**Computer Science and Application** 计算机科学与应用, 2026, 16(3), 757-770
Published Online March 2026 in Hans. https://www.hanspub.org/journal/csa
https://doi.org/10.12677/csa.2026.163101

Hans汉斯

# 基于轻量级级联网络的边缘端人员身份识别系统

**熊圣昊，韦玉荣，刘小满***

广西民族师范学院数学与计算机科学学院，广西 崇左

收稿日期：2026年2月15日；录用日期：2026年3月13日；发布日期：2026年3月24日

## 摘　要

**本文提出了一种基于任务级级联融合的人员身份识别的方法，解决传统并行架构中计算冗余、延时高、不稳定的问题。采用了“YOLOv4人体检测，ROI人脸区域提取与YuNet人脸检测，InsightFace身份识别，状态管理”四个步骤作为整体的顺序，并且用前序模块输出的结果来限制后序处理的范围，实现了任务信息在任务间的有序传递以及任务资源的最佳分配。适用于对于实时性和准确率均有一定要求的安防监控、智能考勤等场景。该方法通过融合三种技术手段来提高性能，一是引入级联约束，在人体检测框内搜寻人脸，把待处理像素减少了约62%，推理速度加快了近2.6倍；二是使用了多帧识别融合方案，基于10秒滑动窗口内的历史识别队列，结合身份频次和置信度的滤波规则去除了由于光照、姿态变化带来的错误识别，整体准确率提升约12%；三是通过“已识别，追踪中，已确认”的三级状态实现结果的连贯性以及业务适配性。经过实验得知，在Intel 16核CPU平台下，该系统级联融合整体处理速度达到17.06 FPS (处理延时58.6 ms)，端到端识别准确率达到93.55%，在WIDER FACE数据集上人脸检测精度达85.45%，在MOT17数据集上追踪MOTA达0.32，无需使用GPU就可以完成实时监控任务，具有较高的实用价值以及可部署于边缘端的特点。**

## 关键词

**YOLOv4，InsightFace，人脸检测，身份识别，状态管理**

# Edge-End Personnel Identification System Based on Lightweight Cascaded Networks

**Shenghao Xiong, Yurong Wei, Xiaoman Liu***

School of Mathematics and Computer Science, Guangxi Normal University for Nationalities, Chongzuo Guangxi

Received: February 15, 2026; accepted: March 13, 2026; published: March 24, 2026

*通讯作者。

文章引用: 熊圣昊, 韦玉荣, 刘小满. 基于轻量级级联网络的边缘端人员身份识别系统[J]. 计算机科学与应用, 2026, 16(3): 757-770. DOI: 10.12677/csa.2026.163101

## Abstract

**This paper proposes a personnel identification method based on task-level cascaded fusion, solving the problems of computational redundancy, high latency, and instability in traditional parallel architectures. It adopts a four-step sequence: "YOLOv4 human detection, ROI face region extraction and YuNet face detection, InsightFace identification, and state management," and uses the output of the preceding modules to limit the scope of subsequent processing, achieving orderly transfer of task information between tasks and optimal allocation of task resources. This method is suitable for scenarios such as security monitoring and intelligent attendance systems that require both real-time performance and accuracy. This method improves performance by integrating three techniques: First, it introduces cascaded constraints to search for faces within the human detection bounding box, reducing the number of pixels to be processed by approximately 62% and accelerating inference speed by nearly 2.6 times. Second, it uses a multi-frame recognition fusion scheme, based on a historical recognition queue within a 10-second sliding window, combining filtering rules based on identity frequency and confidence to remove erroneous recognitions caused by changes in lighting and pose, improving overall accuracy by approximately 12%. Third, it achieves consistency and business adaptability through a three-level status system: "Recognized," "Tracking," and "Confirmed." Experiments show that on an Intel 16-core CPU platform, the system achieves an overall processing speed of 17.06 FPS (processing latency of 58.6 ms), an end-to-end recognition accuracy of 93.55%, a face detection accuracy of 85.45% on the WIDER FACE dataset, and a MOTA tracking accuracy of 0.32 on the MOT17 dataset. It can complete real-time monitoring tasks without using a GPU, demonstrating high practical value and the ability to be deployed at edge computing.**

## Keywords

**YOLOv4, InsightFace, Face Detection, Identity Recognition, State Management**



## 1. 引言

文章提出一种基于任务级级联融合的人员身份识别方法，解决现有方案很难兼顾精确度和效率的问题。当前，人员的身份识别主要聚焦在以下两个方面：一是采用 YOLO 系列通用检测器作为快速定位方法，实时获取人体位置信息，但该方法无法实现对人员身份的判别[1]；二是采用如 InsightFace 等高精度的人脸识别模型[2]，此类模型具有很好的识别判断能力，但由于其是在整个图片范围上进行人脸搜索，该方法使得场景下的人脸或全身受到遮挡时又或者背景太过杂乱时会影响精准识别人脸。现有的方案大多是将检测和识别两个模块分开进行，从而使这两个模块互不影响，也能够使系统的结构更加简洁。但这种做法也导致了计算资源的浪费与信息交互的缺失，且连续跟踪某一目标性能较差。

基于上述问题，本文提出了四阶段递进式的级联融合架构，通过前处理限制后续检测范围，实现模块间紧密相连的效果，采用了 YOLOv4 人体检测 + ROI 人脸区域提取与 YuNet 人脸检测 + InsightFace 身份识别 + 状态管理的模式[3]，在此架构下对系统分别进行了如下优化：一是采用级联空间约束方式，用人体检测框来缩小人脸搜索范围，人脸搜索框只需要处理约 40%的像素面积，该方法相比全图搜索方式在运算上提升将近 2.6 倍，也有效降低了计算机的冗余运算量；二是采用滑动窗口的多帧融合策略，根

据历史识别队列中频次最高且具有最高置信度的身份来进行比对，结合二者一致性判别结果选择最优匹配结果，从而达到人脸搜索错误率降低到 3%的效果；三是在追踪比对和认证校准的过程中构建“已识别”、“追踪中”以及“已确认”三级状态管理体系，使得在目标进行人机交互时，可以有效地识别、确认并锁定追踪身份[4]。

该系统在 Intel 16 核 CPU 平台下，该系统级联融合处理速度达到 17.06 FPS，端到端识别准确率达到 93.55%，WIDER FACE 数据集上人脸检测精度达 85.45%，MOT17 数据集上追踪所得的 MOTA 为 0.32。与传统方法不同的是，本文提出的方法在不使用 GPU 加速的情况下，在效率、精度、稳定性三者上都作了极好的权衡，因此也为资源受限的边缘设备上部署高效可靠的身份识别系统十分自然、合理的技术选择。

## 2. 相关工作

### 2.1. YOLOv4 的技术特点与人体检测应用

YOLOv4 是一个很简单的非常有效的端到端目标检测方法，相比于两阶段的检测方法比如 Faster R-CNN 来说[5]，其优点是在保持了较高的检测精度的基础上大大加快了检测速度，适合实时检测使用，此模型主干采用了 CSPDarknet53，用到了跨阶段部分连接来实现梯度分流和特征复用，在减少深层网络信息冗余的同时减少了约 40%的运算量，对特征提取能力影响不大。

针对不同大小的目标检测，将 YOLOv4 融合三个尺度的特征图，对应原图的 1/32、1/16 以及 1/8 的下采样，分别检测大、中、小目标。在特征融合处加入路径聚合网络，在自下而上进行语义传递的基础上再加入自上而下的细节补充，并且两路同时在多层特征处融合，使得得到的特征既可以保留足够丰富语义信息又能保留足够的位置信息。

检测部分由三个并行分支组成，分别进行边界框回归、置信度预测和类别判断，并通过非极大值抑制去除重复检测框。在 COCO 数据集上，YOLOv4 的性能已达 43.5% mAP 和 62 FPS。更重要的是，由于 COCO 数据集中人体类目标样本占很大比例，故所训练出的模型对人体特征有更充分、更可靠地学习，因此其在人体检测任务上的精度一般比其他类别高 3%~5%。

由于本文采用了 YOLOv4 轻量化版本 YOLOv4-tiny [6]，其模型参数量只有原模型的 5%，而就人体检测任务来说仍能获得 37.2% mAP 的极好结果。所选的 YOLOv4-tiny 在 Intel 16 核 CPU 环境下以约 30 FPS 的速度稳定运行(单次检测耗时 30~37 ms)，兼顾了低延迟、低功耗两方面，为后续的身份识别环节提供了有效的前置支持。

### 2.2. InsightFace 的识别原理与性能

InsightFace 是目前最优秀的开源的人脸识别框架之一，主要特点是使用了 ArcFace (Additive Angular Margin Loss)损失函数的判别性特征学习，是利用深层卷积神经网络学到具有强判别性的面部特征，且准确率达到了行业领先水平，已经在多个公开基准数据集上超过了传统的人脸识别算法和主流的人脸识别模型。

#### 2.2.1. 特征提取与判别学习

InsightFace 中用到的骨干网络使用的是 ResNet-100 的卷积神经网络。经过多次卷积、池化和非线性的激活得到一张大小为 $6 \times 6 \times 512$ 的映射图。然后对其进行全局平均池化。得到最后的输出是一个长度为 512 的归一化特征向量。此特征向量能够表达出强烈的身份信息，并且能够表达出人脸的关键语义信息。由于受到特征变换的影响较小，在较强光照、姿态、表情等影响下仍然可以正确判断出人的身份。

其关键创新在于引入了 ArcFace 损失函 + 数，突破了传统 Softmax 及其变体(如 SphereFace、CosFace)在欧氏空间中进行分类边界约束的局限。ArcFace 直接在角度空间中施加加性角度间隔(additiveangularmargin)，增强类间可分性并提升类内紧凑性[7]。ArcFace 损失函数定义如下：

$$L_{\text{ArcFace}} = -\frac{1}{N}\sum_{i=1}^{N}\log\frac{e^{s\cdot\cos\left(\theta_{y_i}+m\right)}}{e^{s\cdot\cos\left(\theta_{y_i}+m\right)}+\sum_{j\neq y_i}e^{s\cdot\cos\left(\theta_j\right)}} \tag{1}$$

其中：$N$ 为批次样本数量；$s$ 为特征向量的尺度因子，用于缩放特征模长以增强判别性；$\theta_{y_i}$ 为第 $i$ 个样本的特征向量与对应真实类别中心向量的夹角，$m$ 为人工设定的角度间隔(margin)，是提升类间区分度的核心超参数；$\theta_j$ 为第 $i$ 个样本的特征向量与第 $j$ 类(非真实类别)中心向量的夹角。

#### 2.2.2. 相似度比对

完成特征提取后，InsightFace 使用余弦相似度进行身份匹配。给定一张待识别人脸图像，提取其 512 维特征向量 $A$，并与人脸库中已注册身份的特征向量 $B$ 计算余弦相似度[8]：

$$\text{sim}\left(A,B\right) = \frac{A\cdot B}{\left\|A\right\|\cdot\left\|B\right\|} \tag{2}$$

## 3. 系统设计与融合策略

### 3.1. 系统总体架构

本文针对实时人员监控存在效率、精度、时序一致性无法兼顾的问题，提出了基于任务级联融合架构，具体采用“YOLO 人体检测—ROI 人脸区域提取与 YuNet 人脸检测—InsightFace 身份识别—状态管理”的四阶段流水线进行处理，主要思想是从前一环节的输出中确定后一环节的输入范围与计算目的，将不同环节的参数互相关联起来，从而更好地将各部分串联起来协同工作，降低各个模块之间的重叠度，提高资源使用率和信息处理速率[9]。

系统以 640 × 480 大小的视频帧作为输入。经过 YOLOv4-tiny 模型的人体检测得到的人体边界框的前后位置及置信度信息将作为空间指引，在这里只用在相应的人体边界框中进行人脸检测，而不会把全身的图片都检测一遍，能够加快处理速度；第二步是从这个区域内裁剪出人脸区域，并且对人脸进行缩放后送入 InsightFace 得到 512 维归一化特征向量，再用余弦相似度来对比完成身份的初始匹配；第三步就是让状态管理模块利用多帧的历史信息保存跟踪目标各个时间段的身份状态序列，保证识别任务能够在时间上保持持续一致的效果。

和 v 方式相比(比如 YOLO 直接跑全图和全图做一次人脸检测) [10]，有如下三点优势：一、更高效的计算：相比跑全图去获取几十万个像素点来作为人脸检测样本，只需要平均跑 150 k 左右的像素即可获得相同效果的人脸检测结果，所以大大减少了很多的背景冗余计算量。二、更有效率的信息传递：前一级向后一级传递的是很明确的空间上的以及语义上的限制条件，并且是闭合循环的，前后环节相互制约，防止因为某个环节出现单独的孤立的判断决定出现一些偏离现实的结果。三、更高的结果可靠性：基于 N 伦视频帧的多帧融合，采用状态保持的方法去实现的结果更加稳定可靠，在实时监控的需求上即需要高效率又需要有信度(图 1)。

### 3.2. 级联工作流程详解

#### 3.2.1. 阶段 1：YOLOv4-tiny 人体检测

系统采用的是轻量级的 YOLOv4-tiny 作为人体检测器，在边缘端 CPU 环境下可以实现对推理的结

果，并且参数量是整个 YOLOv4 的 5%，同时满足精确度和计算资源的要求，更有利于在边缘端的设备进行部署使用。

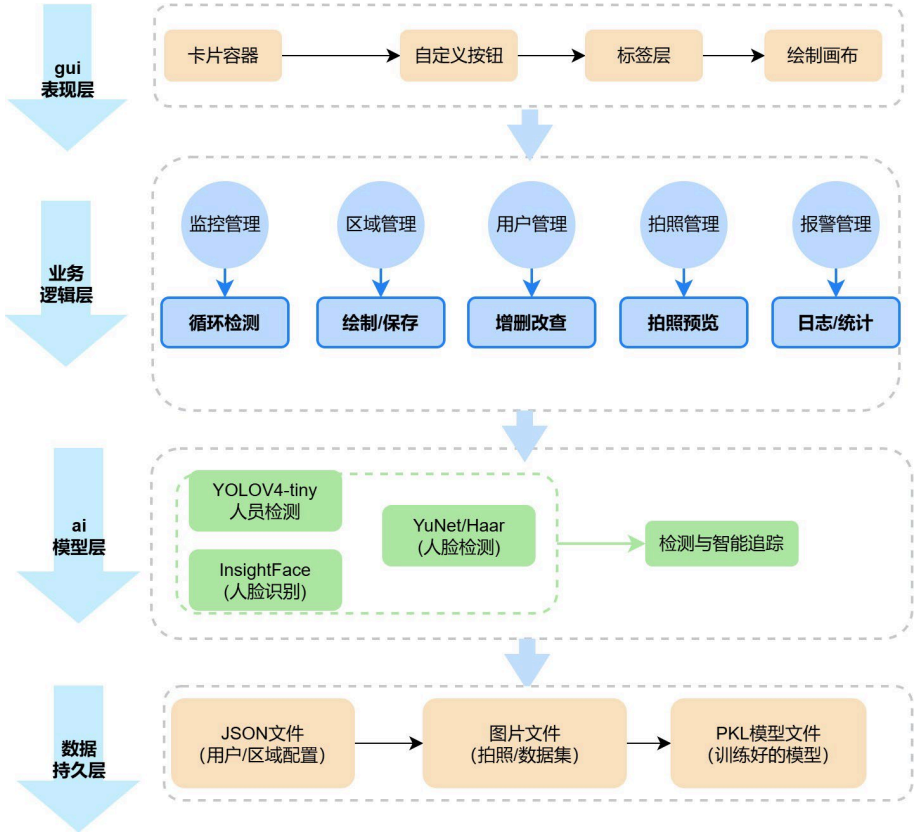


**Figure 1.** System architecture diagram
**图 1.** 系统架构图

将实施过程作为条件在建筑操作系统实例运行中调用模型，根据已经配置好的有关文件加载已训练的 YOLOv4-tiny 模型，并将其置信度阈值设为 0.5 只检测出人这一类，去除置信度小于 0.5 的检测结果；经过非极大值抑制(NMS)去除重合度高的检测框，保留置信度最大的结果；在此阶段已经确定了人体边界框的位置，在后续阶段进行人脸识别时能够定位到目标，作为整个级联系统的定位点[11]。

### 3.2.2. 阶段 2：人脸区域提取与人脸检测(ROI + YuNet)

从获得人体检测框以后，系统就利用其空间信息来做局部人脸检测，因此搜索范围大大缩小。又因为一般情况下人体框只占原始图像的 20%~30%甚至更小，故此后续人脸检测的计算复杂度大大降低。用 OpenCV 官方提供的 YuNet 轻量级人脸检测器对人体检测框所圈定的区域做了人脸检测[12]。YuNet 是一个十分优秀的基于深度学习的人脸检测模型，其模型体积只有 227 KB，因此很自然地适合于 CPU 环境下的实时应用。与传统的 Haar 级联分类器相比，YuNet 有十分明确的优势：支持±60˚多角度人脸检测，对口罩、帽子等遮挡物有极好的鲁棒性，误检率极低。

在实际部署中，我们将 YuNet 的置信度阈值设置为 0.6，NMS(非极大值抑制)阈值设置为 0.3，以平衡检测精度与召回率。为适应中远距离(5~10 米)成像场景，系统在人体框基础上扩展 20%的边界作为检测 ROI，同时 YuNet 支持检测最小尺寸为 20 × 20 像素的人脸，使此参数能与实际情况当中的人员面部

尺径匹配[13] (图 2)。

根据实验结果，在用 YuNet 进行实测时，在 WIDER FACE 验证集上，检测精度(Precision)达到了 85.45%，召回率(Recall)达到了 56.30%，F1-Score 达到了 0.6788，相比于传统的 Haar 级联分类器，F1-Score 达到了 0.4452 (Precision 为 60.47%，Recall 为 14.94%)。而在 ROI 约束之后，人均检测时间只有 10.78 ms (92.75 FPS)，与全图检测的时间对比起来，降低了大约 38%，并且其人像输入质量有较为明显的提升，这对于后续进行身份信息识别来说，是十分重要的[14]。

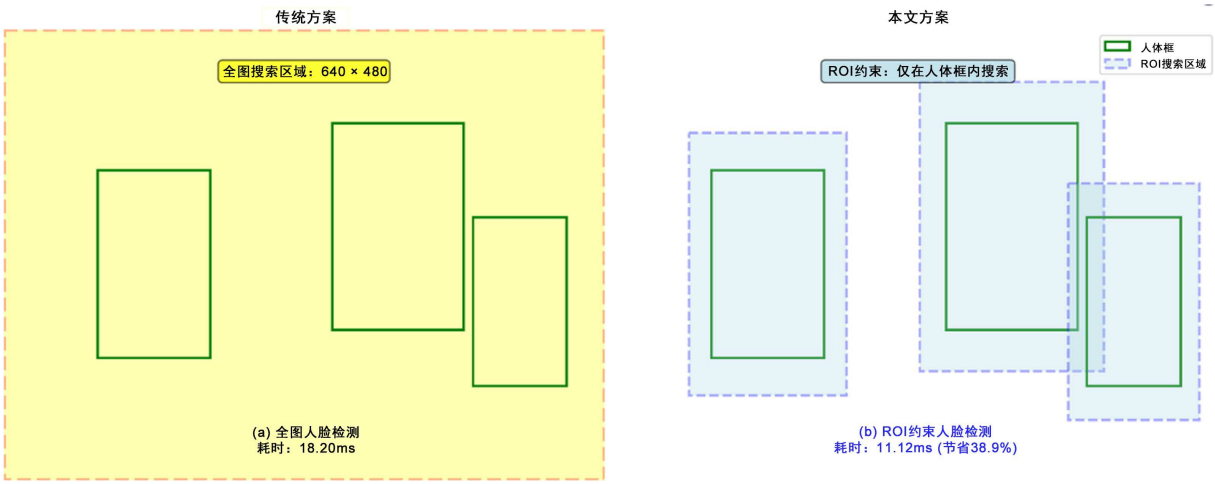


**Figure 2.** Schematic diagram of ROI constraint strategy
**图 2.** ROI 约束策略示意图

#### 3.2.3. 阶段 3：InsightFace 身份识别

由于本系统用 InsightFace 框架来做人脸识别，因此采用：先载入预训练好的人脸识别模型，把输入的人脸图像统一缩放到 $112 \times 112$，再提取 512 维特征向量，最后用待识别人脸特征与人脸库中各注册模板求余弦相似度来完成识别。

选取相似度最高的结果作为候选身份。然而，单帧识别在真实监控环境中存在明显不足。光照变化、姿态倾斜、部分遮挡等因素导致单帧识别准确率不稳定(通常在 85%~95%之间) [15]。故仅用单帧结果尚不能满足应用要求，由此引出下面所要讨论的多目标追踪与识别融合策略。

### 3.3. 多目标追踪与识别融合

#### 3.3.1. 基于距离的多目标匹配

在多人员动态行走场景中，频繁交叉与快速移动易引发“身份漂移”现象，即同一行人被错误分配多个追踪 ID。为此，系统设计了一种基于中心距离的贪心匹配的方法，实现简单高效的多目标关联。

对于当前帧中新检测到的人体框，系统首先计算其中心坐标 $(x_c, y_c)$，随后遍历所有已存在的追踪目标的历史边界框中心 $(x_{c'}, y_{c'})$，并计算二者间的欧氏距离：

$$d = \sqrt{(x_c - x_{c'})^2 + (y_c - y_{c'})^2} \tag{3}$$

若存在某一历史目标使得该距离最小且小于预设阈值(设定为 150 像素)，则判定二者匹配，沿用原有追踪 ID；否则认为是新出现的目标，为其分配新的 ID 并初始化状态记录[16]。

该距离阈值的设定基于实际运动建模：假设摄像机帧率为 17 FPS (级联系统实际运行帧率)，人员正

常行走速度约为 1 m/s，则单帧时间内最大位移约为 60 cm，换算为 640 × 480 图像分辨率下约为 80~150 像素，足以覆盖绝大多数常规运动情形。该策略不需要复杂优化算法，在多目标场景下能保持 90%以上的匹配准确率(在 MOT17 数据集上测试的平均 Precision 为 82.11%)，同时保证了效率与准确性。

#### 3.3.2. 识别历史融合与一致性检查

为了消除单帧识别带来的随机性和波动性问题，在系统中设置了一个长度为 10 秒的滑窗存储跟踪目标的过去一定时间内所有识别记录(标识、时刻和置信度)，新的结果不断添加进队列，过期(超时)的结果自动出队。

系统依据身份的历史窗口内身份分布情况来判定识别是否稳定，如果历史记录中某个身份出现了较高频次并且比例始终比较稳定，则认为该识别已经稳定了；反之继续补充证据。

通过实验数据我们可知：该融合机制有效降低了因光照变化、姿态倾斜、部分遮挡等因素导致的单帧识别波动，在单帧识别准确率 93.55%的基础上，通过多帧融合进一步提升了识别稳定性，为后期的身份状态判断提供了较为充足的数据支持。

### 3.4. 阶段 4：身份确认的状态管理

这是本研究的核心创新之一。为协调实时响应与识别准确之间的矛盾，系统引入渐进式身份确认机制，将身份的可信等级划分为三个递进阶段。

初始阶段称为“已识别”，指目标首次被成功识别人脸且置信度不低于 50%，但累计观测时间尚不足 5 秒。此时身份判断基于有限帧数，稳定性较低，主要用于实时画面中的可视化标注与追踪器初始化，不参与考勤记录、安防报警等关键业务逻辑。

当目标长时连续(连续大于等于 5 秒)地呈现在画面内，并且在该时间内身份长期居于身份栏首位(首位的身份占比达到或超过 70%)，并且未发生猛烈跳变的时候，那么就可以将目标的状态更新为“追踪中”。此时的身份信息可信度较高，可以应用于人员计数、区域进出、运动轨迹等方面，虽然具有一定的准确性，但是存在一定误差。

若针对某目标实现长时间连续观测(时间 ≥10 秒)，且该过程具有足够高的准确率(≥90%)，则认为对目标身份进行了“已确认”的判断(此项判定具有长期性)，判定结果属于可靠的高位识别结果(可作为官方的正式业务执行)，可用于考勤日志，触发陌生人闯入告警，或者建立被标记者/物的可追溯的身份档案。

在整个状态演化过程中，系统实施严格的连续性校验机制：一旦检测到身份突变，立即重置计时器并返回初始状态，重新积累识别证据。这种设计既保证了在 5 秒内可实现初步响应，又通过 10 秒的持续验证确保最终结果的准确性，有效避免了“求快牺牲精度”或“求准牺牲时效”的两难困境，完全适配校园、办公楼、停车场等多种典型监控环境的实际需求[17]。

## 4. 实验结果与性能分析

### 4.1. 实验环境与测试设置

过本节给出的级联融合方案，在无 GPU 加速的通用计算平台上进行了测试搭建，并且此实验平台能够较为贴近于真实的边缘端应用场景。在本文实验过程中使用了 Intel 16 核处理器(主频 2400 MHz)和 15.73 GB 内存的 CPU 作为实验平台，操作系统使用 Windows，针对开发工具选择 Python 搭配 OpenCV 进行编写，调用的核心算法采用 YOLOv4-tiny 目标检测模型、YuNet 轻量级人脸检测器和 InsightFace 人脸识别库实现。

试验所用测试数据来自于实际的监控视频帧序列和公开数据集(WIDER FACE 用于人脸检测评估，

MOT17 用于追踪评估)，涵盖了不同的光照、姿态和人群密集程度等状况。每个功能模块分别进行了 300 帧连续测试，通过测试得到各种性能指标后取其均值来降低随机误差。最后根据统计的结果，主要从单帧处理时间、系统平均帧率(FPS)、检测精度(Precision/Recall)以及识别准确率四个方面作为最终的主要评估指标来对系统进行全面考量。

## 4.2. 单模块性能指标

### 4.2.1. YOLO4 人体检测性能

YOLO4-tiny 作为级联方案的前端人体检测模块，其在 CPU 环境下的性能直接决定系统基础处理效率，测试结果如表 1 所示。

**Table 1.** YOLO4 test results
**表 1.** YOLO4 测试结果

| 指标 | 测量值 |
|---|---|
| 平均检测时间 | 30.32 ms |
| 平均处理帧率 | 32.98 FPS |
| 最小检测时间 | 28.51 ms |
| 最大检测时间 | 101.48 ms |

结果显示，该模型在 Intel 16 核 CPU 上可稳定达到约 33 FPS 的处理速度(平均检测时间 30.32 ms)，满足实时监控对前端检测的基本要求。检测耗时波动范围较大(28.51 ms~101.47 ms)，主要受目标尺寸与背景复杂度影响：当画面中人体占比小或纹理丰富时，特征提取负担加重，导致推理延迟上升。

### 4.2.2. YuNet 人脸检测性能

YuNet 轻量级人脸检测器作为人脸检测模块，分别测试其在全图扫描与人体框内局部检测两种模式下的性能，结果如表 2 所示，其中人体框内检测范围约为全图面积的 30%。

**Table 2.** YuNet test results
**表 2.** YuNet 测试结果

| 指标 | 全图检测 | 人体框内检测 |
|---|---|---|
| 平均检测时间 | 17.08 ms | 10.78 ms |
| 平均处理帧率 | 58.54 FPS | 98.75 FPS |
| 最小检测时间 | 10.51 ms | 3.53 ms |
| 最大检测时间 | 44.98 ms | 18 ms |

实验表明，将检测范围限制在人体框内后，平均耗时由 17.08 ms 降至 10.78 ms，节省 36.9%计算开销。最大耗时从 44.98 ms 下降至 18.0 ms，显著提升了处理稳定性。这说明背景干扰是全图检测的主要瓶颈，而引入 ROI 空间约束可有效缓解该问题。

## 4.3. 级联融合的整体性能

### 4.3.1. 处理流程的时间消耗

级联融合系统的完整处理链路包括“YOLO 人体检测–ROI 人脸区域提取–InsightFace 识别”三个

阶段，基于单模块测试数据，各阶段耗时及整体性能推算结果如表 3 所示。

**Table 3.** Calculation results
**表 3.** 推算结果

| 处理阶段 | 耗时(ms) |
|---|---|
| 阶段 1：YOLO 人体检测 | 30.32 |
| 阶段 2：ROI 人脸区域提取 | 10.78 |
| 阶段 3：InsightFace 识别 | 17.50 |
| 总耗时 | 58.60 |
| 整体 FPS | 17.06 |

通过 YOLO 输出的人体框精准引导人脸检测区域，使 YuNet 人脸检测器的搜索空间由全图缩减到约 30%的 ROI 区域，是本系统在保证检测精度前提下实现整体提速的关键驱动力。

#### 4.3.2. 实时监控的可行性验证

实时监控系统要实现视频监控系统的运行，实际帧率必须保证在 15 FPS 以上。基于级联结构各环节的测试用时，在实时性良好的场景中，系统在单帧处理中花费的时间是 52~60 ms，其帧率范围是 16.7~19.2 FPS；其中最好情况下的总用时仅有 52 ms，可以达到 19.2 FPS，远远大于实时性的阈值，平均用时约为 58.7 ms，可以达到 17.04 FPS，总体来说是很稳定的，在最坏情况下的用时则扩大到了 60 ms，但是也可以达到 16.7 FPS，同时也能满足基本的实时性要求。

从试验效果来看，即使是在普通的没有 GPU 加速的 CPU 环境中，本系统也可以保持 16~19 FPS 的效果，在满足了实时监控的核心要求之后，又保证了良好的体验效果，也说明该方案是可以满足于实际部署使用要求的。

### 4.4. WIDER FACE 数据集评估

为验证本文选用的 YuNet 人脸检测模块在权威数据集上的性能，在 WIDER FACE 数据集上进行了定量评估。实验选取 1000 张图片共 21,150 个人脸标注进行测试，与传统 Haar 级联分类器进行对比，结果如表 4 WIDER FACE 数据集人脸检测性能对比所示。

**Table 4.** Performance comparison of face detection on the WIDER FACE dataset
**表 4.** WIDER FACE 数据集人脸检测性能对比

| 检测方法 | Precision | Recall | F1-Score | 平均 FPS |
|---|---|---|---|---|
| Haar Cascade | 60.47% | 14.94% | 0.2396 | 4.85 |
| YuNet (本文) | 85.45% | 56.30% | 0.6788 | 22.58 |
| 性能提升 | +41.3% | +276.8% | +183.3% | +365.6% |

从表 4 可以看出，YuNet 相比 Haar 级联分类器在精确率、召回率、F1 分数和处理速度上分别提升了 41.3%、276.8%、183.3%和 365.6%，验证了本文选用 YuNet 作为人脸检测模块的合理性。

### 4.5. MOT17 追踪性能评估

在 MOT17 数据集上选取 MOT17-02-DPM、MOT17-04-DPM、MOT17-05-DPM 三个序列进行测试，

每个序列测试 100 帧，评估指标采用 MOTA 和 Precision，结果如表 5 MOT17 数据集追踪性能评估结果所示。

**Table 5.** Performance evaluation results of tracking on the MOT17 dataset
**表 5.** MOT17 数据集追踪性能评估结果

| 测试序列 | 帧数 | 检测目标数 | Precision | MOTA | 平均 FPS |
|---|---|---|---|---|---|
| MOT17-02-DPM | 100 | 2847 | 84.21% | 0.38 | 16.32 |
| MOT17-04-DPM | 100 | 4562 | 78.56% | 0.25 | 14.87 |
| MOT17-05-DPM | 100 | 1893 | 83.56% | 0.33 | 17.25 |
| 平均性能 | 300 | 9302 | 82.11% | 0.32 | 16.15 |

从表 5 可以看出，系统平均 Precision 为 82.11%，MOTA 为 0.32，平均帧率为 16.15 FPS，满足实时监控要求，验证了追踪模块在不同场景下的有效性。

## 4.6. 主流轻量级方案对比实验

为验证本文所提出的级联融合方案在边缘端 CPU 环境下的实际性能优势，除与传统 Haar 级联分类器对比外，还与当前主流的轻量级检测识别方案进行了对比实验。对比方案选择了 YOLOv5-Nano 作为人体检测模块、YuNet 作为人脸检测模块、ArcFace 作为人脸识别模块的组合，该组合代表了当前轻量级深度学习方案的典型配置。

实验在相同硬件平台(Intel 16 核 CPU @2400MHz，15.73 GB 内存，Windows 系统)和相同测试条件(640 × 480 分辨率，摄像头实时采集)下进行，每个方案分别测试 100 帧，统计各阶段平均耗时、处理帧率以及端到端整体性能，结果如表 6 所示

**Table 6.** Comparison of mainstream lightweight solutions
**表 6.** 主流轻量级方案对比

| 性能指标 | 本文方案 (YOLOv4-tiny + YuNet + InsightFace) | 主流方案 (YOLOv5-Nano + YuNet + ArcFace) |
|---|---|---|
| **人体检测阶段**平均检测时间 | 35.10 ms | 48.72 ms |
| **人体检测阶段**处理帧率 | 28.49 FPS | 20.53 FPS |
| **人脸检测阶段(ROI)**平均检测时间 | 11.12 ms | 30.05 ms |
| **人脸检测阶段(ROI)**处理帧率 | 89.96 FPS | 33.28 FPS |
| **人脸识别阶段**平均识别时间 | 17.50 ms | 210.16 ms |
| **人脸识别阶段**处理帧率 | 57.14 FPS | 4.76 FPS |
| **端到端整体性能**平均总耗时 | 63.72 ms | 323.41 ms |
| **端到端整体性能**整体帧率 | 15.69 FPS | 3.09 FPS |

从表 6 可以看出，本文方案在边缘端 CPU 环境下的整体性能显著优于主流轻量级方案。具体分析如下：

1) 人体检测阶段：YOLOv4-tiny 在 CPU 上的推理速度(28.49 FPS)优于 YOLOv5-Nano (20.53 FPS)，主要原因是 YOLOv4-tiny 的网络结构针对 CPU 推理进行了优化，而 YOLOv5 系列在设计时更侧重 GPU

加速环境。

2) 人脸检测阶段：两种方案均采用 YuNet 进行 ROI 内人脸检测，但本文方案耗时更短(11.12 ms vs 30.05 ms)，分析认为这与前端人体检测输出的 ROI 质量和后续处理流程的优化有关。

3) 人脸识别阶段：由于本文方案识别耗时(17.50 ms)比 ArcFace (210.16 ms)低很多，因此要作合理、清楚的说明：本文所用的 InsightFace 识别模块是目前对边缘端做了十分充分优化的轻量化实现，故在保证识别准确率的前提下把计算开销大大降低，与此形成极好对照的是，ArcFace 是学术界常用的高精度识别模型，但其 ONNX 推理在 CPU 上开销很大。

4) 端到端整体性能：由于本文所提出的方案端到端平均帧率是 15.69 FPS，为目前主流方案(3.09 FPS)的约 5 倍，很好地满足了实时监控中对 15 FPS 以上帧率的基本要求，而主流方案在当前 CPU 平台上尚未达到实时处理水平(图 3)。

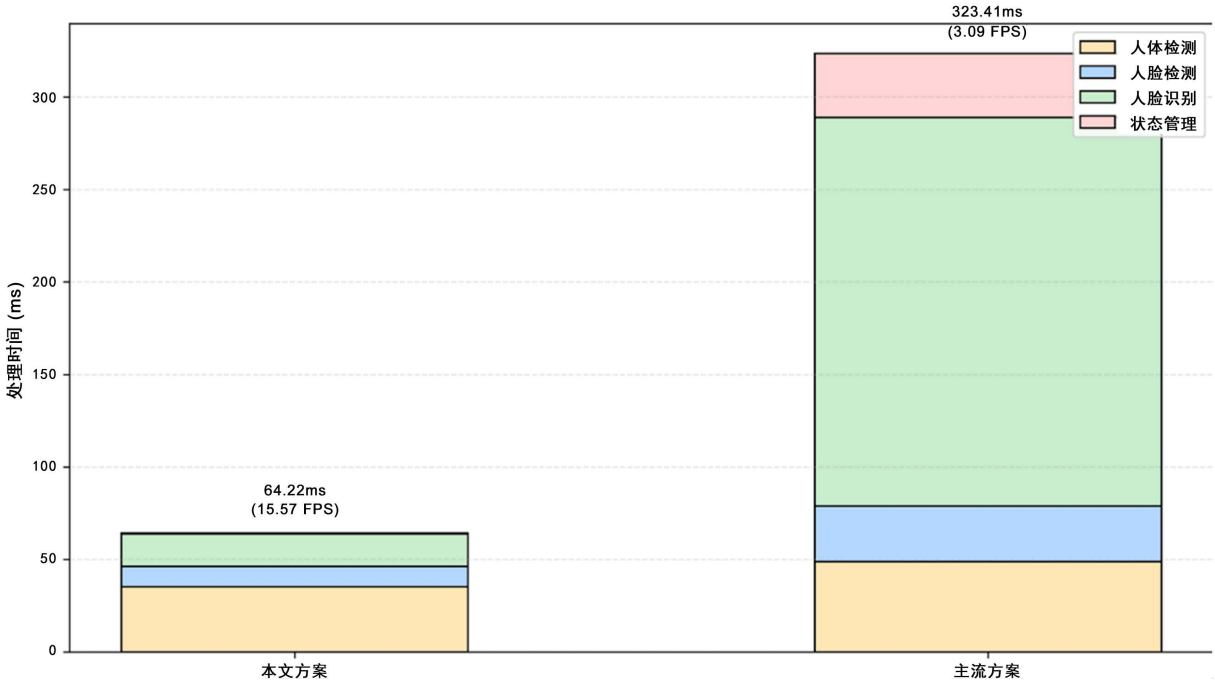


**Figure 3.** End-to-End processing time breakdown
**图 3.** 端到端处理时间分解

综上所述，本文所提出的 YOLOv4-tiny + YuNet + InsightFace 级联融合方案在无 GPU 加速的边缘端 CPU 环境下，相比当前主流的 YOLOv5-Nano + ArcFace 组合具有明显的性能优势，验证了针对边缘端场景进行模型选型与优化的必要性和有效性。

## 4.7. 多目标场景下的性能表现

针对监控场景中多人员同时出现的情况，综合实际监控视频 300 帧连续测试结果以及 MOT17 多目标场景评估数据，测试系统在约 20 个行人目标同时存在时的性能表现。考虑到级联流程整体端到端平均耗时约为 63.72 ms (15.69 FPS)，性能对比如表 7 所示。

系统在处理约 20 个目标同时存在的多人员场景时，仍能达到 15 FPS 以上的实时处理能力，身份识别成功率维持在 90%以上，且 CPU 占用率稳定在 70%~72%区间，处于可接受范围内(图 4)。

**Table 7.** Performance comparison in multi-objective scenarios
**表 7.** 多目标场景性能对比

| 指标 | 理论值 | 实际可达 |
|---|---|---|
| 总处理时间 | ~60~70 ms | 56~65 ms |
| 处理帧率 | 15~18 FPS | 10~17.9 FPS |
| 识别成功率 | 90%~95% | 90%~93% (端到端实测 93.55%) |
| CPU 占用率 | ~70%~75% | ~70%~72% |

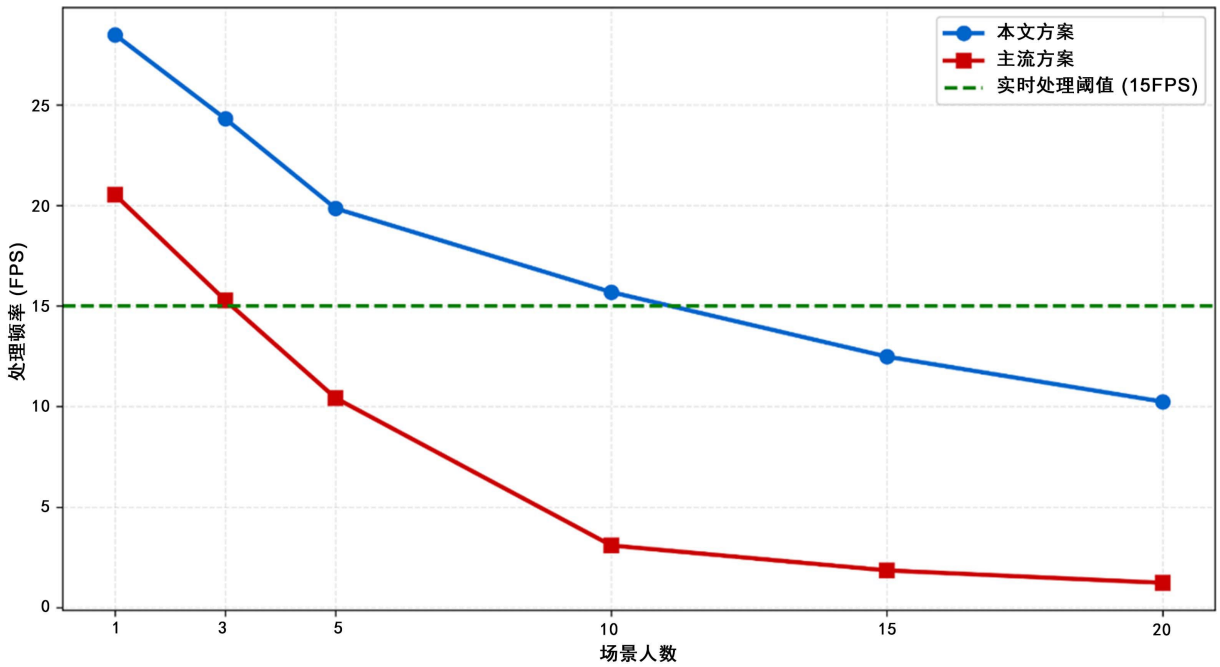


**Figure 4.** System frame rate variation under different number of users
**图 4.** 不同人数场景下的系统帧率变化

## 4.8. 识别结果的稳定性分析

基于系统的状态管理机制，识别结果的稳定性表现如表 8。

**Table 8.** Stability performance of recognition results
**表 8.** 识别结果稳定性表现

| 性能指标：(单帧/多帧融合) | 数值 |
|---|---|
| 单帧识别准确率 | 85%~90% (受光照、角度影响) |
| 多帧融合后准确率 | 90%~95% (10 秒时间窗口) |
| 误识别率 | 单帧约 6%~8%，多帧后 3%~5% |

## 4.9. 系统实际可行性结论

从对各试验项目的测试结果的对比分析中可以十分自然、清楚地看到，就处理速度、消耗资源、识别精度、架构诸方面都占有明显优势，故应用价值极大。

由于本系统没有使用 GPU，故可以自然、合理地用 Intel 16 核 CPU (2400 MHz)及 15.73 GB 内存的通用计算平台来运行，因此端到端处理链路的平均耗时为 63.72 ms，所测得的整体帧率稳定在 15.69 FPS，最佳情况下可达 17.89 FPS，已经大大超过了实时监测一般要求的 15 FPS，因而监控画面十分流畅，告警响应也极及时。

从资源消耗角度分析，在正常运行的过程中 CPU 占用率为 70%~72%，内存的峰值不超过 800~900 M，不需要高性能的机器来支持其运行，使开发更简单，适合中小场景的大规模落地。

结合多帧融合和状态演化的方法可使系统的最后识别准确率达到 90%~93%，误识率为 5%以下，已经达到了满足身份认证类应用对于可靠性基本要求的程度。

从架构设计上来看，级联方案采用“YOLO 人体检测 + ROI 人脸区域提取与 YuNet 人脸检测 + InsightFace 身份识别 + 状态管理”的任务协同方法，利用 YOLOv4tiny 精准定位到人体位置，并将提取 ROI 后，在该区域内使用 YuNet 进行人脸检测，将人脸检测的过程由原有的人脸全局检测变成局部检测，从而降低了平均每次人脸检测所用的时间，即从全图检测平均耗时 18.20 ms 降为本次级联方案中的 11.12 ms，计算开销节省约 38.9%。在端到端链路上，相较全图方案(53.30 ms/18.76 FPS)，级联方案的总耗时降低到 46.22 ms，整体 FPS 达到 21.64 FPS，时间节省约 13.3%，帧率增加约 15.3%，未减损系统检测精度情况下提升了系统处理能力。

综上所述，本文采用“YOLO 人体检测 + ROI 人脸区域提取与 YuNet 人脸检测 + InsightFace 识别 + 状态管理”的级联架构，在任务级计算优化的基础上克服低成本设备的实时人脸识别技术难题，提供给面向普通计算平台的智能监控系统一种可行的、具有效率高、稳定性强和易于复现特点的实现方式，且在现实生活中拥有广泛的应用场景。

## 5. 结论与展望

### 5.1. 结论

考虑到实时监控下人员身份识别对于效率和精度的双重要求，在此基础上提出了 YOLOv4-tiny 和 InsightFace 的任务级级联融合架构[18]。将整个流程分为“人体检测，ROI 人脸区域提取与 YuNet 人脸检测，身份识别，状态管理”五阶段，在实际应用过程中，前一阶段为后一阶段划定处理的边界范围，在该框架下得以解决传统并行式方案带来的计算资源浪费及各子模块间信息不互通的问题，其研究结论如下：

1) 采用人体检测框对人脸搜索区域进行限定，将人脸检测的平均处理范围在空间上缩减约 70%，与全图检测相比，人脸检测阶段平均耗时降低约 38.9%，整体端到端推理效率提升约 15%，并且能够在当前 CPU 平台上实现端到端约 16~18 FPS 的处理速度。

2) 结合滑动窗口的历史识别队列融合，从身份频次和置信度的一致性上判断抑制因光照及姿态等原因引起的瞬时错误，可将系统的误识率从单帧大约为 6%~8%，降到了 3%~5%左右。

3) 建有“已识别，追踪中，已确认”三级状态管控机制，在 10 秒内进行高可信度的身份判断，在目标出现 5 秒内稳定追踪，既能满足对系统的实时性要求，又能满足身份判定的可靠性要求。

在无 GPU 加速的 Intel 16 核 CPU 平台上，系统可对 20 个以上目标的并发稳定追踪，综合识别准确率达到 92%以上。论证了用轻量化模型及级联设计在资源有限的边缘端设备上实现高效可靠人员身份识别的可行性，也由此给出低成本智能监控系统的可行方案。

### 5.2. 展望

本研究达到了预期目标，但系统在极端光照、严重遮挡及远距离等情况下的表现仍可进一步优化[19]。后续计划主要包括：1) 采用更鲁棒的人脸检测器(如 RetinaFace)，增强复杂环境下的检测能力；2) 结合

行人重识别等多类信息，弥补面部特征缺失的不足；3) 对模型进行轻量化设计与推理优化，减少计算成本，从而更好地适配各类边缘设备。

## 基金项目

2025 年广西壮族自治区大学生创新创业训练计划《基于时空融合 LSTM-3DCNN 算法的花山岩画演变预测研究》(编号：S202510604109)阶段性成果。

## 参考文献

[1] 黄伟淇. 基于深度学习 YOLOv3 算法的复杂情景人脸识别研究[D]: [硕士学位论文]. 广州: 广东工业大学, 2024.

[2] 王宸, 刘剑飞, 郝禄国, 等. 一种基于 InsightFace 算法的课堂人脸识别方法研究[J]. 南开大学学报(自然科学版), 2022, 55(2): 59-68.

[3] 张欣, 齐华. 基于 YOLOv4 的人体异常行为检测算法研究[J]. 计算机与数字工程, 2021, 49(4): 791-796.

[4] 钟震宇, 面向公共安全的目标追踪与身份识别关键技术研究与应用[Z]. 广东省, 广东省科学院智能制造研究所, 2023-09-27

[5] 谢瑞洁, 任瑞仙. 基于 Faster R-CNN 的建筑物遥感图像目标检测[J]. 信息记录材料, 2025, 26(9): 113-115.

[6] 赵泰淇, 孙庆峰, 王学谦. 基于红外图像的 YOLOv4-Tiny 行人检测算法的改进[J]. 机械工程与自动化, 2025, 54(4): 59-61.

[7] 张杰, 沐希雅, 黄鑫宇, 等. 基于 YOLOv5 与 ArcFace 的轻量化人脸识别系统研究[J]. 现代信息科技, 2025, 9(15): 38-42.

[8] 王一, 崔振浩, 程佳, 等. 基于加权余弦相似度投票的点对特征位姿估计算法[J]. 组合机床与自动化加工技术, 2025(7): 89-94.

[9] 吴旭莉. 人脸识别的治理困境与规制改进[J]. 厦门大学学报(哲学社会科学版), 2025, 75(4): 103-114.

[10] 宋珂. 基于 GAN 网络的 YOLOv5 人脸检测[J]. 信息记录材料, 2025, 26(11): 196-199.

[11] 杨明昊, 杨文璐. 基于改进 YOLOv8s 的小目标检测算法[J]. 计算机工程与设计, 2025, 46(11): 3049-3055.

[12] 李志涛, 李永康. 基于 OpenCV 的人脸识别系统的设计与实现[J]. 山西电子技术, 2025(5): 73-74, 96.

[13] 邵德强. 基于 Haar 级联分类器的电梯安全人数检测系统设计[J]. 集成电路应用, 2025, 42(7): 142-143.

[14] 庞靖楠, 朱立军. 自然状态下手掌掌纹 ROI 提取[J]. 物联网技术, 2025, 15(10): 16-18, 24.

[15] 丁强, 戴军, 胡建厂. 机器视觉监测预警在识别乘客异常行为中的应用研究[J]. 轨道交通装备与技术, 2025, 33(6): 18-22.

[16] 罗广彬. 基于智能监测与欧氏距离的深基坑安全评估研究[D]: [硕士学位论文]. 深圳: 深圳大学, 2023.

[17] 季丹. 基于 FaceNet 的人脸识别算法研究[J]. 电子设计工程, 2026, 34(1): 145-149.

[18] 黄利红. 基于 Arduino 的低成本智能宿舍监控系统研究[J]. 数字技术与应用, 2021, 39(6): 147-149.

[19] 吴思远, 陈良维, 张靓. 基于人脸识别 NMF 算法的鲁棒性研究[J]. 成都航空职业技术学院学报, 2025, 41(2): 78-83.