基于小目标增强金字塔的高精度定位手部姿态估计

谢青江[1]，周玲珑[2]，耿玉娟[2]，朱洲森[2]，廖雪花[1,*]

[1]四川师范大学计算机科学学院，四川成都 610101
[2]四川师范大学物理与电子工程学院，四川成都 610101

摘要：手部关节姿态评估是手部关节活动度司法鉴定中的关键环节，在司法鉴定领域具有重要意义。为克服传统方法对专业医师高度依赖的局限，本文提出了一种基于计算机视觉的改进型手部姿态评估方法。该方法聚焦于手部 21 个关键点，并针对其在图像中尺寸较小、易被遮挡、特征信息有限等问题，构建了优化的检测框架。具体而言，我们设计了小目标增强特征金字塔，以缓解小关键点特征易丢失的难题；引入 SPDConv 模块以保留细节特征；并结合 Omni-FSAK（Omni-Frequency-Spatial Attention Kernel）模块，实现高效的多尺度特征融合，从而显著提升对模糊、遮挡及细小关键点的感知能力。在检测头设计方面，除分类分支、边界框回归分支和关键点预测分支外，我们进一步引入了位置质量校准器（LQC, Localization Quality Calibrator），充分利用分类分支输出与回归分布信息，对预测结果的可靠性进行评估。该机制有效弥补了分类置信度无法准确反映位置质量的不足，使模型在相同网络规模下能够更稳健地输出高质量的检测与关键点预测结果。实验在 FreiHAND 和 KeypointsHand 数据集上验证了所提出方法的有效性，相较基线方法，准确率提升了 3.15%，召回率提升了 1.74%，充分验证了改进方案的可行性。

关键词：手部关键点检测；小目标检测；司法鉴定；姿态估计；定位质量估计

DOI: 10.57237/j.cst.2026.01.001

High-Precision Hand Pose Estimation Based on Small Object Augmented Feature Pyramid

Xie Qingjiang[1], Zhou Linglong[2], Geng Yujuan[2], Zhu Zhousen[2], Liao Xuehua[1,*]

[1]School of Computer Science, Sichuan Normal University, Chengdu 610101, China
[2]School of Physics and Electronic Engineering, Sichuan Normal University, Chengdu 610101, China

Abstract: Hand joint pose assessment is a critical step in the forensic identification of range of motion, holding significant importance in the field of forensic science. To overcome the limitations of traditional methods that rely heavily on professional physicians, this paper proposes an improved hand pose assessment method based on computer vision. Focusing on 21 hand keypoints, the method addresses challenges such as their small size in images, susceptibility to occlusion, and limited feature information by constructing an optimized detection framework. Specifically, a

基金项目：国家社会科学基金一般项目 (20BMZ092).

*通信作者：廖雪花，liaoxuehua@sicnu.edu.cn

收稿日期: 2025-11-24; 接受日期: 2025-12-26; 在线出版日期: 2026-01-14

Small Object Augmented Feature Pyramid is designed to alleviate the difficulty of losing small keypoint features; the SPDConv module is introduced to preserve detailed features; and the Omni-Frequency-Spatial Attention Kernel (Omni-FSAK) module is incorporated to achieve efficient multi-scale feature fusion, thereby significantly enhancing the perception capability for blurred, occluded, and tiny keypoints. In terms of detection head design, in addition to the classification branch, bounding box regression branch, and keypoint prediction branch, we further introduce a Localization Quality Calibrator (LQC). This mechanism fully utilizes the outputs from the classification branch and regression distribution information to evaluate the reliability of prediction results. It effectively compensates for the insufficiency of classification confidence in accurately reflecting localization quality, enabling the model to output high-quality detection and keypoint prediction results more robustly under the same network scale. Experiments on the FreiHAND and KeypointsHand datasets verify the effectiveness of the proposed method. Compared with the baseline method, the accuracy is improved by 3.15% and the recall rate is increased by 1.74%, fully validating the feasibility of the improved scheme.

Keywords: Hand Keypoint Detection; Small Object Detection; Forensic Identification; Pose Estimation; Localization Quality Estimation

1 引言

手部作为人体最精细、最灵活的运动器官，其功能状态直接关系到个体的日常生活自理能力、工作能力乃至整体生活质量。手部关节活动度是衡量手部功能的核心量化指标，临床实践中广泛采用的评估方法仍以人工测量为主，由康复医师或治疗师手动操作。然而，此类方法耗时费力、效率低下，其测量结果易受到操作者经验、主观判断等多种因素的干扰，导致可重复性差、一致性低，难以满足现代康复医学对精细化、高频次、动态化监测的需求。随着人工智能与计算机视觉技术的飞速发展，利用视觉方法实现手部关节活动度的自动化、无接触式测量，已成为康复工程与生物医学工程领域的研究热点与前沿方向，有望克服传统方法的固有缺陷，为临床提供一种更为客观、高效、标准化的评估新范式。

在手部关键点检测任务中，21 个关键点（包括指尖、指节和掌心关节等）是最常用的标准[1]。然而，这些关键点往往尺寸微小、位置密集[2]，且在自然场景中容易受到遮挡、模糊和光照变化的干扰[3]，使得检测过程充满挑战。以轻量化的 YOLOv11n 为例，其通过大步长下采样获得高层语义特征，从而兼顾效率与速度，但在此过程中，小目标特征往往因多次池化和卷积操作而被严重稀释甚至丢失，导致模型难以在深层特征图中有效表征指尖与关节等微小目标的细节信息[4]。这一瓶颈直接限制了手部关键点检测的上限精度。

针对上述问题，学界提出了多种改进思路。Romero [5]等人通过将手部与身体协同建模，提升了手部姿态估计的真实性；Mokdad [6]等人利用图卷积网络建模双手与物体间的相互依赖关系，增强了交互场景下存在遮挡的姿态识别能力；侯贝贝[7]等人提出了一种基于图卷积网络的关节角度测量方法；Chen [8]等人通过级联裸手姿态估计与关节抖动抑制策略，提高了手部姿态估计的鲁棒性；针对手部姿态识别任务，任建华等人[9]提出基于掩码提示和注意力机制的手部姿态估计方法 HMCA，通过语义分割生成手部掩码图，并设计并行注意力模块与多路残差模块，有效提升了复杂手势识别能力，降低了计算复杂度，解决了多手和遮挡问题；郑泉石等人[10]则提出基于自适应预测的 2D 人体姿态估计方法，通过自适应预测 Query 的关注区域和关键点分布，提升了姿态估计的准确性。Kadam [1]等人利用 YOLOv8-pose 网络实现了手部检测、分类和姿态估计的端到端处理，为 YOLO 架构在手部姿态估计中的应用奠定了基础。相关研究还包括刘佳等人[11]提出的基于 YOLOv3-HM 的估计方法，以及倪广兴等人[12]开发的融合改进 YOLOv5 及 Mediapipe 的手势识别系统。此外，针对手部关键点检测中的密集目标、边缘遮挡和复杂背景等挑战，相关研究通过引入轻量级模块、注意力机制和改进损失函数等方法，提升了手部姿态估计的性能[13-15]。这些研究为手部姿态估计技术的发展提供了重要参考，但在小目标关键点检

测、特征保留和预测可靠性等方面仍有提升空间。

基于此，本文提出了一种针对 YOLOv11-Pose 模型的改进方案，在保持轻量化结构的同时提升手部关键点检测精度。在研究中，我们设计了小目标增强特征金字塔，通过优化特征融合策略来增强对小尺寸关键点的感知能力；引入 SPDConv 模块以替代传统卷积操作，从而在降低分辨率的同时保留全部空间细节，有效缓解关键点特征丢失的问题。其次，引入 Omni-FSAK（Omni-Frequency-Spatial Attention Kernel）模块，利用全局上下文、大感受野与局部细节三分支结构，协同捕获不同尺度的语义与细节特征，并嵌入 CSP 结构以降低计算冗余，从而显著提升对模糊、遮挡和细粒度关键点的感知能力。最后，在检测头中引入位置质量校准器（LQC, Localization Quality Calibrator），通过结合分类分支的输出与边界框回归的分布特征，对预测框的空间质量进行建模，动态修正分类置信度，使得候选框得分更好地反映实际定位精度。这一设计间接提升了关键点分支的预测可靠性，增强了整体姿态估计的鲁棒性。

为验证本文方法的有效性，本文在 FreiHAND [16] 与 KeypointsHand 两个公开手部关键点数据集上进行了系统实验。实验结果表明，所提出的改进模型在保持轻量化结构的前提下，有效提升了检测与关键点估计的性能，相较基线模型 mAP50 提升了 3.15 个百分点，召回率提升了 1.74 个百分点。进一步的消融实验也验证了各改进模块的独立贡献与协同效果，证明了所提出方法在精度与速度平衡方面的优越性。

2 基本原理

2.1 YOLOv11 算法介绍

YOLO（You Only Look Once）作为一种 one-stage 目标检测算法，仅需单次网络前向传播即可同时完成图像中物体类别的识别与边界框的定位。YOLOv11 是 Ultralytics 公司推出的最新一代检测模型，它在继承 YOLOv8 优良特性的基础上进行了全面升级，能够高效支持图像分类、目标检测及姿态估计等多样化任务。该版本进一步优化了核心架构，通过引入改进的骨干网络与更先进的检测头设计，配合重新设计的损失函数，显著提升了模型在不同硬件平台（包括 CPU 和 GPU）上的推理效率与检测精度。为了适应差异化的应用需求，YOLOv11 延续了 n、s、m、l、x 五种不同尺寸的模型配置。其中，YOLOv11n 凭借最少的参数量和最浅的网络深度，成为轻量级部署的理想选择；而 YOLOv11-Pose 则是基于该架构专门针对人体关键点任务扩展的变体，在保持 YOLO 算法“即看即识”的高效性同时，实现了对手部等目标姿态的精准估计。

2.2 YOLOv11n-pose 网络结构

YOLOv11n-pose 按照网络结构可分为：输入、主干、颈部和检测头这 4 部分。

输入是模型的起点，负责接收待处理的图像数据。主干是模型的基础，负责从输入图像中提取多层次、深浅结合的视觉特征，由高效的标准卷积、RepNCSPELAN4 [17] 模块和 SPPF [18]模块组成。主干网络提取的这些特征是后续网络层进行手部目标检测与关键点定位的基础。颈部网络位于主干网络和头部网络之间，它的主要作用是对主干网络提取的不同尺度的特征图进行高效融合和增强。由于手部目标在图像中通常尺寸较小且姿态多变，因此通过结合特征金字塔网络和路径聚合网络的双向融合结构，来提升模型对手部目标的感知能力和关键点定位的精度。检测头是模型的决策部分，负责产生最终的检测结果。它根据颈部网络提供的融合特征图，对每个特征点进行分类（判断是否为手部）、定位（边界框预测）以及关键点回归（手部关键点坐标预测）。

YOLOv11n-pose 模型输入的是一张尺寸为 640×640 的图像，然后利用卷积神经网络将输入图像转换成特征图，在每个网格中，会预测边界框的位置、大小、手部存在的概率以及手部关键点的位置，并根据预测结果生成手部检测结果与关键点坐标。YOLOv11n-pose（手部关键点检测版本）的网络结构如下图 1 所示：

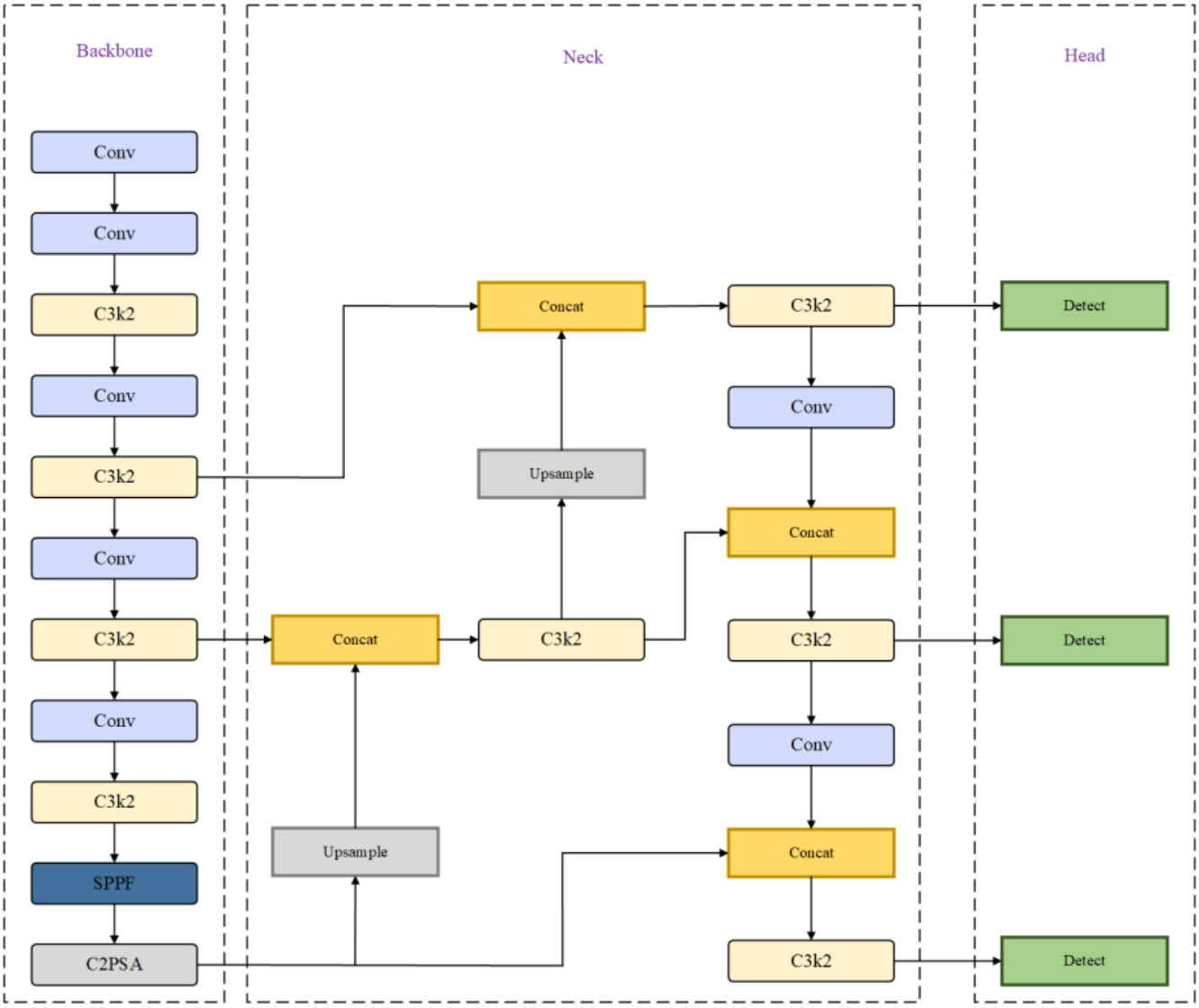

图 1 YOLOv11n-pose 网络结构图

3 改进 YOLOv11n 网络结构

3.1 改进网络结构

在手部关键点检测任务中，指尖、指间关节等关键点尺寸微小，其在图像中的特征信息本就有限。当使用 YOLOv11n 等轻量化网络进行检测时，其固有的大步长下采样策略会导致这些微小目标在经过多次池化和卷积后，特征信息被严重稀释甚至丢失。尤其是在较深的特征层（如 P3、P4、P5），网络难以有效学习和表征这些已经模糊不清的小目标特征，从而限制了检测精度的上限。为解决这一问题，当前研究中一种常见的思路是引入 P2 检测层。P2 检测层因其更高的分辨率（下采样倍数更小），能够保留更丰富的细节信息，从而有效捕捉手部关键点的细微特征。此外，通过高效的特征融合机制，P2 层能将包含精确位置信息的浅层特征与富含语义信息的深层特征相结合，显著提升了对微小关键点的定位与识别能力。然而，在 YOLOv11n 这类追求极致效率的模型中增加 P2 检测层，不可避免地会带来计算量和参数量的激增，导致模型推理速度下降，后处理耗时增加，这与手部实时交互、康复评估等应用场景对低延迟的严苛要求相悖。这种“精度-速度”的矛盾，促使研究者们探索更为高效的特征金字塔结构。本文正是基于此背景，在 YOLOv11n 原有的 PAFPN（Path Aggregation Feature Pyramid Network）[19]结构上进行改进，提出了一种轻量化的 MEFP（Micro-Enhanced Feature Pyramid）模块，旨在

以最小的计算开销，显著增强模型对手部微小关键点的特征感知与定位能力。以下是 MEFP 模块在 YOLOv11n 手部关键点检测网络中的具体实现过程，其模块结构如图 2 所示。

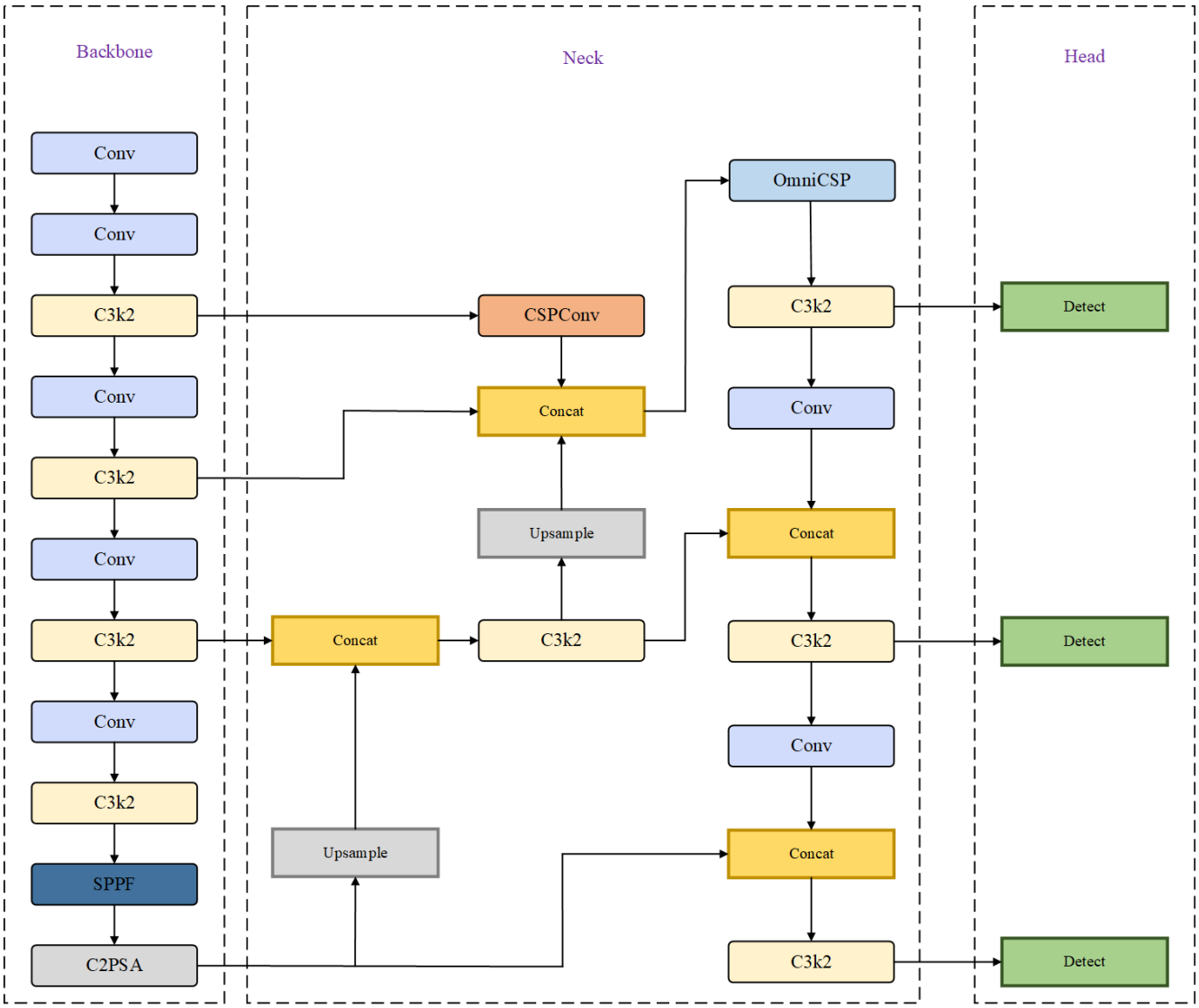

图 2 改进 YOLOv11n-pose 网络结构

3.2 CSPConv 模块

在手部关键点检测任务中，骨干网络提取的 P2 层高分辨率特征图（160×160）是定位指尖、关节等微小目标的基石。然而，在将其与 P3 层特征融合时，传统的下采样方法（如带步长的卷积）会因空间分辨率的急剧下降而破坏这些宝贵的细节信息。YOLO 最初主要用于目标检测任务，特征提取可以忽略一些无效的细节信息。但是与目标检测不同，手部关键点检测需要最大程度保留特征的细节信息。为克服这一瓶颈，本文采用 SPDConv 模块，提供 P2 层特征图的细节信息，并融合 P3 层实现无损的特征尺度转换来实现姿态评估对小目标的学习。

SPDConv 采用“空间到深度”的变换策略。SPDConv 由空间到深度（SPD）层和非步长卷积层组成，它并非简单地丢弃或聚合空间信息，而是通过一个巧妙的切片与拼接操作，将特征图的空间维度信息（如 160×160）编码为通道维度信息（如 4C1）。具体而言，一个 $S \times S \times C1$ 的特征图被分解为 4 个 $S/2 \times S/2 \times C1$ 的子图，并在通道轴上堆叠，形成 $S/2 \times S/2 \times 4C1$ 的新特征表示。这一过程在降低特征图空间尺寸的同时，完整保留了所有原

始像素的细节，该模块的结构示意图如图 4 所示。

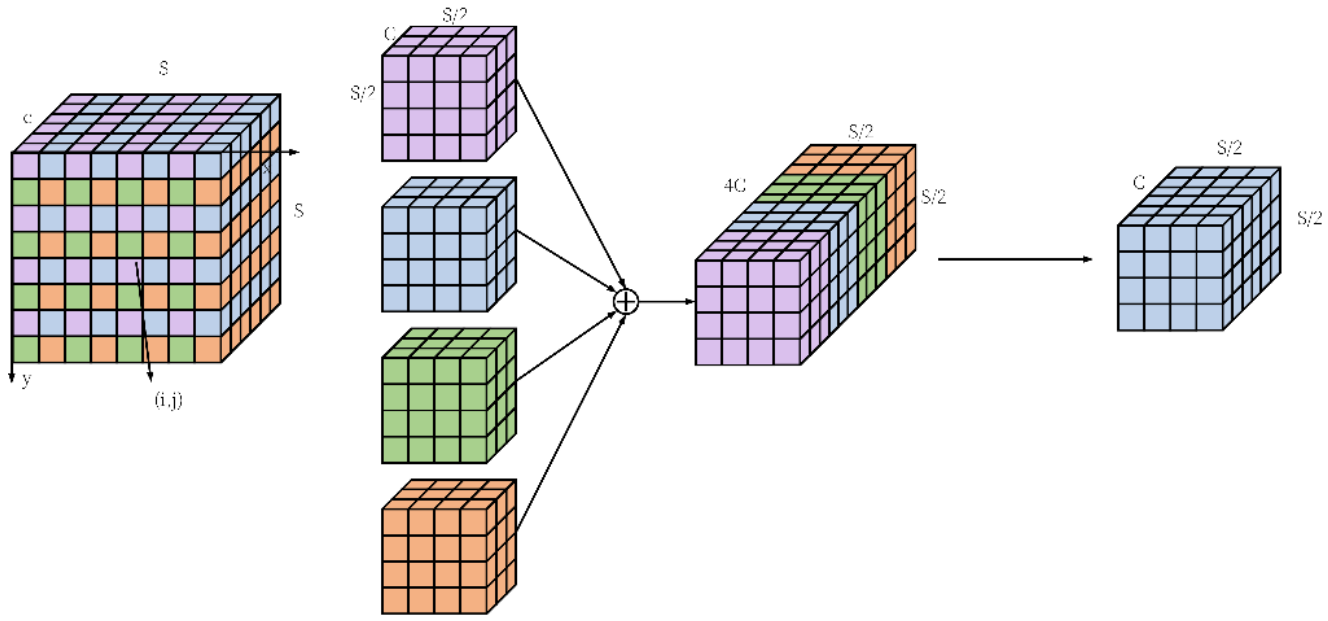

图 3 CSPConv 结构图

随后，一个步长为 1 的非步长卷积被应用于变换后的特征，其主要目的是对新增的通道维度进行特征学习与整合，而非进行空间压缩。通过这种方式，SPDConv 成功地将 P2 层富含细节的高分辨率特征，无损地适配到 P3 层的尺度，为后续的特征金字塔融合提供了高质量的输入，确保了微小目标特征的有效传递，并在上下文范围内进行信息聚合，提高模型的检测精度。

3.3 Omni-FSAK 模块

在经过 CSPConv 模块初步融合与增强后，得到的特征图能够兼顾低层细节与中层语义信息，为后续检测与回归任务提供了较好的特征表达。然而，这类融合特征在建模长程依赖与大感受野信息时仍存在局限：传统卷积核的局部感受野限制使得模型在面对小目标检测或关键点回归等任务时，往往难以充分捕获全局上下文关系，导致预测结果对复杂场景的适应性不足。为缓解这一问题，我们在检测头中引入了 Omni-FSAK 模块。

Omni-FSAK 模块采用 CSP（Cross Stage Partial）结构思想进行特征分流再融合，先用一层 1×1 的投影将通道分成两部分，一部分走轻量但拥有超大卷积核的 OmniKernel 分支以捕获宽广上下文，另一部分作为 identity 保留原始细节，然后再将两部分在通道维度拼接并通过 1×1 投影融合，既扩大了有效感受野又控制了计算量。其结构图如图 4 所示。

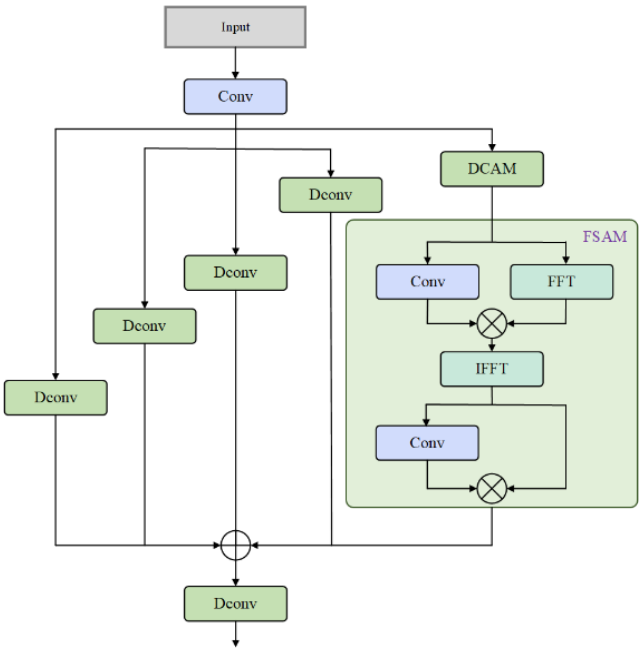

图 4 Omni-FSAK 模块结构图

模块内部可用如下形式精确描述。对输入特征 $x \in \mathbb{R}^{B\times reg\times H\times W}$，先经过第一个 1×1 映射得到 $z = cv1(x)$ 并按比例 e 分裂为两部分：

$$z=[z^{ok};z^{id}],\quad z^{ok}\in\mathbb{R}^{B\times C_e\times H\times W};\quad C_e=e\cdot C',\quad z^{id}\in\mathbb{R}^{B\times(C'-C_e)\times H\times W} \tag{1}$$

其中C'为cv1的输出通道数。对较窄的z^{ok}送入 OmniKernel：定义$M(\cdot)$表示 OmniKernel 的映射，则模块的输出为：

$$y=\mathrm{cv}2\big(\mathrm{concat}(M(z^{ok}),z^{\mathrm{id}})\big) \tag{2}$$

OmniKernel 的核心计算可用如下符号化表达。令$u=\mathrm{in_conv}(z^{\mathrm{ok}})=\phi(W_{\mathrm{in}}*z^{\mathrm{ok}})$（其中$\phi=\mathrm{GELU}$），并设大核宽度为k（代码中$\mathrm{k}=31$），则深度可分离卷积分支给出局部细节分支：

$$d_{1\times k}=\mathrm{DW}_{1\times k}(u),d_{k\times 1}=\mathrm{DW}_{k\times 1}(u),d_{k\times k}=\mathrm{DW}_{k\times k}(u),d_{k\times k}=\mathrm{DW}_{1\times 1}(u) \tag{3}$$

这里的DW表示 depthwise 卷积（groups＝通道数），从而能以较低参数量实现大感受野近似。为引入频域的全局语义交互，先计算一个频域通道加权（FCA，frequency channel attention）：对u做全局平均得到通道描述并通过点卷积生成频域权重a_{f}：

$$a_f=W_{\mathrm{fac}}\big(\mathrm{GAP}(u)\big),\qquad U(\omega)=\mathcal{F}\{u\}(\omega) \tag{4}$$

在频域点乘后逆变换得到频域增强特征并取幅值：

$$U'(\omega)=a_f\odot U(\omega),\qquad u_{\mathrm{fca}}=\left|\mathcal{F}^{-1}\{U'(\omega)\}\right| \tag{5}$$

随后对$\mathrm{u}_{\mathrm{fca}}$做一次空间/通道的缩放注意（SCA，spatial/channel attention）：先用全局池化与 1×1 映射产生缩放系数a_{s}，再做逐通道相乘并通过 FGM（特征门控模块）细化：

$$a_s=W_{\mathrm{sca}}\big(\mathrm{GAP}(u_{\mathrm{fca}})\big),\qquad u_{\mathrm{sca}}=\mathrm{FGM}\big(a_s\odot u_{\mathrm{fca}}\big) \tag{6}$$

此分支专注于保留和增强指尖、关节等微小目标的细粒度特征。它通过轻量化的逐点深度卷积，对特征图进行像素级的精细调整，确保关键细节信息在深层网络传递中不被稀释。最终将所有分支残差化相加并激活、投影回输出通道：

$$s=u+d_{1\times k}+d_{k\times 1}+d_{k\times k}+d_{1\times 1}+u_{\mathrm{sca}},\qquad s'=\mathrm{Re}\,LU(s),\qquad y_{\mathrm{ok}}=W_{\mathrm{out}}*s' \tag{7}$$

因 Omni Kernel 的总体映射为$\mathrm{M}(z^{ok})=y_{ok}$，而CSP Omni Kernel通过与identity通道的拼接与1×1融合保证了信息的低开销保留与高感受野增强的兼顾。

完成Omni-FSAK处理后，特征图通过上采样操作（如最近邻插值或转置卷积）进行尺度恢复。此举旨在将低分辨率、高语义的深层特征图提升至与高分辨率、高细节的浅层特征图相匹配的尺度，为后续的特征融合奠定基础。最终，这些经过上采样的特征将与骨干网络中对应尺度的浅层特征进行跨层连接与融合。这一步至关重要，它将深层丰富的语义信息与浅层精确的位置信息有机结合，从而生成一个既包含高级语义又保留精细细节的最终特征图。

最后，该融合特征图被送入解耦头进行最终处理。解耦头将分类任务（判断是否为关键点）与回归任务（定位关键点坐标）分离处理，有效避免了任务间的相互干扰，提升了模型的收敛速度与定位精度。综上所述，通过 CSPConv、Omni-FSAK、上采样与跨层融合、解耦头这一系列精心设计的处理流程，本文方法能够在不显著增加模型计算负担的前提下，显著提升 YOLOv11n-pose 对手部关键点的检测性能。

3.4 检测头改进

在关键点检测任务中，传统的 Pose Head 主要依靠分类分支、边界框回归分支以及关键点预测分支来完成检测，这种解耦式的独立预测方式在处理目标时具有较高效率，但存在一个固有问题，这种结构在实际应用中虽然能够实现目标检测与姿态估计的结合，但分类得分与边界框质量往往缺乏一致性。模型在得出一个较高的类别置信度，但其对应的边界框若存在明显的偏移或者回归不准时，这导致在 NMS 阶段候选框筛选依赖的分数往往偏离实际的定位精度和关键点回归，会直接影响后续关键点的预测精度和整体任务的性能。原始模型检测头的结构如下图 5 所示：

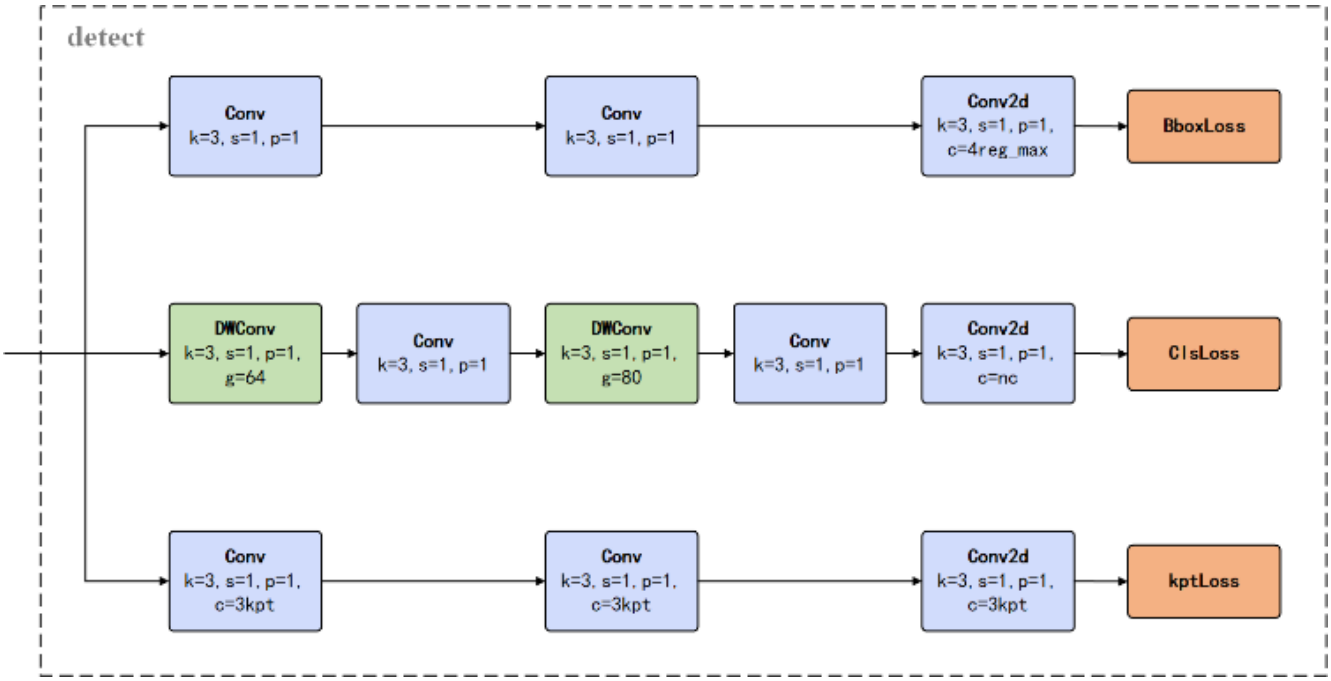

图 5 检测头 Detect 结构图

为解决上述问题，我们在检测头中引入了位置质量估计器（LQC）模块。该模块通过对回归分布的统计建模，学习预测框的几何可靠性，从而得到一个额外的质量分数 $q(b\mid x)$。具体而言，回归分布输出为：

$$P=\mathrm{Softmax}\left(\mathrm{reshape}\left(pred\_corners,(B,reg\_max,4,H,W)\right)\right) \tag{8}$$

其中 $P\in\mathbb{R}^{B\times reg_max\times 4\times H\times W}$ 表示每个边界框四个边的离散分布。随后，我们选取前 k 个概率值并拼接均值作为统计特征：

$$S=\mathrm{Concat}\left(\mathrm{TopK}(P,k),\mathrm{Mean}(\mathrm{TopK}(P,k))\right) \tag{9}$$

并通过一个多层感知机（MLP）映射为质量分数修正量：

$$q(b\mid x)=\mathrm{MLP}(S) \tag{10}$$

最终，分类分数被动态调整为：

$$s'(c\mid x)=s(c\mid x)+q(b\mid x) \tag{11}$$

使得候选框的排序与其空间位置精度更加一致。

在损失函数主要由 三部分损失组成：边界框回归损失 L_{box}，采用 DFL（Distribution Focal Loss），分类损失 L_{cls}，采用 BCE 或 VariFocal Loss；关键点回归损失 L_{kpt}，采用 L1 或 MSE。

引入 LQC 模块后，额外增加了一个位置质量损失 L_q，它通过回归预测质量分数与真实 IoU 的差异来优化，使得模型在学习分类时也兼顾位置置信度。整体目标函数可写为：

$$L=\lambda_{box}L_{box}+\lambda_{cls}L_{cls}+\lambda_{kpt}L_{kpt}+\lambda_q L_q \tag{12}$$

当这种机制被应用到 Pose Head 中时，虽然 LQC 并不直接改变关键点分支的计算过程，但由于关键点预测往往依赖于目标边界框的准确性和可靠性，LQC 通过提高检测框质量，实际上减少了低质量框进入关键点预测阶段的概率。这意味着，在同样的网络规模下， LQC 能够更稳健地输出高质量的检测与关键点结果，从而提升姿态估计的整体表现。

因此，相比原始的 YOLOv11-Pose，加入 LQC 的版本在结构上增加了边界框分布质量评估的环节，在功能上增强了检测分数与定位质量之间的一致性，在效果上则提升了目标检测与关键点检测任务中的整体精度和鲁棒性。这种改进尤其适合轻量化网络场景以及对小目标和细粒度关键点检测有较高要求的应用环境。

改进检测头模型为解决原始模型检测头计算量大的问题，通过减少卷积操作的数量、实现参数共享、简化模型结构等方式，使得 YOLOv8 模型在处理速度、内存使用和计算效率上都得到了提升，同时保持了检测性能。具体改进后的检测头结构如下图 6 所示。

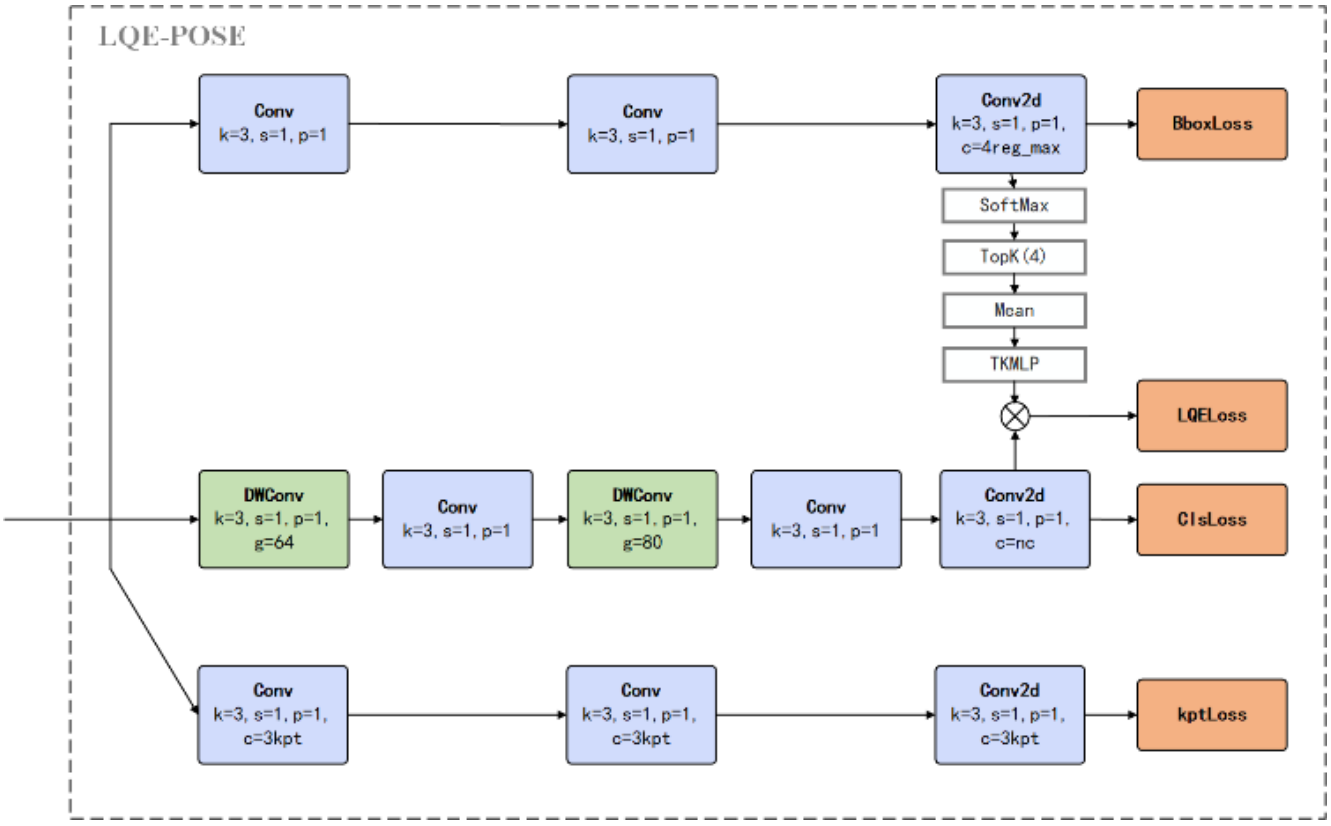

图 6 改进后的 Detect 结构图

4 实验与结果分析

4.1 实验数据集收集与处理

本实验的数据集由 FreiHand 数据集和 keypointsHand 数据集抽取部分后共同构成。

FeiHAND 是由弗莱堡大学和 Adobe 研究院于 2019 年联合发布的一个 3D 手部姿态数据集，专为手部识别研究设计。该数据集记录了 32 名受试者执行的各种手部动作。FreiHAND 特别适合在实验室环境下的手部识别研究，样本采用了三种不同的后处理策略进行数据增强，但评估样本保持了原始状态，这种设计有助于评估算法的泛化能力，确保模型不仅在训练数据上表现良好，也能适应新的、未见过的手部姿态和环境。

keypoints Hand 数据集是由慕尼黑工业大学（Technical University of Munich, TUM）的研究团队于 2021 年提出的一个手部关键点数据集。这个数据集专注于手部关键点检测和姿态估计任务。数据集与 Ultralytics YOLO11 格式兼容，包含 26,768 张用关键点注释的手部图像，使其适合训练像 Ultralytics YOLO

这样的模型以进行姿势估计任务。注释是使用 Google MediaPipe 库生成的，确保了高准确率和一致性。

将以上两个数据集按本文研究内容的需求选取后，按照 8:1:1 的比例随机划分为训练集、验证集和测试集，分别包含 8244 张、1032 张和 1032 张图片；并将其对应标签文件格式都转换为 YOLO 格式，以便实验使用。

4.2 实验环境及参数配置

本研究在 Windows11 操作系统下搭建了实验平台，搭配 Nvidia Tesla P100 显卡，Pytorch1.12 深度学习框架以及 Python3.9 编程语言。通过大量实验得出，模型训练的迭代轮次为 300 时可达到最佳收敛状态，设置初始学习率为 0.01、动量因子为 0.937，有助于模型训练的稳定性，设置 flipud 和 mixup 为 0.5，对图片进行反转混合，增加训练数据的多样性和模型的泛化能力。为了验证实验的科学性和有效性，设置了消融实验和对比实验，以探究改进模型与其他模型对 YOLOv8 模型的性能影响。实验参数设置如下表 1 所示：

表 1 实验训练参数配置表

参数名称	参数值
Learning Rate	0.01
Epochs	300
Batch Size	32
Image Size	640 x 640
Momentum	0.937
Flipud	0.5
Mxiup	0.5

4.3 评估指标

为了全面评估所提模型在手部姿态估计任务中的综合性能，本研究采用精确率（Precision）、召回率（Recall）、平均精度均值（mAP@0.5）作为核心评价指标[20]。各项指标的定义与计算方式如下：

精确率用于衡量模型预测为正样本中真正为正样本的比例，反映了模型预测的准确性，其计算公式为：

$$\text{Precision} = \frac{TP}{TP + FP} \tag{13}$$

其中，TP（True Positives）表示被正确检测出的目标数量，FP（False Positives）表示被错误检测出的目标数量。精确率越高，说明模型误检越少，预测结果越可靠。

召回率表示模型检测出的正样本占所有实际正样本的比例，反映了模型对目标的覆盖能力，其计算公式为：

$$\text{Recall} = \frac{TP}{TP + FN} \tag{14}$$

其中，FN（False Negatives）表示未被模型检测出的实际目标数量。召回率越高，说明模型漏检越少，对目标的识别越全面。

平均精度均值（mAP@0.5）是综合评估模型在多个类别上检测性能的关键指标，其计算方式为对各类别平均精度（AP）求均值：

$$\text{mAP@0.5} = \frac{1}{c}\sum_{i=1}^{c} \text{AP}_i \tag{15}$$

其中，c 表示检测类别数，m 表示目标类别的数量，AP 为某一类别 P-R 曲线下的面积。 AP_i 是第 i 个类别的平均精度，T 是 IOU 阈值的数量，AP_t 是在 IOU 阈值为 t 时的平均精度。

4.4 消融实验

本文提出了 3 个改进点，为验证本文算法改进的模块均具有有效性，以 YOLOv11n-pose 网络为基线，进行消融实验，结果如表 2 所示。通过逐步引入 CSPConv、OmniCSP 和 LQC 三个改进模块，验证各模块及其组合对模型性能的影响。基线模型不包含任何改进模块时，mAP50 为 63.36%。单独使用各模块时，OmniCSP 表现最佳，使 mAP50 提升至 64.32%，而 CSPConv 和 LQC 分别提升至 63.91%和 63.83%。两两组合实验中，OmniCSP 与 LQC 的组合效果最为显著，mAP50 达到 65.32%，CSPConv 与 OmniCSP 组合次之，为 64.89%。当三个模块全部集成时，模型性能达到最优，mAP50 提升至 66.51%，相比基线模型提高了 3.15 个百分点。实验结果清晰地表明，三个改进模块均能有效提升模型性能，且存在明显的协同效应，特别是 OmniCSP 与 LQC 的组合对性能提升贡献较大，验证了这些改进模块的有效性和互补性。所有改进方案在提高精确率的同时都伴随着召回率的轻微下降，这表明模型倾向于牺牲一定的召回率以减少误检，这种策略对于手部姿态估计在司法鉴定、康复评估等对误检敏感的应用中，这种权衡是合理且有价值的。

表 2 消融实验结果

实验	CSPConv	OmniCSP	LQC	mAP50/%	P/%	Recall/%
YOLOv11n-pose	—	—	—	63.36	62.06	63.67
1	√	—	—	63.91	65.19	61.06
2	—	√	—	64.32	65.43	61.06
3	—	—	√	63.83	65.34	61.53
4	√	√	—	64.89	65.91	62.06
5	√	—	√	64.67	65.61	61.82
6	—	√	√	65.32	66.41	62.53
7	√	√	√	66.51	67.21	63.80

4.5 对比实验

为验证本文改进网络模型的检测效果，将改进的 YOLOv11s-pose 与 YOLOv11s-pose、YOLOv10s-pose、YOLOv8s-pose、keypoint-RCNN 等目标检测算法在同一实验环境和条件下进行训练，并对测试集检测效果进行比较，各算法检测对比结果如下表所示。

表 3 对比实验结果

模型	mAP50%
YOLOv11s-pose	63.36
YOLOv8s-pose	62.94
YOLOv10s-pose	62.05
keypoint-RCNN	48.43
ours	66.51

由表可得，本文改进的模型在 AP%指标上达到 66.51%，显著优于基线模型 YOLOv11s-pose 的 63.36%，同时也超过了 YOLOv8s-pose 的 62.94% 和 YOLOv10s-pose 的 62.05%。与传统方法 keypoint-RCNN 相比，本文模型更是表现出明显优势，其 AP%值仅为 38.43%，远低于本文提出的改进算法。实验结果表明，通过引入 CSPConv、OmniCSP 和 LQC 三个改进模块，有效提升了模型的检测性能。综合考虑检测精度、参数量和计算量等指标，本文改进的 YOLOv11s-pose 模型与多种目标检测算法相比具有更优越的性能，验证了所提改进方法的有效性和实用性。

4.6 可视化分析

为验证本研究的模型改进实际效果，将 YOLOv11n-pose 模型与我们改进模型的检测热力图结果进行可视化分析比较，比较结果展示如图 7 所示，原始图像为第一排(a)所展示，图片依次为全掌展开、全掌并拢、示指微伸，指捻物品、示指屈曲，YOLOv11n-pose 模型检测热力图展示如(b)，本研究的改进模型展示结果为第三排(c)所示。

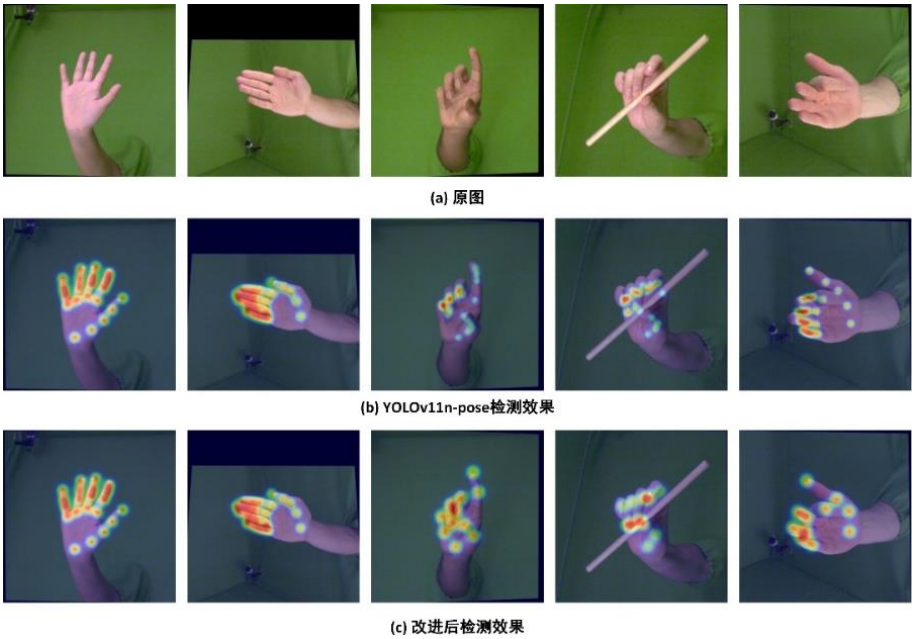

图 7 改进前后模型检测结构可视化对比

由图 7 所示在手掌完全展示时，两个模型均能准确的捕捉到手指的各个关键点，热力图中关节点准确的标注在了对应原图关节上。而非手掌完全展示时，YOLOv11n-pose 基础模型的预测结果会容易出现误判，如在示指微伸时，由于存在另外三个手指遮挡手掌的情况，其手指的关键表标注集中在了近指关节周围，而远指关节关键点标注存在明显偏移。在手指捻起物体时，YOLOv11n-pose 基础模型对手指指尖关键的预测存在一定的偏差，而改进算法较为准确的识别了指尖，受遮挡物影响较小。在示指屈曲时，两者对除示指外的指尖关节判断均正确，而在示指上存在差异，改进模型较为正确的定位到了屈曲的示指。由此从实际应用角度证明了本文的改进算法的在手部关节姿态评估有效性。

5 结论与展望

本文针对手部关键点检测与姿态估计任务中存在的小目标特征易丢失、遮挡影响显著以及分类置信度与定位质量不一致等问题，在 YOLOv11n-pose 的基础上提出了一种改进方法。本文设计了 CSPConv 模块，利用空间到深度的无损变换方式有效保留高分辨率细节特征；提出了 Omni-FSAK 模块，通过全局上下文、大感受野和局部细节三分支结构实现对多尺度信息的高效建模；在检测头中加入位置质量校准（LQC）模块，使分类分数与定位精度更加一致，减少低质量框对关键点预测的干扰。实验结果表明，相较于改进前的模型改进后的 YOLOv11n-pose 模型的整体性能显著优于基线模型，mAP50 提升了 3.15%，召回率提升了 1.74%，同时在复杂姿态和遮挡场景下表现更为稳健。

尽管本文提出的改进方法在检测精度与关键点定位能力方面取得了较大提升，但仍存在一定的不足。随着模型结构的复杂化，参数量和计算量较基线模型有所增加，可能对资源受限的终端设备部署带来挑战。后续研究将探索更高效的轻量化设计，结合模型剪枝、知识蒸馏等技术，在保证精度的同时降低推理延迟，提升模型的实用性；针对实际应用需求，研究跨领域泛化能力，提高模型在司法鉴定、康复评估及人机交互等多种场景中的适应性，为未来研究人体关节活动度手部关节活动检测做进一步。

参考文献

[1] Kadam P, Fang G, Amirabdollahian F, et al. Hand Pose Detection Using YOLOv8-pose [C] // 2024 IEEE Conference on Engineering Informatics (ICEI). IEEE, 2024: 1-6.

[2] Che W, Zhang H, Wu B, et al. SFG-YOLOv8: efficient and lightweight small-feature gesture keypoint detector [J]. Journal of King Saud University Computer and Information Sciences, 2025, 37(4): 44.

[3] Myanganbayar B, Mata C, Dekel G, et al. Partially occluded hands: A challenging new dataset for single-image hand pose estimation [C] // Asian Conference on Computer Vision. Cham: Springer International Publishing, 2018: 85-98.

[4] Jia X, Feng J, Liang B. Hand keypoint detection with super resolution [C] // Proceedings of the 7th International Conference on Cyber Security and Information Engineering. 2022: 398-401.

[5] Romero J, Tzionas D, Black M J. Embodied hands: Modeling and capturing hands and bodies together [J]. arXiv preprint arXiv: 2201.02610, 2022.

[6] JING L, WANG B. EMNet: Edge-guided multi-level network for salient object detection in low-light images [J]. Image and Vision Computing, 2024, 143: 104933.

[7] 侯贝贝.基于手部姿态估计的手关节角度测量系统研究[D]. 中国科学院大学 (中国科学院西安光学精密机械研究所), 2024. https://doi.org/10.27605/d.cnki.gkxgs.2024.000060

[8] Chen M, Shuang F, Li S, et al. Robust dexterous hand control strategy cascading bare hand pose estimation and joint jitter suppression [J]. Robotics and Autonomous Systems, 2025: 105189.

[9] 任建华,曹佳惠,贾迪.基于掩码提示和注意力的手部姿态估计[J/OL].计算机应用, 1-11 [2025-09-29]. https://doi.org/10.11772/j.issn.1001-9081.2024111715

[10] 郑泉石,金城.基于自适应预测的 2D 人体姿态估计[J].计算机科学, 2023, 50(S2): 162-168. https://doi.org/CNKI:SUN:JSJA.0.2023-S2-023

[11] 刘佳, 石豪, 陈大鹏,等. 基于 YOLOv3-HM 的手部姿态估计方法研究 [J]. 测控技术, 2023(004): 60-66, 87.

[12] 倪广兴, 徐华, 王超. 融合改进 YOLOv5 及 Mediapipe 的手势识别研究 [J]. 计算机工程与应用, 2024, 60(07): 108-118.

[13] Ding J, Niu S, Nie Z, et al. Research on human posture estimation algorithm based on YOLO-Pose [J]. Sensors, 2024, 24(10): 3036.

[14] Zhu X, Lyu S, Wang X, et al. TPH-YOLOv5: Improved YOLOv5 based on transformer prediction head for object detection on drone-captured scenarios [C] // Proceedings of the IEEE/CVF international conference on computer vision. 2021: 2778-2788.

[15] Guan X, Shen H, Nyatega C O, et al. Repeated Cross-Scale Structure-Induced Feature Fusion Network for 2D Hand Pose Estimation [J]. Entropy, 2023, 25(5): 724.

[16] Zimmermann C, Ceylan D, Yang J, et al. Freihand: A dataset for markerless capture of hand pose and shape from single rgb images [C] // Proceedings of the IEEE/CVF international conference on computer vision. 2019: 813-822.

[17] Zhang Y, Du S, He H. Rotator-YOLOv5: Improved YOLOv5 for Vehicle and Vessel Detection in UAV Images [C] // 2024 Fourth International Conference on Digital Data Processing (DDP). IEEE, 2024: 156-161.

[18] Zhao P. SPFFNet: Strip perception and feature fusion spatial pyramid pooling for fabric defect detection [J]. arXiv preprint arXiv: 2502.01445, 2025.

[19] Yang G, Lei J, Zhu Z, et al. AFPN: Asymptotic feature pyramid network for object detection [C] // 2023 IEEE International Conference on Systems, Man, and Cybernetics (SMC). IEEE, 2023: 2184-2189.

[20] 李少东, 罗凯, 黄远智, 等. 复杂交互场景下融合关节遮挡信息的手部姿态估计研究 [J]. 计算机学报, 2025, 48(05): 1212-1231.