DOI:10.19651/j.cnki.emt.2519163

改进 YOLOv11 的模糊监控异常检测算法*

刘玉蕾 褚丽莉 李 波
(辽宁工业大学电子与信息工程学院 锦州 121000)

摘 要:针对模糊监控与复杂路况导致的异常行为检测精度不足的问题,本文提出一种多模块协同优化的 YOLOv11 改进模型。首先,采用 Dynamic Sample 替代颈部网络的传统上采样,提升目标定位与识别精度;其次,在骨干网络末层集成重新设计的多窗口注意力机制,增强模糊视频中异常特征的捕捉能力并抑制噪声干扰;最后,引入轻量型网络 ShuffleNetV2 作为主干网络,在保持特征表达能力的同时,将模型参数量显著降低。实验结果表明,在 UCF101 和 UCF Crime 数据集上,通过引入 Dynamic Sample 模块与多窗口注意力机制,本文模型比原始 YOLOv11 模型的 mAP50、mAP50-95 分别提高 8.5%、13.1%,有效减少漏判与误判现象;结合轻量化 ShuffleNetV2,成功将模型参数量从 2.58 M 压缩至 0.82 M。综合结果显示,改进后的 YOLOv11 模型能够更好地满足交通监控等实时场景需求,兼顾检测效率与准确性,具备广泛的应用潜力。

关键词:YOLOv11;异常行为检测;窗口注意力;ShuffleNetV2;轻量化
中图分类号:TN249.7;TP391.41 **文献标识码**:A **国家标准学科分类代码**:510.4010

Optimized YOLOv11 for blurry surveillance anomaly detection

Liu Yulei Chu Lili Li Bo
(School of Electronic and Information Engineering, Liaoning University of Technology, Jinzhou 121000, China)

Abstract: To address the low accuracy in anomaly behavior detection caused by blurry surveillance and complex road conditions, this paper proposes an optimized YOLOv11 model with multi-module collaboration. First, Dynamic Sample replaces traditional upsampling in the neck network to enhance target localization and recognition precision. Second, a redesigned Multi-Window Attention module is integrated into the final layer of the backbone network, improving the capture of anomaly features in blurry videos while suppressing noise interference. Finally, the lightweight ShuffleNetV2 is adopted as the backbone, significantly reducing model parameters while preserving feature representation capability. Through the introduction of Dynamic Sample module and Multi-Window Attention module, experimental results on the UCF101 and UCF Crime datasets demonstrate that our model improves mAP50 and mAP50-95 by 8.5% and 13.1%, respectively, compared to the original YOLOv11, effectively mitigating false negatives and false positives. By combining ShuffleNetV2, the model's parameter count is reduced from 2.58 M to 0.82 M. Overall, the optimized YOLOv11 model better meets the demands of real-time scenarios such as traffic surveillance, balancing detection efficiency and accuracy with broad application potential.

Keywords: YOLOv11;anomalous behavior detection;window attention;ShuffleNetV2;lightweight

0 引 言

随着城市化进程的加速和人们安全意识提升,视频监控系统在各个领域得到了广泛应用,如城市安防、智能家居等。但传统监控依赖人工判读或简单算法,存在诸多局限性。人工监控效率低下,人们易产生疲劳和疏忽,难以应对海量的视频数据[1]。简单的运动检测算法仅能识别简单的运动变化,无法准确理解和分析复杂的行为模式,如异常行为检测、人员身份识别等。

人们对异常行为研究较早,传统方法中最初采用光流法检测视频中的行为,但光流法仅利用灰度梯度表征特征,无法区分旋转与平移变换带来的相似变化。多人 2D 姿态估计算法通过检测人体关键点(如关节位置)及其连接关系[2],提高了行为识别的准确率,但此算法侧重于精准的运

收稿日期:2025-06-22
*基金项目:辽宁省教育厅高等学校基本科研项目(JYTMS20230862)资助

动学分析(如体操评分)。CNN 通过深层语义特征捕捉的能力,适用于开放环境下的复杂时间推理(如城市安防),但其性能在模糊监控场景中仍受限于特征丢失与噪声干扰[3]。与此同时,Transformer 架构凭借自注意力机制在自然语言处理领域取得成功,并逐渐扩展至计算机视觉领域。Vision Transformer 将图像分割为块序列,通过跨补丁交互捕捉远程空间上下文,直接用在视频时会遗漏动态变化的时序信息;时空 Transformer 同时处理空间维度和时间维度,但训练数据的时序标注成本较高;Swin Transformer[4]通过窗口注意力机制有效捕捉时序依赖与空间特征,支持实时推理且满足安防监控的低延迟需求。尽管现有目标检测算法对异常行为检测已取得较好成果,但直接迁移至监控任务仍面临两大挑战:其一,监控视频中异常行为主体占比小(通常小于画面面积的 5%);其二,复杂场景(如交通路口、工业园区)中存在大量背景噪声,导致误检率高达 40%以上[5-6]。YOLOv11 在 Darknet-19 骨干网络基础上引入深度可分离卷积,在保持实时性的同时将 COCO 数据集上的 mAP 显著提升。YOLOv11 通过改进的多尺度特征融合策略增强对微小异常行为的感知能力,适用于检测小目标占比高、背景干扰复杂的监控场景。

综上,本文以 YOLOv11 为基线模型,提出 3 项改进策略:1)引入 Swin Transformer 内置的多窗口注意力机制,可以增强模糊视频中关键行为特征的提取能力,减少噪声和无关信息的干扰。2)以动态上采样器(dynamic sample,DySample)替代传统上采样层,DySample 通过点重采样的方法避免耗时的动态卷积和额外的子网络生成动态内核[7],显著减少信息丢失和噪声引入。3)采用轻量型网络 ShuffleNetV2 作为模型主干,在保持特征表达能力的同时,将模型参数量和计算量显著降低。ShuffleNetV2 的自适应分组卷积可根据输入数据动态调整计算量,在监控场景下自动提升特征提取精度,进一步增强模型适应性。理论分析表明,上述改进通过多模块协同优化(特征提取-上采样-主干网络)可有效平衡检测精度与计算效率。

1 YOLOv11 网络及其改进模型

YOLOv11 是最新一代目标检测算法,它延续 Backbone-Neck-Head 三段式架构,采用 C3k2 动态卷积模块和 C2PSA 注意力融合模块,增强了特征提取能力[8]。C3k2 模块支持并行卷积设计和灵活参数配置,C2PSA 模块通过多头注意力机制提升特征提取效果。此外,YOLOv11 在检测头部分引入深度可分离卷积,减少参数量和计算量,提升推理速度。

针对监控安防的实际需求,本文提出一种基于改进 YOLOv11 的视频异常行为分析模型,它在主干网络(backbone)和颈部(neck)方面进行了显著的改进。首先,采用 DySample 取代颈部网络的传统上采样,提高 YOLOv11 模型在目标检测任务的性能、更精准地定位和识别目标物体;其次,将 Swin Transformer 模型的多窗口注意力机制重新设计并融合到本文模型骨干网络的末层中,有助于在模糊的监控视频中更准确地捕捉人体行为的关键特征,减少噪声和无关信息的干扰;最后,采用轻量型网络 ShuffleNetV2 作为模型主干,在保持特征表达能力的同时,将模型参数量和计算量显著降低。具体网络结构如图 1 所示。

2 改进 YOLOv11 网络结构

2.1 DySample 动态上采样

YOLOv11 模型采用的传统 UpSample 上采样方法[9],其方法存在信息丢失与噪声引入问题,尤其在监控视角下的模糊及复杂背景场景中,难以保留特征的细节与语义信息,导致目标检测漏检率与误检率显著上升[10]。

为解决上述问题,本文引入一种超轻量化且高效的动态上采样模块 DySample。DySample 采用基于点采样的策略[11],无需额外的高分辨率特征输入,通过双线性初始化生成均匀分布的初始采样点,并结合偏移量机制动态调整采样位置,实现位置感知的上采样[12],其上采样过程如图 2 所示。

给定一个 $C\times H\times W$ 的特征图 $\boldsymbol{X}$ 和一个 $2g\times sH\times sW$ 的点采样集 $\boldsymbol{S}$,其中 $2g$ 表示 x 和 y 坐标,$grid\_sample$ 函数使点采样集 $\boldsymbol{S}$ 中的位置对 $\boldsymbol{X}$ 重新采样,生成大小为 $C\times sH\times sW$ 的特征图 $\boldsymbol{X}'$,这一上采样过程如式(1)所示。

$$\boldsymbol{X}'=grid\_sample(\boldsymbol{X},\boldsymbol{S}) \tag{1}$$

式中:$\boldsymbol{X}$ 为输入特征,$\boldsymbol{X}'$ 为上采样特征,$\boldsymbol{S}$ 为采样集。

采样点生成器生成采样集 $\boldsymbol{S}$[13],给定上采样比例因子 s 和形状大小为 $C\times H\times W$ 的特征图 $\boldsymbol{X}$[14],使用输入和输出通道数为 C 和 $2gs^2$ 的线性层来生成大小为 $2gs^2\times H\times W$ 的偏移量 $\boldsymbol{O}$,然后通过像素重组将其重塑为 $2g\times sH\times sW$。

采样集 $\boldsymbol{S}$ 就是偏移量 $\boldsymbol{O}$ 与原始网格采样 $\boldsymbol{G}$ 之和,运算过程如式(2)所示。

$$\boldsymbol{S}=\boldsymbol{G}+\boldsymbol{O} \tag{2}$$

DySample 动态上采样模块使用点采样集 $\boldsymbol{S}$(如图 3 所示),为上采样计算提供动态权重插值,实现了更精确的特征图重建[15],尤其适合监控视角模糊且背景复杂的检测任务。

2.2 多窗口注意力机制 MWA

多窗口注意力机制(multi-window attention,MWA)通过窗口划分来限制注意力计算的范围,降低计算复杂度。该模块的设计灵感源自 Swin Transformer 模型,该模型采用了窗口多头自注意力(W-MSA)和移位窗口多头自注意力(SW-MSA)实现局部与全局特征的平衡。多窗口注意力机制的核心在于构建分层化的特征感知体系。其技术架构包含两个关键组件:W-MSA 和 SW-MSA,二者协同实现局部特征聚合与全局信息融合。

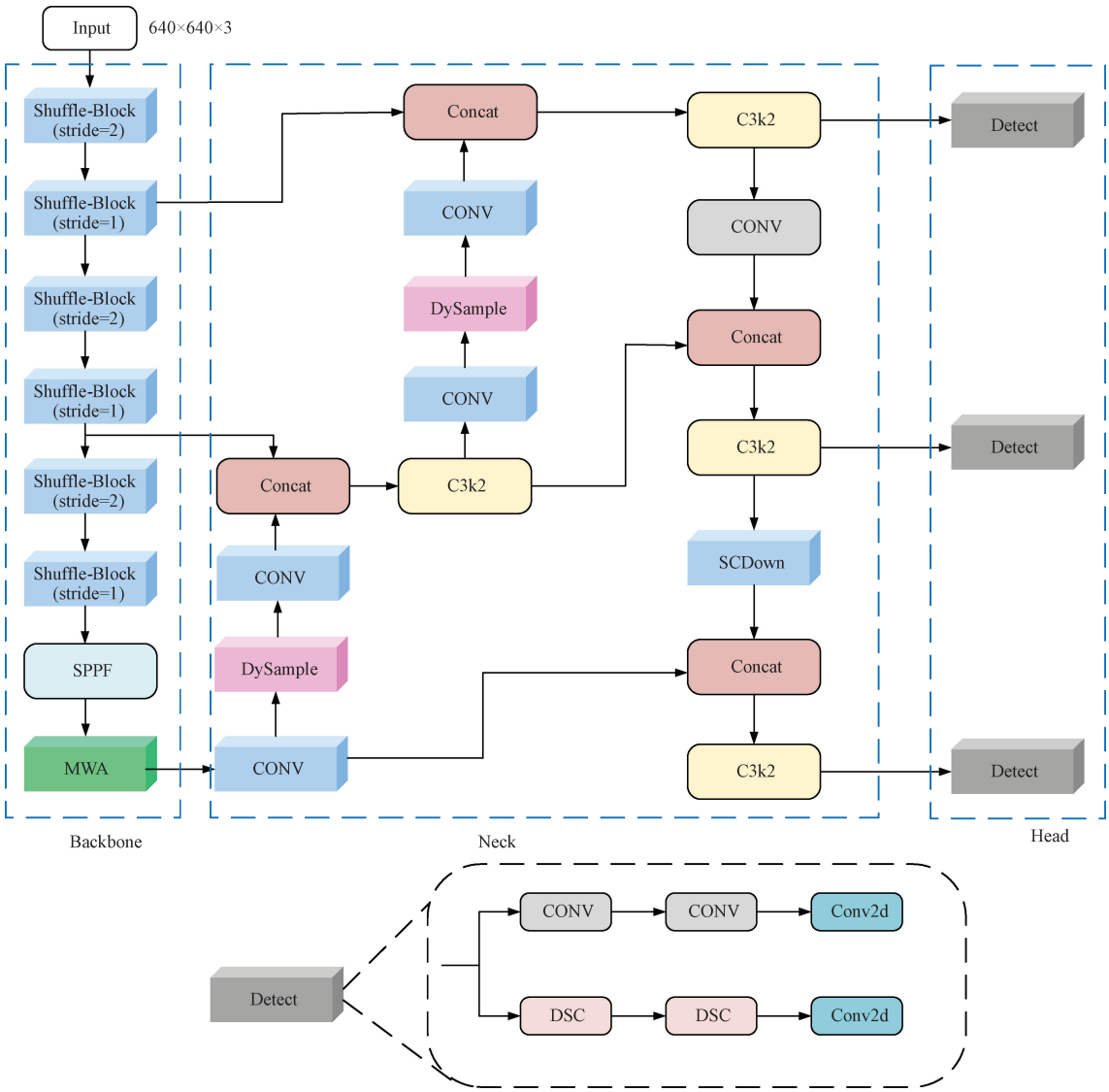


图 1　改进 YOLOv11 网络结构

Fig. 1　Improvement of YOLOv11 network structure

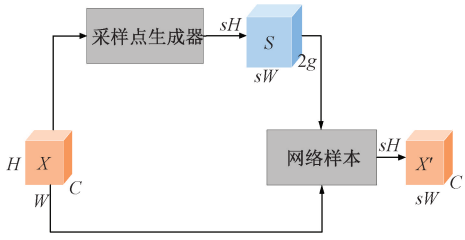


图 2　DySample 结构

Fig. 2　DySample structure

W-MSA 模块采用非重叠窗口划分策略[16]，将输入特征图分割为固定大小的局部窗口（如 7×7）。每个窗口内独立执行自注意力计算，计算公式为：

$$Attention(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V}) = Softmax\left(\frac{\boldsymbol{Q}\boldsymbol{K}^{\mathrm{T}}}{\sqrt{\boldsymbol{d}_k}}+\boldsymbol{B}\right)\boldsymbol{V} \tag{3}$$

式中：$\boldsymbol{B}$ 为相对位置编码矩阵，d_k 为查询向量维度。此设计使计算复杂度从全局注意力的 $O(N^2)$ 降至 $O(N\times M^2)$（M 为窗口尺寸）。$\boldsymbol{Q}$、$\boldsymbol{K}$、$\boldsymbol{V}$ 分别代表查询（Query）、键（Key）、值（Value）向量。

SW-MSA 模块通过引入窗口位移操作实现跨窗口信息交互[17]，增强模型的感受野。具体实现中，窗口在水平和垂直方向各移动 $M/2$ 像素，生成重叠窗口。

MWA 模块的结构如图 4 所示，该模块融合了多层感知机（MLP）、归一化层（Norm）、线性注意力机制（Linear Attention）以及卷积层（Conv）。MLP 是对输入数据进行非线性变换，提取更高级的特征表示[18]。Norm 有助于加速模型训练，提高训练稳定性，防止梯度消失或爆炸问题。Linear Attention 通过线性变换来计算注意力权重[19]。

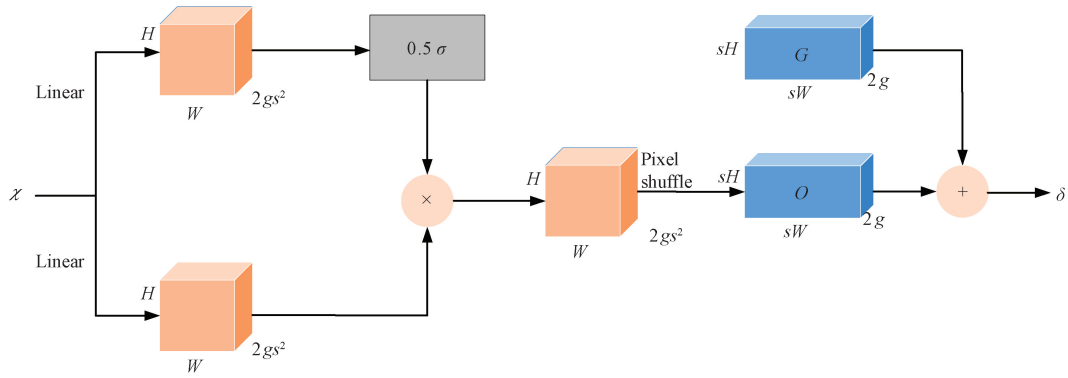


图 3 DySample 模块点采样集的计算

Fig. 3 Calculation of the point sampling set of the DySample module

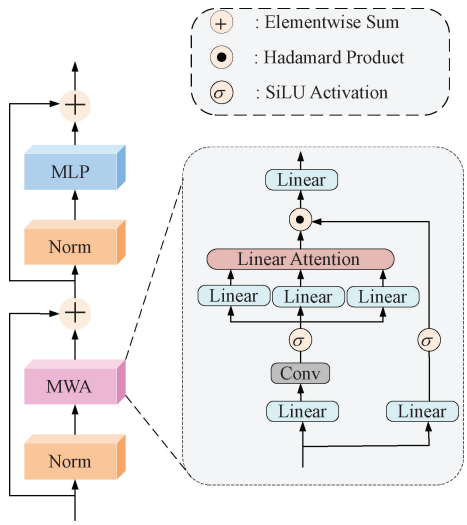


图 4 多窗口注意力结构

Fig. 4 Multi-window attention structure

MWA 模块的窗口划分策略是基于局部性先验假设:异常行为特征在视频帧中通常呈现空间连续性。通过非重叠窗口(W-MSA)与移位窗口(SW-MSA)的交替堆叠,实现分层感知。底层 7×7 的窗口捕获肢体动作、局部姿态等细粒度特征。通过 $M/2$ 像素位移($M=7$ 时位移 3 像素),构建重叠感知区域,解决非重叠窗口的边界信息割裂问题。窗口尺寸采用 7×7,使复杂度控制在 $O(49N)$,显著优于全局注意力(N 为特征图像素数)。在 UCF101 数据集的消融实验验证窗口尺寸会显著影响 mAP50-95 和 GFLOPs:5×5 窗口的 mAP50-95 值为 61,GFLOPs 为 4.7;7×7 窗口的 mAP50-95 值为 65.1,GFLOPs 为 6.2;9×9 窗口的 mAP50-95 值为 63.3,GFLOPs 为 7.7。当尺寸从 5×5 升级到 7×7 时,模型获得最大增益(mAP50-95 + 4.1),计算量增加 31.9%;进一步扩大到 9×9 时则导致精度下降 1.8,计算量增加 24.2%,所以 7×7 为本文的最优平衡点。

2.3 轻量型网络 ShuffleNetV2

ShuffleNet 的核心创新在于点群卷积和通道混洗:使用了新的操作点群卷积(GConv)和通道混洗(Channel Shuffle)的协同设计[20],通过降低计算复杂度与优化特征交互,实现轻量化网络架构[21]。ShuffleNet 通过使用组卷积降低参数量及使用 Channel Shuffle 实现不同组之间的信息交流,进而对残差网络结构(ResNet)进行改进。

ShuffleNetV2 网络进一步优化结构,引入通道分割(Channel Split)操作,并移除分组卷积,从而最小化 MAC(存储模型空间)[22]。其基本模块包含两种核心单元,如图 5 所示。图 5(a)展示了一个标准的瓶颈单元,使用了深度可分离卷积(DWConv)来融合特征。在此基础上,引入了 GConv 和 Channel Shuffle 操作,以增强特征的表达能力。图 5(b)展示了适用于空间下采样的 ShuffleNet 单元,使用步长为 2 的平均池化(AVG Pool)、DWConv 和 GConv 操作减少模型计算量从而提高模型的效率。通过连接操作(Concat)来合并特征[23],最后通过 Channel Shuffle 操作增强特征的多样性,有助于提升模型的性能。这一调整使得特征图的空间大小减半,通道数翻倍,显著降低了计算复杂度与参数量。

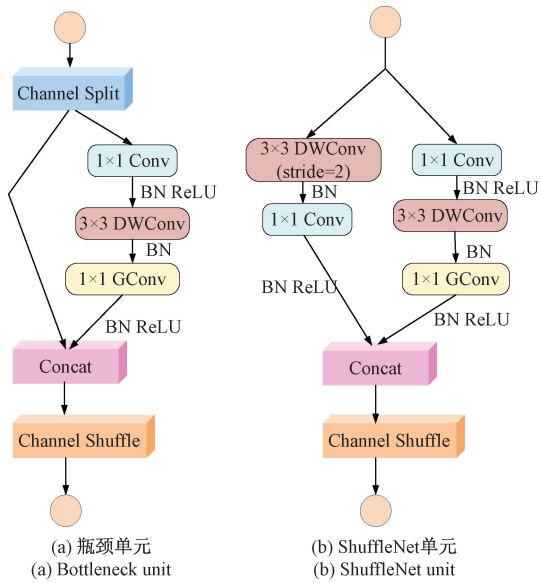


图 5 ShuffleNetV2 网络结构

Fig. 5 ShuffleNetV2 network architecture

3　实验结果与分析

3.1　实验数据集

UCF101 数据集是动作识别领域的景点基准，包含 13 320 个视频片段，涵盖 101 种人类活动，且每类有相似背景或人物、并具有光照变化及遮挡等特性，适用于评估模型在复杂场景下的泛化能力。UCF Crime 数据集专注于监控视频中的异常行为检测[24]。该数据集包含 1 900 个视频片段，涵盖 13 类异常事件（如抢劫、盗窃等），视频源自真实监控场景，存在复杂背景、昼夜光线变化及低分辨率等挑战，是评估模型在低质量监控视频中鲁棒性的重要基准。

针对公路场景的异常情况，本文定义了 7 类人类异常行为（爆炸、打架斗殴、逮捕、盗窃抢劫、故意破坏财物、交通事故、纵火）。协助 Xanylabeling 自动标注工具进行初步标注后、再进行人工审核并修正标注结果，从而确保数据标注的准确性。在 UCF101 数据集和 UCF Crime 数据集上累计标注图片 11 712 张，具体分布如表 1 所示。本文的消融实验和对比实验均基于该公路场景异常行为数据集进行实验验证。

表 1　公路场景的异常行为数据集

Table 1　A dataset of all anomalous behaviors for a highway scene

原标签	新标签	对应数量
Explosion	爆炸	1 131
Abuse＋Assault＋Fighting	打架斗殴	1 793
Arrest	逮捕	848
Burglary＋Robbery	盗窃抢劫	1 495
Vandalism	故意破坏财物	1 402
RoadAccidents	交通事故	4 130
Arson	纵火	913
总计数量	—	11 712

3.2　实验环境和参数

本研究以 YOLOv11 作为主要基准模型，实验环境配置如表 2 所示。训练阶段对比了 SGD 与 Adam 优化器，具体参数设置包括：初始学习率为 0.01，动量设为 0.937，权重衰减系数设为 0.000 5，批大小（batch size）设置为 8，并运行 100 个 epochs。输入图像分辨率统一为 640×640，数据增强操作采用默认配置。为确保实验的一致性，所有消融实验与对比实验皆基于上述设置，且未使用预训练权重，其他参数如表 2 所示。

表 2　实验硬件环境

Table 2　Experimental hardware environment

参数	实验环境
CPU	AMD Radeon(TM) Graphics
GPU	NVIDIA GeForce RTX4060 Laptop GPU
CUDA	11.8
Python	3.9.19
PyTorch	2.0.0＋cu118

3.3　评价指标

本文采用 6 个维度综合评估模型的检测性能：精确率（precision，P）、召回率（recall，R）、F1 分数（F1）、平均精确率均值（mAP）、模型参数量（Params）、浮点数运算量（FLOPs）以及训练时长（time）。具体公式如下：

$$P=\frac{\boldsymbol{TP}}{\boldsymbol{TP}+\boldsymbol{FP}} \tag{4}$$

$$R=\frac{\boldsymbol{TP}}{\boldsymbol{TP}+\boldsymbol{FN}} \tag{5}$$

$$F1=\frac{2\cdot P\cdot R}{P+R} \tag{6}$$

$$AP=\int_0^1 P(r)\mathrm{d}r \tag{7}$$

$$mAP=\frac{1}{n}\sum_{i=1}^{n}AP_i \tag{8}$$

式中：$\boldsymbol{T}$ 和 $\boldsymbol{F}$ 代表样本分类是否正确，$\boldsymbol{P}$ 和 $\boldsymbol{N}$ 代表样本被预测为正样本或负样本。n 代表目标类别数，$P(r)$ 表示基于 P 和 R 的曲线。

3.4　消融实验

为了验证本文算法的有效性，采用 UCF101 数据集和 UCF Crime 数据集对基线模型进行训练，并通过模块组合的方式开展消融实验。实验结果如表 3 和 4 所示。

通过表 3 可以看出：1）依次引入 CARAFE 和 DySample 模块改进上采样，mAP50 相比基线模型分别提高了 1.1%、1.4%，mAP50-95 分别提高了 0.2%、2.1%。DySample 模块在检测性能和训练耗时上表现更优。2）将 Swintransformer 模型、ShuffleNetV2 模块替换到模型的骨干网络中，mAP50 分别提高了 5.4%、1.1%，mAP50-95 分别提高了 7.4%、－1.5%，可见 ShuffleNetV2 模块显著降低参数量（参数量从 2.58 M 降低到 0.93 M），但 ShuffleNetV2 模块的通道分割和 MAC 最小化设计会导致 mAP50-95 下降 1.5%。3）在骨干网络末层加入 MWA 模块后，mAP50 和 mAP50-95 均提高了 3.8%，验证了注意力机制对检测精度的提升作用。4）同时引入 ShuffleNetV2 和 MWA 模块后，mAP50 提高了 1.4%，mAP50-95 提高了 2%，参数量从 2.58 M 降低到 0.81 M，表明轻量化与精度增强模块具有良好兼容性。5）引入 ShuffleNetV2、MWA 及 DySample 模块（即本文算法）后，mAP50 提高了 8.5%，mAP50-95 提高了 13.1%，参数量从 2.58 M 降低到 0.82 M。实验表明，本文算法在模型复杂度、参数量和检测精度取得了良好平衡，验证了其在公路场景异常行为检测中的有效性。

表 3 消融实验(优化器为 SGD,训练 100 epoch)

Table 3 Ablation experiment (optimizer is SGD, train 100 epochs)

实验	P/%	R/%	F1/%	mAP50/%	mAP 50-95)/%	Parameters/M	GFLOPs=FLOPs/G	训练时间/h
yolov11	86.2	80.6	83.3	88.3	61.3	2.583 517	6.3	4.103
yolov11+CARAFE	87.4	81.6	84.4	89.4	61.5	2.723 621	6.6	6.111
yolov11+DySample	89.0	82.1	85.4	89.7	63.4	2.595 869	6.3	4.187
yolov11+Swintransformer	92.9	86.5	89.6	93.7	68.7	2.506 018	6.0	5.472
yolov11+Swintransformer+DySample	94.2	91.3	92.7	96.1	71.3	2.518 370	6.0	5.798
yolov11+ShuffleNetV2	85.2	82.3	83.7	89.4	59.8	0.926 229	3.0	5.019
yolov11+MWA	87.9	86.1	87.0	92.1	65.1	2.469 341	6.2	8.513
yolov11+ShuffleNetV2+MWA	89.0	82.2	85.5	89.7	63.3	0.812 053	2.9	4.847
本文算法	94.1	93.1	93.6	96.8	74.4	0.816 213	2.9	5.265

表 4 消融实验(优化器为 Adam,训练 100 epoch)

Table 4 Ablation experiment (optimizer is Adam, train 100 epochs)

实验	P/%	R/%	F1/%	mAP50/%	mAP 50-95)/%	Parameters/M	GFLOPs=FLOPs/G	训练时间/h
yolov11	84.6	79.4	81.9	87.4	59.3	2.583 517	6.3	5.439
yolov11+CARAFE	86.0	77.7	81.6	87.2	58.7	2.723 621	6.6	5.395
yolov11+DySample	88.8	82.0	85.3	91.4	63.1	2.595 869	6.3	4.219
yolov11+Swintransformer	91.7	83.7	87.5	91.7	64.6	2.506 018	6.0	7.227
yolov11+Swintransformer+DySample	91.5	87.7	89.6	94.2	68.2	2.518 370	6.0	7.713
yolov11+ShuffleNetV2	87.4	74.4	80.4	84.8	55.8	0.926 229	3.0	5.122
yolov11+MWA	89.2	82.6	85.8	90.1	64.2	2.469 341	6.2	5.958
yolov11+ShuffleNetV2+MWA	89.4	82.5	85.8	90.0	64.0	0.812 053	2.9	5.188
本文算法	94.1	91.3	92.7	96.1	70.6	0.816 213	2.9	4.315

本文采用分层优化的精度-效率权衡策略。ShuffleNetV2 作为骨干网络承担主要的效率优化任务,能显著降低模型参数量和计算复杂度。为补偿轻量化网络的特征表达能力限制,本文引入 MWA 模块,通过局部与全局特征融合提升 mAP50-95 达 2%,有效弥补 ShuffleNetV2 的精度损失。同时,DySample 模块进一步优化特征重建精度。本文分层优化的权衡策略避免单一模块承担过重负担,最终实现参数量从 2.58 M 降至 0.82 M 的同时,mAP50-95 整体提升 13.1%。

为系统评估不同优化器(SGD 与 Adam)对模型性能的影响,本研究分别采用两者进行消融实验,结果如表 3(SGD 优化器)和表 4(Adam 优化器)所示。基于表 4 数据,可得出以下结论:1)依次引入 CARAFE 和 DySample 模块改进上采样,mAP50 相比基线模型分别提高了 −0.2%、4%,mAP50-95 分别提高了 −0.6%、3.8%。DySample 模块在检测性能和效率上均优于 CARAFE 模块,故采用 DySample 模块来改进模型的上采样。2)将 Swintransformer 模型、ShuffleNetV2 模块依次替换到模型的骨干网络中,mAP50 分别提高了 2.7%、−2.6%,mAP50-95 分别提高了 8.9%、−3.5%,ShuffleNetV2 模块将参数量从 2.58 M 降低到 0.93 M,ShuffleNetV2 模块虽显著降低参数量,但检测精度下降;需结合 MWA 以提升检测精度。3)在骨干网络末层加入 MWA 模块后,mAP50 提高了 2.7%,mAP50-95 提高了 4.9%,因此在 SGD 或 Adam 优化器下、MWA 模块均能有效提升检测精度。4)同时引入 ShuffleNetV2 和 MWA 模块后,mAP50 提高了 2.6%,mAP50-95 提高了 4.7%,参数量从 2.58 M 降低到 0.81 M,证明检测精度和参数量都得到了优化。5)本文模型的 mAP50 相比基线模型提高了 8.7%,mAP50-95 提高了 11.3%,参数量从 2.58 M 降低到 0.82 M。可见无论采用 SGD 还是 Adam 优化器,本文算法均能显著降低参数量并提升目标检测精度。

3.5 对比试验

为了进一步验证本文算法在公路场景下行人异常行为检测的有效性,将本文算法与经典主流模型进行对比实验,结果如表 5、6 所示。

表 5　对比实验(优化器为 SGD,训练 100 epoch)

Table 5　Contrast experiment (optimizer is SGD, train 100 epochs)

实验	P/%	R/%	F1/%	mAP50/%	mAP 50-95)/%	Parameters/M	GFLOPs=FLOPs/G	训练时间/h
CNN	66.8	91.5	77.2	93.9	82.1	25.56	4.1	0.620
SSD	76.3	79.8	78.0	85.1	63.9	24.5	87.9	2.720
FasterRCNN	91.7	68.9	78.7	78.6	44.4	41.5	182.5	5.280
yolov11	86.2	80.6	83.3	88.3	61.3	2.583 517	6.3	4.103
本文算法	94.1	93.1	93.6	96.8	74.4	0.816 213	2.9	5.265

表 6　对比实验(优化器为 Adam,训练 100 epoch)

Table 6　Contrast experiment (optimizer is Adam, train 100 epochs)

实验	P/%	R/%	F1/%	mAP50/%	mAP 50-95)/%	Parameters/M	GFLOPs=FLOPs/G	训练时间/h
CNN	80.6	71.2	75.6	78.0	51.6	25.56	4.1	0.620
SSD	80.7	63.7	71.2	70.9	42.3	24.5	87.9	3.050
FasterRCNN	80.1	75.3	77.6	81.8	56.8	41.5	182.5	5.280
yolov11	84.6	79.4	81.9	87.4	59.3	2.583 517	6.3	5.439
本文算法	94.1	91.3	92.7	96.1	70.6	0.816 213	2.9	4.315

根据表 5 的实验结果,本文算法在公路场景异常行为数据集上综合性能表现优异,具体结论如下:1)与 CNN 相比,本文算法的 mAP50 提升了 2.9%,mAP50-95 降低了 7.7%,参数量由 25.56 M 降至 0.82 M,本文算法轻量化优势显著。SGD 优化器使 CNN 更关注定位精度,但 CNN 模型本身分类能力不足导致 P、R 较低,从而导致 CNN 虽然目标定位表现较好但分类置信度校准不佳。本文模型综合表现优。2)与 SSD 相比,本文算法的 mAP50 提升了 11.7%,mAP50-95 提升了 10.5%,参数量由 24.5 M 降至 0.82 M,本文算法在轻量化和检测精度上均表现优异。3)与 FasterRCNN 相比,本文算法的 mAP50 提升了 18.2%,mAP50-95 提升了 30%,参数量由 41.5 M 降至 0.82 M,本文算法的轻量化和检测精度均显著优于 FasterRCNN 模型。4)与基线模型 YOLOv11 相比,本文模型的 mAP50 提升了 8.5%,mAP50-95 提升了 13.1%,参数量从 2.58 M 降低到 0.82 M。从表 5 可得,本文算法在显著降低参数量的同时提升目标检测精度,验证了算法在轻量化与检测性能上的双重优势。

为系统评估不同优化器(SGD 与 Adam)对模型性能的影响,本研究采用 Adam 优化器进行对比实验,实验结果如表 6 所示。根据表 6 数据,本文算法在公路场景异常行为数据集上的综合性能表现优异,具体结论如下:1)与 CNN 相比,本文算法 mAP50 提升了 18.1%,mAP50-95 提升了 19%,CNN 参数量高达 25.56 M 而本文只需 0.82 M,本文算法轻量化优势显著。2)与 SSD 相比,本文算法 mAP50 提升了 25.2%,mAP50-95 提升了 28.3%,SSD 参数量高达 24.5 M 而本文只需 0.82 M,本文算法在轻量化和检测精度上均表现优异。3)与 FasterRCNN 相比,本文算法 mAP50 提升了 14.3%,mAP50-95 提升了 13.8%,SSD 参数量高达 41.5 M 而本文只需 0.82 M,本文算法在轻量化和检测精度均显著优于 FasterRCNN 模型。4)与基线模型 YOLOv11 相比,本文算法 mAP50 提升了 8.7%,mAP50-95 提升了 11.3%,且参数量从 2.58 M 降低到 0.82 M。对比表 5、表 6 各模型的实验结果,无论采用 SGD 还是 Adam 优化器,本文算法均能显著降低参数量并提升目标检测精度,验证了算法在轻量化与检测性能上的双重优势。

3.6　可视化结果验证

为了可视化分析各模块对 YOLOv11 模型检测性能的影响,本文对比了原始 YOLOv11、采用 CARAFE 或 DySample 改进上采样、引用 Swintransformer、MWA、ShuffleNetV2 以及本文算法在训练 100 个 epochs 时的 mAP50 值。图 6 验证了本文算法在提升模型检测精度方面的有效性。

图 7 展示了原始 YOLOv11 模型与本文改进模型在公路场景异常行为数据集上的检测对比结果。具体分析如下:从图 7(a)、(b)和(c)可观察到,在较为清晰公路场景下,原始 YOLOv11 模型对异常行为的识别能力较弱,而本文模型能更精准地定位并标注目标,检测精度显著优于原始模型。对比图 7(d)、(e)和(f),原始 YOLOv11 模型在检测背景复杂的情况时存在误判现象(如将石柱误标为“打架斗殴”),而本文模型通过优化特征提取与注意力机制,有效提升了目标类别的识别准确性。从图 7(g)、(h)和(i)可以发现,在背景模糊情况下,原始 YOLOv11 模型发生漏

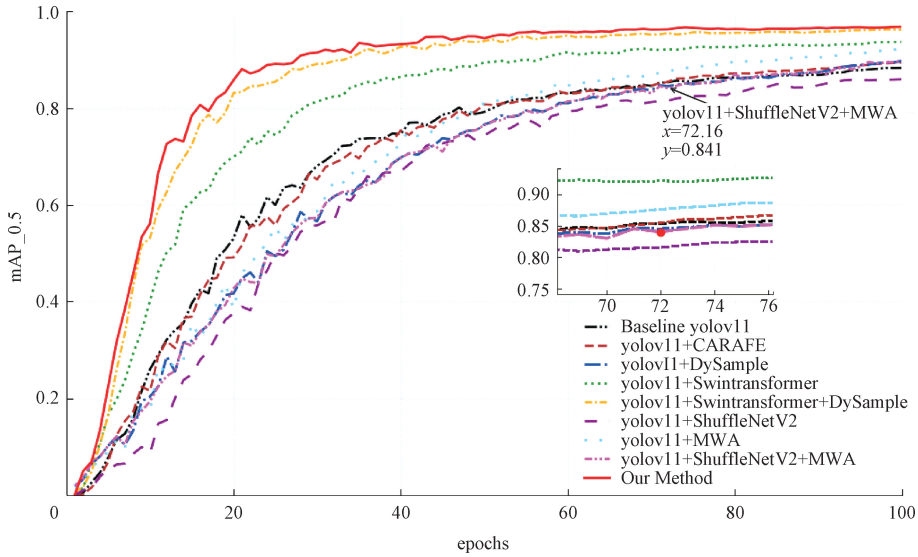
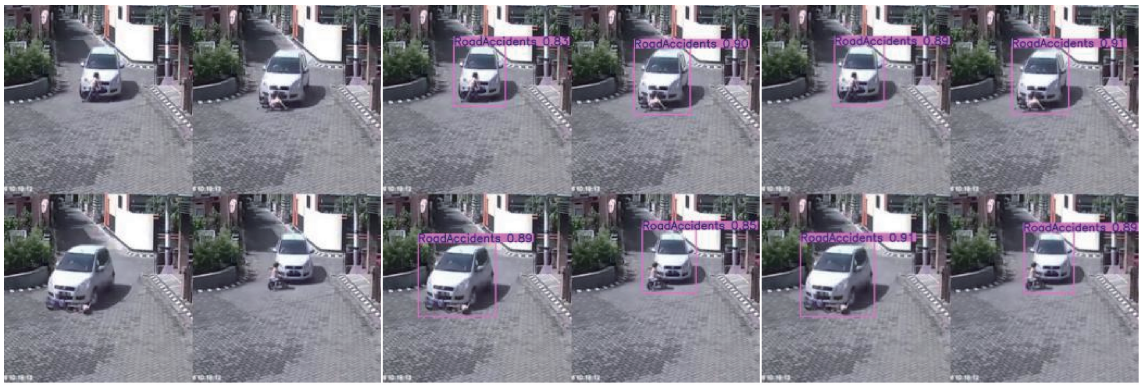


图 6 mAP50 效果对比

Fig. 6 Comparison of mAP50 results

标现象(如未完全识别火焰区域),而本文模型能更精准地 定位火焰位置和燃烧区域。



(a) 交通事故原图
(a) Original image of traffic accident

(b) YOLOv11检测交通事故
(b) YOLOv11 detects traffic accident

(c) 本文模型检测交通事故
(c) Proposed model detects traffic accident

(d) 打架斗殴原图
(d) Original image of fight

(e) YOLOv11检测打架斗殴
(e) YOLOv11 detects fight

(f) 本文模型检测打架斗殴
(f) Proposed model detects fight

(g) 纵火原图
(g) Original image of arson

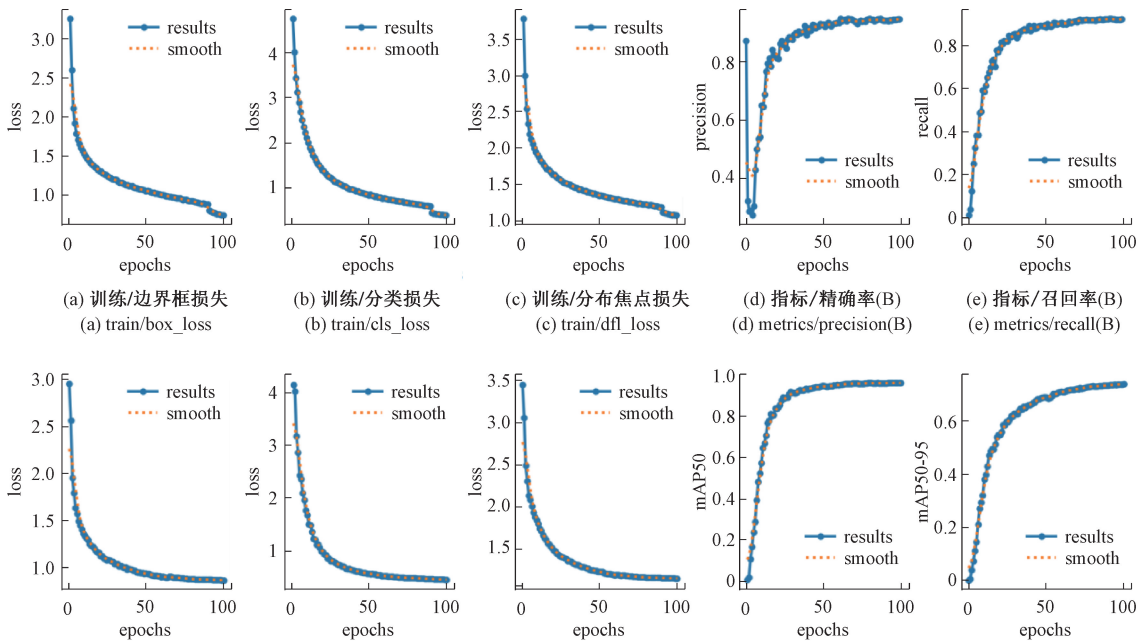
(h) YOLOv11检测纵火
(h) YOLOv11 detects arson

(i) 本文模型检测纵火
(i) Proposed model detects arson

图 7　结果对比

Fig. 7　Comparison of results

图 8 是模型表现的综合评估图，展示了训练集(train)和验证集(val)在损失函数、精度、召回率、mAP 等指标的变化趋势。具体分析如下：1)图 8(a)和(f)是衡量边界框预测的准确度：box_loss 值随训练过程逐渐降低，表明模型对边界框的预测精度不断提升。2)图 8(b)与(g)表示分类损失(即分类预测的准确度)：cls_loss 在训练过程中逐渐减少并最终趋于平稳，表明类别预测的损失逐渐下降并趋近于 0，分类准确性增强。3)图 8(c)与(h)：分布损失，关注目标位置与形状的预测：dfl_loss 随训练过程逐渐降低并趋于稳定，表明模型对目标位置和形状的预测精度提升。4)图 8(d)与(e)展示了精度和召回率：P 和 R 都很高，持续上升并逐渐接近于 1，表明模型在分类任务上表现良好。5)图 8(i)与(j)可以看出，模型在所有动作检测的 mAP 值高于 0.9，表明改进后的模型显著提升了检测准确性。



(a) 训练/边界框损失
(a) train/box_loss

(b) 训练/分类损失
(b) train/cls_loss

(c) 训练/分布焦点损失
(c) train/dfl_loss

(d) 指标/精确率(B)
(d) metrics/precision(B)

(e) 指标/召回率(B)
(e) metrics/recall(B)

(f) 验证/边界框损失
(f) val/box_loss

(g) 验证/分类损失
(g) val/cls_loss

(h) 验证/分布焦点损失
(h) val/dfl_loss

(i) 指标/mAP50(B)
(i) metrics/mAP50(B)

(j) 指标/mAP50-95(B)
(j) metrics/mAP50-95(B)

图 8　模型评估结果

Fig. 8　Model evaluation results

4　结　　论

本文提出了一种基于改进 YOLOv11 的目标检测算法，针对监控视角下公路场景的复杂背景与模糊视角问题，提出以下优化策略：引入 DySample 模块和 MWA 模块，提升模型整体检测性能；引入 ShuffleNetV2 作为模型主干，将模型参数量显著降低。实验结果表明，本文算法较基线模型的 mAP50 提升 8.5%、mAP50-95 提升

13.1%,同时参数量显著降低。与其他检测算法相比,本文算法具有更低的参数量和更高的检测精度,适合实时监控场景应用。未来将继续优化算法,提高模型的实时检测性能,并部署到公路的摄像头上。

参考文献

[1] LIU Z, LIN Y, CAO Y, et al. Swin transformer: Hierarchical vision transformer using shifted windows[C]. IEEE/CVF International Conference on Computer Vision (ICCV). Piscataway: IEEE Press, 2021: 9992-10002.

[2] 杨傲雷,周应宏,杨帮华,等. 基于 Transformer 的三维人体姿态估计及其动作达成度评估[J]. 仪器仪表学报,2024,45(4):136-144.
YANG AO L, ZHOU Y H, YANG B H, et al. 3D human pose estimation based on Transformer and its action achievement evaluation[J]. Chinese Journal of Scientific Instrument, 2024, 45(4): 136-144.

[3] PEREZ M, KOT A C, ROCHA A. Detection of real-world fights in surveillance videoss [C]. IEEE International Conference on Acoustics, Speech and Signal Processing. Veszprem: IEEE, 2019: 2662-2666.

[4] 张英迪,史泽林,王欢,等. 基于融合 Swin Transformer 网络的腰椎解剖区域自动分割方法[J]. 信息与控制,2025,54(3):390-400.
ZHANG Y D, SHI Z L, WANG H, et al. Automatic segmentation method of lumbar anatomical region based on fused Swin Transformer network [J]. Information and Control, 2025, 54(3): 390-400.

[5] 叶彦斐,胡龙癸,张成龙. 基于改进 YOLOv8n-Pose 的轨道作业人员跨轨安全动作识别[J]. 国外电子测量技术,2024,43(8):181-188.
YE Y F, HU L K, ZHANG CH L. Safety action recognition of track workers crossing rails based on improved YOLOv8n-Pose [J]. Foreign Electronic Measurement Technology, 2024, 43(8): 181-188.

[6] CHEN J, KAO SH, HE H, et al. Run, don't walk: chasing higher FLOPS for faster neural networks[C]. IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023: 12021-12031.

[7] WANG C Y, BOCHKOVSKIY A, LIAO H Y M. YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors[C]. IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023: 7464-7475.

[8] GEMMEKE J F, ELLIS D P, FREEDMAN D, et al. Audio set: An ontology and human-labeled dataset for audio eventss[C]. IEEE International Conference on Acoustics, Speech and Signal Processing. New Orleans: IEEE, 2017: 776-780.

[9] WU T, TANG S, ZHANG R, et al. Cgnet: A light-weight context guided network for semantic segmentation [J]. IEEE Transactions on Image Processing, 2020, 30: 1169-1179.

[10] MIAO S, HOU Y, GAO Z, et al. A central difference graph convolutional operator for skeleton-based action recognition[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2022, 32(7): 4893-4899.

[11] YE M L, ZHOU H Y, LI J J. Forest fire detection algorithm based on an improved Swin Transformer [J]. Journal of Central South University of Forestry & Technology, 2022, 42: 101-110.

[12] SULTANI W, CHEN C, SHAH M. Real-world anomaly detection in surveillance videos [C]. IEEE Conference on Computer Vision and Pattern Recognition, 2018: 6479-6488.

[13] 赵云涛,黄哲辉. 面向特钢车间内物料实时跟踪的钢管目标检测算法研究[J]. 电子测量与仪器学报,2024,38(11):210-218.
ZHAO Y T, HUANG ZH H. Research on steel pipe target detection algorithm for real-time material tracking in special steel workshop [J]. Journal of Electronic Measurement and Instrumentation, 2024, 38(11): 210-218.

[14] 季星宇,黄陈蓉,姚军财,等. 结合分支特征和排斥损失的绝缘子检测研究[J]. 计算机工程与应用,2025,61(7):141-152.
JI X Y, HUANG CH R, YAO J C, et al. Insulator detection research combining branch features and repulsion loss [J]. Computer Engineering and Applications, 2025, 61(7): 141-152.

[15] 赵登阁,智敏. 用于人体动作识别的多尺度时空图卷积算法[J]. 计算机科学与探索,2023,17(3):719-732.
ZHAO D G, ZHI M. Multi-scale spatial-temporal graph convolution algorithm for human action recognition [J]. Journal of Frontiers of Computer Science and Technology, 2023, 17(3): 719-732.

[16] MUNRO J, DAMEN D. Multi-modal domain adaptation for fine-grained action recognitions [C]. IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020: 122-132.

[17] FEICHTENHOFER C. X3D: Expanding architectures for efficient video recognitions [C]. IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020: 203-213.

[18] 黄展原,李兵,李庚浩.基于视频和人体姿态估计的老年人摔倒监测研究[J].计算机工程与科学,2021,43(5):883-890.
HUANG ZH Y, LI B, LI G H. Research on fall monitoring for elderly based on video and human pose estimation[J]. Computer Engineering & Science, 2021, 43(5): 883-890.

[19] ZHONG Z, HOU Z, LIANG J, et al. Multimodal cooperative self-attention network for action recognition[J]. IET Image Processing, 2023, 17(6): 1775-1783.

[20] 李焱,董仕豪,张家伟,等.融合通道注意力的跨尺度 Transformer 图像超分辨率重建[J].中国图象图形学报,2025,30(3):784-797.
LI Y, DONG SH H, ZHANG J W, et al. Cross-scale transformer image super-resolution reconstruction with channel attention[J]. Journal of Image and Graphics, 2025, 30(3): 784-797.

[21] NGUYEN T, PHAM D T, VU H, et al. A robust and efficient method for skeleton-based human action recognition and its application for cross-dataset evaluation[J]. IET Computer Vision, 2022, 16(8): 709-726.

[22] 肖恒树,李军营,梁虹,等.改进 YOLOv8 的轻量化烟叶计数检测算法[J].电子测量技术,2025,48(8):177-186.
XIAO H SH, LI J Y, LIANG H, et al. Lightweight tobacco leaf counting detection algorithm based on improved YOLOv8[J]. Electronic Measurement Technology, 2025, 48(8): 177-186.

[23] 李晶晶,黄章进,邹露.基于运动引导图卷积网络的人体动作识别[J].计算机辅助设计与图形学学报,2024,36(7):1077-1086.
LI J J, HUANG ZH J, ZOU L. Human action recognition based on motion-guided graph convolutional network[J]. Journal of Computer-Aided Design & Computer Graphics, 2024, 36(7): 1077-1086.

[24] DETONE D, MALISIEWICZ T, RABINOVICH A. Superpoint: Self-supervised interest point detection and descriptions[C]. IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2018: 224-236.

作者简介

刘玉蕾,硕士,主要研究方向为信号与信息处理技术。
E-mail: 1115974565@qq.com

褚丽莉(通信作者),博士,教授,主要研究方向为信息与编码理论。
E-mail: chulili902@126.com

李波,博士,教授,主要研究方向为智能信息处理。
E-mail:leebo@yeah.net