〈无损检测〉

基于跨模态多层次特征融合的电力设备检测算法

刘善峰[1]，毛万登[1]，李苗苗[1]，周千凯[2]，邹文杰[3]，鲍 华[3]
（1. 国网河南省电力公司电力科学研究院，河南 郑州 450018；2. 国网驻马店确山县电力公司，河南 驻马店 463200；
3. 安徽大学 电气工程及其自动化学院，安徽 合肥 230601）

摘要：针对复杂环境下电力设备检测算法鲁棒性较低和小目标检测不准确的问题，本文提出一种基于自适应融合和自注意力增强的跨模态多层次特征融合算法。首先构建双流特征提取网络，提取可见光图像和红外图像的多层级目标表征。通过引入自适应融合模块，捕捉可见光分支和红外分支两种模态下的互补特征，并进一步利用基于 Transformer 的自注意力机制来增强互补特征的语义空间信息。最后通过不同尺度下的深层特征来实现目标的精确定位。本文在自建电力设备数据集上进行充分实验，实验结果表明，所提算法的平均精确率均值（mAP50）可以达到 91.7%，相较于单一可见光支路和单一红外支路，分别提升了 3.5%和 3.9%，有效地实现了跨模态信息的融合。与当前主流目标检测算法相比，展现出较高的鲁棒性。

关键词：目标检测；跨模态；电力设备；自适应融合；自注意力机制
中图分类号：TP391.4 **文献标志码**：A **文章编号**：1001-8891(2025)07-0884-11

Cross-Modal Multilevel Feature Fusion-Based Algorithm for Power-Equipment Detection

LIU Shanfeng[1]，MAO Wandeng[1]，LI Miaomiao[1]，ZHOU Qiankai[2]，ZOU Wenjie[3]，BAO Hua[3]
(1. *Electric Power Research Institute of State Grid Henan Electric Power Company, Zhengzhou* 450018, *China*;
2. *State Grid Zhumadian Queshan Electric Power Compony, Zhumadian* 463200, *China*;
3. *School of Electrical Engineering and Automation, Anhui University, Hefei* 230601, *China*)

Abstract: A novel cross-modal multilevel feature fusion algorithm based on adaptive fusion and self-attention enhancement is proposed to address the low robustness of power-equipment detection algorithms and inaccurate small-target detection in complex environments. The algorithm begins by constructing a dual-stream feature-extraction network to extract multilevel target representations from visible-light and infrared images. An adaptive fusion module is introduced to capture complementary features from both the visible-light and infrared branches. Furthermore, a self-attention mechanism based on a Transformer is employed to enhance the semantic spatial information of the complementary features. Finally, precise target localization is achieved by utilizing deep features at different scales. Experimental evaluations were conducted on a custom-developed power-equipment dataset, and the results show that the proposed algorithm achieved an average precision mean value of 91.7%. Compared with using only the visible-light or infrared branch separately, the algorithm shows improvements of 3.5% and 3.9%, respectively, thus effectively achieving cross-modal information fusion. Compared with current mainstream object-detection algorithms, it exhibits superior robustness.

Key words: object detection, cross-modal, power equipment, adaptive fusion, self-attention mechanism

收稿日期：2023-12-13；**修订日期**：2024-02-29.
作者简介：刘善峰（1986-），男，博士，高级工程师，主要从事大数据与人工智能方面的研究。E-mail: 179555872@qq.com。
通信作者：鲍华（1978-），男，博士，副教授，硕士生导师，主要从事计算机视觉方面的研究。E-mail：baohua@ahu.edu.cn。
基金项目：安徽省科技厅自然基金面上项目（1908085MF217）；安徽省教育厅重点项目（2023AH050085）。

0 引言

变电站的智能化巡检[1]已经成为了电力系统的重要发展方向，现有巡检方式主要是通过巡检机器人[2-4]，将拍摄到的电力设备图像通过视频传输到监控系统，并由专业人员进行分析与诊断[5]。该方式完全依赖于专业人员的检查，难以做到实时监测。为此，引入高精度目标检测技术对变电站电力设备进行精准定位与识别，对提高变电站安全性和稳定性具有重要意义。

近年来，基于深度学习的目标检测算法[6-8]在变电站电力设备检测任务中取得了显著进展。相较于传统算法[9-10]，深度学习相关算法能够更准确地检测电力设备的故障，应对复杂环境的挑战。特别地，由于注意力机制[11-13]可以有选择地自动学习图像中的特征表示，从而能更好地适应变电站场景中电力设备的多样性和复杂性，因此被广泛应用于电力设备检测算法[14-17]中。例如，郑婷婷等人[18]将协同注意力模块融入主干网络，改进了特征融合模块PANet[19]，并在原模型上增加了一个检测头，提高了模型在复杂场景下的检测能力，减少了小目标的漏检和错测。

尽管单一光源图像下的电力设备检测已经取得了一定的效果，但模型精度依然受限于单一光源图像的质量。随着模态融合理论的不断发展，越来越多的研究结果表明，可见光图像和红外图像的融合能够弥补单一光源在检测任务上的不足[20-21]。通过在合适的位置并采用恰当的融合算法，可以有效整合两种模态的信息，从而提升检测精度。Wagner等人[22]基于CNN（Convolutional Neural Network）提出了不同模态之间早期融合和后期融合两种检测框架，实验发现后期融合效果要优于早期融合。在此基础上，Liu等[23]进一步研究了早期融合、中期融合、后期融合以及置信度融合对模型性能的影响，得出中期融合的效果要优于其他3种融合方式。因此中期融合方式成为了跨模态融合的主要方法。马野等人[24]建立了一种端到端的神经网络模型，利用特征权重网络可以自动地为红外图像和可见光图像分配权重，在M3FD[25]数据集上获得了更优的目标检测效果。施政等人[26]基于YOLO算法构建了双流特征提取网络，设计出结合注意力机制的模态加权融合来代替级联的融合方法以提取有效特征，在行人检测上取得了比单一模态更高的AP（Average Precision）值。邝楚文等人[27]通过构建基于红外与可见光的双支路特征提取网络，并引入自适应特征融合模块来实现不同模态图像信息的互补，在实际变电站场景下的检测结果要高于单一可见光支路和单一红外支路的结果，展现了基于注意力机制的跨模态融合在变电站场景下检测的巨大潜力。上述研究均表现出基于注意力机制的融合方式已经成为不同模态融合的主流方法。此外，随着基于Transformer的架构在自然语言处理上的大获成功，Dosovitskiy等人[28]将Transformer引入计算机视觉领域，凭借Transformer结构特有的长距离依赖和全局上下文的建模关系，在图像分类中取得了出色的结果，同时吸引了大批学者对Transformer架构在计算机视觉领域的应用研究。例如，Zhu等人[29]基于YOLOv5网络，将基于Transformer的预测头来代替原有的预测头，并在原有网络结构上增加了一个预测头来处理目标的大尺度方差，相较于YOLOv5网络，AP值提高了约7%。Prakash等人[30]通过将基于Transformer的自注意力机制作为跨模态的融合模块提出了适用于自动驾驶场景下的TransFuser网络结构。该方法将不同模态的信息展平，拼接成一个长序列，作为融合模块的输入，利用Transformer可以自动学习不同模态的权重机制，在涉及复杂场景的城市环境中取得了优越的性能。

因此，本文利用注意力机制自动学习跨模态图像特征表示以及Transformer结构捕捉全局上下文的出色能力，提出了一种基于自适应融合与自注意力增强的跨模态图像融合目标检测方法。通过构建双流特征提取主干网络，可以获取不同模态下的多尺度独有特征。利用自适应融合模块对不同尺度的可见光图像特征和红外图像特征进行通道特征图的自适应权重选择，放大有效信息并抑制噪声，改善小目标的漏检问题。进一步引入基于Transformer的自注意力机制，增强特征的上下文信息和空间关系，提高模型的表达能力和抗干扰能力。与此同时，采用特征金字塔网络和路径聚合网络相结合的检测模式，优化融合特征的语义信息与目标的定位信息，实现变电站电力设备的精准定位与识别。

1 基于跨模态多层次特征融合的电力设备检测方法

1.1 整体框架

本文所设计的跨模态多层级特征融合检测网络整体结构如图1所示。该网络的主体结构由双流特征提取主干网络、自适应融合模块（Adaptive Fusion Module, AFM）、自注意力增强模块（Self-Attention Enhancement Module, SAE）和检测模块4部分构成。

具体地，本文使用两个参数配置完全相同的CSPDarkNet53[31]网络来构建双流特征提取主干网络，以充分提取可见光模态和红外模态多层级下的独有

特征，分别表示为 Fea_v 和 Fea_i。通过这种方式，能够有效减少不同模态之间的干扰，并且充分利用每个模态所提供的信息。为获取具有跨模态信息的互补特征，本文将两个模态的多层级特征 Fea_v 和 Fea_i 分别输入到自适应融合模块中，通过调整不同通道的权重，可以使模型更关注具有判别力的通道，同时抑制不必要的噪声和冗余的通道。为了获取具有全局依赖的增强特征，将自适应融合模块输出的 3 个尺寸特征进一步分别输入到自注意力增强模块中，以建立长距离的依赖关系，提高网络对上下文的感知能力，并最终得到增强特征 Fea_0、Fea_1 和 Fea_2。此外，为了充分利用多尺度特征的信息，本文的检测模块采用特征金字塔（Feature Pyramid Networks，FPN）和路径聚合网络（Path Aggregation Network，PAN）相结合的网络结构，以增强网络的检测和分类能力。

1.2 自适应融合模块

目前对于跨模态特征的融合，大多是基于不同模态特征的简单加和或拼接，这种处理方式较为粗糙，并不能充分地挖掘不同模态下的互补信息，还可能会引入噪声。因此为有效融合可见光图像和红外图像提取到的特征，本文提出了适用于跨模态融合的自适应融合模块，通过自适应学习各个特征之间的权重或相互关系，可以将不同特征的优势进行有效融合，以获得更具表达能力和区分度的特征表示，实现可见光和红外图像信息的互补。自适应融合模块结构如图 2 所示。

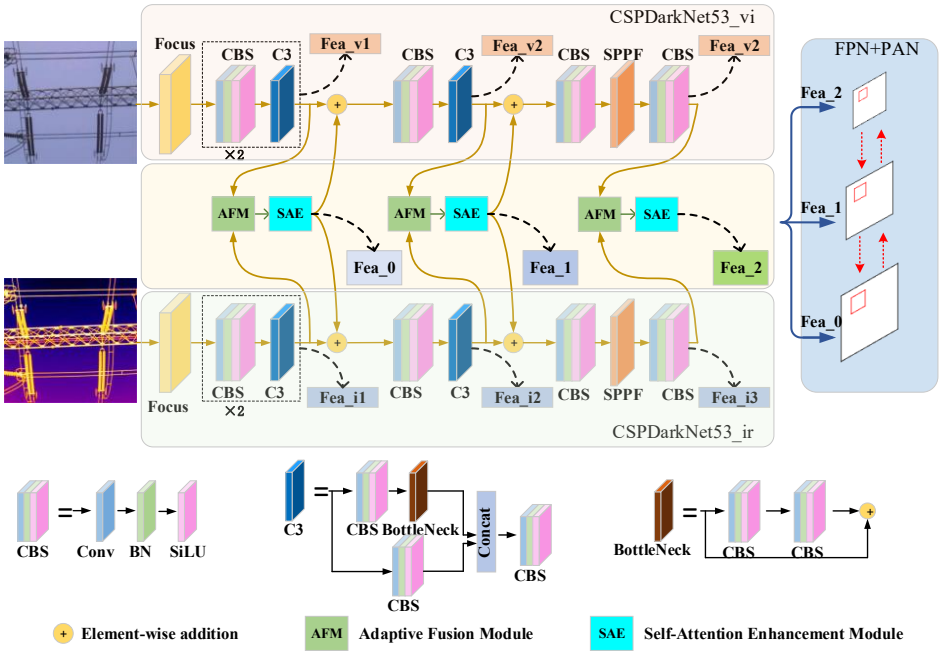


图 1 跨模态多层级特征融合目标检测框架

Fig.1 Cross-Modal multilevel feature fusion object detection framework

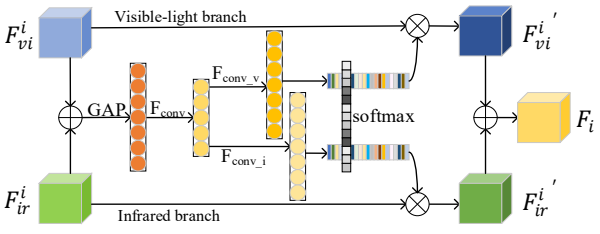


图 2 自适应融合模块

Fig.2 Adaptive fusion module

为保留更丰富的目标信息，将可见光和红外分支提取到的特征（F_{vi}^{i}，$F_{ir}^{i}\in R^{H\times W\times C}$）进行元素级相加融合，以获得具有丰富细节信息的融合特征图。随后对融合的特征图做全局平均池化（GAP），汇总空间信息的同时，也避免了过拟合的问题，全局平均池化原理如图 3 所示，得到全局通道权重$\omega^{i}\in R^{1\times 1\times C}$：

$$\omega^{i}=\frac{1}{H\times W}\sum_{h=1}^{H}\sum_{w=1}^{W}\left(F_{vi}^{i}+F_{ir}^{i}\right) \tag{1}$$

式中：F_{vi}^{i}和 F_{ir}^{i}表示双流特征提取网络分别从可见光图像和红外图像提取到的第 i 层特征。为了减少全局表示ω^{i}的信息量，引入一个 1×1 的卷积操作对其进行压缩，得到中间向量 $T\in R^{1\times 1\times d}$：

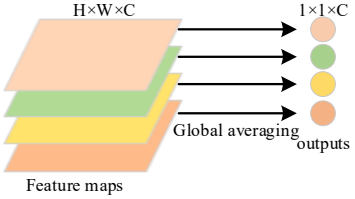


图 3 基于空间压缩的全局平均池化策略

Fig.3 The strategy of global average pooling based on spatial compression

$$T=\rho\left(\theta\left(F_{\text{conv}}\left(\omega^{i}\right)\right)\right) \tag{2}$$

式中：ρ和θ分别表示 ReLU 激活函数和批归一化；F_{conv}表示 1×1 的卷积；d 表示 F_{conv} 的通道维度，本文将 d 设置为 32。为了得到最终不同分支特征图的权重，将中间向量 T 分别通过 1×1 的卷积，同时将维度从 d 提高到原始输入的维度 C，再经由 softmax 激活函数得到归一化的权重比例ω_{vi}^{i}和ω_{ir}^{i}。将该权重分别与原可见光图像和红外图像特征相乘，得到自适应特征 $F_{\text{vi}}^{i\prime}$ 和 $F_{\text{ir}}^{i\prime}$，最后两个自适应特征进行元素级相加，获得最终的互补特征 F_i：

$$\omega_{\text{vi}}^{i}=\text{softmax}\left(F_{\text{conv\_v}}\left(T\right)\right) \tag{3}$$

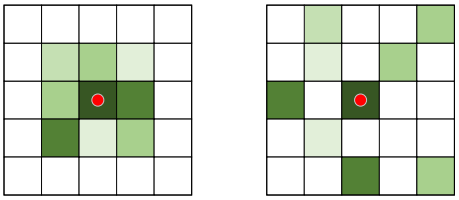
$$\omega_{\text{ir}}^{i}=\text{softmax}\left(F_{\text{conv\_i}}\left(T\right)\right) \tag{4}$$

$$F_i=F_{\text{vi}}^{i}*\omega_{\text{vi}}^{i}+F_{\text{ir}}^{i}*\omega_{\text{ir}}^{i} \tag{5}$$

式中：$F_{\text{conv_v}}$和 $F_{\text{conv_i}}$分别代表 1×1 的卷积；*代表元素级相乘操作。

1.3 基于 Transformer 的自注意力增强模块

普通卷积对输入特征图的运算受限于卷积核的大小，只能处理一定范围内的特征，不能获取全局信息，而基于 Transformer 的自注意力机制可以利用相关矩阵对特征图的每个位置进行加权计算，自动选取有效特征，提高模型的表达能力和对全局信息的捕捉能力，如图 4 所示。



(a) 卷积　(b) 自注意力机制

(a) Convolution　(b) self-attention mechanism

图 4　卷积与自注意力机制的区别

Fig.4　The difference between convolution and self-attention mechanism

因此，本文利用自注意力机制对互补特征进行信息增强。自注意力增强模块主要包括位置编码（Position Embedding）、两个层归一化（Layer Normal）、多头自注意力机制（Multi-Head Self-Attention）和多层感知机（MLP）。具体组成如图 5 所示。

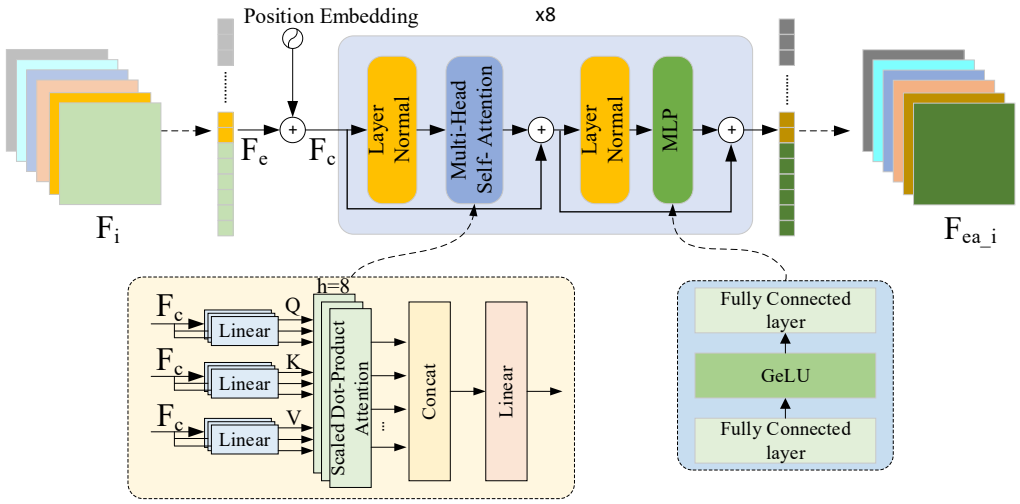


图 5　自注意力增强模块

Fig.5　Self-attention enhancement module

首先对特征 F_i 展平以及维度变换，得到特征 F_{e}（尺寸为 $HW\times C$）。接着，将相同尺寸的位置编码信息与特征 F_{e} 相加，保留原有特征的空间位置信息，得到特征 F_{c}。然后通过层归一化，将通道特征的数值限制在固定范围并呈现相似的尺度分布，加快网络的训练和收敛的速度。进一步地，多头自注意力机制对归一化后的输入特征进行全局的位置权重计算，帮助模型关注其重要的上下文信息和全局依赖关系。其多头自注意力（MSA）计算公式为：

$$\begin{aligned}Z'&=\text{MultiHead}(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V})\\&=\text{Concat}(\text{head}_1,\cdots,\text{head}_h)\boldsymbol{W}_{\text{o}}\end{aligned} \tag{6}$$

式中：$\boldsymbol{Q}$、$\boldsymbol{K}$、$\boldsymbol{V}$ 分别表示查询向量、键向量和值向量，是对输入特征 F_{c} 的空间映射，其值的计算方法如式(7)～式(9)。h 表示头的数量，在本文中取 $h=8$，$\boldsymbol{W}_{\text{o}}$ 是输出变换矩阵，多头自注意力可以形成多个子语义空间，让模型关注不同维度的语义空间信息。每个头的输出 head_i 可以表示为式(10)：

$$\boldsymbol{Q}=F_{\text{c}}\boldsymbol{W}^{Q} \tag{7}$$

$$\boldsymbol{K}=F_c\boldsymbol{W}^K \tag{8}$$

$$\boldsymbol{V}=F_c\boldsymbol{W}^V \tag{9}$$

$$\text{head}_i=\text{Attention}(\boldsymbol{Q}w_i^Q,\ \boldsymbol{K}w_i^K,\ \boldsymbol{V}w_i^V) \tag{10}$$

式中：$\boldsymbol{W}^Q$、$\boldsymbol{W}^K$、$\boldsymbol{W}^V$ 分别是第 i 个头的查询、键、值变换矩阵。Attention 是注意力计算的函数，在 MSA 中，一般使用自注意力机制来进行计算，如式(11)所示：

$$\text{Attention}(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V})=\text{softmax}\left(\frac{\boldsymbol{Q}\boldsymbol{K}^{\mathrm{T}}}{\sqrt{d_k}}\right)\boldsymbol{V} \tag{11}$$

式中：d_k 是键向量的维度，softmax 对相似度进行归一化，将每个键向量的权重计算出来，然后将权重乘以值向量，最后进行加权求和得到注意力的输出。

多头自注意力机制的输出特征通过残差结构与经过位置编码的输入特征相加，并再次对特征进行层归一化处理。然后 MLP 层对经过 LN 层后的特征进行非线性建模，增强模型的拟合能力。最后，将 MLP 的输出与第一次残差结构的输出相加，得到输出特征 O，如式(12)所示。并将输出特征输入到下一结构。

$$O=\text{MLP}(\text{LN}(Z'+F_c))+Z'+F_c \tag{12}$$

式中：LN 表示 Layer Normal 层；MLP 表示多层感知机，由两个全连接层以及中间的 GELU 激活函数组成。

1.4 检测模块

为充分利用双流主干网络提取到的特征，本文的检测模块由自顶向下的特征金字塔和自底向上的路径聚合网络组成，如图 6 所示。

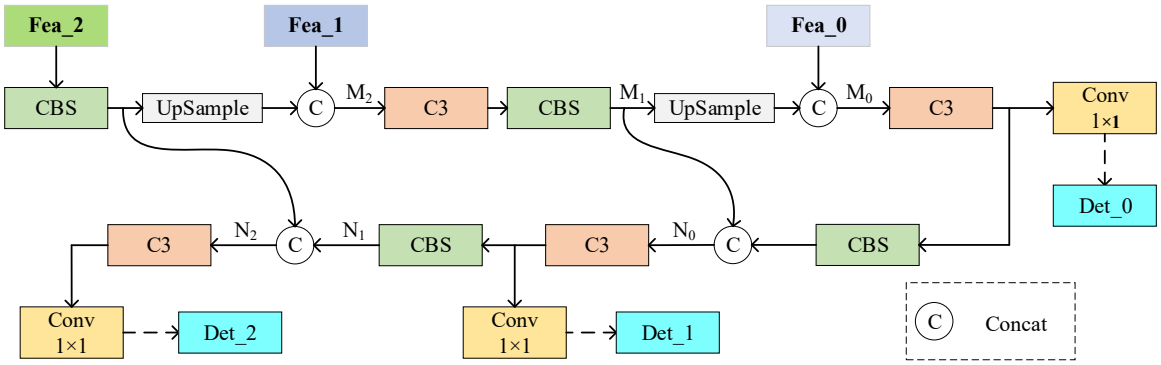


图 6 检测模块

Fig.6 Detection module

特征金字塔的顶层、中间层和底层分别是深层次特征 Fea_2、中间层特征 Fea_1 和浅层特征 Fea_0。深层次特征经过 CBS 模块和上采样模块后与中间层特征在通道上进行拼接，得到拼接特征 M_2。与传统的 PANet[19]不同，考虑到特征的直接加和可能会引入噪声信息，因此本文采用特征拼接来代替特征的加和。M_2 通过 C3 模块与 CBS 模块的顺序传播，得到融合特征 M_1。将融合了深层次特征与中间层特征的 M_1 经过上采样再与浅层特征进行通道拼接，得到具有 3 个不同层次的拼接特征 M_0，如式(13)所示：

$$M_0=f_C(F_{\text{ea\_0}},f_{US}(M_1)) \tag{13}$$

式中：f_C 代表拼接操作；f_{US} 代表上采样。

M_0 经过 C3 模块对特征进行整合后，再通过 1×1 的卷积调整通道数，得到输出 $D_{\text{et\_0}}$，如式(14)所示：

$$D_{\text{et\_0}}=f_{\text{conv}}(f_{C3}(M_0)) \tag{14}$$

式中：f_{C3} 代表 C3 模块；f_{conv} 代表卷积核大小为 1×1 的卷积。

为了将具有更多细节和定位信息的浅层特征通过自底向上的方式传递给深层特征，M_0 顺序通过 C3 模块与 CBS 模块进行特征融合后，与 M_1 在通道上进行拼接，得到特征 N_0，如式(15)所示：

$$N_0=f_{US}(M_1,f_{C3C}(M_0)) \tag{15}$$

式中：f_{C3C} 代表 C3 模块与 CBS 模块的顺序连接。

将拼接特征 N_0 经过 C3 模块与 1×1 的卷积后得到输出 Det_1。另一方面将 N_0 通过 C3 模块与 CBS 模块得到整合特征 N_1。将 N_1 与原始的深层次特征进行拼接得到具有更多语义信息的拼接特征 N_2，N_2 经过 C3 模块的特征整合后通过 1×1 的卷积得到输出 Det_2。

1.5 损失函数

由于目标检测需要兼顾多个任务，所以本实验的损失函数主要由分类损失（L_{cls}）、置信度损失（L_{conf}）和回归框损失（L_{box}）组成，如式(16)所示。分类损失和置信度损失均采用二元交叉熵函数来计算其损失。对于回归框损失，采用 Complete IoU 损失算法，如式(17)所示。具体的损失函数参考基于 YOLO 网络[33]的计算方式：

$$L_{\text{total}}=L_{\text{cls}}+L_{\text{conf}}+L_{\text{box}} \tag{16}$$

$$L_{\text{box}} = 1 - \text{IoU} + \frac{\rho^2\left(b, b^{\text{gt}}\right)}{c^2} + \alpha \cdot \tau \tag{17}$$

式中：IoU 代表预测框与真实框的重叠面积；$\rho(b,b^{\text{gt}})$为真实框中心点与预测框中心点之间的欧式距离；c为真实框与预测框最小包围矩形的对角线长度；α为τ的影响因子，τ为真实框与预测框的宽高比相似度。

2 实验及结果分析

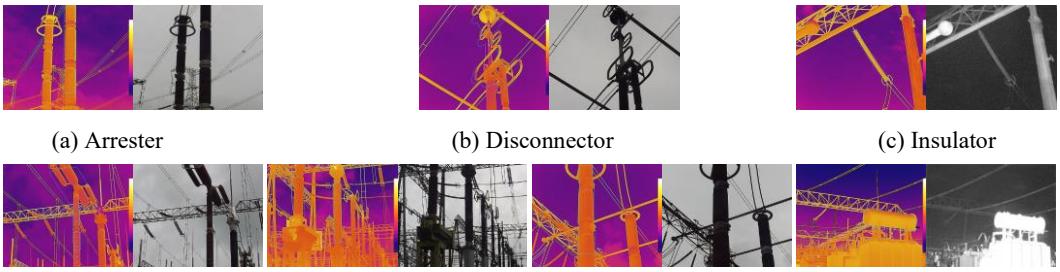
2.1 数据集与预处理

通过巡检机器人搭载的可见光和红外相机对变电站电力设备进行图像采集，由于两个相机处于不同的空间位置，且拍摄到的图片范围也不一样，所以需对采集到的可见光和红外图片进行裁剪、校准等一系列的预处理操作，使其图片内的目标对齐。实验主要采集了 500 kV 变电站下的避雷器（arrester）、隔离开关（disconnector）、绝缘子（insulator）、断路器（breaker）、电流互感器（current transformer）、电压互感器（voltage transformer）、油枕（conservator）7 种常见的变电站设备图像。依据数据质量，从中挑选出质量较好的原始图片，通过随机旋转、平移、缩放、翻转等预处理，将数据集扩充至 2155 对，建立了一个 500 kV 下变电站设备目标检测数据集 TSE500。数据集种类和数量分别如表 1 和图 7 所示。利用公开的标注工具 labelimg 对图片中的待检测目标进行精确地手动标注，并统一调整图片尺寸为 640×512。对处理以后的图片按照 7:2:1 的比例划分训练集、验证集和测试集，其中训练集 1508 对，验证集 431 对，测试集 216 对。

表 1 数据集不同类别数量

Table 1 The number of different categories in the dataset

Classes	Numbers
Arrester	210
Disconnector	500
Insulator	340
Breaker	325
Current transformer	330
Voltage transformer	230
Conservator	220



(a) Arrester (b) Disconnector (c) Insulator

(d) Breaker (e) Current transformer (f) Voltage transformer (g) Conservator

图 7 训练数据集部分样本

Fig.7 Partial samples of the training dataset

2.2 实验配置及参数

所有实验均在 Ubuntu18.04 LTS 操作系统上完成，在基于 Pytorch 的深度学习框架上完成实验的训练、验证及测试。详细实验硬件及软件配置为：CPU 为 Intel i7-10700KF、GPU 为 Nvidia GeForce RTX 3060、显存 12 GB、CUDA11.3、Python3.7。此外，本实验采用随机梯度下降（Stochastic Gradient Descent，SGD）来优化模型，初始学习率设置为 0.01，动量为 0.937，权重衰减为 0.0005。Epochs 设置为 200，batch-size 设置为 4。

2.3 评价指标

使用目标检测中常用的精确率（Precision）、召回率（Recall）、平均精确率均值（mAP）、模型参数量（Params）以及精确率-召回率（*P-R* 曲线）等作为评价指标。相较于准确率和召回率，平均精确率均值指标综合考虑了不同类别的预测结果在不同 IoU 阈值下的精确率表现，可以更加全面地衡量目标检测算法的准确性和鲁棒性。

本实验记 IoU 阈值在 0.5 的平均精确率均值为 mAP50，IoU 阈值在 0.75 的平均精确率均值为 mAP75，IoU 从 0.5 到 0.95，步长为 0.05 的平均精确率均值为 mAP.95。

2.4 对比实验

为了验证本方法的优越性，与目前其他 6 种常用的目标检测算法进行比较，包括单模态目标检测算法 SSD、Faster R-CNN、YOLOv4、YOLOv8，跨模态目标检测算法 CFT[32]和 UA-CMDet[33]。单模态目标检测算法使用 TSE500 中的可见光图像数据进行实验，单模态与跨模态实验配置及训练超参数均与本文模型保持一致。实验结果如表 2 所示。

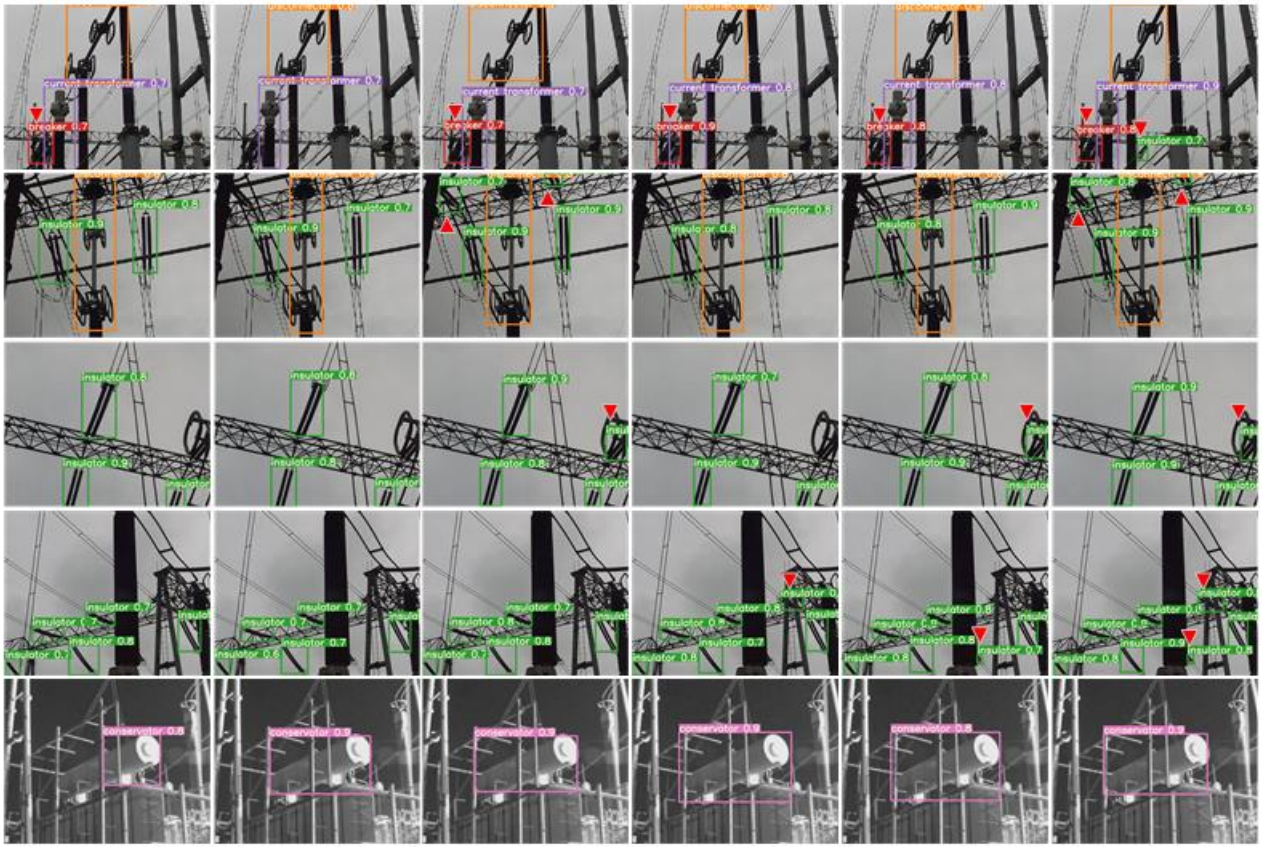
表 2 不同检测算法的定量比较

Table 2 Comparative quantitative analysis of different detection algorithms

Module	Backbone	mAP50/%	mAP75/%	mAP.95/%	Params/M
SSD	VGG16	82.1	68.1	50.3	24.01
Faster R-CNN	ResNet50	83.8	-	53.8	136.75
YOLOv4	CSP DarkNet53	85.5	68.3	55.6	63.95
YOLOv8	-	88.8	-	59.4	11.14
CFT	CSP DarkNet53	88.9	63.9	56.9	206.03
UA-CMDet	ResNet-FPN	87.6	-	57.2	-
Ours	CSP DarkNet53	91.7	74.9	61.5	44.62

定量分析：从表 2 中可以看出，本文提出的双流网络模型取得了最好的检测效果。对于主干网络同为 CSP DarkNet53 的单模态目标检测算法 YOLOv4，其 mAP50、mAP75、mAP.95 比本文模型分别低了 6.2%、6.6%和 5.9%，且本文模型的参数量（44.62 M）也低于 YOLOv4 网络（63.95 M）的参数量。与目前最新的单模态目标检测算法 YOLOv8 相比，虽然本文模型的参数量较大，但 mAP50 和 mAP95 比 YOLOv8 分别高了 2.9%和 2.1%。与当前的多模态目标检测算法 CFT 和 UA-CMDet 相比，本文的模型也取得了最优的结果。从对比实验来看，本文的算法模型在实际变电站场景下的检测效果具有先进性，同时参数量也较为均衡。

定性分析：检测结果的定性对比如图 8 所示。图中的红色三角形指向不同网络易漏检的目标。第一行和第四行中只有本文提出的算法检测出了全部红色三角形标示的待检目标。在第二行如此复杂的背景下，只有 YOLOv8 和本文所提的算法成功检测出了目标物体。而在其他行，本文提出的算法也表现出了优异的性能，不仅精确检测出目标位置，而且其精确度也普遍高于其他先进算法，展现出较高的鲁棒性。



(a) SSD (b) Faster R-CNN (c) YOLOv8 (d) CFT (e) UA-CMDet (f) Ours

图 8 不同算法的检测效果

Fig.8 Detection performance of different algorithms

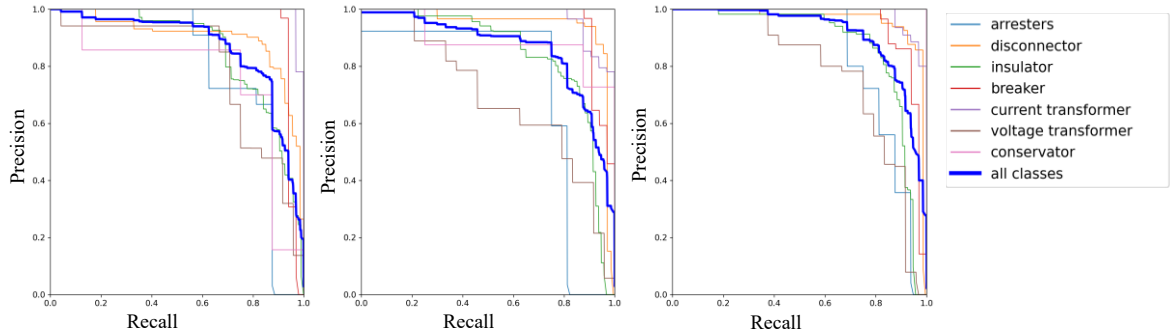
2.5 消融实验

为了验证所提模块的有效性，在所构建的变电站设备数据集 TSE500 上面进行训练，并取训练过程中效果最好的训练参数进行测试。使用精确率、召回率、平均精确率均值和模型参数量作为消融实验的评价指标。首先对单可见光支路、单红外支路与本文的双流分支的实验结果做消融对比实验。训练过程中 *P-R* 曲线变化如图 9 所示。定量的实验结果和定性的检测效果分别如表 3 和图 10 所示。单可见光支路和单红外支路的实验配置及参数均与本文的双流分支实验保持一致。

定量分析：表 3 中加粗的结果为最优结果，ir 和 vi 分别代表单红外支路和单可见光支路网络。从中可以看出，引入跨模态融合后，最能反映检测性能的 mAP50 的值可以达到 91.7%，高于单独使用红外图像（87.8%）或可见光图像（88.2%）的目标检测结果。具体地，绝缘子在单红外支路的 mAP50 检测结果为 84.8%，断路器和隔离开关在单可见光支路的 mAP50 检测结果分别为 95.4%和 92.2%，而经过跨模态融合之后，mAP50 分别提升了 2.6%、1.9%和 3.1%。说明融合算法可以有效的借鉴两个模态的信息，从而实现性能的有效提升。

定性分析：图 9 展现了训练过程中单可见光支路、单红外支路以及跨模态融合的 *P-R* 曲线，其曲线包围的面积代表 mAP50 的值，蓝色曲线代表所有类别的 mAP50 加权平均值。训练过程中跨模态融合的 *P-R* 曲线与坐标轴所包围的面积最大，进一步说明基于跨模态融合的算法可以充分利用不同模态的互补信息，获得更加具有判别性的特征。此外，图 10 展示了本文所提跨模态算法相较于单支路算法的在实际检测中的效果。第一行和第二行中分别是单红外图像和单可见光图像漏检了绝缘子。而第三行中的隔离开关只在跨模态融合算法中被检测了出来。定性实验结果也表明，所提融合算法在实际检测中充分利用了不同模态的有用信息。

为了进一步分析关键模块对算法性能的影响，本文对自适应融合模块（Adaptive Fusion Module，AFM）、自注意力增强模块（Self-Attention Enhancement Module，SAE）进行了消融实验，定量结果和定性对比分别如表 4 和图 11 所示。为了有效验证模块的作用，需要保证所有的实验都是在双流主干网络上进行，所以将用简单的元素相加替换所有的自适应融合模块和自注意力增强模块，将其作为基本的融合方式，送入检测模块。该网络结构作为双流网络的基准模型，记作 baseline＋Add。每组实验均使用相同的超参数以及训练技巧。



(a) vi_PR (b) ir_PR (c) our_PR

图 9 模型训练指标 *P-R* 曲线对比

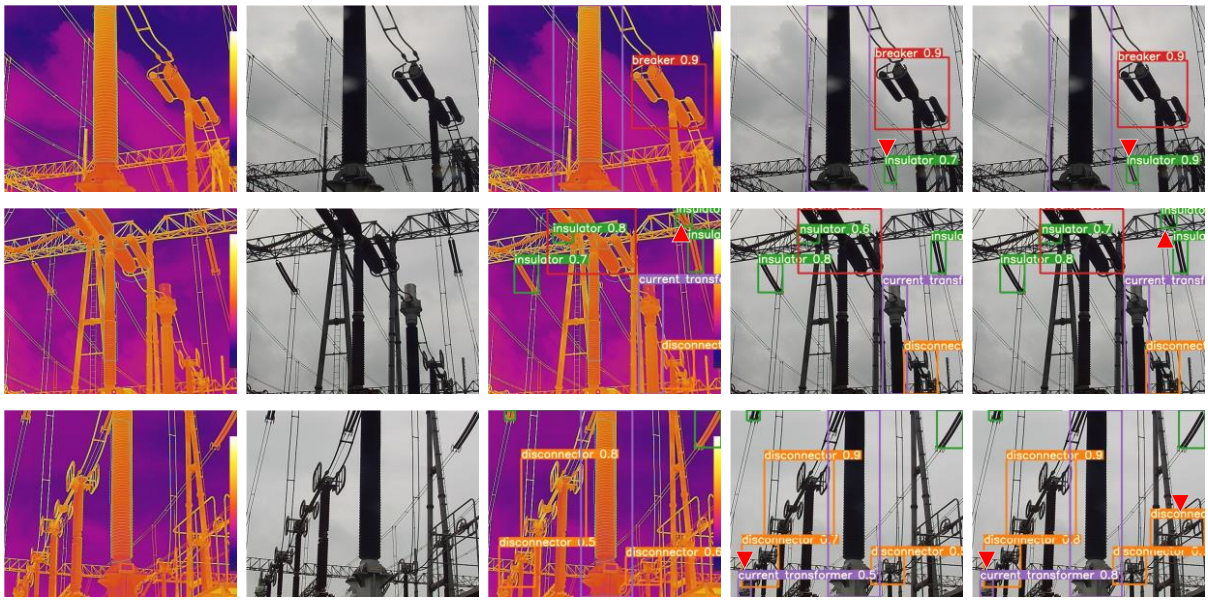
Fig.9 Comparison of *P-R* curves for model training

表 3 单支路与跨模态网络定量结果对比

Table 3 Quantitative comparison of results between single-branch and cross-modal networks

%

	Precision			Recall			mAP50		
	ir	vi	Ours	ir	vi	Ours	ir	vi	Ours
All	**87.2**	81.7	85.6	81.9	84.2	**88.5**	87.8	88.2	**91.7**
Arrester	90.8	86.1	**91.8**	68.8	70.1	**75.0**	85.0	81.1	**86.9**
Disconnector	**91.5**	85.5	88.9	82.1	88.0	**92.8**	90.8	92.2	**95.3**
Insulator	**76.1**	71.8	75.2	75.5	79.2	**85.1**	84.8	84.5	**87.4**
Breaker	**92.7**	86.5	90.2	90.0	90.7	**90.9**	93.2	95.4	**97.3**
Current transformer	**86.1**	82.1	**86.1**	94.1	97.4	**96.9**	**98.0**	96.2	97.8
Voltage transformer	73.0	71.2	75.7	75.0	76.5	79.1	76.5	79.9	82.9
Conservator	100	88.7	100	87.9	87.5	100	86.6	88.2	94.5



(a) 红外图 (b) 可见光图 (c) 红外分支检测 (d) 可见光分支检测 (e) 本文算法

(a) Infrared image (b) Visible light image (c) Infrared branch detection (d) Visible branch detection (e) Ours

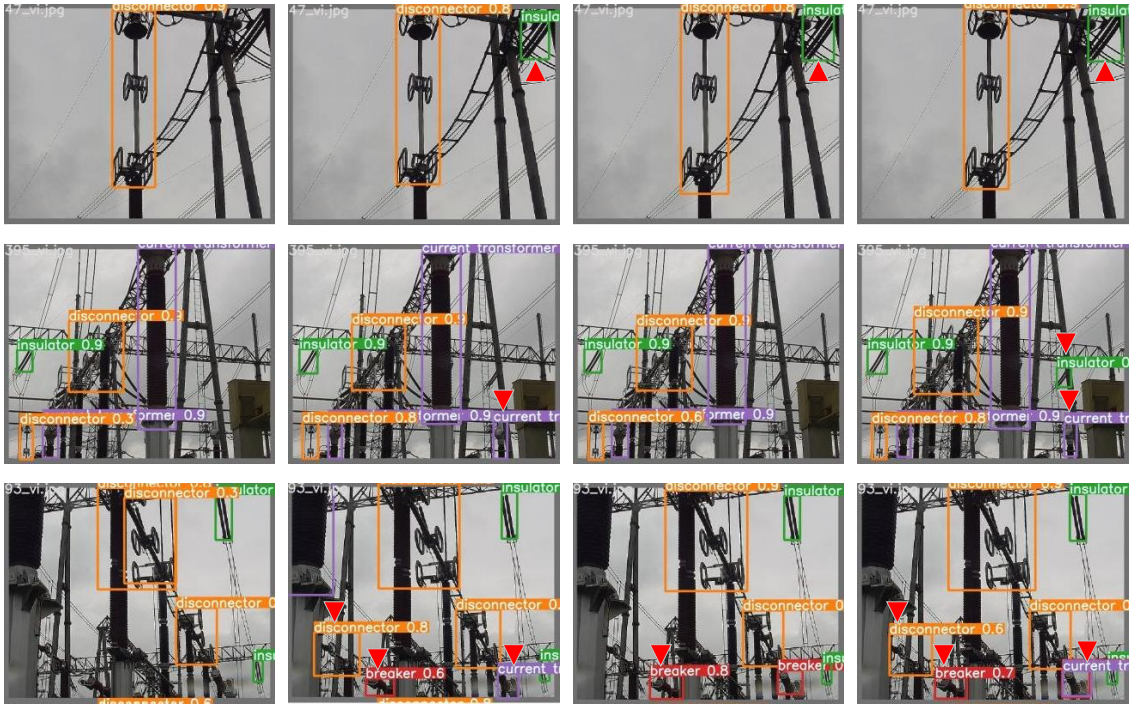
图 10 单支路与跨模态融合目标检测结果定性比较

Fig.10 Qualitative comparison of object detection results between single-modality and cross-modal fusion

表 4 关键模块消融实验的定量结果

Table 4 The quantitative results of the key module ablation experiments

Module	Precision/%	Recall/%	mAP50/%	mAP75/%	mAP.95/%	Params/M
Baseline+Add	82.1	83.5	88.3	62.3	54.5	11.29
Baseline+AFM	85.8	90.6	91.2	70.8	60.9	11.38
Baseline+Add+SAE	84.5	87.5	90.3	68.9	57.9	44.53
Baseline+AFM+SAE	85.6	88.5	91.7	74.9	61.5	44.62



(a) Baseline+Add (b) Baseline+AFM (c) Baseline+Add+SAE (d) Baseline+AFM+SAE

图 11 小目标和复杂背景下的消融实验检测效果

Fig.11 Detection performance in small object and complex backgrounds

定量分析：实验结果显示使用简单的元素相加（baseline＋Add）作为融合方法时，最终的 mAP50、mAP75 和 mAP.95 分别达到了 88.3%，62.3%和 54.5%，将简单相加的融合方式替换成自适应融合（baseline＋AFM）后分别提升了 2.9%，8.5%和 6.4%，而参数量只增加了 0.09 M。可以看出，引入自适应融合模块可以充分利用模态的互补信息并抑制噪声。进一步将自适应融合模块移除，用简单相加的融合方式和自注意力增强模块对特征进行融合增强（baseline＋Add＋SAE），相较于只用简单相加的融合方式，其 mAP50、mAP75 和 mAP.95 也分别提升了 2%，6.6%和 3.4%，参数量增加了 33.24 M。从中可以得出，自注意力增强模块可以增强其融合特征，有助于更好地判别特征，但是会带来参数量的激增。

最后，将简单相加融合模块移除，替换成自适应融合模块，并引入自注意力增强模块（baseline＋AFM＋SAE），其精准率、召回率、mAP50、mAP75 和 mAP.95 相较于简单相加的融合方式分别提升了 3.5%、5%、3.4%、12.6%和 7%，参数量达到了 44.62 M，大幅度地提升了网络的整体性能。

定性分析：为了更加直观地呈现本算法模块在小目标和复杂环境下检测的有效性，选取了部分图像改进过程中的检测结果进行定性对比，如图 11 所示。

第一行中的绝缘子（insulator）背景十分复杂，在只基于简单相加（Add）融合后并没有被检测出，而经过本文设计的自适应融合模块（AFM）或自注意力增强模块（ASE）融合后，都成功地检测出绝缘子。第二行右侧的绝缘子目标较小，只有本文提出的方法成功检测出目标。与此同时，第二行的电流互感器（current transformer）和第三行的隔离开关（disconnector）只在含有 AFM 的网络结构中被检测出来，说明 AFM 模块能够赋予有用信息更高的权重并抑制噪声信息。与此同时，ASE 进一步提高了目标检测的置信度，表现出优秀的鲁棒性。

进一步地，为了探究自适应融合模块（AFM）的中间向量 d 的维度对网络性能的影响，本文将 d 取多个值，并进行消融实验进行定量对比，实验结果如表 5 所示。d 值在 32 的时候，其 mAP50 的值最高，为 91.7%，参数量为 44.62 M，因此本文实验的中间向量维度取为 32。

表 5　d 不同取值的定量实验结果

Table 5　The quantitative results with different values of d

d	8	16	32	64	128
mAP50/%	88.4	90.9	91.7	91.4	91.2
Params/M	44.55	44.57	44.62	44.71	44.88

3　结论

针对复杂环境下小目标易漏检和鲁棒性较低的问题，本文建立了一种端到端的双流神经网络模型。该模型通过构建双流特征提取网络获取多层次特征，并将其输入自适应融合模块进行特征融合，以捕捉可见光图像和红外图像在多个层级上的互补特征。同时，利用基于 Transformer 的自注意力机制对互补特征进行全局上下文建模，提取深层次特征。此外，为了适应不同大小的变电站电力设备目标，本文采用特征金字塔和路径聚合网络的组合对目标进行定位与识别。通过在实际变电站检测场景中的实验结果表明，本文所提网络可以有效整合跨模态的信息，减少小目标的漏检率，并提高目标检测的鲁棒性，可以较好地应用于变电站设备的检测任务。尽管本文所提网络在一定程度上提高了目标检测的效果，但也引入了较多的参数，增加了模型的复杂度。如何减少模型的参数量，从而提高检测速度，可以作为今后的工作内容。

参考文献：

[1] 李凡, 董臣, 韩利群, 等. 智能变电站关键技术及功能研究[J]. **电力勘测设计**, 2023(8): 63-66. DOI:10.13500/j.dlkcsj.issn1671-9913.2023.08.011.
LI F, DONG C, HAN L Q, et al. Key technologies and application scenarios of intelligent substation[J]. *Electric Power Survey & Design*, 2023(8): 63-66. DOI:10.13500/j.dlkcsj.issn1671-9913.2023.08.011.

[2] 张春晓, 陆志浩, 刘相财. 智慧变电站联合巡检技术及其应用[J]. **电力系统保护与控制**, 2021, **49**(9): 158-164. DOI:10.19783/j.cnki.pspc.201045.
ZHANG C X, LU Z H, LIU X C. Joint inspection technology and its application in a smart substation[J]. *Power System Protection and Control*, 2021, **49**(9): 158-164. DOI:10.19783/j.cnki.pspc.201045.

[3] 刘涛, 王淑灵, 詹乃军. 多机器人路径规划的安全性验证[J]. **软件学报**, 2017, **28**(5): 1118-1127. DOI:10.13328/j.cnki.jos.005207.
LIU T, WANG S L, ZHAN R J. Safety verification of trajectory planning for multiple robots[J]. *Journal of Software*, 2017, **28**(5): 1118-1127. DOI:10.13328/j.cnki.jos.005207.

[4] 杨旭东, 黄玉柱, 李继刚, 等. 变电站巡检机器人研究现状综述[J]. **山东电力技术**, 2015, **42**(1): 30-34.
YANG X D, HUANG Y Z, LI J G, et.al. Research status review of robots applied in substations for equipment inspection[J]. *Shandong Electric Power*, 2015, **42**(1): 30-34.

[5] 冯振新, 周东国, 江翼, 等. 基于改进 MSER 算法的电力设备红外故障区域提取方法[J]. **电力系统保护与控制**, 2019, **47**(5): 123-128.
FENG Z X, ZHOU D G, JIANG Y, et al. Fault region extraction using improved MSER algorithm with application to the electrical system[J]. *Power System Protection and Control*, 2019, **47**(5): 123-128.

[6] Redmon J, Farhadi A. YOLOv3: An incremental improvement[J]. arXiv preprint arXiv:1804.02767, 2018.

[7] 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 & Machine Intelligence*, 2017, **39**(6): 1137-1149. DOI:10.1109/TPAMI.2016.2577031.

[8] TAO Xian, ZHANG Dapeng, WANG Zihao, et al. Detection of power line insulator defects using aerial images analyzed with convolutional neural networks[J]. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 2020, **50**(4): 1486-1498.

[9] 姚建刚, 付鹏, 李唐兵, 等. 基于红外图像的绝缘子串自动提取和状态识别[J]. **湖南大学学报**: 自然科学版, 2015, **42**(2): 74-80. DOI:10.16339/j.cnki.hdxbzkb.2015.02.012.
YAO J G, FU P, LI T B, et al. Algorithm research of automatically extracting the area of insulator from infrared image and state identification[J]. *Journal of Hunan University*: Natural Sciences, 2015, **42**(2): 74-80. DOI:10.16339/j.cnki.hdxbzkb.2015.02.012.

[10] 赵洪山, 张则言, 孟航, 等. 基于高压绝缘套管纹理特征的红外目标检测[J]. **红外技术**, 2021, **43**(3): 258-265.
ZHAO H S, ZHANG Z Y, MENG H, et al. Infrared target detection of high voltage insulation bushing based on textural features[J]. *Infrared Technology*, 2021, **43**(3): 258-265.

[11] Woo S, Park J, Lee J Y, et al. CBAM: convolutional block attention module[C]//*Proceedings of the European Conference on Computer Vision* (ECCV), 2018: 3-19.

[12] HU J, SHEN L, SUN G. Squeeze-and-excitation networks[C]// *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018: 7132-7141.

[13] LI X, WANG W, HU X, et al. Selective kernel networks[C]// *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019: 510-519.

[14] 伍艺佳, 华雄, 王丽蓉, 等. 基于注意力机制学习的变电设备缺陷检测方法[J]. **计算机与现代化**, 2021(2): 7-12, 17.
WU Y J, HUA X, WANG L R, et al. Method of substation equipment defect detection based on attention mechanism learning[J]. *Computer and Modernization*, 2021(2): 7-12, 17.

[15] 熊凯飞, 樊绍胜, 吴静. 基于改进YOLOv4的雾天变电站电力设备识别方法[J]. **无线电工程**, 2022, **52**(8): 1504-1512.
XIONG K F, FAN S S, WU J. Identification method of substation power equipment in foggy weathers based on improved YOLOv4[J]. *Radio Engineering*, 2022, **52**(8): 1504-1512.

[16] 于豪, 蒋锦霞, 赖晓翰, 等. 基于自适应感受野的电力设备表面缺陷检测方法[J]. **系统仿真学报**, 2023, **35**(7): 1572-1580. DOI:10.16182/j.issn1004731x.joss.22-0278.
YU H, JIANG J X, LAI X H, et al. Surface defect detection of power equipment using adaptive receptive field network[J]. *Journal of System Simulation*, 2023, **35**(7): 1572-1580. DOI:10.16182/j.issn1004731x.joss.22-0278.

[17] 国腾飞, 张则言, 付宏财, 等. 融合注意力机制的轻量级红外高压套管识别算法[J]. **计算机与现代化**, 2022(1): 70-76.
GUO T F, ZHANG Z Y, FU H C, et al. Lightweight infrared HV bushing identification algorithm based on attention mechanism[J]. *Computer and Modernization*, 2022(1): 70-76.

[18] 郑婷婷, 周浩, 王秋忆. 基于改进 YOLOv5 的电力设备检测算法[J]. **电子测量技术**, 2023, **46**(4): 155-160. DOI:10.19651/j.cnki.emt.2210570.
ZHENG T T, ZHOU H, WANG Q Y. Power equipment detection algorithm based on improved YOLOv5[J]. *Electronic Measurement Technology*, 2023, **46**(4): 155-160. DOI:10.19651/j.cnki.emt.2210570.

[19] LIU S, QI L, QIN H, et al. Path aggregation network for instance segmentation[C]//*Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018: 8759-8768.

[20] 舒子婷, 张泽斌, 宋尧哲, 等. 基于改进 YOLOv5 的低光照图像目标检测[J]. **激光与光电子学进展**, 2023, **60**(4): 77-84.
SHU Z T, ZHANG Z B, SONG Y Z. Low-light image object detection based on improved YOLOv5 algorithm [J]. *Laser & Optoelectronics Progress*, 2023, **60**(4): 77-84.

[21] Redmon J, Farhadi A. YOLOv3: An incremental improvement[J/OL]. [2018-04-8]. https://arxiv.org/ abs/1804.02767.

[22] Wagner J, Fischer V, Herman M, et al. Multispectral pedestrian detection using deep fusion convolutional neural networks[C]//*ESANN*, 2016, **587**: 509-514.

[23] LIU J, ZHANG S, WANG S, et al. Multispectral deep neural networks for pedestrian detection[J]. arXiv preprint arXiv:1611.02644, 2016.

[24] 马野, 吴振宇, 姜徐. 基于红外图像与可见光图像特征融合的目标检测算法[J]. **导弹与航天运载技术**, 2022(5): 83-87.
MA Y, WU Z Y, JIANG X. Object detection based on feature fusion of infrared and visible images[J]. *Missiles and Space Vehicles*, 2022(5): 83-87.

[25] LIU J, FAN X, HUANG Z, et al. Target-aware dual adversarial learning and a multi-scenario multi-modality benchmark to fuse infrared and visible for object detection[C]//*Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022: 5802-5811.

[26] 施政, 毛力, 孙俊. 基于 YOLO 的多模态加权融合行人检测算法[J]. **计算机工程**, 2021, **47**(8): 234-242. DOI:10.19678/j.issn.1000-3428.0058745.
SHI Z, MAO L, SUN J. YOLO-based multi-modal weighted fusion pedestrian detection algorithm[J]. *Computer Engineering*, 2021, **47**(8): 234-242. DOI:10.19678/j.issn.1000-3428.0058745.

[27] 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.
KUANG C W, HE W. Object detection algorithm based on infrared and visible light images[J]. *Infrared Technology*, 2022, **44**(9): 912-919.

[28] Dosovitskiy A, Beyer L, Kolesnikov A, et al. An image is worth 16×16 words: Transformers for image recognition at scale[J]. arXiv preprint arXiv:2010.11929, 2020.

[29] ZHU X, LYU S, WANG X, et al. TPH-YOLOv5: Improved YOLOv5 based on transformer prediction head for object detection on drone-captured scenarios[C]//*Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021: 2778-2788.

[30] Prakash A, Chitta K, Geiger A. Multi-modal fusion transformer for end-to-end autonomous driving[C]//*Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021: 7077-7087.

[31] Bochkovskiy A, WANG C Y, LIAO H Y M. YOLOv4: optimal speed and accuracy of object detection[J/OL]. arXiv preprint arXiv:2004.10934.

[32] FANG Qingyun, HAN Dapeng, WANG Zhaokui. Cross-modality fusion transformer for multispectral object detection[J]. arXiv preprint arXiv:2111.00273, 2021.

[33] SUN Y, CAO B, ZHU P, et al. Drone-based RGB-infrared cross-modality vehicle detection via uncertainty-aware learning[J]. *IEEE Transactions on Circuits and Systems for Video Technology*, 2022, **32**(10): 6700-6713.