文章编号：1003－1251(2024)03－0018－08

基于驾驶场景的高效多模态融合检测方法

李东宇，王绪娜，高宏伟

（沈阳理工大学 自动化与电气工程学院，沈阳 110159）

摘 要：目标检测是自动驾驶中重要的组成部分。为解决在弱光条件下单一的可见光图像不能满足实际驾驶场景检测的需求并进一步提高检测精度，提出一种用于红外和可见光融合图像的交通场景检测网络，简称AM-YOLOv5。AM-YOLOv5中改进的Repvgg结构可以提升对融合图像特征的学习能力。此外，在主干网络末端引入自注意力机制并提出一种新的空间金字塔模块(SimSPPFCSPC)充分获取信息；为提升网络推理速度，在颈部网络的前端使用一种全新的卷积(GS卷积)。实验结果表明，AM-YOLOv5在FLIR数据集融合图像上的$mAP_{0.5}$达到了69.35%，与原YOLOv5s相比，在没有牺牲推理速度的前提下，检测精度提升了1.66%。

关 键 词：目标检测；多模态融合；驾驶场景；融合图像

中图分类号：TP391.41 **文献标志码**：A **DOI**：10.3969/j.issn.1003－1251.2024.03.003

Efficient Multi-modal Fusion Detection Method Based on Driving Scenes

LI Dongyu，WANG Xuna，GAO Hongwei

（Shenyang Ligong University，Shenyang 110159，China）

Abstract：Target detection is an important component in autonomous driving. In order to solve the problem that a single visible image cannot meet the demand of actual driving scene detection under low light conditions and to improve the detection accuracy，a traffic scene target detection network for fused infrared and visible images is proposed，which is called AM-YOLOv5 for short. The improved Repvgg structure in AM-YOLOv5 can enhance the ability of learning features of fused images. In addition，a self-attention mechanism is introduced and a new spatial pyramid module(SimSPPFCSPC) is proposed at the end of the backbone network to obtain sufficient information. To improve the inference speed of the network，a new convolution(GS convolution) is used at the front of the neck. Experimental results show that AM-YOLOv5 achieves 69.35% $mAP_{0.5}$ on the fusion image of FLIR dataset，the detection accuracy is improved by 1.66% compared with the original YOLOv5s，without any sacrifice in inference speed.

Key words：object detection；multi-modal fusion；driving scenes；image fusion

智能交通、物联网和人工智能等科学技术的不断革新推动了自动驾驶的迅猛发展。自动驾驶车辆需要对周围环境进行实时感知和识别，以便做出正确的驾驶决策。目标检测算法从车载摄像头拍摄的可见光图像中提取当前驾驶场景的相关信息，包括障碍位置、其他车辆信息和行人信息等[1]。在日常驾驶环境中，通常存在着复杂且实时变化的道路情况，并且在弱光条件下可见光图

收稿日期：2023－03－28

基金项目：辽宁省重点科技创新基地联合开放基金项目(2021－KF－12－05)

作者简介：李东宇(1999—)，男，硕士研究生；高宏伟(1978—)，通信作者，男，教授，博士，研究方向为计算机视觉测量、图像处理与识别等。

像包含的目标信息不足[2],依靠单一视觉传感器的信息不能满足全天候检测的需求,于是针对多模态图像的目标检测任务逐渐受到关注。

可见光图像包含丰富的细节信息但容易受到环境光线影响;红外图像可以突出目标信息,抗干扰能力强,但环境细节信息不足。而可见光图像和红外图像两种模态融合得到的融合图像同时具有两种模态的特点信息,对于目标检测任务有很大的增益,同时相比于可见光 - 激光雷达的多模态融合方式可以保留更多的视觉信息且方便部署。然而可见光红外融合图像具有的特征信息比单一可见光图像和红外图像复杂,对目标检测算法在准确度和推理速度方面提出了更高要求。

Redmon 等[3]提出的 YOLO 系列算法,尤其是 YOLOv5,在目标检测方面具有优异的性能。YOLOv5 的主干网络采用 CSPDarknet53 结构,颈部网络部分采用的是 PANet 结构,检测头为 Yolo head。由于其兼顾了准确性和较快的检测速度,已成为单阶段目标检测算法的代表,尽管 YOLOv5 性能已经很优秀,但在处理复杂的交通环境中可见光红外融合图像信息时,原始网络的特征学习能力不足。

本文选取 YOLOv5 作为基础网络并针对日常自动驾驶任务需求进行改进,简称为 AM-YOLOv5,解决复杂驾驶场景融合图像的目标检测任务。算法在特征提取和学习方面加入了更高效的结构和模块,将多分支学习结构与一种改进的空间金字塔模块(SimSPPFCSPC)引入到主干网络中,实现了多分支的学习能力,在提高特征学习能力的同时不影响推理速度,提高计算效率。同时,对于图片中目标可能被遮挡的问题,本文将多注意力 C3 模块引入到主干网络中,使网络充分获得全局信息和丰富的上下文信息。此外,本文在网络的颈部使用了 GS 卷积,优化网络参数以提高效率,并使网络能够更好地学习多尺度特征。

1　AM-YOLOv5

1.1　整体网络结构

AM-YOLOv5 在网络的主干部分将改进的 Repvgg 模块(Replite)与 YOLOv5 整合在一起,末端引入自注意力机制,加入了更高效的多注意 C3 模块(C3TR)和 SimSPPFCSPC。在此基础上,网络颈部的前端普通卷积被 GS 卷积替代。AM-YOLOv5 算法的结构如图 1 所示。

1.2　高效特征提取主干网络

1.2.1　改进的 Repvgg 模块

主干网络设计对于自动驾驶场景多模态融合图像的目标检测至关重要。多分支结构的训练网络[4]虽然能获得较高的性能提升,但因为内存占用增加导致推理速度下降。普通单路径结构的网络由于没有分支,特征提取能力偏弱,但在操作完成后可以立即释放输入占用的内存,通过牺牲性能可实现较高的推理效率。Ding 等[5]提出的结构重参数化可有效解决该问题,如图 2(a)所示,训练结构由多个 Repvgg 模块组成,每个 Repvgg 模块一般由一个 3 ×3 卷积,一个 1 ×1 卷积和一个恒等分支组成。推理结构如图 2(b)所示,原来的多分支结构经过重参数化操作等价变成了由多个 3 ×3 卷积组成的单路经结构。通常在无重参数化的卷积神经网络算法中,卷积层参数在训练和推理时不会发生变化,卷积层参数相关公式如下。

$$P_c = (K^2 \times C + b) \times N \tag{1}$$

$$W_{out} = \frac{W_{in} + 2p - w}{s} + 1 \tag{2}$$

$$H_{out} = \frac{H_{in} + 2p - h}{s} + 1 \tag{3}$$

$$D_{out} = k \tag{4}$$

式中:P_c 是卷积层所有参数量;K 代表卷积核尺寸;C 代表输入通道数;N 代表过滤器数量;$K^2 \times C$ 代表权重数量;b 代表偏置数量;W_{in} 和 H_{in} 是输入层参数;W_{out}、H_{out} 和 D_{out} 是经过卷积后的输出层参数;k 是卷积核个数;p 是填充大小;s 是步长;w 和 h 是卷积核参数。

重参数化意味着卷积层参数发生相应改变,将卷积层与批量归一化(BN)层融合并对恒等分支转换如下。

$$W_i' = \frac{\gamma_i}{\sigma_i} W_i \tag{5}$$

$$b_i' = -\frac{\mu_i \gamma_i}{\sigma_i} + \beta_i \tag{6}$$

式中:W_i 是转换前的卷积层参数;σ_i 代表 BN 层标卷积准差;μ_i 代表 BN 层累计平均值;γ_i 和 β_i 分别代表 BN 层学习的比例因子和偏差;W_i' 和 b_i' 分别代表融合后卷积层的权重与偏置,i 代表操作范围为所有通道。

具有残差和串联连接的层不应存在恒等分支,会导致不同特征图存在更多的梯度多样性[6],所以本研究提出了改进的 Repvgg 模块,去除

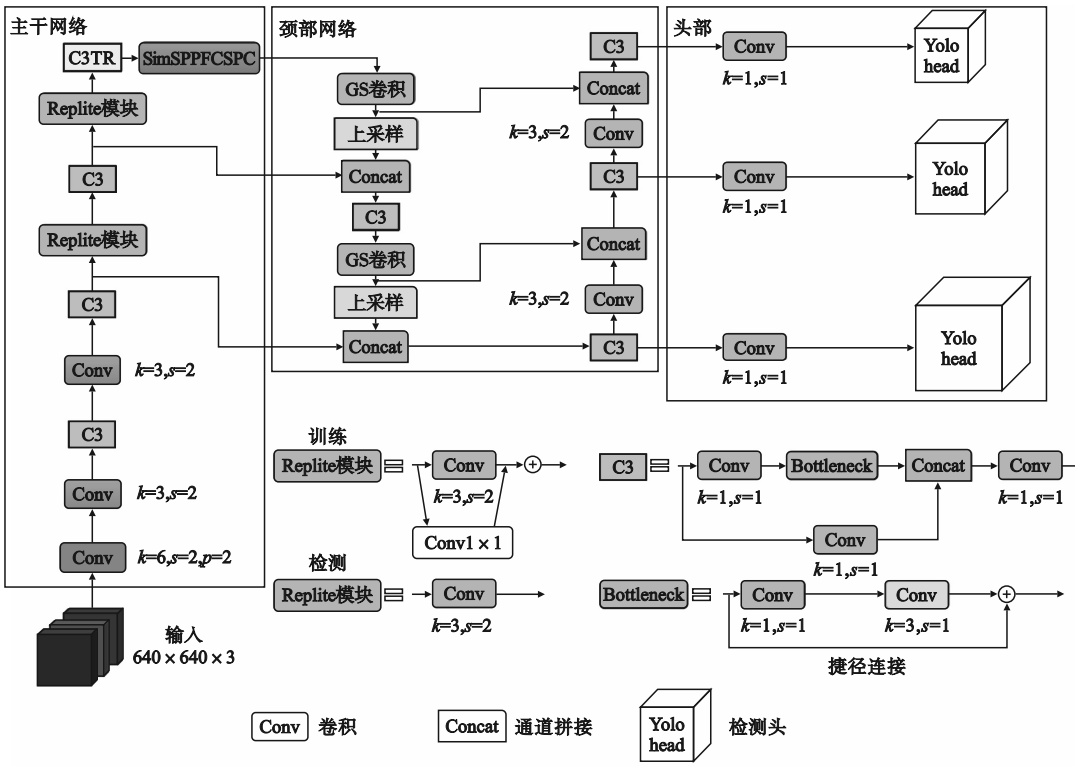


图 1 AM-YOLOv5 结构

Fig. 1 The structure of AM-YOLOv5

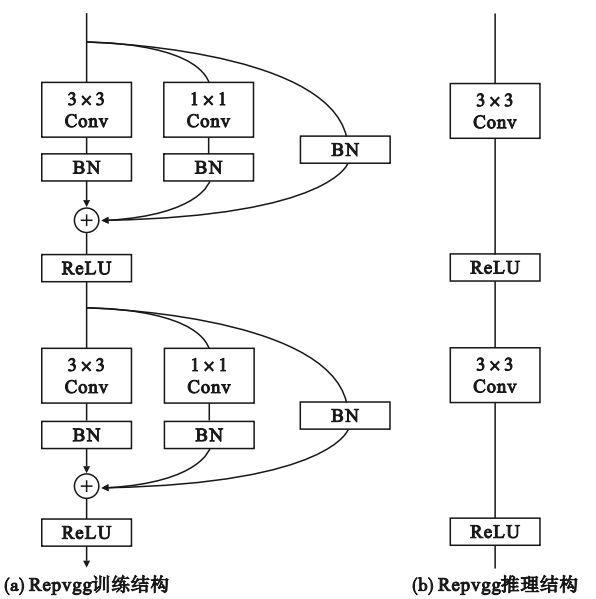


图 2 Repvgg 在不同任务时的结构

Fig. 2 The structure of Repvgg under different tasks

Repvgg 模块的恒等分支,只保留两个卷积分支,并命名为 Replite 模块,用其替换 YOLOv5 主干网络中末端的两个 3×3 卷积块,以实现多分支训练结构与单路径推理结构的转换,不明显增加推理速度的同时提升网络对于细节特征的学习能力。训练阶段如图 3(a)所示,由 3×3 卷积和 1×1 卷积组成的多分支结构可以被看作是多浅层模型的集成,有利于解决梯度消失问题,提高模型的准确度和训练时的算法性能。推理阶段采用与 Repvgg 同样的重参数化方法将原来的 1×1 卷积融入到 3×3 的卷积中,变回单路径结构,如图 3(b)所示。

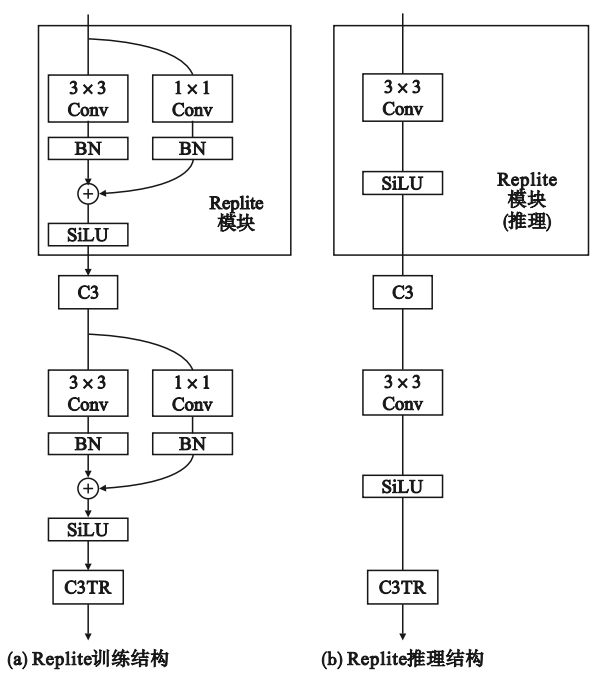


图 3 带有 Replite 模块的主干网络在不同任务时的结构

Fig. 3 The structure of the backbone network with Replite modules under different tasks

1.2.2　多注意力 C3 模块

为使网络能够充分获得全局信息和丰富的上下文信息,提高对融合图像的检测能力,本研究受到 Vision Transformer 的启发,将自注意力机制引入 YOLOv5,以提高在实际场景中检测被遮挡物体的能力[7]。Transformer 模块结构如图 4 所示,由输入嵌入、位置编码、Transformer 编码器组成。每个 Transformer 编码器[8]包含一个多头注意力模块、一个前馈层以及残差连接,多头注意力模块可以更多地关注像素点并获得上下文信息,多头注意力处理过程表示如下。

$$\text{MultiHead}(Q,K,V)=\text{Concat}(head_1,\cdots,head_h)\boldsymbol{W}^O \tag{7}$$

$$head_i=\text{Attention}(\boldsymbol{Q}\boldsymbol{W}_i^Q,\boldsymbol{K}\boldsymbol{W}_i^K,\boldsymbol{V}\boldsymbol{W}_i^V) \tag{8}$$

式中:MultiHead 代表多头注意力操作;Attention 代表注意力操作;Concat 代表拼接操作;$\boldsymbol{W}_i^Q\in\mathbf{R}^{d_{\text{model}}\times d_k}$;$\boldsymbol{W}_i^K\in\mathbf{R}^{d_{\text{model}}\times d_k}$;$\boldsymbol{W}_i^V\in\mathbf{R}^{d_{\text{model}}\times d_v}$($1\leqslant i\leqslant h$);$\mathbf{W}^O$ 代表权重矩阵且 $\boldsymbol{W}^O\in\mathbf{R}^{h\times d_v\times d_{\text{model}}}$;$d_{\text{model}}$代表模型的维度;$d_k$ 代表内容信息的维度;d_v 代表信息的维度;K、V、Q 分别代表内容信息、信息本身和输入信息;h 代表头部或平行的注意力层数。

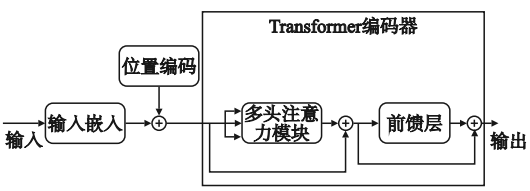


图 4　Transformer 模块结构

Fig. 4　The structure of Transformer module

位置编码由一个全连接层构成,以保留位置信息。输入嵌入通过重塑、展开等操作将二维图像处理成一个序列,输入 Transformer 编码器。最终的嵌入因子通过位置编码与输入嵌入相加得到,再送入 Transformer 编码器。

本文将 C3 模块与 Transformer 结合提出了多注意力 C3 模块(C3TR)且只有网络中主干末端的 C3 模块被替换成 C3TR,即将 C3 模块中 Bottleneck 模块用 Transformer 模块取代,C3TR 结构如图 5 所示。通过在分辨率较低的主干网络末端部署 Transformer 模块,网络可以获得更丰富的上下文信息和更好的全局信息,以提高大型物体的检测精度并影响所有的检测头。此外,部署在主干网络末端不会占用过多内存,并在提高检测效率的同时降低模型复杂度。

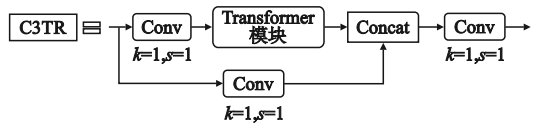


图 5　C3TR 结构

Fig. 5　The strueture of C3TR

1.2.3　改进的空间金字塔模块

本研究提出了一个改进的空间金字塔模块 SimSPPFCSPC,其基于 SPPF 和 SPPCSPC[6]的结构,将两者相结合并针对自动驾驶场景融合图像检测进一步优化以达到更好的特征学习能力,SimSPPFCSPC 的结构如图 6 所示。特征在输入 SimSPPFCSPC 模块后被分到两个分支进行处理,一个分支利用标准卷积对特征进行常规处理,另一个分支利用快速空间金字塔(SPPF)结构对特征进行处理,最后合并两个分支,此结构相比于按序处理减少了计算量并提高了精度。此外,在 SPPF 分支中使用 3×3 的最大池化层,可以在保证学习对象特征的同时节约计算成本,在一定程度上减轻了过拟合的风险。

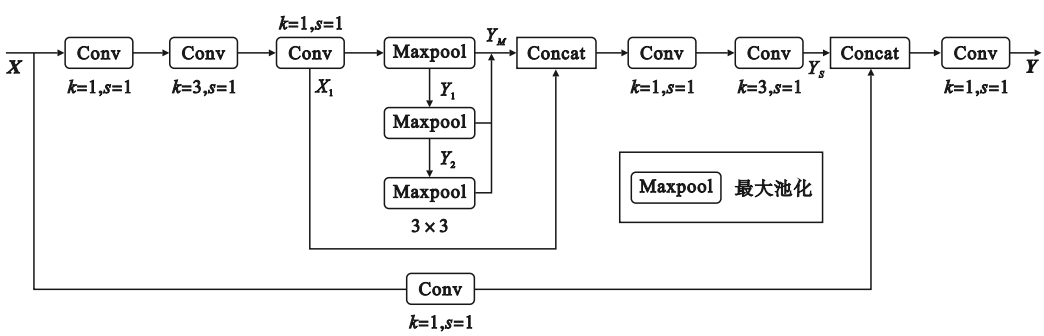


图 6　SimSPPFCSPC 结构

Fig. 6　The structure of SimSPPFCSPC

假设 SimSPPFCSPC 模块的输入为 $\boldsymbol{X}\in\mathbf{R}^{C\times W\times H}$,模块的输出为 $\boldsymbol{Y}\in\mathbf{R}^{C\times W\times H}$,$C$ 是特征图的通道数,W 和 H 分别代表特征图的宽和高,卷积层为 $\text{Conv}_i(i=1,\cdots,7)$,拼接操作层为 Concat,

则 SimSPPFCSPC 的输出 $\boldsymbol{Y}$ 可以表示为

$$\boldsymbol{Y}=\mathrm{Conv}_7(\mathrm{Concat}(\mathrm{Conv}_2(\boldsymbol{X}),Y_{\mathrm{S}})) \quad (9)$$

$$Y_{\mathrm{M}}=(\mathrm{MXP}_1(X_1),\mathrm{MXP}_2(Y_1),\mathrm{MXP}_3(Y_2)) \quad (10)$$

$$Y_{\mathrm{S}}=\mathrm{Conv}_6(\mathrm{Conv}_5(\mathrm{Concat}(Y_{\mathrm{M}},X_1))) \quad (11)$$

$$X_1=\mathrm{Conv}_4(\mathrm{Conv}_3(\mathrm{Conv}_1(\boldsymbol{X}))) \quad (12)$$

式中：Y_{S} 是 SPPF 分支的输出；X_1 是 SPPF 结构的输入；最大池化层 $\mathrm{MXP}_i(i=1,2,3)$ 的内核尺寸为 3×3；Y_1 是第一个最大池化层 MXP_1 处理的输出；Y_2 是第二个最大池化层 MXP_2 处理的输出；Y_{M} 是所有最大池化层分支输出。

1.3 性能平衡的颈部网络

在自动驾驶场景融合图像目标检测算法的实际应用中，平衡模型的精度与速度是十分必要的。为进一步提升网络的速度，本文考虑了在颈部网络部署当前常用的深度可分离卷积（DSC）[9-10]替换标准卷积以降低模型的参数量。深度可分离卷积的结构如图 7 所示。

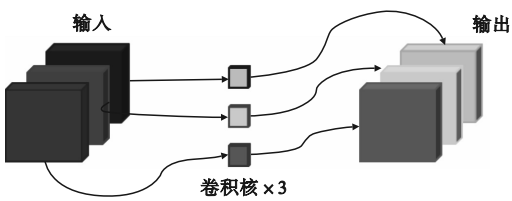


图 7 DSC 结构

Fig. 7 The structure of DSC

图像的通道在进行卷积的过程中被分离，每个通道进行单独的运算，降低了卷积层的参数量。然而，各通道之间的信息缺失导致其特征融合能力变弱，不利于针对融合图像的模型训练和检测。本文在网络中引入了 GS 卷积[11]，GS 卷积的结构如图 8 所示，GS 卷积同时结合了标准卷积与深度可分离卷积中的逐通道卷积（DW 卷积），最后采用打乱操作重分配信息，从而促进了模型对更多正确特征的学习，使特征信息得到充分利用。通过使用 GS 卷积，可以在速度和精度之间取得良好的平衡，减少网络的计算量（FLOPs）和参数量，而不牺牲性能。

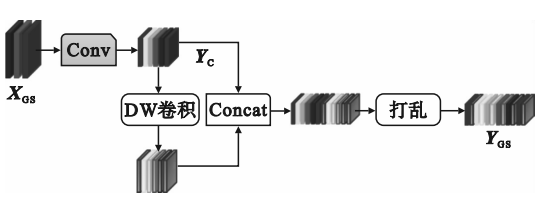


图 8 GS 卷积结构

Fig. 8 GS convolutional structure

假设 GS 卷积的输入为 $\boldsymbol{X}_{\mathrm{GS}}\in\mathbf{R}^{C\times W\times H}$，输出为 $\boldsymbol{Y}_{\mathrm{GS}}\in\mathbf{R}^{C\times W\times H}$，$C$ 是特征图的通道数，W 和 H 分别代表特征图的宽和高，普通卷积层为 Conv，逐通道卷积层为 DWC，拼接操作层为 Concat，打乱操作表示为 SHF，则 GS 卷积的输出 $\boldsymbol{Y}_{\mathrm{GS}}$ 可以表示为

$$\boldsymbol{Y}_{\mathrm{GS}}=\mathrm{SHF}(\mathrm{Concat}(\mathrm{Conv}(\boldsymbol{X}_{\mathrm{GS}}),\mathrm{DWC}(Y_{\mathrm{C}}))) \quad (13)$$

式中：Y_{C} 是普通卷积的输出，Conv 和 DWC 的卷积核参数分别为 $C\times C'\times1\times1$ 和 $C\times C'\times5\times1$，此处 $C'=C/2$。

本文将颈部网络部分中 FPN 结构的 1×1 卷积替换成了 GS 卷积，以努力平衡模型在实际应用中部署时的准确性和速度，而且 GS 卷积的部署仅限于颈部网络部分，因为在卷积神经网络中，空间信息是逐渐传输到通道上的，由于特征图的空间压缩和通道扩展，传输过程中可能会导致语义信息的丢失，而 GS 卷积可以保留存在通道之间的隐藏连接从而保留部分语义信息，但是全阶段部署可能会造成数据流阻碍以及推理时间的增加，为了缓解此问题本文采用了部分部署的策略。

2 实验结果和分析

2.1 数据集

FLIR 数据集是近年发布，用于自动驾驶领域神经网络训练，包含可见光图像和标注过的红外图像的目标检测数据集（拍摄于白天和黑夜情况下），共计 14 000 张。本研究为实现多模态自动驾驶场景，充分验证所提出算法的提升效果，使用对齐的 FLIR 数据集[12]（https://paperswithcode.com/dataset/flir-aligned），选取 4 489 可见光－红外图像对，保证了白天和黑夜道路情况的均匀分布。所有图像对采用基于非下采样剪切波变换（NSST）图像融合算法[13]进行融合，得到 FLIR 数据集融合图像作为实验的数据集，3 476 张作为训练集，1 013 张作为验证集和测试集。其中只使用三种常用的标注类别：汽车，自行车和人，融合图像标注信息与可见光图像和红外图像一致，数据集中每张图像的大小均为 640×512 像素。

2.2 评价指标

本文对提出的 AM-YOLOv5 算法与其他目标检测代表算法进行性能对比实验和消融研究，采用的评价指标包括：平均精度均值（mAP）、计算量（FLOPs）、参数量和推理时间。mAP 是所包含种类的 AP 平均值，数值越大代表模型准确度越高，FLOPs 代表模型的复杂程度，推理时间代表模型在推理单张图片时所用的时间，数值越小代表

推理速度越快。

2.3　实验细节

在 Matlab 软件环境下调用 NSST 算法得到 FLIR 数据集融合图像。

在自动驾驶场景融合图像目标检测实验中,网络的输入为 640×640×3,本文将 FLIR 数据集融合图像输入到训练网络进行端对端的权重模型训练,训练完成后对相应图像进行检测。所有训练建立在预训练过的 YOLOv5s 权重基础之上,初始学习率设置为 0.01,设置动量为 0.937,训练批次大小为 8,设置权重衰减为 0.000 5,训练的迭代次数为 100 次,其余参数设置遵循 YOLOv5 默认的设置。

本研究在 Intel i7-11800 CPU、32 GB 运行内存、NVIDIA RTX3070 GPU 的硬件环境以及 Windows 11、matlab R2020b、Opencv3.4.10 图像处理视觉库和 Pytorch 1.10.1 软件环境下完成。

2.4　与其他检测器性能对比

在 FLIR 数据集融合图像上进行不同检测算法训练,结果对比如表 1 所示。本文算法 $mAP_{0.5}$ 达到了 69.35%,比原 YOLOv5s 提升了约 1.66%,同时超越了一些代表性算法,如 SSD[14]、CenterNet[15] 和 Faster-RCNN[16] 算法的 $mAP_{0.5}$,涨幅都在 14% 以上。同时高于 YOLOXs 的 $mAP_{0.5}$,表明 YOLOXs 对于融合图像效果欠佳。

表 1　本文算法与其他目标检测算法在 FLIR 数据集融合图像上的对比实验结果

Table 1　Comparisons of experimental results of the algorithm in this paper and other target detection algorithms on fused images of FLIR dataset

检测网络	$mAP_{0.5}$/%	FLOPs/10^9 s^{-1}	参数量/MB
SSD	51.86	62.8	26.29
Faster-RCNN	52.84	939.6	28.48
CenterNet	55.47	69.9	32.67
YOLOXs	66.99	21.7	8.05
YOLOv5s	67.69	15.8	7.02
AM-YOLOv5	69.35	21.2	13.41

图 9 展示了 YOLOv5s 和 AM-YOLOv5 在 FLIR 数据集融合图像上的可视化检测结果对比,其中(a)列图像中的黄色框代表 YOLOv5s 未检测出来的对象,绿色框代表 YOLOv5s 错检出的对象。AM-YOLOv5 可以检测出 YOLOv5s 在远距离以及光线不强的情况下没有检测出的行人、自行车和汽车等目标,同时算法可以检测到被遮挡的物体,提高了检测精度,有效减少错检。



(a) YOLOv5s　　(b) AM-YOLOv5

图 9　YOLOv5s 和 AM-YOLOv5 在 FLIR 数据集融合图像上的检测结果对比

Fig. 9　Comparisons of detection results of YOLOv5s and AM-YOLOv5 on fused images of FLIR dataset

2.5　消融研究

为进一步体现本文算法对于多模态自动驾驶场景图像检测性能的提升,依然采用 FLIR 数据集融合图像,以 YOLOv5s 为基础,添加、替换不同组件,对比结果如表 2 所示。

1)添加改进的 Repvgg 模块的影响。网络在 FLIR 数据集上 $mAP_{0.5}$ 由 67.69% 提升到了 68.50%。多分支结构会使整体参数量稍微增大,是由于在推理过程中融合了分支,检测时的推理时间没有增加反而有所下降,为增加精确度而采用 Replite 模块是值得的。

2)添加多注意力 C3 模块的影响。单独添加 C3TR 模块,$mAP_{0.5}$ 比原 YOLOv5s 提高 0.88%,参数量和 FLOPs 增加很少,在 B 组配置的基础上叠加此模块后精度提升了 0.41%。因配置此模块的位置恰当,不会占用过多内存,参数量没有发生明显变化,FLOPs 数减少了 $0.2\times10^9 s^{-1}$,推理时间与未添加此模块时极度接近,检测效率有效提高,证明添加 C3TR 模块有效。

3)添加改进的空间金字塔模块的影响。单独添加 SimSPPFCSPC 模块,$mAP_{0.5}$ 比原始模型提高 0.61%,网络在 F 组配置上叠加此模块之后模型的 $mAP_{0.5}$ 提高至了 69.27%。因为结合了两个分支,参数量和FLOPs数量有了比较明显的上涨。

表 2 在 FLIR 数据集融合图像上消融研究实验结果对比

Table 2 Comparisons of experimental results of ablation studies on fused images of FLIR dataset

	方法	$mAP_{0.5}$/%	$mAP_{0.5:0.95}$/%	FLOPs/10^9 s^{-1}	参数量/MB	推理时间/ms
A	YOLOv5s	67.69	32.31	15.8	7.02	9.3
B	A + Replite	68.50	33.02	16.3	7.06	7.9
C	A + C3TR	68.57	33.18	16.1	7.06	8.8
D	A + SimSPPFCSPC	68.30	32.49	21.5	13.48	9.3
E	A + GSConv	67.97	33.15	16.2	6.98	8.3
F	B + C3TR	68.91	33.15	16.1	7.06	8.7
G	F + SimSPPFCSPC	69.27	32.74	21.3	13.49	9.5
H	G + GSConv	69.35	32.77	21.2	13.41	9.1

此种结构对于按序处理结构已经降低了计算量且提升了模型精度,推理时间没有大幅提升,此模块的增益程度更大。

4)添加 GS 卷积的影响。单独添加 GS 卷积,网络在没有牺牲精度的情况下参数量有所下降,得益于 GS 卷积轻量化的结构以及接近标准卷积效果的特点。在 G 组配置基础上叠加此模块后,模型的 FLOPs 数量降低了 0.1×10^9 s^{-1},参数量也降低了 0.08 MB,在 FLIR 数据集上推理时间有少许降低。因为局部使用 GS 卷积,特征被充分利用,语义信息没有过多丢失,模型准确度有了小幅提升。

上述实验数据表明,本文网络所引入的模块对于检测性能的提升达到了预期效果。

2.6 AM-YOLOv5 检测结果分析

本文将 FLIR 数据集可见光灰度图像和融合图像的训练集分别放入 AM-YOLOv5 网络内进行训练,保存训练好的权重参数,并对测试集图像进行检测,部分结果如图 10 所示。

本文列出了具有代表性的四组实验结果,由第一组结果可以看出,融合图像因为具有红外图像的优点,所有行人在图像中被显著高亮且均被检测,行人被成功识别。此外,融合图像可以克服灯光反射造成的过度曝光现象并且削弱闪光影响,如第二组结果中对向车辆灯光通过地面反射到行人所在区域,融合图像中行人被成功识别,可见光图像则相反。第三组图像是由暗光隧道穿越到日光环境中,在可见光图像上隧道内的车辆可以被识别,但是远处道路几乎没有车辆的迹象,而融合图像中远处强光的目标和环境细节也可见。最后一组在正常日光条件下,所有图像均可以表现出良好的检测性能,但是融合图像上被检测的目标置信度更高。



(a) 可见光图像 (b) 融合图像

图 10 在 FLIR 数据集可见光图像和融合图像上目标检测对比结果

Fig. 10 Comparison results of target detection performed on visible images and fused images of FLIR dataset

通过结合可见光和红外图像的优势,融合图像对目标检测的优势得到了明确的体现,同时本文提出的算法表现出了良好的检测性能。

3 结论

本文提出了一种应用于多模态自动驾驶场景

的高性能目标检测算法 AM-YOLOv5。AM-YOLOv5 的主干网络采用了 Replite 模块,实现了多分支训练结构和单路推理结构的转换,在提高精度的同时没有影响速度;C3TR 模块和 SimSPPFCSPC 模块的加入提升了网络的计算效率,进一步提高了精度;颈部网络引用的全新卷积较好地平衡了网络的精度和速度。AM-YOLOv5 在 FLIR 数据集融合图像上的检测性能相比于原始的 YOLOv5s,$mAP_{0.5}$ 提升了1.66%,整体参数量有些许增加,但是没有牺牲推理速度,基本符合预期结果。

参考文献(References):

[1] 王麒. 基于深度学习的自动驾驶感知算法[D]. 杭州:浙江大学,2022.

[2] 祝文斌,苑晶,朱书豪,等. 低光照场景下基于序列增强的移动机器人人体检测与姿态识别[J]. 机器人,2022,44(3):299-309.
ZHU W B,YUAN J,ZHU S H,et al. Sequence-enhancement-based human detection and posture recognition of mobile robots in low illumination scenes[J]. Robot,2022,44(3):299-309. (in Chinese)

[3] REDMON J,DIVVALA S,GIRSHICK R,et al. You only look once:unified,real-time object detection[C]//2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Las Vegas,NV,USA:IEEE,2016:779-788.

[4] NORKOBIL SAYDIRASULOVICH S,ABDUSALOMOV A,JAMIL M K,et al. A YOLOv6-based impro-ved fire detection approach for smart city environme-nts[J]. Sensors,2023,23(6):3161.

[5] DING X H,ZHANG X Y,MA N N,et al. RepVGG:making VGG-style ConvNets great again[C]//2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Nashville,TN,USA:IEEE,2021:13733-13742.

[6] JIANG K L,XIE T Y,YAN R,et al. An attention mechanism-improved YOLOv7 object detection algor-ithm for hemp duck count estimation[J]. Agriculture,2022,12(10):1659.

[7] 于楠晶,范晓飚,邓天民,等. 基于多头自注意力的复杂背景船舶检测算法[J]. 浙江大学学报(工学版),2022,56(12):2392-2402.
YU N J,FAN X B,DENG T M,et al. Ship detection algorithm in complex backgrounds via multi-head self-attention[J]. Journal of Zhejiang University(Engineering Science),2022,56(12):2392-2402. (in Chinese)

[8] VASWANI A,SHAZEER N,PARMAR N,et al. Attention is all you need[C]//Proceedings of the 31st International Conference on Neural Information Processing Systems. Long Beach,California,USA:ACM,2017:6000-6010.

[9] CHOLLET F. Xception:deep learning with depthwise separable convolutions[C]//2017 IEEE Conference on Computer Vision and Pattern Recognition(CVPR). Honolulu,HI,USA:IEEE,2017:1800-1807.

[10] 杨小冈,高凡,卢瑞涛,等. 基于改进 YOLOv5 的轻量化航空目标检测方法[J]. 信息与控制,2022,51(3):361-368.
YANG X G,GAO F,LU R T,et al. Lightweight aerial object detection method based on improved YOLOv5[J]. Information and Control,2022,51(3):361-368. (in Chinese)

[11] HU J E,WANG Z B,CHANG M J,et al. PSG-Yolov5:a paradigm for traffic sign detection and recognition algorithm based on deep learning[J]. Symmetry,2022,14(11):2262.

[12] ZHANG H,FROMONT E,LEFEVRE S,et al. Mul-tispectral fusion for object detection with cyclic fuse-and-refine blocks[C]//2020 IEEE International Confer-ence on Image Processing(ICIP). Abu Dhabi,United Arab Emirates:IEEE,2020:276-280.

[13] 张全. 基于 NSST 的红外与可见光图像融合算法研究[D]. 西安:西安电子科技大学,2020.

[14] LIU W,ANGUELOV D,ERHAN D,et al. SSD:single shot MultiBox detector[C]//European Conference on Computer Vision. Cham:Springer,2016:21-37.

[15] DUAN K W,BAI S,XIE L X,et al. CenterNet:keypoint triplets for object detection[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. Seoul,Korea(South):IEEE,2020:6568-6577.

[16] REN S Q,HE K M,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.

(责任编辑:和晓军)