基于特征交互与自适应分组融合的多模态目标检测

叶志晖[1]，武 健[1]，赵晓忠[1]，王文娟[1]，邵新光[2]
（1. 浙江中烟工业有限责任公司，浙江 杭州 310008；2. 浙江大学 工程师学院，浙江 杭州 310058）

摘要：为提升目标检测方法在复杂场景下的检测效果，将深度学习算法与多模态信息融合技术相结合，提出了一种基于特征交互与自适应分组融合的多模态目标检测模型。模型采用红外和可见光目标图像为输入，以 PP-LCNet 网络为基础构建对称双支路特征提取结构，并引入特征交互模块，保证不同模态目标特征在提取过程中的信息互补；其次，设计二值化分组注意力机制，利用全局池化结合 Sign 函数将交互模块的输出特征以所属目标类别进行特征分组，再分别采用空间注意力机制增强各特征组中的目标信息；最后，基于分组增强后的特征，提取不同尺度下的同类特征组，通过自适应加权方式由深至浅进行多尺度融合，并根据融合后的各尺度特征实现目标预测。实验结果表明，所提方法在多模态特征交互、关键特征增强以及多尺度融合方面都有较大的提升作用，并且在复杂场景下，模型也具有更高的鲁棒性，可以更好地适用于不同场景中。

关键词：多模态；目标检测；特征交互；二值化分组；自适应融合
中图分类号：TP391.41 **文献标志码：**A **文章编号：**1001-8891(2025)04-0468-07

Multimodal Object Detection Based on Feature Interaction and Adaptive Grouping Fusion

YE Zhihui[1]，WU Jian[1]，ZHAO Xiaozhong[1]，WANG Wenjuan[1]，SHAO Xinguang[2]
(1. *China Tobacco Zhejiang Industrial Co. LTD., Hangzhou* 310008, *China*;
2. *Polytechnic Instiute, Zhejiang University, Hangzhou* 310058, *China*)

Abstract: To improve the performance of object detection methods in complex scenes, a multimodal object detection model based on feature interaction and adaptive grouping fusion is proposed by combining deep learning algorithms with multimodal information fusion technology. The model uses infrared and visible object images as inputs, constructs a symmetrical dual-branch feature extraction structure based on the PP-LCNet network, and introduces a feature interaction module to ensure complementary information between different modal object features during the extraction process. Secondly, a binary grouping attention mechanism was designed. Global pooling combined with the sign function was used to group the output features of the interaction module into their respective object categories, and spatial attention mechanisms were used to enhance the object information in each group of features. Finally, based on the group-enhanced features, similar feature groups at different scales were extracted, and multi-scale fusion was carried out through adaptive weighting from deep to shallow. Object prediction was then achieved based on the fused features at each scale. The experimental results show that the proposed method significantly improves multimodal feature interaction, key feature enhancement, and multi-scale fusion. Moreover, in complex scenarios, the model exhibits higher robustness and can be better applied to different scenarios.

Key words: multimodal, object detection, feature interaction, binary grouping, adaptive fusion

0 引言

目标检测作为机器视觉的核心问题之一，旨在利用特征提取、分析等手段识别出感兴趣物体，并标注其所在图像中的位置[1-2]。近年来，随着计算机视觉技术与人工智能算法的不断突破，目标检测技术也取得

收稿日期：2023-10-25；**修订日期：**2023-11-21.
作者简介：叶志晖（1986-），男，汉族，浙江平阳人，高级工程师。研究方向：模式识别、信息技术、智能制造。E-mail: 1703389865@qq.com。
基金项目：国家自然科学基金（62002320）。

了长足的进步，被广泛应用于交通、医疗、安防、物流、家居等行业[3-4]。然而，随着检测环境以及目标数据愈发复杂、多样化，基于单模态的目标检测方法已难以满足实际需求[5-6]。因此，为解决单一传感器信息量少、鲁棒性差等局限，提升目标检测方法在实际应用中的泛化性及鲁棒性，融合目标多模态特征检测技术逐渐成为当前研究热点。

目前，多模态目标检测主要基于红外和可见光图像信息，通过红外相机和可见光相机分别捕获目标后，利用不同的融合策略实现多模态信息互补，再基于融合后的信息对目标进行预测[7-8]。而对于融合检测策略，现有研究可大致分为像素级、特征级以及决策级三个方向[9]，像素级融合主要是在输入数据层处理，如 Wu 等人[10]利用基于梯度的残差密集块来分别强化红外和可见光图像细节特征后进行拼接融合，再利用深度神经网络模型进行检测。特征级融合则是在目标特征提取过程中进行融合，如解宇敏等人[11]以 YOLOv5 为基本框架分别提取红外和可见光目标特征后，再将特征进行卷积融合后进行检测。而决策级融合则是对检测结果的合并，如宁大海等人[12]利用红外模型与可见光模型对目标分别检测后，根据目标预测框不同模态下的重叠程度进行融合。综上可见，现有的多模态检测方法虽取得了一定的进展[13-14]，但在目标多模态信息互补、多尺度融合、特征增强等方面仍然存在一定的局限。

针对上述问题，本文在总结前人多模态目标检测工作前提下，提出了一种新的基于特征级融合的多模态目标检测网络。网络以双支路结构分别提取红外和可见光图像特征，并通过交叉融合结构实现多模态特征交互；根据融合特征，设计二值化分组注意力结构来分别提升各组特征中的目标信息，再基于同类特征组进行多尺度融合后对目标类别及位置进行预测。模型针对各个目标特征分别增强，并配合自适应多模态、多尺度融合的方式使其可以在复杂场景中表现出更优的检测性能。

1 多模态目标检测模型

1.1 整体结构

为充分利用红外和可见光图像中的目标特征信息，模型以双支路 PP-LCNet 特征提取结构结合 YOLOV4 多尺度检测头为基础骨干结构，结合特征交互、分组、融合等策略实现多模态目标检测，如图 1 所示。模型首先利用 PP-LCNet 高效率卷积神经网络分别提取红外和可见光图像特征，同时针对网络降维处特征引入了交叉融合模块，保证多模态信息提取时相互补充；其次，将所提各维度多模态特征通过全局池化、二值激活、二值卷积等操作将特征进行类别分组，并结合空间注意力机制对各组目标空间位置信息进行增强；然后，根据分组特征提取同类目标对应的多尺度特征，并将深层特征逐步上采样后与相应浅层特征进行自适应融合；最后，基于多尺度融合后的特征对目标进行预测，再利用 soft NMS 算法筛选出最优目标。

1.2 多模态特征交叉融合提取

为保证多模态目标信息间的交互，设计了如图 2 所示的双支路特征交叉融合结构。

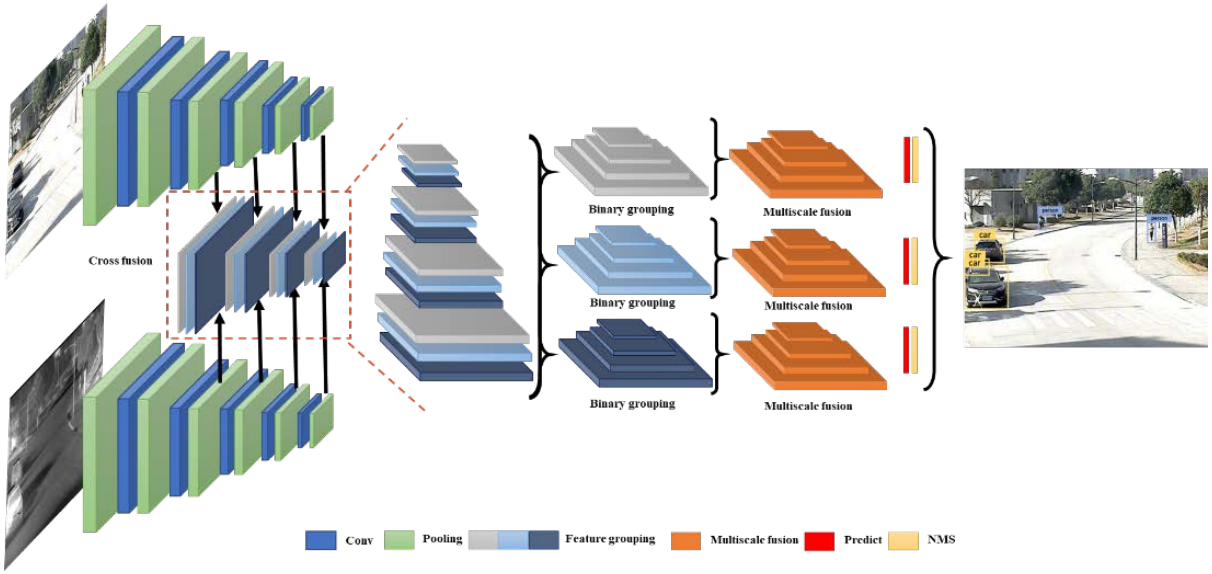


图 1 多模态目标检测整体结构

Fig.1 Multimodal object detection of overall structure

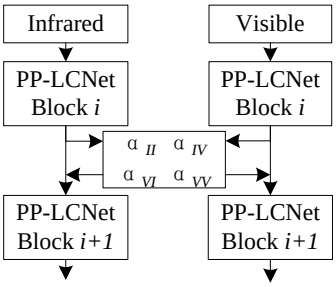


图 2 特征交叉融合结构

Fig.2 Feature cross fusion structure

所提多模态特征提取结构主要分为基础特征提取和多模态特征交互两部分。基础特征提取针对红外和可见光的输入图像，采用高效卷积神经网络 PP-LCNet[15]的骨干结构，分别对两个模态下的目标边缘细节以及抽象语义特征进行逐步提取，图中 i 取值[1, 5]。而特征交互部分则针对基础特征提取结构在不同阶段下提取的特征，引入自适应加权参数，将目标在不同模态下的特征进行交叉融合互补。特征交叉融合具体计算过程如式(1)所示，通过训练的方式不断调整加权参数，进而保证交互特征的有效性，同时丰富后续特征提取的多样性。

$$\begin{bmatrix} y_{\mathrm{I}}^{i} \\ y_{\mathrm{V}}^{i} \end{bmatrix} = \begin{bmatrix} \alpha_{\mathrm{II}} & \alpha_{\mathrm{IV}} \\ \alpha_{\mathrm{VI}} & \alpha_{\mathrm{VV}} \end{bmatrix} \begin{bmatrix} x_{\mathrm{I}}^{i} \\ x_{\mathrm{V}}^{i} \end{bmatrix} \tag{1}$$

训练时权重优化过程如式(2)(3)所示：

$$\begin{bmatrix} \dfrac{\partial L}{x_{\mathrm{I}}^{i}} \\ \dfrac{\partial L}{x_{\mathrm{V}}^{i}} \end{bmatrix} = \begin{bmatrix} \alpha_{\mathrm{II}} & \alpha_{\mathrm{IV}} \\ \alpha_{\mathrm{VI}} & \alpha_{\mathrm{VV}} \end{bmatrix} \begin{bmatrix} \dfrac{\partial L}{y_{\mathrm{I}}^{i}} \\ \dfrac{\partial L}{y_{\mathrm{V}}^{i}} \end{bmatrix} \tag{2}$$

$$\begin{bmatrix} \dfrac{\partial L}{\partial \alpha_{\mathrm{II}}} & \dfrac{\partial L}{\partial \alpha_{\mathrm{IV}}} \\ \dfrac{\partial L}{\partial \alpha_{\mathrm{VI}}} & \dfrac{\partial L}{\partial \alpha_{\mathrm{VV}}} \end{bmatrix} = \begin{bmatrix} \dfrac{\partial L}{y_{\mathrm{I}}^{i}} x_{\mathrm{I}}^{i} & \dfrac{\partial L}{y_{\mathrm{V}}^{i}} x_{\mathrm{I}}^{i} \\ \dfrac{\partial L}{y_{\mathrm{I}}^{i}} x_{\mathrm{V}}^{i} & \dfrac{\partial L}{y_{\mathrm{V}}^{i}} x_{\mathrm{V}}^{i} \end{bmatrix} \tag{3}$$

式中：x_{I}^{i}、x_{V}^{i} 分别指输入的红外和可见光在 i 通道上的特征；$\alpha_{\mathrm{II}}, \alpha_{\mathrm{IV}}, \alpha_{\mathrm{VI}}, \alpha_{\mathrm{VV}}$ 为特征交叉融合的权重参数，各参数取值[0, 1]区间，且满足$\alpha_{\mathrm{II}}+\alpha_{\mathrm{IV}}=1$，$\alpha_{\mathrm{VI}}+\alpha_{\mathrm{VV}}=1$；$y_{\mathrm{I}}^{i}$、$y_{\mathrm{V}}^{i}$ 为交叉融合后的红外和可见光特征，∂表示求导数操作。由上式可见，若$\alpha_{\mathrm{II}}=\alpha_{\mathrm{VV}}=1$，$\alpha_{\mathrm{IV}}=\alpha_{\mathrm{VI}}=0$，则表示对应 i 通道特征不进行交叉融合，通过训练调节交互参数，进而达到最优的融合结果。

1.3 二值化分组注意力模块

二值化分组注意力模块主要作用在于提升目标所关联特征的权重，基本结构如图 3 所示。

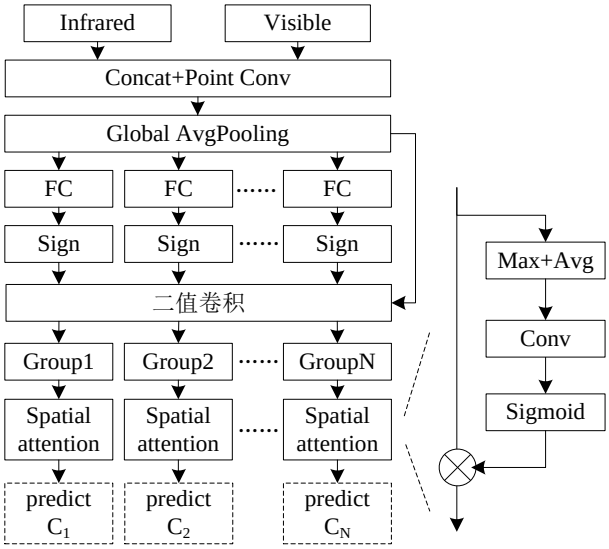


图 3 二值化分组注意力结构

Fig.3 Binary grouping attention structure

所提注意力结构以特征交叉后的红外和可见光特征作为输入，将两个模态特征拼接后通过点卷积进行融合，再利用全局平均池化将每个通道特征转换为一个具有全局感受野的实数；然后，采用 N 组全连接操作将池化后的一维特征扩增成 N 组，N 为目标类别数；其次，利用如式(4)所示的 Sign 二值函数将每组特征转化为 0 和 1 后与输入特征相乘，进而实现对每类目标对应的特征通道进行分组；最后，将每个特征组中的特征归一化至相同维度后，利用最值、均值配置卷积、Sigmoid 函数构建空间位置注意力机制来分别对每组特征进行增强，进而实现更有针对性的目标特征增强。同时，在实际训练过程中，为更好地指导特征分组，将 0-1 分组后特征结合辅助损失来分别预测目标类别。并且，由于 Sign 函数梯度为零，故训练时的误差反向传播采用 Htanh 进行替代，计算方式如式(5)所示。

$$\mathrm{Sign}(x)=\begin{cases} 1 & x>0 \\ 0 & x\le 0 \end{cases} \tag{4}$$

$$\mathrm{Htanh}(x)=\max(-1,\ \min(1,x)) \tag{5}$$

式中：x 为全局平均池化后的特征信息；Sign 为前向传播函数，取值为 1 时则表示提取对应通道特征，反之则舍弃；Htanh 为反向传播函数，取值区间[−1, 1]，有效避免了 Sign 倒数为 0 的限制。

1.4 自适应多尺度分组融合

多尺度结构主要是综合目标在不同维度下的特征，使浅层细节信息与深层语义信息充分融合，提升模型对不同大小目标的识别效果。为降低特征融合时

不同目标间的相互干扰，所提多尺度结构在注意力模块基础上，采用了分组策略来对同类多尺度特征进行自适应融合，结构如图 4 所示。

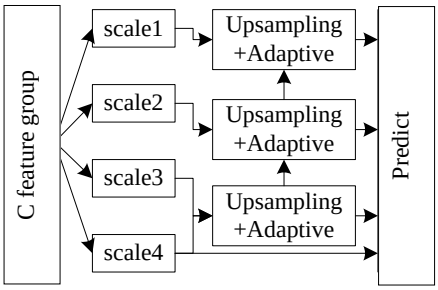


图 4　类别 C 特征组自适应多尺度融合

Fig.4　Class C feature group adaptive multi-scale fusion

该结构以注意力模块增强后的特征组作为输入，根据特征对应的目标类别，提取同类目标在不同尺度下的特征组，将其由深层到浅层逐步上采样与对应尺度特征进行自适应加权融合，融合计算方式如(6)所示。以此类推，将每个类别目标对应的特征组进行多尺度融合。最后，基于融合后的特征，借鉴 YOLOV4 检测头结构实现对各目标类别及位置的预测，并结合 softNMS 算法提取最优检测结果。

$$y_{ic}^{s}=\alpha_{ic}x_{ic}^{d\to s}+\beta_{ic}x_{ic}^{s} \tag{6}$$

误差反向传播时如式(7)所示：

$$\frac{\partial L}{\partial \alpha_{ic}}=\frac{\partial L}{\partial y_{ic}^{s}}\frac{\partial y_{ic}^{s}}{\partial \alpha_{ic}}=x_{ic}^{d\to s}\frac{\partial L}{\partial y_{ic}^{s}} \tag{7}$$

式中：i 表示类别；c 表示特征通道；s 表示浅层特征；d 表示深层特征；x_{ic} 表示属于的浅层 c 通道输入特征；$x_{ic}^{d\to s}$ 表示将类别 i 的深层 c 通道特征上采样至浅层尺度；α_{ic} 和 β_{ic} 为对应的自适应参数，且参数满足 $\alpha_{ic},\beta_{ic}\in[0,1]$，$\alpha_{ic}+\beta_{ic}=1$；$y_{ic}^{s}$ 表示类别 i 在 s 层自适应融合后的特征。

2　实验结果与分析

为验证所提方法的有效性，实验采用 Ubuntu18.04.3 操作系统，NVIDIA GeForce RTX 2070 显卡以及 Intel Core i7-10750H 的 CPU 作为测试平台。模型基于 Pytorch 深度学习框架进行搭建，训练测试时采用 KAIST、RGBT、FLIR 公开标准红外-可见光数据集，各数据集详细信息如表 1 所示。

实验将数据集中图像以 7:1:2 构建训练验证测试集，训练时超参数设置如表 2 所示，并通过目标检测精度（mAP、mAP_s、mAP_m、mAP_l）和每秒处理图像数（fps）等指标对模型的精度与效率进行评估。

2.1　可行性实验

对于所提结构的可行性及有效性验证，实验以双支路 PP-LCNet 特征提取结构结合 YOLOV4 多尺度检测头作为基础结构，利用 KAIST 数据集来依次对交叉模块、分组注意力模块以及多尺度融合模块进行训练测试。针对特征交叉模块，实验主要考虑了多模态特征叠加融合（Add）、拼接融合（Concat）以及交叉融合（Cross）3 种策略，并分别进行了对比测试，结果如表 3 所示。

表 1　数据集具体信息

Table 1　Dataset details information

	KAIST	FLIR	RGBT
Quantity	3600	14000	5000
Size	640×480	512×512	640×480
Categories	6	5	11

表 2　超参数设置

Table 2　Hyperparameter setting

Hyperparameters	Value
Batch size	4
Initial learning rate	0.01
Momentum parameter	0.95
Weight attenuation coefficient	0.0005
Initialization strategy	Gaussian
Adjustment of learning rate	Sequential LR
Optimization	Adam
Position loss	CIou loss
Category loss	Cross Entropy

表 3　多模态特征融合策略对比

Table 3　Comparison of multimodal feature fusion strategies

Network	Efficiency /(fps)	Accuracy/(%)			
		mAP	mAP_s	mAP_m	mAP_l
Add	38	76.8	57.1	77.2	86.1
Concat	36	77.5	58.3	77.8	86.9
Cross	38	77.3	58.2	77.7	86.7

根据不同策略的对比结果可见，叠加融合的方式虽然丰富了目标特征，但也引入了不同模态下的噪声信息，故检测精度相对较低；拼接融合采用点卷积操作来融合所有通道特征，虽充分利用了全局信息，但也引入了较多冗余计算。而交叉融合方式较好地综合了叠加与拼接融合的优势，通过可训练加权参数来自适应融合不同模态下的目标特征，使得检测精度与效率都达到相对较优的结果。对于所提二值化分组注意

力模块，实验分别与传统 ECA[16]、CBAM[17]等结构进行了对比，并可视化展示了 Block3 特征层增强效果，结果如表 4 和图 5 所示。

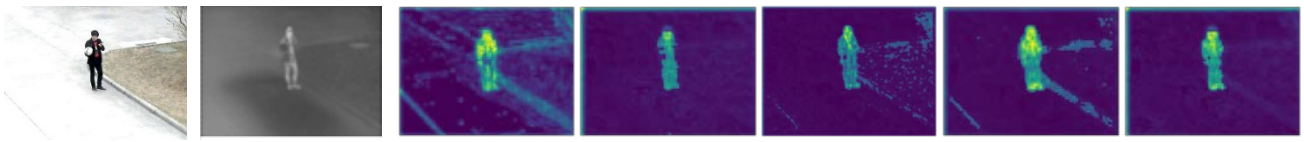
表 4 注意力机制对比

Table 4 Comparison of attention mechanisms

Network	Efficiency /(fps)	Accuracy/(%)			
		mAP	mAP_s	mAP_m	mAP_l
Lack attention	38	77.3	58.2	77.7	86.7
ECA[16]	37	78.2	58.9	78.6	87.8
CBAM[17]	36	78.4	59.3	78.8	88.1
Group attention	35	79.3	60.2	79.6	89.0

通过不同注意力机制的对比结果可见，传统注意力结构基于所有特征信息进行建模，虽能提升目标相关特征权重，但其针对性相对较弱，无法聚焦目标关键特征。而所提分组注意力结构以可训练的二值卷积来区分不同目标特征，虽引入了一定计算量，但却可以更精确地针对目标特征进行增强。对于多尺度结构，实验分别对比测试了分组与不分组情况下多尺度融合效果，结果如表 5 所示。其中，不分组即每个维度不区分特征组所属类别，将该维度所有特征组与其他维度自适应融合。

根据表 5 对比结果可以看出，分组与不分组对于特征自适应融合时的计算量基本相同，但分组后的特征在自适应融合时避免了不同目标间的相互干扰。同时，在目标预测时基于分组特征的结构将多目标检测转变成了单目标检测，有效降低了模型的复杂度，进而检测效果更佳。而对于整体模型的有效性验证，实验分别与当前基于像素级（LAINet[10]）、特征级（CSPNet[11]）以及决策级（DFNet[12]）的多模态目标检测进行对比，结果见表 6 和图 6 所示。



(a) 可见光 (b) 红外 (c) 可见光特征 (d) 红外特征 (e) ECA (f) CBAM (g) 分组注意力
(a) Visible light (b) Infrared (c) Visible light features (d) Infrared features (e) ECA (f) CBAM (g) Group attention

图 5 Block3 层特征增强效果对比

Fig.5 Comparison of Block3 layer feature enhancement effects

表 5 多尺度分组融合结构对比

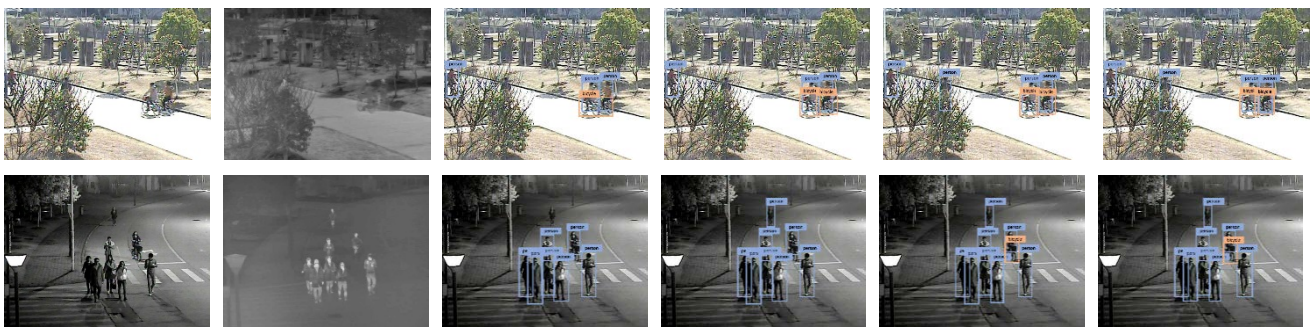
Table 5 Comparison of multi-scale grouping fusion structures

Network	Efficiency /(fps)	Accuracy/(%)			
		mAP	mAP_s	mAP_m	mAP_l
Non grouped	35	79.3	60.2	79.6	89.0
Group	35	79.8	60.8	80.2	89.4

表 6 多模态目标检测模型对比

Table 6 Comparison of multimodal object detection models

Network	Efficiency /(fps)	Accuracy/(%)			
		mAP	mAP_s	mAP_m	mAP_l
LAINet[10]	41	77.7	58.6	78.3	88.2
CSPNet[11]	37	78.9	59.7	79.6	89.7
DFNet[12]	33	78.4	59.3	79.0	88.6
Ours	35	79.8	60.8	80.2	89.4



(a) 可见光 (b) 红外 (c) LAINet[10] (d) CSPNet[11] (e) DFNet [12] (f) 本文
(a) Visible light (b) Infrared (c) LAINet[10] (d) CSPNet[11] (e) DFNet [12] (f) Ours

图 6 KAIST 数据集检测结果对比

Fig.6 Comparison of detection results in KAIST dataset

根据上述结果可以看出，由于像素级方式在不同模态图像融合时，噪声干扰相对严重，故容易造成小目标漏检。而决策级方式由于综合了所有模态下的预测结果，使得预测框成倍增加，进而导致误检率较高。相比之下，基于特征级融合的目标检测方法较好地协调了两者劣势，可以更好地利用不同模态下的目标特征，使得检测精度相对更高。而所提方法在特征级融合检测的基础上考虑了对不同目标分组融合的策略，较好地降低了目标间相互影响以及噪声信息的干扰，进而进一步提升了目标检测效果。

2.2 鲁棒性实验

为进一步验证本文算法在不同场景下的鲁棒性，实验分别利用了相对复杂的 RGBT 和 FLIR 数据集对各个多模态目标检测模型进行了对比测试，结果如表 7、表 8 和图 7 所示。其中，表 7 和表 8 中 dr 指标表示各个方法 mAP 精度相对于 KAIST 数据集下的 mAP 精度差距。

根据上述测试结果可见，由于数据集对应的场景复杂度较高，且待检测目标类别增多，使得各模型的检测精度都有一定的下降。而通过对比各模型的精度变化可以看出，所提模型在两个数据集下的精度下降率都最低。由此可见，模型具有更高的鲁棒性及稳定性，可以更好地适用于复杂场景多模态目标检测任务中。

表 7 RGBT 数据集测试结果对比

Table 7 Comparison of RGBT dataset test results

Network	dr	Efficiency /(fps)	Accuracy/(%)			
			mAP	mAP_s	mAP_m	mAP_l
LAINet[10]	2.4	40	75.3	56.5	76.0	86.1
CSPNet[11]	1.9	36	77.0	57.9	77.8	87.8
DFNet[12]	2.1	32	76.3	57.2	77.2	86.9
Ours	1.3	33	78.5	59.3	78.7	88.5

表 8 FLIR 数据集测试结果

Table 8 Comparison of FLIR dataset test results

Network	dr	Efficiency /(fps)	Accuracy/(%)			
			mAP	mAP_s	mAP_m	mAP_l
LAINet[10]	3.5	38	74.2	54.9	75.1	84.8
CSPNet[11]	3.1	34	75.8	56.4	76.7	86.6
DFNet[12]	3.0	29	75.4	56.2	76.3	85.9
Ours	2.4	30	77.4	58.1	77.6	87.1

(a) 可见光 (b) 红外 (c) LAINet[10] (d) CSPNet[11] (e) DFNet [12] (f)本文

(a) Visible light (b) Infrared (c) LAINet[10] (d) CSPNet[11] (e) DFNet [12] (f) Ours

图 7 RGBT 和 FLIR 数据集检测结果（前两行：RGBT；后两行：FLIR）

Fig.7 RGBT and FLIR dataset detection results(first two rows: RGBT; second two rows: FLIR)

3 结束语

本文针对现阶段基于红外及可见光图像的多模态目标检测方法存在的不足，从特征交互以及特征分组融合两个角度展开深入研究，提出了一种基于特征级融合的多模态目标检测模型。模型利用 PP-LCNet

网络作为特征提取骨干结构，以并列交互的方式逐步提取红外和可见光目标信息；同时，针对多模态交互特征引入了分组策略，根据特征关联目标类别对其进行分组，并利用注意力机制对每组特征中目标空间位置信息进行增强；最后，将不同尺度下的同类特征组以逐层采样和自适应加权的方式实现多尺度特征融合，并基于融合后的特征来对对应类别目标进行预测。通过在 KAIST、FLIR、RGBT 公开标准数据集上的实验结果有效验证了所提多模态目标检测方法的有效性和鲁棒性，可以准确高效地实现全天候目标检测。尽管所提方法取得了较好的效果，但网络对特征分组的操作较为繁琐，下一步工作将针对特征分组问题进行持续优化，同时进一步探索所提检测方法在多源异构数据下的可行性。

参考文献：

[1] 孙涵, 刘译善, 林昱涵. 基于深度学习的显著性目标检测综述[J]. **数据采集与处理**, 2023, **38**(1): 21-50.
SUN Han, LIU Yishan, LIN Yuhan. A review of salient object detection based on deep learning[J]. *Data Collection and Processing*, 2023, **38**(1): 21-50.

[2] KANG J, Tariq S, Oh H, et al. A survey of deep learning-based object detection methods and datasets for overhead imagery[J]. *IEEE Access*, 2022, **10**: 20118-20134.

[3] 张静, 农昌瑞, 杨智勇. 基于卷积神经网络的目标检测算法综述[J]. **兵器装备工程学报**, 2022, **43**(6): 37-47.
ZHANG Jing, NONG Changrui, YANG Zhiyong. Overview of object detection algorithms based on convolutional neural networks[J]. *Journal of Weapon Equipment Engineering*, 2022, **43**(6): 37-47.

[4] JIAO L, ZHANG F, LIU F, et al. A survey of deep learning-based object detection[J]. *IEEE Access*, 2019, **7**: 128837-128868.

[5] 汪鹏, 张大蔚. 基于多源信息和时空约束的移动目标检测算法[J]. **信息技术与信息化**, 2022(11): 82-85.
WANG Peng, ZHANG Dawei. Mobile object detection algorithm based on multi-source information and spatiotemporal constraints[J]. *Information Technology and Informatization*, 2022(11): 82-85.

[6] YAO X, ZHAO S, XU P, et al. Multi-source domain adaptation for object detection[C]//*Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021: 3273-3282.

[7] 陶洋, 祝小钧, 杨柳. 基于皮尔逊相关系数和信息熵的多传感器数据融合[J]. **小型微型计算机系统**, 2023, **44**(5): 1075-1080.
TAO Yang, ZHU Xiaojun, YANG Liu. Multi sensor data fusion based on Pearson correlation coefficient and information entropy[J]. *Small Microcomputer System*, 2023, **44**(5): 1075-1080.

[8] LI H, WU X J, Kittler J. RFN-Nest: an end-to-end residual fusion network for infrared and visible images[J]. *Information Fusion*, 2021, **73**: 72-86.

[9] YANG Y, LIU J, HUANG S, et al. Infrared and visible image fusion via texture conditional generative adversarial network[J]. *IEEE Transactions on Circuits and Systems for Video Technology*, 2021, **31**(12): 4771-4783.

[10] WU J, SHEN T, WANG Q, et al. Local adaptive illumination-driven input-level fusion for infrared and visible object detection[J]. *Remote Sensing*, 2023, **15**(3): 660.

[11] 解宇敏, 张浪文, 余孝源, 等. 可见光－红外特征交互与融合的 YOLOv5 目标检测算法[J]. **控制理论与应用**, 2024, **41**(5): 914-922.
XIE Yumin, ZHANG Langwen, YU Xiaoyuan, et al. YOLOv5 object detection algorithm with visible-infrared feature interaction and fusion[J]. *Control Theory and Technology*, 2024, **41**(5): 914-922.

[12] 宁大海, 郑晟. 可见光和红外图像决策级融合目标检测算法[J]. **红外技术**, 2023, **45**(3): 282-291.
NING Dahai, ZHENG Sheng. Decision level fusion target detection algorithm for visible light and infrared images[J]. *Infrared Technology*, 2023, **45**(3): 282-291.

[13] 吴泽, 缪小冬, 李伟文, 等. 基于红外可见光融合的低能见度道路目标检测算法[J]. **红外技术**, 2022, **44**(11): 1154-1160.
WU Ze, MIAO Xiaodong, LI Weiwen, et al. Low visibility road target detection algorithm based on infrared visible light fusion[J]. *Infrared Technology*, 2022, **44**(11): 1154-1160.

[14] BAO C, CAO J, HAO Q, et al. Dual-YOLO architecture from infrared and visible images for object detection[J]. *Sensors*, 2023, **23**(6): 2934.

[15] CUI C, GAO T, WEI S, et al. PP-LCNet: a lightweight CPU convolutional neural network[J]. arXiv preprint arXiv, 2021: 2109.15099.

[16] WANG Q, WU B, ZHU P, et al. ECA-Net: efficient channel attention for deep convolutional neural networks[C]//2020 *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020: 11534-11542.

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