基于多模态可见光和红外图像融合的船舶检测方法

于梦源，刘向阳
（河海大学数学学院，江苏 南京 211100）

摘　要： 单一模态图像在全天候的船舶检测中易受光照、天气等环境影响，导致船舶检测精度低、漏检率高。为此，提出了一种融合可见光与红外图像信息的船舶检测方法 VIF-RTDETR。该方法根据可见光图像丰富的细节和颜色信息以及红外图像在低光照环境下的稳定表现，构建了四通道输入模型；设计可见光与红外图像信息的融合模块 VIF，实现了不同模态信息的互补融合，使得在检测网络中更加合理利用两种模态的信息；在主干 Backbone 特征提取网络中结合通道注意力，为通道动态分配不同的权重，以增强通道的特征表达能力来进一步优化特征提取能力。此外，为进一步提升船舶检测中船舶小目标的检测性能，设计了一种加权的边界框损失函数，使模型能够有效地关注不同尺寸目标的特征表达，提高模型在不同目标尺寸下的检测精度。实验结果表明，在船舶可见光和红外数据集上，该模型的检测精度 $AP_{0.5:0.95}$、$AP_{0.5}$ 分别达到了 78.3%、98.5%，相对于单一模态的可见光和红外模型的 $AP_{0.5:0.95}$ 分别提升了 4.7、9.2 百分点；召回率 $AR_{0.5:0.95}$ 达到了 85.2%，相对于单一模态模型分别提升了 3.1、7.3 百分点，显著提高船舶的检测精度且降低漏检情况。

关键词： 船舶检测；可见光和红外图像；全天候检测；双模态融合；注意力机制
中图分类号： TP391　　**文献标志码：** A　　**DOI：** 10.19678/j.issn.1000-3428.0070436

Ship Detection Method Based on Multimodal Visible and Infrared Image Fusion

YU Mengyuan, LIU Xiangyang
(School of Mathematics, Hohai University, Nanjing 211100, Jiangsu, China)

【Abstract】 Single-modal images are easily affected by light, weather, and other environmental conditions in all-weather ship detection. This leads to a low ship detection accuracy and high leakage rate. To address these issues, this paper proposes a ship detection method, VIF-RTDETR, which fuses visible light and infrared image information. The method fully utilizes the rich details and color information of visible images and the stable performance of infrared images in low-light environments, and constructs a four-channel input model. The complementary fusion of varied modal information is realized by designing the fusion module VIF such that it makes more reasonable use of the information from the two modalities (visible light and infrared) in the detection network. The channel attention in the backbone feature extraction network is combined to further optimize the feature extraction capability by dynamically assigning different weights to the channels, thereby enhancing the feature expression capability of the channels. To further enhance the detection performance of small targets in ship detection, a weighted bounding box loss function is designed so that the model can effectively focus on the feature expression of targets of different sizes and improve the detection accuracy under different target sizes. The experimental results show that in the visible and infrared datasets for the ships, the detection precision $AP_{0.5:0.95}$, $AP_{0.5}$ of the model reaches 78.3% and 98.5%, respectively, reflecting improvements by 4.7 and 9.2 percentage points relative to $AP_{0.5:0.95}$ the single-modal visible and infrared models. Further, the recall rate $AR_{0.5:0.95}$ reaches 85.2%, reflecting improvements by 3.1 and 7.3 percentage points relative to the single-modal visible and infrared models, respectively. Thus, the findings contribute to significantly improving the precision of ship detection and reducing the leakage rate.

【Key words】 ship detection; visible and infrared images; all-weather detection; multimodal fusion; attention mechanism

0　引言

近年来，随着海洋经济的不断发展，海洋监控和海事安全[1]等领域对港口和航道监测的需求日益增长。船舶作为海洋与河流的主要运输载体，是海洋经济的重要组成部分。船舶的正确检测[2]有助于保障航运安全和海洋经济事业，同时能为海洋运输以及监控船舶活动提供一定的技术保证。因此，对船

基金项目： 云南省重大科技专项(202002AE090010)。
作者简介： 于梦源，女，硕士，主研方向为目标检测、图像处理；刘向阳(通信作者)，副教授、博士。
收稿日期： 2024-10-08　　**修回日期：** 2025-01-06
E-mail： liuxy@hhu.edu.cn

舶进行精准定位和检测识别具有重要意义。

传统船舶检测方法的特点是通过手工提取图像特征,对船舶进行分类和定位,如 YKSEL 等[3]通过轮廓图像提取了船舶特征,但在环境背景复杂、船体差异小的情况下效果不理想。随着深度学习技术的广泛应用,基于深度学习的目标检测算法成为研究热点。与传统方法相比,深度学习算法无需人工手动提取特征,为船舶检测[4]提供了技术支持。目前,ZHAO 等[5]提出了船舶的嵌入式检测系统,将基于深度学习的目标检测算法应用在实际场景下的船舶检测。DONG 等[6]提出了基于注意力机制的遥感船舶图像的目标检测算法,通过空间注意力信息检测船舶目标。HAN 等[7]提出对于特征的多尺度提取模块,增强了算法对图像特征的提取能力,进一步提高了检测精度。牛为华和郭迅[8]在 YOLOv8 下通过改进卷积模块更精确地提取不规则形状物体的特征,增强了检测的精度。由于 Transformer 的高性能以及自身特点,因此研究者尝试将 Transformer 引入到计算机视觉领域。DETR 于 2020 年被 CARION 等[9]提出,为在目标检测任务中应用 Transformer 技术提供了重要的基础,其后,涌现出了许多 DETR 系列算法[10-16]被成功应用在目标检测任务中。

大多数船舶检测方法主要依赖于单一的可见光或红外图像[17],目前已有学者结合不同模态图像进行融合对船舶进行检测。如 LU 等[18]融合自动识别系统(AIS)和视觉图像,通过识别船舶目标与相机之间的距离,达到更准确的检测。赵炜东等[19]在 YOLOv5[20]基础上融合可见光和红外卫星图像,通过空间特征与谱段特征的联合提取来提高检测性能。可见光图像[21]包含丰富的色彩和细节信息,但易受光照、天气等环境因素的影响,如在大雾、夜晚等场景下无法获得准确的检测结果。红外图像[22]由红外传感器探测物体的热辐射对物体进行成像,能够在低光照和恶劣天气(如大雾、雨雪)条件下拥有较好的成像效果,然而红外图像缺少颜色和细节信息,在船舶检测中也有一定的局限性。

为克服这些局限性,本文提出一种充分利用船舶可见光和红外的特征信息进行船舶检测的方法。由于 RT-DETR 模型具有优越的特征交互性和准确性,因此本文在当前较先进的目标检测算法 RT-DETR 基础上进行改进,进一步实现可见光和红外图像信息的互补来检测船舶。本文的具体工作为:1)构建同一时刻船舶的可见光和红外图像数据集,将可见光和红外图像同时作为模型的输入,根据两种模态的互补信息,从而减小不同光照和场景变化对检测性能的影响;2)为了实现全天候的船舶检测,设计了可见光与红外图像信息融合(VIF)模块,该模块实现了模态间的互补融合,确保了两模态特征信息的合理利用;3)在主干 Backbone 提取网络中结合通道注意力机制,通过模型训练动态更新通道权重,进一步提高特征提取能力;4)针对船舶小目标识别不精确的问题,提出了赋有权重的边界框损失,增强模型对不同面积大小目标的敏感度。本文通过这种信息融合的方法在模型检测中提供了更全面的信息,能够实现更精确和可靠的船舶检测,为海事安全等应用提供技术支持。

1 RT-DETR 算法模型结构

RT-DETR 模型[23]的提出是基于 Transformer 的目标检测算法 DETR,该算法模型简化了目标检测的流程,实现了端到端的目标检测。RT-DETR 模型从以下两点对 DETR 模型进行了改进:1)设计一种高效的混合编码器来取代原来的 Transformer 编码器,通过解耦多尺度特征的尺度内交互和尺度间融合,使得编码器更有效地处理不同尺度的特征;2)采用交并比(IoU)感知的查询选择机制来选择固定数量的图像特征以优化解码器查询的初始化。

RT-DETR 模型结构如图 1 所示(彩色效果见《计算机工程》官网 HTML 版,下同)。该模型将主干特征提取网络最后三个阶段的输出{S3,S4,S5}作为编码器的输入,再通过混合编码器进行内尺度交互(AIFI)和跨尺度特征融合模块(CFFM)操作,提高多尺度特征提取能力。其中,AIFI 是指将 S5 经过 Transformer 编码器进行特征提取后再与 S3、S4 进行特征融合。CFFM 是指将多尺度特征进行上下交互特征融合。随后,IoU 感知的查询选择机制选取固定数量的图像特征作为解码器的初始查询,最终生成目标边框和置信度。该模型的混合编码结构和 IoU 感知的查询选择机制能对图像和目标信息进行丰富的特征提取及预测,将这个模型用于船舶检测,以进一步提高性能和检测效果。

2 基于双模态可见光与红外融合的船舶检测

针对可见光图像色彩和细节信息丰富以及红外图像在低光照和恶劣天气条件下成像效果较好的特点,本文设计了可见光与红外图像融合的全天候船舶检测方法,充分利用两种模态图像的互补优势,提升船舶检测性能,其网络结构如图 2 所示。

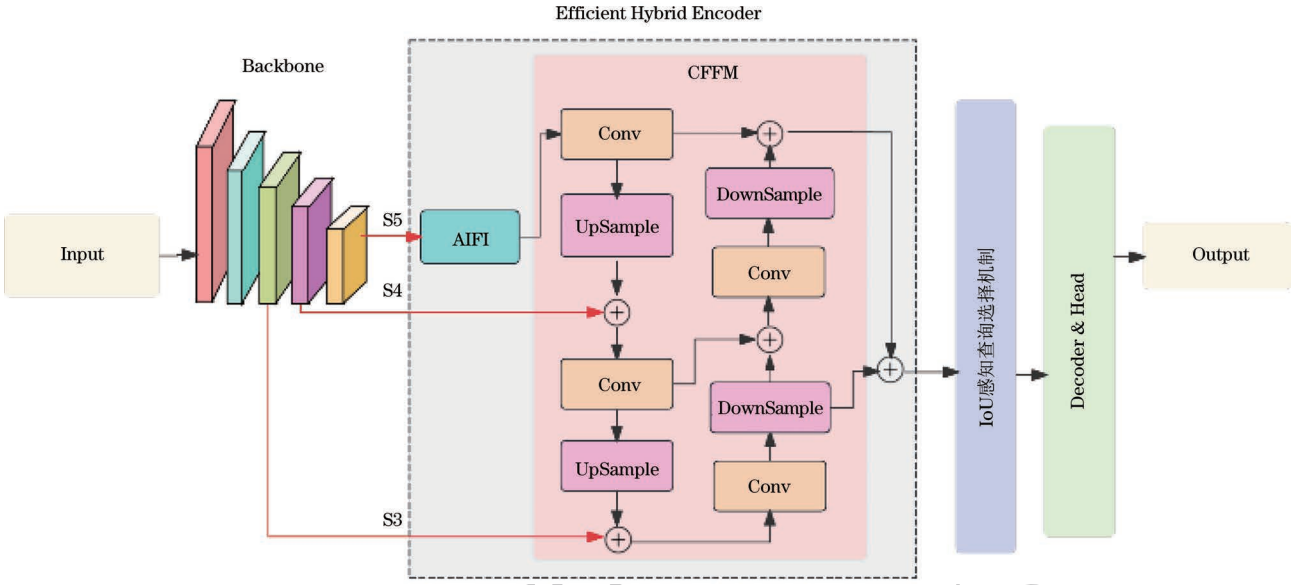

图 1 RT-DETR 网络结构

Fig. 1 RT-DETR network structure

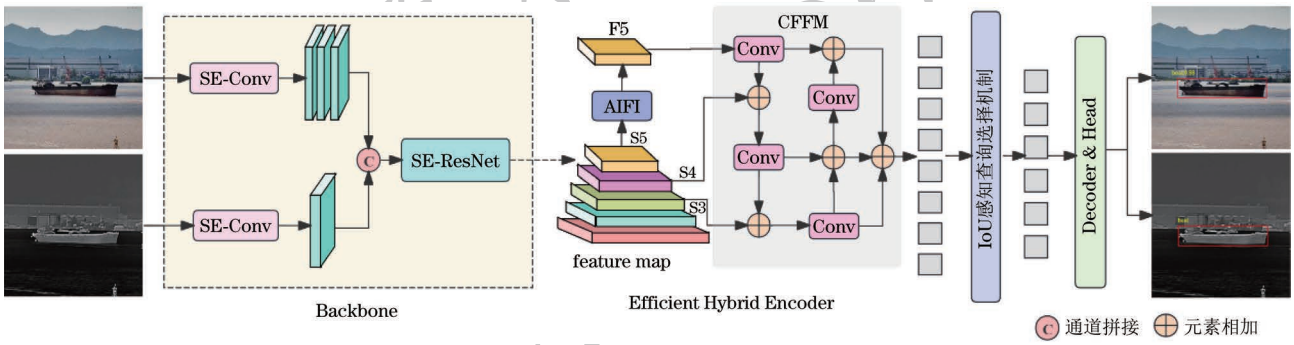

图 2 多模态检测网络结构

Fig. 2 Multimodal detection network structure

本文模型选择 RT-DETR 作为基础模型来实现可见光和红外图像融合的船舶检测。VIF-RTDETR 网络结构主要包括四个部分：首先是船舶的可见光和红外图像的融合模块，将可见光和红外通道分别注入通道注意力机制，通过动态更新通道权重使得模型全面利用不同模态的特征；其次是 Backbone 主干特征提取阶段，将可见光和红外图像作为模型主干提取阶段的输入，为了全面利用不同的特征，同时也在主干特征提取网络中结合通道注意力机制进行动态调整；然后是编码阶段，通过不同尺度的特征交互与融合的编码器，将可见光和红外特征进行多尺度特征融合；最后是编码和预测阶段，通过 IoU 感知查询选择机制和带有预测头的解码器进行迭代优化，获得船舶的目标框和预测分数。

2.1 多模态融合模块

在不同环境下，船舶的可见光和红外两模态图像信息有着不同的特点。可见光图像在晴朗环境下信息完备而红外在光照弱条件下信息丰富。为了合理和全面地利用船舶的可见光和红外图像信息，本文设计了多模态的 VIF 融合模块。该模块充分考虑了可见光和红外图像特征的空间信息和通道信息，实现了对两种模态信息的互补融合。VIF 融合模块如图 3 所示。

为考虑可见光和红外图像特征的空间信息，VIF 模块先对输入的可见光特征 x_{RGB} 和红外特征 x_{IR} 分别经过 3×3 的卷积操作，通过捕捉两模态的局部空间结构和细节信息，提取出可见光和红外图像的初步局部空间特征。然后继续采用最大池化操作进行降采样，用于减少特征图的空间维度和保留最重要的信息，进而得到两模态更为重要的空间信息特征 F^1_{RGB}、F^1_{IR}，计算公式如式(1)所示：

$$\begin{cases} F^1_{\mathrm{RGB}} = \mathrm{Maxpool}(\mathrm{Conv}(x_{\mathrm{RGB}})) \\ F^1_{\mathrm{IR}} = \mathrm{Maxpool}(\mathrm{Conv}(x_{\mathrm{IR}})) \end{cases} \tag{1}$$

式中：x_{RGB} 和 x_{IR} 分别表示输入的可见光与红外的原始特征；Conv( ) 表示 3×3 的卷积操作；Maxpool( )表示最大池化操作。

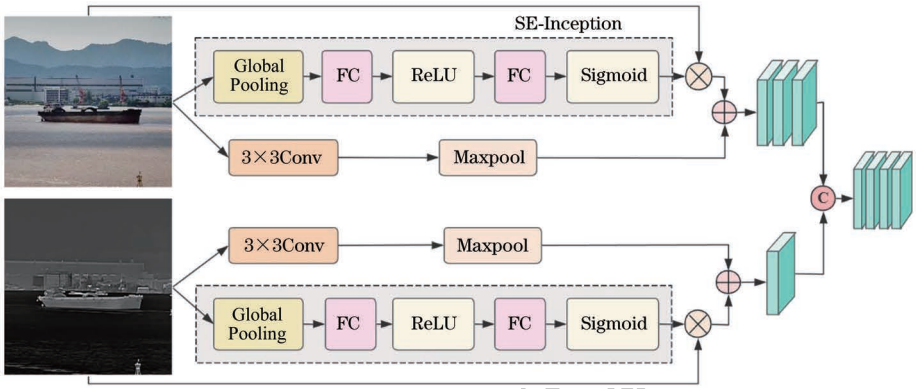

图 3 可见光和红外融合模块

Fig. 3 Visible and infrared fusion module

同时为了增强两模态的通道表达能力，分别利用 SE[24] 模块对可见光和红外图像的原始特征 x_{RGB}、x_{IR} 进行通道注意力处理。首先通过 Squeeze 操作对输入特征 x_{RGB}、x_{IR} 进行全局平均池化，得到对应的特征向量 $\boldsymbol{x}^1_{RGB}$、$\boldsymbol{x}^1_{IR}$，这个操作将大小为 $H\times W\times C$ 的原始特征压缩为 $1\times1\times C$，把每个二维通道变成了一个具有全局感受野的数值；然后通过两个全连接层获取通道间的相关性，并在两个全连接层之后分别使用 ReLU 和 Sigmoid 激活函数，前者激活函数通过引入非线性变换将通道数减少，产生更加紧凑的特征表示，从而增强模型的表达能力，后者激活函数将得到的特征映射为通道重要性权重，输出范围在[0,1]之间，进而得到带有注意力机制的权重参数，经过第一个全连接层，通道数由 C 变为 C/r，减少参数量，经过第二个全连接层，将特征通道数恢复到 C 个；最后将通道权重与两模态的原始特征逐元素相乘得到具有通道信息的特征 F^2_{RGB}、F^2_{IR}，从而实现对通道的自适应加权，以进一步增强模态内部的通道表示能力，计算公式如式(2)所示：

$$\begin{cases}F^2_{RGB}=F_{RGB}\otimes\sigma_2(W_2\sigma_1(W_1\boldsymbol{x}^1_{RGB}))\\F^2_{IR}=F_{IR}\otimes\sigma_2(W_2\sigma_1(W_1\boldsymbol{x}^1_{IR}))\end{cases}\tag{2}$$

式中：W_1 和 W_2 分别表示经过第一个和第二个全连接层的操作；σ_1 和 σ_2 分别表示 ReLU 激活函数和 Sigmoid 激活函数；$\otimes$ 表示将元素对应相乘的操作。

在对可见光和红外图像特征的空间信息和通道信息进行提取后，为了更加有效地整合不同模态的空间特性以及通道信息，VIF 模块采用合并的操作，将来自可见光模态的融合特征与红外模态的融合特征进行合并。具体操作为先将可见光空间特征和通道特征 F^1_{RGB}、F^2_{RGB} 与红外空间特征和通道特征 F^1_{IR}、F^2_{IR} 分别进行合并操作，再将每种模态的空间信息和通道信息整合到一起，从而获得更加丰富的特征表示。两模态的最终特征 F^*_{RGB} 和 F^*_{IR} 如式(3)所示：

$$\begin{cases}F^*_{RGB}=F^1_{RGB}\oplus F^2_{RGB}\\F^*_{IR}=F^1_{IR}\oplus F^2_{IR}\end{cases}\tag{3}$$

式中：$\oplus$ 表示将特征沿维度进行合并操作。

最后将两模态最终特征沿通道维度进行合并操作，使得两种模态信息进行融合，得到 F_C，如式(4)所示：

$$F_C=F^*_{RGB}\oplus F^*_{IR}\tag{4}$$

通过以上 VIF 模块操作，能够有效整合可见光和红外图像的多层次、多维度信息。红外图像能够弥补可见光图像在低光照或不清晰环境下的局限性，而可见光图像则在清晰度和细节表现上具有优势。两者的融合能够提升模型对物体的感知能力，并且有效提升船舶检测的稳定性与准确性。

2.2 多模态 Backbone 特征提取方法

本文的 Backbone 特征提取网络选择 ResNet-50[25] 为基础网络。ResNet-50 网络设计了残差块的结构，在保持较高特征提取能力的同时有效缓解了梯度消失问题，具有良好的特征提取能力。由于可见光和红外图像具有不同的特征分布和信息特性，并且不同的模态在特定条件下可能表现出不同的优劣势，因此仅使用 ResNet-50 网络进行特征提取并不能有效地利用两种模态图像融合后的信息。多模态 Backbone 特征提取结构如图 4 所示。

为了进一步提高对可见光和红外图像特征的合理利用和特征提取，本文同时将通道注意力机制融入到特征提取网络中。基于动态建模特征通道的重要性，使得在特征提取过程中网络能够更加关注对目标任务关键的通道信息，进而实现船舶可见光和红外图像信息的合理利用。

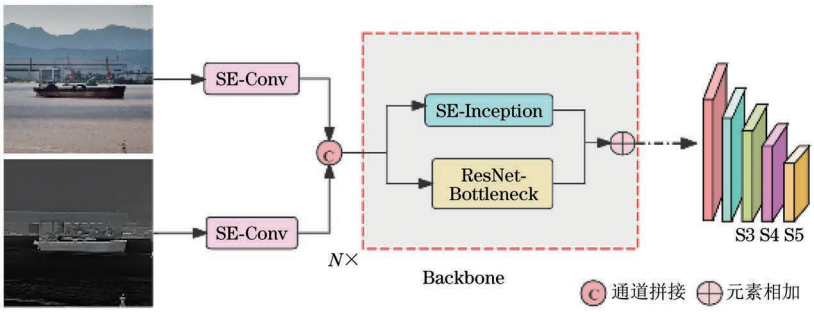

图 4 Backbone 特征提取结构

Fig. 4 Backbone feature extraction structure

本文的多模态 Backbone 特征提取结构能够有效提取出细粒度的特征，并且提取出的多模态特征具有更高的语义表达能力，从而提高了检测任务的性能。具体操作如下：对于融合后的可见光和红外图像特征，先将融合后的特征经过通道注意力机制，动态计算出每个通道对目标任务的重要性，并根据计算结果对通道特征进行加权调整。这个操作使得网络生成精确的通道权重，对可见光与红外融合的特征按照不同权重进行加权，从而增强网络对两模态特征的选择性。同时，将调整后的特征通过 Bottleneck 块进行高层语义信息的提取与表达，增强提取深层信息的能力。这种设计不仅增强了网络对关键特征的关注能力，还能有效抑制冗余或不相关信息，从而为后续的检测任务提供更加精准和丰富的多模态特征表达。

2.3 改进的损失函数

对于船舶检测中较小的船舶目标，小目标往往更难检测，因此提高不同大小目标的船舶检测精度是一项重要工作。基线算法采用三种损失加权来计算损失，包括分类损失、边界框的回归损失和 GIoU 损失[26]。其中，GIoU 的损失计算公式如式(5)所示：

$$L_{\text{loss\_GIoU}}=1-R_{\text{IoU}}+\frac{C-(A\cup B)}{C} \tag{5}$$

式中：A、B 分别表示预测框与真实框；C 表示预测框和真实框的最小外接矩形的面积，用于衡量预测框和真实框的空间包围关系；R_{IoU} 表示 A 和 B 的交并比。GIoU 损失通过加入外接矩形面积，解决了预测框和真实框没有交集时梯度为零的问题，使优化更加稳定。

本文在损失函数中加入了针对不同大小目标的损失，以增强模型对不同大小目标的敏感程度。具体方法为：通过面积来区分小目标，对不同大小目标的 GIoU 损失赋予不同的权重，然后将分类损失、边界框的回归损失和赋予权重的 GIoU 损失组合在一起，得到最终的总损失。计算公式如式(6)和式(7)所示：

$$L=L_{\text{loss\_vlf}}+L_{L_1\text{\_loss}}+\alpha L_{\text{loss\_GIoU}} \tag{6}$$

$$\alpha=2+\frac{3}{1+\mathrm{e}^{-\left(\frac{w\cdot h}{\gamma}-1\right)}} \tag{7}$$

式中：$L_{\text{loss\_vlf}}$ 表示分类损失；$L_{L_1\text{\_Loss}}$ 表示边界框回归损失；$L_{\text{loss\_GIoU}}$ 表示 GIoU 损失；α 表示考虑面积信息的权重；w、h 分别为目标框的宽和高；γ 为小目标的阈值，设定 $\gamma=900$，即当目标框面积 $w\cdot h\leqslant 900$ 时为小目标。

通过在损失函数中增加不同目标大小的权重项，使得当目标框为小目标时，加大其损失项，使模型在训练过程中更加关注小目标，促使模型学习到更多关于小目标的细节信息和位置信息，从而提高船舶检测中较小船舶目标的检测精度。

3 实验结果及分析

3.1 数据集

本文所使用的数据集来自于某一港口摄像头所捕捉到船舶的真实数据，该摄像头获取的是同一时刻的可见光和红外船舶图像。在获取图像后，对这两种模态的数据进行对齐操作，使得两者尺寸大小均为 640×640。该数据集共有 6 970 张图像，其中可见光和红外图像船舶图像各 3 485 张，包含有晴天、阴天、大雾和夜晚各种场景下的船舶图像。数据集包含各种类型的图像如图 5 所示。可见光图像具有 RGB 通道，红外图像具有单一的灰度通道。在大雾、夜晚等光照差的情况下，船舶的红外图像包含的信息比可见光更丰富。相反，在晴朗光照好的情况下红外图像所包含的船舶信息更少，可见光图像包含了丰富的纹理和细节信息。

基于对齐后的可见光和红外船舶图像，使用 labelimg 对图像进行标注，并且两种模态的船舶图像共用相同的标注信息。为了进一步提高数据的表达能力，本文对可见光和红外数据集进行了数据增

(a)晴天 (b)阴天 (c)大雾 (d)夜晚

图 5 数据集中船舶图像部分示例

Fig.5 Partial examples of the ship images on the dataset

强处理。采用的数据增强策略从几何变换和像素调整这两个方面展开，包括水平翻转、调整对比度和亮度等操作。通过数据增强操作能够有效增加训练数据的多样性，提高模型的泛化能力和鲁棒性。由于模型输入的数据集标注格式需要文件后缀为 json 的 COCO 格式，因此需将 Labelimg 标注得到的 xml 文件转化为 json 格式，之后再将训练集、验证集和测试集按照 8∶1∶1的比例来划分得到模型所使用的数据集。

3.2 实验环境及参数配置

本实验是在 PyTorch 深度学习框架下进行的，所使用的 PyTorch 版本为 2.3.0 版本。代码使用的语言为 Python 3.8 版本。实验过程中使用 GPU 加速，在模型训练过程中，同时要配置 onnx，其版本为 1.14.0。根据以往神经网络训练经验，实验过程中的参数配置如表 1 所示。

表 1 训练参数配置

Table 1 Configuration of training parameters

实验参数	相关说明
learning_ rate 为 0.000 01	学习率为 0.000 01
Epoch 为 200	训练周期设置为 200
Batch Size 为 4	批量处理大小为 4
weight_decay 为 0.000 1	权重衰减系数为 0.000 1

在实验中，采用参数学习率为 0.000 01，训练周期为 200，批量处理大小为 4，权重衰减系数为 0.000 1 进行训练。其中，低的学习率有助于在复杂网络训练中实现更稳定的梯度下降以及避免出现梯度爆炸的问题。训练周期设置为 200 是为了确保模型有足够的时间对数据特征进行学习，低的权重衰减系数是为了防止其对训练数据产生过拟合，通过正则化来约束模型的复杂度。将基于可见光与红外融合的船舶检测分别与基于可见光的船舶检测和基于红外的船舶检测进行对比实验，实验环境等配置均相同。

3.3 评价指标

船舶检测是一项目标检测任务，本文采用目标检测中常用的指标对模型的检测性能进行评估。将平均精度(AP)、平均召回率(AR)、F1 值(F1，F_1)、Params 作为本文模型检测效果的评价指标。

精确率(P)是在图片经过识别后，正样本的正确识别数量所占的比例，召回率(R)是在验证集中正确识别为正样本的所有正样本样例数占比。精确率和召回率的计算分别如式(8)和式(9)所示：

$$P=\frac{N_{TP}}{N_{TP}+N_{FP}} \tag{8}$$

$$R=\frac{N_{TP}}{N_{TP}+N_{FN}} \tag{9}$$

式中：N_{TP} 表示正样本的正确识别数量；N_{FP} 表示负样本错误被识别为正样本的数量；N_{FN} 表示正样本错误被识别为负样本的数量。

AP 表示 P-R 曲线下的面积，可以用来衡量目标检测算法的性能。划分 AP 值的范围是 0～1，值越高表示目标检测模型的性能越好。AR 是在每个图像中检测到固定数量的最大召回率，它表示模型能够正确识别出所有实际存在目标的比例，也可以衡量漏检情况。划分 AR 值的范围也是 0～1，值越高表示目标检测模型的性能越好。

F1 值是一个平衡精确率和召回率的评估方法。高精确率意味着模型检测到的对象大多数是正确的，而高召回率意味着模型能够检测到大部分实际存在的对象。F1 值能够同时考虑这两个方面，为模型的性能提供了全面的评估。其计算公式如式(10)所示：

$$F_1=2\cdot\frac{P\cdot R}{P+R} \tag{10}$$

Params 表示模型中需要通过训练优化的参数量，参数量衡量了模型的复杂度和表达能力。使用 Params 作为一项评价指标，有助于定量评价模型的性能和计算复杂度。较少的参数量通常意味着更高的效率和较低的计算成本，而较大的参数量则可能对应更强的表达能力，但需要更多的训练时间和计算资源。

3.4 船舶检测结果分析

3.4.1 Transformer 解码器深度选择

在该模型中，通过实验调整解码器的深度，以分析模型参数和深度对网络性能的影响。解码器深度测试结果如表 2 所示，加粗表示最优数据(下同)。

表 2 解码器深度测试

Table 2 Decoder depth tests

Layer	$AP_{0.5:0.95}$	$AP_{0.5}$	$AP_{0.75}$	$AR_{0.5:0.95}$	F1	Params
2	0.770	0.984	0.909	0.828	0.797	**40.7×10^6**
4	0.774	0.980	0.910	**0.859**	0.814	42.9×10^6
6	**0.783**	**0.985**	**0.917**	0.852	**0.817**	45.2×10^6
8	0.771	0.978	0.916	0.850	0.808	47.4×10^6

从表 2 可以看出，随着解码器深度的增加，模型参数量和计算复杂度呈线性增长。较深的解码器层数能够更好地捕捉和融合多模态信息，从而提升模型的表达能力，特别是在复杂场景下对目标检测的准确性有所提高。然而，过大的模型深度也会导致推理速度减慢和计算资源消耗增加。因此，实验通过权衡性能提升与计算开销，在深度 2～8 层之间确定最佳解码器层数，从而在模型效果和计算效率之间实现平衡。本文通过解码器深度测试的实验结果将解码器深度设置为 6 层。

3.4.2 对比实验

为了对改进后模型的检测性能进行总体评估，本文在船舶的可见光和红外数据集上进行实验，使用 VIF-RTDETR 算法与其他六种不同的算法模型进行对比实验，并使用 $AP_{0.5:0.95}$、$AP_{0.5}$、$AR_{0.5:0.95}$ 和 F1 值等指标来评价模型的检测效果。实验结果如表 3 所示，不同模型融合可见光和红外的训练 AP 值对比如图 6 所示。

表 3 不同算法的实验结果

Table 3 Experimental results among different algorithms

算法	数据集	$AP_{0.5:0.95}$	$AP_{0.5}$	$AP_{0.75}$	$AR_{0.5:0.95}$	F1	Params
DETR	RGB	0.676	0.968	0.802	0.761	0.715	**41.2×10^6**
	IR	0.630	0.948	0.735	0.746	0.683	**41.2×10^6**
	RGB+IR	0.676	0.968	0.802	0.761	0.715	**41.2×10^6**
DINO-DETR[27]	RGB	0.739	0.983	0.878	0.832	0.782	46.6×10^6
	IR	0.650	0.973	0.742	0.766	0.703	46.6×10^6
	RGB+IR	0.739	0.978	0.889	**0.853**	0.791	46.6×10^6
DN-DETR[28]	RGB	0.689	0.975	0.811	0.780	0.731	43.4×10^6
	IR	0.564	0.938	0.625	0.716	0.630	43.4×10^6
	RGB+IR	0.689	0.975	0.811	0.780	0.731	43.4×10^6
Salience-DETR[29]	RGB	0.748	0.978	0.880	0.813	0.779	49.2×10^6
	IR	0.677	0.959	0.797	0.781	0.725	49.2×10^6
	RGB+IR	0.761	0.984	0.897	0.821	0.790	49.2×10^6
Relation-DETR[30]	RGB	0.724	0.979	0.839	0.814	0.766	41.3×10^6
	IR	0.622	0.939	0.689	0.762	0.684	41.3×10^6
	RGB+IR	0.739	0.969	0.849	0.834	0.783	41.3×10^6
MS-DETR[31]	RGB	0.715	0.977	0.857	0.791	0.751	53.5×10^6
	IR	0.538	0.907	0.552	0.725	0.617	53.3×10^6
	RGB+IR	0.724	0.970	0.846	0.814	0.766	53.3×10^6
VIF-RTDETR	RGB	0.736	0.967	0.827	0.821	0.732	42.7×10^6
	IR	0.691	0.958	0.863	0.779	0.776	42.7×10^6
	RGB+IR	**0.783**	**0.985**	**0.917**	0.852	**0.817**	45.2×10^6

从表 3 对比实验结果以及图 6 中不同模型 AP 值比较可以看出，使用不同模型分别对船舶的红外图像或可见光图像进行检测时，精度较低；而同时使用可见光和红外图像进行检测时，精度高于单一模态图像的检测精度。这是由于同时使用可见光和红外图像综合了两种模态下的船舶信息，避免了不同场景下（如可见光在阴天、大雾等恶劣天气下）船舶形态不清晰等问题产生的影响，因此使用两种图像信息的检测效果最好。

在使用两种模态数据进行检测的不同模型中，本文提出的 VIF-RTDETR 模型取得了最高的检测精度。相对于单独使用可见光图像检测的 $AP_{0.5}$ 提

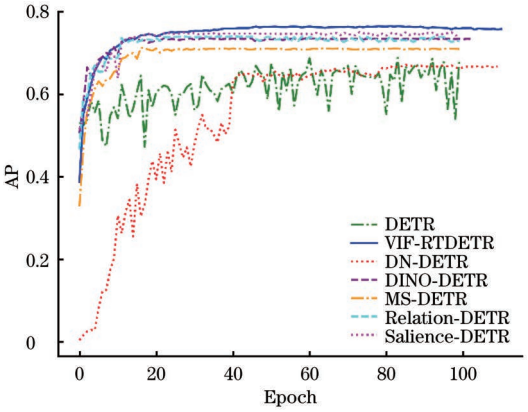

图 6 不同模型融合可见光和红外训练 AP 值对比

Fig. 6 Comparison of training AP values among different models fusing visible and infrared light

高了 1.8 百分点，$AP_{0.5:0.95}$ 提高了 4.7 百分点，召回率 $AR_{0.5:0.95}$ 提高了 3.1 百分点，F1 值提高了 8.5 百分点；相对于单独使用红外图像检测的 $AP_{0.5}$ 提高了 2.7 百分点，$AP_{0.5:0.95}$ 提高了 9.2 百分点，$AR_{0.5:0.95}$ 提升了 7.3 百分点，F1 值提高了 4.1 百分点。因此，VIF-RTDETR 模型显著提升了检测精度并有效减少了漏检情况。

为了直观地观察模型检测的效果，本文使用不同的模型以及多模态可见光和红外 VIF-RTDETR 模型对船舶进行检测，并将部分检测结果进行可视化。通过将检测结果可视化，清晰展示出了模型在不同光照条件下对船舶目标的检测性能。不同模型和 VIF-RTDETR 模型的部分检测结果如图 7 所示。

图 7 不同场景下部分检测结果

Fig. 7 Partial detection results in different scenario

在可视化检测结果中，本文选取了检测精度较高的 3 个检测模型与 VIF-RTDETR 方法进行可视化比较，实验所采用数据为不同场景状态下的可见光和红外图像，分别展示出了阴天、有雾、晴天和夜晚的检测效果。图 7(a)和图 7(b)分别为可见光和红外图像原图；图 7(c)、图 7(d)和图 7(e)分别为使用 Salience-DETR、DINO-DETR 和 Relation-DETR 模型检测的可视化图，图 7(f)是本文提出的 VIF-RTDETR 模型进行检测的可视化结果图。

从图 7 的检测结果可以看出，使用 VIF-RTDETR 模型对船舶进行检测，能够将图像中含有的船舶目标均检测出来，并且有较高的检测精度。而使用 Salience-DETR、DINO-DETR 和 Relation-DETR 模型进行检测时存在船舶目标漏检的情况，并且检测出部分船舶目标的精度不高。其中，DINO-DETR 和 Relation-DETR 存在对船舶小目标以及在夜晚环境下船舶漏检的情况，这可能是由于在特征提取中对小目标的提取能力不强，使得船舶目标

的特征提取不够准确，从而导致了漏检。Salience-DETR 在有雾的环境下，对船舶检测的精度不够高。由图 7 可以看出，Salience-DETR 在有雾环境下对图中左上角船舶的检测精度为 68%，而 VIF-RTDETR 模型对其检测精度为 85%，这展示了在有雾条件下 VIF-RTDETR 模型具有较好的检测效果。因此，使用本文提出的 VIF-RTDETR 模型对于不同环境条件均有着较好的检测效果，并且减少了漏检情况，可以用于全天候的船舶检测。

3.4.3 消融实验

为了验证引入注意力机制的 Backbone 特征提取网络和加入损失模块的有效性，在船舶的可见光和红外图像的数据集上进行消融实验，消融实验结果如表 4 所示。

表 4 消融实验结果

Table 4 Results of ablation experiment

SE-Backbone	Loss	$AP_{0.5:0.95}$	$AP_{0.5}$	$AP_{0.75}$	$AR_{0.5:0.95}$	F1	Params
—	—	0.771	0.984	0.899	0.841	0.804	**42.7×10^6**
√	—	0.775	0.984	0.913	0.839	0.806	45.2×10^6
—	√	0.777	0.983	0.904	0.851	0.812	**42.7×10^6**
√	√	**0.783**	**0.985**	**0.917**	**0.852**	**0.817**	45.2×10^6

通过消融实验结果可以看出，引入注意力机制的 Backbone 特征提取网络和加入损失模块的检测结果均有提升。其中，加入注意力机制的 Backbone 特征提取网络使得模型的 $AP_{0.75}$ 提高了 1.4 百分点，这是因为加入通道注意力机制的特征提取网络能够实现可见光和红外这两种模态数据集特征信息的合理利用。加入改进的损失函数模块使得模型 $AP_{0.5:0.95}$、$AR_{0.5:0.95}$ 分别提高了 0.6 和 1.0 百分点，这是由于当目标框为小目标时，改进的损失函数加大了小目标损失项，使得网络模型更加关注小目标的检测。RT-DETR 的模型参数量为 42.7×10^6，加入 SE-Backbone 后模型参数量增加到了 45.2×10^6，这是因为在模块中加入的通道注意力模块引入了参数量，模型参数量的增多使得模型特征提取能力上升，同时会增加计算开销，要权衡模型的性能提升与资源消耗。从表 4 可以看出，共同加入注意力机制的 Backbone 特征提取网络和改进的损失函数模块，使得模型的精度和召回率都有所提升，这说明本文算法对船舶的红外及可见光融合的检测有显著的提升效果。

4 结束语

针对单一模态图像下船舶检测精度较低和漏检率高的问题，本文提出了一种充分融合可见光和红外信息的船舶目标检测方法 VIF-RTDETR。该方法根据可见光图像具有丰富的色彩和细节信息，以及红外图像在低光照和恶劣天气条件下具有成像效果较好的特点，从多模态融合方法、Backbone 特征提取网络、损失函数等方面对模型进行改进，提高了船舶的检测性能。实验结果表明，所提的 VIF-RTDETR 模型相对于仅使用可见光和红外的单一模态模型均有很好的检测效果，可以实现全天候的船舶检测，显著提高了船舶检测的性能，并且降低了漏检情况。后续将进一步扩充船舶的可见光和红外数据集，并结合更先进的多模态融合方法，提高模型的检测效果和泛化能力。

参考文献

[1] 宋志娜，眭海刚，李永成. 高分辨率可见光遥感图像舰船目标检测综述[J]. 武汉大学学报(信息科学版)，2021，46(11)：1703-1715.
SONG Z N, SUI H G, LI Y C. A survey on ship detection technology in high-resolution optical remote sensing images [J]. Geomatics and Information Science of Wuhan University, 2021, 46(11): 1703-1715. (in Chinese)

[2] 马啸，邵利民，金鑫，等. 舰船目标识别技术研究进展[J]. 科技导报，2019，37(24)：65-78.
MA X, SHAO L M, JIN X, et al. Advances in ship target recognition technology[J]. Science & Technology Review, 2019, 37(24): 65-78. (in Chinese)

[3] YÜKSEL G K, YALıTUNA B, TARTAR Ö F, et al. Ship recognition and classification using silhouettes extracted from optical images[C]//Proceedings of the 24th Signal Processing and Communication Application Conference (SIU). Zonguldak, Turkey: IEEE Press, 2016: 1617-1620.

[4] 樊怡颖，禹维. 基于时序图像的双分支 SAR 图像船舶检测方法[J]. 计算机工程，2025，51(12)：31-42.
FAN Y Y, GUO W. Dual-branch SAR image ship detection method based on time-series images [J]. Computer Engineering, 2025, 51(12): 31-42. (in Chinese)

[5] ZHAO H W, ZHANG W S, SUN H Y, et al. Embedded deep learning for ship detection and recognition[J]. Future Internet, 2019, 11(2): 53.

[6] DONG Y X, CHEN F K, HAN S, et al. Ship object detection of remote sensing image based on visual attention [J]. Remote Sensing, 2021, 13(16): 3192.

[7] HAN W X, KUERBAN A, YANG Y C, et al. Multi-vision network for accurate and real-time small object detection in optical remote sensing images [J]. IEEE Geoscience and Remote Sensing Letters, 2022, 19: 6001205.

[8] 牛为华，郭迅. 基于改进 YOLOv8 的船舰遥感图像旋转目标检测算法[J]. 图学学报，2024，45(4)：726-735.
NIU W H, GUO X. Rotating target detection algorithm in

ship remote sensing images based on YOLOv8[J]. Journal of Graphics, 2024, 45(4): 726-735. (in Chinese)
[9] CARION N, MASSA F, SYNNAEVE G, et al. End-to-end object detection with transformers[C]// Proceedings of European Conference on Computer Vision. Berlin, Germany: Springer, 2020: 213-229.
[10] ZHU X Z, SU W J, LU L W, et al. Deformable DETR: deformable transformers for end-to-end object detection[EB/OL]. [2024-09-01]. https://arxiv.org/pdf/2010.04159.
[11] DAI Z G, CAI B L, LIN Y G, et al. UP-DETR: unsupervised pre-training for object detection with transformers[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Nashville, USA: IEEE Press, 2021: 1601-1610.
[12] MENG D P, CHEN X K, FAN Z J, et al. Conditional DETR for fast training convergence[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). Montreal, Canada: IEEE Press, 2022: 3631-3640.
[13] ROH B, SHIN J W, SHIN W, et al. Sparse DETR: efficient end-to-end object detection with learnable sparsity[EB/OL]. [2024-09-01]. https://arxiv.org/pdf/2111.14330.
[14] 何智杰, 肖玮,柯学良,等. 基于改进 RT-DETR 的相似背景干扰场景目标检测算法[J]. 光电工程, 2025, 52(10): 250132.
HE Z J, XIAO W, KE X L, et al. Object detection algorithm based on improved RT-DETR in similar background interference scenario[J]. Opto-Electronic Engineering, 2025, 52(10): 250132. (in Chinese)
[15] LIU S L, LI F, ZHANG H, et al. DAB-DETR: dynamic anchor boxes are better queries for DETR[EB/OL]. [2024-09-01]. https://arxiv.org/pdf/2201.12329.
[16] 李士博, 肖振久, 曲海成, 等. 面向 SAR 图像舰船检测的多粒度特征与形位相似度量方法[J]. 光电工程, 2025, 52(2): 240254.
LI S B, XIAO Z J, QU H C, et al. Multi-granularity feature and shape-position similarity metric method for ship detection in SAR images[J]. Opto-Electronic Engineering, 2025, 52(2): 240254. (in Chinese)
[17] 陈振. 红外舰船检测与目标识别方法研究[D]. 哈尔滨: 哈尔滨工程大学, 2016.
CHEN Z. Methods of infrared ship detection and target recognition[D]. Harbin: Harbin Engineering University, 2016. (in Chinese)
[18] LU Y Q, MA H J, SMART E, et al. Fusion of camera-based vessel detection and AIS for maritime surveillance[C]// Proceedings of the 26th International Conference on Automation and Computing (ICAC). Portsmouth, United Kingdom: IEEE Press, 2021: 1-6.
[19] 赵炜东, 郭鹏宇, 刘勇, 等. 基于可见光与红外卫星图像融合的舰船目标检测[J]. 上海航天(中英文), 2023, 40(1): 44-52.
ZHAO W D, GUO P Y, LIU Y, et al. Ship detection based on visible and infrared satellite image fusion[J]. Aerospace Shanghai (Chinese & English), 2023, 40(1): 44-52. (in Chinese)
[20] REDMON J, DIVVALA S, GIRSHICK R, et al. You only look once: unified, real-time object detection[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Las Vegas, USA: IEEE Press, 2016: 779-788.
[21] 王昱婷, 刘志明, 万亚平, 等. 基于可见光与红外图像的弱光条件下目标检测[J]. 计算机工程, 2024, 50(8): 270-281.
WANG Y T, LIU Z M, WAN Y P, et al. Target detection under low light conditions based on visible and infrared images[J]. Computer Engineering, 2024, 50(8): 270-281. (in Chinese)
[22] 李海军, 孔繁程, 林云. 基于改进 YOLOv5s 的红外舰船检测算法[J]. 系统工程与电子技术, 2023, 45(8): 2415-2422.
LI H J, KONG F C, LIN Y. Infrared ship detection algorithm based on improved YOLOv5s[J]. Systems Engineering and Electronics, 2023, 45(8): 2415-2422. (in Chinese)
[23] ZHAO Y A, LV W Y, XU S L, et al. DETRs beat YOLOs on real-time object detection[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE Press, 2024: 16965-16974.
[24] HU J, SHEN L, SUN G. Squeeze-and-excitation networks[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE Press, 2018: 7132-7141.
[25] HE K M, ZHANG X Y, REN S Q, et al. Deep residual learning for image recognition[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Las Vegas, USA: IEEE Press, 2016: 770-778.
[26] REZATOFIGHI H, TSOI N, GWAK J, et al. Generalized intersection over union: a metric and a loss for bounding box regression[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA: IEEE Press, 2020: 658-666.
[27] ZHANG H, LI F, LIU S L, et al. DINO: DETR with improved denoising anchor boxes for end-to-end object detection[EB/OL]. [2024-09-01]. https://arxiv.org/pdf/2203.03605.
[28] LI F, ZHANG H, LIU S L, et al. DN-DETR: accelerate DETR training by introducing query denoising[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). New Orleans, USA: IEEE Press, 2022: 13609-13617.
[29] HOU X Q, LIU M Q, ZHANG S L, et al. Salience DETR: enhancing detection transformer with hierarchical salience filtering refinement[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle, USA: IEEE Press, 2024: 17574-17583.
[30] HOU X Q, LIU M Q, ZHANG S L, et al. Relation DETR: exploring explicit position relation prior for object detection[EB/OL]. [2024-09-01]. https://arxiv.org/pdf/2407.11699.
[31] ZHAO C Y, SUN Y F, WANG W H, et al. MS-DETR: efficient DETR training with mixed supervision[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle, USA: IEEE Press, 2024: 17027-17036.

文字编辑 薛晋栋
栏目编辑 宋 圆