文章编号:1001-5078(2023)09-1364-11　　·红外技术及应用·

基于双模态特征增强的目标检测算法研究与应用

王文霞[1],张　文[2],何　凯[3]
(1. 太原师范学院网络信息中心,山西 太原 030619;2. 北京邮电大学信息与通信工程学院,北京 100080;
3. 中国空间技术研究院西安分院,陕西 西安 710100)

摘　要:为提升目标检测算法在复杂环境下的精确性和实用性,将多源信息和深度学习技术相结合,提出了一种基于双模态特征增强的目标检测方法。该方法以红外和可见光图像作为输入,利用颜色空间转换、边缘提取、直方图均衡化等传统图像处理方法丰富图像信息,达到数据增强效果;特征提取部分采用卷积神经网络结构分别提取目标红外及可见光信息,并设计混合注意力机制分别从通道和空间位置角度提升有效特征权重;同时,针对目标双模态信息,引入了自适应交叉融合结构,提高特征多样性;最后,利用交替上下采样将目标全局和局部特征充分融合,并以自主选择方式提取目标相关特征实现检测。通过在标准数据集以及实际场景数据集上的实验结果表明,所提方法有效融合并增强了目标多模态特征,提升了目标检测效果,并能较好的应用于电网场景中,辅助机器人完成目标设备检测。

关键词:双模态;特征增强;目标检测;混合注意力;自适应融合;多尺度检测
中图分类号:TP391.41;TN219　　**文献标识码:**A　　**DOI:**10.3969/j.issn.1001-5078.2023.09.010

Research and application of object detection algorithm based on bimodal feature enhancement

WANG Wen-xia[1], ZHANG Wen[2], HE Kai[3]
(1. Network Information Center of Taiyuan Normal University, Taiyuan 030619, China;
2. Beijing University of Posts and Telecommunications, Beijing 100080, China;
3. Xi'an Branch of China Academy of Space Technology, Xi'an 710100, China)

Abstract: In order to improve the accuracy and practicability of object detection algorithm in complex environments, an object detection method based on bimodal feature enhancement is proposed by combining multi-source information and deep learning technology. This method takes infrared and visible images as input and traditional image processing methods are used such as color space conversion, edge extraction, and histogram equalization to enrich image information and achieve data enhancement effects. In the feature extraction part, the convolutional neural network structure is used to extract the infrared and visible light information of the object respectively, and a hybrid attention mechanism is designed to enhance the effective feature weight from the channel and spatial position respectively. At the same time, an adaptive cross fusion structure is introduced to enhance the feature diversity for the object bimodal information. Finally, the global and local features of the object are fully fused by alternating up and down sampling, and the relevant features of the object are extracted in an autonomous way to achieve detection. The experimental results on standard datasets and the actual real scene datasets show that the proposed method effectively fuses and enhances the multi-modal features of the object, improves the object detection effect, and can be better applied to the power grid scene to as-

基金项目:国家自然科学基金项目(No. 62071058)资助。
作者简介:王文霞(1979 -),女,工程师,研究方向为图像处理,云计算,云存储技术。E-mail:lmclw13@sina.com
收稿日期:2022-10-31

sist the robot to complete object equipment detection.

Keywords: bimodal; feature enhancement; object detection; mixed attention; adaptive fusion; multiscale detection

1 引 言

目标检测作为计算机视觉领域的三大任务之一,被广泛应用于自动驾驶、视频监控、电力巡检等场景中[1]。所谓目标检测主要根据目标特征对图像或视频中的目标进行分类并定位[2],前期的目标检测思路主要利用人工设计特征(HOG、Haar、DPM、LBP 等)结合浅层分类器(SVM、Adaboost 等)方式实现检测[3],虽然有较高的计算效率,但特征设计过程复杂,且检测效果较差,应用场景有限。而随着人工智能和计算机技术的发展,基于深度神经网络的目标检测以自主学习的方式提取特征,有效避免了人工设计特征的局限,并逐渐成为了目标检测主流方向[4]。虽然深度学习技术有效提升了目标检测效果,但现有大多数方法主要利用单一红外或可见光图像进行目标检测,对于日益复杂的检测场景仍存在诸多困难[5]。因此,设计一种融合目标多模态特征的检测方法不仅能推动深度学习技术的发展,也能加速目标检测方法落地实际应用。

对于深度学习多模态目标检测算法的研究,目前,已有部分学者进行了相应的探索。顾晶晶等人[6]针对遥感图像中的小目标检测,设计了一种基于红外和可见光平衡多模态深度模型,通过融目标浅层特征后再利用 YOLOv4 深层网络实现小目标检测,但仅融合浅层特征的方式无法充分利用目标双模态信息,且浅层特征融合也会引入较多噪声。邝楚文等人[7]提出了一种自适应的特征融合方法,将红外图像多维度特征以自主加权的方式融入可见光网络中,弥补可见光信息的局限,提升检测效果。该方式虽丰富了特征信息但缺乏对特征有效性的关注,容易导致融入较多无效特征。Banuls 等人[8]利用深层神经网络分别对红外和可见光图像进行目标检测后将检测结果融合,再利用非极大值抑制算法筛选出最优目标框,但网络仅从决策层进行融合,并考虑特征层面融合,故检测精度的提升有限。Ma 等人[9]提出了一种基于显著目标检测的红外可见光融合网络,利用目标掩码来突出红外和可见光图像中的关键信息,以隐式的方式来融合增强目标特征,提升网络对显著目标的检测,但该方式对于特征信息较少的小目标检测效果较差。可见,现阶段的红外及可见光融合检测方法都相对存在一定局限,目标检测性能仍有较大提升空间。

针对上述目标检测方法存在的不足,本文从特征多样性、注意力以及多尺度等角度,提出了一种基于双模态特征增强的目标检测方法。该方法首先通过多种传统图像处理技术分别处理红外和可见光图像,丰富输入图像信息;其次,利用双支路深层卷积神经网络提取目标双模态特征,并设计混合注意力机制提升可见光目标类别信息以及红外目标空间位置信息;然后,以自适应交叉融合的方式使红外和可见光目标信息相互补充,增加特征多样性。最后,针对不同维度的目标特征,设计了特征交替采样以及自主选择结构,充分融合目标深层和浅层特征同时降低了不同维度特征之间相互干扰,保障网络准确高效地实现目标检测。

2 目标检测结构设计

2.1 整体结构

所提双模态目标检测网络整体结构如图 1 所示,主要分为图像增强、特征提取、混合注意力、自适应交叉融合、多尺度检测结构几个模块。图像增强采用颜色空间转化、边缘检测、滤波等方法对红外和可见光原图进行处理,以丰富输入图像的信息;特征提取利用卷积、激活、池化等操作构建双支路的深层神经网络,分别提取红外和可见光目标特征;混合注意力结构主要从特征显著性角度,以多种方式获取全局上下文信息,增强目标空间位置信息以及所属类别信息;自适应交叉融合则通过自适应加权方式将对应维度的红外特征和可见光特征交叉融合,使目标双模态信息相互补充;而多尺度检测则针对不同维度目标,通过深度到浅层再浅层到深层的交替采样融合方式充分捕获目标全局及局部特征,并综合各维度特征,以自主选择方法提取目标相关特征,提升网络尺度不变性;最后,利用单步检测器结合非极大值抑制算法实现目标的识别定位。

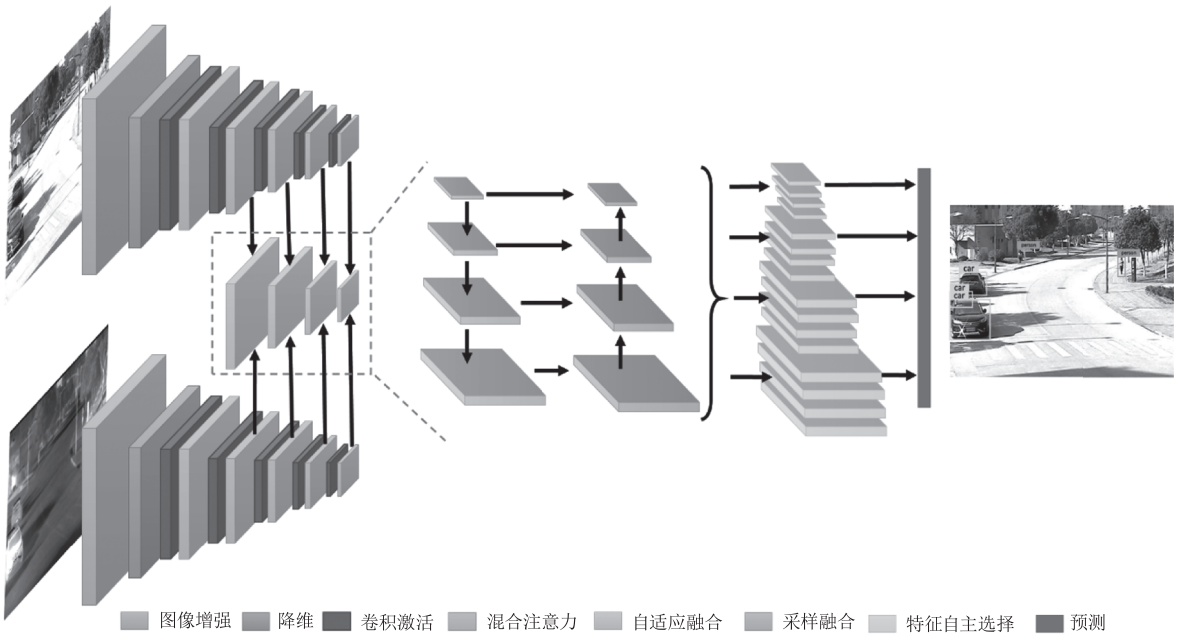


图 1 双模态目标检测网络框架

Fig. 1 Bimodal object detection network framework

2.2 特征提取

特征提取是计算机视觉任务的关键步骤之一,将输入数据通过降维、卷积等方式挖掘出与目标相关的有效信息,供后续任务模块使用[10-11]。现阶段大多数基于深度学习的目标检测网络[12-15]主要针对单一输入源,通常只需要一条骨干网络进行特征提取,而所提方法针对目标双模态特征,故采用了对称双支路网络结构提取特征。而支路基本结构主要在综合现有网络基础上,通过丰富输入信息并引入高效率特征提取单元完成构建,支路基础结构如表 1 所示。

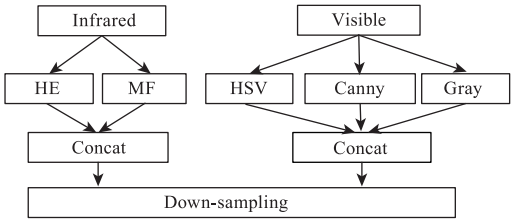
所提特征提取支路结构主要由图像增强部分、降维采样操作以及一系列卷积模块组成。图像增强(Processing)主要利用传统图像预处理方式分别对红外和可见光原图进行处理,如图 2(a)所示。由于红外图像包含较多的目标位置信息,故采用直方图均衡化(Histogram Equalization,HE)、均值滤波(Mean Filter,MF)等方法增强;而可见光图像包含较多细节信息,故采用了颜色空间转化(HSV)、边缘提取(Canny)、灰度转化(Gray)等方式增强。降维采样操作(Down sampling)如图 2(b)所示,主要对增强后的输入图像进行降维,减少后续模块计算量。为避免降维过程造成信息丢失,分别采用了步长为 2 的标准卷积、深度可分离卷积、平均池化以及最大池化操作进行降维。而卷积模块(Block)则

表 1 特征提取支路基础结构

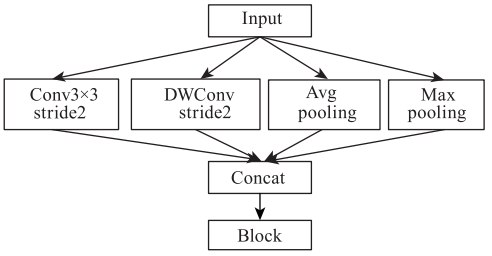
Tab. 1 Feature extraction branch base structure

网络模块	操作层	重复次数	输出维度
Source	RGB,3	1	448×448
Visible Processing	HSV,3 Canny,1 Gray,1	1	448×448
Infrared Processing	HE,1 MF,1	1	448×448
Down-sampling	Conv 3×3,8 DWconv 3×3,8 Max pooling 2×2,8 Avg pooling 2×2,8	1	224×224
Block 1	DWconv 3×3,24 H-Swish Residual Hybrid Attention	1	112×112
Block 2	DWconv 3×3,40 H-Swish Residual Hybrid Attention	2	56×56
Block 3	DWconv 3×3,80 H-Swish Residual Hybrid Attention	3	28×28
Block 4	DWconv 3×3,112 H-Swish Residual Hybrid Attention	3	14×14
Block 5	DWconv 3×3,160 H-Swish Residual Hybrid Attention	2	7×7

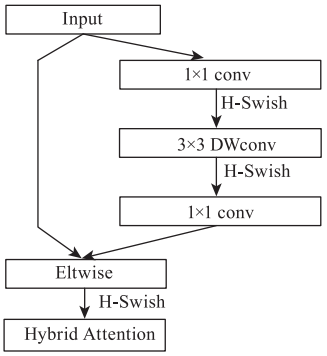
是特征提取的基本单元，采用深度可分离卷积结合 H-Swish 激活函数的方式来使所构建的特征提取网络保持轻量化，并利用 1 ×1 的卷积核实现跨通道的信息交互，再通过残差连接操作来缓解深层网络训练时梯度消失等问题，基本结构如图 2(c)所示。为避免双支路网络引入过多参数，网络特征通道数相对较少，虽可能造成部分信息丢失，但也减少了冗余特征，且缺失的特征也可通过双模态特征融合得到补充。



(a) 图像增强模块



(b) 降维采样模块



(c) Block卷积块

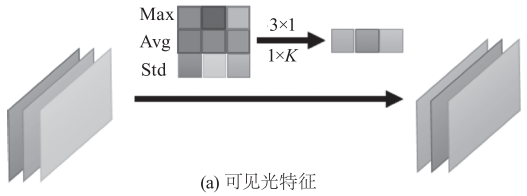
图 2　特征提取子模块

Fig. 2 Feature extraction sub modules

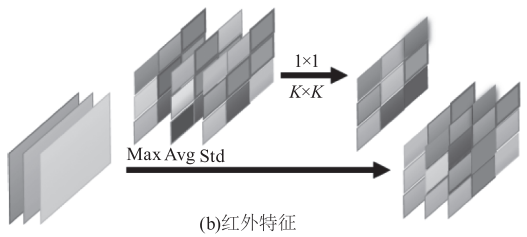
2.3　特征增强与融合

通过特征提取可以分别获取红外和可见光图像特征，但所提特征对目标关注度较低，且未充分利用目标多模态信息。因此，本文设计了混合注意力机制和自适应交叉融合结构来进一步增强目标特征信息。

混合注意力机制主要从目标类别和目标所在图像空间位置角度提升有效特征的权重，考虑到可见光图像中包含丰富的细节和纹理信息，可以较好地区分不同类别目标；而红外图像根据目标发射的热辐射成像，有效屏蔽了背景信息并突出了目标空间位置。因此，所提混合注意力机制以每个 Block 输出特征作为输入，对于可见光支路，注意力结构从特征通道入手，通过最大值、均值以及标准差三个维度充分获取每个通道目标类别的全局信息，并通过 1 ×1 的点卷积融合特全局征后利用 $K \times 1$ 的一维卷积提升通道间的信息交互。最后经 Sigmoid 函数归一化后与对应通道相乘，提升目标类别通道信息权重，并降低背景通道干扰。注意力结构如图 3(a)所示。而对于红外支路，混合注意力机制则聚焦空间位置特征，将所有通道在同一位置的特征作为输入，利用与通道类似的操作计算出每个位置的权重进行加权，增强目标所在位置的特征信息，如图 3(b)结构所示。混合注意力机制权重计算方式如式(1) ~ (4)所示。



(a) 可见光特征



(b)红外特征

图 3　混合注意力机制

Fig. 3 Mixed attention mechanism

$$X = [\mathrm{Max}(x), \mathrm{Avg}(x), \mathrm{Std}(x)] \tag{1}$$

$$W_{\text{visible}} = \sigma(Conv_{K\times1}(Conv_{3\times1}(X))) \tag{2}$$

$$W_{\text{infrared}} = \sigma(Conv_{K\times K}(Conv_{1\times1}(X))) \tag{3}$$

$$K = \frac{1}{2}(\log_2(d) + 1)_{\text{odd}} \tag{4}$$

其中，x 表示输入的通道 / 空间位置信息；X 表示分别计算最大值、均值和标准差；$Conv_{3\times1}$ 表示核为 3 × 1 的卷积操作；同理，$Conv_{K\times1}$ 表示 $K \times 1$ 的卷积操作；σ 表示利用 sigmoid 函数进行归一化；W_{visible} 表示可见光特征权重；W_{infrared} 表示红外特征权重；d 表示输入特征信息维度；odd 表示计算值

取奇数。

自适应交叉融合主要是将红外和可见光支路所提的各维度特征信息进行融合互补，进一步丰富目标特征。由于现有的多模态特征融合大都采用直接相加或拼接的方式，尽管也能提升特征多样性，但引入了较多噪声信息。因此，所提特征融合结构引入了可训练的自适应参数，通过自主加权的方式将红外和可见光对应特征信息进行融合，计算方式如(5)(6)所示，训练时参数调整过程如式(7)所示：

$$y_i = \alpha_i x_i^I + \beta_i x_i^V \tag{5}$$

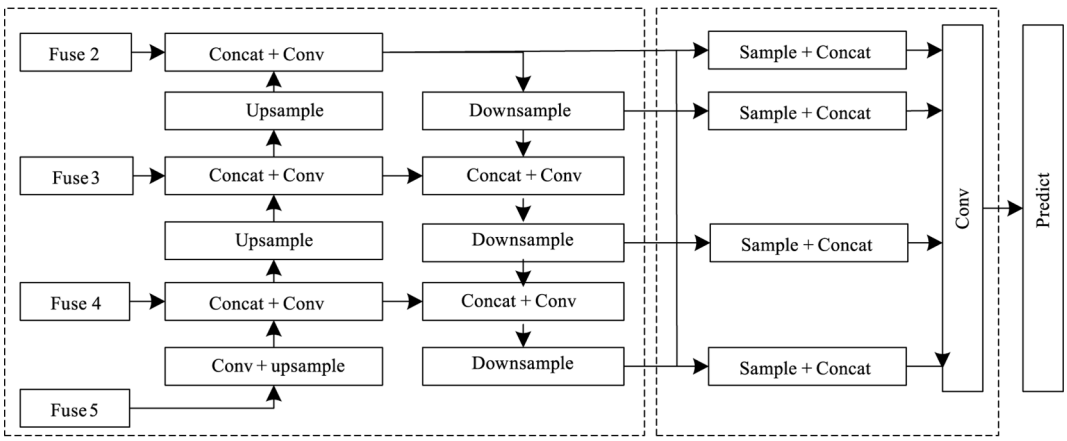
$$\alpha_i, \beta_i \in [0,1] \quad \alpha_i + \beta_i = 1 \tag{6}$$

$$\begin{cases} \dfrac{\partial L}{\partial \alpha_i} = \dfrac{\partial y_i}{\partial \alpha_i}\dfrac{\partial L}{\partial y_i} = \dfrac{\partial L}{\partial y_i} x_i^I \\ \dfrac{\partial L}{\partial \beta_i} = \dfrac{\partial y_i}{\partial \beta_i}\dfrac{\partial L}{\partial y_i} = \dfrac{\partial L}{\partial y_i} x_i^V \end{cases} \tag{7}$$

式中，i 表示特征通道；I 表示红外支路；V 表示可见光支路；x_i^I 表示红外支路第 i 个通道特征；x_i^V 表示可见光支路第 i 个通道特征；α_i 和 β_i 分别为红外和可见光第 i 通道特征的自适应加权参数；L 表示误差；∂ 表示计算偏导数。

2.4 多尺度检测

通过特征提取、注意力增强、多模态融合模块可以由浅到深逐步获取红外和可见光图像局部以及全局特征。而在实际检测任务中，目标的形状和尺寸通常大小不一，若仅用特征提取结构最后一层的输出进行预测，容易导致目标漏检情况。因此，为提升不同大小目标检测的准确性，设计了多尺度特征检测结构，利用多个维度的红外和可见光融合特征，以交替上采样和下采样的方式将深层抽象类别信息与浅层边缘细节信息充分融合，并通过自主选择的方式提取目标关联维度特征进行预测，如图 4 所示。



(a) 交替采样 (b) 特征自主选择

图 4 多尺度检测结构

Fig. 4 Multiscale detection structure

多尺度检测结构主要分为交替采样和自主选择两部分，交替采样部分将红外和可见光交叉融合后的特征作为输入，通过上采样操作将深层特征升维至相邻特征维度后进行 Eltwise 融合，依次升维融合至最浅层，使浅层特征中包含深层信息；同理，为使深层特征中融入浅层信息，将浅层特征再依次进行下采样降维并融合，如图 4(a)所示。特征自主选择主要考虑到不同大小目标通常集中在部分特征层，因此，为充分利用目标所关联特征层信息，同时避免其他层的影响，自主选择结构将不同维度的特征统一采样至相应维度，再通过加权融合的方式提取出目标关联的特征进行预测，如图 4(b)所示。加权计算方式如式(8)、(9)所示。

$$y_c^l = \alpha_c^l x_c^{5\to l} + \beta_c^l x_c^{4\to l} + \chi_c^l x_c^{3\to l} + \delta_c^l x_c^{2\to l} \tag{8}$$

$$\begin{cases} \alpha_c^l, \beta_c^l, \chi_c^l, \delta_c^l, \theta_l^{ij} \in [0,1] \\ \alpha_c^l + \beta_c^l + \chi_c^l + \delta_c^l = 1 \end{cases} \tag{9}$$

式中，c 表示通道；$x_c^{5\to l}$ 表示将交替采样后 fuse5 对应的特征层采样至 l 层维度；α_c^l 表示 $x_c^{5\to l}$ 特征层权重；$\beta_c^l x_c^{4\to l}$、$\chi_c^l x_c^{3\to l}$、$\delta_c^l x_c^{2\to l}$ 同理；根据公式可以看出，当权重参数为 0 时，表示忽略该通道特征，反之为 1 时则

选择该层特征。

3 实验与结果分析

为验证所提目标检测网络的有效性和实用性，实验利用标准数据集以及实际电网设备数据集进行训练测试。所提网络基于 PyTorch 深度学习框架进行搭建，实验平台采用 NVIDIA Jetson Xavier NX AI 边缘计算设备，网络训练过程中超参数配置如表 2 所示。

表 2 实验环境及超参数设置

Tab. 2 Experimental environment and hyperparameter setting

超参数	值
批处理大小	4
初始学习率	0.001
初始化权重	Gaussian
学习率调整	StepLR
动量参数	0.9
权值衰减系数	0.005
网络优化策略	Adam
目标类别损失函数	Cross Entropy Loss
目标位置损失函数	CIou Loss

对于所提目标检测网络的性能评估采用均值平均精度(mAP)和每秒处理图像帧数(FPS)来衡量，计算公式如式(10)~(11)所示。同时，为评估不同尺度目标的识别效果，分别以 mAP_l、mAP_m、mAP_s 来表示大中小目标的检测精度。其中，大中小目标划分借鉴文献[16]设置，以目标标注框中像素数量 32^2 和 96^2 为边界划分目标。

$$mAP = \sum AP_c / N_{\text{Class}} = \left(\sum P_c / N_{\text{image\_c}}\right) / N_{\text{Class}} \tag{10}$$

$$FPS = N_{\text{image}} / \sum_{i}^{N_{\text{image}}} T_i \tag{11}$$

式中，N_{class} 表示目标类别总数；$N_{\text{image\_c}}$ 表示包含 C 类别的图像数；P_c 表示一张图像中 C 类别的识别精度；AP_c 表示所有图像中 C 类目标的平均精度；T_i 表示网络处理第 i 张图像消耗的时间，N_{image} 表示目标检测的总图像。

3.1 标准数据集实验

为验证所提双模态目标检测方法的有效性，实验首先利用了李成龙教授团队公开的标准数据集 RGBT[17] 进行训练测试。该数据集主要由标定好的红外和可见光相机对上百个场景下的目标采集构成，包含不同时间段、不同天气、不同光照强度下的红外和可见光图像对约 210000 张，目标种类约 20 多种。由于数据集中的图像多从连续的视频帧中提取出来，重复度较高，且部分目标数量较少。因此，为更好的验证所提方法，本文只从中筛选出约 10000 张重复率较低的图像对，并确定了 8 类目标，各目标占比如表 3 所示。将图像尺寸调整为 512 × 512 大小后以 7∶1∶2 的比例随机划分训练、验证、测试集进行实验。

表 3 数据集各目标占比

Tab. 3 Proportion of each target in the dataset

目标	占比/%	目标	占比/%
person	25	umbrella	9
bicycle	13	dog	10
car	20	motocycle	11
kite	8	toy	4

实验首先针对支路的基础网络进行训练测试，基础网络即输入为原图、特征提取结构无注意力模块、检测部分为 FPN 网络检测结构。由于红外支路和可见光支路基本对称，故只对可见光支路进行了测试，并将测试结果与目前主流的轻量级目标检测网络进行了对比，对比结果如表 4 所示。

表 4 可见光支路基础网络测试对比

Tab. 4 Comparison of basic network of visible light branch

网络	*FPS*	测试精度/%			
		mAP	mAP_s	mAP_m	mAP_l
Shuffle Netv2[18]	39	71.0	49.8	71.6	78.9
Mobile Netv3[19]	34	72.1	51.0	72.9	80.1
Efficient Netv2[14]	32	71.6	50.7	72.5	79.4
Ghost Net[20]	35	72.7	51.6	73.2	80.6
BaseNet	46	68.3	45.6	69.3	76.4

由表4可以看出,由于所提网络针对目标双模态信息,为保证双支路结构的高效性,支路构建采用了较少的特征通道来保障检测效率,但也损失了部分特征信息,使检测效果较差。而主流的轻量级主要针对单源输入,网络结构相对双源网络的支路更为复杂,提取信息更多,故精度相对较高,但效率较低。为丰富支路特征信息,引入了图像增强模块,针对该模块的有效性验证,实验通过依次引入不同图像处理方法来对比检测精度变化,结果如表5所示。

表5 图像增强模块测试对比

Tab. 5 Test comparison of image enhancement module

	origin	HSV	Gray	Canny	MF	HE	FPS	mIoU
可见光	√	/	/	/	/	/	46	68.3
	√	√	/	/	/	/	46	68.5
	√	√	√	/	/	/	45	68.7
	√	√	√	√	/	/	44	69.0
	√	√	√	√	√	/	44	69.1
	√	√	√	√	√	√	43	69.1
红外	√	/	/	/	/	/	46	64.0
	√	/	/	√	/	/	44	64.1
	√	/	/	√	√	/	44	64.3
	√	/	/	√	√	√	43	64.5

根据表5可以看出,不同的预处理方法对红外和可见光支路检测精度的影响也各不相同(红外图像为灰度图,无法进行HSV和Gray处理)。其中,对于可见光图像,HSV颜色空间转换、Canny边缘提取以及Gray灰度转化方法提升较大;而对于红外图像,直方图均衡化、均值滤波等方法的提升效果更佳。为进一步提升特征提取过程中关键特征贡献,设计了混合注意力模块,针对该模块的有效性验证,实验基于可见光支路网络,分别对比了不同注意力机制对检测性能的影响,结果如表6所示。其中,混合注意力在特征通道和特征空间位置上同时使用。

表6 注意力机制对比

Tab. 6 Comparison of attention mechanism

网络	*FPS*	测试精度/%			
		mAP	mAP_s	mAP_m	mAP_l
可见光支路	44	69.0	46.1	69.9	77.2
SE注意力	43	69.3	46.5	70.3	77.9
CBAM注意力	39	70.4	47.9	71.1	78.5
混合注意力	41	70.9	48.5	71.6	79.0

由上表可见,SE注意力机制仅针对通道特征,故在效率上相对较高,但精度提升相对较少;CBAM注意力机制虽同时考虑通道和空间位置特征,但仅通过特征最大值来表示全局信息过于局限,且全连接方式也引入了较多计算量;而所提混合注意力机制以多种方式对全局信息建模,并利用一维卷积替代全连接,提升精度的同时也保障了计算效率。为进一步验证通道和空间混合注意力对红外和可见光特征的影响,实验对比了不同模态特征在不同注意力下网络性能变化,结果如表7所示。

表7 通道和空间注意力对比

Tab. 7 Comparison of channel and spatial attention

网络	通道	空间	FPS	mAP
红外支路	√	/	43	64.8
	/	√	42	65.3
	√	√	41	65.4
可见光支路	√	/	43	70.8
	/	√	42	70.2
	√	√	41	70.9

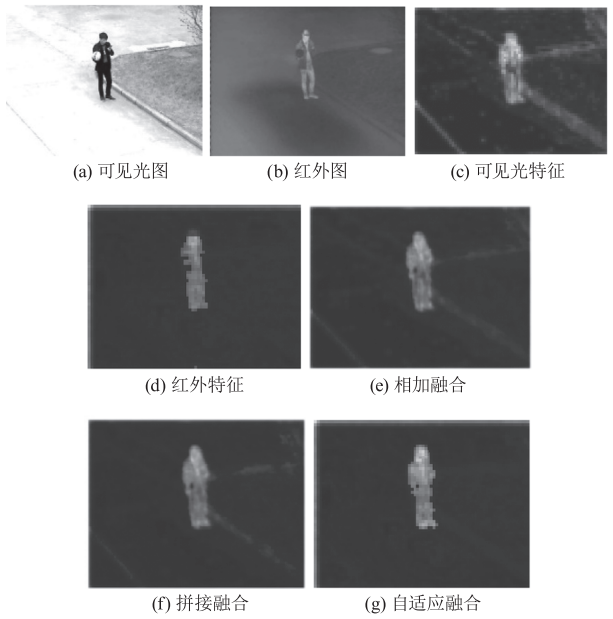
可见,通道混合注意力可以较好的聚焦可见光特征,空间位置混合注意力则更适用于增强红外特征信息。因此,为保证网络整体效率,在红外支路中仅适用了空间位置注意力机制,而可见光支路中也

只使用了通道注意力机制。由于图像增强和注意力机制都是针对单支路特征，而多模态特征有效融合才能体现出双支路网络的优越性。对于所提自适应融合模块的可行性验证，实验分别测试对比了不同融合方式下目标检测精度的变化，并通过特征可视化进一步体现所提融合模块的有效性，实验结果如表 8 和图 5 所示。

表 8　双模态特征融合对比

Tab. 8 Comparison of bimodal feature fusion

网络	*FPS*	测试精度/%			
		mAP	mAP_s	mAP_m	mAP_l
红外支路	42	65. 3	42. 8	64. 6	73. 1
可见光支路	43	70. 8	48. 4	71. 0	78. 7
相加融合	23	73. 5	52. 6	74. 1	81. 4
拼接融合	23	74. 1	53. 8	74. 9	82. 7
自适应融合	23	74. 9	54. 7	75. 3	83. 8



(a) 可见光图　(b) 红外图　(c) 可见光特征

(d) 红外特征　(e) 相加融合

(f) 拼接融合　(g) 自适应融合

图 5　Block3 层特征融合可视化对比

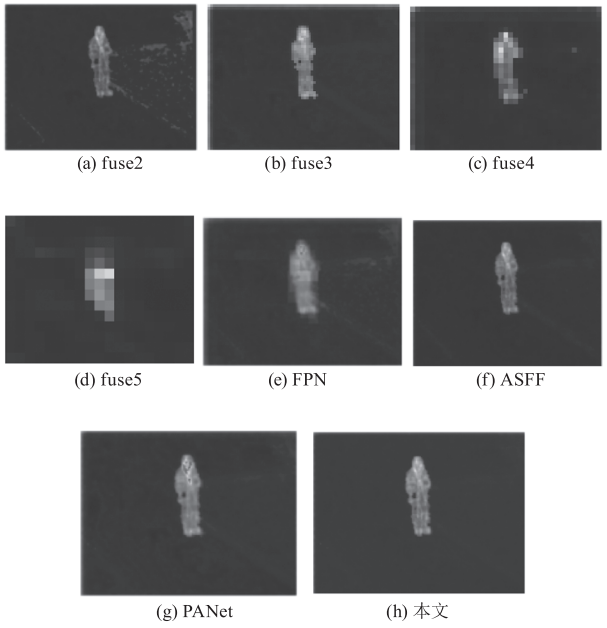
Fig. 5 Visual comparison of feature fusion of Block3 layer

根据表 8 结果可以看出，相较于直接相加和拼接的融合方式，自适应融合对目标检测精度提升最大。同时，从特征可视化效果中也可看出，所提方法在丰富目标信息的同时有效避免了无效特征的干扰，而相加和拼接方式虽然也增强了目标信息，但也引入了较多的噪声。对于多尺度检测结构的验证，实验分别与当前主流的多尺度方法 FPN、ASFF 以及 PANet 进行了对比，实验结果如表 9 所示。同时，为进一步体现所提多尺度结构的有效性，将 fuse3 层对应维度的多尺度特征图进行可视化展示，如图 6 所示。

表 9　多尺度结构对比

Tab. 9 Multiscale structure comparison

网络	*FPS*	测试精度/%			
		mAP	mAP_s	mAP_m	mAP_l
FPN	23	74. 9	54. 7	75. 3	83. 8
ASFF	21	76. 2	55. 9	76. 7	84. 8
PANet	22	75. 9	55. 2	76. 3	84. 4
本文	20	76. 7	56. 5	77. 3	85. 4



(a) fuse2　(b) fuse3　(c) fuse4

(d) fuse5　(e) FPN　(f) ASFF

(g) PANet　(h) 本文

图 6　多尺度特征融合可视化对比

Fig. 6 Visual comparison of multi-scale feature fusion

根据上述实验结果可以看出，传统的 FPN 结构在不同维度特征融合时引入了较多无效信息，且不同维度目标特征容易相互干扰；ASFF 和 PANet 结构虽然在一定程度上缓解了不同维度特征间的信息干扰，但仍存在一定局限；而所提方法在尽可能保证网络效率的同时充分吸取了现有多尺度结构优势，使网络对不同大小目标的检测效果都有较大改善，检测效果也达到了最优。综上实验结果有效验证了所提各个模块的可行性，而对于整个网络的有效性验证，实验与同类型红外和可见光目标检测方法进行了对比，实验结果如表 10 所示，检测效果如图 7 所示。

表10 同类型网络测试对比

Tab. 10 Comparison of network tests of the same type

网络	*FPS*	测试精度/%			
		mAP	mAP_s	mAP_m	mAP_l
文献[6]	22	73.7	52.7	75.6	82.5
文献[7]	16	76.1	55.8	77.3	84.6
文献[8]	14	74.3	53.4	76.9	83.8
本文	20	76.7	56.5	77.3	85.4

(a) 可见光图 (b) 红外图 (c) 文献[6]

(d) 文献[7] (e)文献[8] (f)本文

图7 同类型网络检测效果对比

Fig. 7 Comparison of network detection effects of the same type

根据上述对比结果可见,文献[6]方法通过融合浅层特征的方式虽然效率较高,但红外和可见光信息融合有限,目标检测效果相对较差;文献[7]虽采用了将红外特征融入可见光网络中来丰富目标网络特征信息,但融合过程相对简单,对目标多模态以及多尺度信息提取不够充分,检测精度提升相对有限;文献[8]通过分别检测再融合检测结果的方式过于冗余,且仅是对检测结果的融合,忽略了特性互补性,故在检测精度及效率上都表现一般;而所提网络从多个角度来增强目标特征,并利用自适应的融合方式来实现目标不同模态、不同维度特征的互补,进而使网络整体检测效果达到最优。

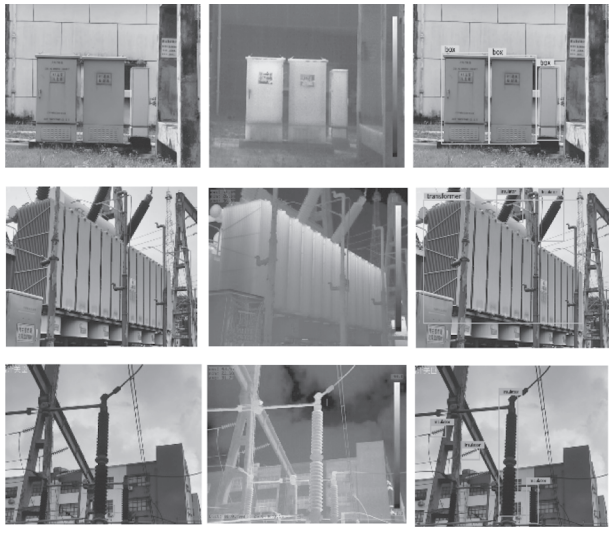
3.2 实际场景实验

为进一步验证所提网络在实际场景中的检测效果,实验利用真实电网环境下的设备来测试目标检测网络的实用性。对于实验数据集的构建,主要通过巡检机器人搭载的红外和可见光相机进行采集(红外和可见光相机通过水平标定后再利用裁剪使图像对达到像素级对齐)。为更好的验证所提网络,所采集的设备图像涵盖了不同光照、不同天气等情况,并且设备间存在尺寸差距较大的目标。实验共筛选了约4000组大小为512×512的图像对,包含变压器、冷控箱、断路器、绝缘子等6类目标,通过Labelimg工具对图像中各目标进行人工标注后以7∶1∶2的比例随机划分训练验证和测试集后进行训练测试,实验结果如表11和图8所示。

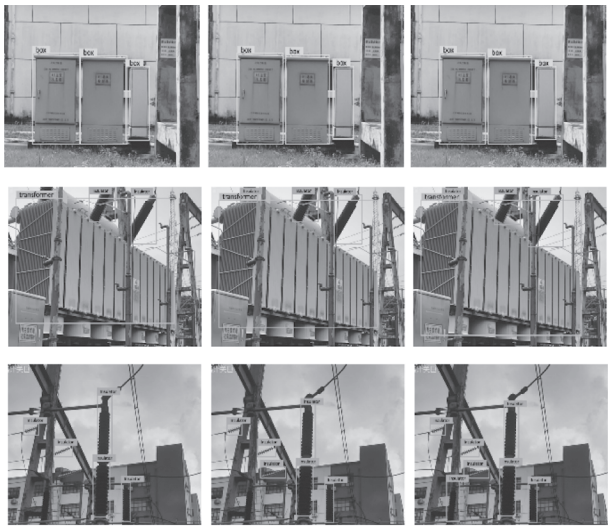
表11 实际电网设备测试对比

Tab. 11 Test comparison of actual power grid equipment

网络	*FPS*	测试精度/%			
		mAP	mAP_s	mAP_m	mAP_l
文献[6]	23	84.7	64.9	85.8	92.3
文献[7]	17	86.9	66.4	87.4	94.1
文献[8]	14	85.9	65.7	86.7	93.0
本文	21	87.8	67.1	87.9	94.4



(a) 原图 (b) 标注图 (c) 文献[6]



(d) 文献[7] (e)文献[8] (f)本文

图 8 电网设备目标检测效果

Fig. 8 Object detection effect of power grid equipment

根据实际场景中的检测结果可以看出,由于数据集中目标的复杂程度相对较低,各目标检测精度都有一定提升,而与同类型方法相比,所提双模态目标检测方法在较大和较小目标的检测中表现更佳,且网络整体检测精度仍保持最优,有效验证了该方法泛化性以及落地实际应用的可行性。尽管检测效率未达到最高,但在实际电网设备目标检测过程中,巡检机器人行驶速度相对较慢,所提方法基本可以满足其实时检测的需求。

4 结 语

针对目前红外和可见光双模态目标检测方法存在的不足,本文从图像输入、特征提取、特征融合、多尺度特征角度进行了深入分析,提出了一种基于特征增强的目标检测网络。该网络针对红外和可见光图像,利用颜色空间转换、边缘提取、均值滤波等方法分别对原始图像进行预处理操作,丰富网络输入信息。在特征提取阶段,采用了相对较少的特征通道来降低冗余信息提取,保障整体网络效率,并设计了混合注意力机制,从可见光通道特征和红外空间位置特征角度提升目标相关特征贡献。同时,为充分互补目标双模态信息,引入了自适应交叉融合结构,通过训练的方式自适应加权红外和可见光特征,降低了背景信息的干扰。目标检测阶段,对于不同尺度目标,采用了两次采样充分融合目标深层和浅层信息,并以自主选择的方式提取目标相关维度特征进行预测,避免了各目标特征间相互干扰。通过实验表明,所提方法有效增强了输入图像信息、特征多样性信息以及多尺度特征信息,并且与同类型方法相比,该方法也体现出较高的鲁棒性和实用性,可以准确高效完成目标检测。虽然所提方法检测效果较优,但效率上仍有较大的提升空间,在后续工作中将探索模型剪枝和知识蒸馏方法进一步优化网络。

参考文献:

[1] LI Kecen, WANG Xiaoqiang, LIN Hao, et al. Survey of one - stage small object detection methods in deep learning[J]. Journal of Frontiers of Computer Science & Technology, 2022, 16(1): 41 - 58. (in Chinese)
李科岑,王晓强,林浩,等. 深度学习中的单阶段小目标检测方法综述[J]. 计算机科学与探索, 2022, 16(1): 41 - 58.

[2] LI Zhangwei, HU Anshun, WANG Xiaofei. Survey of vision based object detection methods[J]. Computer Engineering and Applications, 2020, 56(8): 1 - 9. (in Chinese)
李章维,胡安顺,王晓飞. 基于视觉的目标检测方法综述[J]. 计算机工程与应用, 2020, 56(8): 1 - 9.

[3] Khan A A, Laghari A A, Awan S A. Machine learning in computer vision: a review[J]. EAI Endorsed Transactions on Scalable Information Systems, 2021, 08(32): 4 - 14.

[4] Zaidi S, Ansari M, Aslam A, et al. A survey of modern deep learning based object detection models[J]. Digital Signal Processing, 2022, 126: 103514.

[5] SONG Wenshu, HOU Jianmin, CUI Yuy. A method of intelligent target detection based on multi-source information fusion[J]. Video Engineering, 2021, 45(6): 101 - 105. (in Chinese)
宋文姝,侯建民,崔雨勇. 基于多源信息融合的智能目标检测技术[J]. 电视技术, 2021, 45(6): 101 - 105.

[6] 顾晶晶,胡俊. 基于遥感图像的多模态小目标检测方法及系统: 中国, CN 202210072288. 8[P]. 2022 - 01 - 21.

[7] KUANG Chuwen, HE Wang. Object detection algorithm based on infrared and visible light images[J]. Infrared Technology, 2022, 44(9): 912 - 919. (in Chinese)
邝楚文,何望. 基于红外与可见光图像的目标检测算法[J]. 红外技术, 2022, 44(9): 912 - 919.

[8] Banuls A, Mandow A, Vazquez-Martin R, et al. Object detection from thermal infrared and visible light cameras in search and rescue scenes[C]//2020 IEEE International Symposium on Safety, Security, and Rescue Robotics (SSRR). IEEE, 2020: 380 – 386.

[9] Ma J, Tang L, Xu M, et al. STDFusionNet: an infrared and visible image fusion network based on salient target detection[J]. IEEE Transactions on Instrumentation and Measurement, 2021, 70: 1 – 13.

[10] XU Degang, WANG Lu, LI Fan. Review of typical object detection algorithms for deep learning[J]. Computer Engineering and Applications, 2021, 57(8): 10 – 25. (in Chinese)
许德刚, 王露, 李凡. 深度学习的典型目标检测算法研究综述[J]. 计算机工程与应用, 2021, 57(8): 10 – 25.

[11] Elahe A, Shruthi G, Ratnajit M, et al. A comprehensive study of real-time object detection networks across multiple domains: a survey[J]. arXiv Preprint arXiv, 2022, 2208: 10895.

[12] Ren S, He K, 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.

[13] Wang C, Bochkovskiy A, Liao H. YOLOv7: trainable bag-of-freebies sets new state-of-the-art for real-time object detectors[J]. arXiv Preprint arXiv, 2022: 2207. 02696.

[14] Tan M, Le Q. Efficientnetv2: smaller models and faster training [C]//International Conference on Machine Learning. PMLR, 2021: 10096 – 10106.

[15] Zhou X, Wang D, Philipp K. Objects as points[J]. arXiv Preprint arXiv, 2019: 1904. 07850.

[16] Lin T Y, Maire M, Belongie S, et al. Microsoft coco: common objects in context [C]//European Conference on Computer Vision. Springer, Cham, 2014: 740 – 755.

[17] Li C, Zhao N, Lu Y, et al. Weighted sparse representation regularized graph learning for RGB-T object tracking [C]//Acm on Multimedia Conference. ACM, 2017: 1856 – 1864.

[18] Ma N, Zhang X, Zheng H T, et al. Shufflenet v2: practical guidelines for efficient cnn architecture design[C]//Proceedings of the European Conference on Computer Vision (ECCV), 2018: 116 – 131.

[19] Howard A, Sandler M, Chu G, et al. Searching for mobilenetv3[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019: 1314 – 1324.

[20] Han K, Wang Y, Tian Q, et al. Ghostnet: more features from cheap operations[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020: 1580 – 1589.