基于改进 YOLOv8 复杂街道场景下的红外目标检测算法

洪 俐，曾祥进
（武汉工程大学 计算机科学与工程学院，湖北 武汉 430205）

摘要：针对复杂街道背景下的红外图像因遮挡、缺乏纹理细节等因素而导致目标误检、漏检的问题，提出一种复杂街道场景下的红外目标检测算法。以 YOLOv8n 作为基线模型，首先，通过设计多分支卷积结构，以强化特征提取和特征表达，利用结构重参数化实现训练和推理阶段解耦，提高模型推理速度，同时引入全局自注意力估计来加快注意力的计算，将时间复杂度降为 $O(n)$，使得卷积核注意力实现动态同一。其次，结合深度可分离卷积和可变形卷积的优势，对上采样结果与主干网络的输出特征进行特征融合之后，引入显著信息感知的可变形卷积注意力门控机制，提高融合特征的语义信息丰富度。最后，替换定位损失函数为高效交并比，分别计算预测框和真实框的长、宽影响因子，加速收敛速度。在 Flir 数据集上进行验证实验，改进算法的平均精度均值达到 79.5%，相较于 YOLOv8n 算法提高了 3.9%，验证了所提算法在复杂街道背景下的红外目标检测上的优越性。

关键词：红外目标；街道场景；WIoU；全局自注意力估计；可变形卷积
中图分类号：TP391.4 **文献标识码**：A **文章编号**：1001-8891(2025)05-0591-10

Infrared Target Detection Algorithm Based on Improved YOLOv8 in Complex Street Scenes

HONG Li，ZENG Xiangjin
(*School of Computer Science and Engineering, Wuhan Institute of Technology, Wuhan* 430205, *China*)

Abstract: Aiming at the problem of target misdetection and missed detection in infrared images under complex street backgrounds due to factors such as occlusion and lack of texture details, this paper proposes an infrared target detection algorithm for complex street scenes. Using YOLOv8n as the baseline model, firstly, a multi branch convolutional structure is designed to enhance feature extraction and expression. Structural reparameterization is used to decouple the training and inference stages, improve the inference speed of the model, and global self attention estimation is introduced to accelerate the calculation of attention. The time complexity is reduced to $O(n)$, enabling the convolutional kernel attention to achieve dynamic identity. Secondly, combining the advantages of depthwise separable convolution and deformable convolution, after feature fusion between the upsampling results and the output features of the backbone network, a salient information aware deformable convolution attention gating mechanism is introduced to improve the semantic information richness of the fused features. Finally, An efficient intersection and union ratio replace the localization loss function, calculate the length and width influence factors of the predicted box and the true box separately, and accelerate the convergence speed. Validation experiments were conducted on the Flir dataset, and the average accuracy of the improved algorithm reached 79.5%, which is 3.9% higher than the YOLOv8n algorithm. This validates the superiority of the proposed algorithm in infrared target detection under complex street backgrounds.

Key words: infrared targets, street scenes, WIoU, global self-attention estimation, deformable convolution

收稿日期：2023-12-28；**修订日期**：2024-01-24.
作者简介：洪俐（1998−），男，硕士研究生，研究方向为机器视觉。E-mail：1292286139@qq.com。
通信作者：曾祥进（1977−），男，博士，副教授，硕士生导师。研究方向为智能机器人控制、机器视觉、运动控制。E-mail：xjzeng21@163.com。
基金项目：国家自然科学基金（61502354）；湖北省湖北三峡实验室创新基金（SC215001）。

0 引言

近年来，随着人工智能技术的不断发展，越来越多的领域中都有目标检测的应用。目前许多学者的目标检测主要研究背景为大气可见光环境；而红外光环境下的目标检测研究较少。可见光图像成像容易受周边环境影响，在可见度差的雪天、浓雾和烟尘等恶劣环境下的目标检测十分困难。而红外光在浓雾、暴雪和雾霾严重等恶劣环境下依然能获取目标信息。目前红外图像成像技术被广泛应用于自动驾驶[1-2]、农业生产[3-5]、道路视频监控[6]等各个领域。在自动驾驶领域内，目标检测技术能够帮助车载计算机分析复杂街边道路的情况，减少交通事故的发生。然而，红外图像也存在分辨率低，只有灰度颜色信息、纹理特征少等缺陷，同时街道复杂背景也给红外图像目标检测带来了一定挑战。

基于传统的目标检测方法主要存在两个缺陷：缺乏自适应性和实时性，无法同时满足检测速度和检测精度的要求。主要原因是传统目标检测采用滑动窗口策略和手动选取特征。而深度学习技术能够很好地解决上述问题。基于深度学习的目标检测方法主要分为两种。分别是二阶段检测方法与一阶段检测方法。前者的思想是先提取图像中的特征生成候选框，随后通过算法对生成的不同候选框进行分类和筛选，最后对筛选后的候选框进行回归得到最终的检测结果框，代表方法有 R-CNN（Region-based Convolutional Neural Network）系列[7-9]、FPN（Feature Pyramid Network）等；该类方法检测精度高、效果好但实时性较差。一阶段方法则取消了生成候选框的步骤，改为对图像进行特征提取后直接进行回归预测得到目标位置和类别，由于该步骤会对同一目标区域产生大量检测框，因此后续通过非极大值抑制的方法筛选检测框，代表方法有 YOLO[10-12]、SSD[13]、RetinaNet[14]等；该类方法实时性较好，检测效果也能满足要求。

为了解决红外图像背景复杂、纹理特征不足等问题，Li[15]等人设计了基于混合池化模块的改进快速空间池化金字塔，该方法使用条形池化方法替换传统池化方法，同时在注意力的基础上添加了水平和垂直两个方向上的全局池化操作，实验结果表明，该方法能排除背景干扰拥有较高的检测准确率。Jiang[16]等人为了解决当红外图像中存在较多弱小目标，导致检测性能下降的问题，提出了一种轻量化的红外目标检测模型 YOLO-IDSTD，该模型优化了主干网络的同时增加了改进的感受野增强模块，在红外弱小目标检测任务中得较好的检测速度和准确率。Cai[17]等人基于 YOLOv3 检测网络，通过将跨阶段局部模块、Focus 结构和空间金字塔池化等结构组合以进行特征提取，并且采用多路径聚合思想来优化融合网络，提高特征利用效率。Chen[18]等人基于 YOLOv7 提出一种轻量化红外图像目标检测算法 ITB-YOLO，通过调整网络感受野，改变多尺度融合关系，增大浅层网络特征层权重。将 ELAN 网络部分卷积改进为 PConv，进一步实现网络轻量化目标。

上述研究工作通过不同的方法提升了红外目标的检测效果，但是对比目前的先进算法，其采用的基线模型和改进模块在检测实时性和准确性方面仍有较大改进空间。针对以上问题，基于最新 YOLOv8 算法，本文提出了一种基于全局自注意力估计的多分支卷积结构和显著信息感知的可变形注意力门控机制的复杂街道红外目标检测算法。

1 YOLOv8 网络介绍

目前，YOLOv8 是 YOLO 系列中的最新模型，基于 YOLOv5 进行改进，设计新型网络架构，引入一些优化方法和高效操作，进一步提升模型的性能与扩展性。以深度和宽度为标准，可分为 YOLOv8n、YOLOv8s、YOL0v8m、YOLOv8l 和 YOLOv8x，模型参数量和计算量随着精度的提升大幅度提高，以满足不同场景的需求。由于街道场景下的红外目标检测的实时性需求较高，因此本文选用 YOLOv8n 作为基线模型进行改进。

YOLOv8 一般由输入端、主干网络、颈部网络、预测头 4 个部分组成。输入端，使用自适应图片缩放，以满足不同的输入尺寸，同时使用 mosaic 数据增强来提升模型的鲁棒性。主干网络由 CBS 模块、C2f 模块和 SPPF 模块组成，CBS 模块即卷积、批归一化和 SiLU 激活函数，C2f 模块相较于 C3 模块增加跳层连接和额外的切片操作，使模型的梯度流更丰富，SPPF 模块通过池化和卷积操作进行特征融合，自适应地融合不同尺度的特征信息，从而增强模型的特征提取能力。颈部网络对主干网络输出的有效特征进行深层次提取和融合，使用 PAN 和 FPN 的组合结构以自顶向下与自底向上的跨层连接方式，充分融合深浅层特征。预测头使用解耦头结构，将检测与分类分离，根据分类和回归的分数加权的结果来确定正负样本，以有效提升模型性能。YOLOv8 模型结构如图 1 所示。

2 改进 YOLOv8 的网络结构

本文在 YOLOv8n 的基础上，提出了一种基于全局注意力估计和显著信息感知的复杂街道场景下的红外目标检测模型，如图 2 所示。

首先设计一种基于全局自注意力估计的多分支卷积结构来改进 Backbone 主干结构，增大感受野、增强获取全局上下文的能力。同时在网络中嵌入显著信息感知的注意力门控机制，增强网络的特征融合能力，与初始网络相比，该方法能够适应复杂街道背景下的红外目标检测能力，增强了网络的目标检测效果。主要改进如下：

1）构建一种基于全局自注意力估计的多分支卷积结构，利用像素偏移自注意力方法来估计全局自注意力。该结构在训练阶段分为两个分支，一个是标准卷积分支，另一个是全局注意力估计。通过在自注意机制中引入卷积，能够聚合全局上下文信息，增大模型的感受野，增强了网络的特征提取能力以及强局部依赖，使得特征信息更加丰富和完整；在推理阶段，将两个分支合并，加快推理速度。

2）在上采样结果与主干网络的输出特征进行特征融合之后，引入可变形卷积注意力门控机制，提高模型对融合特征的语义信息丰富度。

3）替换损失函数为 WIoU。削弱几何因素的惩罚，提高模型的泛化能力，并在训练中后期，将较小的梯度增益分配给低质量锚框，以减少有害梯度，提高模型的定位性能。

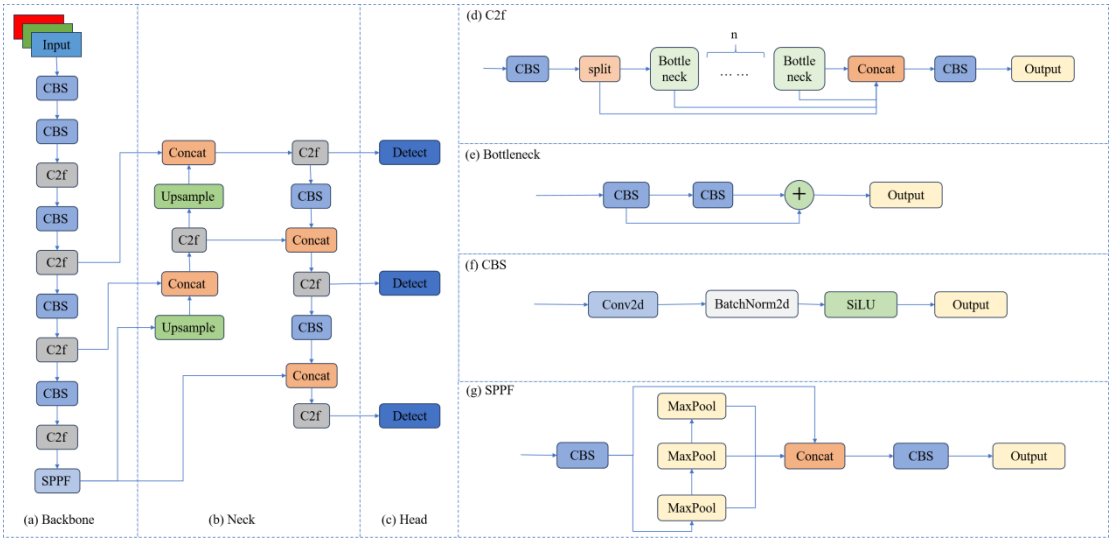


图 1 YOLOv8 网络结构

Fig.1 YOLOv8 network structure

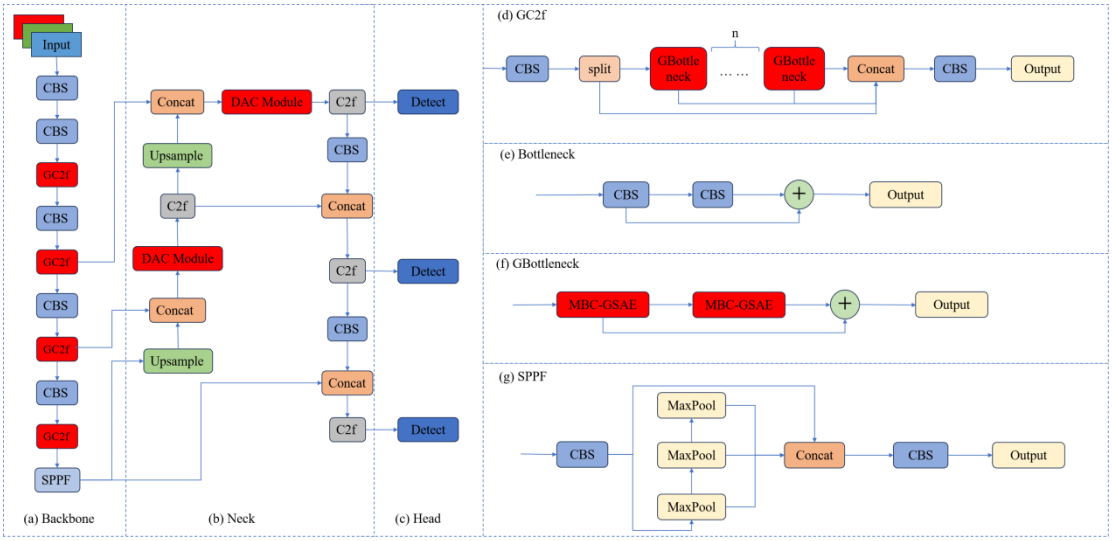


图 2 改进后的 YOLOv8 网络结构

Fig.2 Improved YOLOv8 network structure

2.1 基于全局自注意力估计的多分支卷积结构

在深度神经网络中，卷积操作通常关注输入图像的局部感受野，导致模型难以有效地理解图像的全局上下文信息。与卷积操作相比，自注意力能够处理长距离依赖、上下文信息、强局部相关性等复杂问题。然而自注意力机制需要更长的训练周期和复杂的训练技巧。考虑到这一点，有学者[19-20]提出在自注意机制中引入卷积，以提高其鲁棒性。根据这一思想，本文提出了一种基于全局自注意力估计的多分支卷积结构（Multi-branch convolution based on global self attention estimation, MBC-GSAE）。该结构分为两部分：全局自注意力估计和自注意力与卷积的动态结构。

2.1.1 全局自注意力估计

全局自注意力是最原始的注意力机制，其优势来自全局性。然而，它的计算复杂度为 $O(n^2)$，这使得其在视觉任务中的应用极为有限。为了降低该部分的计算开销，本文利用 COSA（Cell Offset Self-Attention）方法进行全局自注意力估计（Global Self-Attention Estimation）。COSA 方法包含两个部分：相似性计算（特征矢量的点乘）和聚合操作（基于点乘得到的相似性加权聚合全局特征）。

对于输入位置 x_0，其对应的 attention logits 为：

$$s_0=\sum_{x_t\in\Omega}\alpha_t\langle x_0,x_t\rangle=\underbrace{\sum_{x_j\in A}\alpha_j\langle x_0,x_j\rangle}_{\text{Local Region}}+\underbrace{\sum_{x_j\in(\Omega\setminus A)}\alpha_j\langle x_0,x_j\rangle}_{\text{Non-local Region}},\alpha_t=\boldsymbol{w}^q\boldsymbol{w}^k\boldsymbol{w}^v x_t \tag{1}$$

式(1)中：Ω为图像全局区域；A 是以 x_0 为中心的局部区域；$\Omega\setminus A$ 表示Ω中的除去 A 的剩余区域，即非局部区域；$\boldsymbol{w}^q$、$\boldsymbol{w}^k$ 和 $\boldsymbol{w}^v$ 为 3 个线性变换矩阵，$\boldsymbol{w}^q$ 用来寻找与输入位置 x_0 相关的信息、$\boldsymbol{w}^k$ 确定不同位置之间的相关性、$\boldsymbol{w}^v$ 提供每个位置的具体信息。由于图像自身具有较强的马可夫性质，x_0 可以用其局部区域内的像素近似地线性表示：

$$x_0\approx\sum\nolimits_{x_k\in\mathring{A}}\beta_k x_k \tag{2}$$

式(2)中：β_k 为线性权重，$\mathring{A}$ 是局部邻域 A 的外围边界区域。式(2)可替换为：

$$\sum_{x_i\in(\Omega\setminus A)}\alpha_i\langle x_0,x_i\rangle\approx\sum_{x_i\in(\Omega\setminus A)}\alpha_i\left\langle\sum_{x_k\in\mathring{A}}\beta_k x_k,x_i\right\rangle=\sum_{x_i\in(\Omega\setminus A)}\sum_{x_k\in\mathring{A}}\alpha_i\beta_k\langle x_k,x_i\rangle \tag{3}$$

在式(3)中，当前位置的特征 x_0 被局部近似并替换成了特定窗口内的特征组合。为了保持一般性，将非局部区域运算中被排除在外的“局部区域”的系数α设为 0 得到表达式：

$$\sum_{x_i\in(\Omega\setminus A)}\sum_{x_k\in\mathring{A}}\alpha_i\beta_k\langle x_k,x_i\rangle=\sum_{x_i\in\Omega}\sum_{x_k\in\mathring{A}}\alpha_i\beta_k\langle x_k,x_i\rangle \tag{4}$$

由于图像的马尔可夫性质，可以假设对于参考点 x_0（以及附近的位置 x_k）和远离参考点 x_i 的位置的交互是非常弱的，简化上式得到：

$$\sum_{x_i\in\Omega}\sum_{x_k\in\mathring{A}}\alpha_i\beta_k\langle x_k,x_i\rangle=\sum_{x_k\in\mathring{A}}\sum_{x_i\in U(x_k)}\alpha_i\beta_k\langle x_k,x_i\rangle \tag{5}$$

x_i 被进一步聚合到 x_0 的临近点的局部区域 x_k 内。

$$\sum_{x_i\in\Omega}\alpha_i\langle x_0,x_i\rangle\approx\sum_{x_i\in A}\alpha_i\cdot 1\langle x_0,x_i\rangle+\sum_{x_k\in\mathring{A}}\sum_{x_i\in U(x_k)}\alpha_i\beta_k\langle x_k,x_i\rangle=\sum_{x_k\in A}\sum_{x_i\in U(x_k)}\alpha_i\beta_k\langle x_k,x_i\rangle=\sum_{x_k\in A}\beta_k\sum_{x_i\in U(x_k)}\boldsymbol{w}^q\boldsymbol{w}^k\boldsymbol{w}^v x_i\langle x_k,x_i\rangle \tag{6}$$

全局自注意力估计的具体步骤为：首先，对特征图进行空间偏移，沿着给定的 8 个方向各偏移 L 个像素位置；然后，通过两者之间的点乘得到变换特征，在该操作中构建了邻域内的特征点之间的上下文关系，然后对该特征在局部区域进行加权求和，以整合局部特征，最终得到近似自注意力机制处理后的特征图。通过上述的分层堆叠方式，可以不断地将局部的上下文关系传播到全局区域，从而实现全局的自注意力估计。移位、像素点乘积和卷积加权求和的复杂度为 $O(n)$，因此 COSA 是一个具有 $O(n)$时间复杂度的算子。图 3 为 COSA 流程处理图。

2.1.2 自注意力与卷积的动态结构

卷积凭借局部与各向同性的归纳偏置具有平移不变性，然而卷积的局部特性使其无法构建长距离关系。相反，自注意力机制则忽视上述归纳偏置，专注于在没有明确模型假设的情况下从数据集中发现自然模式，使得自注意力有着非常大的自由度来探索复杂关系（比如长距离依赖、各向异性、强局部相关等），但由于缺乏局部先验，该机制需要大规模的训练数据。为了利用不同模型假设，进而同时利用两者的优化特性、注意力范畴以及内容依赖性；本文将卷积引入到自注意力中以改善其鲁棒性，同时在训练阶段和推理阶段结合卷积和自注意力的各自优势，实现二者的动态同一，具体操作为：

1）在训练阶段，所提模块为多分支结构：一个分支为标准卷积，一个分支为本文所提全局自注意力估计。在该阶段，多分支结构模块的公式如下：

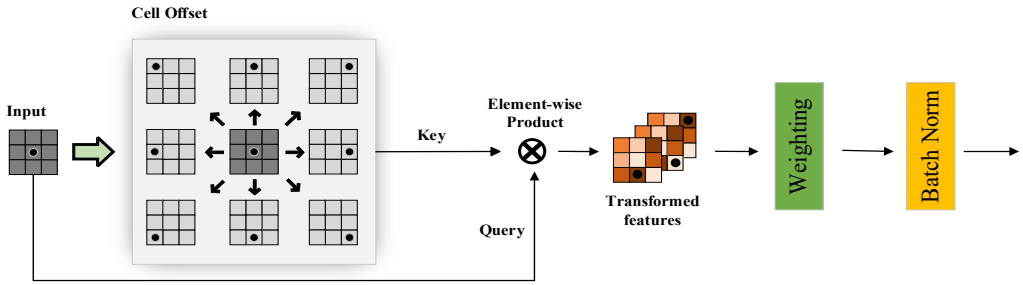


图 3 COSA 流程处理

Fig.3 COSA process processing

$$y_0 = \underbrace{\sum_{x_i \in A} \alpha_i \langle x_0, x_i \rangle}_{\text{COSA Branch}} + \underbrace{\sum_{x_i \in A} w^c x_i + b^c}_{\text{Conv Branch}} = \sum_{x_i \in A} \boldsymbol{w}^q \boldsymbol{w}^k \boldsymbol{w}^v x_i \langle x_0, x_i \rangle + \sum_{x_i \in A} w^c x_i + b^c \tag{7}$$

式(7)中：w^c 表示卷积权值；b^c 表示其对应的偏置值。

2）在推理阶段，将两个分支合并为单一卷积操作。具体公式为：

$$y_0 = \sum_{x_i \in A} \left(\boldsymbol{w}^q \boldsymbol{w}^k \boldsymbol{w}^v \langle x_0, x_i \rangle + w^c \right) x_i + b^c = \sum_{x_i \in A} \left(w^A (x_0, x_i) + w^c \right) x_i + b^c \tag{8}$$

$$w^A (x_0, x_i) = \boldsymbol{w}^q \boldsymbol{w}^k \boldsymbol{w}^v \langle x_0, x_i \rangle$$

式(8)中：$w^A(x_0,x_i)$表示从 COSA Self-Attention 分支计算得到与内容相关的动态系数。w^c 表示从卷积分支继承来与内容无关的静态系数，在训练阶段完成后将被固定。通过该操作，在推理阶段可将多分支结构转化为单一动态卷积算子。该算子权值由需要动态计算的自注意力特征图和经过训练并固定的卷积权值组成。

MBC-GSAE 成功统一了卷积和自注意力，在保持高效性的同时显著提升了模型性能，整体结构如图 4 所示。

2.2 显著信息感知的可变形注意力门控机制

通常注意机制可分为通道注意机制、空间注意机制和通道-空间混合注意机制。注意力模块堆叠方法在计算注意力感知特征图中的注意力权重之前，一般都会采用平均池化操作来进行特定注意力的计算。该部分计算难以提供密集注意力，同时增加了网络整体的计算开销。本文引入 DAC（Deformable Attention for Conspicuous Information）注意力门控机制，通过结合深度可分离卷积和可变形卷积的优势，从整体上能够有效地集中和增加模型对显著图像区域的关注度，在提高计算效率的同时不降低 YOLOv8n 模型性能。因此，在上采样结果与主干网络的输出特征进行特征融合之后，引入可变形卷积注意力门控机制，提高了模型对融合特征的语义信息丰富度，有效改善了复杂街道场景下的红外目标检测。DAC 注意力门机制主要包含深度可分离卷积（Depthwise Separable Convolution，DSC）和可变形卷积（Deformable Convolution, DC）两个方面，如图 5 所示。

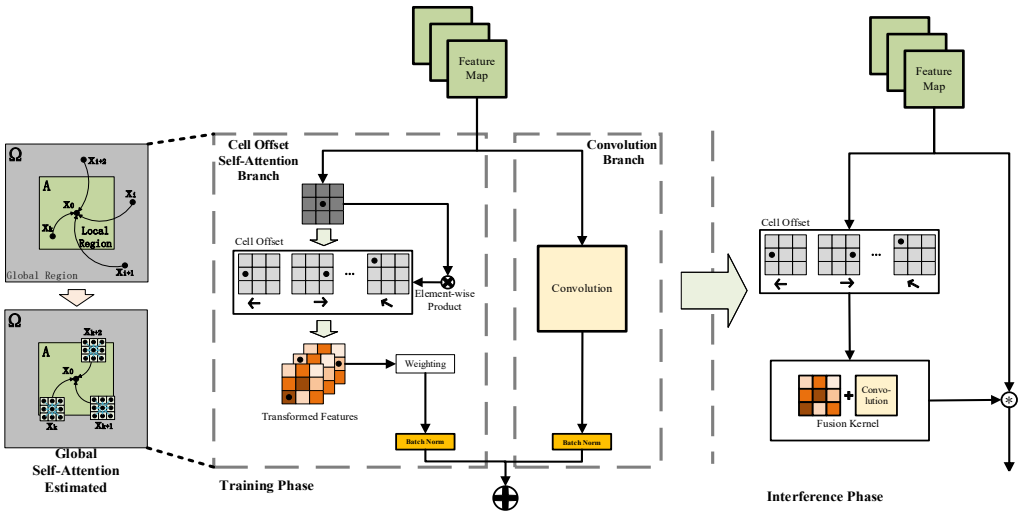


图 4 MBC-GSAE 结构

Fig.4 MBC-GSAE structural diagram

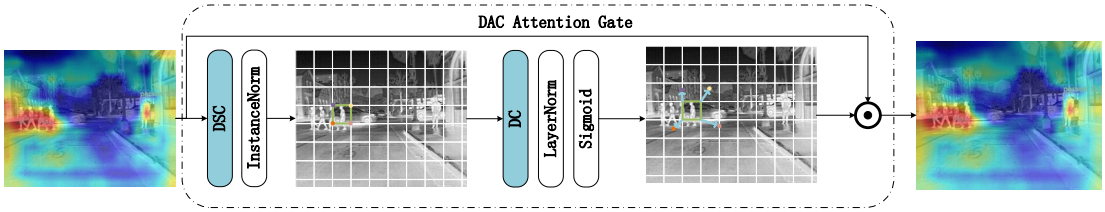


图 5 DAC 结构

Fig.5 DAC structure

2.2.1 通道压缩

本文使用深度可分离卷积操作作为瓶颈层。该操作减少了特征映射中的通道数，将它们从 c 通道转换为 $a\times c$ 通道，其中 $0<a<1$。选择尺寸缩减参数 a 是为了平衡计算效率和精度，默认值为 0.2。在瓶颈层之后，应用一个规范化层，该层包含实例规范化（IntanceNorm，IN）和层规范化（LayerNorm，LN），接着是 GELU 非线性激活。这些操作增强了特征的表征能力，使得注意力机制更加有效。特征压缩表达式如下：

$$X_c = \mathrm{GELU}\left(\mathrm{InstanceNorm}\left(XW_1\right)\right) \tag{9}$$

式(9)中：X 为输入特征，W_1 为深度可分离卷积；X_c 为压缩后的特征数据；GELU 为激活函数。

2.2.2 可变形注意力门控

压缩后的特征数据 X_c 表示特征上下文，然后通过可变形卷积传递该数据，该卷积使用动态网格（偏移量Δp）代替规则网格，从而更加关注相关图像区域。可变形卷积核的操作表达式如下：

$$\mathrm{deform}(p)=\sum_{k=1}^{K} w_k \cdot w_p \cdot X\left(p_{\mathrm{ref},k}+\Delta p_k\right) \tag{10}$$

式(10)中：K 为核的大小，其权值与 YOLOv8 的正则核一样，在 p_{ref} 的固定参考点上施加。Δp 是一个可训练的参数，通过该参数可以找到内核最相关的特征。w_p 是另一个介于 0 和 1 之间的可训练参数。Δp 和 w_p 的值取决于核函数所作用的特征。

在可变形卷积之后，将应用层归一化，然后通过 Sigmoid 激活函数σ，将该卷积操作的通道数从 $a\times c$ 还原为原始输入 c：

$$A=\sigma(\mathrm{LayerNorm}(\mathrm{deform}(X_c))) \tag{11}$$

式(11)中：A 表示注意力门。该门控制来自特征映射的信息流，门张量中的每个元素的值在 0～1 之间。这些值决定了特征映射的哪些部分被强调或过滤掉。最后，将原始输入张量与前一步获得的注意张量进行逐点乘法：

$$X_{\mathrm{out}}=X\odot A \tag{12}$$

式(12)中：⊙表示逐点相乘，X_{out} 是乘法结果，也是下一层 YOLO 模型的输入。将 DAC 注意力门控机制整合到 YOLOv8 模型中并不需要改变 YOLOv8 的主干架构。

2.3 损失函数设计

YOLOv8n 使用的回归损失函数是 CIoU 和 DFL 的组合，其中 CIoU 定义如式(13)所示：

$$I_{\mathrm{CIoU}}=1-\mathrm{IoU} \tag{13}$$

式(13)中：IoU 是预测框与真实框的交并比；$\mathrm{R_{CIoU}}$ 为惩罚项：

$$R_{\mathrm{CIoU}}=\frac{\rho^2(b,b^{\mathrm{gt}})}{c^2}+\tau\nu \tag{14}$$

式(14)中：ρ代表计算两点间的欧氏距离；b、b^{gt} 分别代表预测框和真实框的中心点；c 代表能够同时包含预测框与真实框的最小包围框的对角线距离；τ 是权重系数：

$$\tau=\frac{\nu}{(1-\mathrm{IoU})+\nu} \tag{15}$$

式(15)中：ν是用来度量长宽比的相似性，定义为：

$$\nu=\frac{4}{\pi^2}\left(\arctan\frac{w^{\mathrm{gt}}}{h^{\mathrm{gt}}}-\arctan\frac{w}{h}\right)^2 \tag{16}$$

式(16)中：w、h、w^{gt}、h^{gt} 分别为标注框宽、高和真实框宽、高。

但 CIoU 存在一定缺陷，ν长宽比描述的是相对值，存在一定模糊，并且未考虑难易样本平衡问题，但由于训练数据不可避免地包含低质量示例，几何因素（距离和长宽比）将加重对低质量示例的惩罚，静态聚焦机制不能区分标注框质量高低，模型的泛化性能较弱。本文通过使用 WIoU 利用动态非单调聚焦的梯度增益分配策略，可以在训练的不同阶段，做出最符合当前情况的梯度增益分配策略。在训练早期保存高质量锚框，削弱几何因素的惩罚，使模型获得更好的泛化能力，并在训练中后期，WIoU 将较小的梯度增益分配给低质量锚框，以减少有害梯度，提高模型的定位性能。WIoU 定义如下：

$$L_{\mathrm{WIoU\_v3}}=\gamma\cdot R_{\mathrm{WIoU}}\cdot L_{\mathrm{IoU}} \tag{17}$$

$$R_{\mathrm{WIoU}}=\exp\left(\frac{(x-x_{\mathrm{gt}})^2+(y-y_{\mathrm{gt}})^2}{(W_{\mathrm{g}}^2+H_{\mathrm{g}}^2)}\right) \tag{18}$$

$$\gamma=\frac{\beta}{\delta\alpha^{\beta-\delta}} \tag{19}$$

$$\beta=\frac{L_{\mathrm{IoU}}^{*}}{\overline{L_{\mathrm{IoU}}}}\in[0,+\infty) \tag{20}$$

式中：R_{WIoU} 为预测框和真实框中心点之间的归一化距离，γ 是非单调聚焦系数；W_{g}、H_{g} 为最小包围框的宽、高；α和 δ 是预先设定的常数；β是离群度，用来表示锚框质量的高低，其数值大小与锚框质量高低成反比；L_{IoU}^{*} 为真实边界框的损失；$\overline{L_{\mathrm{IoU}}}$ 为平均边界框的损失。动态非单调聚焦机制利用离群度β评估锚框的质量，该机制可以避免过度关注简单或困难样本，平衡训练过程。

3 实验结果

3.1 数据集

本文采用的是 FLIR 公司提供的红外场景数据集 FLIR_ADAS_1_3[21]，该数据集于 2018 年 7 月发行，包含了 2017 年 11 月～2018 年 5 月美国加利福尼亚州的圣塔芭芭拉市的街道等场景一天内不同时间的红外图片。在这个数据集中含 14452 张热图像，并且含有多种类别，分别为行人（person）、汽车（car）、自行车（bicycle）、狗（dog）等。在本文数据集中狗以及其他类别的数量过少，所以本文只对其余 3 个类别做验证。筛选后的数据集包含 10228 张红外图像。对于训练数据，本次实验除了使用 Mosaic 数据增强以外，还随机使用 Mixup 方法。其主要思想是在每个 epoch 中随机选取两张图像，以一定概率融合生成新图像，从而实现数据扩充，增强模型鲁棒性。

同时，为了验证本文算法的普适性，使用公开数据集 PASCAL VOC2007[22]对本文算法进行测试；该数据集是目标检测的通用数据集，共包含 9963 张图片，20 种类别和 24640 个目标，这些目标大多来自日常生活场景。

3.2 评价指标

本文采用平均精度均值（mean Average Precision, mAP）、个别类别的平均精度（Average Precison, AP）、浮点运算次数（Floating Point Operations, FLOPs）和模型大小（Size）来评价本算法的精度。平均精度指以召回率 R 为 x 轴，查准率 P 为 y 轴所绘制的曲线围成的面积，其公式如式(21)所示：

$$\mathrm{AP}=\int_0^1 P(R)\mathrm{d}R \tag{21}$$

一般来说，一个模型对一个类别的检测率越高，那么 AP 值就越大。但是对于整个数据集进行多分类时，则采用平均精度均值进行评价，平均精度指对数据集中所有类别的 AP 进行叠加后再求其平均值，用来评估模型在整个数据集上检测性能的好坏，其计算公式如式(21)所示：

$$\mathrm{mAP}=\frac{1}{m}\sum_{i=1}^{m}\mathrm{AP}_i \tag{22}$$

式(22)中：AP_i 是不同类别的检测准确率；m 表示检测总类别数。

3.3 实验环境与参数配置

在所有的检测模型中，本文采用训练集占总数据集的 80%、验证集占总数据集的 10%、测试集占总数据集的 10%的比例进行训练。实验参数具体设置为：使用 SGD 优化器在训练时输入图片大小为 512×640，batch size 为 12，初始学习率为 0.01，使用标准 PASCAL-VOC 评价指标，即预测框与真实框的 IoU 大于等于 0.5，所有实验的 epoch 均为 300。本次实验的环境配置如表 1 所示。

表 1 实验环境配置

Table 1 Experimental environment configuration

Name	Environment Configuration
Operating System	Windows10
CPU	Intel 12400F
GPU	NVIDIA RTX 4070 12GB
Framework	Pytorch1.9.0 + CUDA12.2 +cuDNN8.9.6
Languages	Python3.9

3.4 对比实验

为了验证改进 YOLOv8n 算法的有效性，本文将所提算法与目前其他主流的单阶段目标检测模型在 FLIR 数据集上进行性能比较。通过表 2 不同模型的对比实验结果可以看出，本文所提出的 Improved-YOLOv8n 模型与 YOLOv5s、YOLOv7-tiny、YOLOv8n 模型相比，平均精度分别提升了 0.9、1.0 和 3.9 个百分点。与初始 YOLOv8n 相比，在 3 类检测目标中，IMPROVED-YOLOv8n 模型对自行车的检测精度提升最大，提升了 9.5%，说明融入的 MBCS 和 DAC 注意力门控机制对于水平方向上数量较少且有遮挡的自行车目标，具有更强的特征聚合能力和更高的提取能力。与文献[16, 23-24]模型相比，模型大小分别减少 11.5%、68.4%和 39.2%，在综合轻量化指标与平均检测精度二者分析中，IMPROVED-YOLOv8n 模型在实现轻量化的同时，兼顾了模型的检测性能。

为了进一步分析本文算法的综合性能，在目标检测通用数据集 PASCAL VOC2007 上将本文算法与主流深度学习目标检测算法：Faster R-CNN、SSD、YOLOv5s、YOLOv7 和主流传统机器学习目标检测算法：DPM-v5[25]、DPM-CF[26]、Fastest-DPM[27]进行对比实验，结果如表 3 所示。在 VOC2007 数据集中，IMPROVED-YOLOv8n 的 mAP@0.5 分别比 YOLOv8n、YOLOV5s 和 FSSD 高 2.6%、8.0%和 1%，FPS 比 YOLOv8n 仅下降 4 帧，这证明了本文对 YOLOv8n 网络改进的有效性和实时性。同时，本文模型大小比 YOLOv8n 仅增加 0.56 MB，且远小于其他主流深度学习目标检测算法模型。对比主流传统目标检测算法，本文算法的 mAP 相较于 DPM-v5 提升了 147.3%，FPS 相较于传统算法中 FPS 最高的 Fastest-DPM 提升了 263.2%。在 3 种传统机器学习目标检测算法中，FPS 最高为 28.6，未达到 30 FPS 的实时性要求；且 mAP 最高为 32.1。本文模型相较于传统目标检测算法有明显优势。

3.5 消融实验

本文通过消融实验来验证每个模块对本文模型的贡献度。通过分别添加 MBC-GSAE 模块、DAC 模块、WIoU 损失函数到原始模型 YOLOv8n 中，其得到的结果如表 4 所示。检测结果如图 6 所示，从图中可以看出，YOLOv8n 检测到的目标置信度偏低，本文方法可以进行更加准确的定位，检测到的目标具有更高的置信度。

表 2 各实验对比结果

Table 2 Comparison of experimental results

Models	FLOPs/G	Size/MB	AP			mAP(IoU=0.5)/%	FPS
			Car/%	Bicycle/%	Person/%		
YOLOv5s	15.8	13.76	90.3	62.6	83.0	78.6	80.4
YOLO-IDSTD[16]	3.0	7.36	83.1	44.8	72.4	66.8	-
FEID-YOLO[23]	-	20.62	76.5	36.6	58.7	57.3	-
YOLOv7-tiny	13.0	11.72	90.1	61.5	83.8	78.5	108.2
MSC-YOLO	5.9	4.63	89.2	62.3	83.1	78.2	96.3
FS-YOLOv5s[24]	-	10.72	89.1	59.2	81.5	76.6	-
YOLOv8n	8.9	5.96	89.3	56.8	81.3	75.6	117.6
IMPROVED-YOLOv8n	9.6	6.52	90.2	66.3	82.1	79.5	114.1

表 3 不同模型在 VOC 2007 数据集上的对比结果

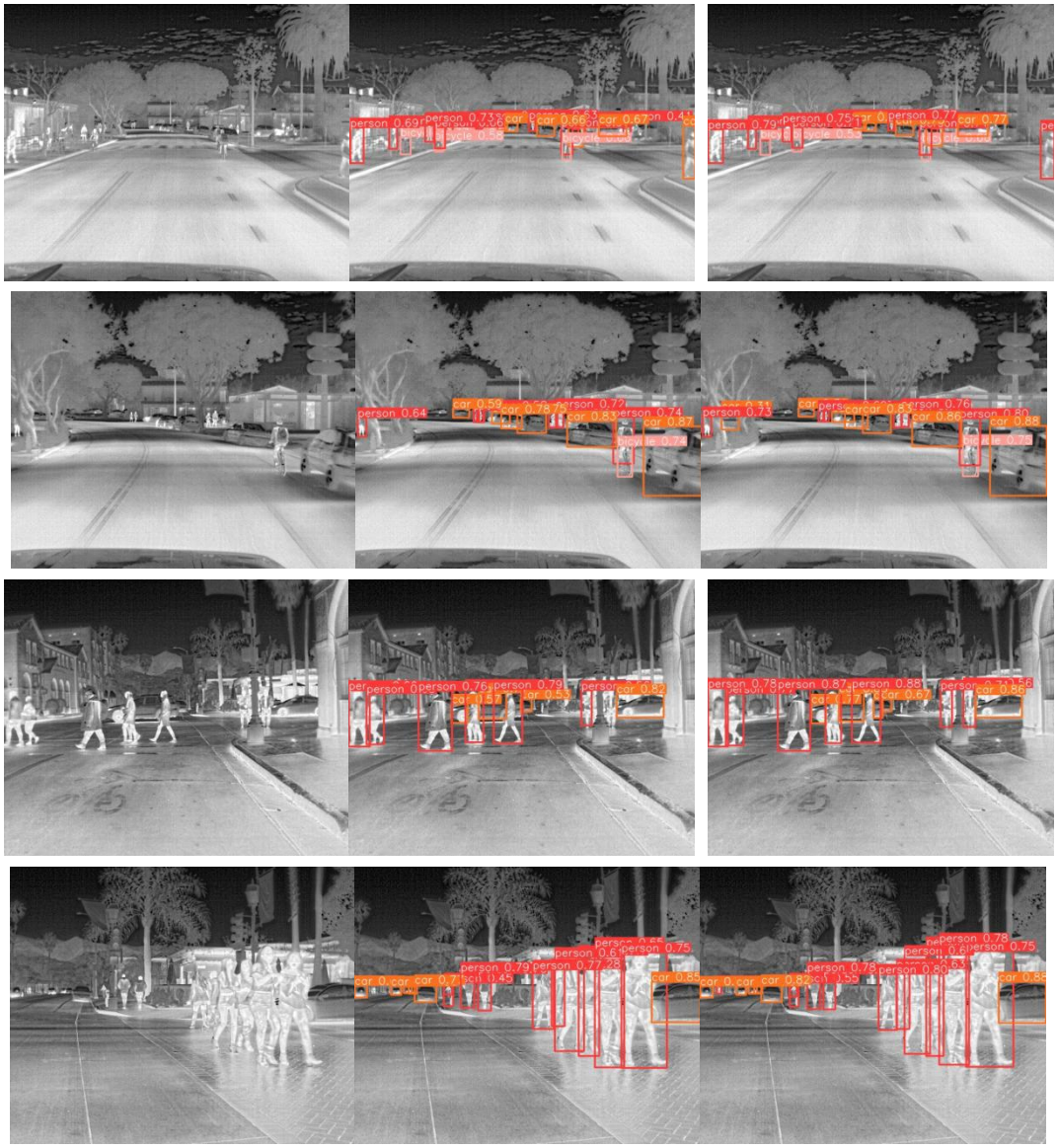
Table 3 Comparison results of different models on the VOC 2007 dataset

Models	Input image size	Size/MB	mAP(IoU=0.5)/%	FPS
DPM-v5[25]	-	-	32.1	0.7
DPM-CF[26]	-	-	30.6	5.2
Fastest-DPM[27]	-	-	30.4	28.6
Faster R-CNN(VGG)	600*1000	462	81.5	13.5
SSD(VGG)	512*512	105.8	77.2	49.5
DSSD(ResNet101)	321*321	490.3	78.4	9.5
FSSD(VGG)	300*300	-	78.6	68.5
YOLOv5s	544*544	28.8	73.5	76.2
YOLOv8n	512*640	5.96	76.8	104.3
IMPROVED-YOLOv8n	512*640	6.52	79.4	100.7

Table 4 消融实验

Table 4 Ablation experiment

Models	MBC-GSAE	DAC	WIoU	Car/%	Bicycle/%	Person/%	$mAP_{0.5}$/%
YOLOv8-n				89.3	56.8	81.3	75.6
	√			89.6	61.7	81.6	77.6
	√	√		89.8	64.9	81.8	78.8
	√	√	√	90.2	66.3	82.1	79.5



(a) Original (b) YOLOv8n (c) IMPROVED-YOLOv8n

图 6 原图、YOLOv8n 以及改进 YOLOv8n 检测结果对比

Fig.6 Comparison of the original image, YOLOv8n and improved YOLOv8n detection results

4 结语

为了提高对红外图像目标的检测精度，基于 YOLOv8 算法，本文提出了基于全局自注意力估计和显著信息感知的注意力门控机制的复杂街道场景下的红外目标检测算法。通过在卷积中引入全局自注意力机制，增强特征的全局相关性和长距离特征。同时，引入显著信息感知的可变形注意力门控机制，通过结合深度可分离卷积和可变形卷积的优势，从整体上能够有效地集中和增加模型对显著图像区域的关注度。最后，替换损失函数为 WIoU，利用动态非单调 FM 的梯度增益分配策略，可以在训练的不同阶段，做出最符合当前情况的梯度增益分配策略。在 FLIR 数据集上的实验结果表明，本文模型对于复杂街道场景下的红外目标拥有较好的检测性能。在未来的工作中，将针对更小的红外点目标进行算法优化，构建检测性更强并且鲁棒性更强的高精度红外目标检测算法。

参考文献：

[1] 楼哲航, 罗素云. 基于 YOLOX 和 Swin Transformer 的车载红外目标检测[J]. **红外技术**, 2022, **44**(11): 1167-1175.
LOU Zhehang, LUO Suyun. Vehicle infrared target detection based on YOLOX and swin transformer[J]. *Infrared Technology*, 2022, **44**(11): 1167-1175.

[2] DAI X, YUAN X, WEI X. TIRNet: Object detection in thermal infrared images for autonomous driving [J]. *Applied Intelligence*, 2020, **51**(3): 1244-1261.

[3] 易诗, 李欣荣, 吴志娟, 等. 基于红外热成像与改进 YOLOV3 的夜间野兔监测方法[J]. **农业工程学报**, 2019, **35**(19): 223-229.
YI Shi, LI Xinrong, WU Zhijuan, et al. Night hare detection method based on infrared thermal imaging and improved YOLOV3[J]. *Transactions of the Chinese Society of Agricultural Engineering*.2019, **35**(19): 223-229.

[4] 刘晓文, 曾雪婷, 李涛, 等. 基于改进 YOLO v7 的生猪群体体温热红外自动检测方法[J]. **农业机械学报**, 2023, **54**(S1): 267-274.
LIU Xiaowen, ZENG Xueting, LI TAO, et al. Automatic detection method of body temperature in herd of pigs based on ilmproved YOLOv7[J]. *Transactions of the Chinese Society for Agricultural Machinery*, 2023, **54**(S1): 267-274.

[5] 刘刚, 冯彦坤, 康熙. 基于改进 YOLO v4 的生猪耳根温度热红外视频检测方法[J]. **农业机械学报**, 2023, **54**(2): 240-248.
LIU GANG, FENG Yankun, KANG XI. Detection method of pig ear root temperature based on improved YOLO v4[J]. *Transactions of the Chinese Society for Agricultural Machinery,* 2023, **54**(2): 240-248.

[6] ZHANG H, LUO C, WANG Q, et al. A novel infrared video surveillance system using deep learning based techniques [J]. *Multimedia Tools and Applications*, 2018: **77**(20): 26657-26676.

[7] Girshick R, Donahue J, Darrell T, et al. Rich feature hierarchies for accurate object detection and semantic segmentation[C]//*Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014: 580-587.

[8] Girshick R. Fast R-CNN[C]//*Proceedings of the IEEE International Conference on Computer Vision*, 2015: 1440-1448.

[9] REN S Q, HE K M, GIRSHICK R, et al. Faster R-CNN：towards realtime object detection with region proposal networks[J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017, **39**(6): 1137-1149.

[10] Redmon J, Divvala S, Girshick R, et al. You only look once: unified, realtime object detection[C]//*Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016: 779-788.

[11] Redmon J, Farhadi A. YOLO9000: better, faster, stronger[C]// *Proceedings of* 2017 *IEEE Conference on Computer Vision and Pattern Recognition*, 2017: 6517-6525.

[12] Redmon J, Farhadi A. Yolov3: An incremental improvement[J/OL]. arXiv preprint arXiv: 1804.02767, https://arxiv.org/abs/1804.02767.

[13] LIU W, Anguelov D, Erhan D, et al. SSD: single shot multibox detector[C]//*Computer Vision–ECCV Proceedings,* 2016: 21-37.

[14] LIN T Y, Goyal P, Girshick R, et al. Focal loss for dense object detection[C]//*Proceedings of the IEEE International Conference on Computer Vision*, 2017: 2980-2988.

[15] 李强龙, 周新文, 位梦恩, 等. 基于条形池化和注意力机制的街道场景红外目标检测算法[J/OL]. **计算机工程**: 1-13, [2023-05-20]. Doi:10.19678/j.issn.1000-3428.0065481.
LI Qianglong, ZHOU Xinwen, WEI Meng'en, et al. Infrared target detection algorithm based on strip pooling and attention mechanism in street scene[J/OL]. *Computer Engineering*: 1-13, [2023-05-20]. Doi:10.19678/j.issn.1000-3428.0065481.

[16] 蒋昕昊, 蔡伟, 杨志勇, 等. 基于 YOLO-IDSTD 算法的红外弱小目标检测[J]. **红外与激光工程**, 2022, **51**(3): 502-511.
JIANG Xinhao, CAI Wei, YANG Zhiyong, et al. Infrared dim and small target detection based on YOLO-IDSTD algorithm[J]. *Infrared and Laser Engineering*, 2022, **51**(3): 502-511.

[17] 陈永麟, 王恒涛, 张上. 基于 YOLO v7 的轻量级红外目标检测算法[J]. **红外技术**, 2024, **46**(12): 1380-1389.
CHEN Yonglin, WANG Hengtao, ZHANG Shang. Lightweight infrared target detection algorithm based on YOLOv7[J]. *Infrared Technology*, 2024, **46**(12): 1380-1389.

[18] 蔡伟, 徐佩伟, 杨志勇, 等. 复杂背景下红外图像弱小目标检测[J]. **应用光学**, 2021, **42**(4): 643-650.
CAI Wei, XU Peiwei, YANG Zhiyong, et al. Dim-small targets detection of infrared images in complex backgrounds[J]. *Journal of Applied Optics*, 2021, **42**(4): 643-650.

[19] WU Haiping, XIAO Bin, Noel Codella, et al. CvT: Introducing convolutions to vision transformers[J/OL]. arXiv:2103.15808, https://doi.org/10.48550/arXiv.2103.15808.

[20] Irwan Bello, Barret Zoph, Quoc Le, et al. Attention augmented convolutional networks[C]// *IEEE International Conference on Computer Vision*, 2019: 3286-3295.

[21] ZHANG H, Fromont E, Lefevre S, et al. Multispectral fusion for object detection with cyclic fuse-and-refine blocks[C]*//IEEE International Conference on Image Processing*, 2020: 276-280.

[22] 邓姗姗, 黄慧, 马燕. 基于改进 Faster R-CNN 的小目标检测算法[J]. **计算机工程与科学**, 2023, **45**(5): 869-877.
DENG Shanshan, HUANG Hui, MA Yan. A small object detection algorithm based on improved Faster R-CNN[J]. *Computer Engineering and Science*, 2023, **45**(5): 869-877.

[23] 郭勇, 张凯. 基于特征增强的快速红外目标检测[J]. **无线电工程**, 2023, **53**(1): 47-55.
GUO Yong, ZHANG Kai. Fast infrared object detection based on feature enhancement[J]. *Radio Engineering*, 2023, **53**(1): 47-55.

[24] 黄磊, 杨媛, 杨成煜, 等. FS-YOLOv5: 轻量化红外目标检测方法[J]. **计算机工程与应用**, 2023, **59**(9): 215-224.
HUANG Lei, YANG Yuan, YANG Chengyu, et al. FS-YOLOv5: lightweight infrared rode target detection method[J]. *Computer Engineering and Applications*, 2023, **59**(9): 215-224.

[25] Girshick R, Felzenszwalb P, FMcAllester D. Object Detection with Discriminatively Trained Part Based Models[J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2010, **32**(9): 1627-1645.

[26] Pedersoli M, Vedaldi A, Gonz`alez J, et al. A coarse-to-fine approach for fast deformable object detection[J]. *Pattern Recognition,* 2015, **48**(5): 1844-1853,.

[27] YAN J, LEI Z, WEN L, et al. The fastest deformable part model for object detection[C]//*Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2014: 2497-2504.