智能系统学报
CAAI TRANSACTIONS ON INTELLIGENT SYSTEMS

结合多尺度大核卷积的红外图像人体检测算法

邵煜潇, 鲁涛, 王震宇, 彭勇杰, 姚巍

引用本文:
邵煜潇, 鲁涛, 王震宇, 等. 结合多尺度大核卷积的红外图像人体检测算法[J]. 智能系统学报, 2025, 20(4): 787-799.
SHAO Yuxiao, LU Tao, WANG Zhenyu, et al. Human detection algorithm in infrared images combining multi-scale large kernel convolution[J]. *CAAI Transactions on Intelligent Systems*, 2025, 20(4): 787-799.

在线阅读 View online: https://dx.doi.org/10.11992/tis.202404027

DOI: 10.11992/tis.202404027
网络出版地址: https://link.cnki.net/urlid/23.1538.tp.20250224.0853.002

结合多尺度大核卷积的红外图像人体检测算法

邵煜潇[1], 鲁涛[2], 王震宇[1], 彭勇杰[1], 姚巍[1]
（1. 华北电力大学 控制与计算机工程学院, 北京 102206; 2. 中国科学院自动化研究所 多模态人工智能系统全国重点实验室, 北京 100190）

摘　要: 针对废墟环境下红外图像人体检测任务中存在的图像分辨率低且人体特征不明显的问题, 基于YOLO 框架设计了一种包含重参数化 (re-parameterization) 和多尺度大核卷积 (multi-scale large kernel convolution) 的红外图像人体检测网络 RML-YOLO(re-parameterization multi-scale large kernel convolution)。该网络通过空间和通道重构注意力模块, 将注意值集中到对检测任务更重要的区域。通过 Sobel 算子强化边缘特征, 提高对不同姿态人体的检测能力。RML-YOLO 的有效性在自制数据集上得到验证。在只有 1.8×10^6 可学习参数的情况下, 模型的 AP_{50} 和 $AP_{50\text{-}75}$ 分别达到了 91.2% 和 87.3%, 与参数量相近的 YOLOv8-n 相比分别提高了 4.4% 和 5.3%。结果表明, RML-YOLO 显著提高了利用红外图像进行废墟环境下人体检测的精度。

关键词: 红外图像; 目标检测; 重构注意力; 多尺度特征; 大核卷积; 卷积神经网络; 特征提取; 重参数化
中图分类号: TP391.4　　**文献标志码**: A　　**文章编号**: 1673−4785(2025)04−0787−13

中文引用格式: 邵煜潇, 鲁涛, 王震宇, 等. 结合多尺度大核卷积的红外图像人体检测算法 [J]. 智能系统学报, 2025, 20(4): 787–799.
英文引用格式: SHAO Yuxiao, LU Tao, WANG Zhenyu, et al. Human detection algorithm in infrared images combining multi-scale large kernel convolution[J]. CAAI transactions on intelligent systems, 2025, 20(4): 787–799.

Human detection algorithm in infrared images combining multi-scale large kernel convolution

SHAO Yuxiao[1], LU Tao[2], WANG Zhenyu[1], PENG Yongjie[1], YAO Wei[1]
(1. The School of Control and Computer Engineering, North China Electric Power University, Beijing 102206, China; 2. The State Key Laboratory of Multimodal Artificial Intelligence Systems, Institute of Automation, Chinese Academy of Science, Beijing 100190, China)

Abstract: Aiming at the problems of low image resolution and inconspicuous human features in the human detection task of infrared images under the ruins environment, an infrared image human detection network re-parameterization multi-scale large kernel convolution(RML-YOLO) is designed based on the YOLO framework, which includes re-parameterization and multi-scale large kernel convolution. The network, RML-YOLO, reconfigures the spatial and channel reconstruction attention module to focus on regions that are more important for the detection task. Edge features are strengthened by the Sobel operator to improve the detection ability of human with different poses. The validity of RML-YOLO is verified on a homegrown dataset. With only 1.8×10^6 learnable parameters, the AP_{50} and $AP_{50\text{-}75}$ of the model reach 91.2% and 87.3%, respectively, which are improved by 4.4% and 5.3% compared with YOLOv8-n with similar number of parameters. The results show that RML-YOLO significantly improves the accuracy of human detection in the ruins environment using infrared images.

Keywords: infrared image; object detection; reconstruction attention; multi-scale feature; large kernel convolution; convolutional neural network; feature extraction; re-parameterization

自然灾害频发给人类生命安全带来巨大威胁。灾后被困人员的搜救工作对于挽救生命损失具有重大意义[1]。目前生命搜救装备普遍体积较大且只能基于废墟表面进行生命探测, 搜索效率低, 难以有效定位掩埋人员[2]; 搜救机器人体型小巧, 可深入废墟内部定位被困人员位置, 成为生命搜救的新型装备[3]。

在废墟内部, 红外图像基于物体热辐射成像, 不受环境中光照和烟尘等条件的影响, 更适合用于废墟环境下的人体检测。但是受制于体积、算

收稿日期: 2024−04−22.　**网络出版日期**: 2025−02−24.
通信作者: 王震宇. E-mail: zywang@ncepu.edu.cn.

力等影响，机器人所携带的红外设备捕获的图像存在图像分辨率低和缺少纹理特征等缺点[4]，这增加了人体检测的难度。传统基于频域滤波[5]或稀疏表示[6]的目标检测算法识别速度慢、泛化能力较弱，在实时检测中难以充分提取目标特征。基于神经网络[7]的单阶段目标检测方法 YOLO (you only look once) 系列[8-14]和两阶段目标检测方法 Faster R-CNN(faster region-based convolutional neural network) 等[15-18]在各种任务中表现出较高的性能。搜救机器人属于边缘设备，且搜救任务对实时性要求较高。因此，轻量的单阶段目标检测是本文研究的关键。

在废墟掩埋环境条件下，人体可能被遮挡且姿态不确定，在红外图像中的特征不够明显。为了增强特征的提取和融合，本文引入了注意力机制以增强网络捕获关键信息的能力。通道注意力最早由 Hu 等[19]通过 SENet(squeeze-and-excitation network) 提出。CBAM(convolutional block attention module)[20]以复杂的方式获取通道注意力的空间信息，将通道和空间注意力以不同的方式组合使用。Li 等[21]提出的 SKNet(selective kernel network) 可以动态调整感受野的大小。Qin 等[22]改进了 SE(squeeze-and-excitation) 模块以获得更强大的表示。不同于以往的工作，本文分别对空间和通道注意力进行重构以实现特征之间的信息关联，在特征提取的不同阶段采用不同的注意力模块，避免了特征冗余。

此外，多尺度特征对于目标检测任务非常重要。He 等[23]提出了空间金字塔池化网络 (spatial pyramid pooling network, SPP-Net)，利用空间金字塔池化层来增强模型的多尺度能力。YOLOv3[9]引入了特征金字塔网络 (feature pyramid network, FPN)[24]和 PAFPN(path aggregation feature pyramid network)[25]，在颈部进行多尺度特征融合。YOLO-MS[26]在目标检测模型中引入大核卷积来学习丰富的多尺度特征。本文同样利用大核卷积解决人体目标的多尺度问题。虽然大核卷积及其变体[27-29]已经取得了优异的性能，但 RepVGG(re-parameterized visual geometry group)[30]和 RepLKNet(re-parameterized large kernel network)[31]表明，大核卷积的潜力还没有被充分发掘，对大核卷积的使用策略进行合理地设计，能够进一步突破卷积神经网络的性能极限。因此，本文不只引入大核卷积，还为其提供一个小内核分支，使非常大的内核能够捕获小规模的模式，且不影响推理速度。

综上，本文针对废墟环境下搜救机器人红外人体搜索需求，提出了一种实时目标检测网络 RML-YOLO(re-parameterization multi-scale large kernel convolution)。本文主要贡献如下：

1）设计了一种多尺度特征提取模块 RML-Block。引入大核卷积并用一个小核重新参数化大核卷积。在 RML-Block 各个分支之间的特征转移过程中加入重构注意力。此外，利用 Sobel 算子增强红外图像边缘信息。RML-Block 的多分支结构能够聚合来自不同层次的特征以增强多尺度表示，以更少的参数提取出更好的人体特征。

2）提出了一种单阶段目标检测网络 RML-YOLO。使用 RML-Block 作为基本特征提取块并添加残差连接[32]。使用大小不同的卷积核，并根据网络的阶段数选择使用空间或通道重构注意力。在特征融合部分，对来自不同层的特征进行加权融合。RML-YOLO 显著提高了红外图像人体检测任务的准确性。

3）制作了废墟环境下的红外图像人体检测数据集 SARHuman。与现有的红外图像行人检测数据集不同，SARHuman 考虑了废墟环境下人体变形、姿态变化和杂物遮挡等特点，为人员搜救相关研究提供数据基础。

1　红外图像人体检测算法

红外图像分辨率低、缺少纹理特征，这会影响检测器的性能。而在废墟环境中，由于人体姿态多变、废墟内部遮挡严重等问题，检测难度大大增加。为了解决这一问题，本文提出了一种实时红外图像人体检测网络 RML-YOLO，具有网络参数量小、推理速度快等特点。

1.1　RML-YOLO 的总体结构

RML-YOLO 的总体结构如图 1 所示，主要分为主干（Backbone）、颈部（Neck）和头部（Head）共 3 部分，图 1 中的 K 表示当前阶段 RML-Block 中卷积核大小。主干部分由 4 个阶段组成，每个阶段各包括一个用于下采样的卷积层和一个 RML-Block，每一阶段的分辨率是前一阶段的一半，通道数增加一倍。随着网络深度的增加，RML-Block 的卷积核大小也逐渐增加，用于提取多尺度特征。最后，通过 SPPF(spatial pyramid pooling-fast) 块[11]进行初步的特征融合。在颈部，通过 PAFPN[25]构建特征金字塔，使多尺度信息之间进行充分的融合。为了让网络能够学习到不同尺度特征的重要程度，特征在赋予可学习的权重之后再融合[33]。颈部的基本构建块为 C2f，将简化后的空间和通道重构注意力插入 C2f 块，实现更好的特征融合和快速推理。头部则与 YOLOv8保持相同。

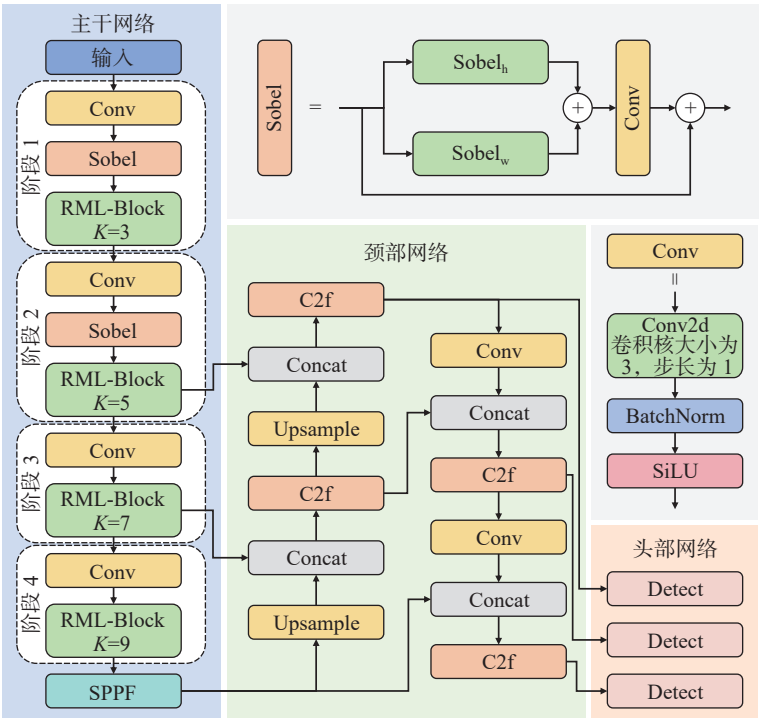


图 1　RML-YOLO 网络结构

Fig. 1　RML-YOLO network structure

1.2　RML-Block

图 1 中所示的 RML-Block 旨在提取多尺度特征。Res2Net[34] 是一个具有多尺度特征提取能力的强大架构，它聚合来自不同层次的特征以增强多尺度表示；YOLO-MS[26] 将大核卷积合并到 Res2Net 中，并用倒置瓶颈层（inverted bottleneck layer）[35] 取代标准的 3×3 卷积来享受大核卷积。

基于 YOLO-MS，如图 2所示，本文提出 RML-Block，充分利用大核卷积提取多尺度特征。

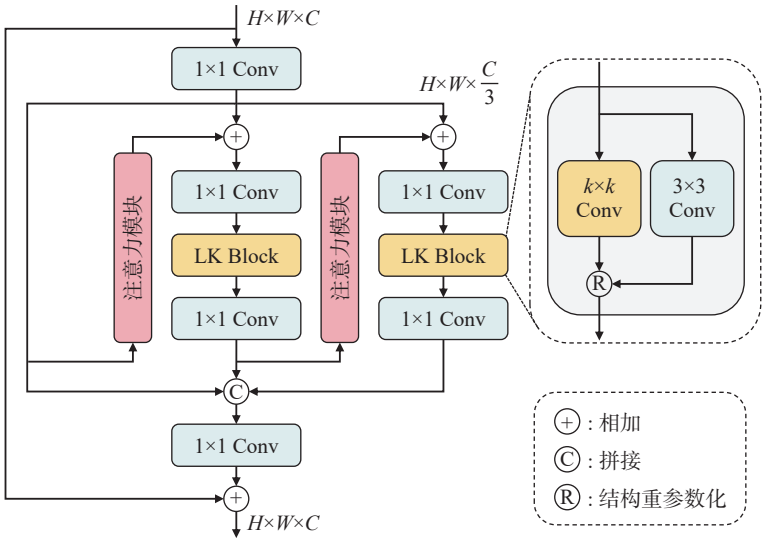


图 2　RML-Block 结构

Fig. 2　Structure of RML-Block

对于输入特征 $\boldsymbol{X}\in\mathbf{R}^{H\times W\times C}$，经过 1×1 的卷积后，在通道维度将其均匀地分为 3 个特征子集 $\boldsymbol{X}_i$，其中 $i\in 1,2,3$。除 $\boldsymbol{X}_1$ 外，每个特征子图 $\boldsymbol{X}_i$ 都要经过一个倒置瓶颈层。倒置瓶颈层首先通过 1×1 的卷积层将特征映射到高维空间，得到更多对象表示；利用大核卷积块（large kernel convolution block，LK Block）处理各通道的特征映射，得到 $\boldsymbol{X}_i\in\mathbf{R}^{H\times W\times 2C}$；最后经过一个 1×1 的卷积层将通道数量还原，这个过程记为 $\mathbf{I}_i(\cdot)$。特征子集 $\boldsymbol{X}_i$ 与 $\mathbf{I}_{i-1}(\cdot)$ 的输出 $\boldsymbol{Y}_{i-1}$ 通过一个空间或通道重构注意力 $A(\cdot)$ 相

加，然后送入$\mathrm{I}_i(\cdot)$，这样可以得到所有处理过后的特征映射。这个过程可以表示为

$$\boldsymbol{Y}_i=\begin{cases}\boldsymbol{X}_i, & i=1\\ \mathrm{I}_i(\boldsymbol{X}_i+A(\boldsymbol{Y}_{i-1})), & i>1\end{cases}$$

所有特征映射拼接起来，利用 1×1 的卷积在特征之间进行交互。最后，通过残差连接关联不同阶段特征，加快模型收敛速度。

简单地增加内核大小会降低精度[36]。因此，本文构建一个与大核卷积平行的 3×3 卷积分支（如图 3 所示）。训练结束后将分支中 3×3 卷积核和 BN 层的参数合并到大核中，这样推理时就没有分支了。

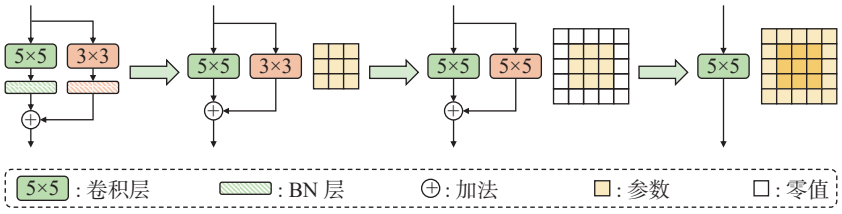


图 3　将小内核（3×3）重新参数化为大内核（5×5）的过程

Fig. 3　Process of re-parameterizing a small kernel (3×3) into a large one (5×5)

1.2.1　Sobel 算子

废墟环境下人体被遮挡情况严重，这个问题在红外图像中更加明显。对于红外图像来说，人体的边缘特征是检测被遮挡人体的重要信息。Sobel 算子是一阶导数的边缘检测算子，它结合了高斯平滑和微分求导，通过计算梯度逼近来找到边缘。本文分别在水平方向和垂直方向上使用 Sobel 算子强化边缘信息，并使用残差增强信息流。这个过程可表示为

$$\boldsymbol{X}_{\mathrm{out}}=\mathrm{F}_3(\mathrm{Sobel}_{\mathrm{h}}(\boldsymbol{X})+\mathrm{Sobel}_{\mathrm{w}}(\boldsymbol{X}))+\boldsymbol{X}$$

式中：$\mathrm{Sobel}_{\mathrm{h}}$和$\mathrm{Sobel}_{\mathrm{w}}$分别为垂直方向和水平方向的 Sobel 操作，$\mathrm{F}_3$为卷积核大小为 3×3 的卷积层。RML-YOLO 只在前 2 个阶段 RML-Block 的输入特征进行 Sobel 操作，主要考虑到浅层网络更关注如边缘、角点等细节信息，而深层的网络更关注语义信息。这样既能够强化网络对边缘细节的敏感度，又能避免边缘信息对语义提取的影响。

1.2.2　空间重构注意力

本文设计了空间重构注意力 (spatial reconstruction attention, SRA)，为每个像素生成一个权重，利用这个权重来引导网络聚焦感兴趣的区域。空间重构注意力的结构如图 4 所示。

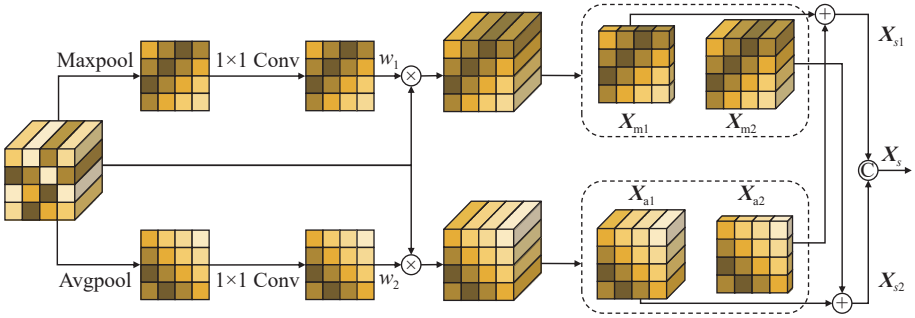


图 4　空间重构注意力

Fig. 4　Spatial reconstruction attention

为了同时利用输入特征的全局和局部信息，首先沿通道维度分别进行平均池化和最大池化操作；然后分别经过一个 1×1 的卷积核，进一步对权重进行调整。将得到的权重分别与特征图相乘，得到 2 个加权后的特征图。这个过程可以表示为

$$\boldsymbol{X}_{\mathrm{a}}=\mathrm{F}_1\left(\mathrm{Avgpool}(\boldsymbol{X})\right)\cdot\boldsymbol{X}$$

$$\boldsymbol{X}_{\mathrm{m}}=\mathrm{F}_1\left(\mathrm{Maxpool}(\boldsymbol{X})\right)\cdot\boldsymbol{X}$$

式中：$\boldsymbol{X}_{\mathrm{a}}$和$\boldsymbol{X}_{\mathrm{m}}$分别为由平均池化和最大池化操作得到的特征图，$\mathrm{F}_1$为卷积核大小为 1×1 的卷积层。最后，为了增强特征的多样性以及全局和局部信息的互补性，将 2 个特征图进行重构操作。具体为，2 个特征图按照通道维度分割得到 4 个子特征图，采用交叉相加和拼接，得到最终的特征图$\boldsymbol{X}_s$。重构的过程可以表示为

$$\boldsymbol{X}_{s1}=\boldsymbol{X}_{\mathrm{m1}}+\boldsymbol{X}_{\mathrm{a2}}$$

$$\boldsymbol{X}_{s2}=\boldsymbol{X}_{\mathrm{m2}}+\boldsymbol{X}_{\mathrm{a1}}$$

$$\boldsymbol{X}_s=\mathrm{concat}(\boldsymbol{X}_{s1},\boldsymbol{X}_{s2})$$

式中$\boldsymbol{X}_{\mathrm{m1}}$、$\boldsymbol{X}_{\mathrm{m2}}$、$\boldsymbol{X}_{\mathrm{a1}}$、$\boldsymbol{X}_{\mathrm{a2}}$为 4 个子特征图。浅层网络的特征分辨率高，保留了更多的空间信息，因此空间重构注意力应用在主干网络前 2 个阶段

的 RML-Block 中。

1.2.3　通道重构注意力

通道重构注意力 (channel reconstruction attention, CRA) 根据 SE 模块[19] 和 SK 模块[21] 进行设计。SE 模块通过全局平均池化压缩空间依赖性来学习通道特定描述符，然后使用 2 个全连接层和一个 sigmoid 函数来重新缩放输入特征映射，以突出显示有用的通道。为了避免较高的模型复杂度，SE 模块的全连接层会减少通道数量，这会导致信息丢失。为了克服这个问题，本文采用 1×1 的卷积核代替减少通道数量的全连接层。在保留信道信息之外，还可捕获通道之间的关系以获得全局表示。SK 模块分别通过核大小为 3×3 和 5×5 的卷积层对输入特征进行处理，并融合特征图，有利于多尺度目标的检测。

通道重构注意力具体结构如图 5 所示。

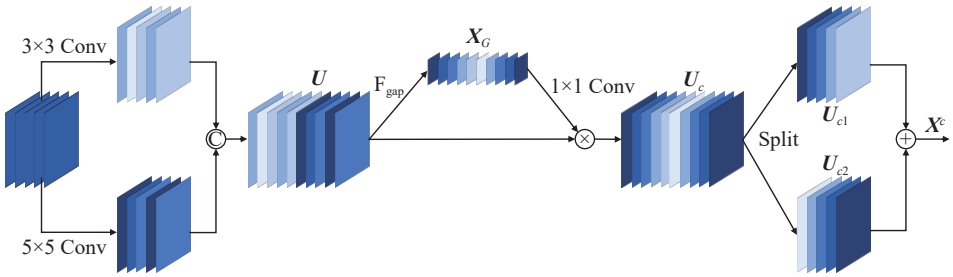


图 5　通道重构注意力

Fig. 5　Channel reconstruction attention

首先利用 3×3 和 5×5 这 2 种大小不同的卷积核分别对输入特征 $\boldsymbol{X}\in\mathbf{R}^{H\times W\times C}$ 进行处理，得到输出向量 $\tilde{\boldsymbol{U}}$ 和 $\hat{\boldsymbol{U}}$；将 2 个向量拼接得到 $\boldsymbol{U}$，然后对 $\boldsymbol{U}$ 进行全局平均池化得到 $\boldsymbol{X}_G\in\mathbf{R}^{1\times1\times2C}$；最后使用 1×1 的卷积层对 $\boldsymbol{X}_G$ 进行处理，获得全局表示。通道注意力也进行重构处理。将得到的权值与 $\boldsymbol{U}$ 相乘，得到通道矫正后的特征映射 $\boldsymbol{U}_c$，然后将这个特征映射在通道维度上拆成两部分并相加，得到最终的特征图 $\boldsymbol{X}_c$。这个过程可以表示为

$$\boldsymbol{U}=\mathrm{concat}(\mathrm{F}_3(\boldsymbol{X}),\mathrm{F}_5(\boldsymbol{X}))$$

$$\boldsymbol{X}_G=\mathrm{F}_{\mathrm{gap}}(\boldsymbol{U})$$

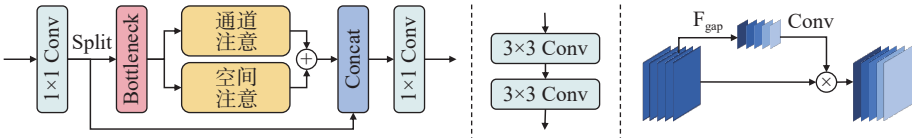
$$\boldsymbol{U}_c=\mathrm{F}_1(\boldsymbol{X}_G)\cdot\boldsymbol{U}$$

$$\boldsymbol{X}_c=\boldsymbol{U}_{c1}+\boldsymbol{U}_{c2}$$

式中 $\mathrm{F}_{\mathrm{gap}}(\cdot)$ 为全局平均池化操作。网络深层特征具有丰富的语义信息，在通道上具有较强的区分性。因此通道重构注意力在主干网络后 2 个阶段的 RML-Block 中使用，让网络更关注语义信息丰富的通道。

1.3　颈部结构

RML-YOLO 颈部采用 PAFPN 构建特征金字塔，依次经过自下而上和自上而下 2 条路径，在每一层都会有一次特征融合。本文在特征融合时增加一个可学习的权重，让不同特征之间的重要性得到动态调整。YOLOv8 颈部的 C2f 作为特征融合块（如图 6(a) 所示），并把空间和通道重构注意力并联后插入到 C2f 中。为了实现快速推理，颈部所使用的通道重构注意力进行简化（如图 6(c) 所示），对输入特征图直接进行全局平均池化，然后通过 1×1 的卷积层获得所有通道之间的关系，再将得到的权值与输入特征图相乘得到校正后的特征向量。



(a) 颈部的 C2f 模块　(b) Bottleneck　(c) 简化后的通道重构注意力

图 6　颈部结构实现细节

Fig. 6　Detail of neck structure realization

1.4　头部结构

头部结构保持和 YOLOv8 相同，2 条并行的分支分别提取类别特征和位置特征，然后用 2 个 1×1 的卷积层完成分类和定位任务。YOLOv8 采用无锚框检测机制，不需要预定义候选框，提高了检测速度。

2　人体检测实验结果与分析

本节对本文所提 RML-YOLO 进行性能验证及分析。首先介绍了使用的数据集和实验设置；

然后，将模型性能与主流方法进行性能比较并且进行了消融实验。

2.1　红外图像人体检测数据集

为了提高实验结果的准确性，本文使用被动红外相机手动拍摄并筛选了 7 700 张图像构建实验数据集。

为了模拟废墟内部光线不足的特点，数据集大部分图像在夜晚拍摄，并利用杂乱岩石环境模拟建筑废墟。每张图像中有 1~3 个人体目标，且姿态不同。图像中的被拍摄目标会被石块等杂物遮挡，以此模拟废墟环境实际情况。由远及近地对目标进行拍摄，保证数据集中会有不同尺度的人体目标。此外，考虑到能够深入废墟搜索的搜救机器人体积不会太大，能够搭载的红外相机分辨率有限，因此采用分辨率为 320 像素×320 像素的菲利尔 (forward-looking infrared imaging, FLIR) 红外相机[37]进行图像拍摄。

为了提高数据集质量，采集的红外图像进行人工筛选。首先删除模糊图像；其次保证不同背景图像数量分布均匀；最终，形成一个包含 5 189 张红外图像的数据集 SARHuman。数据集划分为 3 742 张训练图像和 1 447 张验证图像，训练集和验证集中图像的背景不同。数据集的样本图像如图 7 所示。

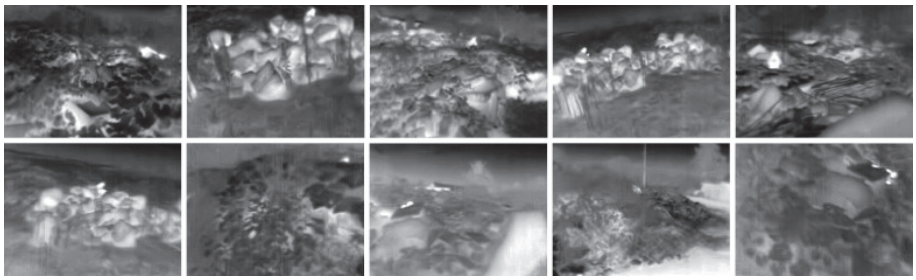


图 7　红外图像人体检测数据集不同背景的样本

Fig. 7　Samples with different backgrounds in the infrared image human detection dataset

2.2　实验设置

RML-YOLO 输入为 320 像素×320 像素的红外图像，对其进行随机水平反转和随机擦除等操作。并采用 Mosaic[10] 数据增强，丰富了图像背景。网络基于 PyTorch 框架构建。考虑到网络的规模和任务的复杂性，初始学习率设为 1×10^{-2}，以加快模型的收敛速度，权值衰减参数设置为 1×10^{-2}，旨在有效防止过拟合现象的发生。模型在 NVIDIA GeForce RTX 4 090 平台上训练，批次大小设置为 64，这一选择考虑了显存限制和训练效率的平衡。模型在验证集上迭代 50 轮后性能若没有改善，训练过程将停止以避免过拟合。

使用以下指标来评估方法性能：

AP_{50}：计算平均精度 (average precision, AP) 时，使用 0.5 作为交并比 (intersection over union, IoU) 的阈值。

$AP_{50\text{-}75}$：IoU 不同阈值精度的平均值。它提供了对不同 IoU 阈值下预测边界框更准确的评估。不同的是，本文将最高阈值设为 0.75（而非 0.95），主要考虑到实际应用时 IoU 阈值为 0.95 的条件过于苛刻，会导致很多合理的检测结果被忽略。

Params：模型的参数量，衡量模型的大小。

FLOPs：浮点运算数，衡量模型的计算复杂度。

Latency：模型处理图像的延时，用于评估模型的实时处理能力。

2.3　算法性能

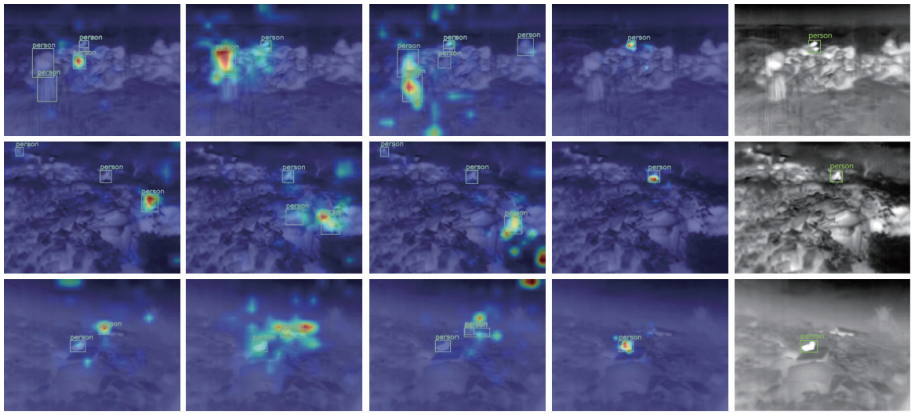
本文将 RML-YOLO 与主流检测网络进行比较。根据网络的类型和大小，本文将各网络分类进行实验。表 1 显示，两阶段目标检测器 Faster R-CNN 检测精度较低，检测速度较慢，难以实现人体检测的实时性。在单阶段检测器中，RML-YOLO 参数数量最少，但是精度却是最高的。在红外图像人体检测任务中，RML-YOLO 达到了 91.2% 的 AP_{50} 和 87.3% 的 $AP_{50\text{-}75}$，分别比 YOLOv8-n 提高了 4.4% 和 5.3%，参数量却只有 YOLOv8-n 的 2/3。即使是与更大的模型相比，如 RTMDet-m 和 YOLOv8-m，性能也占优。在推理速度方面，RML-YOLO 的 Latency 仅比 YOLO 系列的最小规模模型要大一些。上述结果表明，RML-YOLO 有效地针对了红外图像的特点，克服了废墟环境下人体检测的困难。且其轻量化设计，更容易部署到边缘设备上使用。

表 1　与最先进的实时检测器的比较

Table 1　Comparison with state-of-the-art real-time object detectors

模型	AP_{50}/%	$AP_{50\text{-}75}$/%	Params/10^6	FLOPs/10^9	Latency/ms
Faster R-CNN[18]	83.1	79.2	41.6	206.7	6.9
YOLOv5-n[11]	86.1	81.6	2.5	7.1	0.8
YOLOv6-n[13]	86.6	81.2	4.2	11.9	0.8
YOLOv8-n[12]	86.8	82.0	3.2	8.9	0.7
YOLOv5-s[11]	87.0	81.7	9.1	24.0	0.9
YOLOX-s[36]	87.3	81.0	10.8	29.7	1.0
YOLOv6-s[13]	87.2	82.3	16.3	44.1	0.9
YOLOv8-s[12]	87.6	83.2	11.2	28.8	0.9
SSD[38]	84.0	80.1	26.3	101.9	2.8
RTMDet-m[39]	87.4	83.2	28.5	87.7	1.4
YOLOv5-m[11]	87.7	83.0	25.0	64.4	1.3
YOLOv6-m[13]	87.3	83.0	52.0	161.5	1.5
YOLOv8-m[12]	87.3	82.9	25.9	79.3	1.4
YOLOv8-l[12]	89.1	84.2	43.6	165.3	1.8
RML-YOLO	91.2	87.3	1.8	14.7	0.9

利用 Grad-CAM[40] 生成热力图，用来评估检测网络所关注的焦点。结果如图 8 所示，不论人体目标大小或遮挡情况如何，RML-YOLO 对图像中的目标都有较强的响应能力。



(a) YOLOv5　(b) YOLOv6　(c) YOLOv8　(d) RML-YOLO　(e) 原图像

图 8　与 YOLO 系列模型的可视化对比结果

Fig. 8　Visualization comparison results with YOLO series models

此外，为了进一步验证模型的泛化能力，如表 2 所示，在 2 个具有代表性的红外数据集上进行了实验。LLVIP 数据集[41] 专为低光照视觉任务设计，包含 30 976 张可见光与红外配对图像，涵盖 24 个夜间场景和 2 个白天场景，取其中的红外图像进行检测；FLIR 数据集主要应用于自动驾驶领域，包含超过 14 000 张图像，涵盖多种目标类别，考虑到本文的研究重点，仅针对 Person 类进行了检测效果的比较。可以看出，RML-YOLO 在这 2 个数据集上的性能与 YOLO 的 s 系列模型相当，这证明本文模型在普通环境下的红外图像数据集中仍能保持良好的检测效果。

表 2　RML-YOLO 在公开数据集上的检测结果

Table 2　RML-YOLO detection results on public datasets

模型	FLIR/%		LLVIP/%		Params/10^6
	AP_{50}	AP_{50-75}	AP_{50}	AP_{50-75}	
Faster R-CNN	56.9	44.2	95.3	82.1	41.6
YOLOv5-n	81.9	69.8	92.4	80.1	2.5
YOLOv6-n	81.4	69.9	91.0	79.7	4.2
YOLOv8-n	82.2	69.9	92.5	81.1	3.2
YOLOv5-s	84.9	71.8	94.5	81.9	9.1
YOLOv6-s	83.6	70.4	93.5	81.7	10.8
YOLOX-s	85.0	72.3	93.6	80.5	16.3
YOLOv8-s	85.3	72.7	94.9	81.9	11.2
SSD	64.2	50.1	92.6	79.6	26.3
RTMDet-m	85.0	72.2	94.8	81.4	28.5
RML-YOLO	84.8	71.9	94.4	81.0	1.8

2.4　消融实验

本文进行了一系列消融实验验证不同部件的性能，以此进一步评估模型的有效性和可行性。

2.4.1　内核大小的影响以及重参数化分析

首先，将主干部分不同阶段的卷积核大小均设置为 3、5、7 和 9 进行 4 次实验；然后，对不同阶段卷积核大小的组合方式进行了验证，分别采用了 [9,7,5,3] 和 [3,5,9,11] 共 2 种设置。

实验结果如表 3 所示。可以看出，[3,5,7,9] 的设置方式性能最好。简单地增加卷积核的大小，性能并没有明显提升。对于其他组合策略，如果在浅层采用大内核，深层采用小内核，AP_{50} 下降了 4.3%，AP_{50-75} 下降了 4.8%。这表明网络的深层需要大内核来感受全局特征，以检测出大目标。此外，[3,5,7,9] 的设置方式在参数量和计算复杂度方面也有着很大的优势，这表明在特征提取的不同阶段采用不同大小的卷积核，可以提高卷积核的利用率。

表 3　大核卷积不同组合策略的性能比较

Table 3　Comparison of different combinatorial strategies for large kernel convolution in terms of performance

$[k_1,k_2,k_3,k_4]$	AP_{50}/%	AP_{50-75}/%	Params/10^6	FLOPs/10^9
[3,3,3,3]	87.1	82.3	0.9	3.2
[5,5,5,5]	86.1	80.8	1.2	9.9
[7,7,7,7]	89.6	84.5	1.6	19.8
[9,9,9,9]	85.8	79.9	2.0	29.6
[9,7,5,3]	86.9	82.5	1.0	14.7
[3,5,9,11]	88.7	84.0	2.3	19.7
[3,5,7,9]	91.2	87.3	1.8	14.7

为了测试重参数化对于大核卷积的影响，分别将后 3 个阶段 RML-Block 中大核卷积的重参数化分支去掉进行 3 次实验。结果如表 4 所示，加粗的阶段未进行重参数化。结果显示大内核不进行重参数化会降低精度，并且内核越大，精度降得越多。

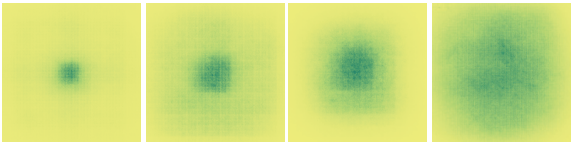
表 4　重参数化的消融研究

Table 4　Ablation study of re-parameterization

模型	AP_{50}/%	AP_{50-75}/%	Latency/ms
[3,**5**,7,9]	90.3	85.8	1.1
[3,5,**7**,9]	89.7	85.5	1.2
[3,5,7,**9**]	88.7	83.7	1.2
本文方法	91.2	87.3	0.9

2.4.2　有效感受野分析

为了说明大核卷积有效性原因，对模型进行有效感受野 (effective receptive field, ERF) 分析。Luo 等[42] 提到，ERF 指特征图某个元素在输入图像上映射区域的大小，要比理论感受野小得多。本文计算了 RML-YOLO 第 4 阶段的 ERF，并且与 YOLOv8 等模型的对应阶段做了对比。图 9 显示，RML-YOLO 的暗区分布最广，有效感受野比其他 3 个模型要大。



(a) YOLOv5　(b) YOLOv6　(c) YOLOv8　(d) RML-YOLO

图 9　RML-YOLO 与其他 YOLO 系列模型在类似参数数量下的有效感受野比较

Fig. 9　Comparison of the effective receptive field between RML-YOLO and other YOLO series models at similar number of parameters

2.4.3　重构注意力模块

通道重构注意力是由 SE 模块和 SK 模块改进而来。如表 5 所示，通道重构注意力模块的 AP_{50} 较 SE 模块提升了 1.6%、较 SK 模块提升了 1.1%。

表 5　通道重构注意力与 SE 模块和 SK 模块的效果比较

Table 5　Comparison of the effects of channel reconstruction attention with SE and SK modules

模型	AP_{50}/%	$AP_{50\text{-}75}$/%	Param/10^6
RML-YOLO(+SE[19])	89.6	86.0	1.7
RML-YOLO(+SK[21])	90.1	85.5	1.8
本文方法	91.2	87.3	1.8

为了验证重构注意力模块对网络特征提取能力的提升，本文将 RML-YOLO 的空间重构注意力 (SRA) 或通道重构注意力 (CRA) 去掉进行消融实验。如表 6 所示，可以看出空间和通道重构注意力都会提升网络性能，将两者同时使用对网络性能的提升最大。

表 6　重构注意力模块的消融研究

Table 6　Ablation study of reconstruction attention modules

SRA	CRA	AP_{50}/%	$AP_{50\text{-}75}$/%	Param/10^6	FLOPs/10^9
		90.1	86.0	1.7	14.2
√		90.3	86.1	1.7	14.2
	√	90.3	86.0	1.8	14.7
√	√	91.2	87.3	1.8	14.7

随后本文将注意力模块的重构过程剔除，与带重构过程的注意力模块进行比较。具体来说，对于空间重构注意力，在得到特征图 $\boldsymbol{X}_m$ 和 $\boldsymbol{X}_a$ 之后，直接将 2 张特征图相加。对于通道重构注意力，去掉特征图拼接后再计算权重的流程。如图 10 所示，分别对 2 张特征图计算权重，将得到的权重分别与对应特征图相乘，再将 2 张特征图相加，此时 2 个特征图的注意力权重是相互独立的。表 7 给出实验对比的结果，重构操作能够提高注意力块的性能。

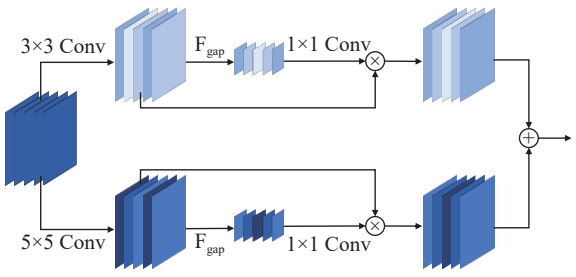


图 10　移除了重构过程的通道注意力

Fig. 10　Remove the reconfiguration process from the channel reconstruction attention

表 7　重构过程对注意力模块性能的影响

Table 7　Impact of the reconstruction process on the performance of the attention module　%

重构过程	AP_{50}	$AP_{50\text{-}75}$
	88.3	84.2
√	91.2	87.3

2.4.4　注意力模块使用策略

在特征提取部分，网络浅层采用空间重构注意力，网络深层采用通道重构注意力。在特征融合块中，空间重构注意力和简化后的通道重构注意力并联后进行使用。表 8 中的 I、II 表示阶段数，从表 8 中可以看出，与只采用空间重构注意力和只采用通道重构注意力相比，本文的使用策略性能最佳。表 9 表明在特征融合阶段，将两者并行使用的效果比将两者串行使用的效果要好。

表 8　骨干网络中注意力组合方式的比较

Table 8　Comparison of attention mixes in backbone

I	II	AP_{50}/%	$AP_{50\text{-}75}$/%	FLOPs/10^9	Latency/ms
S	S	87.1	85.0	13.4	1.0
C	C	86.7	84.3	15.8	1.1
S∩C	S∩C	85.0	80.7	16.0	1.1
S∪C	S∪C	84.8	80.6	16.0	1.3
C	S	85.6	81.5	14.8	0.9
S	C	91.2	87.3	14.7	0.9

表 9　特征融合过程中注意力模块使用策略消融研究

Table 9　Ablation study of attention block usage strategies during feature fusion

模型	AP_{50}/%	$AP_{50\text{-}75}$/%	FLOPs/10^9	Latency/ms
C2f (+S∩C)	88.0	84.2	14.2	1.0
C2f (+C∩S)	88.9	84.9	14.2	0.9
C2f (+S∪C)	91.2	87.3	14.7	0.9

2.4.5　特征融合

对 RML-Block 进行简化，并作为颈部的特征融合块。具体操作为将卷积核更换为 Depth-wise 卷积，卷积核大小设为 3 并去除重参数化过程。如表 10 所示，AP_{50} 和 $AP_{50\text{-}75}$ 分别下降到 88.7% 和 84.1%。由此表明，多分支的结构不利于特征的融合。此外，表 11 的结果表明了可学习权重的重要性。添加可学习权重使 AP_{50} 和 $AP_{50\text{-}75}$ 分别提升 0.2% 和 1.2%。

表 10　简化 RML-Block 和 C2f 特征融合性能比较

Table 10　Comparison of feature fusion performance of simplified RML-Block and C2f

特征融合块	AP_{50}/%	$AP_{50\text{-}75}$/%	FLOPs/10^9	Latency/ms
简化RML	88.7	84.1	15.1	1.1
C2f	91.2	87.3	14.7	0.9

表 11　可学习权重对性能的影响
Table 11　Impact of learnable weights on performance　%

可学习权重	AP_{50}	$AP_{50\text{-}75}$
	91.0	86.1
√	91.2	87.3

2.4.6　Sobel 算子

本文实验比较了 Sobel 算子对网络性能的影响，结果如表 12 所示。罗马数字表示阶段数，√表示该阶段引入了 Sobel 算子。结果显示，在特征提取前 2 个阶段加入 Sobel 算子能够使网络检测的 AP_{50} 和 $AP_{50\text{-}75}$ 分别提高 1.7% 和 0.9%，这表明 Sobel 算子对边缘细节的强化是有效的。但在网络深层引入 Sobel 算子，检测精度会下降，说明 Sobel 算子会干扰网络深层对语义特征的提取。

表 12　Sobel 算子的消融研究
Table 12　Ablation study of the Sobel operator　%

I	II	III	IV	AP_{50}	$AP_{50\text{-}75}$
				89.5	86.4
√				90.4	87.0
√	√	√		88.8	85.6
√	√	√	√	87.0	83.1
√	√			**91.2**	**87.3**

2.5　不同环境下的检测效果分析

除了验证模型内部各个结构的有效性，本文还验证了不同使用环境对检测效果的影响，具体包括遮挡程度不同以及光照强度不同 2 种情况。为了突出展示 2 种环境因素对红外检测的影响，在本节中仅将参数量相近的模型与本方法进行比较。

2.5.1　遮挡程度

验证集图像根据遮挡程度划分为 3 个不同的难度等级，分组情况如表 13 所示。测试 RML-YOLO 在简单、中等以及困难 3 个等级下的检测情况。并与 2.3 节中提到的其他方法进行比较，结果如表 14 所示。

RML-YOLO 在不同难度级别的验证集上展现出了显著的性能优势。与参数量相近的 YOLOv8-n 相比，RML-YOLO 在 3 个不同遮挡程度的验证集上的 AP_{50} 分别提高了 3.5%、4.4% 和 5.4%。这一结果同时表明，随着遮挡程度的增加，RML-YOLO 的性能优势也更加凸显。

表 13　数据集按遮挡程度不同划分的结果
Table 13　Results of dataset segmentation by degree of occlusion

难度	遮挡程度/%	数据量/幅
简单	< 30	531
中等	30~60	455
困难	≥60	461

表 14　各方法在不同遮挡程度下的检测结果
Table 14　Detection results of each method at different levels of shading

模型	简单/%		中等/%		困难/%		Params/10^6
	AP_{50}	$AP_{50\text{-}75}$	AP_{50}	$AP_{50\text{-}75}$	AP_{50}	$AP_{50\text{-}75}$	
Faster R-CNN	88.2	86.3	82.4	78.1	77.9	72.1	41.6
YOLOv5-n	87.4	83.6	86.3	83.3	84.4	77.6	2.5
YOLOv6-n	89.2	83.2	86.3	81.4	83.9	78.7	4.2
YOLOv8-n	88.3	85.6	86.7	83.3	85.2	76.6	3.2
YOLOv5-s	87.4	85.7	87.3	81.4	86.2	77.4	9.1
YOLOv8-s	88.6	87.2	88.1	83.3	85.9	78.5	11.2
SSD	85.6	81.8	83.3	81.0	82.9	78.5	26.3
RML-YOLO	91.8	89.4	91.1	87.4	90.6	84.8	1.8

2.5.2　光照

白天光照强烈的情况下，人体周围环境物体表面的温度会上升。从红外图像上来看，这会导致人体和其他热源在亮度上变得难以区分，具体如图 11 所示。从图 11 中可以看出，周围环境的亮度增加会影响对人体的检测。特别是当人体周围有杂乱的石块等物体时，人体与这些物体之间的特征相似度较高，大大增加了检测难度。为了验证光照强度对废墟环境下人体检测任务的影响，从验证集中筛选出了人体周围存在高亮度杂

物的图像, 并测试 RML-YOLO 和基准模型的性能, 结果如表 15 所示。从表 15 中可以看出, 在光照强烈时, RML-YOLO 的性能优于其他基准模型。与 YOLOv8-n 相比, RML-YOLO 在光照强烈验证集上的 AP_{50} 和 $AP_{50\text{-}75}$ 分别提高了 5.6% 和 7.8%。原因在于利用 Sobel 算子对边缘特征进行强化以及有针对性地设计特征提取的架构, 可以更完整地提取出人体的特征, 从而在周围有相似目标干扰时也能正确地检测出人体。

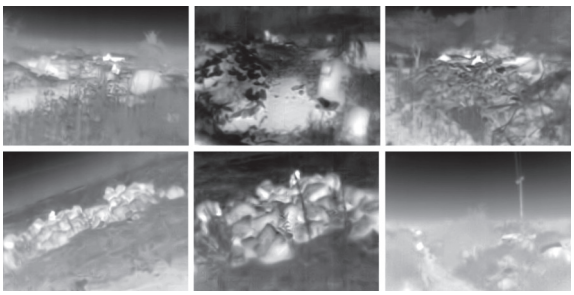


图 11　光照强烈时的数据集图像示例

Fig. 11　Example of an image of the dataset when the light is strong

表 15　光照强度对模型性能的影响

Table 15　Effect of light intensity on model performance

模型	光照强烈数据(592幅)/%		其他数据(855幅)/%		Params/10^6
	AP_{50}	$AP_{50\text{-}75}$	AP_{50}	$AP_{50\text{-}75}$	
Faster R-CNN	80.2	77.3	87.3	81.9	41.6
YOLOv5-n	84.9	79.2	87.8	85.1	2.5
YOLOv6-n	83.9	79.2	90.5	84.1	4.2
YOLOv8-n	84.5	78.3	90.1	87.3	3.2
YOLOv5-s	84.4	77.2	90.8	88.2	9.1
YOLOv8-s	85.1	80.0	91.2	87.8	11.2
SSD	82.3	78.2	86.5	82.8	26.3
RML-YOLO	90.1	86.1	92.8	89.0	1.8

3　结束语

在废墟搜救的背景下, 本文基于 YOLOv8 提出了一种用于红外图像人体检测的模型 RML-YOLO。使用内核大小不同的卷积提取多尺度特征, 解决图像中人体大小不同的问题。为了优化大核卷积的使用, 引入结构重参数化, 为大核卷积添加了并行小核分支。RML-YOLO 使用空间重构注意力和通道重构注意力, 让网络能够关注到有利于人体检测的区域。为了有效检测被遮挡的人体, 在浅层网络引入 Sobel 算子强化边缘特征。实验结果表明, RML-YOLO 的 AP_{50} 和 $AP_{50\text{-}75}$ 分别达到了 91.2% 和 87.3%, 并且参数量小于主要模型(1.8×10^6)。

尽管 RML-YOLO 在小参数量下取得了优异性能, 但是推理速度还略慢于 YOLOv8。在后续的研究中, 可以从以下 3 个方面开展: 1)进一步精简网络, RML-YOLO 的多分支结构会导致推理速度减慢, 可以通过实验找出效果不明显的残差连接并进行删减; 2)在特征融合部分, PAFPN 两条路径的特征融合虽然有着很好的效果, 但是会加深网络的深度, 对推理速度会有影响, 需要在大核卷积和特征融合间做权衡; 3)进一步优化大核卷积的使用, 在不影响精度的情况下, 可将大核卷积的卷积方式替换为 Depth-wise 卷积, 或者用小核卷积通过稀疏或移位等操作来等效大核卷积。

参考文献:

[1] 高荣伟. 人类应对“气候紧急状态”, 需快速强力行动[J]. 世界文化, 2021(4): 4−7.
GAO Rongwei. To cope with the “climate emergency”, human beings need to act quickly and forcefully[J]. World culture, 2021(4): 4−7.

[2] 郑学召, 杨卓瑞, 郭军, 等. 灾后救援生命探测仪的现状和发展趋势[J]. 工矿自动化, 2023, 49(6): 104−111.
ZHENG Xuezhao, YANG Zhuorui, GUO Jun, et al. The current status and development trend of post-disaster rescue life detectors[J]. Journal of mine automation, 2023, 49(6): 104−111.

[3] 苏卫华, 吴航, 张西正, 等. 救援机器人研究起源、发展历程与问题[J]. 军事医学, 2014, 38(12): 981−985.
SU Weihua, WU Hang, ZHANG Xizheng, et al. Rescue robot research: origin, development and future[J]. Military medical sciences, 2014, 38(12): 981−985.

[4] 曲海成, 王宇萍, 谢梦婷, 等. 结合亮度感知与密集卷积的红外与可见光图像融合[J]. 智能系统学报, 2022, 17(3): 643−652.
QU Haicheng, WANG Yuping, XIE Mengting, et al. Infrared and visible image fusion combined with brightness perception and dense convolution[J]. CAAI transactions on intelligent systems, 2022, 17(3): 643−652.

[5] 张铭津, 周楠, 李云松. 平滑交互式压缩网络的红外小目标检测算法[J]. 西安电子科技大学学报, 2024, 51(4): 1−14.
ZHANG Mingjin, ZHOU Nan, LI Yunsong. Smooth interactive compression network for infrared small target detection[J]. Journal of Xidian University, 2024, 51(4): 1−14.

[6] 吴一非, 杨瑞, 吕其深, 等. 红外与可见光图像融合: 统计分析, 深度学习方法和未来展望[J]. 激光与光电子学进展, 2024, 61(14): 42−60.
WU Yifei, YANG Rui, LYU Qishen, et al. Infrared and visible image fusion: statistical analysis, deep learning methods and future prospects[J]. Laser & optoelectronics progress, 2024, 61(14): 42−60.

[7] LECUN Y, BOTTOU L, BENGIO Y, et al. Gradient-based learning applied to document recognition[J]. Proceedings of the IEEE, 1998, 86(11): 2278−2324.

[8] REDMON J, DIVVALA S, GIRSHICK R, et al. You only look once: unified, real-time object detection[C]//IEEE/CVF Conference on Computer Vision and Pattern Recognition. Las Vegas: IEEE, 2016: 779−788.

[9] REDMON J, FARHADI A. YOLOv3: an incremental improvement[EB/OL]. (2018−04−08)[2024−04−01]. https://arxiv.org/abs/1804.02767.

[10] BOCHKOVSKIY A, WANG C Y, LIAO H M. YOLOv4: optimal speed and accuracy of object detection[EB/OL]. (2020−04−23)[2024−04−01]. https://arxiv.org/abs/2004.10934.

[11] JOCHER G. Ultralytics YOLOv5[EB/OL]. (2022−11−22)[2024−04−01]. https://github.com/ultralytics/yolov5.

[12] JOCHER G, CHAURASIA A, QIU Jing. Ultralytics YOLOv8[EB/OL]. (2023−01−22) [2024−04−01]. https://github.com/ultralytics/ultralytics.

[13] LI Chuyin, LI Lu, JIANG Hongliang, et al. YOLOv6: a single-stage object detection framework for industrial applications[EB/OL]. (2022−09−07)[2024−04−01]. https://arxiv.org/abs/2209.02976.

[14] WANG C Y, BOCHKOVSKIY A, LIAO H 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. Vancouver: IEEE, 2023: 7464−7475.

[15] HE Kaiming, GKIOXARI G, DOLLÁR P, et al. Mask R-CNN[C]//IEEE/CVF International Conference on Computer Vision. Venice: IEEE, 2017: 2980−2988.

[16] GIRSHICK R, DONAHUE J, DARRELL T, et al. Rich feature hierarchies for accurate object detection and semantic segmentation[C]//IEEE/CVF Conference on Computer Vision and Pattern Recognition. Columbus: IEEE, 2014: 580−587.

[17] GIRSHICK R. Fast R-CNN[C]//IEEE/CVF International Conference on Computer Vision. Santiago: IEEE, 2015: 1440−1448.

[18] REN Shaoqing, HE Kaiming, GIRSHICK R, et al. Faster R-CNN: towards real-time object detection with region proposal networks[J]. IEEE transactions on pattern analysis and machine intelligence, 2017, 39(6): 1137−1149.

[19] HU Jie, SHEN Li, SUN Gang. Squeeze-and-excitation networks[C]//IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018: 7132−7141.

[20] WOO S, PARK J, LEE J Y, et al. CBAM: convolutional block attention module[C]//European Conference on Computer Vision. Munich: Springer, 2018: 3−19.

[21] LI Xiang, WANG Wenhai, HU Xiaolin, et al. Selective kernel networks[C]//IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2019: 510−519.

[22] QIN Zequn, ZHANG Pengyi, WU Fei, et al. FcaNet: frequency channel attention networks[C]//IEEE/CVF International Conference on Computer Vision. Montreal: IEEE, 2021: 763−772.

[23] HE Kaiming, ZHANG Xiangyu, REN Shaoqing, et al. Spatial pyramid pooling in deep convolutional networks for visual recognition[C]//European Conference on Computer Vision. Zurich: Springer, 2014: 346−361.

[24] LIN T Y, DOLLÁR P, GIRSHICK R, et al. Feature pyramid networks for object detection[C]//IEEE/CVF Conference on Computer Vision and Pattern Recognition. Honolulu: IEEE, 2017: 936−944.

[25] LIU Shu, QI Lu, QIN Haifang, et al. Path aggregation network for instance segmentation[C]//IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018: 8759−8768.

[26] CHEN Yuming, YUAN Xinbin, WANG Jiabao, et al. YOLO-MS: rethinking multi-scale representation learning for real-time object detection[J]. IEEE transactions on pattern analysis and machine intelligence, 2025, 47(6): 4240−4252.

[27] SIMONYAN K, ZISSERMAN A. Very deep convolu-

tional networks for large-scale image recognition[C]//International Conference on Learning Representations. San Diego: ICLR, 2014.

[28] LIU Ze, LIN Yutong, CAO Yue, et al. Swin transformer: hierarchical vision transformer using shifted windows[C]//IEEE/CVF International Conference on Computer Vision. Montreal: IEEE, 2021: 9992−10002.

[29] LIU Zhuang, MAO Hanzi, WU Chaoyuan, et al. A ConvNet for the 2020s[C]//IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans: IEEE, 2022: 11966−11976.

[30] DING Xiaohan, ZHANG Xiangyu, MA Ningning, et al. RepVGG: making VGG-style ConvNets great again[C]//IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville: IEEE, 2021: 13733−13742.

[31] DING Xiaohan, ZHANG Xiangyu, HAN Jungong, et al. Scaling up your kernels to 31 × 31: revisiting large kernel design in CNNs[C]//IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans: IEEE, 2022: 11953−11965.

[32] HE Kaiming, ZHANG Xiangyu, REN Shaoqing, et al. Deep residual learning for image recognition[C]//IEEE/CVF Conference on Computer Vision and Pattern Recognition. Las Vegas: IEEE, 2016: 770−778.

[33] TAN Mingxing, PANG Ruoming, LE Q V. EfficientDet: scalable and efficient object detection[C]//IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle: IEEE, 2020: 10781−10790.

[34] GAO Shanghua, CHENG Mingming, ZHAO Kai, et al. Res2Net: a new multi-scale backbone architecture[J]. IEEE transactions on pattern analysis and machine intelligence, 2021, 43(2): 652−662.

[35] SANDLER M, HOWARD A, ZHU Menglong, et al. MobileNetV2: inverted residuals and linear bottlenecks[C]//IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018: 4510−4520.

[36] GE Zheng, LIU Songtao, WANG Feng, et al. YOLOX: exceeding YOLO series in 2021[EB/OL]. (2021−08−06)[2024−04−01]. https://arxiv.org/abs/2107.08430.

[37] 菲力尔. FLIR ONE Pro 红外热像仪[EB/OL]. (2018−01−01)[2024−04−01]. https://www.flir.cn/products/flir-one-pro/?vertical=condition%20monitoring&segment=solutions.
TELEDYNE FILR. FLIR ONE prothermal imaging camera[EB/OL]. (2018−01−01)[2024−04−01]. https://www.flir.cn/products/flir-one-pro/?vertical=condition%20monitoring&segment=solutions.

[38] LIU Wei, ANGUELOV D, ERHAN D, et al. SSD: single shot MultiBox detector[C]//European Conference on Computer Vision. Amsterdam: Springer, 2016: 21−37.

[39] LYU Chengqi, ZHANG Wenwei, HUANG Haian, et al. RTMDet: an empirical study of designing real-time object detectors[EB/OL]. (2022−12−16)[2024−04−01]. https://arxiv.org/abs/2212.07784.

[40] SELVARAJU R R, COGSWELL M, DAS A, et al. Grad-CAM: visual explanations from deep networks via gradient-based localization[C]//IEEE/CVF International Conference on Computer Vision. Venice: IEEE, 2017: 618−626.

[41] JIA Xinyu, ZHU Chuang, LI Minzhen, et al. LLVIP: a visible-infrared paired dataset for low-light vision[C]//IEEE/CVF International Conference on Computer Vision Workshops. Montreal: IEEE, 2021: 3489−3497.

[42] LUO Wenjie, LI Yujia, URTASUN R, et al. Understanding the effective receptive field in deep convolutional neural networks[C]//Neural Information Processing Systems. Long Beach: MIT Press, 2016: 29.

作者简介：

邵煜潇，硕士研究生，主要研究方向为计算机视觉和模式识别。E-mail：yx_shao@ncepu.edu.cn。

鲁涛，副研究员，主要研究方向为智能机器人控制、人机交互、操作技能学习、模仿学习。发表学术论文 50 余篇，授权国家发明专利 20 项。E-mail：tao.lu@ia.ac.cn。

王震宇，教授，博士生导师，主要研究方向为模式识别、计算机视觉。主持国家自然科学基金等科研项目 5 项，2019 年获吴文俊人工智能科学技术奖。E-mail：zywang@ncepu.edu.cn。