计算机科学
COMPUTER SCIENCE

EvR-DETR:融合事件与RGB图像的轻量级端到端目标检测

周秉泉, 蒋杰, 陈江民, 詹礼新

引用本文

周秉泉, 蒋杰, 陈江民, 詹礼新. EvR-DETR:融合事件与RGB图像的轻量级端到端目标检测[J]. 计算机科学, 2026, 53(1): 153-162.

ZHOU Bingquan, JIANG Jie, CHEN Jiangmin, ZHAN Lixin. EvR-DETR:Event-RGB Fusion for Lightweight End-to-End Object Detection [J]. Computer Science, 2026, 53(1): 153-162.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.250300021

EvR-DETR:融合事件与 RGB 图像的轻量级端到端目标检测

周秉泉　蒋　杰　陈江民　詹礼新
国防科技大学系统工程学院　长沙 410073
(bingquanzhou@nudt.edn.cn)

摘　要　基于神经脉冲信号的事件摄像机可以提供光线变化的信息,以弥补传统 RGB 相机目标检测在恶劣环境性能下降的缺陷。然而,传统融合事件相机的现有方法存在模型参数大和非端到端训练方法的问题,限制了模态融合的有效性。因此,提出了一种事件与 RGB 信息融合的轻量级端到端对象检测框架,基于两种模态各级尺度特征进行不同细粒度的信息融合,同时基于重参数化卷积实现轻量级的融合模块并进行端到端的训练,从而提升模型对于两种模态互补信息的提取能力,以克服自动驾驶中具有挑战性的不利环境。所提出的模型在大规模数据集 PKU-SOD 上进行了测试,该数据集提供了低光、高速运动模糊与正常光照环境下车辆行驶的视觉数据。实验结果表明,与此前的多模态目标检测框架相比,所提方法在模型参数量上大幅下降,并提升了目标检测的准确率与推理速度,表现出优于现有方法的性能。

关键词:目标检测;仿生相机;自动驾驶;深度学习;端到端目标检测;事件相目标机检测;轻量化目标检测
中图分类号　TP391

EvR-DETR:Event-RGB Fusion for Lightweight End-to-End Object Detection

ZHOU Bingquan,JIANG Jie,CHEN Jiangmin and ZHAN Lixin
College of System Engineering,National University of Defense Technology,Changsha 410073,China

Abstract　Event cameras based on neuromorphic spike signals can provide information about illumination changes,compensating for the performance degradation of traditional RGB cameras in object detection under adverse environments. However,existing methods fusing event cameras with conventional cameras suffer from large model parameters and non-end-to-end training approaches,which restrict the effectiveness of modality fusion. To address this,this paper proposes a lightweight end-to-end object detection framework that integrates event and RGB information through multi-granularity fusion of multi-scale features across different network levels. By implementing lightweight fusion modules with reparameterized convolutions and enabling end-to-end training,the proposed framework enhances the model's capability to extract complementary information from both modalities,overcoming challenging conditions in autonomous driving. Evaluated on the large-scale PKU-SOD dataset containing vehicular visual data under low-light,high-speed motion blur,and normal illumination scenarios,the proposed method significantly reduces model parameters compared to state-of-the-art multimodal approaches while improving detection accuracy and inference speed,demonstrating superior performance over existing methods.

Keywords　Object detection,Neuromorphic camera,Autonomous driving,Deep learning,End-to-end object detection,Event-basedobject detection,Light-weight object detection

1　引言

目标检测[1]能够对图像中的物体进行类别的识别并标记物体所处的位置。近年来,基于深度学习的目标检测在诸多领域的应用得到了长足发展。在自动驾驶等无人系统的应用中,目标检测通常提供车辆周边的物体类别信息与位置标记信息,起着至关重要的作用。然而,传统相机在特定曝光时间内捕获光子以生成图像的局限性,物体检测经常会遇到由极端照明和高速运动引起的模糊性用。这阻碍了自动驾驶应用中对象检测的性能[2-3]。而事件像机的出现为这些问题提供了解决方案[4-5]。

事件像机可以在每个像素点位置对光强的变化进行相应,并输出此时刻光强变化的信息,如图 1 所示,每一个点是该像素位置产生光强变化所输出的信号,红色表示光强增加,蓝色则表示光强减弱。这使得事件相机能够对运动物体感知敏感,尤其是复杂环境中物体的边缘信息。基于事件相机的目标检测(Event-based Object Detection)[6]算法也因此提出。不同于此前基于 RGB 静态图像的算法,基于事件相机的目标检测算法通过对时空点数据进行特征提取,从这些光照变化点数据中获取物体的类别信息与位置信息。

到稿日期:2025-03-04　返修日期:2025-05-08
通信作者:蒋杰(jiejiang@nudt.edu.cn)

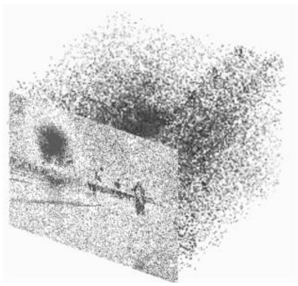

图 1 事件数据可视化(电子版为彩图)

Fig. 1 Visualization of event data

基于事件相机的目标检测的关键挑战在于两点。一是事件相机本身的光强敏感性易引入无关的噪点信息,如图 2 中低光与高速模糊场景,在对运动物体进行敏感感知之外,背景的细微光强变化也被事件相机捕捉,带来诸多的噪点。二是事件数据缺乏纹理信息,因此在静止或缓慢运动这样光强变化微弱的场景下,事件相机无法很好地捕捉信息。如图 2 的静止场景所示,摄像机与前车均处于静止状态,未发生光强的变化,使得事件相机仅产生了少量事件数据,算法在此种场景下无法对目标进行有效检测。因此,如何利用 RGB 图像的纹理信息与事件数据的光强变化信息,将二者进行有效融合以获取两种模态潜在的互补特征,成为指导算法在不同环境下均能进行有效目标检测的关键。

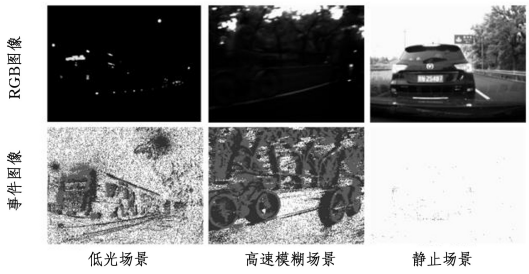

图 2 各场景下 RGB 图像与事件图像的比对

Fig. 2 Comparison of RGB andevent image in different scenarios

如图 3(a)所示,早期利用 RGB 图像与事件数据进行多模态目标检测的方法,利用卷积神经网络(Convolutional Neural Network,CNN)方法进行特征提取,通过对两种模态特征的置信度的比较,选择应用的特征进行边界框与分类的解码[7-8],这一方法在最终的特征选区中通常涉及较多人工设计的选择算法,且不存在对两种模态特征的融合操作。如图 3(b)所示,去除人工设计的选择算子,基于拼接的方法[9]将两种模态特征进行拼接输送到特征金字塔网络(Feature Pyramid Network,FPN),利用特征金字塔来分析两种模态之间的相关特征,使得网络的一体性更强。然而,这种方法缺乏对长期的全局上下文的依赖分析,使得两种模态之间的全局特征聚合与分析受到限制。近年来,视觉 Transformer[10]的出现使图像处理具备全局信息特征提取的能力,一些研究也证明了 Transformer 在 RGB 事件多模态目标检测中应用的可行性。如图 3(c)所示,一些方法通过卷积神经网络进行特征提取,对每一层级的特征采用包含 Transformer 结构的融合模块进行融合[11-12],这类方法的特点是在特征提取的中间阶段进行多模态特征融合。

而另外一类方法则是在特征提取的末期阶段进行特征融合,如图 3(d)所示,这类方法通过基于自注意力机制的模块对两种模态的上下文长依赖信息进行全局聚合,再通过基于交叉注意力的融合模块对提取的最高层级的全局特征进行融合[13]。然而,不管是中间阶段融合还是末期融合,它们的特征融合均是同层级特征融合,忽略了不同层级、不同尺度的特征之间的融合交互。具体而言,中间阶段融合仅对特征提取的相同层级的特征进行融合,而末期融合仅对最高层级的特征进行融合,且二者均采用了大量的 Transfor-mer 结构。其核心自注意力机制是计算所有输入位置之间的相关性,对于序列长度为 n 的序列,计算复杂度为 $O(n^2)$,且 Transformer 使用的多头注意力机制中每个头都有独立的 $\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V}$ 矩阵,导致这些模型均有着非常大的参数量,计算复杂度高。例如 SODFormer[11]的参数量超过 8000 万,且计算复杂度高、梯度回归慢,训练时仅能对每个基于 Transformer 的模块进行单独训练后,再对融合模块进行训练。这种非端到端的训练方式限制了模型对两种模态的提取能力,同时大参数量也限制了模型在自动驾驶中的边缘部署。

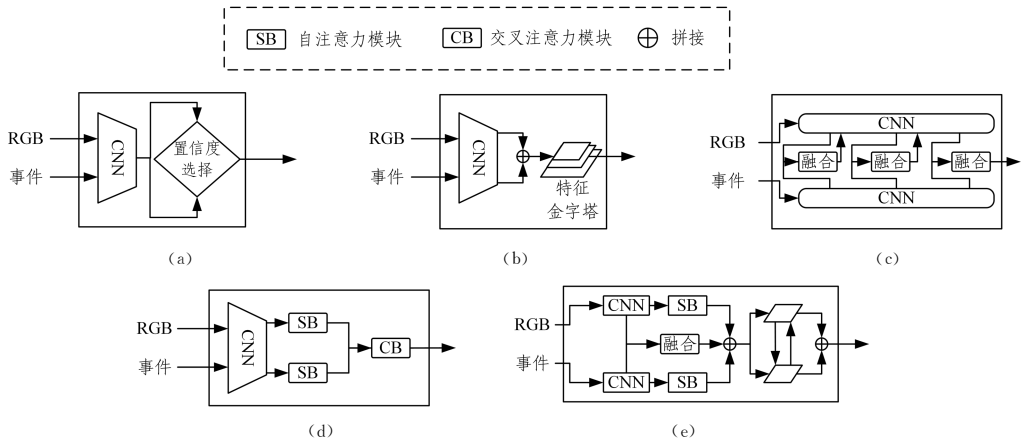

图 3 不同融合 RGB 图像与事件数据进行目标检测的算法对比

Fig. 3 Comparison of algorithms for fusing RGB images and event data for object detection

RT-DETR[14]是一种基于 Transfomrer 的目标检测框架。通过基于单模态的多尺度融合与简化 DN-DETR 解码器结构,RT-DETR 在实现计算效率提升的同时保持了模型的检测性能。受 RT-DETR 的启发,本文通过探索 RGB 图像与

事件数据两种模态在不同尺度层级的特征相关性与特征互补性，来充分利用两种模态数据的潜在信息，并减少计算冗余。具体而言，本文设计了一种具有特征层次化聚合和多模态跨尺度融合的轻量级端到端目标检测框架 EvR-DETR（Event-RGB DEtectionTRansformer），将 RT-DETR 拓展为一个多模态的目标检测 Transformer 结构，并结合中期融合与末期融合的优势，对同层级进行不同颗粒度的融合，如图 3(e)所示，通过基于 Transoformer 的设计能够弥补卷积神经网络无法进行有效全局感受野分析的问题，并能够对多模态的潜在信息进行上下文分析，提取高密度信息的特征。为了实现这一点，本文构造了一个多模态跨尺度融合模块（Multi-modal Cross-scale Fusion Module，MCFM），其目的是通过对两种模态不同尺度层级的特征进行交互式融合，来充分利用这些特征相关、互补的信息，从而提高特征信息的利用程度。同时，这一模块尽可能减少了自注意力机制的使用，以降低模型参数量与计算复杂度。此外，本文还提出了一个特征尺度分解空间编码器（Feature Scale-decompose Spatial Encoder，FSSE），通过提取卷积神经网络不同层级的特征，对包含较少信息的特征进行粗略融合，对高密度信息的特征进行全局上下文分析，从而减少了 Transformer 的冗余使用，同时确保特征信息能被充分发掘。本文的主要贡献如下：

1）开发了一种轻量级的融合 RGB 图像与事件数据的端到端 Transformer 目标检测框架，以克服自动驾驶环境感知在不利环境下的检测失效问题。

2）提出了特征尺度分解空间编码器（FSSE）和双阶段的多模态跨尺度交互融合模块（MCFM），以利用两种异构的模态在不同尺度层次特征之间的潜在空间关系，提高模型的运行效率与检测准确率。

3）在大规模多模态数据集 PKU-SOD 上进行了训练，并针对不同环境场景进行广泛测试实验。与之前的 SOTA 模型 SODFormer 相比，本文的检测方法将参数数量减少近 50%，准确率提高 2.5%，推理速度提高 20%以上。

2 方法描述

2.1 动机与概述

考虑到传统基于 RGB 图像的目标检测在低光、高速运动产生模糊的情况下会发生检测失灵的情形，构建融合 RGB 图像与事件数据的方法成为一个重要的解决思路。而如何充分利用这两种异构数据潜在的相关信息与互补信息则是多模态融合的关键。以往的基于卷积神经网络的方法[7,9]进行了初步的尝试与探索，这些方法取得了一定的成效，但是整体直接沿用传统 RGB 目标检测的结构，对特征采取直接聚合的方式，缺乏细致的融合与对不同模态特征的全局信息关注。目前一些工作[11-13]基于 Transformer 建立了融合架构，但也存在一定的局限性：1）这些方法缺乏对两种模态不同层级之间特征的交互融合，使得不同颗粒度信息不能彼此融合；2）对各层级特征使用自注意力机制造成了冗余计算，计算效率不高；3）端到端训练策略的缺失与人工选择算法的参与限制了模型对目标检测信息的关注。

为了解决上述问题，本文提出端到端融合目标检测模型 EvR-DETR，其基于端到端检测 Transformer（End-to-End DEtectionTRansformer）[15]）设计。该框架能够利用基于 Transformer 的全局信息分析能力，减少人工选择算法的参与，实现端到端的学习训练与检测。随后，本文构建了一个多模态跨尺度融合模块，通过对两种不同模态不同层次的信息进行交互式融合的策略，对不同颗粒度的相关互补信息进行有效聚合。同时，对不同尺度特征进行分解处理的策略被引入空间编码器中。如图 4 所示，这一框架能够同时处理视频流和与视频流相对应的事件流，这两种模态的数据均由 DAVIS[16]摄像机生成。通过事件数据构建事件图表示，RGB 图像和事件图经过两个参数共享的骨干网络（如 ResNet50[17]），以得到它们相应模态的特征。这些特征随后被特征尺度分解空间编码器进行层级分解并进行粗融合，以获得初步的融合特征。此后，这些初步融合的特征通过具备两个阶段的多模态跨尺度融合模块进行交互式融合，使得不同尺度的特征中的不同颗粒度信息能够最大限度地进行相关性与互补性分析，以获得最终的融合特征。这些最终的融合特征由一个 DETR 风格的解码器（Decoder）进行解码，得到目标物体的分类标签与物体的位置信息。

2.2 算法结构

为了充分利用这两种异构数据潜在的相关信息与互补信息，从而缓解目标检测在低光和高速运动模糊的情形下检测失灵的问题，本文提出了一个具有特征层次化聚合和多模态跨尺度融合的轻量级端到端目标检测框架（Event-RGB Detection Transformer，EvR-DETR），总体框架如图 4 所示。

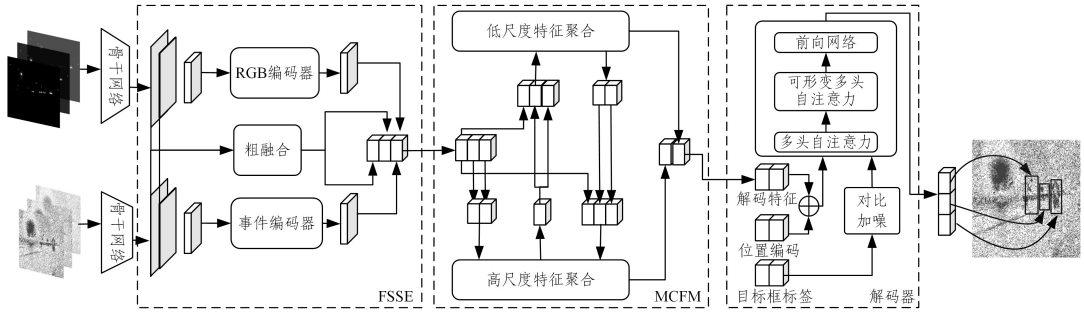

图 4 具有特征层次化聚合和多模态跨尺度融合的轻量级端到端目标检测框架

Fig. 4 Lightweight end-to-end object detection framework with feature hierarchical aggregation and multi-modal cross-scale fusion

该网络框架是一个双流的多模态检测 Transformer 结构，输入为 RGB 图像和事件数据。该算法的结构由特征提

取模块、特征尺度分解空间编码器、多模态跨尺度交互融合模块和基于 Transformer 检测解码器组成。具体而言，基于 Transformer 的空间编码器对最高层级的特征进行全局感受野获取，不同模态的低层级特征进行预聚合处理，减少 Transformer 中的自注意力机制对冗杂信息的处理。双阶段的多模态跨尺度特征融合模块对不同模态、不同层级的特征进行交互式融合，增强了不同颗粒度信息之间的关联与互补以及多模态特征的表示。

1)事件预处理。与大多数融合 RGB 图像与事件数据的目标检测方法相似，本文对时空域的点数据进行预处理，使其能够被卷积神经网络更好地提取特征。具体而言，事件流表示为集合$\mathcal{E}=\{e_k\}_{k=1}^{N}$，其中每个事件定义为 $e_k=(x_k,y_k,t_k,p_k)$，满足 $x_k,y_k\in\mathbb{Z}^2$（离散像素坐标）、$t_k\in\mathbb{R}^+$（时间戳）、$p_k\in\{+1,-1\}$（极性，表示亮度增减）。首先，通过选定时间窗口$\Delta T=[t_{\text{start}},t_{\text{end}})$，提取子事件流$\mathcal{E}_{\Delta T}=\{e_k \mid t_k\in\Delta T\}$，并依据事件的极性进行统计生成事件图像(Event Image)，具体表示如下：

$$\boldsymbol{I}_{\text{Event}}(i,j,c)=\begin{cases}\boldsymbol{C}_{\text{on}}(c), & \exists\, e_k\in\mathcal{E}_{\text{on}},y_k=i\wedge x_k=j\\ \boldsymbol{C}_{\text{off}}(c), & \exists\, e_k\in\mathcal{E}_{\text{off}},y_k=i\wedge x_k=j\\ I^{(0)}_{\text{Event}}, & \text{otherwise}\end{cases}\tag{1}$$

$$\boldsymbol{C}_{\text{on}}(c)=c\cdot[0,0,255]\tag{2}$$

$$\boldsymbol{C}_{\text{off}}(c)=c\cdot[255,0,0]\tag{3}$$

其中，$\boldsymbol{I}^{(0)}_{\text{Event}}$表示初始全白图像张量；$\mathcal{E}_{\text{on}}$，$\mathcal{E}_{\text{off}}$表示正/负事件子集；$\boldsymbol{C}_{\text{on}}$，$\boldsymbol{C}_{\text{off}}$表示与位置坐标无关的颜色编码向量；$c$ 表示事件中的极性大小；i,j 表示张量索引，对应图像坐标系 y 和 x 轴。

2)特征层次分解空间编码器。参照检测 Transformer 的方法，本文将 RGB 图像 $\boldsymbol{I}_{\text{RGB}}\in\mathbb{R}^{H_{\text{R}}\times W_{\text{R}}\times 3}$ 和 $\boldsymbol{I}_{\text{Event}}\in\mathbb{R}^{H_{\text{E}}\times W_{\text{E}}\times 3}$ 作为输入，利用训练好的 ResNet-50 网络提取 RGB 图像特征集合$\mathcal{F}_{\text{RGB}}=\{\boldsymbol{F}^h_{\text{R}},\boldsymbol{F}^m_{\text{R}},\boldsymbol{F}^l_{\text{R}}\}$与事件图像特征集合$\mathcal{F}_{\text{Event}}=\{\boldsymbol{F}^h_{\text{E}},\boldsymbol{F}^m_{\text{E}},\boldsymbol{F}^l_{\text{E}}\}$。为了充分利用各个层级的特征并减少计算冗余，对两种模态的高级特征进行了全局特征分析，对中级特征与低级特征进行了聚合与局部分析。FSSE 通过对不同层级不同模态的特征进行预融合，并利用不同的上下文分析方法对信息进行进一步提取，能够降低空间编码器的计算复杂度，并提取不同颗粒度信息以供后续的融合模块进行操作。FSSE 的具体细节将在 2.3 节中说明。

3)跨尺度融合模块。近年来，基于融合 RGB 图像与事件图像的目标检测取得了长足进步，其中融合模块是提升两种异构模态进行信息互补的关键。以往基于特征金字塔的单向融合方法限制了信息的交互，而基于交叉注意力的方法则带来了计算复杂度的大幅升高与参数量的上升。为了解决上述问题，对两种模态的不同尺度层级的特征进行有效融合，本文基于多尺度特征融合交互设计了多模态融合模块。该模块由两个融合阶段构成，每个融合阶段由两个融合卷积模块构成。首先，将 FSSE 生成的多模态多尺度特征$\{\hat{\boldsymbol{F}}^h_{R+E},\hat{\boldsymbol{F}}^m_{R+E},\hat{\boldsymbol{F}}^l_{R+E}\}$输入模块中，通过高层级聚合阶段对中高层级特征进行融合，再与低层级特征在低层级聚合阶段进行融合，随后将输出的特征进行交互式融合，最终得到细致融合后的融合特征。这种通过两个阶段不断交互融合的方法能够确保各个尺度的特征以及各个细粒度的信息能够得到充分的交互，使得潜在的目标检测相关信息、两种模态的互补信息能够得到更加深入的分析。MCFM 的详细信息将在第 2.4 节中说明。

4)检测 Transformer 解码器。为了对多模态的检测框架进行端到端训练，本文采取 DETR[14] 的二分图匹配的设计。这一设计虽然减少了人工干预并降低了模型复杂性，但造成了模型的收敛较慢的问题。多模态数据进一步加剧了模型的收敛问题，为了能够加速收敛过程，本文在解码器设计上遵循 DN-DETR[18] 的设计，通过在训练过程中向模型输入带有噪声的标注信息（如噪声边界框、噪声类别标签等），让模型学习如何从这些噪声输入中恢复出真实的标注信息，帮助模型更快地学习到有效的特征表示，从而加速收敛。

2.3 特征尺度分解空间编码器

Transformer 结构被应用于目标检测任务中，通过多头注意力机制能够有效捕捉特征，为目标检测提供全局信息，同时，基于 Transformer 的 RGB-事件目标检测也展现出注意力机制在多模态信息捕获上的优势。然而，当前的 RGB 图像-事件多模态特征提取方法[10]存在以下局限性：1)由于新模态的引入，对新模态数据进行特征编码会使模型的参数量激增，且注意力机制的运用会导致模型的计算复杂度升高，最终导致基于 Transformer 的多模态检测模型无法有效收敛，只能够分模块训练而非端到端训练，极大延长了模型训练时间；2)由于进行特征提取的骨干网络是基于 RGB 图像数据集进行训练的，对事件图像的特征不能进行有效感知，非端到端的训练方法使得骨干网络的微调不能有效地结合 RGB 图像与事件图像的共同损失进行梯度回传，使得骨干网络的特征提取能力受限；3)基于经典 DETR 的方法仅使用最高级特征，而忽略了低级特征的使用，导致低尺度层级中的信息未能被充分利用。针对上述局限，本文提出了针对多模态的特征尺度分解空间编码器，具体结构如图 5 所示。

1)高级特征聚合。高层级特征通常包含更抽象的语义信息，其颗粒度较粗，适合通过 Transformer 的多头自注意力机制(Multi-Head Self-Attention，MHSA)进行处理。MHSA 能够捕捉全局依赖关系，提取模态内和模态间的长程关联，从而增强高层特征的语义表达能力。对高层特征进行独立处理后再拼接，能够保留各自模态的独特信息，同时通过拼接实现模态间的信息互补。具体而言，对于 RGB 图像高级特征$\boldsymbol{C}^f_1\in\mathbb{R}^{\frac{H}{8}\times\frac{W}{8}\times D^f_1}$与事件图像高级特征$\boldsymbol{C}^e_1\in\mathbb{R}^{\frac{H}{8}\times\frac{W}{8}\times D^e_1}$，使用多头自注意力机制进行全局特征分析与初步拼接融合：

$$\hat{\boldsymbol{S}}^f_1=MHSA(\boldsymbol{C}^f_1+\mathcal{P}^f)\tag{4}$$

$$\hat{\boldsymbol{S}}^e_1=MHSA(\boldsymbol{C}^e_1+\mathcal{P}^e)\tag{5}$$

$$\hat{\boldsymbol{S}}^{e+f}_1=Concat(\hat{\boldsymbol{S}}^f_1,\hat{\boldsymbol{S}}^e_1)\in\mathbb{R}^{\frac{H}{8}\times\frac{W}{8}\times 2D_h}\tag{6}$$

其中，$Concat(\cdot)$为沿通道维度拼接操作，$MHSA(\cdot)$为多头自注意力机制，$\mathcal{P}\in\mathbb{R}^{\frac{H}{8}\times\frac{W}{8}\times D_1}$为固定正弦编码。

2)中低级特征局部粗融合。对于中低层级特征，其信息颗粒度较细，主要包含两种模态的局部细节和结构信息。通过拼接与卷积操作处理中低层级特征，能够有效融合两种模态的局部特征，卷积操作进一步提取局部空间模式，增强特征的判别能力。这种设计既保留了低层级特征的细节信息，又

通过卷积实现了模态间的局部交互，提升了特征的鲁棒性和表达能力。具体地，对中级特征与低级特征进行拼接与卷积前馈操作：

$$S_2^{e+f}=Concat(C_2^f,C_2^e)\in\mathbb{R}^{\frac{H}{4}\times\frac{W}{4}\times 2D_2} \tag{7}$$

$$\hat{S}_2^{e+f}=Conv(FFN(S_2^{e+f})) \tag{8}$$

$$S_3^{e+f}=Concat(C_3^f,C_3^e)\in\mathbb{R}^{\frac{H}{4}\times\frac{W}{4}\times 2D_3} \tag{9}$$

$$\hat{S}_3^{e+f}=Conv(FFN(S_3^{e+f})) \tag{10}$$

其中，$Concat(\cdot)$为沿通道维度拼接操作；$Conv(\cdot)$是卷积核大小为 3 且激活函数为 $ReLU(\cdot)$的单层卷积；$FFN(\cdot)$为全连接的全连接多层感知机；$\hat{S}_2^{e+f}$，$\hat{S}_3^{e+f}$ 为中尺度与低尺度的粗融合特征。

通过分层处理，针对不同颗粒度的信息采用适配的操作，高层特征通过 Transformer 捕捉全局语义，低层特征通过卷积提取局部细节，FSSE 对 RGB 图像与事件图像特征中的多粒度信息进行了高效融合与利用。

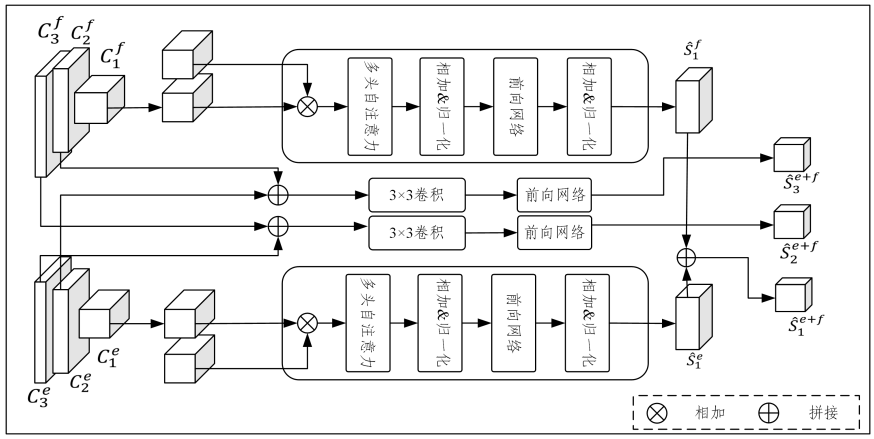

图 5 特征尺度分解空间编码器

Fig. 5 Feature Scale decomposition Space Encoder(FSSE)

2.4 多模块跨尺度融合模块

多模态目标检测的关键挑战在于信息融合，即如何将事件数据与 RGB 数据这两种异构的模态数据背后潜藏的信息进行有效结合，以此提升目标检测在各个场景的有效性与鲁棒性。本文着力于发掘多模态信息中不同层级特征、不同颗粒度信息之间的相关性与互补性。此前基于 Transformer 的融合策略虽展现了一定的效果，但基于注意力的方法带来的计算复杂度升高限制了模型在边缘应用上的部署，这与事件相机低延迟、低功耗的特点背道而驰。相比之下，基于卷积神经网络的方法能够带来更快的响应速度。为了能够利用卷积达成更加充分的多模态融合，本文提出一种多模态跨尺度的融合方法，通过对此前 FSSE 粗略融合的特征进行不同尺度的交互式融合，探索多模态特征中潜藏的更深层次的语义信息。具体如图 6 所示。

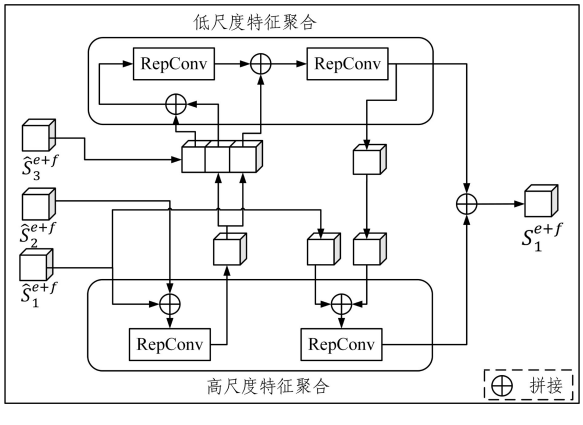

图 6 多模态跨尺度融合模块

Fig. 6 Multi-modal Cross-scale Fusion Module(MCFM)

具体而言，高层次融合阶段会将 $\hat{S}_1^{e+f}$ 与 $\hat{S}_2^{e+f}$ 进行聚合；低层次阶段则将 $\hat{S}_3^{e+f}$ 的输出与高层次阶段的输出进行融合。这两阶段之间进行进一步交互，并将最后融合的特征输出给解码器。这一过程可表示为：

$$\mathcal{F}_1^{high}=RepConv(Concat(\hat{S}_1^{e+f},\hat{S}_2^{e+f})) \tag{11}$$

$$\mathcal{F}_2^{high}=RepConv(Concat(\hat{S}_1^{e+f},\mathcal{F}_2^{low})) \tag{12}$$

$$\mathcal{F}_1^{low}=RepConv(Concat(\mathcal{F}_1^{high},\hat{S}_3^{e+f})) \tag{13}$$

$$\mathcal{F}_2^{low}=RepConv(Concat(\mathcal{F}_1^{low},\mathcal{F}_1^{high})) \tag{14}$$

$$S_1^{e+f}=Concat(\mathcal{F}_2^{low},\mathcal{F}_2^{high}) \tag{15}$$

其中，RepConv[19] 将不同尺度的特征进行融合。RepConv 的结构如图 7 所示，在训练阶段，RepConv 使用多分支的卷积层，在推理阶段，这些分支的参数会被重参数化到主分支上，等效为一个 3×3 的卷积核，这样的结构能够提高推理速度与效率，并保持良好的精度。

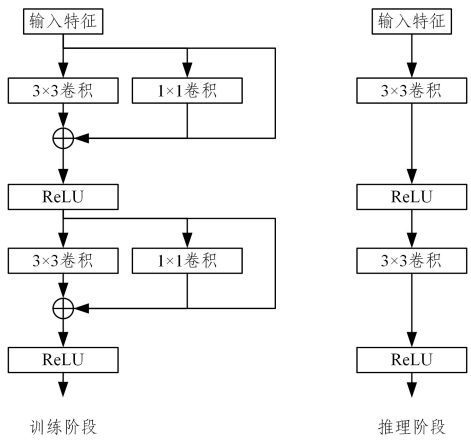

图 7 RepConv 重参数化示意图

Fig. 7 RepConv reparameterization diagram

由于本文的融合模块并未像 SODFormer 一样使用注意力结构，因此本文的融合模块并没有对模型的推理速度造成严重影响。具体的分析对比在下文中给出。双阶段的设计能够确保充分利用来自两种异构模态的不同层级之间的特征。

2.5 模型优化

本文的模型框架在进行损失函数计算时采用计算目标物体的分类损失与目标位置的损失，损失函数的计算式如下：

$$\mathcal{L}=\mathcal{L}_{\text{class}}+\mathcal{L}_{\text{box}} \tag{16}$$

$$\mathcal{L}_{\text{class}}=-\sum_{i=1}^{N}\log p_{\hat{y}_i}(c_i) \tag{17}$$

$$\mathcal{L}_{\text{box}}=\lambda_{L1}\cdot\|b_i-\hat{b}_i\|_1+\lambda_{\text{GIoU}}\cdot(1-GIoU(b_i,\hat{b}_i)) \tag{18}$$

其中，$\mathcal{L}_{\text{class}}$采用改进的 Focal Loss，能够对未匹配的预测施加空类别惩罚，降低虚警率；N 是预测目标物体的数量；$p_{\hat{y}_i}$ 是第 i 个物体的预测类别概率；c_i 是目标物体的类别真值；λ_{L1} 和 λ_{GIoU}是平衡 L1 损失与 GIoU 损失权重的超参；b_i 是预测的目标框，通过预测目标框与真实标签框的差异比较，能够对目标框坐标进行直接约束；GIoU 是测量预测框与真实框之间重叠程度的指标，能够解决边界框重叠与尺寸敏感性问题。通过二分图匹配将预测框与真实标注框进行最优配对，能够避免传统 NMS 后处理，实现端到端的优化。

3 实验与分析

3.1 实验设置

本文在 PKU-SOD 数据集上进行训练，对于低光与运动模糊的界定，采用了 PKU-SOD 数据集的设定。对于光照强弱，将场景细节清晰的片段界定为正常光照，而场景存在显著光照不足的则界定为低光场景。对于运动场景，通过人眼观察判定，将目标运动未导致 RGB 帧出现显著运动模糊的图像定义为正常速度，而目标运动导致 RGB 帧出现可见运动模糊的视频片段则被界定为运动模糊情形。在 PKU-SOD 数据集中，正常光照占比 92%，低光占比 8%，高速场景占比 13%，正常速度占比 87%。

本文采用在 ImageNet 上进行预训练的 ResNet-50 作为模型的骨干网络进行特征预提取。在提取不同尺度层级特征时，将 ResNet 的第 1，2，3 层分别作为高、中、低级特征。为了提升计算速度与效率，本文 FSSE 中，Transformer 的层数设置为 1，Transformer 中的多层感知机层的隐藏维度设置为输入维度的 4 倍。训练过程中，基础学习率设置为 2×10^{-4}，使用 AdamW 优化器，针对不同模块使用不同学习率和权重衰减策略，将骨干网络 ResNet350 的学习率分别设为 1×10^{-5}，并使用正则表达式对 FSSE 和解码器中的 bias 以及归一化层的 weight 设置权重衰减为 0。在 PKU-SOD 数据集上训练 20 个 epoch。此外，为了让模型更好地学习到多尺度信息，本文对输入的 RGB 图像与事件图像进行了多尺度变换。本文实验是在 PyTorch 中实现的，并在 Nvidia Tesla V100 GPU 上进行。

3.2 实验结果与分析

本文模型在大型自动驾驶数据集 PKU-SOD 上进行评估，基于 ResNet-50 骨干网络，与当前几种先进方法从多个维度进行了比较。

3.2.1 对比方法

本文的对比方法包含两类：1）本文设计的初衷是希望以接近单模态方法的参数量达到更优的检测准确率，因此将本文方法与单模态上的经典方法[19]与最优方法[13-14]进行比较，以验证本文多模态方法相比单模态方法的优越性；2）为了验证本文方法在减小参数量的情形下依旧能保持与其他多模态方法的竞争力甚至超越已有的多模态方法，本文与目前的事件-RGB 图像多模态方法[7-8]以及现有的多模态 SOTA 方法[13]进行对比。

3.2.2 定量分析

这一小节中，本文方法与此前在 PKU-SOD 数据集上进行训练测试的工作进行了多个维度的对比，本文选择了模型的参数数量、mAP_{50} 与模型的推理速度作为模型的测验指标，比较结果如表 1 所列。

表 1 与其他目标检测基准模型在 PKU-SOD 上的性能对比

Table 1 Performance comparison with other object detection benchmark models on PKU-SOD

模态	方法	参数量	mAP_{50}/%	推理时间/ms
纯事件模态	SSD-events[21]	$>60\times10^6$	22.1	**7.2**
	Faster-RCNN[22]	41×10^6	25.1	74.5
	Deformable DETR[21]	**39×10^6**	30.7	21.6
	Spatial-Temporal DETR[13]	42×10^6	33.4	25.0
	RT-DETR[14]	42×10^6	35.5	20.8
纯 RGB 模态	Faster-RCNN[23]	41×10^6	44.3	75.2
	YOLOv3[20]	$>60\times10^6$	42.6	7.9
	Deformable DETR[21]	**39×10^6**	46.1	21.5
	Spatial-Temporal DETR[13]	42×10^6	48.9	24.9
	RT-DETR[14]	42×10^6	48.6	20.7
RGB 事件多模态	MFEPD[7]	$>60\times10^6$	43.8	8.2
	JDF[8]	$>60\times10^6$	44.2	8.3
	SODFormer[13]	$>80\times10^6$	50.4	39.7
	EvR-DETR(本文方法)	44×10^6	**52.9**	31.4

为了更好地凸显本文方法的优越性，可视化了本文与对比基线方法的各项性能参数，如图 8 所示，其中横坐标代表模型的参数量，纵坐标代表模型的推理准确率，气泡的大小代表模型的推理时间。可以看到，本文的多模态方法的参数量与推理速度保持在合理范围内并实现了最高的准确率，且与此前准确率最高的方法相比大大减少了参数量并提高了推理速度。

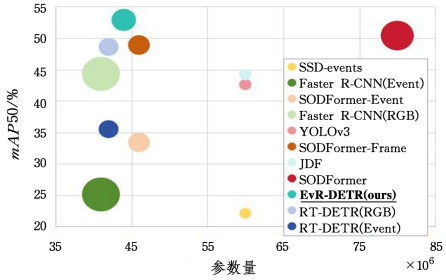

图 8 各项性能参数对比气泡图

Fig. 8 Chart of performance parameter comparison bubble

1）模型参数。本文方法仅对每个模态的最高尺度特征使用 Transformer 进行全局特征提取，因此并未使用大量的 Attention 结构。此外，本文整体使用 Transformer 高阶特征处理与 RepConv 全阶特征融合的结构，很好地降低了 EvR-DETR 的参数量。通常而言，模型参数量的下降会带来性能随之下降的问题，但本文使用的跨尺度的多模态融合策略高效利用了特征信息，弥补了性能问题。与此前的性能最优（State of Art，SOTA）模型 SODFormer 相比，本文模型的参数量下降了近 50%，并且没有检测准确率的性能妥协。

2）检测准确率。基于 RGB＋事件模态融合检测方法的检测准确率通常要高于基于 RGB 帧或纯事件模态的单模态方法。这表明物体的纹理信息对于目标检测任务来说至关重要，而事件所带来的光照变化信息能够有效弥补 RGB 帧在恶劣条件（弱光、模糊）情形下的缺失信息。同时，由于 EvR-DETR 是跨尺度多模态特征融合的设计，而非 SODFormer 中仅对最高尺度特征进行融合的设计，因此它能够从不同尺度的特征中更充分地提取到信息。

此前的 SOTA 方法 SODFormer 由于参数量过大且计算复杂度高，采用了一种对每个模块进行单独训练，并在训练融合模块时对特征提取模块进行参数冻结的训练方法，导致特征提取器不能够提取适合于融合的最优特征。而 EvR-DETR 的端到端训练方式能够有效解决这一问题，使得 EvR-DETR 能够在减少近一半参数量的情况下达到更高的检测准确率。

3）推理时间。基于单阶段 Transformer 的 FSSE 与基于卷积的 MCFM 大幅减少了 Attention 结构的应用，使得本文 EvR-DETR 的推理速度相比此前的 SOTA 模型有了明显的提升。与 SODFormer 相比，这一混合不同机制的结构提升了近 20%的推理速度。

3.2.3 定性分析

基于 ResNet-50 的单模态与多模态目标检测可视化结果如图 9 所示。可视化图中 4 列分别代表事件 RGB 图像、事件可视化图像、真实标签和预测结果。针对不同的光照场景，本文方法能够保持检测的准确性与鲁棒性，从而在低光、高速模糊这样的场景下获得有效的检测结果。

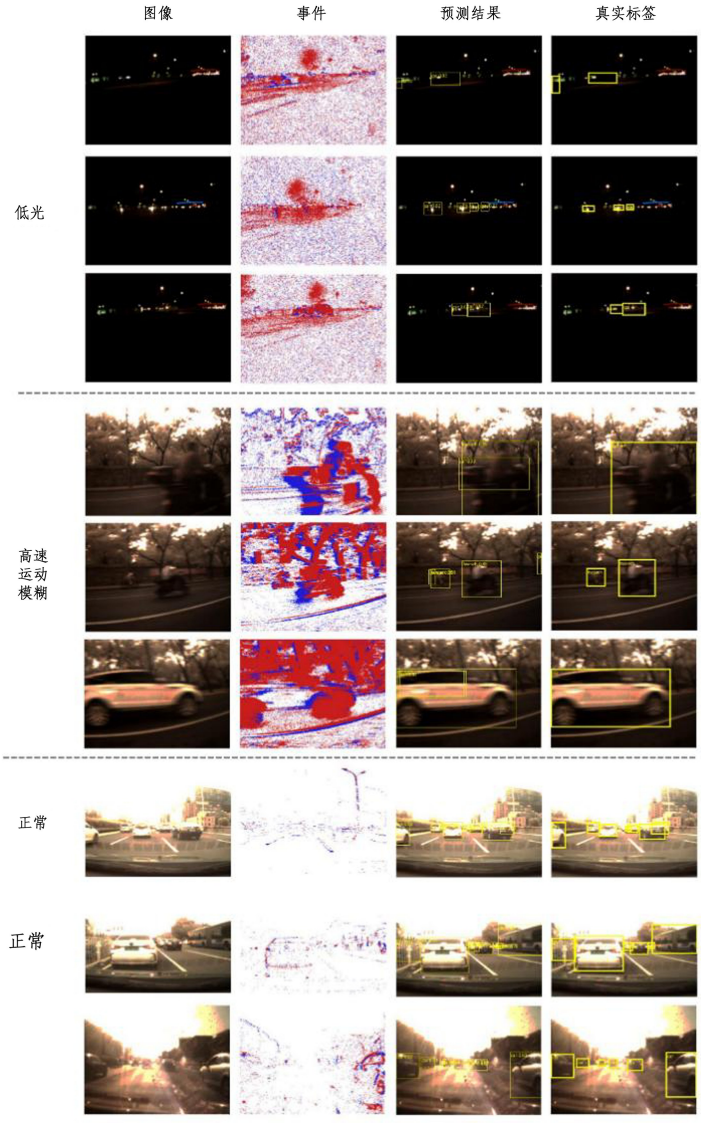

图 9 不同场景下目标检测定性结果对比

Fig. 9 Qualitative object detection results comparison in various scenarios

3.2.4 与其他检测 Transformer 方法的对比

由于本文方法是基于端到端检测 Transformer 的方法，因此也在多个指标上对比了本文方法与其他基于检测 Transformer 的方法，如表 2 所列。

可以看到，本文方法的每秒 10 亿次的浮点运算数略高于单模态方法，但准确率显著低于目前的 SOTA 工作 SODFormer，相比其降低了近 40%。同时，本文方法在 mAP_S，mAP_M，mAP_L 上均高于其他现有的基于端到端检测 Transformer 的方法，在小目标、中目标以及大目标指标上领先目前的单模态 SOTA 方法 RT-DETR，并显著超越融合方法 SODFormer，表明了本文方法在这 3 种大小目标的检测上的方法优越性。

表 2 DETR 类模型相关指标对比

Table 2 Comparison of related metrics for DETR-based models

(%)

模态	方法	GFLOP	mAP_S	mAP_M	mAP_L
纯事件模态	Deformable DETR[21]	31.5	5.7	19.8	32.9
	Spatial-Temporal DETR[13]	31.9	6.5	21.0	34.8
	RT-DETR[14]	20.6	7.4	26.5	45.0
纯 RGB 模态	Deformable DETR[21]	31.5	13.4	25.3	40.3
	Spatial-Temporal DETR[13]	31.9	14.2	26.4	41.7
	RT-DETR[14]	**20.6**	17.5	33.7	46.5
事件 RGB 多模态	SODFormer[13]	62.5	14.4	28.5	45.4
	EvR-DETR(本文方法)	36.8	**18.1**	**40.1**	**55.6**

3.3 消融实验

针对不同模态的使用策略与融合策略，本文进行具体实验与进一步分析。为了更好地对比，本文采用具有单模态跨尺度融合模块的模型 RT-DETR[23] 作为比较模型来探究模态使用策略的影响。为了探究 MCFM 的有效性，本文建立了多模态早期融合的模型以作比较，这一模型先对两种模态的输入数据进行平均融合，再将平均融合的数据输入 RT-DETR 中。通过不同场景下模型的表现，观察多模态融合策略的影响。本文选取了 mAP，mAP_{50}，mAP_{75}，mAP_S，mAP_M，mAP_L 作为对比分析的指标，以验证本文多模态融合策略的有效性。

3.4 早期融合网络

由于本文网络采用了一种多模态跨尺度融合的策略，模型一方面利用多尺度的特征信息，另一方面利用不同模态的互补信息。为了探究模型性能提升是否单纯是由事件图的边缘信息所带来的，而非融合模块充分利用事件信息的潜空间中的相应特征，本文基于 FSSE 的主干与平均融合方法[23]设计了一种早期融合网络。这一网络通过平均融合将 RGB 帧与事件图结合，以达到引入边缘信息的作用，再将结合后的图像输入以 FSSE 主干网络为基础的目标检测网络中。

与单模态的目标检测相比，平均融合后的图像结合了 RGB 图像的纹理信息与事件图像带来的边缘信息，在低光环境下能够为目标检测带来增益。但同时，较为粗暴的平均融合并未利用两种模态的深层特征信息，且事件图像中的事件噪点容易为评价融合图像的特征提取带来影响，并最终影响目标检测的效果。

1)低光场景下的消融对比

如表 3 所列，在低光场景下，单模态的事件流在晚期融合中结合了事件流和常规帧数据的优势，事件流在动态目标检测(mAP_L：58.5% vs. 52.4%)上表现突出，而常规帧数据在静态纹理信息上更为可靠，二者互补，显著提升了整体性能(mAP：21.2% vs. 15.2%)。其次，与早期融合相比，晚期融合在中等目标(mAP_M：21.6% vs. 20.7%)和小目标(mAP_S：16.6% vs. 12.9%)上表现更优，表明延迟特征交互能更好地保留模态特异性：事件数据捕捉运动轮廓，帧数据补充细节纹理，二者在高层语义融合时产生更强的互补效应。此外，晚期融合在整体精度(mAP：21.2% vs. 18.6%)和关键指标(mAP_{50}：52.1% vs. 42.9%)上均优于早期融合，说明其能更有效地整合多模态信息，避免早期融合可能导致的特征混淆。综上，晚期融合通过延迟交互和多模态互补，显著提升了低光场景下的检测性能。

表 3 低光场景下模态种类及融合阶段的消融实验结果

Table 3 Ablation study results on modality types and fusion stages in low-light scenarios

(%)

场景	模态		融合阶段		mAP	mAP_{50}	mAP_{75}	mAP_S	mAP_M	mAP_L
	事件	RGB	事件	RGB						
低光场景	√		—	—	13.6	34.4	8.8	7.7	16.3	52.4
		√	—	—	15.2	36.1	11.5	13.4	16.4	39.9
	√	√	√		18.6	42.9	13.7	12.9	20.7	56.8
	√	√		√	21.2	52.1	15.0	16.6	21.6	58.5

2)高速场景下的消融对比

在高速运动模糊场景下，晚期融合相较于早期融合和其他方法表现出显著的优越性，如表 4 所列。具体而言，晚期融合在 mAP 指标上达到了 22.0%，相较于早期融合的 19.7%提升了 2.3 个百分点。这一提升主要归因于晚期融合能够在特征提取后充分利用事件相机和 RGB 相机的互补信息，从而更有效地捕捉目标的动态和静态特征。事件相机的高时间分辨率能够捕捉快速运动的目标，而 RGB 相机则提供了丰富的纹理信息，两者结合，显著提高了检测精度。此外，晚期融合在 mAP，mAP_{50}，mAP_{75}，mAP_S，mAP_M，mAP_L 等指标上也均有明显提升，分别达到了46.6%，17.7%，13.4%，29.1%和 46.2%，相较于早期

融合分别提升了 4.8 个百分点、1.9 个百分点、2.6 个百分点、2.1 个百分点和 0.7 个百分点。这些提升表明，晚期融合在处理高速运动模糊场景时，能够更好地利用两种模态的优势，提高目标检测的准确性和鲁棒性。

表 4 高速运动场景的模态种类及融合阶段的消融实验结果

Table 4 Ablation study results on modality types and fusion stages in high-speed motion scenarios

(%)

场景	模态		融合阶段		mAP	mAP_{50}	mAP_{75}	mAP_S	mAP_M	mAP_L
	事件	RGB	事件	RGB						
	√		—	—	13.5	30.8	10.2	5.6	20.1	33.7
高速		√	—	—	18.1	38.7	14.6	11.2	24.3	41.1
场景	√	√	√		19.7	41.8	15.8	10.8	27.0	45.5
	√	√		√	22.0	46.6	17.7	13.4	29.1	46.2

3）正常场景下的消融对比

在正常场景下，晚期融合相较于早期融合及其他方法展现出显著优势。如表 5 所列，晚期融合的 mAP 达到 27.3%，相比早期融合的 25.7% 有所提升。尤其在 mAP_L 指标上，晚期融合取得 23.4% 的优异成绩，优于早期融合的 22.2%，这表明晚期融合在高精度定位方面更具优势。此外，晚期融合在 mAP_S（小目标检测）上达到 19.4%，超越早期融合的 17.5%，凸显出其在小目标检测方面的卓越能力，通过有效结合事件相机的动态特性和 RGB 相机的静态纹理信息，显著提高了目标检测的整体性能和定位精度。

表 5 正常场景下模态种类及融合阶段的消融实验结果

Table 5 Ablation study results on modality types and fusion stages in normal scenarios

(%)

场景	模态		融合阶段		mAP	mAP_{50}	mAP_{75}	mAP_S	mAP_M	mAP_L
	事件	RGB	事件	RGB						
	√		—	—	17.0	37.5	13.1	8.2	30.7	57.9
正常		√	—	—	27.2	**55.8**	23.0	20.7	39.9	57.8
场景	√	√	√		25.7	52.6	22.2	17.5	38.0	65.4
	√	√		√	**27.3**	55.5	**23.4**	**19.4**	**45.2**	**68.4**

4）总体消融对比

在整体场景中，如表 6 所列，晚期融合相较于早期融合及其他方法表现出显著的优势。具体而言，晚期融合的 mAP 达到了 25.4%，相较于早期融合的 23.2% 提升了 2.2 个百分点。这一提升主要归因于晚期融合能够在特征提取后充分利用事件相机和 RGB 相机的互补信息，从而更有效地捕捉目标的动态和静态特征。事件相机的高时间分辨率能够捕捉快速运动的目标，而 RGB 相机则提供了丰富的纹理信息，两者结合显著提高了检测精度。此外，晚期融合在 mAP，mAP_{50}，mAP_{75}，mAP_S，mAP_M，mAP_L 等指标上也均有明显提升，分别达到了 52.9%，21.1%，18.1%，40.1% 和 55.6%，相较于早期融合分别提升了 4.4 个百分点、1.8 个百分点、2.6 个百分点、6.7 个百分点和 1.6 个百分点。这些提升表明，晚期融合在处理不同场景时，能够更好地利用两种模态的优势，提高目标检测的准确性和鲁棒性。

表 6 全部场景下模态种类及融合阶段的消融实验结果

Table 6 Ablation study results on modality types and fusion stages in all scenarios

(%)

场景	模态		融合阶段		mAP	mAP_{50}	mAP_{75}	mAP_S	mAP_M	mAP_L
	事件	RGB	事件	RGB						
	√		—	—	15.6	35.5	11.7	7.4	26.5	45.0
全部		√	—	—	23.2	48.6	19.2	17.5	33.7	46.5
场景	√	√	√		23.2	48.5	19.3	15.5	33.4	54.0
	√	√		√	**25.4**	**52.9**	**21.1**	**18.1**	**40.1**	**55.6**

结束语 本文提出了一种轻量级的多模态端到端目标检测框架，这一框架能够利用 RGB 图像与事件数据进行跨尺度融合，充分利用 RGB 图像中的纹理信息与事件数据的光照变化时空信息。与单模态的对比模型相比，本文方法显著提升了低光环境以及高速运动环境下的检测性能，弥补了低光带来的光照信息缺失以及高速运动带来的图像模糊造成的性能损失，整体上提升了自动驾驶环境下各个场景的检测性能。与其他多模态方法相比，本文方法极大地减少了模型的参数量，并提升了检测的准确率以及检测的推理速度。在大规模数据集 PKU-SOD 上的测试充分证明了本文方法在恶劣环境下进行自动驾驶环境目标检测的有效性。

参 考 文 献

[1] LIU L，OUYANG W，WANG X，et al. Deep Learning for Generic Object Detection：A Survey [J]. International Journal of Computer Vision，2020，128(2)：261-318.

[2] GALLEGO G，DELBRÜCK T，ORCHARD G，et al. Event-Based Vision：A Survey [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence，2022，44(1)：154-180.

[3] CHEN G，CAO H，CONRADT J，et al. Event-Based Neuromorphic Vision for Autonomous Driving：A Paradigm Shift for Bio-Inspired Visual Sensing and Perception [J]. IEEE Signal Pro-

cessing Magazine, 2020, 37(4): 34-49.
[4] BERNER R, BRANDLI C, YANG M, et al. A latency sparseoutput vision sensor for mobile applications [C]// 2013 Symposium on VLSI Circuits. 2013: C186-C187.
[5] WANG L, LIU Z, SHI D X, et al. Fusion Tracker: Single-object Tracking Framework Fusing Image Features and Event Features [J]. Computer Science, 2023, 50(10): 96-103.
[6] JIANG J, ZHOU B, ZHOU T, et al. Deep Event-Based Object Detection in Autonomous Driving: A Survey [C]// 2024 10th International Conference on Big Data and Information Analytics (BigDIA). 2024: 447-454.
[7] JIANG Z, XIA P, HUANG K, et al. Mixed Frame-/Event-Driven Fast Pedestrian Detection [C]// 2019 International Conference on Robotics and Automation(ICRA). 2019: 8332-8338.
[8] LI J, DONG S, YU Z, et al. Event-Based Vision Enhanced: A Joint Detection Framework in Autonomous Driving [C]// 2019 IEEE International Conference on Multimedia and Expo (ICME). 2019: 1396-1401.
[9] TOMY A, PAIGWAR A, MANN K S, et al. Fusing Event-based and RGB camera for Robust Object Detection in Adverse Conditions [C]// 2022 International Conference on Robotics and Automation(ICRA). 2022: 933-939.
[10] DOSOVITSKIY A, BEYER L, KOLESNIKOV A, et al. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale[C]// International Conference on Learning Representations. 2021.
[11] LIU M, QI N, SHI Y, et al. An Attention Fusion Network For Event-Based Vehicle Object Detection[C]// IEEE International Conference on Image Processing. IEEE, 2021.
[12] ZHOU Z, WU Z, BOUTTEAU R, et al. RGB-Event Fusion for Moving Object Detection in Autonomous Driving [C]// 2023 IEEE International Conference on Robotics and Automation (ICRA). 2023: 7808-7815.
[13] LI D, TIAN Y, LI J. SODFormer: Streaming Object Detection With Transformer Using Events and Frames [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2023, 45(11): 14020-14037.
[14] ZHAO Y, LYU W, XU S, et al. DETRs Beat YOLOs on Real-time Object Detection [C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2024: 16965-16974.
[15] CARION N, MASSA F, SYNNAEVE G, et al. End-to-End Object Detection with Transformers [C]// Computer Vision—ECCV 2020. Cham: Springer, 2020: 213-229.
[16] BERNER R, BRANDLI C, YANG M, et al. A 240×180 10mW 12us latency sparseoutput vision sensor for mobile applications [C]// 2013 Symposium on VLSI Circuits. 2013: C186-C187.
[17] HE K, ZHANG X, REN S, et al. Deep Residual Learning for Image Recognition [C]// 2016 IEEE Conference on Computer Vision and Pattern Recognition(CVPR). 2016: 770-778.
[18] LI F, ZHANG H, LIU S, et al. DN-DETR: Accelerate DETR Training by Introducing Query DeNoising [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2024, 46(4): 2239-2251.
[19] DING X, ZHANG X, MA N, et al. RepVGG: Making VGG-style ConvNets Great Again [C]// 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition(CVPR). 2021: 13728-13737.
[20] ZHU X, SU W, LU L, et al. Deformable DETR: Deformable Transformers for End-to-End Object Detection[J]. arXiv: 2010.04159, 2020.
[21] IACONO M, WEBER S, GLOVER A, et al. Towards Event-Driven Object Detection with Off-the-Shelf Deep Learning [C]// 2018 IEEE/RSJ International Conference on Intelligent Robots and Systems(IROS). 2018: 1-9.
[22] 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, 2017, 39(6): 1137-1149.
[23] LI H, WU X J, KITTLER J. MDLatLRR: A Novel Decomposition Method for Infrared and Visible Image Fusion [J]. IEEE Transactions on Image Processing, 2020, 29: 4733-4746.

ZHOU Bingquan, born in 2000, postgraduate, professor. His main research interests include computer vision and event-based vision.

JIANG Jie, born in 1974, Ph.D, professor. His main research interests include artificial intelligence and deep learning, visualization and visual analytics, virtual reality and intelligent interaction.

(责任编辑:何杨)