基于多注意力机制的红外与可见光图像夜间目标检测

黎瑞虹[1]，付志涛[1]，张韶琛[1]，张 健[1]，王雷光[2]
（1. 昆明理工大学 国土资源工程学院，云南 昆明 650093；
2. 西南林业大学 森林生态大数据国家林业和草原重点实验室，云南 昆明 650024）

摘要：目标检测一直是计算机视觉领域的研究热点，YOLO 系列目标检测模型已广泛应用于多个领域。然而，目前关于目标检测的图像数据大多是基于单一类型传感器，难以完整地表征成像场景，且检测到的目标所包含有用信息具有局限性，尤其是在低照度、夜晚、雨雾等条件下，目标检测更加困难。为了更好地检测夜间目标，本文提出了一种结合 CBAM 注意力机制与 Transformer 的多注意力机制的红外与可见光图像夜间目标检测方法，通过添加 Transformer 来获取丰富的局部和上下文信息，通过添加 CBAM 注意力机制来减少误检。为了验证方法的有效性，本文选取了 5 种当前主流的目标检测算法在公开红外目标检测数据集上进行测试，本文方法与原始 YOLO v7 相比，mAP 从 62.6%提升至 71.5%。本文还制作了一个用于夜间目标检测红外-可见光融合目标检测数据集。在该数据集上与原始 YOLOv7 相比，mAP 从 79.90%提升至 94.80%，效果非常显著。

关键词：多注意力；夜间目标检测；红外与可见光图像；YOLOv7
中图分类号：TP391.4 **文献标识码：**A **文章编号：**1001-8891(2024)12-1371-09

Nighttime Object Detection in Infrared and Visible Images Based on Multi-Attention Mechanism

LI Ruihong[1]，FU Zhitao[1]，ZHANG Shaochen[1]，ZHANG Jian[1]，WANG Leiguang[2]
(1. *Faculty of Land Resources Engineering*, *Kunming University of Science and Technology*, *Kunming* 650093, *China*;
2. *Key Laboratory of State Forestry and Grassland Administration on Forestry and Ecological Big Data, Southwest Forestry University*, *Kunming* 650024, *China*)

Abstract Object detection has long been a research hotspot in the field of computer vision, and the YOLO series of object detection models is widely used in numerous fields. However, most current image data for object detection are based on a single type of sensor, which makes it difficult to fully characterize the imaging scene. The detected objects contain limited useful information, especially under conditions of low illumination, night, rain, and fog. To improve nighttime object detection, our study proposed a multi-attention mechanism for infrared and visible images. This mechanism combines the CBAM attention mechanism with a Transformer to obtain rich local and contextual information and reduce false detections. To verify the effectiveness of the method, five current mainstream object detection algorithms were selected and tested on a public infrared object detection dataset. The mAP of the proposed method improved from 62.6% to 71.5% compared to the original YOLOv7. This study also produced an infrared–visible fusion dataset for nighttime object detection. On this dataset, the mAP improved significantly from 79.90% to 94.80% compared to the original YOLOv7.

Key words: multi-attention mechanism, night object detection, infrared and visible images, YOLOv7

收稿日期：2023-07-23；**修订日期：**2023-08-16；
作者简介：黎瑞虹（1998-），女，硕士研究生，主要从事目标检测、图像融合方面研究。E-mail：july_lrh@163.com。
通信作者：付志涛（1982-），男，副教授，博士，主要从事多源遥感图像处理与大数据分析研究。E-mail：zhitaofu@126.com。
基金项目：国家自然科学重点基金项目（41961053）；云南省重大科技专项（202202AD080010）；云南省科技厅基础研究计划面上项目（202301AT070463、202201AT070164）；森林生态大数据国家林业和草原局重点实验室开放基金重点项目（2022-BDK-01）；“兴滇英才支持计划”项目（KKRD202221041）。

0 引言

目标检测作为计算机视觉非常重要的一个研究领域，它的主要任务是对图像或视频中的目标进行识别和定位。随着相关技术的不断发展，目标检测技术也在不断进步，现在已经广泛应用于多个领域，例如实例分割[1]、图像标注[2]和目标跟踪[3]等。

对于夜间图像的目标检测一直是当前研究的热点和难点。主要是夜间图像视觉特征模糊，同时还容易受各种噪声的干扰，目标物体的边缘细节也很难被检测出来。虽然红外图像能够在夜晚、昏暗的场景下不受光照的干扰，夜间场景下的目标检测大多利用红外图像[4-5]，然而，单一的红外图像依然不能解决红外图像缺乏纹理、边缘模糊以及检测精度低等问题。考虑到红外与可见光图像的信号来自不同的传感器，从不同方面提供场景与物体信息，即可见光图像捕捉反射光，而红外图像捕捉热辐射，这两者之间的融合比单传感器图像能提供更加丰富、有效的信息[6]。因此，即使在夜间场景检测中可见光所包含的信息也是有利于目标检测，同时还能获得丰富的信息（相较于图1(a)，图1(b)在1，2中能清晰看到车牌号和车身等细节信息，同时可以检测到更多的目标，如3，4，5）。基于此，学者们展开了基于红外与可见光图像融合的目标检测研究。Ma等人[7]提出了一种基于显著目标检测的红外与可见光图像融合网络，它能够保留异源图像中各自的优点，能够更准确地检测和识别目标。该方法虽然能够通过融合图像进一步表征待检测目标，但此方法只是注重于图像融合本身，后续并没有将融合图像进一步用于目标检测。Chen等人[8]提出了一种面向可见光和SAR图像的决策级融合算法来实现基于多源融合的目标检测，结果表明，该算法的检测性能优于基于单一传感器图像的目标检测。同时，该方法具有更少的误检和漏检，可是没有提供用于目标检测的数据集和目标检测的算法。

综上，目前基于融合图像目标检测已经逐渐发展起来，相关研究人员已经关注到融合图像的实际应用，仍然存在数据集不足、检测精度不高等问题，并且针对夜间可见光与红外图像的目标检测算法也较少。针对以上问题，本文将夜间红外、可见光异源图像融合作为数据集，并对YOLOv7目标检测网络进行改进，通过引入CBAM（convolutional block attention module）注意力机制以及Transformer，提出了一种基于多注意力机制的YOLOv7夜间目标检测网络，来提高网络的特征提取能力和捕获局部信息的能力。通过在FLIR数据集和自制的可见光与红外融合图像数据集上进行对比实验，验证了本文方法的有效性。

(a) 红外图像检测结果 (b) 融合图像检测结果

(a) Infrared image detection results (b) Fusion image detection results

图1 红外与融合图像目标检测对比

Fig. 1 Comparison of infrared and fused image object detection

1 算法原理

1.1 YOLOv7

YOLOv7[9]是目前主流的目标检测算法之一，具有高检测精度，并且能够快速处理大量的图像。此外，它还能够满足不同环境下的检测需求。YOLOv7 网络结构如图 2 所示，检测思路与 YOLOv4[10]、YOLOv5 大致相同。创新之处在于：①Backbone 加入 multi concat block 进行特征提取，使网络的连接结构更加的紧密；②在 SPP（spatial pyramid pooling）结构中引入 CSP（cross stage partial），进一步扩大感受野等。

1.2 本文整体网络结构

本文提出一种结合 CBAM[11]注意力机制与 Transformer[12]的多注意力机制的红外与可见光目标检测，用于夜间场景下的行人、车辆等目标检测。整体框架如图 3 所示。通过在 YOLOv7 网络基础上加入 Transformer Encoder 和 CBAM 注意力机制模块来提升目标网络的精度。具体地：①在 3 个检测头上，分别加入了一个 CBAM 注意力机制模块来提高网络对图像的特征提取，CBAM 能够在通道与空间维度进行注意力增强；②在 CBAM 的上层以及 backbone 的末层，我们加入了 transformer encoder 结构，提高网络对局部特征的提取能力；③在 backbone 的末端也将相同的 transformer encoder 结构加入其中，来提升网络的检测精度。

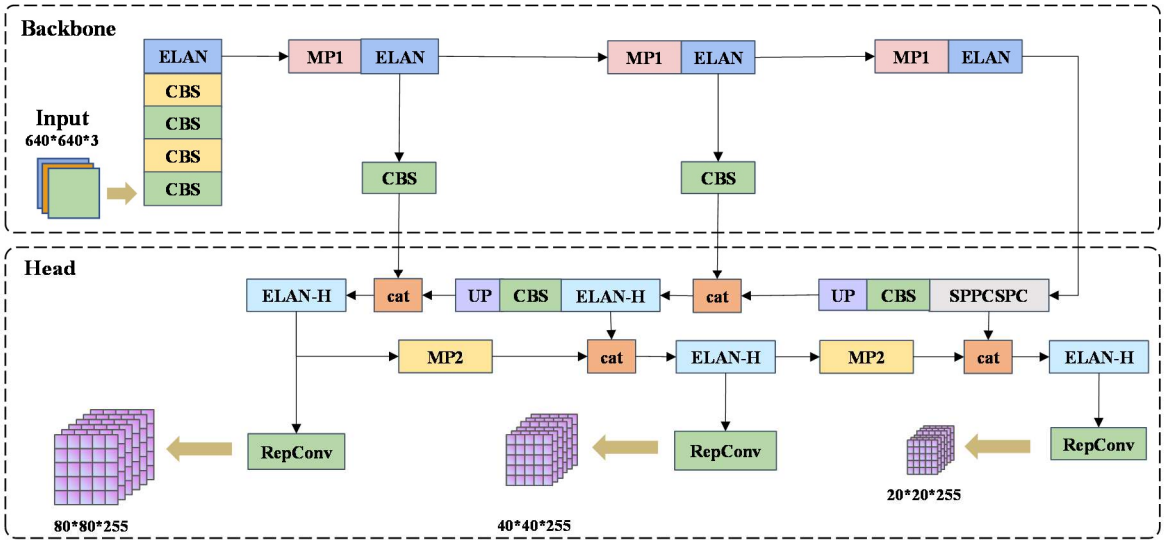


图 2 YOLOv7 网络结构

Fig. 2 YOLOv7 network structure diagram

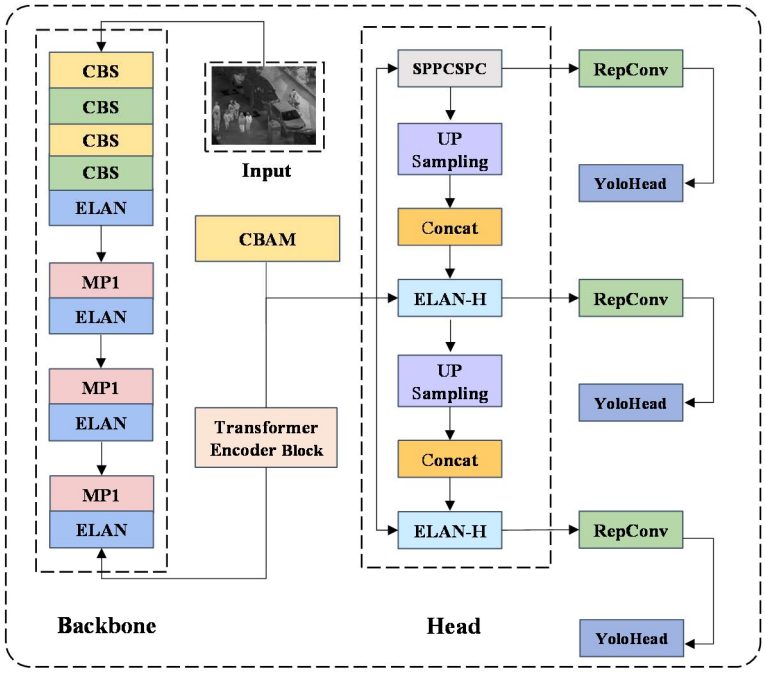


图 3 结合 CBAM 注意力机制与 Transformer 多注意力机制目标检测网络

Fig. 3 Object detection network combining CBAM attention mechanism and Transformer multi-attention mechanism

1.3 CBAM注意力机制模块

为了提高网络的特征提取能力，我们增加了 CBAM 注意力机制模块。如图 4 所示，CBAM 包括 CAM（channel attention module）和 SAM（spatial attention module）两个子模块，分别从通道和空间两个维度完成注意力增强[11]。此外，融合 CBAM 注意力机制模块不仅能够节约网络参数和计算能力，同时作为即插即用的模块，能够直接应用到已有的 YOLOv7 网络架构中。

由于 CBAM 模块先经过通道维度后经过空间维度，能够获得较低错误率的同时具有更高的准确率。因此，本文将 CBAM 模块放置于 YOLOv7 网络的检测头上，让改进后的网络更加关注重点区域目标，同时抵御目标检测过程中的混淆信息。

CBAM 结构中通道注意力模块具体计算如式(1)所示：

$$M_c = \sigma[\mathrm{MLP}(\mathrm{Avg}P(F))] + \mathrm{MLP}(\mathrm{MaxPool}(F)) = \sigma[W_1(W_0(F_{avg}^c))] + W_1(W_0(F_{max}^c)) \quad (1)$$

式中：σ为激活函数；W_0、W_1为MLP权值 $W_0 \in R^{C/r \times C}$，$W_1 \in R^{C \times C/r}$。

空间注意力模块（Spatial Attention Module）具体计算如式(2)所示：

$$M_s(F) = \sigma[(f^{7\times7}([\mathrm{AvgPool}(F);\mathrm{MaxPool}(F)]) = \sigma[(f^{7\times7}([F_{avg}^s; F_{max}^s]) \quad (2)$$

式中：σ为激活函数；$f^{7\times7}$是 alter size 为 7×7 卷积操作。

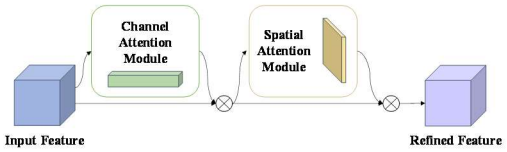


图 4 CBAM 注意力机制模块

Fig. 4 CBAM attention mechanism block

1.4 Transformer Encoder 结构

为了能够捕获全局信息和丰富的上下文信息，本文还在 CBAM 的上层以及 Backbone 的末端加入了 Transformer Encoder 结构，它是由多个编码器块（Encoder Block）组成，每个编码器块包含一个 Self-Attention 层和一个全连接前馈网络层，具体结构如图 5 所示。Self-Attention 层是 Transformer 的核心部分，它能够根据输入序列的内容自动计算序列中各个位置的重要性权重，并用这些权重来对序列进行加权求和并生成上下文表示。全连接前馈网络则用于处理 Self-Attention 层的输出，从而更好地提取信息。这些层可以通过堆叠来形成一个深层的 Transformer Encoder，以便逐步提取序列中的更高层次、抽象的特征信息[12]。这种自注意力机制能够在不同位置对输入序列进行推理，从而更好地捕捉序列之间的关系和依赖关系。

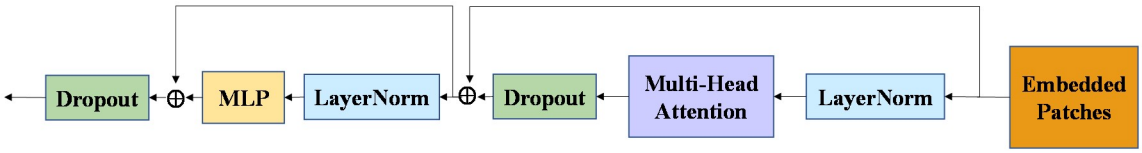


图 5 Transformer encoder 结构

Fig. 5 Transformer encoder structure

自注意力机制具体计算如式(3)所示：

$$\mathrm{Attention}(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V}) = \mathrm{soft\,max}(\frac{\boldsymbol{Q}\boldsymbol{K}^{\mathrm{T}}}{\sqrt{d_k}}) \quad (3)$$

式中：$\boldsymbol{Q}\boldsymbol{K}^{\mathrm{T}}$是注意力矩阵，$\boldsymbol{Q}$、$\boldsymbol{K}$、$\boldsymbol{V}$为线性映射后形成的 3 个矩阵。

Feed Forward 具体计算如式(4)所示：

$$X_{hidden} = \mathrm{Activate}[L(L(X_{attention}))] \quad (4)$$

式中：L 表示两层线性映射激活函数。

2 实验结果与分析

2.1 实验基础

2.1.1 实验环境

本文实验基于 PyTorch 框架，GPU 为 NVIDIA GTX3090，Windows 10 系统。使用 Adam 优化算法来训练网络模型。具体地，在初始训练阶段学习率为 0.01，通过余弦函数来降低学习率，最终学习率为 0.1。我们在 FLIR 数据集上训练 200 轮，用时约 8 h；而在自制数据集上同样训练 200 轮，花费约 4 h。

2.1.2 数据与预处理

为了对我们数据的实用性以及我们算法的有效性进行验证，在对比实验中我们选取 FLIR 红外目标检测数据作为对比实验数据集，包括白天和夜晚场景，已过滤筛选为 3 个类别，即行人、自行车、汽车，可用于 YOLO 网络训练，共包含 8862 张训练集，1366 张测试集数据。

目前，基于红外-可见光融合图像的用于夜间目标检测的数据集还较少，因此本文制作了一个夜间红外-可见光融合目标检测数据集，制作过程如图 6 所示，应用于夜间融合图像关于行人、汽车以及自行车的检测。在深度学习方面，具有代表性的融合图像算法有 FusionGAN[13]、GTF[14]、U2Fusion[15]以及 DenseFuse[16]等，对比结果如图 7 所示。根据 LLVIP 中提供的实验结果以及数据，本实验采用 DenseFuse 融合方法，从 LLVIP 数据集中挑选包块十字路口、人行道等多场景的图像进行融合。

数据集融合了 1009 张可见光-红外夜间图像，同时将融合后图像利用 labeling 标注软件进行人工手动标注，标注类别为行人、车辆以及自行车，图 8 所示即为标注实例。标注结束后，将 1009 张图像随机抽取分为训练集（756 张）、验证集（202 张）与测试集（51 张）。

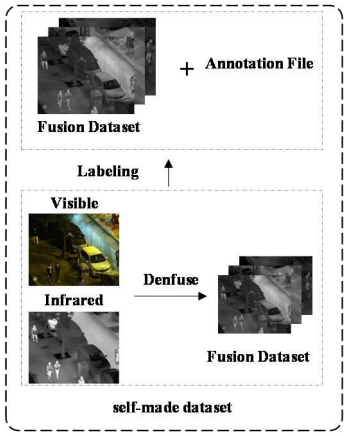


图 6 数据集制作

Fig. 6 Self-annotated dataset making

2.1.3 评价指标

1）精确度

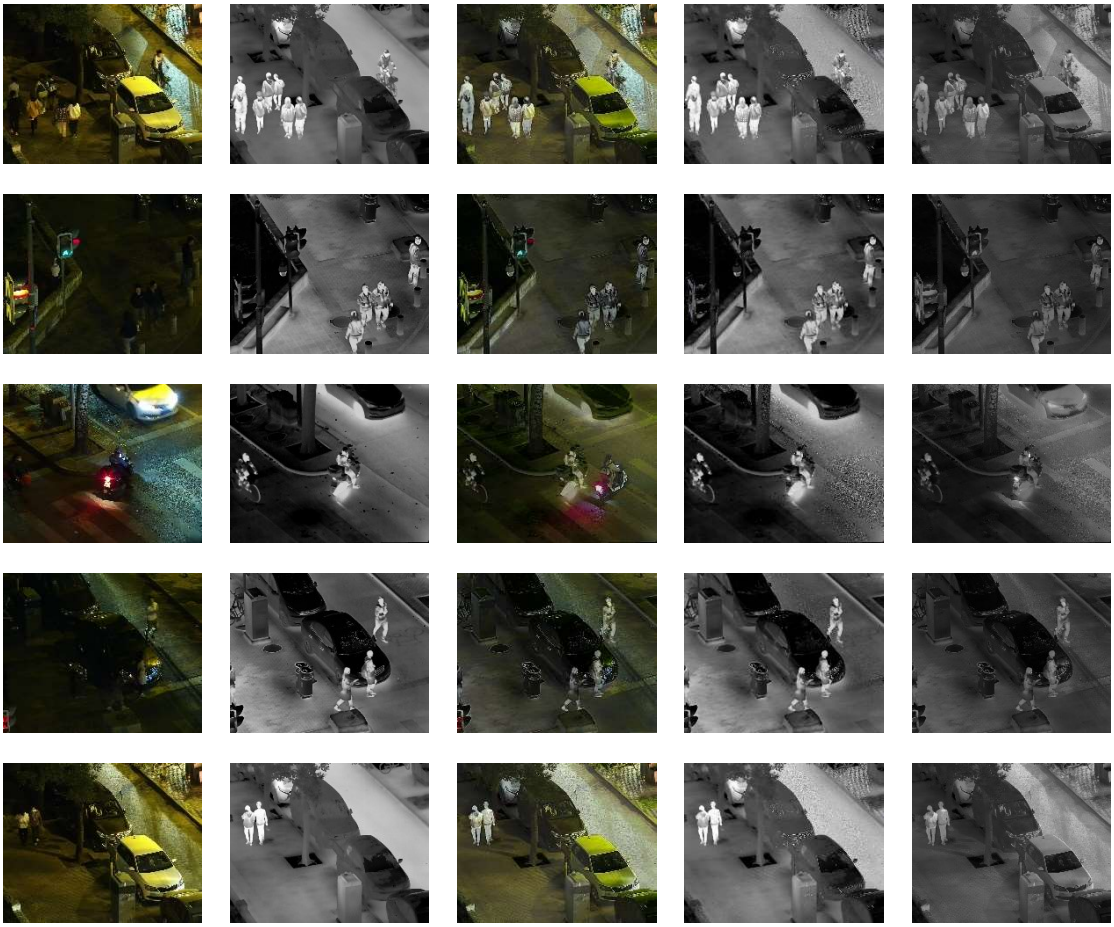
精确度（Precision，P）的结果是真正例（True Positive，TP）与真正例与假正例（False Positive，FP）之和的比值。P 的定义公式如下：

$$P=\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}} \tag{5}$$

2）召回率

召回率（Recall，R）的结果是真正例与真正例和假负例（False Negative, FN）之和的比值。R 的定义公式如下：

$$P=\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}} \tag{6}$$



(a) VI (b) IR (c) IFCNN (d) GTF (e) DenseFuse

图 7 融合算法对比实验

Fig. 7 Fusion algorithm comparison test

图 8 数据集标注场景实例

Fig. 8 Example of dataset annotation scenario

$$\mathrm{mAP}=\frac{\sum_{j=1}^{S}\mathrm{AP}(j)}{S} \tag{7}$$

2.2 消融实验

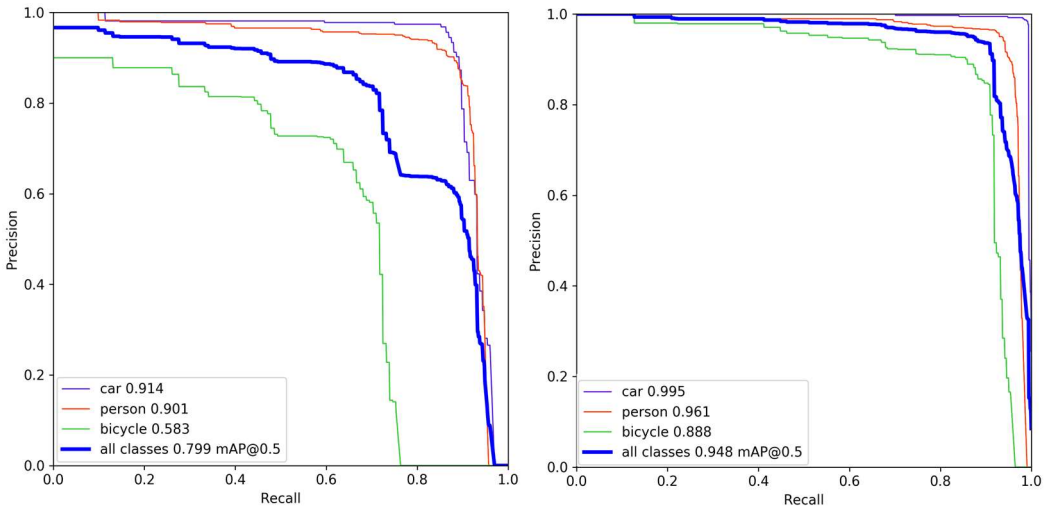
为了验证本文算法的有效性，本文在相同实验条件下对模块的不同加入位置进行精确度评估，输入图像的分辨率为 640×640。

实验结果如表 1 所示。在自制数据集上，我们的算法展现了较好的效果。从表 1 消融实验可知，B 中通过引入 CBAM 注意力机制，提升了 car 类和 bicycle 类的检测精度，模型的 mAP 相比于 A 中原始的 YOLOv7 提高了 1.4%。C 和 D 分别为 YOLOv7 网络中不同位置加入得到结果。其中，C 即 Transformer Encoder（TE）为添加到 backbone 中部实验结果，D 为添加到 backbone 末端与 CBAM 上的实验结果。结果表明，mAP@0.5 较 A 提升 10.5%，能够获得更好的检测性能。E 为本文提出的将 CBAM 和 Transformer Encoder 结合引入的网络，结果表明除了 *R*，其他指标表现都为最优。图 9 为自制数据集在 YOLOv7 网络与我们网络上的 *PR* 曲线图，总的来说，与基线 YOLOv7 相比，能明显看出我们的网络在 *PR* 图中，不同的类别和总体 mAP 曲线都更接近点(1,1)。

表 1 本文网络消融实验

Table 1 Ablation experiment of our network

	Method	*P*	*R*	mAP@0.5	Car（mAP）	Person（mAP）	Bicycle（mAP）
A	YOLOv7	96.30%	89.00%	79.90%	91.40%	90.10%	58.30%
B	YOLOv7＋CBAM	95.80%	92.00%	81.30%	92.60%	88.60%	63.50%
C	YOLOv7＋TE(1)	87.60%	89.00%	67.40%	85.20%	84.90%	32.22%
D	YOLOv7＋TE(2)	95.00%	**99.00%**	90.40%	98.30%	91.90%	81.00%
E	Ours	**96.50%**	98.00%	**94.80%**	**99.50%**	**96.10%**	**88.80%**



(a) YOLOv7 *PR* 曲线 (b) 本文网络 *PR* 曲线

(a) PR curves of YOLOv7 (b) PR curves of ours

图 9 *PR* 曲线对比

Fig. 9 Comparison of *PR* curves

2.3 对比实验

本文使用 FLIR 数据集和自制数据集上进行了测试,还选取目标检测领域 5 种代表算法与本文方法进行对比,分别是 faster-RCNN[17]、SSD[18]、YOLOv5s、tph-YOLOv5[19]与 YOLOv7x[9]。FLIR 数据集上实验结果如表 2 所示,检测示例如图 10 所示。在 FLIR 数据集上虽然我们的网络在自行车的检验精度略低于个别网络,但我们在车辆以及行人的检测上都高于对比算法,同时我们的算法较 YOLOv7 算法在 FLIR 数据集上 mAP@0.5 提升了 8.90%,总体精度都优于本文选取的目标检测对比算法。分析原因,推测是其他网络对于夜间情况下的细节特征提取较弱,难以在黑暗下精准定位识别出目标。关于自行车检测效果略差的原因,我们考虑到是有遮挡以及网络对小目标的学习能力还不太强,因为在我们研究中发现,FLIR 中自行车大多数存在部分遮挡情况,同时在图片中的目标露出部分也较小,并且在数据集中自行车的标注数量也远远少于行人以及车辆的数量,这也是影响网络检验精度的重要原因之一。

为了进一步验证本文网络的有效性,我们也在自制数据集上进行实验,实验结果如表 3。我们的自制数据在本文所改进的目标检测网络上所展现的效果较其他算法都有明显的优越性,相比 faster-RCNN 算法,mAP@0.5 高出 5.08%,比 YOLOv7x 高出 13.50%。对比算法表现略差主要是由于在夜晚条件下,网络难以在黑暗环境下把目标与背景分开,同时他们也对目标的边缘提取能力较差,对昏暗场景中的目标检测比较困难,以及对局部特征的检测还远远不够。同时相较于 SSD,我们的网络的 mAP@0.5 高出 12.03%。因为 SDD 网络属于 one-stage 目标检测算法,本文网络为 two-stage 算法经过了初筛,就会展现出更好的效果,检测结果更准确。在自制数据集上预测结果如图 11 所示,可以明显看出在夜晚情况下我们的检测算法精确度比较高,能够完整、准确地识别出目标物体。

综上所述,本文选取了5种具有代表性的目标检测算法进行实验,验证本文所提方法的优越性,可视化结果对比如图12所示。通过实验说明针对自己所制作的夜间红外-可见光融合数据集和公开的FLIR红外数据集上mAP@0.5均能得到较优的结果。由此可知本文所提算法具有一定的泛化能力,不仅在公开数据集上适用,在自制数据集上也同样适用。

表 2 不同检测算法在 FLIR 数据集上的对比实验

Table 2 Comparative experiment of different detection algorithms on FLIR dataset

	P	*R*	mAP@0.5	Car（mAP）	Person（mAP）	Bicycle（mAP）
Faster-RCNN	34.01%	79.58%	63.99%	70.14%	44.10%	**77.73%**
SSD	82.91%	23.04%	43.93%	50.05%	21.85%	59.88%
YOLOv5s	89.30%	90.00%	68.10%	79.00%	76.60%	48.70%
YOLOv7	95.60%	75.00%	62.60%	74.80%	75.10%	37.80%
tph-YOLOv5	**96.40%**	**91.00%**	68.60%	78.40%	75.00%	52.60%
Ours	**96.40%**	90.00%	**71.50%**	**82.00%**	**80.90%**	51.60%

图 10 本文网络在 FLIR 数据集上预测结果图（左图为 GT，右图为预测结果图）

Fig. 10 Prediction result graph of our network on FLIR dataset(GT(left), Prediction result(right))

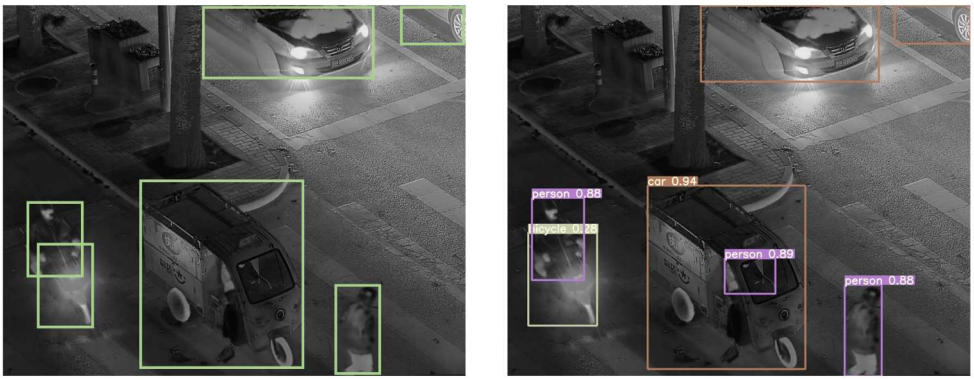


图 11 本文网络在自制数据集上预测结果（左图为 GT，右图为预测结果图）

Fig. 11 Prediction results of our network on the self-made dataset(GT(left), Prediction result(right))

表 3 不同检测算法在自制数据集上的对比实验

Table 3 Comparative experiment of different detection algorithms the on self-made dataset

	P	*R*	mAP@0.5	Car（mAP）	Person（mAP）	Bicycle（mAP）
faster-RCNN	72.00%	91.51%	89.72%	96.97%	87.12%	85.08%
SSD	90.92%	69.91%	82.77%	96.11%	78.08%	74.11%
YOLOv5s	89.30%	90.00%	68.10%	79.00%	76.60%	48.70%
tph-YOLOv5	95.20%	**98.00%**	94.30%	98.70%	95.80%	88.40%
YOLOv7x	**98.70%**	91.00%	81.30%	93.20%	89.80%	60.90%
Ours	96.50%	**98.00%**	**94.80%**	**99.50%**	**96.10%**	**88.80%**

faster-RCNN SSD YOLOv5s

YOLOv7x tph-YOLOv5 Ours

图 12 可视化结果对比

Fig. 12 Comparison of visualization results

3 结论

本文提出了一种结合 CBAM 注意力机制与 Transformer 多注意力机制的红外与可见光图像夜间目标检测方法，在 YOLOv7 网络结构的原有基础上结合了 CBAM 与 Transformer 来优化网络结构，提升网络的目标检测精度与准确度。本文还制作了一个红外-可见光融合目标检测数据集，用于夜间行人、车辆与自行车的检测，为夜间图像融合再应用于目标检测提供了数据支持。我们的算法不论是在公开 FLIR 红外数据集或自制融合数据集上都展现出很好的效果。此外，为了验证本文所提算法的有效性，还将其与 faster-RCNN、SSD、YOLOv5s、tph-YOLOv5、YOLOv7x 进行对比，相比之下我们的网络在精确率、

召回率以及 mAP 上都有较好效果。

下一步的研究将侧重于利用已有数据集训练权重,研究对于红外-可见光融合数据的实时目标检测,进而将融合后检测应用到实际的工作场景之中,比如夜间检测、人体测温、缺陷检测等。

参考文献:

[1] Hafiz A M, Bhat G M. A survey on instance segmentation: state of the art[J]. *International Journal of Multimedia Information Retrieval*, 2020, **9**(3): 171-189.

[2] ZHANG D, Islam M M, LU G. A review on automatic image annotation techniques[J]. *Pattern Recognition*, 2012, **45**(1): 346-362.

[3] Souza É L, Nakamura E F, Pazzi R W. Object tracking for sensor networks: a survey[J]. *ACM Computing Surveys* (CSUR), 2016, **49**(2): 1-31.

[4] YAO H, ZHANG Y, JIAN H, et al. Nighttime pedestrian detection based on fore-background contrast learning[J]. *Knowledge-Based Systems*, 2023, **275**: 110719.

[5] Polukhin A, Gordienko Y, Jervan G, et al. Object detection for rescue operations by high-altitude infrared thermal imaging collected by unmanned aerial vehicles[C]//*Iberian Conference on Pattern Recognition and Image Analysis*. Cham: Springer Nature Switzerland, 2023: 490-504.

[6] MA J, MA Y, LI C. Infrared and visible image fusion methods and applications: a survey[J]. *Information Fusion*, 2019, **45**: 153-178.

[7] 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.

[8] CHEN J, XU X, ZHANG J, et al. Ship target detection algorithm based on decision-level fusion of visible and SAR images[J]. *IEEE Journal on Miniaturization for Air and Space Systems*, 2023, **4**(3): 242-249.

[9] WANG C Y, Bochkovskiy A, LIAO H Y M. YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors[C]// *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023: 7464-7475.

[10] WANG C Y, LIAO H Y M. Yolov4: Optimal speed and accuracy of object detection[J]. arXiv preprint arXiv:2004.10934, 2020.

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

[12] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[J]. *Advances in Neural Information Processing Systems*, 2017, **30**: 5998-6008.

[13] MA J, YU W, LIANG P, et al. FusionGAN: a generative adversarial network for infrared and visible image fusion[J]. *Information Fusion*, 2019, **48**: 11-26.

[14] MA J, CHEN C, LI C, et al. Infrared and visible image fusion via gradient transfer and total variation minimization[J]. *Information Fusion*, 2016, **31**: 100-109.

[15] XU H, MA J, JIAGN J, et al. U2Fusion: a unified unsupervised image fusion network[J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence,* 2020, **44**(1): 502-518.

[16] LI H, WU X J. DenseFuse: a fusion approach to infrared and visible images[J]. *IEEE Transactions on Image Processing*, 2018, **28**(5): 2614-2623.

[17] 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*, 2016, **39**(6): 1137-1149.

[18] LIU W, Anguelov D, Erhan D, et al. Ssd: single shot multibox detector[C]//C*omputer Vision–ECCV* 2016: 14*th European Conference*, Proceedings, Part I 14. Springer International Publishing, 2016: 21-37.

[19] 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.