Pure Mathematics 理论数学, 2024, 14(7), 1-14
Published Online July 2024 in Hans. https://www.hanspub.org/journal/pm
https://doi.org/10.12677/pm.2024.147265

Hans汉斯

改进YOLOv7算法的工业安全多目标检测研究

孟祥龙

上海理工大学，光电信息与计算机工程学院，上海

收稿日期：2024年3月27日；录用日期：2024年5月6日；发布日期：2024年7月3日

摘　要

在工厂阀室内，安全问题一直是一项关注的重点，为了防止工厂阀室内员工未佩戴安全帽、阀室内抽烟、打电话分心等异常行为对工厂设备和员工造成伤害，目标检测技术常常被应用于工业安全场景下进行员工异常行为的识别，从而保证工业生产安全进行。目标检测技术作为工业安全监控的核心组成部分，其在真实场景下应用、算法识别性能尤为关键。为了进一步研究应用于工业生产安全领域的人体异常行为识别算法，本文进行了如下工作：(1) 本文提出了一个近两万张工厂阀室背景下的数据集，数据集共2万多张真实工厂阀室下的监控拍摄图片。将其按照训练集(验证集)和测试集按照4:1的比例进行划分。为了数据集的实用性，对风险等级最高的8种异常行为以及三种安全行为进行数据标注形成了11类标签的USDA数据集，并且为了提高模型的鲁棒性，还对USDA数据集进行数据增强扩充，使其更具有在真实场景下应用的实用性。(2) 本文提出了改进的SD-YOLOv7模型进行异常行为的识别。在该模型中，首先在核心网络Backbone内集成了Squeeze-and-Excitation Networks (SENet)的注意力机制。其中，SENet引入了一个创新的特征重标定方法，它能够自动学习并分配各个特征通道的权重，从而提高关键特征通道的权重值，通过自学习的方式自动获取每个特征通道的权重，增大目标重要特征通道的权重值，同时加入可变形卷积DCNv4来替换传统卷积，以更好地捕捉各种角度和姿态的目标以加强SD-YOLOv7模型在工厂监控下在不同角度下的目标检测能力，此外还提出了一种符合工厂阀室场景特征的加权损失函数FA loss。本模型在自建的2万多张图像数据集下与传统YOLOv7、YOLOv9等模型做了对比实验。结果表明本模型在召回率、平均精确率(mAP)比传统YOLOv7、YOLOv9等模型有较大提升，改进后的SD-YOLOv7算法在增加较少复杂度的情况下明显提升了算法的性能。此外，该模型已经成功部署在边缘设备上，可以成功地为合作单位进行实时检测。本研究为工业安全监控领域提供了一个高度实用的数据集和一种高效的目标检测模型，未来将探索在更多实际工业场景中的应用。

关键词

工厂阀室，注意力机制，可变形卷积，边缘设备

Research on the Improvement of YOLOv7 Algorithm for Industrial Safety Multi-Target Detection

文章引用: 孟祥龙. 改进 YOLOv7 算法的工业安全多目标检测研究[J]. 理论数学, 2024, 14(7): 1-14.
DOI: 10.12677/pm.2024.147265

Xianglong Meng

School of Optoelectronic Information and Computer Engineering, University of Shanghai for Science and Technology, Shanghai

Received: Mar. 27th, 2024; accepted: May 6th, 2024; published: Jul. 3rd, 2024

Abstract

Safety has always been a focal point within factory valve rooms. To prevent abnormal behaviors such as employees not wearing safety helmets, smoking within the valve room, or being distracted by phone calls from causing harm to factory equipment and personnel, target detection technology is often utilized in industrial safety scenarios for the recognition of employee abnormal behaviors, thereby ensuring the safety of industrial production. Target detection technology, as a core component of industrial safety monitoring, is particularly crucial for its application in real scenarios and the performance of algorithm recognition. To further study the human abnormal behavior recognition algorithms applied in the field of industrial production safety, this paper has conducted the following work: (1) This paper proposes a dataset of nearly 20,000 images set against the backdrop of a factory valve room, consisting of over 20,000 real surveillance photos taken in actual factory valve room settings. The dataset is divided into training (validation) and testing sets in a 4:1 ratio. To enhance the practicality of the dataset, data annotation was conducted for the 8 highest-risk abnormal behaviors and three safe behaviors, forming an 11-class labeled USDA dataset. Additionally, to improve the robustness of the model, data augmentation was applied to the USDA dataset to make it more applicable in real-world scenarios. (2) This paper introduces an improved SD-YOLOv7 model for the recognition of abnormal behaviors. In this model, the Squeeze-and-Excitation Networks (SENet) attention mechanism was first integrated into the core network Backbone. SENet introduces an innovative feature recalibration method that can automatically learn and allocate weights to each feature channel, thereby enhancing the weight of key feature channels. It automatically acquires the weights of each feature channel through self-learning, increasing the weight of important target feature channels. At the same time, deformable convolution DCNv4 is used to replace traditional convolution to better capture targets at various angles and postures, thereby strengthening the target detection capabilities of the SD-YOLOv7 model in factory monitoring under different angles. Furthermore, a weighted loss function, FA loss, was proposed that is tailored to the characteristics of the factory valve room scenario. This model was compared with traditional YOLOv7, YOLOv9, and other models on a self-built dataset of over 20,000 images. The results indicate that this model has significantly improved in recall rate and mean average precision (mAP) compared to traditional YOLOv7, YOLOv9, and other models, with the improved SD-YOLOv7 algorithm significantly enhancing the performance of the algorithm with minimal increase in complexity. Additionally, the model has been successfully deployed on edge devices and can successfully perform real-time detection for cooperative units. This study provides a highly practical dataset and an efficient target detection model for the field of industrial safety monitoring, and future work will explore its application in more actual industrial scenarios.

Keywords

Factory Valve Room, Attention Mechanism, Deformable Convolution, Edge Device

1. 引言

近年来，随着全球工业化进程的加速，工业安全问题日益凸显，成为影响社会稳定和经济发展的重要因素。工业安全事故不仅对工人的生命安全构成直接威胁，还可能导致巨大的经济损失和环境破坏。特别是在化工、石油、采矿等高风险行业中，异常行为的发生的后果非常严重[1]，这些行为往往源于工人的安全意识不足、安全培训不到位或管理层的监督不力。例如，在工厂阀室内，未佩戴安全帽可能导致头部受伤，吸烟可能引发火灾或爆炸[2]，而打电话分心则可能使工人无法及时响应紧急情况。这些看似日常的行为，一旦发生意外，后果往往是灾难性的。根据国际劳工组织(ILO)的报告，每年约有 270 万工人死于职业事故和职业病，而这些事故中有相当一部分是由于上述等异常行为引起的[3]。此外，工业事故还可能导致生产中断、设备损坏、产品损失和法律责任，给企业带来重大的经济负担。对于社会而言，工业事故不仅会造成人员伤亡和财产损失，还可能引发公众对工业安全的不信任，影响社会稳定[4]。

为了应对这一挑战，各国政府和企业已经开始采取一系列措施，如制定严格的安全法规、加强安全培训、提高安全设施投入等。然而，仅仅依靠传统的安全管理措施已经难以满足日益复杂的工业环境需求。因此，利用先进的技术手段，如人工智能和深度学习，来提高工业安全监控的智能化水平，已成为行业发展的新趋势。通过引入智能化的安全监控系统，可以实时监测工人的行为，及时发现并预防异常行为的发生[5]。接下来介绍国内外在工业安全监控领域的研究现状。

目标检测是计算机视觉领域的重要研究方向之一，旨在从图像或视频中自动检测并识别出特定类别的物体目标[6]，其广泛应用于智能交通、智能安防、无人驾驶等领域。在国外，YOLO、SSD、Faster R-CNN、Mask R-CNN 等算法已成为目标检测领域的代表性算法，它们在准确率、速度等方面都取得了不俗的表现。在国内，研究者们也取得了不少进展，如 CFE、RefineDet、HRNet、ATSS 等算法在特定场景下表现出了优异的性能[7]，其中 HRNet 在多个目标检测竞赛中取得了不俗的成绩。

行为识别(Action Recognition)是计算机视觉中极其重要也非常活跃的研究方向，它已经被研究了数十年。因为人们可以用动作(行为)来处理事情、表达感情，因此行为识别有非常广泛但又未被充分解决的应用领域，例如智能监控系统、人机交互、虚拟现实、机器人等。以往的方法中都使用 RGB 图像序列，深度图像序列，视频或者这些模态的特定融合(例如 RGB+光流)，也取得了超出预期的结果[8]。然而，和骨架数据(人体关节和骨头的一种拓扑表示)相比，前述模态会产生更多的计算消耗，且在面对复杂背景以及人体尺度变化、视角变化和运动速度变化时鲁棒性不足。此外，像 Microsoft Kinect 这样的传感器和一些先进的人体姿态估计算法都可以更轻松地获得准确的 3D 骨架(关键点)数据[9]。行为识别方法相对于目标检测方法表现较差，比如因摸嘴巴和抽烟的行为很类似，难以辨别，以及对边缘设备的性能要求较高，于是放弃了研究，选择了基于目标检测的人体异常行为识别。

本文提出了一种改进版的 YOLOv7 算法 SD-YOLOv7，在自建数据集上进行实验，结果显示本模型在检测工厂阀室内工人未佩戴安全帽、未着劳保服装等异常行为方面，SD-YOLOv7 与基线模型相比性能有明显提升。

2. 相关技术

2.1. YOLOv7 网络介绍

YOLOv7 是 YOLO 系列中最先进的新型目标检测器[10]。它是在目标检测领域测测速度非常快、非常准确的实时目标检测器，YOLOv7 通过引入多项架构改革提高了速度和准确性。与 Scaled YOLOv4 类似，YOLOv7 主干不使用 ImageNet 预训练的主干。相反，模型完全使用 COCO 数据集进行训练。YOLOv7 相较于之前的版本改进点主要是有引入扩展高效层聚合网络(E-ELAN)、进行了基于串联模型的模型缩放、加入了可训练的 BoF 以及计划重新参数化卷积等结果表明，模型获得了 56.8%的平均精度，所有

YOLOv7 模型在 5 FPS 到 160 FPS 范围内的速度和精度都超过了之前的目标检测器。YOLOv7 最大的贡献是再一次平衡好了参数量、计算量和性能之间的矛盾，主干网络 backbone 最开始的部分是 stem 层，通过三层 3x3 的卷积核输出一个二倍降采样的特征图，YOLOv7 再用一层步长为 2 的卷积得到 4 倍降采样图，然后接了 ELAN 模块处理这个 4 倍降采样特征图，随后 YOLOv7 算法再对该 4 倍降采样的特征图再进行降采样操作，不过，不同于之前的步长为 2 的卷积，YOLOv7 这里稍微设计得精细了一些，左分支主要采用 maxpooling (MP)来实现空间降采样，并紧跟一个 1 × 1 卷积压缩通道；右边先用 1 × 1 卷积压缩通道，然后再用步长为 2 的 3 × 3 卷积完成降采样，最后，将两个分支的结果合并，随后，ELAN 模块再对被降采样的特征图进行处理。然后重复堆叠这两块，直到最后的 YOLOv7 的 backbone 构建完成。

2.2. 卷积神经网络

卷积神经网络(Convolutional Neural Networks, CNNs)是深度学习领域中一种具有革命性意义的网络结构，它在图像识别、视频分析和自然语言处理等多个领域均取得了卓越的成就[11]。CNN 的核心思想是利用卷积层自动从输入数据中学习空间层次结构的特征，这一机制使得 CNN 在处理具有网格结构的数据(如图像)时表现出独特的优势。

卷积层是 CNN 的核心组件，它通过卷积运算提取输入数据的局部特征。在数学上，卷积可以定义为两个函数(卷积核和输入信号)的积分或求和。在 CNN 中，卷积核是一个小的权重矩阵，它在输入数据(如图像)上滑动，计算核与输入数据局部区域的点积，从而生成特征图(feature map)。这个过程可以表示为：

$$(f*g)(t)=\int_{-\infty}^{\infty} f(\tau)g(t-\tau)\mathrm{d}\tau \tag{2.1}$$

其中，f 是输入信号，g 是卷积核，$*$ 表示卷积运算，t 是卷积核滑动的位置。卷积操作的优势在于它能够捕捉输入数据的空间相关性，同时通过共享卷积核的权重实现参数共享，这大大减少了模型的复杂度。此外，卷积层通常伴随着非线性激活函数(如 ReLU)，以增强模型的表达能力。

池化层(Pooling Layer)是 CNN 中的另一个关键组件，它负责降低特征图的空间维度，从而减少计算量和防止过拟合。最常见的池化操作是最大池化(Max Pooling)和平均池化(Average Pooling)。池化层通过在特征图上滑动一个固定大小的窗口，并提取窗口内的最大值或平均值作为输出，实现下采样。这一过程不仅减少了数据的空间尺寸，还保留了重要的特征信息。CNN 通常由多个卷积层和池化层交替堆叠而成，形成深度网络结构。在网络的深层，低级特征(如边缘和纹理)逐渐被组合成更高级的语义特征(如物体的部分和整体结构)。这种层次化的特征表示不仅符合人类视觉系统的工作原理，而且能够有效地处理图像中的平移、旋转和缩放变化。

在 CNN 的末端，全连接层(Fully Connected Layer)将前面层次结构中提取的特征映射到最终的输出，如分类标签。全连接层的每个神经元与前一层的所有激活值相连，通过学习权重和偏置，实现从特征空间到输出空间的映射。在多分类问题中，通常在全连接层后接一个 softmax 层，它将网络输出转换为概率分布，从而进行类别预测。CNN 的训练涉及前向传播和反向传播两个过程。在前向传播中，输入数据通过网络，生成输出；在反向传播中，根据输出与真实标签的差异计算损失函数的梯度，并通过网络传播回权重，更新网络参数。梯度下降及其变体(如 SGD、Adam 等)是最常用的优化算法，用于最小化损失函数，训练过程中还常常结合正则化技术(如 L1、L2 正则化和 Dropout)来提高模型的泛化能力[12]。CNN 的标准结构虽然强大，但研究人员一直在探索新的变体来提高性能和适应性。例如，残差网络(ResNet)通过引入残差连接来解决深度网络训练的困难，使得网络能够学习到更深层次的特征[13]。此外，Inception 网络通过多尺度的卷积核来捕捉不同尺度的特征，提高了模型对图像的理解能力。这些变体不仅在图像分类任务中取得了突破，也为其他视觉任务，如目标检测和语义分割，提供了新的思路。

2.3. 注意力机制

注意力机制(Attention Mechanism)是一种模拟人类视觉注意力的计算模型，它允许深度学习网络在处理信息时对输入数据的某些部分进行聚焦。在卷积神经网络中，注意力机制通过引入额外的加权层来实现，这些层能够学习输入特征图中各个区域的重要性，并据此分配处理资源。

注意力机制的基本原理是通过对输入特征图 F 中的每个区域进行加权求和，生成一个加权特征表示 F'：

$$F' = \sum_{i=1}^{N} w_i \cdot F_i \tag{2.2}$$

其中，N 是特征图中区域的数量，w_i 是第 i 个区域的注意力权重，通常由模型学习得到。注意力权重 w_i 的计算通常依赖于输入特征 F_i 和一个可学习的权重向量 $\mathbf{v}$ 之间的相似度评分，计算如下：

$$w_i = \frac{\exp\left(\mathbf{v}^{\top} \cdot \mathbf{s}\left(F_i\right)\right)}{\sum_{j=1}^{N} \exp\left(\mathbf{v}^{\top} \cdot \mathbf{s}\left(F_j\right)\right)} \tag{2.3}$$

这里，$\mathbf{s}(\cdot)$ 是一个用于提取特征表示的函数，可以是一个全连接层或者特定的特征提取网络。权重向量 $\mathbf{v}$ 通过训练过程学习得到，确保模型能够关注对当前任务最重要的特征。

注意力机制的应用显著提升了模型对关键信息的捕捉能力，尤其是在目标检测任务中，它能够帮助模型更好地识别和定位目标。此外，注意力机制还增强了模型对于上下文信息的利用，从而在复杂的视觉场景中提高了识别的准确性和鲁棒性。在 YOLOv7 中，注意力机制的集成可以通过 Squeeze-and-Excitation (SE) 模块实现，该模块通过显式地建模通道间的关系来提高特征的表达能力。SE 模块的核心思想是利用全局平均池化(Global Average Pooling, GAP)来获取通道的统计信息，并通过两个全连接层(FC)来学习每个通道的重要性权重：

$$z_c = \sigma\left(\mathbf{W}_2 \sigma\left(\mathbf{W}_1 \cdot \frac{1}{H \cdot W} \sum_{i=1}^{H} \sum_{j=1}^{W} F_c\left(i, j\right)\right)\right) \tag{2.4}$$

$$F_c' = z_c \cdot F_c \tag{2.5}$$

这里，$F_c(i, j)$ 表示在通道 c 上的 (i, j) 位置的特征值，$H \cdot W$ 是特征图的元素总数，$\mathbf{W}_1$ 和 $\mathbf{W}_2$ 是 SE 模块中的全连接层的权重矩阵，σ 是激活函数。

通过这种方式，YOLOv7 能够有效地提高对目标的检测精度，尤其是在复杂的工业场景中，注意力机制的引入显著提升了模型对异常行为的识别能力。

3. 改进的 YOLOv7 模型

随着工业自动化和智能化的快速发展，工厂阀室内的人体异常行为识别成为工业安全监控领域的核心问题。为了提高检测的准确性和鲁棒性，本章提出了一种改进的 YOLOv7 算法，即 SD-YOLOv7。该算法通过融合 Squeeze-and-Excitation Networks 的注意力机制和可变形卷积 DCNv4，显著提升了模型在复杂工业场景下的表现。此外，本章还将介绍 SD-YOLOv7 模型的实验设计，包括数据集的构建、数据增强策略、评估指标的选择以及实验结果的详细分析。本文将展示如何通过自定义的加权损失函数 FA loss，进一步优化模型以适应工厂阀室场景的特定需求。改进的 YOLOv7 模型 SD-YOLOv7 在核心网络 Backbone 内集成了 Squeeze-and-Excitation Networks 的注意力机制。同时在部分 CBS 模块改为引入了可变形卷积 DCNv4 的 DCBS，以更好地捕捉各种角度和姿态的目标。提高模型在工厂监控下在不同角度下的目标检测能力，并且提出了符合工厂阀室场景特征的加权损失函数 FA loss。SD-YOLOv7 模型如图 1 所示。

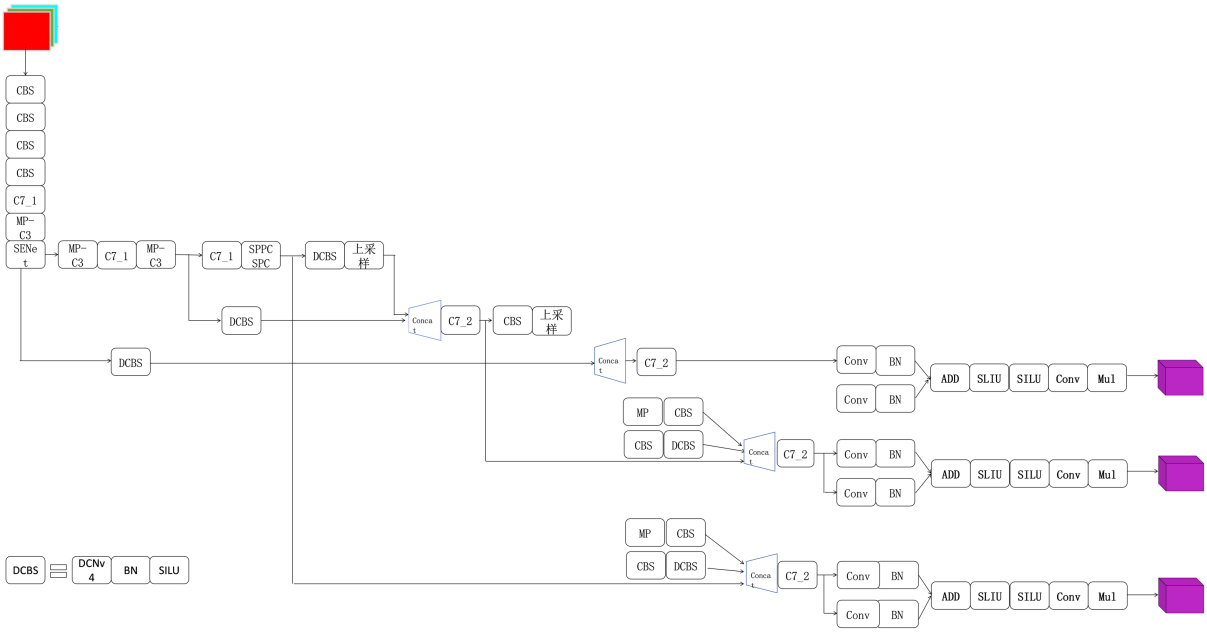

Figure 1. SD-YOLOv7 network structure
图 1. SD-YOLOv7 网络结构

3.1. 主干网络引入注意力机制

基于人类视觉系统的研究，注意力机制旨在从大量数据中筛选出对特定任务最为关键的信息。本研究针对香烟识别任务引入了注意力机制，优先处理与香烟相关的特征，从而使模型集中分析与识别香烟有关的关键区域。研究表明，将注意力机制纳入模型设计中可以有效提高其性能。而这种机制的整合位置相对灵活，关键在于能否促进模型效能的提升。本文特别采用了 Squeeze-and-Excitation Networks 的架构于模型的通道层级中，并将其嵌入到 YOLOv7 网络架构之中，获得了显著的实验成果。SENet 通过引入一种革新的特征重调方法，实现了对特征通道权重的自动学习与分配，以此强调对检测任务有益的特征并抑制那些较为无关的特征。通过这种方式，SENet 增强了卷积神经网络(CNN)内部通道之间的相互联系，通过自适应学习分配权重，有效地提升了模型对关键特征的关注度并降低了不相关特征的干扰。SENet 包含两个关键模块：Squeeze 模块和 Excitation 模块，如图 2 所示。在 Squeeze 模块中，它通过使用全局平均池化(global average pooling)将每个通道的特征图转换为一个单独的数字，然后将该数字输入到一个小型的全连接神经网络(FCN)中。这个 FCN 可以学习通道之间的关系，并为每个通道分配一个权重。在 Excitation 模块中，每个通道的特征图都被乘以相应的权重，从而增强有用的通道，抑制无用的通道。

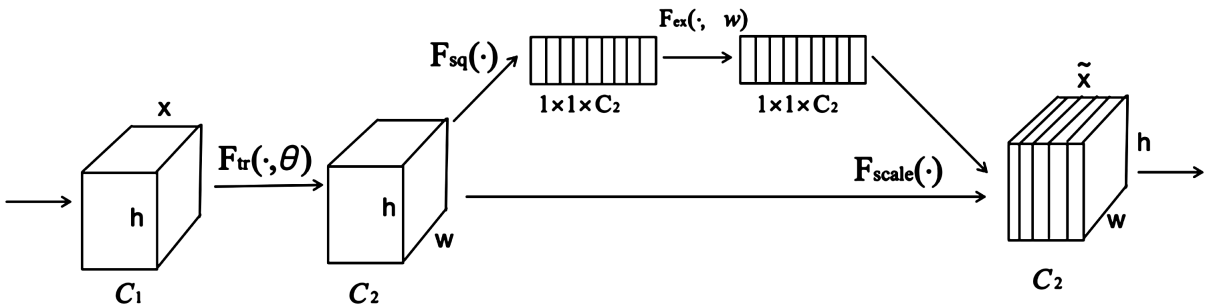

Figure 2. SENet model diagram
图 2. SENet 模型图

在 SD-YOLOv7 模型中将 SENet 注意力机制加入 MP-C3 模块之后，如图 3 所示，处理后的特征图被输入到下一层网络中。通过 SENet 注意力机制，可以增强本模型对香烟、手机、安全帽重要特征的关注度，从而提高网络的表现能力。YOLOv7 算法能够更加准确地识别出工厂阀室内的异常行为，从而有效提高工厂设备和员工的安全性。

设输入的特征图为 $X \in \boldsymbol{R}^{H\times W\times C}$，经过全局平均池化得到全局描述向量 $z \in \boldsymbol{R}^{C}$。然后，将全局描述向量 z 分别输入到两个全连接层中，得到它们的输出 s 和 t：

$$s = f_1(z;W_1),\ t = f_2(z;W_2) \tag{3.1}$$

其中，W_1 和 W_2 是全连接层的权重参数，$f_1(\cdot)$ 和 $f_2(\cdot)$ 是激活函数。然后，将s通过 Sigmoid 函数得到权重系数 $s' \in \boldsymbol{R}^{C}$，即 $s' = \sigma(s)$。接着，将 t 经过 ReLU 函数激活，得到激活输出 $t' \in \boldsymbol{R}^{C}$，即 $t' = relu(t)$。最后，将 s' 和 t' 相乘，得到最终权重 $z' \in \boldsymbol{R}^{C}$，即 $z' = s' \odot t'$。其中，$\odot$ 表示逐元素相乘。最终，将原始特征图 X 和得到的权重 z' 相乘得到加权特征图 $Y \in \boldsymbol{R}^{H\times W\times C}$，即 $Y_{i,j,c} = X_{i,j,c} \times z'_c$，引入 SENet 模块后的 SD-YOLOv7 局部模型图如图 3 所示。

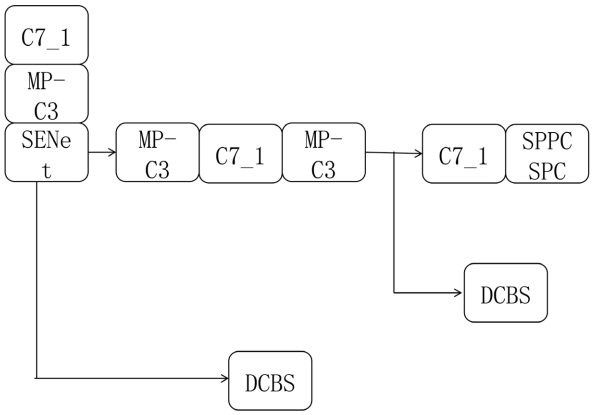

Figure 3. SD-YOLOv7 local model diagram after adding SENet
图 3. 加入 SENet 后的 SD-YOLOv7 局部模型图

3.2. 加入可变形卷积

在深度学习的卷积神经网络中，传统的卷积层通常具有固定的卷积核，这意味着它们以固定的方式从输入数据中提取特征。然而，这种方法在处理具有复杂几何变换的数据时可能不够灵活。为了解决这一问题，在模型中引入了可变形卷积(Deformable Convolution)。

可变形卷积是一种增强版的卷积运算，它通过添加额外的偏移量来允许卷积核适应输入特征的空间变化，从而提供了更好的几何适应性。这些偏移量是可学习的参数，使得卷积核能够自由地变形以更好地对齐和捕捉到目标物体的关键特征，尤其是当物体出现旋转、缩放或其他非刚性变换时[14]。

在 SD-YOLOv7 模型中，加入了可变形卷积 DCNv4 的新的 DCBS 模块替换原模型的 CBS 模块，如图 4 所示，以增强模型对不易捕捉的目标，如手机、香烟、安全带、烟雾等的检测能力。这是因为这些目标可能会在图像中以各种姿态和角度出现，传统的卷积核可能难以捕捉到它们的全部特征。通过可变形卷积的应用，模型显示出了更高的灵活性和准确性。实验结果表明，加入可变形卷积后的 SD-YOLOv7 模型在复杂场景的目标检测任务上具有更强的鲁棒性，特别是在处理不规则形状和姿态变化的目标时。DCNv4 通过两个关键增强解决了其前身 DCNv3 的局限性：去除空间聚合中的 softmax 归一化，增强空间聚合的动态性和表现力；优化内存访问以最小化冗余操作以提高速度[15]。与 DCNv3 相比，这些改进显

著加快了收敛速度，并大幅提高了处理速度，其中 DCNv4 的转发速度是 DCNv3 的三倍以上。DCNv4 在各种任务中表现出卓越的性能，包括图像分类、实例和语义分割，尤其是图像生成。因此，可变形卷积的引入显著提升了模型在工厂阀室安全监控场景中的表现，使其能够更准确地识别和定位潜在的安全风险。

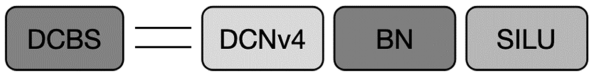

Figure 4. DCBS module after replacing DCNv4
图 4. 替换 DCNv4 后的 DCBS 模块

3.3. FA 损失

YOLOv7 采用了一种复合损失函数来训练模型，以优化检测性能。损失函数包括几个主要部分：坐标损失(L_{loc})、对象置信度损失(L_{obj})和类别损失(L_{cls})。这种设计旨在同时优化模型对目标位置、存在性及其类别的预测能力。通过给这三种损失赋予不同的权重，从而得到总损失。

由于模型对劳保服装、香烟等不同大小的目标进行分散预测，因此为了平衡不同尺度特征之间的关系，防止某些特征单方面决定预测结果，从而使某些特征的预测结果失效。本模型将检测的目标根据大小划分为 tiny、small、medium、large 等级，并且将不同等级的损失函数划分为不同的权重，以经过实验，通过针对不同的特征设计了不同的权重并进行了实验，收敛性最好的 FA Loss 如下：

$$Loss = \lambda_1 L_{cls} + \lambda_2 L_{obj} + \lambda_3 L_{loc} \tag{3.2}$$

$$L_{obj} = 5.6L_{obj}^{tiny} + 3.2L_{obj}^{small} + 1.3L_{obj}^{medium} + 0.5L_{obj}^{large} \tag{3.3}$$

3.4. 实验部分

本文采用的图像处理单元(GPU)为英伟达RTX 8000 GPU，并在rk3399ProD上测试，依赖pytorch1.7.0、python3.7，框架为 SD-YOLOv7。本文共设计了 6 组实验，分别在自制数据集下测试 SD-YOLOv7 与其他主流模型对目标检测性能结果，来综合判断本文提出的改进模型的性能。本文的实验流程图如图 5 所示。

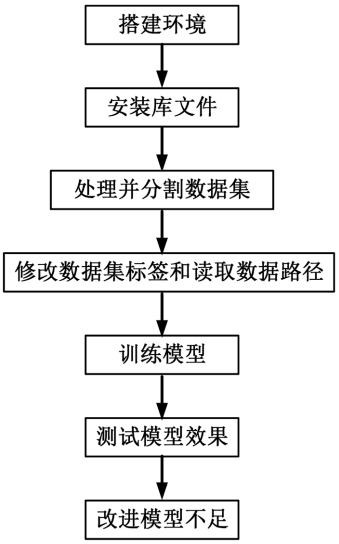

Figure 5. Experimental flow chart
图 5. 实验流程图

3.4.1. 数据集制作

本文数据集为自制的工厂阀室背景下的数据集，数据集共 2 万多张真实工厂阀室下的监控拍摄图片照片。将其按照训练集(验证集)和测试集按照 4:1 的比例进行划分。为了模型的实用性，数据集均来自天然气阀室的真实工作场景。在对异常行为进行分类时，放弃了使用单一目标分类或 CNN 自动特征提取方法，因为这些方法有时会提取出一些人类难以解读的行为特征。在对从业人员进行问卷调查后，将异常行为划分为不同的风险等级。调查包括 60 种危险行为，对象是 150 多名员工。本文对风险等级最高的 8 种异常行为以及三种安全行为进行数据标注形成了 11 类标签的 USDA 数据集。具体分类如下：未着劳保服装、着劳保服装(安全行为)、未戴安全帽、佩戴安全帽(安全行为)、着火、打电话、抽烟、烟雾、未系安全带、系安全带(安全行为)、爬墙。数据集示例如图 6 所示。

Figure 6. Dataset example
图 6. 数据集示例

表 1 展示了不同特征的标签计数。此外，本研究采用了 Github 上可用的开源标记工具，YOLO Mark，以对香烟目标进行数据标注。标注所用的参数格式为一个五元组(class, x, y, w, h)，其中 class 代表待检测目标的种类；(x, y)定位了标注框的中心点坐标；而 w 和 h 分别代表了标注框的宽度与高度。在进行标注之前，需要对 x、y、w、h 的数值进行归一化处理。

Table 1. Number of labels in the data set
表 1. 数据集各标签数量

标签	类别	扩充前标签数	扩充后标签数
No-cloth	未着劳保服	5893	18,042
Cloth	着劳保服	672	3187
No-hat	未戴安全帽	4762	16,212
Hat	戴安全帽	893	4267
Fire	火焰	2541	5203
Phone	手机	1982	4622

续表

Cigarette	香烟	2431	4752
Smoke	烟雾	1209	3381
No-belt	未系安全带	5982	19,391
Belt	系安全带	721	3564
Climb wall	攀爬	994	2254

3.4.2. 数据增强

通常，模型训练是离线训练。因此，本文希望利用这一特点设计出更好的训练方法，在不增加推理成本的情况下获得更高的精度。数据增强通常用于物体检测。增强方法大致可分为两类[16]，一类是增加图像的数量，如 Randomizataflip、RandomCrop；另一类是提高图像的鲁棒性[17]，如 Normalize、CutMix、Mixup 和 Brig-htnessJust。

由于火灾和烟雾的图像较少，标签类型在数据集中表现出数据不平衡。以及对于低照度情况下，特征能见度较低，为了解决这些问题，对数据集使用了一些增强方法来增加火灾和烟雾图像的数量。同时，由于数据集取自监控视频，图像之间存在较大的相似性。因此，使用了 Mosaic 和 CutMix 来提高模型的鲁棒性，使用了高斯模糊来提高模型的泛化能力和抗噪声能力，如图 7 所示，以便于模型在实际应用场景中能够准确识别工业现场的危险行为。接下来介绍本工作中采用的数据增强方法。

CutMix：移除图像的一部分，并将另一幅图像中相同大小的部分添加到原始图像中。

Mosaic：通过数据增强将四幅随机图像拼接在一起。

高斯模糊：将图像进行模糊处理。

Random Horizonta Flip：随机水平翻转图像。

Figure 7. Data enhancement renderings
图 7. 数据增强效果图

3.4.3. 评估指标

为了评估模型的性能，采用了平均精确率(mean average precision, mAP)和召回率(recall)两个指标。召回率计算的是模型正确识别的正例图片数量与实际正例图片总数之比，它的值反映了模型找到所有相关实例的能力。而 mAP 是评估排名任务的常用指标，它计算的是模型在每个查询上的平均精确率，考虑了排名和相关性，因此在检测多类别对象时特别有用。在实验中，mAP 的高值表明模型在不同查询上保持了高精确率，而较高的召回率说明模型能够找到大多数真实正例。

$$\text{Recall} = \frac{TP}{TP + FN} \tag{3.4}$$

$$\text{mAP} = \frac{1}{Q}\sum_{q=1}^{Q}\frac{1}{M_q}\sum_{l=1}^{M_q} P(k) \times rel(k) \tag{3.5}$$

其中，TP 是真正例的数量，FN 是假负例的数量，Q 表示查询数量，M_q 表示第 q 个查询的相关样本的数量，$P(k)$ 表示截至第 k 个样本的准确率，$rel(k)$ 是一个指示函数，表示第 k 个样本是否与查询相关。

3.4.4. 实验结果分析

所有模型都在英伟达 RTX 8000 GPU 上训练，并在 rk3399ProD 上测试。在训练阶段，SD-YOLOv7 很多权重与 YOLOv7 相同，可以加速收敛，从而节省大量训练时间。考虑到图像数量和收敛速度，进行了 87 轮训练，并使用了 SGD 优化器，模型训练如图 8 所示。初始学习率采用余弦退火算法，使学习率在循环中递减，实验参数设置遵循 YOLOv7 官方网站的初始化参数。

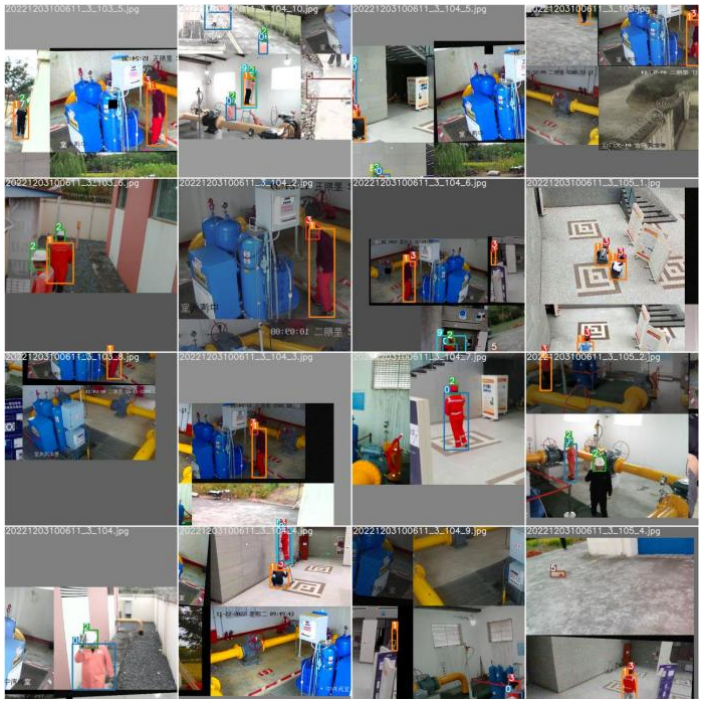

Figure 8. Model training diagram
图 8. 模型训练图

模型 PR 曲线如图 9 所示。PR 曲线是训练结果通过将不同置信度下的 P 值 R 值连点成线所获，从图中可以看出类别“Hat”，“No_Belt”，“Climb_Wall”和“Belt”展示了非常高的精确率和召回率，这意味着模型在检测这些类别时几乎没有错过任何真正例(即这些类别的目标)，同时几乎没有错误地将其他类别的对象错误识别为这些类别。例如，“Hat”类别的精确率接近 1，这表明几乎所有预测为“Hat”的结果确实是佩戴了安全帽。对于“Cloth”和“No_Cloth”类别，观察到它们的精确率和召回率也很高，但略低于“Hat”和“Belt”类别。这可能表明在检测衣物相关的对象时，模型会有少量的错误识别。对于“Phone”和“Smoke”类别的精确率和召回率较低，但仍然处于可接受范围。特别值得注意的是“Cigarette”类别，由于数据集来自于摄像头拍摄的真实环境下的摄像头，受限于摄像头拍摄清晰度和工厂阀室灯光亮度，其精确率显著低于其他类别，这表明模型在预测为“Cigarette”的结果中有大量的假正

例(即被错误标记为香烟的非香烟对象)。“all classes”一项显示了所有类别的平均精确率(mAP)为 0.923，表明在多数类别上模型的整体检测效果是非常好的。然而，“Cigarette”类别的低精确率影响了整体的 mAP 值，这让后续的研究和模型改进中需要特别关注该类别的识别。

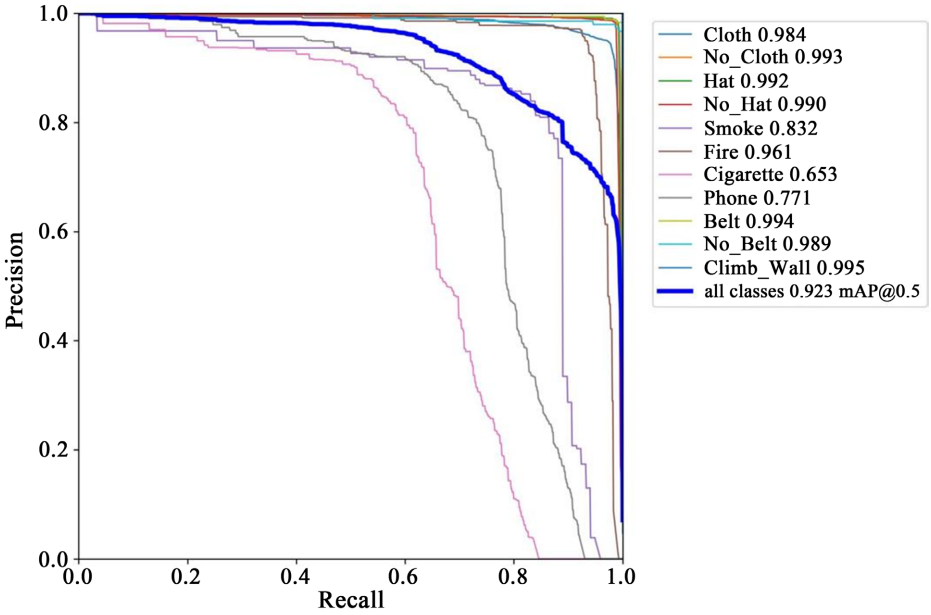

Figure 9. Model PR curve
图 9. 模型 PR 曲线图

为了验证本文所提模型的进展，选择了六种最主流的物体检测模型作为比较基准，包括 Faster-RCNN、DETR、YOLO v5、YOLO v7、YOLO v8 和 YOLO v9，实验结果见表 2。

Table 2. Comparison of experimental results
表 2. 实验结果对比

模型	R	mAP
Faster-RCNN	0.802	0.824
DETR	0.824	0.842
YOLOv5	0.904	0.887
YOLOv7	0.883	0.892
YOLOv8	0.872	0.873
YOLOv9	0.917	0.891
SD-YOLOv7	0.941	0.923

通过表 2 可以看出召回率的提升：SD-YOLOv7 模型在召回率上相较于其他模型实现了显著的提升。特别是与传统的 YOLOv7 模型相比，SD-YOLOv7 的召回率提高了 6.30%，表明其在检测过程中能够识别到更多的真实正样本，减少了漏检的情况。这一提升在工业安全监控中尤为重要，因为漏检可能导致安全隐患的忽视。

(1) 平均精度均值(mAP)的提升：在 mAP 方面，SD-YOLOv7 相较于 YOLOv7 提升了接近 5.82%，与

YOLOv9 相比也有 2.30%的提高。mAP 的提升反映了 SD-YOLOv7 在多类别检测任务中的综合性能得到了增强，这在复杂的工厂阀室环境中尤为关键。

(2) 性能对比：通过与 YOLO 家族的其他模型，包括最新型号 YOLOv9 的对比，SD-YOLOv7 在召回率上仍显示出 4.15%的提高，这进一步证明了 SD-YOLOv7 在实际应用中的有效性和可靠性。

(3) 模型改进点：SD-YOLOv7 在 YOLOv7 的基础上，通过引入 SENet 注意力机制和可变形卷积 DCNv4，显著提高了小目标如手机和香烟的检测性能。这些改进使得 SD-YOLOv7 在多角度和复杂背景下的检测更为准确和鲁棒。

(4) 数据集的贡献：自建的近两万张图像数据集，涵盖了 11 种不同的行为标签，为模型训练和评估提供了丰富的、现实环境下的样本。这些数据集的建立，不仅对本研究至关重要，也对工业安全领域的未来研究具有重要价值。

4. 总结与展望

4.1. 工作总结

本研究针对工业安全监控领域中的关键问题——工厂阀室内的人体异常行为识别，提出了一种改进的 YOLOv7 算法，即 SD-YOLOv7。通过集成 Squeeze-and-Excitation Networks 的注意力机制和可变形卷积 DCNv4，本文提出的模型在复杂监控场景下展现出了卓越的目标检测能力[18]。实验结果表明，SD-YOLOv7 在精准率、召回率和平均精确率 mAP 方面均优于传统的 YOLOv7 和 YOLOv9 模型。此外，模型已成功部署在边缘设备上，为合作单位提供了实时的检测服务，证明了其在实际应用中的有效性和可行性。

4.2. 进一步展望

尽管 SD-YOLOv7 模型在工业安全监控领域取得了显著的成果，但仍存在一些局限性和改进空间[19]。首先，模型在处理极端光照条件和复杂背景时的检测性能仍有待提高。此外，对于某些特定类别的目标，如香烟和烟雾，模型的识别精度尚未达到理想水平，这可能与数据集中这些类别样本的多样性和数量有关。未来研究方向将集中在以下几个关键领域：数据增强和样本平衡[20]、模型鲁棒性[21]、小目标检测[22]、实时性能优化、模型压缩与蒸馏、多模态学习[10]、解释性和可信赖性以及跨场景泛化。通过上述研究方向的深入探索，期望能够进一步提升人体异常行为识别模型的性能，为工业安全监控领域带来更多创新和价值。

基金项目

在本研究的完成过程中，特别感谢国家自然科学基金(批准号：61772342)对本研究项目的资助与支持。

参考文献

[1] Alessandro, B. and Mauro, T. (2023) YOLO-S: A Lightweight and Accurate YOLO-Like Network for Small Target Selection in Aerial Imagery. *Sensors*, **23**, 1865-1865. https://doi.org/10.3390/s23041865

[2] Jia, W., Wang, C., Lin, Q., *et al*. (2022) Adversarial Attacks and Defenses in Deep Learning for Image Recognition: A Survey. *Neurocomputing*, **514**, 162-181. https://doi.org/10.1016/j.neucom.2022.09.004

[3] Du, W., Dash, A., Li, J., *et al*. (2023) Safety in Traffic Management Systems: A Comprehensive Survey. *Designs*, **7**, Article 100. https://doi.org/10.3390/designs7040100

[4] Zhang, Y., Fan, G., Jiang, J., *et al*. (2022) Light-Guided Growth of Gradient Hydrogels with Programmable Geometries and Thermally Responsive Actuations. *ACS Applied Materials & Interfaces*, **14**, 29188-29196. https://doi.org/10.1021/acsami.2c04679

[5] Mu, X. and Antwi-Afari, M.F. (2024) The Applications of Internet of Things (IoT) in Industrial Management: A

Science Mapping Review. *International Journal of Production Research*, **62**, 1928-1952. https://doi.org/10.1080/00207543.2023.2290229

[6] 李丹妮, 栾静, 穆金庆. 基于 YOLOv5 的香烟目标检测算法[J]. 软件导刊, 2023, 22(1): 229-235.

[7] Clinkinbeard, R.N. and Hashemi, N.N. (2024) Supplementation of Deep Neural Networks with Simplified Physics-Based Features to Increase Accuracy of Plate Fundamental Frequency Predictions. *Physica Scripta*, **99**, Article 056010. https://doi.org/10.1088/1402-4896/ad3c77

[8] 胡新荣, 王梦鸽, 刘军平, 等. 基于 Kinect 的人体三维动作实时动态识别[J]. 科学技术与工程, 2020, 20(34): 14133-14137.

[9] 陈俊. 基于手部骨骼点跟踪的大屏交互系统及其去噪算法研究[D]: [硕士学位论文]. 苏州: 苏州大学, 2022. https://doi.org/10.27351/d.cnki.gszhu.2022.002118

[10] 韩晶, 张天鹏, 吕学强. 基于多模态特征与增强对齐的细粒度图像分类[J/OL]. 北京邮电大学学报, 1-6. https://doi.org/10.13190/j.jbupt.2023-140, 2024-04-25.

[11] Sinha, P.K. and Marimuthu, R. (2024) Conglomeration of Deep Neural Network and Quantum Learning for Object Detection: Status Quo Review. *Knowledge-Based Systems*, **288**, Article 111480. https://doi.org/10.1016/j.knosys.2024.111480

[12] Pragadeeswaran, S. and Kannimuthu, S. (2024) Cosine Deep Convolutional Neural Network for Parkinson's Disease Detection and Severity Level Classification Using Hand Drawing Spiral Image in IoT Platform. *Biomedical Signal Processing and Control*, **94**, Article 106220. https://doi.org/10.1016/j.bspc.2024.106220

[13] Xie, Y., Chen, H., Ma, Y., *et al*. (2022) Automated Design of CNN Architecture Based on Efficient Evolutionary Search. *Neurocomputing*, **491**, 160-171. https://doi.org/10.1016/j.neucom.2022.03.046

[14] 游丽萍, 贝绍轶. 结合可变形卷积和注意力机制的目标跟踪方法[J]. 科技与创新, 2024(1): 31-34, 38. https://doi.org/10.15913/j.cnki.kjycx.2024.01.008

[15] 郭宗洋, 刘立东, 蒋东华, 等. 基于语义引导神经网络的人体动作识别算法[J]. 图学学报, 2024, 45(1): 26-34.

[16] Abdellatef, H. and Karam, L.J. (2024) Reduced-Complexity Convolutional Neural Network in the Compressed Domain. *Neural Networks*, **169**, 555-571. https://doi.org/10.1016/j.neunet.2023.10.020

[17] Qi, Q., Xu, Z. and Rani, P. (2023) Big Data Analytics Challenges to Implementing the Intelligent Industrial Internet of Things (IIoT) Systems in Sustainable Manufacturing Operations. *Technological Forecasting & Social Change*, **190**, Article 122401. https://doi.org/10.1016/j.techfore.2023.122401

[18] Nazakat, A., Hussain, M. and Hong, J.-E. (2022) SafeSoCPS: A Composite Safety Analysis Approach for System of Cyber-Physical Systems. *Sensors*, **22**, Article 4474. https://doi.org/10.3390/s22124474

[19] Latif, S., Idrees, Z., Huma, Z., *et al*. (2021) Blockchain Technology for the Industrial Internet of Things: A Comprehensive Survey on Security Challenges, Architectures, Applications, and Future Research Directions. *Transactions on Emerging Telecommunications Technologies*, **32**, e4337. https://doi.org/10.1002/ett.4337

[20] Jung, J.M., Han, D.S. and Kim, J. (2024) Re-Scoring Using Image-Language Similarity for Few-Shot Object Detection. *Computer Vision and Image Understanding*, **241**, Article 103956. https://doi.org/10.1016/j.cviu.2024.103956

[21] 吕俊双. 基于反馈增量学习的鲁棒视觉缺陷检测研究[D]: [硕士学位论文]. 成都: 电子科技大学, 2024. https://doi.org/10.27005/d.cnki.gdzku.2023.001363

[22] 贵向泉, 秦庆松, 孔令旺. 基于改进 YOLOv5s 的小目标检测算法[J]. 计算机工程与设计, 2024, 45(4): 1134-1140. https://doi.org/10.16208/j.issn1000-7024.2024.04.024