基于自适应注意力机制的红外与可见光图像目标检测

赵松璞，杨利萍，赵 昕，彭志远，梁东兴，梁洪军
（深圳市朗驰欣创科技股份有限公司成都分公司，四川 成都 610000）

摘要：针对红外和可见光目标检测方法存在的不足，将深度学习技术与多源目标检测相结合，提出了一种基于自适应注意力机制的目标检测方法。该方法首先以深度可分离卷积为核心构建双源特征提取结构，分别提取红外和可见光目标特征。其次，为充分互补目标多模态信息，设计了自适应注意力机制，以数据驱动的方式加权融合红外和可见光特征，保证特征充分融合的同时降低噪声干扰。最后，针对多尺度目标检测，将自适应注意力机制结合多尺度参数来提取并融合目标全局和局部特征，提升尺度不变性。通过实验表明，所提方法相较于同类型目标检测算法能够准确高效地在复杂场景下实现目标识别和定位，并且在实际变电站设备检测中，该方法也体现出更高的泛化性和鲁棒性，可以有效辅助机器人完成目标检测任务。

关键词：红外与可见光；目标检测；深度学习；自适应注意力机制
中图分类号：TP391.41 **文献标志码：**A **文章编号：**1001-8891(2024)04-0443-09

Object Detection in Visible Light and Infrared Images Based on Adaptive Attention Mechanism

ZHAO Songpu，YANG Liping，ZHAO Xin，PENG Zhiyuan，LIANG Dongxing，LIANG Hongjun
(*Shenzhen Launch Digital Technology Co. Ltd. Chengdu Branch*, *Chengdu* 610000, *China*)

Abstract: To address the shortcomings of infrared and visible light object detection methods, a detection method based on an adaptive attention mechanism that combines deep learning technology with multi-source object detection is proposed. First, a dual-source feature extraction structure is constructed based on deep separable convolution to extract the features of infrared and visible objects. Second, an adaptive attention mechanism is designed to fully complement the multimodal information of the object, and the infrared and visible features are weighted and fused using a data-driven method to ensure the full fusion of features and reduce noise interference. Finally, for multiscale object detection, the adaptive attention mechanism is combined with multiscale parameters to extract and fuse the global and local features of the object to improve the scale invariance. Experiments show that the proposed method can accurately and efficiently achieve target recognition and localization in complex scenarios compared to similar object detection algorithms. Moreover, in actual substation equipment detection, this method also demonstrates higher generalization and robustness, which can effectively assist robots in completing object detection tasks.

Key words: infrared and visible light, object detection, deep learning, adaptive attention mechanisms

0 引言

目标检测技术是机器视觉方向重要研究课题之一，其核心任务是对图像中所关注的目标进行识别，并标注出目标类别及位置[1]。现阶段大多数目标检测方法主要利用目标在单一波段上的成像作为输入源，如红外图像或可见光图像[2]。红外图像根据目标物体热辐射能量进行成像，不依赖于其他光线，可以较好地应用于夜间、烟雾等环境，但图像对比度较低、细节缺失严重[3]；而可见光图像利用目标反射的自然光进行成像，可以较好地获取目标细节和纹理信息，但却容易受到光照强弱、目标反射率等影响[4]。可见，单一传感器获取目标信息时存在一定局限，而随着目标所处环境逐渐复杂化，其局限性也将不断扩大，进而影响目

收稿日期：2022-08-30；**修订日期：**2022-09-28.
作者简介：赵松璞（1973-），男，汉族，陕西西安人，硕士，工程师。研究方向：机器人技术、智能电网、模式识别。E-mail: 1419446206@qq.com。
基金项目：深圳市科技计划项目（JSGG20210802153009029）。

标检测效果[5]。因此，设计一种基于红外和可见光的目标检测方法，不仅可以丰富目标多模态特征，而且对目标检测性能提升以及实际应用价值都有较大的促进作用。

目前，大多数基于红外与可见光的目标检测方法仍采用传统图像处理方法，如引导滤波结合最小加权二乘法[6]、SIFT 结合 BOW（Bag-of-Words）模型[7]、图像低秩和显著信息分解再加权融合[8]等。传统方式通常在特定场景下检测精度较高，但其泛化性较弱，并且对于复杂环境下的目标检测效果较差。而随着深度学习技术以及计算机性能的不断突破，部分研究者开始逐渐将多源目标检测与卷积神经网络相结合，并取得了较好的效果。Hui 等[9]人针对红外和可见光特征融合提出了一种新型深度学习结构，通过稠密编码器丰富所提取的目标特征，再利用解码器对特征进行直接相加融合，虽然提升了检测精度，但稠密连接方式计算量较大，且融合方式比较粗糙。唐聪等[10]人通过在训练好的可见光目标检测网络基础上微调出红外检测网络，间接共享目标特征，并结合红外和可见光网络结果实现目标检测。该方式采用了两个网络实现检测，在一定程度上互补了目标多模态特征，但检测过程繁琐，且对目标信息利用不够充分。Ma 等[11]人提出了一种显著目标检测方法，通过设计显著目标模板来选择性地提取并融合红外热目标特征和可见光纹理结构，实现关键目标识别检测，但该方法只针对显著目标检测和关键点识别，对小目标识别效果较差，且容易受到高频噪声干扰。由此可见，现有红外-可见光目标检测方法在特征提取的有效性、特征融合充分性以及检测方法的鲁棒性和泛化性等方面仍有较大的提升空间。

针对上述红外-可见光目标检测方法存在的不足，本文在总结现有研究基础上，提出了一种基于自适应注意力机制的红外与可见光目标检测方法。该方法以高效率的深度可分离卷积为基础，分别构建红外和可见光特征提取网络，提取目标多模态特征。其次，设计自适应注意力机制结构（adaptive attention mechanisms, AAM），将提取的红外和可见光特征以自主学习的方式加权融合，提升有效特征权重，并丰富目标特征信息。同时，为保证不同大小目标准确识别定位，将融合后的特征同样以自适应注意力机制方式进行多尺度自主叠加，降低不同维度目标相互干扰，保障目标多尺度不变性。

1 目标检测结构设计

1.1 整体结构

所提自适应注意力机制的红外-可见光目标检测方法整体结构如图 1 所示，主要由双源特征提取网络、AAM 特征融合以及多尺度检测 3 部分组成。双源特征提取网络以深度可分离卷积为基本，结合池化、激活、残差等操作，构建成对的深层特征提取结构，分别提取目标红外特征和可见光特征。AAM 特征融合结构采用自适应的通道和空间注意力机制来分别提升目标类别及定位特征权重，并以自主学习的方式将红外和可见光特征进行融合，降低噪声干扰。而多尺度检测将不同层次的融合特征采样至相同维度，并再次利用自适应注意力机制，使网络自主选择目标所处特征层，避免不同层次特征信息相互影响。整个网络以深度可分离卷积保障了特征提取的高效性，并以自适应注意力机制提升了特征融合的有效性以及多尺度检测的准确性。

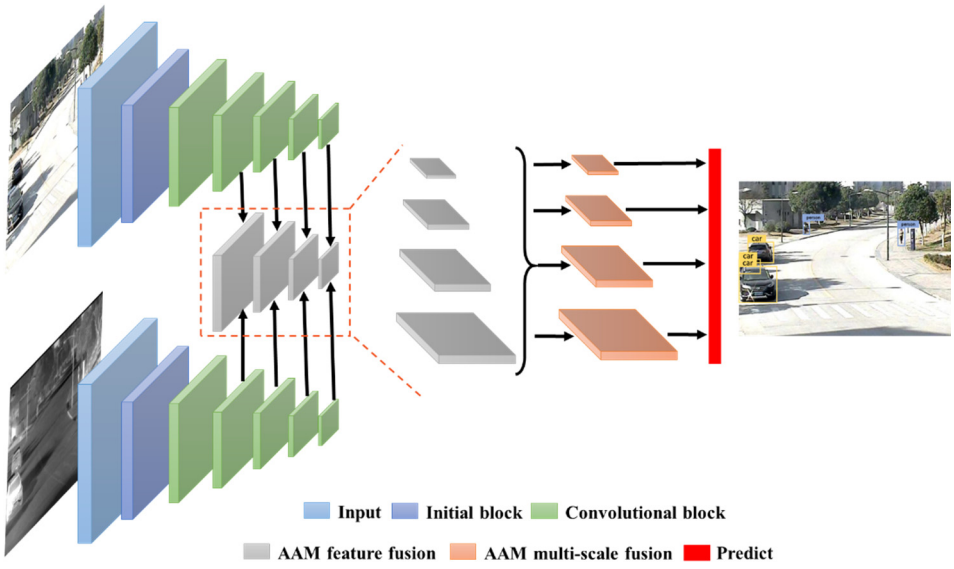


图 1 红外-可见光目标检测整体架构

Fig.1 Overall framework of infrared-visible light object detection

1.2 双源特征提取网络

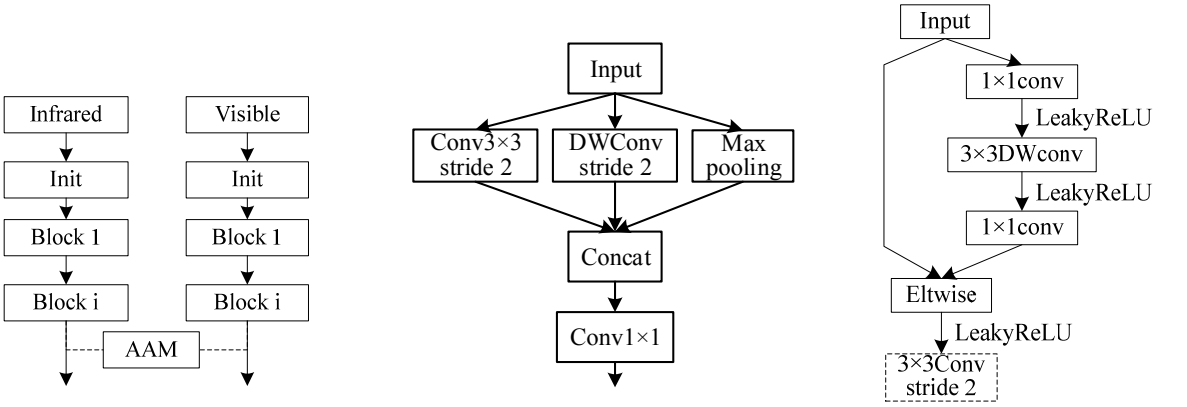
特征提取是计算机视觉任务的关键，所提特征的优劣直接决定了视觉任务的效果[12]。传统的特征提取方法主要根据对目标呈现形态的认知进行建模，如Harris、SIFT（scale-invariant feature transform）、HOG（histogram of oriented gradients）、DMP（deformable parts model）等[13]。虽有较强的理论支撑，但调参过程复杂，且各个算法都针对具体应用，鲁棒性和泛化性都较差。而基于深度学习的卷积神经网络作为当前主流的特征提取方法，采用数据驱动的方式提取特征，避免了人工特征建模的局限，且所提特征可以更好地对目标进行表示[14]。同时，随着近几年深度学习的深入，逐渐沉淀出了一批经典的特征提取网络，如DarkNet[15]、ResNet[16]、MobileNet[17]、AdderNet[18]等。为有效提取目标特征信息，本文借鉴了现有特征提取方法，构建了适用于红外-可见光目标检测的轻量级特征提取网络。

由于输入源为红外和可见光图像，所提特征提取网络采用对称双支路结构，如图2(a)所示，其中，支路详细结构如表1所示。该结构由初始化模块（init）和多个卷积模块（block）串联堆叠组成，初始化模块如图2(b)所示，采用步长为2的3×3标准卷积、3×3深度可分离卷积以及2×2最大池化操作，以并行处理的方式从多个角度提取输入图像特征。该模块主要是尽可能避免目标有效信息丢失的同时降低输入图像维度，并减少噪声干扰。而block卷积模块作为特征提取的关键部分，主要以深度可分离卷积为核心，结合激活函数、残差结构实现对目标由浅到深的提取特征，如图2(c)所示。该模块以深度可分离卷积代替标准卷积，并通过1×1的点卷积调整特征通道数量，有效降低了网络参数量，保障了双支路特征提取结构的计算效率。尽管深度可分离卷积损失了部分特征，但双支路结构的特征互补特性有效弥补了该缺陷。同时，为缓解深层网络训练时梯度消失等问题，引入了残差结构，并以LeakyReLU函数作为激活函数，降低无效神经元的产生，加速网络收敛。其中，block模块内的卷积操作步长都为1，block块最后一层步长为2，如图2(c)虚线部分。

表1 特征提取支路

Table 1 Feature extraction branch

Module	Layer	Repetitions	Output
Input	RGB, 3	1	512×448
Init	Conv 3×3, 10 DWconv 3×3, 3 Max pooling 2×2, 3	1	256×224
Block 1	DWconv 3×3, 32 Residual	1	128×112
Block 2	DWconv 3×3, 64 Residual	2	64×56
Block 3	DWconv 3×3, 128 Residual	3	32×28
Block 4	DWconv 3×3, 256 Residual	3	16×14
Block 5	DWconv 3×3, 512 Residual	2	8×7



(a) 双源头特征提取 (b) 初始化模块 (b) Block 卷积模块

(a) Dual source feature extraction (b) Initialization module (b) Block convolution module

图2 特征提取模块

Fig.2 Feature extraction modules

1.3 AAM 特征融合

对于多源数据的计算机视觉任务，其关键在于信息融合，而特征融合是目前最为常见融合方式之一[19]。现有的特征融合通常采用特征拼接、特征叠加等方式[20]，这种无差别的融合方式在丰富信息的同时也引入了较多无效信息。因此，为提升特征融合的有效性，本文设计了自适应注意力机制的特征融合结构，通过数据驱动的方式自适应调整红外和可见光特

征融合权重，降低无效信息干扰，示意图如图 3 所示。考虑到过浅层特征中噪声较多，网络只选择了 block 2～block 5 的特征进行融合，即 i=2,3,4,5。

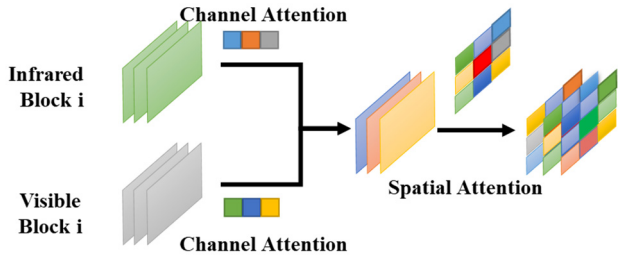


图 3 AAM 特征融合

Fig.3 AAM feature fusion

融合结构以 block 模块的输出作为输入，先通过批量归一化操作规范化红外和可见光特征权重后，再利用自适应注意力机制将两类特征进行融合。而自适应注意力机制又分为通道和空间两个注意力模块，通道注意力针对红外和可见光的每个特征通道进行自适应加权融合，提升目标类别所属特征通道的权重，计算方式如式(1)所示。空间注意力则是针对通道注意力融合后的所有特征通道，对不同空间位置上的特征进行自适应加权，提升目标所处位置权重，计算方式如式(2)所示。

$$y_c=[\alpha_c \ \ \beta_c]\begin{bmatrix}x_c^{\mathrm{V}}\\ x_c^{\mathrm{I}}\end{bmatrix} \tag{1}$$

$$y_s^{ij}=\theta^{ij}y_c^{ij} \tag{2}$$

式中：x_c^{V} 为可见光第 c 个通道特征；x_c^{I} 为红外第 c 个通道特征；α_c 为可见光 c 通道权重；β_c 为红外 c 通道权重；y_c 为红外和可见光 c 通道注意力融合输出；y_c^{ij} 为通道注意力融合后第(i,j)位置的特征；θ^{ij} 为特征图(i,j)位置权重；y_s^{ij} 为空间注意力输出。同时，各权重满足α_c, β_c, $\theta^{ij}\in[0,1]$，且$\alpha_c+\beta_c=1$，训练时通过误差反向传播方式调整各参数权重，如式(3)(4)(5)所示。

$$\frac{\partial L}{\partial y_c^{ij}}=\theta^{ij}\frac{\partial L}{\partial y_s^{ij}},\quad \frac{\partial L}{\partial \theta^{ij}}=\frac{\partial L}{\partial y_s^{ij}}y_c^{ij} \tag{3}$$

$$\begin{bmatrix}\dfrac{\partial L}{\partial x_c^{\mathrm{V}}}\\ \dfrac{\partial L}{\partial x_c^{\mathrm{I}}}\end{bmatrix}=[\alpha_c \ \ \beta_c]\begin{bmatrix}\dfrac{\partial L}{\partial y_c}\\ \dfrac{\partial L}{\partial y_c}\end{bmatrix} \tag{4}$$

$$\frac{\partial L}{\partial \alpha_c}=\frac{\partial L}{\partial y_c}x_c^{\mathrm{V}},\quad \frac{\partial L}{\partial \beta_B}=\frac{\partial L}{\partial y_c}x_c^{\mathrm{I}} \tag{5}$$

式中：L 为训练误差；∂为偏导计算。由上式可以看出，当通道注意力中的α_c为 0 时，其对应可见光特征通道被认为是无效信息，不参与融合；反之，红外特征类似。同理，当空间注意力中θ^{ij}为 0 时，该位置被认为背景。由此可见，当网络训练时，通过误差反向传播自适应调整上述权重参数，可以有效抑制噪声的干燥。

1.4 多尺度检测

特征提取实现了目标特征由浅到深的提取，特征融合丰富了各层次特征信息，而对于不同尺度目标的检测，需要综合多个层次的特征信息。常见的多尺度检测结构主要基于特征金字塔的方式[21]将深层和浅层特征逐层融合检测，但文献[22]研究发现不同尺度的目标通常集中在部分特征层，若将目标主要所在的特征层次与其他层次特征融合，反而会干扰对该尺度目标的检测。因此，为缓解不同层次特征之间相互干扰问题，本文在自适应注意力机制基础上调整输入特征，将其应用于目标多尺度检测结构中，以自主选择的方式实现各尺度目标检测，检测结构如图 4 所示。

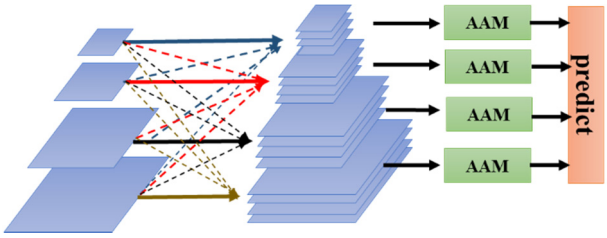


图 4 AAM 多尺度检测

Fig.4 AAM multiscale detection

该结构以红外和可见光融合后的特征作为输入，而不同 block 融合后的特征层维度不同，需要分别将其他层的特征上采样或下采样至当前特征维度，再利用自适应注意力机制进行特征加权，最后，根据加权融合后的多尺度特征进行检测。检测部分综合考虑网络精度与效率后采用 YOLO[15]单阶段检测结合非极大值抑制算法（non-maximum suppression, NMS）实现最终目标定位识别。多尺度特征自适应注意力加权融合计算公式如式(6)(7)所示：

$$y_c^l=\alpha_c^l x_c^{5\to l}+\beta_c^l x_c^{4\to l}+\chi_c^l x_c^{3\to l}+\delta_c^l x_c^{2\to l} \tag{6}$$

$$y_{lc}^{ij}=\theta_l^{ij}y_l^{ij} \tag{7}$$

式中：$x_c^{5\to l}$ 表示 block 5 融合层的第 c 个通道采样至 l 层维度；α_c^l 为 $x_c^{5\to l}$ 特征层对应权重，即 block 5 融合层采样后的第 c 个通道权重；另外 3 个加权参数同理；y_c^l 表示 l 层 c 通道与其他层 c 通道自适应融合输出；θ_l^{ij} 为 l 层特征图(i,j)位置权重；y_l^{ij} 为 l 多尺度通道融合后第(i,j)位置的特征；y_{lc}^{ij} 为 l 层 c 通道空间注意力输出。同时，各参数满足α_c^l, β_c^l, χ_c^l, δ_c^l,$\theta_l^{ij}\in[0,1]$，且$\alpha_c^l+\beta_c^l+\chi_c^l+\delta_c^l=1$。同理，训练时利用误差反向传播自适应调整参数，如式(8)(9)所示：

$$\frac{\partial L}{\partial y_l^{ij}}=\theta_l^{ij}\frac{\partial L}{\partial y_{ls}^{ij}},\quad \frac{\partial L}{\partial \theta_l^{ij}}=\frac{\partial L}{\partial y_{ls}^{ij}}y_l^{ij} \tag{8}$$

$$\begin{aligned}\frac{\partial L}{\partial x_c^5}&=\frac{\partial y_c^5}{\partial x_c^5}\frac{\partial L}{\partial y_c^5}+\frac{\partial x_c^{5\to4}}{\partial x_c^5}\frac{\partial y_c^4}{\partial x_c^{5\to4}}\frac{\partial L}{\partial y_c^4}+\\&\frac{\partial x_c^{5\to3}}{\partial x_c^5}\frac{\partial y_c^3}{\partial x_c^{5\to3}}\frac{\partial L}{\partial y_c^3}+\frac{\partial x_c^{5\to2}}{\partial x_c^5}\frac{\partial y_c^2}{\partial x_c^{5\to2}}\frac{\partial L}{\partial y_c^2}\\&=\alpha_c^5\frac{\partial L}{\partial y_c^5}+\alpha_c^4\frac{\partial L}{\partial y_c^4}+\alpha_c^3\frac{\partial L}{\partial y_c^3}+\alpha_c^2\frac{\partial L}{\partial y_c^2}\end{aligned} \tag{9}$$

式中：$\frac{\partial L}{\partial y_c^5}$ 表示对 block 5 融合层求偏导，block2 3 4 层类似。由上述公式可以看出，当α_c^l为 0 时，block5 层 c 通道特征被忽略，可理解为该通道不存在当前尺度目标。由此可见，通过自适应调整α_c^l, β_c^l, χ_c^l, δ_c^l参数，可以实现网络自主选择目标对应尺度特征。

2 实验与结果分析

为验证所提结构的可行性和实用性，本文利用不同性能的测试平台配合多个场景下的数据集进行实验。为方便与同类型网络对比，实验利用 tensorflow 深度学习框架搭建所提网络，训练时的超参数以及相关策略借鉴文献[14-15]进行设置，如表 2 所示。

表 2 网络训练超参及策略

Table 2 Network training hyperparameter and strategy

Parameter	Value
Batch_Size	4
Base_Lr	0.01
Momentum	0.95
Weight_Decay	0.0005
Learning	step
Optimization	Adam
Loss function	Cross Entropy

对于网络性能评估主要依据检测精度和计算效率两个指标，精度采用目标检测网络最常用的评估指标——均值平均精度（mAP，mean average precision）来衡量，如式(10)所示。同时，为衡量不同尺度目标效果，将图像中目标包围框以像素面积 32^2 和 96^2 分为小中大 3 个尺度，利用 mAP_s、mAP_m、mAP_l 分别进行衡量。而效率则通过计算网络每秒处理的图像数量来衡量，如式(11)所示。

$$\mathrm{mAP}=\frac{\sum \mathrm{AP}_C}{N} \tag{10}$$

$$\mathrm{FPS}=M/\sum_i^M T_i \tag{11}$$

式中：C 为目标类别；AP_C 表示 C 类别目标平均检测精度；N 为目标类别总数；M 表示训练样本数量；T_i 表示处理第 i 张图像时间消耗。

2.1 可行性实验

为验证所提方法各个模块的可行性，实验采用了 RGBT210[23]公开标准数据集，在搭载 NVIDIA TITAN Xp 的主机上进行测试。该数据集涵盖了不同天气、光照、时间段下的二十多类目标，约二十万张红外-可见光图像对，但图像多取自连续视频帧，重复性较高。为避免重复图像影响网络训练效果，从数据集中选择了一万张低重复率的图像，共 10 类目标，并统一图像尺寸为 512×448 后进行训练测试。

实验利用控制变量法来分别测试各个模块，首先测试了所提单源网络的有效性，即只利用可见光图像对单个特征提取支路进行训练测试，并与当前主流的目标检测网络进行对比。其中，3 个网络的检测部分都采用金字塔结构，结果如表 3 所示。

表 3 单源网络测试对比

Table 3 Single source network test comparison

Network	FPS	Accuracy/(%)			
		mAP	mAP_s	mAP_m	mAP_l
YOLO[14]	67	70.6	49.8	73.8	81.1
MobileNet[15]	107	69.7	48.7	72.2	79.5
Ours	121	69.3	48.1	71.9	79.3

由表 3 可以看出，为保证整体目标检测网络计算效率，所提单源特征提取结构尽可能提升了网络效率，与同类网络相比效率达到了最高，但不可避免损失了部分特征，使检测精度较低。为丰富目标特征信息，引入了双源网络结构，针对双源网络结构的特征互补性，本文分别对比了红外、可见光单分支以及不同融合结构的双分支网络。同理，为避免其他因素影响，检测部分也都采用金字塔结构。实验结果如表 4 和图 5 所示。

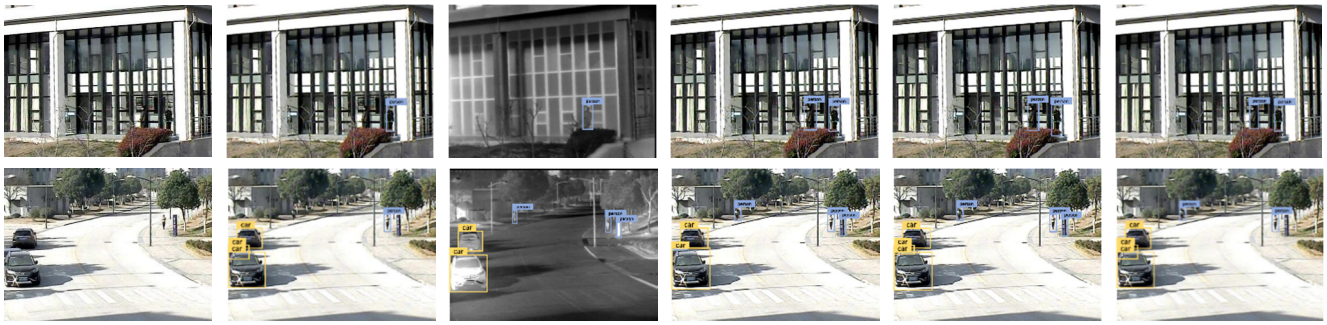
表 4 双源特征融合测试对比

Table 4 Comparison of dual source feature fusion

Network	FPS	Accuracy/(%)			
		mAP	mAP_s	mAP_m	mAP_l
Infrared branch	120	62.1	43.8	65.9	72.8
Visible branch	121	69.3	48.1	71.9	79.3
SE fusion	89	71.2	51.5	74.7	82.4
CBAM fusion	87	72.0	52.4	75.2	83.1
AAM fusion	86	72.6	53.8	75.9	83.6

根据表 4 和图 5 结果可以看出，双支路方式可以更好地互补目标特征信息，对比不同的注意力融合机制，由于 SE 只利用了通道特征，故检测精度提升有限；CBAM 方式虽同时关注了通道和空间位置特征，

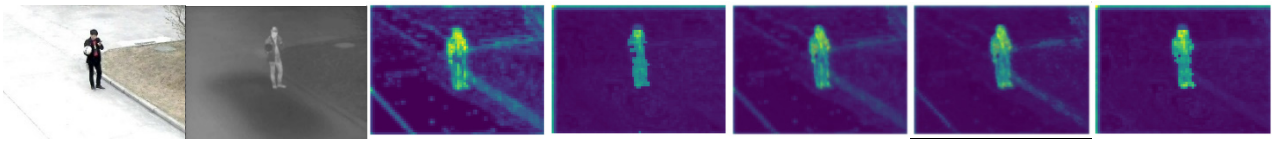
但增强特征的同时也引入了较多噪声，如图 5 第二排中将柱子误识别成行人。而所提 AAM 特征融合方式以自适应的方式可以更好地屏蔽无效信息干扰，进而保障目标检测效果。为进一步提升所提自适应注意力机制说服力，实验可视化了 block 3 输出特征在不同融合方式下的结果。为方便观测，选择了相对简单的场景，如图 6 所示。



(a) 原图 (b) 可见光检测 (c) 红外检测 (d) SE 检测 (e) CBAM 检测 (f) AAM 检测
(a) Original image (b) Visible light detection (c) Infrared detection (d) SE detection (e) CBAM detection (f) AAM detection

图 5 单源与双源网络检测结果对比

Fig.5 Comparison of single source and dual source network detection results



(a) 可见光图 (b) 红外图 (c) 可见光特征 (d) 红外特征 (e) SE 融合 (f) CBAM 融合 (g) AAM 融合
(a) Visible light image (b) Infrared image (c) Visible light features (d) Infrared features (e) SE fusion (f) CBAM fusion (g) AAM fusion

图 6 Block 3 特征融合结果对比

Fig.6 Block 3 feature fusion results comparison

根据可视化结果可以看出，SE 和 CBAM 注意力融合的方式虽然也增强了目标特征，但也引入了其他噪声。而自适应注意力机制则有效地降低了噪声的干扰，进而提升了检测精度。而对于多尺度检测结构则是从目标大小维度方面进一步提升检测效果，为验证该结构的有效性，实验分别对比了所提结构与金字塔结构的多尺度目测检测效果以及 block 3 检测层的可视化效果，实验结果如表 5 和图 7 所示。其中 block 2 融合层指红外和可见光 block 2 特征层 AAM 融合后的特征。

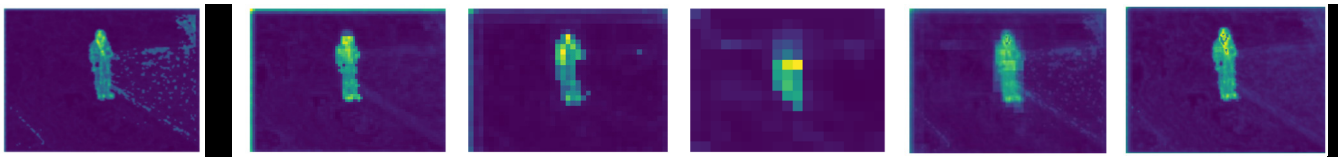
根据表 5 结果可见，所提自适应注意力机制的多尺度检测结构对于不同尺度目标都有一定提升，其中，在小目标的识别精度方面提升较大。同时，根据特征对比结果可以看出，在金字塔结构的深层与浅层信息融合过程中，浅层的特征在深层特征影响下变的相对模糊，而浅层多为小目标特征，由此可推断出小目标受其他层影响较大，而所提结构则较好地降低了其他层的干扰。

表 5 多尺度结构对比

Table 5 Multiscale structure comparison

Network	FPS	Accuracy /(%)			
		mAP	mAP_s	mAP_m	mAP_l
Pyramid multiscale	86	72.6	53.8	75.9	83.6
AAM multiscale	84	73.5	54.9	76.4	84.0

综上数据结果有效验证了各模块的可行性，而对于整个目标检测网络可行性验证，实验将所提方法与同类型红外和可见光目标检测方法进行对比，结果如表 6 所示。



(a) block2 融合层 (b) block3 融合层 (c) block4 层 (d) block5 融合层 (e) 金字塔多尺度 (f) AAM 多尺度
(a) Block2 fusion layer (b) Block3 fusion layer (c) Block4 layer (d) Block5 fusion layer (e) Pyramid multi-scale (f) AAM multi-scale

图 7 多尺度融合结果对比

Fig.7 Comparison of multiscale fusion results

表 6 同类方法测试对比

Table 6 Test comparison of similar methods

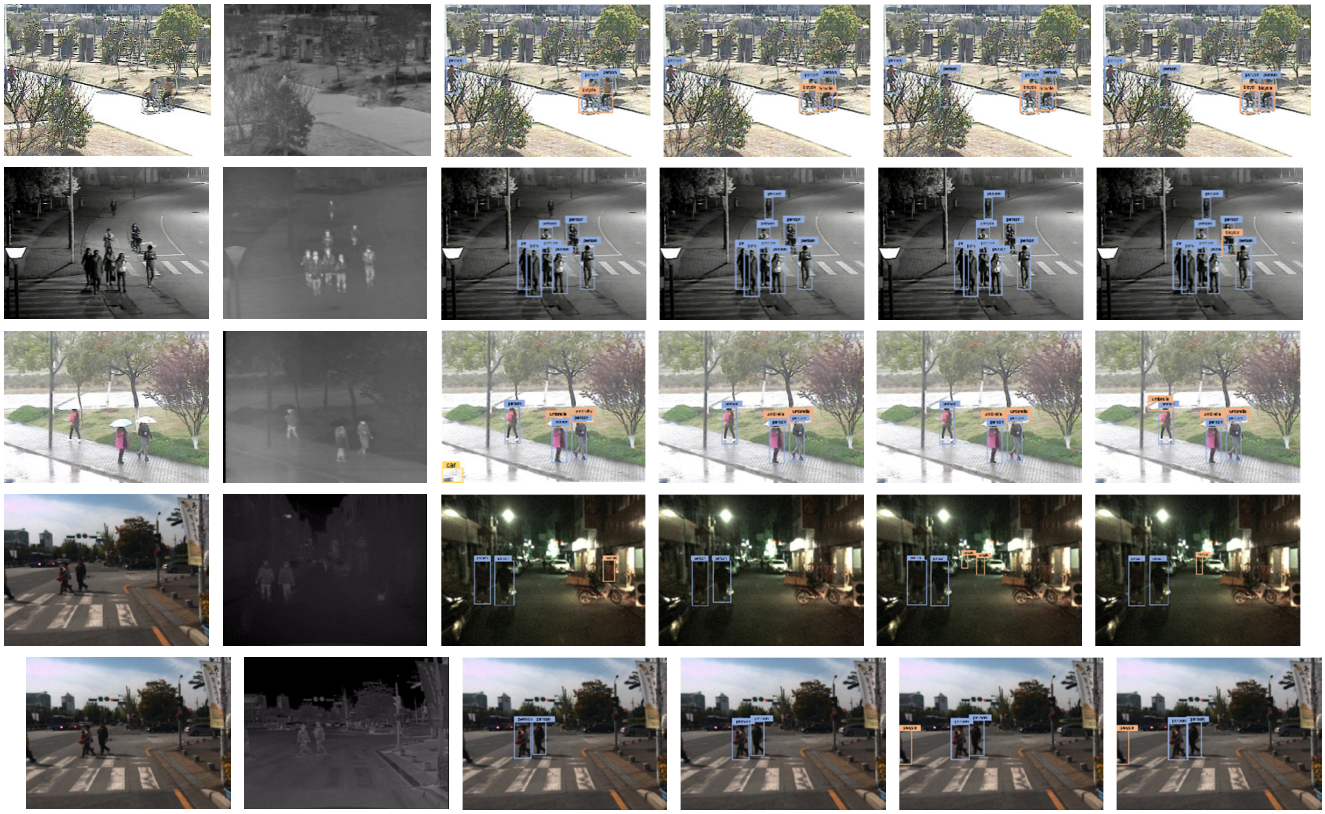
Network	FPS	Accuracy /(%)			
		mAP	mAP_s	mAP_m	mAP_l
Reference [6]	105	67.4	48.3	70.5	77.8
Reference [10]	79	71.3	51.0	73.6	81.2
Reference [24]	73	72.9	53.3	76.1	83.9
Ours	84	73.5	54.9	76.4	84.0

为进一步验证所提方法的鲁棒性，实验利用 KAIST 行人数据集进行测试。该数据集主要为白天和夜晚不同场景下红外可见光图像对，共包含 person、people 和 cyclist 三类目标。由于数据集来源于连续的视频帧，且 cyclist 类别目标较难辨认，故实验只从中筛选出约 7000 张重复率较低的图像，并将 cyclist 类别都转为 person 类别，归一化图像尺寸为 512×448 后，以 7:3 比例构建训练测试集进行实验，实验结果如表 7 所示。所提方法在 RGBT210 和 KAIST 数据集上的目标检测效果如图 8 所示。

表 7 KAIST 数据集测试对比

Table 7 Test comparison of KAIST dataset

Network	FPS	Accuracy /(%)			
		mAP	mAP_s	mAP_m	mAP_l
Reference [6]	107	72.8	50.3	77.6	85.3
Reference [10]	80	77.5	53.1	81.2	88.1
Reference [24]	74	78.6	54.7	73.1	89.9
Ours	85	79.0	55.9	73.4	90.3



(a) 可见光图 (b) 红外图 (c) 文献[6] (d) 文献[10] (e) 文献[24] (f) 本文

(a) Visible light image (b) Infrared image (c) Reference [6] (d) Reference [10] (e) Reference [24] (f)Ours

图 8 红外-可见光网络检测效果对比（前三排：RGBT210；后两排：KAIST）

Fig.8 Comparison of infrared-visible network detection effects(The first three rows: RGBT210; The last two rows: KAIST)

通过上述实验结果可以看出，与传统图像处理方法[6]相比，所提方法检测精度大幅提升，但深层神经网络的大量数据计算也使得检测效率相对较低。与基于目标检测结果融合的深度学习方法[10]相比，所提方法在特征层面融合，可以更好地对不同模态的目标信息进行互补，进而精度也相对更高。而对于同类型基于特征融合的检测方法[24]，所提自适应注意力机制增强噪声抑制和多尺度自主特征选择，使网络在小尺度目标检测中效果更佳。同时，根据图 8 也可以看出，本文所提方法可以较好应用于不同场景，并且在目标遮挡、目标较小、光线变化等复杂场景中也体现出更优的检测效果。

2.2 实际场景实验

通过标准数据集有效验证了所提方法的可行性，为进一步验证在现实场景中的实用性，实验将该网络应用于变电站巡检机器人中，测试其对变电站设备的

检测效果。巡检机器人主要搭载 Jetson Xavier NX 边缘 AI 计算平台，通过机器人自带相机采集了 6 类设备的红外及可见光图像对约 5000 张，图像大小为 512×448，使用 LabelImg 工具进行标注后对所提网络和同类型方法进行训练测试，结果如表 8 和图 9 所示。

表 8 变电站设备检测测试对比

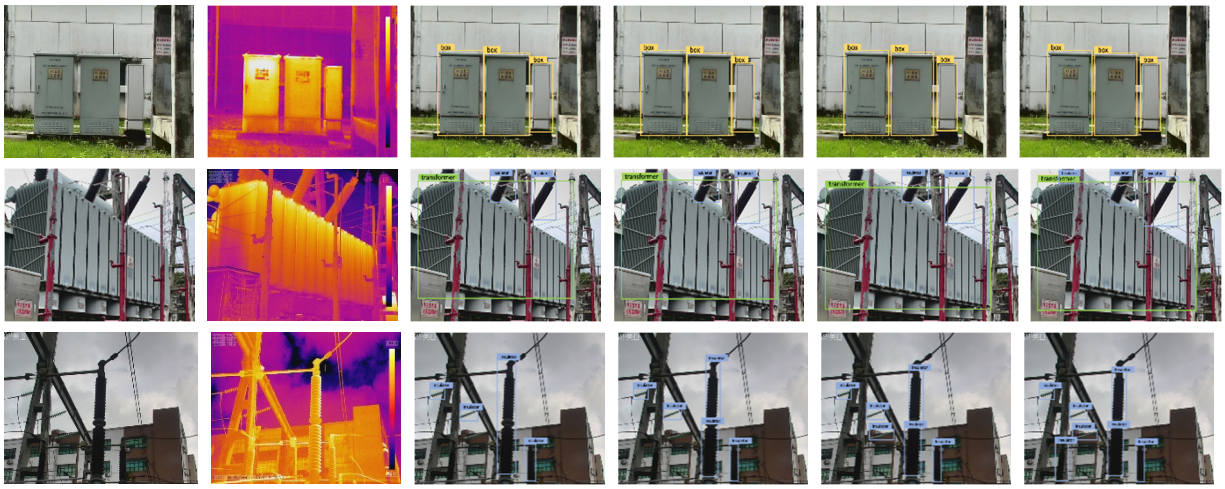
Table 8 Comparison of inspection and test of substation equipment

Network	FPS	Accuracy /(%)			
		mAP	mAP_s	mAP_m	mAP_l
Reference [6]	26	84.2	64.3	86.7	92.1
Reference [10]	18	86.3	66.5	88.2	95.3
Reference [24]	15	88.0	68.2	90.1	96.7
Ours	20	88.5	69.1	90.3	97.0

根据上述实验结果可以看出，由于机器人平台计算性能相对较低，同时，实际场景数据集在目标种类以及场景复杂度上都低于标准数据集，因此，各方法计算效率等比例下降，但检测精度都有较大提升。对于实际变电站设备检测场景中，所提方法与同类方法相比仍保持最优的检测效果，有效验证了该方法的可移植性和泛化性。同时，由图 9 结果也可看出，对于背景简单、目标尺度中等的场景，各方法检测效果都较佳，但对于复杂背景且目标过大或过小时，所提方法则体现出更优的性能。

3 结论

本文针对红外和可见光图像目标检测问题，提出了一种基于自适应注意力机制的目标检测方法。通过深度可分离卷积构建红外和可见光双支路特征提取网络，提取目标多模态特征；其次，设计自适应注意力机制将对应维度的红外和可见光特征进行融合，从特征通道以及空间位置两个角度提升有效特征的显著性。同时，针对多尺度目标，将自适应注意力机制应用于自主选择目标所处特征层，降低其他尺度特征的影响。通过实验表明，所提方法有效互补了红外和可见光特征，提升了目标多尺度识别效果，并抑制了无效特征的干扰。在标准数据集和实际变电站设备检测中，该方法都更优于同类目标检测算法，可以较好地落地实际应用。尽管所提方法在效率上未达到最高，但基本满足巡检机器人实时检测的需求，后续考虑网络剪枝或知识蒸馏等方式优化网络，进一步提升目标检测效率。



(a) 可见光图 (b) 红外图 (c) 文献[6] (d) 文献[10] (e) 文献[24] (f) 本文

(a) Visible light image (b) Infrared image (c) Reference [6] (d) Reference [10] (e) Reference [24] (f)Ours

图 9 变电站设备检测效果对比

Fig.9 Comparison of substation equipment detection effects

参考文献：

[1] 王灿, 卜乐平. 基于卷积神经网络的目标检测算法综述[J]. **舰船电子工程**, 2021, **41**(9): 161-169.
WANG Can, BU Leping. Overview of target detection algorithms based on convolutional neural networks[J]. *Naval Electronic Engineering*, 2021, **41**(9): 161-169.

[2] 郝永平, 曹昭睿, 白帆, 等. 基于兴趣区域掩码卷积神经网络的红外-可见光图像融合与目标识别算法研究[J]. **光子学报**, 2021, **50**(2): 84-98.
HAO Yongping, CAO Zhaorui, BAI Fan, et al Research on infrared visible image fusion and target recognition algorithm based on region of interest mask convolution neural network[J]. *Acta PHOTONICA Sinica*, 2021, **50**(2): 84-98

[3] 刘齐, 王茂军, 高强, 等. 基于红外成像技术的电气设备故障检测[J]. **电测与仪表**, 2019, **56**(10): 122-126.
LIU Qi, WANG Maojun, GAO Qiang, et al Electrical equipment fault detection based on infrared imaging technology[J]. *Electric Measurement and Instrumentation*, 2019, **56**(10): 122-126.

[4] XIA J, LU Y, TAN L, et al. Intelligent fusion of infrared and visible image data based on convolutional sparse representation and improved pulse-coupled neural network[J]. *Computers*, *Materials and Continua*, 2021, **67**(1): 613-624.

[5] 汪勇, 张英, 廖如超, 等. 基于可见光、热红外及激光雷达传感的无人机图像融合方法[J]. **激光杂志**, 2020, **41**(2): 141-145.
WANG Yong, ZHANG Ying, LIAO Ruchao, et al. UAV image fusion method based on visible light, thermal infrared and lidar sensing[J]. *Laser Journal*, 2020, **41**(2): 141-145.

[6] ZHANG S, LI X, ZHANG X, et al. Infrared and visible image fusion based on saliency detection and two-scale transform decomposition[J]. *Infrared Physics & Technology*, 2021, **114**(3): 103626.

[7] 王传洋. 基于红外与可见光图像的电力设备识别的研究[D]. 北京: 华北电力大学, 2017.
WANG Chuanyang. Research on Power Equipment Recognition Based on Infrared and Visible Images[D]. Beijing: North China Electric Power University, 2017.

[8] LI H, WU X J. Infrared and visible image fusion using Latent low-rank representation[J]. Arxiv Preprint Arxiv, 2018: 1804.08992.

[9] HUI L, WU X J. DenseFuse: A fusion approach to infrared and visible images[J]. *IEEE Transactions on Image Processing*, 2018, **28**(5): 2614-2623.

[10] 唐聪, 凌永顺, 杨华, 等. 基于深度学习的红外与可见光决策级融合跟踪[J]. **激光与光电子学进展**, 2019, **56**(7): 209-216.
TANG Cong, LING Yongshun, YANG Hua, et al. Decision-level fusion tracking of infrared and visible light based on deep learning[J]. *Advances in Lasers and Optoelectronics*, 2019, **56**(7): 209-216.

[11] MA J, TANG L, XU M, et al. STDFusionNet: an infrared and visible image fusion network based on salient object detection[J]. *IEEE Transactions on Instrumentation and Measurement*, 2021, **70**: 1-13.

[12] 杨雪鹤, 刘欢喜, 肖建力. 多模态生物特征提取及相关性评价综述[J]. **中国图象图形学报**, 2020, **25**(8): 1529-1538.
YANG Xuehe, LIU Huanxi, XIAO Jianli. A review of multimodal biometric feature extraction and correlation evaluation[J]. *Chinese Journal of Image and Graphics*, 2020, **25**(8): 1529-1538.

[13] WANG Z, XIN Z, HUANG X, et al. Overview of SAR image feature extraction and object recognition[J]. *Springer*, 2021, **234**(4): 69-75.

[14] WEI Z. A summary of research and application of deep learning[J]. *International Core Journal of Engineering*, 2019, **5**(9): 167-169.

[15] Bochkovskiy A, WANG C Y, LIAO H. YOLOv4: Optimal speed and accuracy of object detection[J]. Arxiv Preprint Arxiv, 2020: 2004.10934.

[16] HE K, ZHANG X, REN S, et al. Deep residual learning for image recognition[C]// *IEEE Conference on Computer Vision and Pattern Recognition* (CVPR), 2016: 770-778.

[17] Howard A, Sandler M, Chen B, et al. Searching for MobileNetV3 [C]//*IEEE International Conference on Computer Vision* (ICCV), 2020: 1314-1324.

[18] CHEN H, WANG Y, XU C, et al. AdderNet: Do we really need multiplications in deep learning?[C]// 2020 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (CVPR). *IEEE*, 2020: 1465-1474.

[19] 宋鹏汉, 辛怀声, 刘楠楠. 基于深度学习的海上舰船目标多源特征融合识别[J]. **中国电子科学研究院学报**, 2021, **16**(2): 127-133.
SONG Penghan, XIN Huaisheng, LIU Nannan. Multi-source feature fusion recognition of marine ship targets based on deep learning[J]. *Journal of the Chinese Academy of Electronic Sciences*, 2021, **16**(2): 127-133.

[20] Hassan E. Multiple object tracking using feature fusion in hierarchical LSTMs[J]. *The Journal of Engineering*, 2020(10): 893-899.

[21] LIN T Y, Dollar P, Girshick R, et al. Feature pyramid networks for object detection[C]// *IEEE Conference on Computer Vision and Pattern Recognition* (CVPR), 2017: 936-944.

[22] LIU S, HUANG D, WANG Y. Learning spatial fusion for single-shot object detection[J]. Arxiv Preprint Arxiv, 2019: 1911.09516v1.

[23] 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.

[24] XIAO X, WANG B, MIAO L, et al. Infrared and visible image object detection via focused feature enhancement and cascaded semantic extension[J]. *Remote Sensing*, 2021, **13**(13): 2538.