Artificial Intelligence and Robotics Research 人工智能与机器人研究, 2026, 15(4), 1085-1096
Published Online July 2026 in Hans. https://www.hanspub.org/journal/airr
https://doi.org/10.12677/airr.2026.154098

改进的YOLOv11s无人机手势识别算法

迟　晨，魏立臻，赵菁菁，朱以墨，张丽艳*

大连交通大学，轨道智能工程学院，电子与通信工程系，辽宁 大连

收稿日期：2026年5月4日；录用日期：2026年7月10日；发布日期：2026年7月22日

摘　要

针对标准YOLOv11算法在直接应用于无人机手势识别时，面临手部特征提取不充分、多尺度手势适应能力不足以及边界框定位精度有限等问题，导致检测效果难以满足实际需求。为此，本文从特征提取、多尺度融合及损失函数三个方向对YOLOv11进行改进，提出适用于无人机手势识别任务的改进算法。首先，针对手部特征提取能力不足的问题，在Backbone阶段采用C2PSA_EDFFN模块替换原有的C2f模块。该模块通过增强特征提取过程中的上下文信息交互，有效提升了网络对手部关键特征的捕获能力，改善了对手势区域的鉴别效果。其次，针对不同尺度手势(如远距离小目标手势与近距离大目标手势)的检测适应性问题，在Neck阶段引入CGAFusion模块优化多尺度特征融合策略。该模块能够自适应整合浅层细节信息与深层语义信息，提升了模型对不同尺度手势的检测能力。最后，针对手势边界框定位精度不足的问题，采用PIoU2损失函数替代原有的CIoU损失函数。该损失函数能够更精确地评估预测框与真实框之间的重叠程度，提高了手势目标边界框的回归精度。

关键词

无人机手势识别，多尺度特征融合，注意力特征提取，PIoU2边界框回归

Improved YOLOv11s Algorithm for UAV Gesture Recognition

Chen Chi, Lizhen Wei, Jingjing Zhao, Yimo Zhu, Liyan Zhang*

Department of Electronic Communications, School of Railway Intelligent Engineering, Dalian Jiaotong University, Dalian Liaoning

Received: May 4, 2026; accepted: July 10, 2026; published: July 22, 2026

Abstract

When directly applied to UAV-based hand gesture recognition, the standard YOLOv11 algorithm

*通讯作者。

文章引用: 迟晨, 魏立臻, 赵菁菁, 朱以墨, 张丽艳. 改进的 YOLOv11s 无人机手势识别算法[J]. 人工智能与机器人研究, 2026, 15(4): 1085-1096. DOI: 10.12677/airr.2026.154098

suffers from insufficient hand feature extraction, limited adaptability to multi-scale gestures, and suboptimal bounding box localization accuracy, making it difficult to satisfy practical application requirements. To address these limitations, this paper proposes an improved YOLOv11-based algorithm tailored for UAV hand gesture recognition by enhancing feature extraction, multi-scale feature fusion, and loss function design. Specifically, to improve hand feature representation, the C2PSA_EDFFN module is introduced in the backbone to replace the original C2f module, which enhances contextual information interaction during feature extraction and strengthens the network's ability to capture critical hand features. To better handle gestures at different scales, a CGAFusion module is incorporated into the neck to optimize multi-scale feature fusion by adaptively integrating shallow detailed information with deep semantic features, thereby improving detection performance across varying gesture scales. Furthermore, to enhance bounding box localization accuracy, the PIoU2 loss function is employed to replace the original CIoU loss, enabling a more precise evaluation of the overlap between predicted and ground-truth bounding boxes and improving regression accuracy for gesture localization.

Keywords

UAV-Based Hand Gesture Recognition, Multi-Scale Feature Fusion, Attention-Based Feature Extraction, PIoU2 Bounding Box Regression

1. 引言

随着无人机技术的快速发展，基于视觉的手势识别作为一种自然、高效的人机交互方式，在无人机控制、应急救援及智能巡检等场景中展现出重要应用价值[1] [2]。相较于传统遥控方式，手势交互具有非接触性强、操作直观、交互灵活等优势，能够有效降低操作门槛并提升人机协同效率。然而，在无人机高空拍摄场景下，由于手势目标尺度较小、成像分辨率受限，同时伴随背景复杂多变、光照条件不稳定以及运动模糊等因素的影响，手势识别任务仍面临显著挑战，易出现误检与漏检问题[3]。近年来，基于深度学习的目标检测方法，尤其是 YOLO 系列算法，在手势识别领域取得了良好效果，并在实时性与精度之间实现了较好的平衡[4]。然而，现有方法大多针对通用场景设计，缺乏对无人机视角特性的针对性优化，直接应用于无人机手势识别时，仍存在特征提取不足、多尺度适应能力有限以及定位精度不高等问题，难以满足复杂实际环境下的应用需求[3] [5]。

针对无人机平台下手势目标尺寸较小、姿态变化复杂、背景干扰强以及尺度变化明显等问题，本文提出一种面向无人机手势识别任务的改进 YOLOv11s 算法。由于无人机在空中采集图像时存在视角变化大、目标分辨率低以及远距离拍摄等特点，传统 YOLOv11s 在手部细粒度特征提取、多尺度信息融合以及边界框定位精度方面仍存在一定局限性。为此，本文从特征表达、特征融合以及边界框回归三个方面对模型进行针对性优化。在 Backbone 阶段，引入 C2PSA_EDFFN 模块，以增强网络对手势区域细节信息的建模能力。其中，PSA 机制能够有效提升网络对关键区域的注意力响应，强化手部纹理与边缘特征表达；同时，EDFFN 结合频域特征建模与深度可分离卷积，不仅能够扩大特征感受范围，还能够有效捕获手势目标中的高频细节信息，从而提高复杂背景下小目标手势的特征辨识能力。因此，该模块更加适用于无人机视角下手势目标尺度小、细节信息弱的问题。在 Neck 阶段，设计 CGAFusion 特征融合模块，以提升不同层级语义信息与空间信息之间的交互能力。由于无人机手势目标在不同飞行高度和拍摄距离

下存在显著尺度变化，仅依赖传统特征融合方式难以充分兼顾浅层位置信息与深层语义信息。CGAFusion通过内容引导注意力机制对浅层细节特征与深层语义特征进行自适应权重分配，从而增强模型对多尺度手势目标的表达能力，提高复杂场景下的检测鲁棒性。此外，在边界框回归阶段，采用 PIoU2 损失函数替代传统 CIoU 损失函数。相比 CIoU，PIoU2 能够更加关注预测框与真实框之间的几何匹配关系，并有效缓解复杂场景下由目标遮挡、尺度变化以及姿态变化带来的回归不稳定问题，从而进一步提高模型定位精度与训练稳定性。

实验结果表明，所提出的改进 YOLOv11s 算法在复杂无人机场景下能够有效降低漏检率，并显著提升手势检测精度与鲁棒性，为无人机手势交互系统提供了一种高效且可靠的技术方案。

2. 算法模型架构

2.1. CEP-YOLOv11s 结构图

以 YOLOv11s 网络结构为基础，本文构建了一种面向无人机手势识别任务的改进模型 CEP-YOLOv11s，其整体结构如图 1 所示。针对无人机场景下手势目标尺度变化大、细粒度特征易丢失以及复杂背景干扰严重等问题，本文分别从主干网络、特征融合网络以及边界框回归策略三个方面对 YOLOv11s 进行改进，引入 C2PSA_EDFFN 模块、CGAFusion 模块以及 PIoU2 损失函数，以进一步提升模型的特征提取能力、多尺度特征融合能力以及目标定位精度[5] [6]。

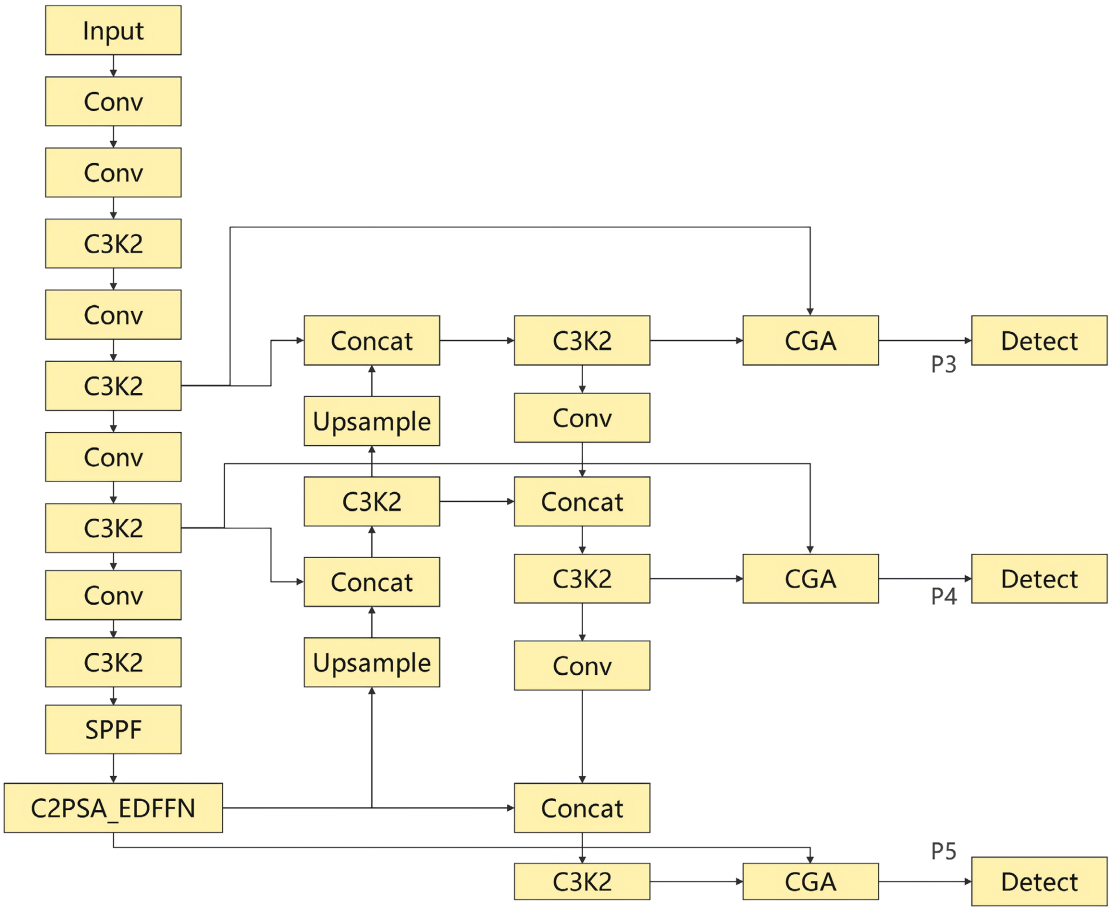


Figure 1. Network architecture of CEP-YOLOv11s
图 1. CEP-YOLOv11s 网络结构图

在主干网络部分，引入 C2PSA_EDFFN 模块对原有特征提取结构进行增强。该模块在 C2PSA 注意力机制基础上结合 EDFFN 频域前馈网络，通过空间域特征建模与频域特征筛选的协同作用，增强模型对手部局部纹理、边缘及高频细节信息的表达能力，从而提高复杂背景下小尺度手势目标的特征辨识能力[7]。在特征融合阶段，设计 CGAFusion 模块以优化多尺度特征融合过程。该模块基于内容引导注意力机制，对浅层空间细节信息与深层语义信息进行自适应权重分配，从而增强不同层级特征之间的信息交互能力，提高模型对不同尺度手势目标的检测鲁棒性[8]。此外，在边界框回归阶段，采用 PIoU2 损失函数替代传统 CIoU 损失函数。相比 CIoU，PIoU2 能够更加有效地建模预测框与真实框之间的空间几何关系，从而提升复杂场景下目标定位的准确性与训练稳定性[9]。

2.2. C2PSA_EDFFN 模块

针对无人机手势识别任务中小尺度手势目标难以检测以及手部局部细节特征提取不足的问题，本文在 YOLOv11s 主干网络的特征提取结构基础上，引入 C2PSA_EDFFN 模块，以增强模型对手势关键区域与细粒度特征的表达能力。其中，C2PSA 模块通过引入自注意力机制，能够有效增强网络对全局依赖关系与关键区域的建模能力[10]；在此基础上，进一步结合高效判别频域前馈网络(Efficient Discriminative Frequency Domain-Based Feedforward Network, EDFFN)，利用频域特征筛选机制强化对手部边缘纹理、高频细节等关键信息的提取能力，从而提升复杂背景下小目标手势的特征辨识能力。此外，该模块在增强特征表达能力的同时，兼顾了网络的轻量化需求，有利于无人机平台下实时手势识别任务的部署与应用。

C2PSA_EDFFN 模块以“局部特征提取 - 注意力增强 - 频域特征建模 - 残差融合”为核心流程，其整体结构如图 2 所示。首先，输入特征 XXX 经过 C2f 模块进行初步特征提取与通道信息融合，以增强局部空间特征表达能力；随后，利用 PSA (Partial Self-Attention)模块对特征进行注意力建模，通过建立特征之间的全局依赖关系，增强模型对手势关键区域的响应能力，从而提高复杂背景下的重要特征表征能力[10]。在此基础上，引入 EDFFN 模块对特征进行频域增强，通过空间域与频域协同建模进一步强化对手部边缘纹理、轮廓结构等细粒度信息的提取能力。完成频域特征建模后，模块通过 Residual Connection 对增强后的特征进行残差融合，以缓解深层网络训练过程中可能出现的梯度退化问题，并增强原始特征信息的保留能力。最后，将融合后的特征与输入分支通过残差加法(Element-Wise Addition)进行特征整合，输出最终增强特征 YYY。通过引入 PSA 与 EDFFN 的协同建模机制，C2PSA_EDFFN 模块在保持较低计算开销的同时，有效提升了模型对无人机场景下小目标手势细节特征与全局语义信息的联合表达能力。

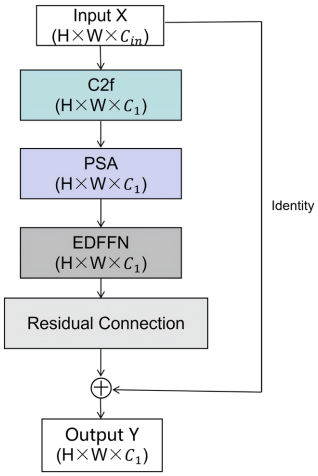


Figure 2. Architecture of the C2PSA_EDFFN module
图 2. C2PSA_EDFFN 模块结构图

EDFFN 模块

EDFFN 模块的核心思想是在前馈网络(FFN)结构中引入频域特征建模机制，并将频域操作置于特征降维阶段后执行，以降低频域变换带来的计算开销。相比传统方法直接在高维特征空间中进行傅里叶变换所产生的高计算复杂度与特征冗余问题，该设计在保持输入输出通道数一致的基础上，有效提升了特征利用效率并降低了模型计算量[11]。具体而言，EDFFN 充分利用 FFN“先升维进行特征扩展、后降维实现信息压缩”的结构特点，在更加紧凑的低维特征空间中开展频域特征建模，从而能够更加有效地增强模型对不同频率信息的表达能力，包括手部边缘纹理等细粒度特征以及整体轮廓结构等语义信息。通过空间域与频域信息的协同建模，EDFFN 进一步增强了模型对复杂背景下小目标手势细节特征的表达能力。

EDFFN 模块以“空间特征提取–频域特征筛选–特征重构”为核心流程。首先，利用 1 × 1 卷积对输入特征进行通道扩展，并结合 GELU 激活函数增强网络的非线性特征表达能力；随后，引入 3 × 3 深度可分离卷积(DWConv)对局部空间信息进行建模，以增强模型对手部边缘纹理等细粒度特征的感知能力；之后，通过 1 × 1 卷积将扩展后的特征压缩至与输入一致的通道维度，以获得更加紧凑的特征表示。在此基础上，采用快速傅里叶变换(FFT)将特征从空间域映射至频域，实现频域特征建模。随后，引入可学习的频域权重矩阵 WWW 对频谱特征进行自适应筛选，以强化对手势识别具有判别意义的频率成分，同时抑制冗余频域信息。最后，通过逆快速傅里叶变换(IFFT)将筛选后的频域特征恢复至空间域，并结合 LayerNorm 归一化操作与残差加法(Element-Wise Addition)实现特征融合，输出同时包含局部细节信息与全局结构信息的增强特征表示，从而提升模型在复杂无人机场景下对小目标手势的特征表达能力，其结构如图 3 所示。

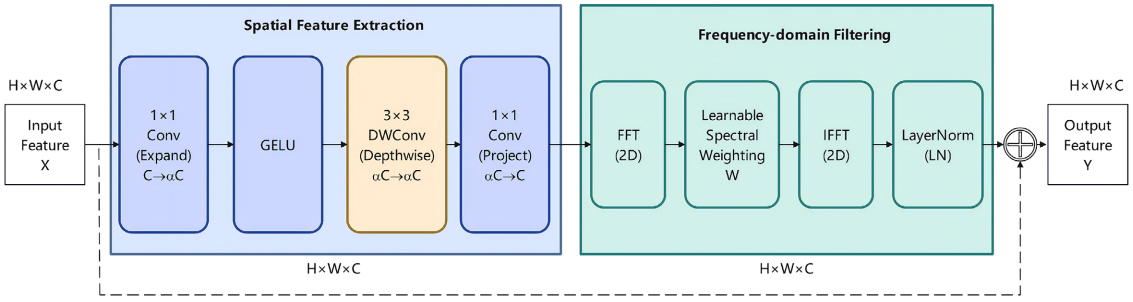


Figure 3. Architecture of the EDFFN module
图 3. EDFFN 模块结构图

EDFFN 的计算过程如式(1)所示：

$$y = \mathrm{FFN}(x) + \mathrm{DWConv}\left(\mathrm{FFN}(x)\right) + x \tag{1}$$

其中，输入特征 $x \in R^{B\times C\times H\times W}$，$B$ 为 batch size，C 为通道数，H、W 为特征图的高和宽，PConv 为 1 × 1 点卷积，GELU 为激活函数，DWConv 为深度可分离卷积，FFT 与 IFFT 分别为快速傅里叶变换及其逆变换，W 为可学习的频域量化矩阵，$\odot$ 为元素级乘法，Y 为最终输出。

通过引入 EDFFN 模块，C2PSA_EDFFN 在保持原有自注意力机制全局建模能力的同时，以较低的计算开销增强了局部频域特征的提取与筛选能力，有效提升了网络对手部关键细节的捕获能力，改善了小目标手势和姿态多变手势的检测效果。

2.3. CGAFusion 模块

CGAFusion 模块的核心思想是利用内容引导注意力机制实现不同层级特征的自适应融合，其结构如

图 4 所示。该模块的设计借鉴了 DEA-Net 中提出的 CGA 模块[12]。对于输入的浅层特征 F_s 与深层特征 F_d，模块首先分别通过 Query Projection 与 Key-Value Projection 对输入特征进行映射，以构建注意力计算所需的查询(Query)、键(Key)和值(Value)特征表示。其中，浅层特征主要包含丰富的空间细节信息，而深层特征则包含更强的语义表达能力。随后，在 Content-guided Attention 模块中，通过内容引导注意力机制对不同层级特征之间的相关性进行建模，从而增强模型对关键区域特征的关注能力[13]。完成注意力交互后，模块通过 Fusion 模块对浅层与深层特征进行拼接融合(Concatenation)，并进一步利用 1 × 1 卷积实现加权特征融合，以增强跨层级特征之间的信息交互能力，最终输出融合后的特征表示，结构图如图 4 所示。

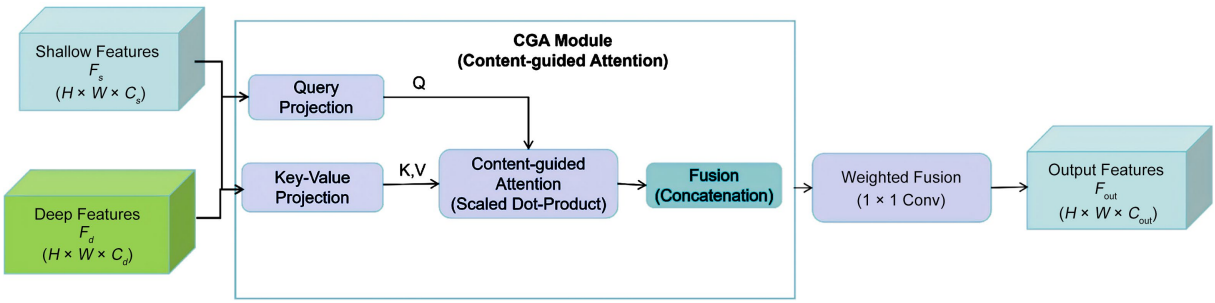


Figure 4. Architecture of CGAFusion
图 4. CGAFusion 结构图

CGAFusion 模块能够有效结合浅层特征中的空间细节信息与深层特征中的高层语义信息，从而提升模型对多尺度手势目标的特征表达能力。相比传统固定权重的特征融合方式，该模块通过内容引导注意力机制实现不同层级特征的动态自适应融合，使模型能够根据手势目标的尺度变化与特征分布自动调整特征响应权重，从而增强复杂无人机场景下的检测鲁棒性。此外，加权融合策略能够进一步缓解浅层特征与深层特征之间的语义差异问题，提高远距离小目标手势的检测性能。

2.4. PIoU2 损失函数

本文采用 PIoU2 (Pixels-IoU v2)损失函数替代 CIoU 损失。PIoU2 损失是一种基于像素级交并比的边界框回归损失函数，其核心思想是将边界框回归任务分解为中心点定位与形状回归两个子任务，分别计算损失后再进行加权融合[14]。

PIoU2 损失的计算公式如式(2)所示：

$$L_{PIoU2} = 1 - \frac{\sum_{(i,j)\in\Omega} 1_{(i,j)\in B\cap B^{gt}} \cdot w_{(i,j)}}{\sum_{(i,j)\in\Omega} 1_{(i,j)\in B\cup B^{gt}} \cdot w_{(i,j)}} + \frac{\lambda}{2}\left(\|\Delta c\|_2^2 + \|\Delta s\|_2^2\right) \tag{2}$$

其中：

B, B^{gt} 分别为预测框与真实框的区域；$1_{(\cdot)}$ 为指示函数，Ω 为图像像素空间；$w_{(i,j)}$ 为像素级权重，用于增强边界区域的重要性；中心点偏移量如式(3)所示；对数空间下的尺寸偏差如式(4)所示；λ 为平衡系数，用于协调像素级损失与几何约束。

$$\Delta c = c - c^{gt} \tag{3}$$

$$\Delta s = \left(\log w - \log w^{gt}, \log h - \log h^{gt}\right) \tag{4}$$

PIoU2 损失将边界框回归分解为三个独立分量：

像素级重叠损失如式(5)所示，该分量通过像素级加权的方式计算预测框与真实框的重叠程度。与传统 IoU 不同，PIoU2 为边界区域的像素分配更高的权重，使模型更加关注手势的边缘轮廓信息，从而提升边界框的贴合精度。

$$L_{pixel}=1-\frac{\sum_{(i,j)\in\Omega}1_{(i,j)\in B\cap B^{gt}}\cdot w_{(i,j)}}{\sum_{(i,j)\in\Omega}1_{(i,j)\in B\cup B^{gt}}\cdot w_{(i,j)}} \tag{5}$$

中心点定位损失如式(6)所示，采用欧氏距离的平方作为损失，直接约束预测框中心点与真实框中心点的对齐程度。该分量对小目标手势的定位尤其重要，能够有效减少中心点偏移带来的检测误差。

$$L_{center}=\left\|c-c^{gt}\right\|_2^2 \tag{6}$$

尺寸回归损失如式(7)所示，采用对数空间下的均方误差作为损失，能够对不同尺度的手势给出一致的梯度尺度，避免了大尺寸手势与小尺寸手势之间的损失不平衡问题。

$$L_{size}=\left(\log w-\log w^{gt}\right)^2+\left(\log h-\log h^{gt}\right)^2 \tag{7}$$

最终的 PIoU2 损失为三者加权和如式(8)所示，其中，μ 和 ν 为超参数，用于平衡不同损失分量的贡献。

$$L_{PIoU2}=L_{pixel}+\mu L_{center}+\nu L_{size} \tag{8}$$

3. 实验与结果分析

3.1. 数据集与预处理

HaGRID (Hand Gesture Recognition Image Dataset)数据集是由 SberDevices 发布的一个专为手势识别设计的大型数据集，该数据集以其丰富多样的手势类别、高质量的图像以及详尽的标注信息而著称[15]。HaGRID 数据集最初包含 18 种常见的通用手势，这些手势覆盖了日常生活中常见的动作。参考 Light-HaGRID 和 NUS-II 公开手势数据集中的手势形态与类别定义，从中筛选与本文 12 类手势形态相近的图像样本。其中，Light-HaGRID 数据集包含 18 类手势，NUS-II 数据集包含 10 类手势，两个数据集有 3 类手势重复。本文从上述数据集中筛选出符合条件的图像，并对筛选后的图像进行统一标注，将其类别映射至本文定义的 12 类手势标签体系中。为弥补公开数据集在无人机实际操控场景上的不足，本文进行了自采集数据补充。采集地点为校内 12 个不同地点，采集设备为相机，采集团队共 4 人，其中 1 人负责拍摄，3 人为手势实验对象。在 15 种不同场景下(涵盖不同距离、光照条件、背景环境、前景遮挡等情况)，对 3 名实验人员的手势进行同步采集。每种场景下每名实验人员依次比出全部 12 类手势，每张图像中同时包含 3 名实验人员的不同手势(共 3 个手势/张)。本文构建的无人机手势识别数据集图像共计 2477 张，涵盖了不同复杂背景的手势图像，每张图像中可能包含多个手势目标，将其手势形状分为 12 个类别，自建数据集来源如表 1 所示。

Table 1. Sources of the self-constructed dataset
表 1. 自建数据集来源

来源	图像数量	标签数量	占比(%)
Light-HaGRID	842	869	32.4
人工采集	1635	1709	67.6

3.2. 实验配置

本文在数据集上使用相同的实验环境和实验参数。实验环境配置如表 2 所示，操作系统为 Ubuntu20.04，GPU 为 NVIDIA GeForce RTX 3090，显存为 24 G，CPU 为 Intel(R) Core(TM) i7-11700KF，实验 Pytorch2.3.0，Cuda11.8，Cudnn8.9.0，Python3.9。实验参数表配置如表 3 所示，输入图像大小 image size 为 640，迭代次数 epoch 为 100，单次批处理 batch size 数量为 64，优化器 optimizer 为 SGD，初始学习率 Lr0 为 0.01。

Table 2. Experimental setup
表 2. 实验环境配置

配置名称	配置信息
操作系统	Ubuntu20.04
CPU	Intel(R) Core(TM) i7-11700KF
GPU	NVIDIA GeForce RTX 3090
显存	24 G
Pytorch	2.3.0
Cuda	11.8
Cudnn	8.9.0
Python	3.9

Table 3. Training parameter settings
表 3. 实验参数设置

参数	参数值
image size	640
epoch	100
batch size	64
optimizer	SGD
Lr0	0.01

3.3. 实验结果与分析

3.3.1. 消融实验分析

为验证各改进模块的有效性，本文基于 YOLOv11s 开展了 8 组消融实验，分别引入 C2PSA_EDFFN 模块、CGAFusion 模块及 PIoU2 损失函数，并对不同组合方案进行对比分析，实验结果如表 4 所示。结果表明，C2PSA_EDFFN 模块通过融合注意力机制与增强型前馈结构，有效提升了模型对多尺度及遮挡手势目标的特征表达能力；CGAFusion 模块通过自适应融合浅层细节与深层语义信息，缓解了尺度变化带来的特征错位问题；PIoU2 损失函数则通过引入更精细的回归约束，提高了边界框定位精度。整体来看，各模块均对模型性能产生积极贡献，验证了所提改进策略的有效性。

Table 4. Ablation study results
表 4. 消融实验结果

Model	C2PSA_EDFFN	CGAFusion	Piou2	P (%)	R (%)	mAP 50 (%)	mAP 50~95 (%)	Params (M)	GFLOPs/G
YOLOv11s				93.7	85.9	93.6	71.6	9.43	21.6

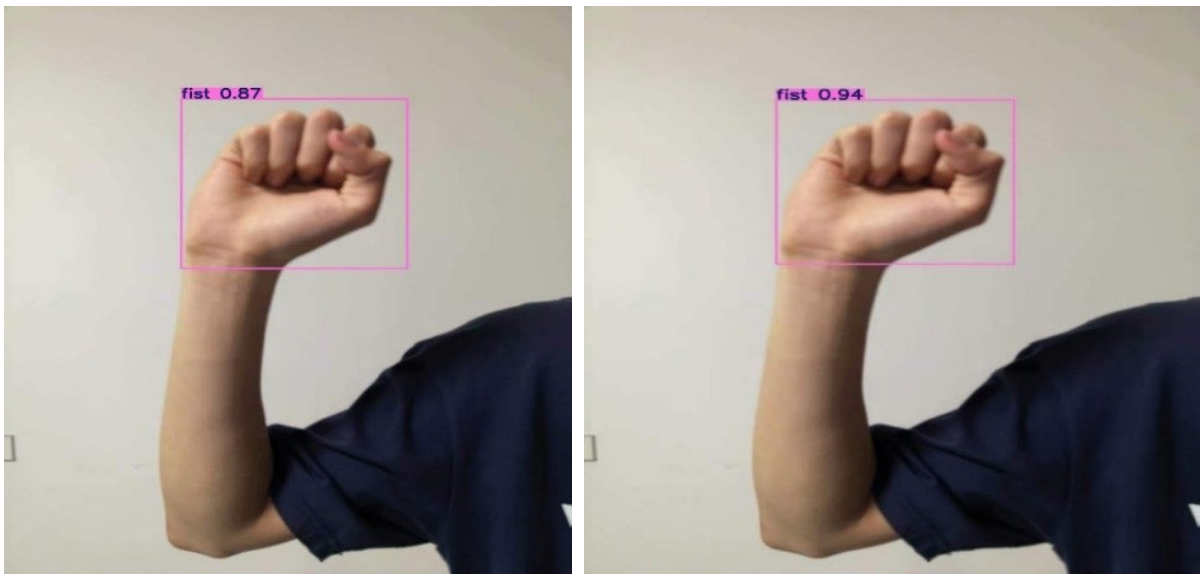
续表

YOLOv11s	√			93.6	90.6	94.3	72.2	9.58	21.7
YOLOv11s		√		93.0	87.1	94.2	71.4	10.81	29.6
YOLOv11s			√	92.6	89.7	94.4	72.9	9.43	21.6
YOLOv11s	√	√		92.3	91.0	94.7	72.8	10.95	29.7
YOLOv11s	√		√	93.2	92.0	95.0	72.3	9.58	21.7
YOLOv11s		√	√	91.5	91.5	95.4	73.1	10.95	29.7
CEP-YOLOv11s	√	√	√	93.5	93.6	96.1	75.5	10.95	29.7

消融实验结果表明，各改进模块均对模型性能产生显著提升。单模块方面，引入 C2PSA_EDFFN 后召回率和 mAP 50 分别提升 4.7%和 0.7%，提升最为明显；引入 CGAFusion 后召回率和 mAP 50 分别提升 1.2%和 0.6%；引入 PIoU2 损失函数后，在不增加模型复杂度的情况下，召回率和 mAP 50~95 分别提升 3.8%和 1.3%。其中，C2PSA_EDFFN 主要提升目标检出能力，PIoU2 在高 IoU 阈值下对定位精度提升更为显著。组合实验中，CGAFusion + PIoU2 表现最佳(mAP 50 95.4%，mAP 50~95 73.1%，召回率 91.5%)。在此基础上进一步引入 C2PSA_EDFFN (完整模型)，mAP 50 和 mAP 50~95 分别提升至 96.1%和 75.5%，较基线提升 2.5%和 3.9%，召回率提升至 93.6% (+7.7%)。尽管参数量与计算量略有增加，但在无人机手势识别任务中换取了更高的检测精度与召回性能。整体来看，三种改进在特征提取、特征融合与损失优化三个层面形成良好的互补与协同作用，显著提升了模型在复杂场景下的检测能力。

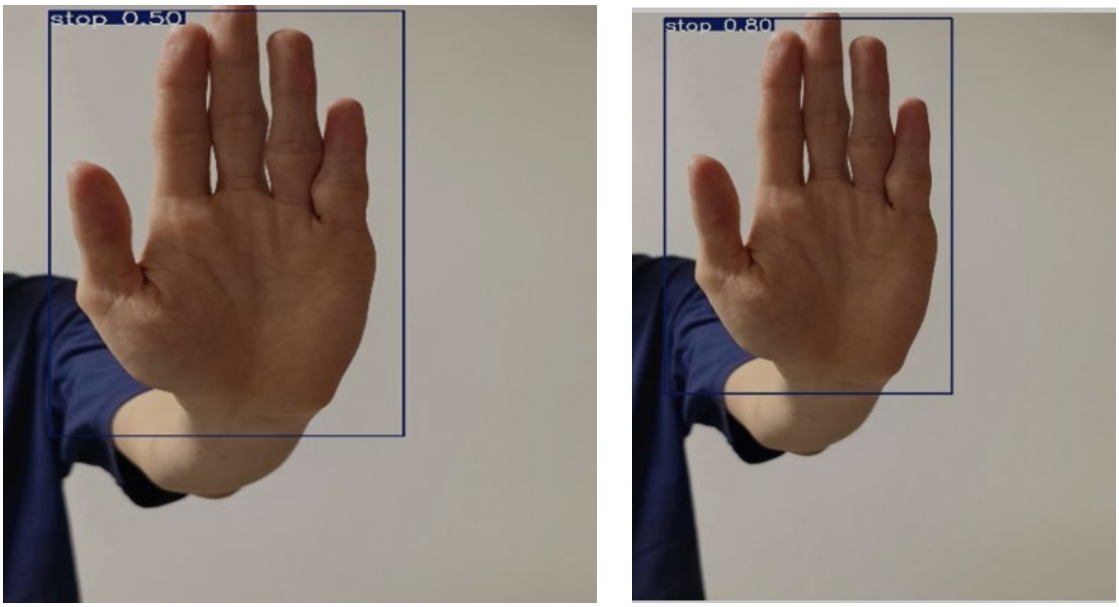
3.3.2. 模型效果对比

为了更直观展现 CEP-YOLOv11s 算法在实际应用中的性能，我们与 YOLOv11s 进行了效果对比，挑选了部分手势图像进行可视化对比，改进前后无人机手势识别对比如图 5、图 6 所示。



(a) YOLOv11s 模糊检测手势　　(b) CEP-YOLOv11s 模糊检测手势

Figure 5. Visual comparison of blurred gesture samples
图 5. 模糊手势对比图



(a) YOLOv11s 大手势　　(b) CEP-YOLOv11s 大手势

Figure 6. Comparison results for large-scale gestures
图 6. 大手势对比图

在图 5 中，可以看到该图整体比较模糊，无人机拍摄过程中这个场景比较容易出现。图中可以看出 YOLOv11s 算法进行手势识别检测时，检测手势的置信度只有 0.52，而 CEP-YOLOv11s 模型检测手势的置信度高达 0.81。这一对比充分说明，本文提出的 CEP-YOLOv11s 模型在无人机常见的运动模糊、低分辨率等退化场景下具有更强的鲁棒性。在图 6 中，可以看到该图手势目标较大，这是无人机飞行距离较近时常见的拍摄场景。图中显示，YOLOv11s 算法在进行手势识别时出现了两个检测结果：一个正确识别出手势，另一个为置信度 0.34 的错误检测框；而 CEP-YOLOv11s 模型仅输出一个检测结果，置信度高达 0.91。这一对比充分说明，在无人机近距离、大目标手势的场景下，本文提出的 CEP-YOLOv11s 模型能够有效抑制虚假检测，避免多框误检，展现出更强的识别可靠性与精度优势。

3.3.3. 对比实验分析

实验所用数据集为上述所构建的无人机手势识别数据集，分析实验结果时精度指标采用精确率(P)、召回率(R)、平均准确率(mAP 50)、模型参数量(Params)和模型推理的浮点数(GFLOPs)。为了更好地展现出本文改进模型的优势，与目标检测算法模型 YOLOv8s、YOLOv9s、YOLOv10s、YOLOv11s、YOLOv12s 和 RT-DETR-l 等进行对比实验。RT-DETR 是百度推出的基于 Transformer 架构的实时目标检测大模型，在检测精度和推理速度上实现了突破性进展，该模型采用创新的混合编码器架构，巧妙结合 CNN 的局部特征提取优势与 Transformer 的全局建模能力。模型统一使用本文自建复杂环境下手势数据集进行训练和验证，实验结果如表 5 所示。

Table 5. Comparative experimental results
表 5. 对比实验结果

Model	P (%)	R (%)	mAP 50 (%)	mAP 50~95 (%)	Params (M)	GFLOPs/G
YOLOv8s	89.2	81.5	86.3	68.4	9.84	23.6
YOLOv9s	88.7	82.9	85.7	67.9	6.3	22.8

续表

YOLOv10s	89.5	83.8	88.1	67.2	8.07	24.8
YOLOv11s	93.7	85.9	93.6	71.6	9.43	21.6
YOLOv12s	91.3	84.2	87.4	70.1	9.10	19.6
RT-DETR-l	90.1	86.0	84.3	66.5	32.8	108.1
CEP-YOLOv11s	93.5	93.6	96.1	75.5	10.95	29.7

从表 4 可以看出，CEP-YOLOv11s 在精确率、召回率、mAP 50 及 mAP 50~95 等指标上均优于对比算法，其中 mAP 50 和 mAP 50~95 分别达到 96.1%和 75.5%，较基线 YOLOv11s 提升 2.5%和 3.9%；召回率提升至 93.6%，较 YOLOv11s 提高 7.7%，显著降低了漏检率。与轻量化模型 YOLOv12s 相比，在参数量略有增加的情况下，mAP 50 和召回率分别提升 8.7%和 9.4%；同时，相较 YOLOv10s、YOLOv9s 及 YOLOv8s 等主流模型，mAP 50 分别提升 8.0%、10.4%和 9.8%。即便对比参数量更大的 RT-DETR-l，本文方法在参数减少约 2/3 的情况下，mAP 50 仍高出 11.8%，表现出良好的精度与效率优势。综合来看，CEP-YOLOv11s 通过优化特征提取与多尺度融合策略，有效提升了复杂无人机场景下手势检测的整体性能，尤其在召回率方面表现突出。

4. 结论

针对无人机手势识别任务中存在的特征提取不足、多尺度适应能力有限以及定位精度不高等问题，本文提出了一种改进的 CEP-YOLOv11s 算法。该方法通过在主干网络中引入 C2PSA_EDFFN 模块增强特征表达能力，在颈部网络中采用 CGAFusion 模块优化多尺度特征融合策略，并在损失函数中引入 PIoU2 以提升边界框回归精度。实验结果表明，所提方法在自建无人机手势数据集上表现优异，在精确率、召回率及 mAP 等指标上均优于多种主流目标检测模型。其中，mAP 50 达到 96.1%，mAP 50~95 达到 75.5%，召回率提升至 93.6%，显著降低了漏检率，验证了各改进模块的有效性及其协同作用。

参考文献

[1] Zhang, B., Zhang, H., Zhen, T., Ji, B., Xie, L., Yan, Y., *et al*. (2024) A Two-Stage Real-Time Gesture Recognition Framework for UAV Control. *IEEE Sensors Journal*, **24**, 24770-24782. https://doi.org/10.1109/jsen.2024.3413787

[2] Fang, W., Lai, G., Yu, Y., Shi, C., Meng, X., Sun, J., *et al*. (2025) Real-Time Human-Drone Interaction via Active Multimodal Gesture Recognition under Limited Field of View in Indoor Environments. *IEEE Robotics and Automation Letters*, **10**, 11705-11712. https://doi.org/10.1109/lra.2025.3615031

[3] Li, S., Wang, Z., Liang, J. and Wang, Y. (2025) Small Object Detection in UAV Scenarios Based on YOLOv5. *Computer Modeling in Engineering & Sciences*, **145**, 3993-4011. https://doi.org/10.32604/cmes.2025.073896

[4] Guo, N., Ding, Y. and Meng, H. (2025) YOLOv8-GR: Real Time Gesture Recognition by Improving of YOLOv8 Algorithm. *Complex & Intelligent Systems*, **11**, Article No. 479. https://doi.org/10.1007/s40747-025-02107-0

[5] Liang, X., Xiang, J., Qin, S., Xiao, Y., Chen, L., Zou, D., *et al*. (2025) Small Target Detection Algorithm Based on Sahi-Improved-Yolov8 for UAV Imagery: A Case Study of Tree Pit Detection. *Smart Agricultural Technology*, **12**, Article 101181. https://doi.org/10.1016/j.atech.2025.101181

[6] Zhang, B., Li, J., Wang, H., *et al*. (2025) Object Detection Model of Vehicle-Road Cooperative Autonomous Driving Based on Improved YOLO11 Algorithm. *PLOS ONE*, **20**, e0329628.

[7] Wang, C., Zhang, K., Ma, J., Chen, Z. and Yu, Y. (2025) Improved YOLOv11n-Based Small Object Detection Algorithm from UAV Perspective. 2025 4*th International Symposium on Computer Applications and Information Technology* (*ISCAIT*), Xi'an, China, 21-23 March 2025, 900-908. https://doi.org/10.1109/iscait64916.2025.11010315

[8] Zheng, L., Wang, Z., Chen, Y., *et al*. (2025) Efficient Discriminative Frequency Domain-Based Feedforward Network for Image Restoration. *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), Nashville, 11-15

June 2025, 25641-25650.

[9] Chen, J., Liu, Y., Zhang, S., *et al*. (2024) Powerful-IoU and PIoU v2: Advanced Boundary Box Regression Losses for Object Detection. *Neural Networks*, **179**, Article ID: 106542.

[10] Zhao, Q. and Zhu, J. (2025) An Improved YOLOv11 Architecture with Multi-Scale Attention and Spatial Fusion for Fine-Grained Residual Detection. *Results in Engineering*, **27**, Article 107061. https://doi.org/10.1016/j.rineng.2025.107061

[11] Kong, L., Dong, J., Ge, J., Li, M. and Pan, J. (2023) Efficient Frequency Domain-Based Transformers for High-Quality Image Deblurring. 2023 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), Vancouver, 17-24 June 2023, 5886-5895. https://doi.org/10.1109/cvpr52729.2023.00570

[12] Chen, Z., He, Z. and Lu, Z.M. (2024) DEA-Net: Single Image Dehazing Based on Detail-Enhanced Convolution and Content-Guided Attention. *IEEE Transactions on Image Processing*, **33**, 1002-1015. https://doi.org/10.1109/tip.2024.3354108

[13] Li, X., Wang, Z., Chen, Y., *et al*. (2025) A Drone Detection Network Based on Multi-Scale Feature Fusion and Content-Guided Attention Mechanism. 2025 8*th International Conference on Artificial Intelligence and Big Data* (*ICAIBD*), Chengdu, 23-26 May 2025, 1-6.

[14] Liu, C., Wang, K., Li, Q., Zhao, F., Zhao, K. and Ma, H. (2024) Powerful-IoU: More Straightforward and Faster Bounding Box Regression Loss with a Nonmonotonic Focusing Mechanism. *Neural Networks*, **170**, 276-284. https://doi.org/10.1016/j.neunet.2023.11.041

[15] Kapitanov, A., Kvanchiani, K., Nagaev, A., *et al*. (2024) HaGRID—HAnd Gesture Recognition Image Dataset. *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision* (*WACV*), Waikoloa, 3-8 January 2024, 4572-4581.