷积网络

一维时序卷积通过聚合局部时间邻域内的姿态－运动特征，在保持序列长度不变的同时生成逐帧密度表示。主模型所有卷积的膨胀率均为 1，并在时间轴上采用对称填充。输出位置 o 处的局部特征可写为：

$$h(t,o) = b(o) + \sum_c \sum_r W(o,c,r)\, x\left(t + r - \lfloor k/2 \rfloor, c\right) \tag{3}$$

式(3)描述了卷积核覆盖范围内局部时间特征的加权聚合过程。图 2 以 $k = 5$、$d = 1$ 为例展示该局部计算关系。

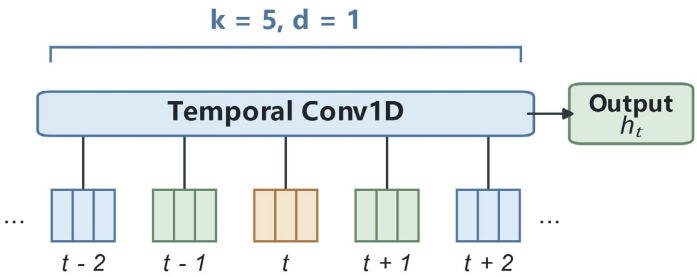


Figure 2. One-dimensional temporal convolution used in the main model. k and d denote the kernel width and dilation rate, respectively. For $k = 5$ and $d = 1$, the output h_t is jointly computed from the local features of five frames from $t - 2$ to $t + 2$. Symmetric padding is used to preserve the temporal resolution

图 2. 主模型采用的一维时序卷积。k 和 d 分别表示卷积核宽度与膨胀率；以 $k = 5$、$d = 1$ 为例，输出 h_t 由 $t - 2$ 至 $t + 2$ 共 5 帧的局部特征共同计算。卷积采用对称填充以保持时间分辨率

主模型以连续局部卷积逐步汇聚时间信息，整体数据流如图 3 所示。起始层以核宽度 7 将 149 维输入映射到 64 个通道，随后串联四个核宽度 5 的卷积块。起始层和各卷积块均采用 8 组 Group Normalization [21]，卷积块进一步使用 GELU 激活[22]和丢弃概率为 0.1 的 Dropout [23]。主模型不进行残差相加；A4 沿相同卷积主路径加入残差连接[24]，仅用于消融比较。最后的 1×1 卷积生成单通道响应 z_t，Softplus 函

数保证预测密度非负：$\hat{y}_t = \ln\left[1+\exp\left(z_t\right)\right]$。

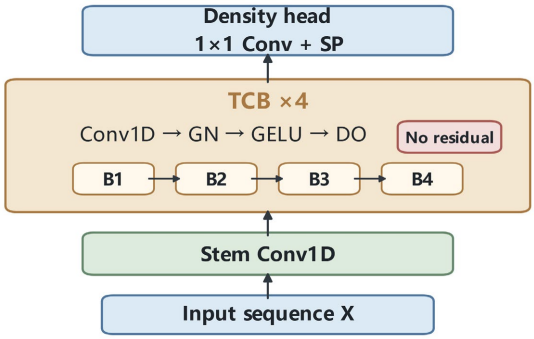


Figure 3. Architecture of the temporal convolutional density-regression network. The input sequence X is first processed by a stem one-dimensional convolution (kernel width 7, 149 → 64), followed by four temporal convolutional blocks (TCBs). Each TCB consists of a 64-channel one-dimensional convolution with kernel width 5 and dilation rate 1, group normalization (GN), GELU activation, and dropout (DO), with no residual connections between blocks. A 1 × 1 convolution and Softplus (SP) finally produce non-negative frame-wise densities

图 3. 时序卷积密度回归网络结构。输入序列 X 首先经起始一维卷积(核宽度 7，149 → 64)，随后依次通过 4 个时序卷积块(TCB)。各 TCB 均由核宽度 5、膨胀率 1 的 64 维一维卷积、组归一化(GN)、GELU 激活与 Dropout(DO)组成，块间不使用残差连接；最后经 1 × 1 卷积与 Softplus(SP)输出逐帧非负密度

主模型的前向过程由起始映射、四个串联时序卷积块和密度读出组成。图 3 展示了各阶段的数据流及卷积块的串联关系；与之对应，表 1 进一步列出各阶段的运算、卷积核宽度、膨胀率和输出通道数，以便复现网络配置。

Table 1. Layer-wise architecture of the main temporal convolutional network

表 1. 主时序卷积网络的逐层结构

阶段	运算	核宽度/膨胀率	输出通道
输入	逐帧姿态与运动特征	—	149
起始层	Conv1D + GroupNorm + GELU	7/1	64
卷积块 1	Conv1D + GroupNorm + GELU + Dropout	5/1	64
卷积块 2	Conv1D + GroupNorm + GELU + Dropout	5/1	64
卷积块 3	Conv1D + GroupNorm + GELU + Dropout	5/1	64
卷积块 4	Conv1D + GroupNorm + GELU + Dropout	5/1	64
输出层	Conv1D + Softplus	1/1	1

3.4. 窗口级训练与完整视频读出

长视频按长度 $L = 160$ 帧、步长 $S = 64$ 帧生成窗口；末尾不足一个窗口时采用尾部对齐并以有效帧掩码排除填充位置。设 m_t 表示有效帧，窗口真实质量为 $q = \sum_t m_t y_t$，预测质量为 $\hat{q} = \sum_t m_t \hat{y}_t$。A0 仅以窗口级 Smooth L1 计数损失训练：

$$L_{\mathrm{A0}} = 0.3\,\mathrm{Smooth\ L1}\left(\hat{q}, q\right) \tag{4}$$

该目标不要求预测曲线在每一帧严格匹配固定高斯形状，而是直接约束局部窗口应包含多少计数质量。为检验帧级形状监督的作用，A1 在式(4)基础上加入权重为 1.0 的有效帧 Smooth L1 密度损失，A2 将计数损失权重置为 0 并仅保留权重为 1.0 的逐帧密度损失；二者只用于消融，不改变 A0 定义。

推理时，设 $W(t)$为覆盖原始帧 t 的窗口集合，同一帧的融合密度为：

$$\bar{y}_t = \frac{1}{|W(t)|}\sum_{w\in W(t)} \hat{y}_t^{(w)},\ \hat{c}^{(v)} = \sum_{t=1}^{T_v} \bar{y}_t \tag{5}$$

先平均再积分可避免重叠帧被重复计数。最终读出不使用峰值阈值、最小峰间距或按视频调节的后处理参数，计数来源可由融合密度逐帧追踪。

窗口融合减少同一帧因上下文位置不同造成的预测差异，视频积分实现任意时长序列的统一读出。局部预测质量对应其对最终次数的贡献，便于追踪持续过计或漏计区间；微小正密度可能沿长序列累积，因此窗口级总量监督具有直接的任务意义。

4. 实验与结果

4.1. 数据集与实验设置

本文将由 134 段跳绳视频构成、并人工标注跳跃次数和相位边界的数据集命名为自建跳绳视频计数数据集 RopeAnno-134。该名称用于标识本文实验数据。RopeAnno-134 共含 19,448 次标注跳跃，固定划分为 94 段训练视频、19 段验证视频和 21 段测试视频。表 2 汇总各划分的视频数、跳跃总数、单视频次数范围、平均时长、平均帧率和平均节奏。所有模型在相同划分上训练，验证集用于选择各随机种子运行的检查点，测试集用于统一评价。

Table 2. Dataset composition and descriptive statistics of RopeAnno-134
表 2. RopeAnno-134 数据集构成与描述统计

划分	视频数	跳跃总数	单视频次数范围	平均时长/s	平均帧率/(帧·s^{-1})	平均节奏/(次/min)
训练集	94	13,786	42~464	69.6042	29.2315	125.3640
验证集	19	2822	70~220	68.6657	27.7806	130.3395
测试集	21	2840	72~216	60.4030	30.7545	134.1243
合计	134	19,448	42~464	—	—	—

三个数据划分在视频时长和平均节奏上呈现出可比的总体范围。在表 2 所示各划分整体统计结果的基础上，图 4 进一步展示了真实计数、视频时长和平均跳绳频率的样本分布。

原始相位标注长度可能为 $2c$、$2c+1$ 或 $2c+2$，其中 c 为人工记录的跳跃次数。对于 $2c$ 标注，程序按相邻低点–高点配对并取帧号均值；对于 $2c+1$ 标注，先丢弃最后一个尾部边界再配对；对于 $2c+2$ 标注，先删除低点–高点间隔最大的整对再配对。派生中心经四舍五入后截断至有效特征帧范围，再用于生成单位质量高斯密度。各划分的标注格式分布见表 3。该步骤只改变标注表示，不改变人工计数总量。

正式只读审计覆盖 RopeAnno-134 全部 134 段视频及其特征文件，未发现文件缺失、非有限特征、派生事件数与人工计数不一致或跨划分完全重复视频等硬错误。审计同时保留 33 条遗留警告，涉及 30 段视频，其中 17 条为原始相位边界重复或逆序，15 条为相位边界超出特征帧范围，1 条为派生中心需要边界截断。实验未重排或人工修正原始相位列，而是统一执行上述冻结转换规则。

训练采用 AdamW 优化器[25]，共 120 轮，批量大小为 8；主要超参数见表 4。为评估随机因素引起的波动，每个配置采用 1001、1002、1003、1004 和 1005 共 5 个随机种子分别训练($n=5$)，并报告跨运行均值 ± 总体标准差。每次运行均在同一 21 段测试视频上分别计算 MAE 和 OBO；重复单位为一个随机种子运行，视频为运行内计算指标的评价单位。各消融臂使用相同训练轮数、数据划分、检查点选择规则和评价代码。

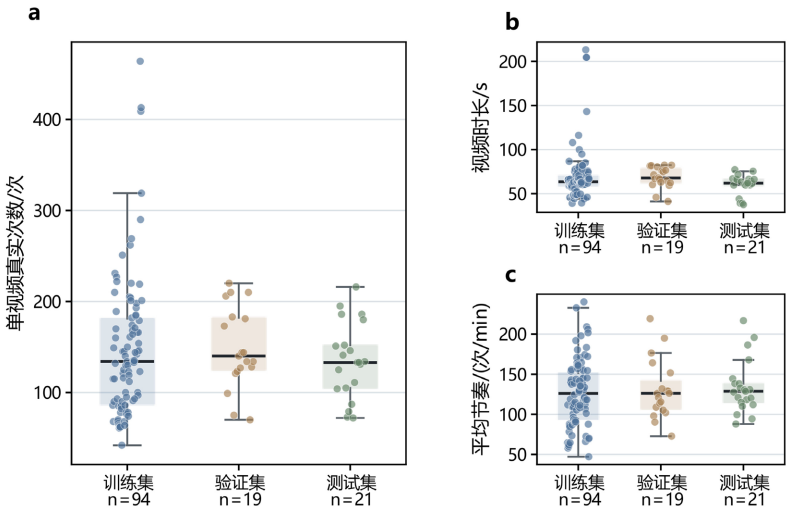


Figure 4. Video-level descriptive statistics of RopeAnno-134. Panel (a) shows the distribution of ground-truth counts for all 134 videos, while panels (b) and (c) show the distributions of video duration and average jump-rope frequency, respectively; colors distinguish the training, validation, and test sets. These statistics characterize the samples in the fixed data split and are not used as evidence of generalization under an independent blind test

图 4. RopeAnno-134 的视频级描述统计。图(a)给出全部 134 段视频的真实计数分布，图(b)和图(c)分别给出视频时长与平均跳绳频率分布；颜色区分训练集、验证集与测试集。上述统计用于说明固定数据划分的样本特征，不作为独立盲测泛化的证据

Table 3. Distribution of phase-annotation formats

表 3. 相位标注格式分布

划分	$2c$	$2c+1$	$2c+2$
训练集	22	71	1
验证集	5	14	0
测试集	4	15	2

Table 4. Training and inference settings

表 4. 训练与推理设置

项目	设置	项目	设置
优化器	AdamW	训练轮数	120
学习率	0.0007	批量大小	8
权重衰减	0.0001	Dropout	0.1
梯度范数上限	1.0	窗口长度	160 帧
窗口步长	64 帧	密度核半径	4 帧
高斯核标准差	2.0	计数损失权重	0.3
随机种子	1001, 1002, 1003, 1004, 1005	学习率调度	无
GroupNorm 分组数	8	逐帧损失权重(A1/A2)	1.0
姿态提取器	YOLOv8*n*-pose	关键点置信度阈值	0.3

实验比较遵循两项一致性原则。其一，当前统一评估阶段的每个运行只依据固定验证集选择检查点，测试结果不用于调整该运行的训练轮数或挑选单次最好结果；但主模型配置是在前期已访问该测试集的研究基础上确定的，因此本文将测试结果限定为固定测试集上的描述性证据。其二，A1~A6 沿用 A0 的数据、优化器、训练预算和完整视频读出，仅改变预先声明的单个概念因素。本文未进行假设检验。总

体标准差只描述 5 个固定随机种子运行之间的离散程度，不代表对新视频总体的抽样不确定性；同种子配对结果用于说明配置差异在当前重复实验中的一致程度。

4.2. 评价指标与主结果

以视频级平均绝对误差(MAE)作为主指标，并以一跳以内准确率(OBO)描述严格命中比例。对 N 个测试视频，两项指标定义为：

$$\mathrm{MAE}=\frac{1}{N}\sum_{v}\left|\hat{c}^{(v)}-c^{(v)}\right|,\ \mathrm{OBO}=\frac{1}{N}\sum_{v}I\left(\left|\hat{c}^{(v)}-c^{(v)}\right|\le 1\right) \tag{6}$$

MAE 越低越好，OBO 越高越好。两者关注不同误差区域：MAE 持续反映误差幅度，OBO 只判断是否处于 1 次误差阈值内。

A0 在 5 个随机种子运行上的测试集 MAE 为 1.1925 ± 0.0867 次，OBO 为 0.5143 ± 0.0762；五次运行的 MAE 范围为 1.0277~1.2782 次，OBO 范围为 0.4286~0.6190。表 5 汇总两项核心指标，表明主模型在不同初始化下保持了较低的平均计数误差。

MAE 反映连续误差幅度，OBO 表示误差不超过 1 次的比例；较低 MAE 不保证每段视频均落入一跳范围。两项指标中 A0 均取得当前消融配置中的最好均值，本文联合报告两项指标，并将固定划分结果限定于该评价条件。

Table 5. Test-set counting results of the main model
表 5. 主模型测试集计数结果

方法	Test MAE/次 ↓	Test OBO ↑
无空洞、无残差时序卷积窗口计数模型(A0)	1.1925 ± 0.0867	0.5143 ± 0.0762

图 5 进一步比较 21 段测试视频的真实次数与跨种子平均预测，并展示 MAE、OBO 及其波动。

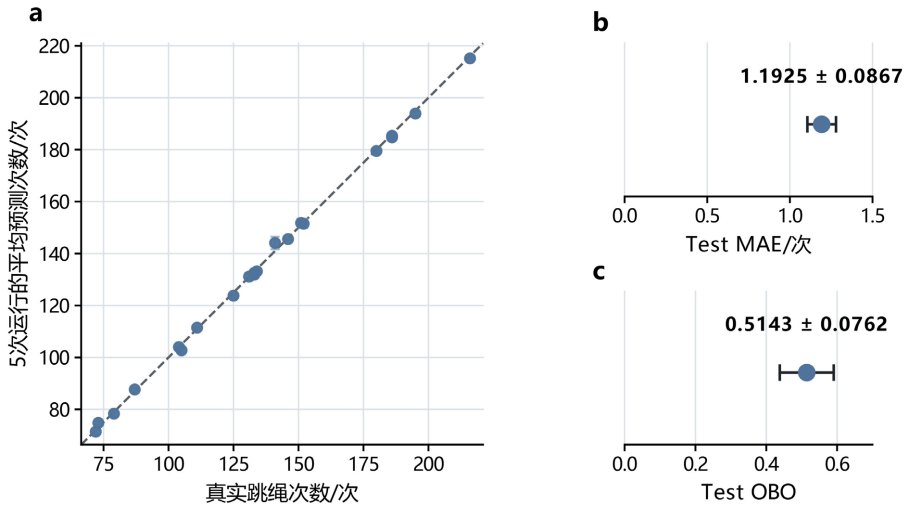


Figure 5. Test-set counting results of the main model. Panel (a) compares the ground-truth counts of 21 test videos with the mean predicted counts across five random-seed runs; vertical error bars denote the population standard deviation of the predicted count for each video. Panels (b) and (c) show MAE and OBO, respectively, where the points and error bars denote the mean and population standard deviation across five random-seed runs (n = 5). These standard deviations do not represent sampling uncertainty for the population of new videos
图 5. 主模型测试集计数结果。图(a)比较 21 段测试视频的真实次数与 5 个随机种子运行的平均预测次数，纵向误差线表示对应视频预测次数的总体标准差；图(b)和图(c)分别给出 MAE 与 OBO，点和误差线表示 5 个随机种子运行(n=5)的均值和总体标准差。这里的标准差均不表示对新视频总体的抽样不确定性

4.3. 单因素消融实验

消融实验以结果最好的 A0 置于首行并将其确定为主模型。A1~A6 均相对 A0 只改变一个概念因素：A1 增加帧级密度损失形成联合监督；A2 用帧级密度损失替换窗口级计数损失；A3 采用膨胀率 1、2、4 和 8 的空洞卷积；A4 增加残差相加；A5 取消窗口重叠；A6 将隐藏通道数减至 32。除 A4 外，其余配置均不使用残差相加。表 6 汇总各配置的 MAE 和 OBO。

Table 6. Single-factor ablation results relative to the main model
表 6. 相对主模型的单因素消融结果

编号	配置说明	Test MAE/次 ↓	Test OBO ↑
A0	**主模型：无空洞卷积，仅窗口级计数损失**	**1.1925 ± 0.0867**	**0.5143 ± 0.0762**
A1	增加逐帧密度损失(联合监督)	1.7850 ± 0.6237	0.4381 ± 0.1061
A2	仅逐帧密度损失	9.5275 ± 1.7543	0.1143 ± 0.0646
A3	采用膨胀率 1、2、4 和 8	1.4074 ± 0.4945	0.5048 ± 0.1367
A4	增加残差相加	1.8970 ± 0.2586	0.4381 ± 0.0762
A5	取消窗口重叠：160 帧步长与分块起点	2.2471 ± 0.6293	0.2952 ± 0.1327
A6	主干通道数由 64 减至 32	1.5165 ± 0.1892	0.4286 ± 0.1313

注：每个配置均为 5 个随机种子运行(n = 5)的测试集均值 ± 总体标准差。A0 为主模型；A1~A6 各自只改变表中所述概念因素。

损失对照给出最明确的差异。A2 移除窗口级计数约束后，MAE 升至 9.5275 ± 1.7543；按 MAE 进行同种子配对时，A2 相对 A0 为 0 胜/0 平/5 负。该结果说明，在当前标签与训练协议下，逐帧拟合固定密度形状不能替代对窗口总质量的直接约束。A1 同时使用两种损失时平均 MAE 也未优于 A0，同种子配对结果为 1 胜/0 平/4 负，提示逐帧形状约束可能与视频级积分目标并不完全一致。上述结果支持“窗口级计数损失是当前系统的关键约束”，但不意味着帧级监督在其他标签形式或定位任务中普遍无效。

结构对照显示，A3 采用递增空洞率后平均 MAE 和波动均高于 A0，同种子配对结果为 2 胜/0 平/3 负；A4 加入残差相加后 MAE 升至 1.8970 ± 0.2586，同种子配对结果为 0 胜/0 平/5 负。由此可见，对当前四块浅层网络，更大的理论时间范围和恒等捷径没有转化为更好的平均视频级积分结果。该结论只针对当前层数、窗口长度和数据规模，不应扩展为对空洞卷积或残差网络的一般否定。

窗口与容量对照进一步说明推理组合方式的重要性。A5 取消重叠后 MAE 为 2.2471 ± 0.6293、OBO 为 0.2952 ± 0.1327；A6 把通道数减半后 MAE 为 1.5165 ± 0.1892。两项对照相对 A0 的同种子 MAE 配对结果均为 0 胜/0 平/5 负。在当前设置下，这些描述性结果提示窗口边界处上下文不足会影响完整视频累计，32 通道也不足以保持 A0 的平均精度；若用于受限设备，还需结合参数量、延迟和能耗重新权衡，而不能仅据计数指标决定部署配置。

图 6 将表 6 的单因素消融结果及跨种子波动可视化，便于比较不同配置。

4.4. 留一被试交叉验证

为进一步考察对未见被试的适应性，作者对全部 134 段视频的人员对应关系进行人工核对，并在补充实验前冻结为 69 个被试组。采用留一被试交叉验证(leave-one-subject-out, LOSO)：每折将一名被试的全部视频作为测试集，其余 68 人按被试编号的确定性哈希排序划分为 54 人训练、14 人验证，三者人员不重叠。标准化参数仅由该折训练视频估计，以验证集 MAE 最小的检查点评价留出被试；各折采用表 4 中的训练设置及 5 个随机种子，共完成 345 次独立训练，每段视频在每个种子下恰好测试一次。

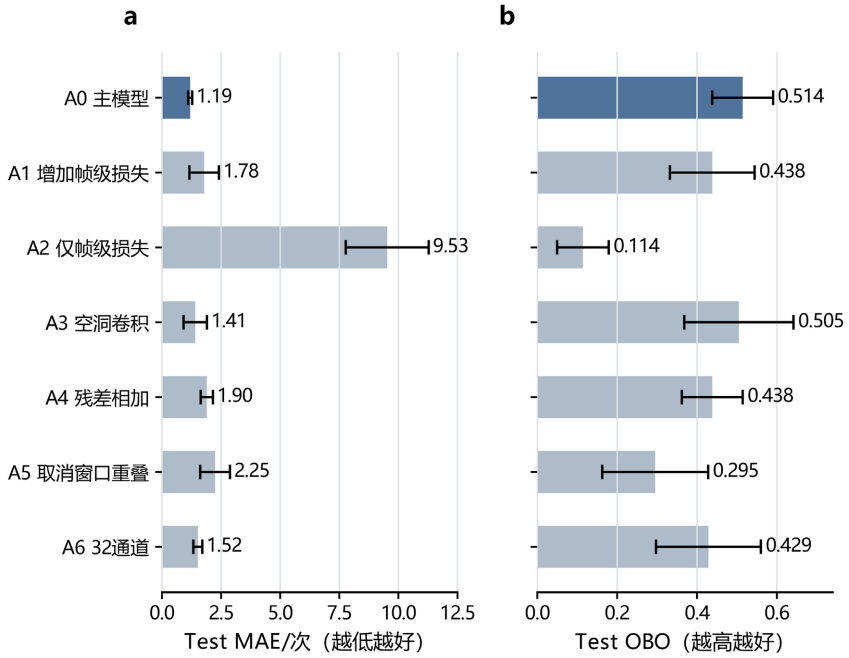


Figure 6. Single-factor ablations of loss, temporal structure, window fusion, and channel width. A0 achieves the best mean MAE and mean OBO; error bars denote the population standard deviations across five random-seed runs (n = 5)
图 6. 损失、时间结构、窗口融合与通道宽度的单因素消融。A0 在平均 MAE 和平均 OBO 上均为最优；误差线表示 5 个随机种子运行(n = 5)的总体标准差

表 7 报告全部 134 段视频的视频微平均结果，以及先对各被试求均值再等权平均的被试宏平均结果。69 人中，46 人的五种子平均 MAE 不超过 3 次，9 人超过 10 次；个别案例存在较大漏计，如 P25 的平均 MAE 为 164.861 次，全部保留在汇总中。上述差异提示，多数被试具有较小计数偏差，但少数困难案例仍影响总体表现。该协议覆盖全部视频，评价范围与原固定划分的 21 段测试视频不同，不能视为同条件优劣比较。

Table 7. Counting results of leave-one-subject-out cross-validation
表 7. 留一被试交叉验证的计数结果

汇总口径	MAE/次 ↓	OBO ↑
视频微平均(134 段)	5.4229 ± 0.3143	0.3552 ± 0.0278
被试宏平均(69 人)	7.7456 ± 0.6120	0.3373 ± 0.0257

注：每个种子均覆盖 69 人、134 段视频。均值 ± 总体标准差以 5 个种子为重复单位(ddof = 0)，不表示新被试总体的置信区间；OBO 按连续预测次数的绝对误差 ≤1 计算，不取整。

4.5. 计算效率

效率测量平台为 Intel Core i5-12400F (6 个计算线程)及 NVIDIA GeForce RTX 4060 Ti(8 GB)，采用 PyTorch 2.12.1+cu126、FP32 并关闭 TF32。表 8 报告 A0 的参数量、卷积计算量及实测速度。单窗口输入为 1 × 160 × 149，卷积计算按一次乘加记 2 次浮点运算，不含归一化、激活与姿态提取器。窗口前向预热 50 次后计时 500 次，输入已驻留对应设备，GPU 计时进行同步。

完整离线流程另从原测试集按视频时长排序选择第 6、11、16 段，各重复 3 次，包含解码、YOLOv8n-pose、特征构建与计数；姿态输入尺寸 640、批量 1，不计模型加载、文件写入和可视化。每遍合并 3 段

视频，以总帧数除以总耗时计算 FPS。

Table 8. Computational cost and inference efficiency on the evaluation platform
表 8. 测试平台上的计算开销与推理效率

指标	CPU	GPU
计数网络参数量	149,697	149,697
单窗口卷积计算量/MFLOPs	47.5955	47.5955
单窗口前向延迟/ms	0.688 ± 0.124	0.757 ± 0.209
完整离线流程吞吐率/(帧/s)	未测	87.06 ± 0.71

注：延迟为 500 次前向的均值 ± 总体标准差；完整流程为 3 遍合并处理的均值 ± 总体标准差。吞吐率反映处理能力，不等同于在线输出延迟。

4.6. 讨论与局限

综合主结果与消融实验，窗口级计数监督、重叠融合与 64 通道主干在当前固定划分下构成了表现较好的组合。逐帧密度保留了局部动作证据与完整视频次数之间的可加关系，便于分析误差积累；但其主要含义是计数贡献，尚不宜直接解释为经校准的起跳或落地概率。

人员独立评估进一步明确了该方法的适用边界。多数被试的平均误差处于较小范围，但少数较大偏差表明，现有姿态表征与模型对不同被试和采集条件的适应性仍有提升空间。这些差异可能与身体运动模式、拍摄视角和姿态观测分布有关，现有诊断尚不足以区分各因素的独立作用。后续可扩充多视角及困难样本，检验视角增强、缺失感知训练和暂停段约束的效果。分组在修订期间回溯核对，数据已用于前期开发，本次评价仍需独立来源数据进一步验证。

系统当前按关键点置信度选择单个人体，未维护跨帧身份，多人同框或交叉遮挡时可能发生目标切换。由于输入特征由姿态估计结果构建，关键点缺失或定位偏差可能通过派生运动特征影响累计计数，真实干扰条件下的适用性仍需进一步评价。因此，单人且姿态可稳定提取的场景仍是当前较适合的应用条件。后续将在其他论文研究中结合姿态质量控制、缺失感知训练及身份跟踪开展进一步改进与验证。

效率结果表明，计数网络本身的计算开销较小，但当前身体尺度依赖视频级统计，卷积采用对称填充，GroupNorm 统计包含窗口时间维，现有实现仍属于离线处理。桌面平台的较高吞吐率为应用提供了计算基础，尚不能等同于低延迟在线计数。实时部署需进一步检验因果特征、滑动缓存和输出滞后，并评价移动设备上的速度与能耗。

5. 结论

本文提出基于姿态特征与一维时序卷积的跳绳视频计数方法，以窗口内非负密度预测、同帧平均及积分实现完整视频计数。固定划分上 5 个种子运行的 MAE 为 1.1925 ± 0.0867 次、OBO 为 0.5143 ± 0.0762；补充 LOSO 的视频微平均 MAE 为 5.4229 ± 0.3143 次，多数被试保持较小偏差，部分困难案例仍有改进空间。效率测量补充了计算成本证据。该方法提供了结构简洁、结果可复核的时序卷积基线，后续可围绕跨被试适应性、姿态质量与在线部署进一步完善。

基金项目

本课题受到“温州大学元宇宙与人工智能研究院”的“重大课题及项目产业化专项资金”(编号：2023103)的资助。

参考文献

[1] Soro, A., Brunner, G., Tanner, S. and Wattenhofer, R. (2019) Recognition and Repetition Counting for Complex Physical Exercises with Deep Learning. *Sensors*, **19**, Article 714. https://doi.org/10.3390/s19030714

[2] Hsu, Y.C., Efstratios, T. and Tsui, K. (2023) Viewpoint-Invariant Exercise Repetition Counting. *Health Information Science and Systems*, **12**, Article No. 1. https://doi.org/10.1007/s13755-023-00258-3

[3] Abedi, A., Bisht, P., Chatterjee, R., Agrawal, R., Sharma, V., Jayagopi, D.B., *et al*. (2023) Rehabilitation Exercise Repetition Segmentation and Counting Using Skeletal Body Joints. 2023 20*th Conference on Robots and Vision* (*CRV*), Montreal, 6-8 June 2023, 288-295. https://doi.org/10.1109/crv60082.2023.00044

[4] Levy, O. and Wolf, L. (2015) Live Repetition Counting. 2015 *IEEE International Conference on Computer Vision* (*ICCV*), Santiago, 7-13 December 2015, 3020-3028. https://doi.org/10.1109/iccv.2015.346

[5] Runia, T.F.H., Snoek, C.G.M. and Smeulders, A.W.M. (2019) Repetition Estimation. *International Journal of Computer Vision*, **127**, 1361-1383. https://doi.org/10.1007/s11263-019-01194-0

[6] Zhang, H., Xu, X., Han, G. and He, S. (2020) Context-Aware and Scale-Insensitive Temporal Repetition Counting. 2020 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), 14-19 June 2020, 667-675. https://doi.org/10.1109/cvpr42600.2020.00075

[7] Dwibedi, D., Aytar, Y., Tompson, J., Sermanet, P. and Zisserman, A. (2020) Counting Out Time: Class Agnostic Video Repetition Counting in the Wild. 2020 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), 13-19 June 2020, 10384-10393. https://doi.org/10.1109/cvpr42600.2020.01040

[8] Hu, H., Dong, S., Zhao, Y., Lian, D., Li, Z. and Gao, S. (2022) TransRAC: Encoding Multi-Scale Temporal Correlation with Transformers for Repetitive Action Counting. 2022 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), New Orleans, 19-24 June 2022, 18991-19000. https://doi.org/10.1109/cvpr52688.2022.01843

[9] Newell, A., Yang, K. and Deng, J. (2016) Stacked Hourglass Networks for Human Pose Estimation. In: Leibe, B., Matas, J., Sebe, N. and Welling, M., Eds., *Computer Vision—ECCV* 2016, Springer, 483-499. https://doi.org/10.1007/978-3-319-46484-8_29

[10] Cao, Z., Hidalgo, G., Simon, T., Wei, S. and Sheikh, Y. (2021) OpenPose: Realtime Multi-Person 2D Pose Estimation Using Part Affinity Fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **43**, 172-186. https://doi.org/10.1109/tpami.2019.2929257

[11] Sun, K., Xiao, B., Liu, D. and Wang, J. (2019) Deep High-Resolution Representation Learning for Human Pose Estimation. 2019 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), Long Beach, 15-20 June 2019, 5686-5696. https://doi.org/10.1109/cvpr.2019.00584

[12] Du, Y., Wang, W. and Wang, L. (2015). Hierarchical Recurrent Neural Network for Skeleton Based Action Recognition. 2015 *IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*), Boston, 7-12 June 2015, 1110-1118. https://doi.org/10.1109/cvpr.2015.7298714

[13] Yan, S., Xiong, Y. and Lin, D. (2018) Spatial Temporal Graph Convolutional Networks for Skeleton-Based Action Recognition. *Proceedings of the AAAI Conference on Artificial Intelligence*, **32**, 7444-7452. https://doi.org/10.1609/aaai.v32i1.12328

[14] Shi, L., Zhang, Y., Cheng, J. and Lu, H. (2019) Two-Stream Adaptive Graph Convolutional Networks for Skeleton-Based Action Recognition. 2019 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), Long Beach, 15-20 June 2019, 12018-12027. https://doi.org/10.1109/cvpr.2019.01230

[15] Simonyan, K. and Zisserman, A. (2014) Two-Stream Convolutional Networks for Action Recognition in Videos. *Advances in Neural Information Processing Systems*, **27**, 568-576.

[16] Wang, L., Xiong, Y., Wang, Z., Qiao, Y., Lin, D., Tang, X., *et al*. (2016) Temporal Segment Networks: Towards Good Practices for Deep Action Recognition. In: Leibe, B., Matas, J., Sebe, N. and Welling, M., Eds., *Computer Vision—ECCV* 2016, Springer, 20-36. https://doi.org/10.1007/978-3-319-46484-8_2

[17] Carreira, J. and Zisserman, A. (2017) Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset. 2017 *IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR)*, Honolulu, 21-26 June 2017, 4724-4733. https://doi.org/10.1109/cvpr.2017.502

[18] Lea, C., Flynn, M.D., Vidal, R., Reiter, A. and Hager, G.D. (2017). Temporal Convolutional Networks for Action Segmentation and Detection. 2017 *IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR)*, Honolulu, 21-26 June 2017, 1003-1012. https://doi.org/10.1109/cvpr.2017.113

[19] Farha, Y.A. and Gall, J. (2019) MS-TCN: Multi-Stage Temporal Convolutional Network for Action Segmentation. 2019 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), Long Beach, 15-20 June 2019, 3575-3584. https://doi.org/10.1109/cvpr.2019.00369

[20] Jocher, G., Chaurasia, A. and Qiu, J. (2023) Ultralytics YOLOv8, Version 8.0.0. https://github.com/ultralytics/ultralytics

[21] Wu, Y. and He, K. (2018) Group Normalization. In: Ferrari, V., Hebert, M., Sminchisescu, C. and Weiss, Y., Eds., *Computer Vision—ECCV* 2018, Springer, 3-19. https://doi.org/10.1007/978-3-030-01261-8_1

[22] Lee, M. (2023) Mathematical Analysis and Performance Evaluation of the GELU Activation Function in Deep Learning. *Journal of Mathematics*, **2023**, Article ID: 4229924. https://doi.org/10.1155/2023/4229924

[23] Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I. and Salakhutdinov, R. (2014) Dropout: A Simple Way to Prevent Neural Networks from Overfitting. *Journal of Machine Learning Research*, **15**, 1929-1958.

[24] He, K., Zhang, X., Ren, S. and Sun, J. (2016) Deep Residual Learning for Image Recognition. 2016 *IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*), Las Vegas, 27-30 June 2016, 770-778. https://doi.org/10.1109/cvpr.2016.90

[25] Loshchilov, I. and Hutter, F. (2019) Decoupled Weight Decay Regularization. arXiv: 1711.05101.