Communications in Transportation Research
https://www.editorialmanager.com/commtr/

Beyond conventional vision: RGB-event fusion for robust object detection in dynamic traffic scenarios

Zhanwen Liu[a,*], Yujing Sun[a], Yang Wang[a] , Nan Yang[a], Shengbo Eben Li[b], Xiangmo Zhao[a]

[a] *School of Information Engineering, Chang'an University, Xi'an, 710000, China*

[b] *School of Vehicle Mobility & College of AI, Tsinghua University, Beijing, 100084, China*

[*] *Corresponding author.*

E-mail address: zwliu@chd.edu.cn

Abstract

The dynamic range limitation is intrinsic to conventional RGB cameras, which reduces global contrast and causes the loss of high-frequency details such as textures and edges in complex, dynamic traffic environments (e.g., nighttime driving or tunnel scenes). This deficiency hinders the extraction of discriminative features and degrades the performance of frame-based traffic object detection. To address this problem, we introduce a bio-inspired event camera integrated with an RGB camera to complement high dynamic range information, and propose a motion cue fusion network (MCFNet), an innovative fusion network that optimally achieves spatiotemporal alignment and develops an adaptive strategy for cross-modal feature fusion, to overcome performance degradation under challenging lighting conditions. Specifically, we design an event correction module (ECM) that temporally aligns asynchronous event streams with their corresponding image frames through optical-flow-based warping. The ECM is jointly optimized with the downstream object detection network to learn task-ware event representations. Subsequently, the event dynamic upsampling module (EDUM) enhances the spatial resolution of event frames to align its distribution with the structures of image pixels, achieving precise spatiotemporal alignment. Finally, the cross-modal mamba fusion module (CMM) employs adaptive feature fusion through a novel cross-modal interlaced scanning mechanism, effectively integrating complementary information for robust detection performance. Experiments conducted on the DSEC-Det and PKU-DAVIS-SOD datasets demonstrate that MCFNet significantly outperforms existing methods in various poor lighting and fast moving traffic scenarios. Notably, on the DSEC-Det dataset, MCFNet achieves a remarkable improvement, surpassing the

best existing methods by 7.4 % in mAP50 and 1.7 % in mAP metrics, respectively. The code is available at https://github.com/Charm11492/MCFNet.

Keywords

dynamic traffic environments, object detection, multimodal fusion, event camera

1 介绍

视觉感知系统是使智能车辆能够感知周围环境并促进决策制定的基础组件 (**???**) 。这些系统必须在复杂动态的边缘情况下（如夜间驾驶和隧道导航）保持对物体检测的强大性能。然而，由于传统 RGB 相机受限于其内在感光元件的窄动态范围，常常难以在这些具有挑战性的条件下捕捉关键的边缘细节和快速的场景变化。如图 1 所示，这导致图像质量下降，削弱了特征提取的辨识能力，进而对基于帧的交通物体检测方法的准确性产生负面影响 (**?**) 。

相比之下，受生物启发的事件相机提供了高动态范围和微秒级的时间分辨率，可以在极端条件下（如低光和过曝光 (**???**) ）实现稳定成像。通过利用 RGB 和事件信息的互补优势，多模态融合方法为复杂交通环境中的视觉感知提供了有前途的途径 (**?**) 。最近，许多物体检测算法被开发出来以利用 RGB 和事件融合，这显著增强了在复杂场景下视觉感知系统的鲁棒性 (**???**) 。这些方法通常包括设计专用的多模态表示分支以从每种模式中提取特征，并实施先进的融合模块以深入整合跨模态信息，从而提高检测性能。然而，RGB 和事件相机之间的异质性导致多模态数据在时空上不一致，这为现有的多模态特征提取和对准方法带来了挑战。首先，在时间维度上，事件数据的微秒级时间分辨率显著高于 RGB 数据的毫秒级分辨率，导致时间上不对齐，使得直接的多模态融合变得不可行 (**?**) 。为了解决这个问题，现有方法通常对事件数据进行时间采样和压缩，以生成与 RGB 数据时间分辨率对齐的事件帧。这些方法大致可以分为两类：事件帧积累方法和运动补偿方法。事件帧积累方法使用事件的统计特征生成类似图像的张量 (**?????**) ，但它们不可避免地受到热噪声和暗电流噪声的影响，从而降低了后续结构特征提取的质量。运动补偿方法利用事件数据的时间动态来对齐连续运动事件，从而增强运动区域并形成高强度边缘，同时通过不连续运动抑制噪声。然而，这些方法 (**??**) 依赖于恒定光照和线性运动的假设，而在具有频繁亮度变化（例如，车辆前灯、街灯）和复杂运动模式的高度动态交通场景中，这些假设经常被违反，导致不准确的运动估计和事件帧质量下降 (**??**) 。其次，在空间维

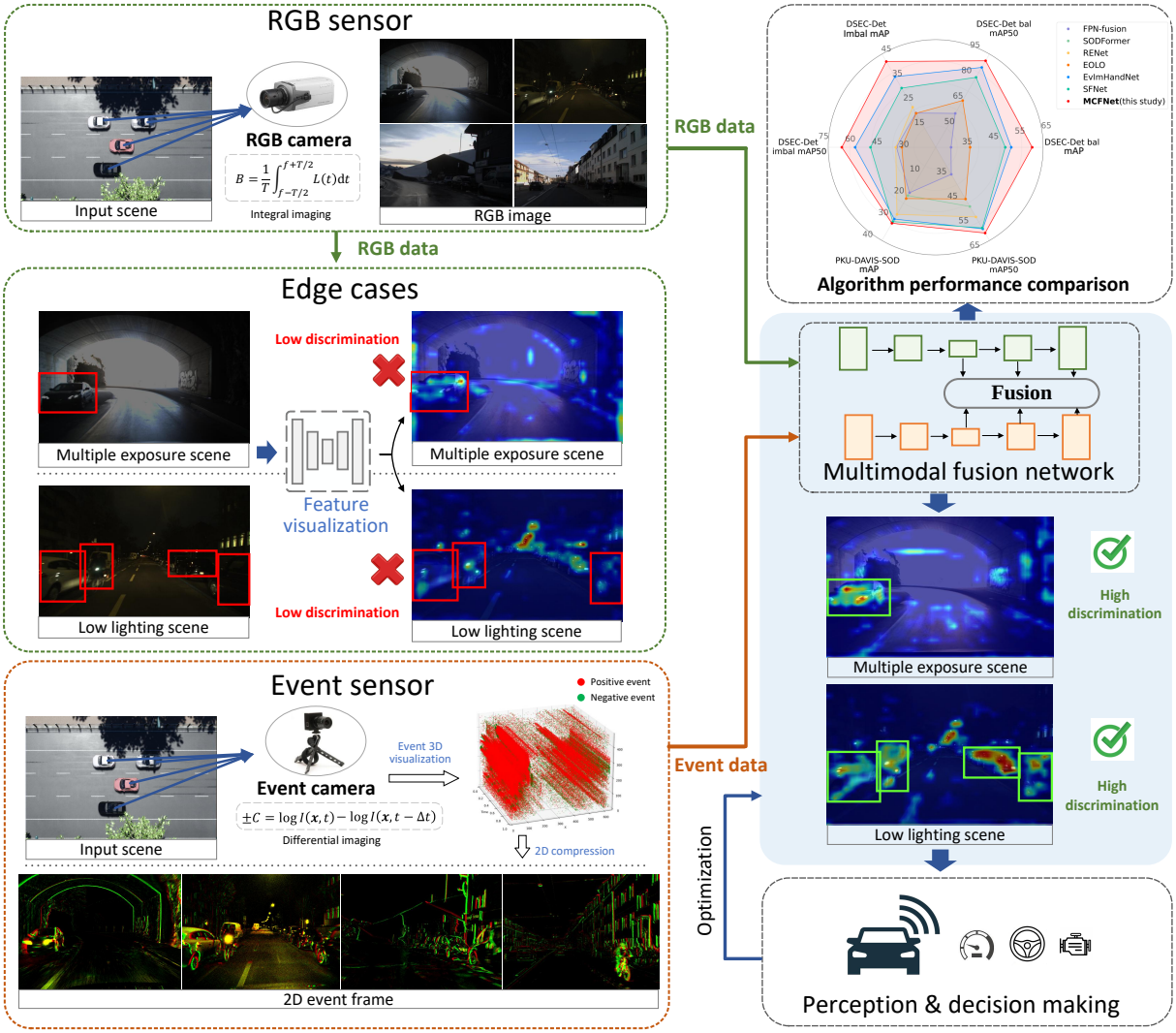

Fig. 1 在边界感知场景中，传统的 RGB 相机因其光敏元件和固定的帧速率而受到限制，导致由于曝光不足而丢失关键信息（左上角），这会导致后续算法的失效。相比之下，事件相机可以在极端条件下捕捉高动态的时空特征（左下角）。为此，我们提出了一种多模态融合算法，该算法联合优化 RGB 数据和事件数据的互补性，以构建跨模态特征表示和决策机制（右下角），从而显著增强智能车辆在复杂场景中的感知鲁棒性和准确性（右上角）。

度上，RGB 数据通常比事件数据具有更高的分辨率，这导致显著的不一致 (**?**) 。现有方法通常通过直接下采样 RGB 数据以匹配事件数据的分辨率来解决这一问题，这进一步加剧了信息损失 (**??**) 。在从两种模态提取表示后，光照分布的空间不一致性导致这两种模态在不同区域上具有不同的可辨别性水平。这需要动态平衡每种模态的贡献，以实现稳健的跨模态特征融合。现有方法通常通过简单的加法或拼接来融合两种模态的特征 (**???**) ，这无法根据动态场景变化自适应调整融合比。最新的方法使用注意机制来捕获互补信息 (**????**) 。然而，它们要么将一种模态视为辅助以增强另一种模态，要么为每种模态构建单独的增强子模块，未能充分利用模态间的协作信息并阻碍模态贡献的精确平衡。

为了应对上述挑战，我们提出了一种新颖的运动线索融合网络 (MCFNet)，该网络由三个主要组件组成：事件校正模块 (ECM) 和事件动态上采样模块 (EDUM)，用于 RGB 与事件数据之间的时空对齐，以生成高质量的像素对像素对齐特征；跨模态曼巴融合模块 (CMM) 用于在模态间进行全局上下文建模以指导自适应融合。具体而言，ECM 估计事件流的运动向量，以在时间上与 RGB 帧时间戳对齐运动事件。为了克服恒定照明和线性运动的限制，采用了一种端到端的联合训练策略，该策略利用高级语义信息来指导模型优化。这使得在高度动态的交通场景中学习场景特征并生成具有高强度边缘和鲁棒性噪声抵抗力的事件帧。在此基础上，引入了 EDUM，通过利用事件特征的空间分布和图像局部平滑特征来进行动态上采样以抑制噪声，从而产生高分辨率和高质量的事件特征。在从两种模态中提取高质量的时空对齐特征之后，CMM 使用跨模态全局上下文交互来感知场景内容并评估每种模态的贡献。这指导了自适应利用事件动态或 RGB 纹理特征进行跨模态融合，充分发挥它们的互补特性。为了验证 MCFNet 的有效性，我们在 DSEC-Det (**?**) 和 PKU-DAVIS-SOD (**?**) 数据集上进行了广泛实验。如图 1 所示，MCFNet 显著优于现有方法。

总之，本研究的主要贡献如下：

1) 在时间对齐中，提出了一种基于自监督学习的事件校正模块（ECM），用于估计全场景光流，为联合训练提供初始值。随后，ECM 与检测网络共同优化，以降低光流中恒定照明和线性运动假设的影响，从而保证高质量的任务感知事件表示。

2) 在空间对齐中，设计了一种新颖的事件动态上采样模块（EDUM），用于提取 RGB 特征的平滑属性，以在事件表示上采样过程中抑制噪声放大，保证更纯净的结构特征。

我们提出了一种由跨模态交错扫描机制驱动的跨模态曼巴融合模块 (CMM)。这一机制促进了深度跨模态特征的交互和全局上下文提取，显著增强了对互补跨模态特征的感知和融合。

4) 我们采用最先进的目标检测方法在 DSEC-Det 和 PKU-DAVIS-SOD 数据集上进行了广泛的训练和测试，结果表明我们提出的 MCFNet 在性能上显著超越了现有的方法。

2 相关工作

本节首先介绍动态交通场景中现有的目标检测算法，然后介绍事件相机的工作原理和事件表示方法。

2.1 动态交通场景中的目标检测

RGB & RGB-雷达融合。**?** 提出了 Mamba YOLO 检测器，通过引入残差门控 (RG) 机制增强状态空间模型 (SSMs) 在交通物体检测中的局部建模能力。这种检测器利用选择性二维 (2D) 扫描来处理高分辨率图像特征，以线性计算复杂度实现竞争性能。YOLOv11 通过增强特征提取和采用轻量化设计进一步改善交通物体检测，在提供高精度和低延迟的同时，这些改进使得在自动驾驶场景中能够快速响应并支持安全决策 (**?**) 。虽然基于图像的物体检测方法提供了出色的实时性能，但 RGB 相机在恶劣天气或光照条件变化下容易导致性能下降，这可能导致检测失败 (**???**) 。为了解决这一限制，部分研究人员引入了免疫光照变化的雷达传感器，并提出将其与 RGB 图像融合，以增强在低光或动态光照环境中的感知鲁棒性。例如，**?** 开发了 RCTrans，一种基于 Transformers 的雷达-摄像头融合模型。它使用密集雷达编码器增强稀疏雷达点云特征，并通过剪枝顺序解码器逐步定位目标。类似地，**?** 集成图像和点云数据设计了一个模块，该模块将遮挡状态估计与最佳外观特征选择相结合，有效缓解了自动驾驶车辆在交通物体跟踪中与遮挡相关的挑战。然而，雷达相对较低的扫描频率（通常为 10-20 Hz）可能导致在高速场景中运动信息丢失或失真，可能损害检测精度并威胁自动驾驶系统的安全稳定运行 (**?**) 。

RGB-事件融合。神经形态相机最近被引入，以克服 RGB 相机和雷达在捕捉运动信息方面的局限性。由于其异步输出，事件相机提供了极高的时间分辨率和广泛的动态范围，使其在复杂和动态的交通环境中表现出色。在使用融合的 RGB 和事件数据进行目标检测的背景下，现有的方法大致可以分为两类：后期融合和中间融合。后期融合方法在决策层面运行。例如，**?** 应用非极大值抑制（NMS）来合并两种模态的检测输出；**?** 采用 Dempster-Shafer 理论来整合事件和帧，以检测车辆和行人；相关研究 (**?**) 建议结合来自每种模态的置信图。然而，这些方法通常由于特征交互不足而受限，未能充分利用 RGB 和事件数据的互补特性。相反，中间融合策略在特征层面上执行整合。**?** 采用简单的事件和基于帧的特征连接来增强性能；SFNet 引入了两个专用的子模块来细化每种模态的特征，以实现有效的跨模态融合 (**?**) ；RENet 设计了一个双向融合模块，以跨空间和通道维度捕获多模态特征 (**?**) ；SODformer 提出了一种基于异步注意力的融合机制 (**?**) ；EOLO 结合了一个带注意力机制的对称融合模块，以对准 RGB 和事件特征 (**?**) 。EvImHandNet (**?**) 利用事件和 RGB 相机的互补优势，通过融合模块实现 3 维（3D）手部网格重建，该模块采用空间注意力机制和全连接层进行加权融合。然而，它在跨模态交互性方面仍

然有限。为了解决这个问题并增强跨模态特征融合，CAFR (**?**) 引入了一个基于自注意力的融合模块，该模块分别计算 rgb_ 交叉注意力 和 事件交叉注意力 。这种双路径设计引入了参数冗余，导致计算开销增加。尽管这些方法有其优点，但它们主要强调跨模态特征交互，而忽视了对两种模态全球上下文信息的联合建模。这种疏忽限制了网络根据不同场景特征自适应利用事件数据中的动态线索和 RGB 图像中丰富纹理的能力。

此外，大多数融合方法假设两种模态具有相同的空间分辨率。然而，广泛使用的传感器如异步时间图像传感器（ATISs）和动态活动像素视觉传感器（DAVISs），其在共享像素位置输出事件流和强度图像，提供的分辨率相对较低，无法满足交通感知任务的需求。更常见的是，现代成像系统在空间上分隔了 RGB 和事件摄像机，使得能够同时获取高分辨率 RGB 图像和事件数据。当对这类数据进行多模态融合时，必须解决分辨率不一致的问题。现有解决方案通常将 RGB 图像下采样以匹配事件数据的分辨率，这会导致高频视觉信息的损失和感知性能的下降。为了解决这一问题，我们建议在融合之前对事件特征进行动态上采样，从而最大化利用丰富的视觉内容并增强融合的效果。

事件相机是一种神经形态传感器，其感光芯片由多个独立操作的比较器组成，通过检测光强度变化来生成事件。具体来说，当某个像素位置的比较器感知到亮度变化并且电压变化超过预设阈值时，它会生成一个事件，其中包括位置、时间戳和极性。此过程为：其中极性表示事件的极性，数值 1 和 -1 分别表示像素光强度的增加或减少；和分别表示在时间和的光照强度。

有效地表示事件数据对于准确建模至关重要。现有的融合框架通常将异步事件转换为密集的、类似图像的表示，以便可以使用成熟的机器学习算法和神经网络架构。这些密集表征可以大致分为四类：基于图像的方法 (**??**) 、基于时间戳的方法 (**??**) 、基于体素的方法 (**?**) 和运动补偿方法 (**?????**) 。基于图像的方法通常依赖于事件统计或极性计数，这通常导致时间信息的重大丢失。基于时间戳的方法，如时间表面 (**?**) 、HATS (**?**) 和 DiST (**?**) ，在事件窗口内应用时间衰减，为较新的事件赋予更高的权重。基于体素的方法，如体素网格 (**?**) ，将原始事件流划分为时间段，并应用插值技术来构建体素化表示。然而，这些方法对热噪声和暗电流噪声非常敏感，这可能降低后续结构信息提取的质量。

最近的运动补偿方法 (**??????**) 显示了很有前景的结果，可以分为基于模型的、自监督的和监督学习的方法。基于模型的技术 (**???**) 通过模拟运动并基于对比度最大化来优化模型。自监督方法通常通过假设亮度恒定和线性运动来估计光流，这在涉及复杂运动或变化光照（例如，车

辆前灯和路灯）的真实世界条件下常常被违背，这导致重建帧中的伪影和噪声。为了克服这些限制，一些研究 (**??**) 结合递归结构来增强时间建模，并减轻过于简单化假设的缺点。然而，这些方法通常增加了训练的复杂性，并且其任务解耦的设计限制了光流网络在语义丰富场景中建模复杂运动的能力。另一方面，监督方法需要真实的光流，而大多数目标检测数据集中通常缺乏这种数据，并且在真实应用中不切实际，因此在此不作讨论。

3 提出的方法

本节首先在第 3.1 节介绍 Mamba 的原理，然后在第 3.2 节描述所提方法的整体架构。第 **??** 节介绍 ECM 的详细信息，揭示其内部工作原理。第 **??** 节展示了 EDUM 的设计。最后，第 3.3 节介绍了 CMM。

3.1 SSMs 和 mamba 的介绍

Mamba 代表了一种新颖的序列建模架构，基本上基于选择性 SSMs (**?**) 。SSM 是一种线性模型，旨在刻画系统随时间变化的动态行为，将一维（1D）输入信号 $x(t) \in R$ 映射到 N 维的潜在状态 $h(t) \in R^N$ ，然后再将它们投射回一维输出信号 $y(t)$ 。其数学公式表达为

$$\begin{cases} h'(t) = Ah(t) + Bx(t) \\ y(t) = Ch(t) + Dx(t) \end{cases} \tag{1}$$

，其中 $A \in R^{N\times N}$ 是状态转移矩阵，$B \in R^{N\times 1}$ 是从输入到状态的映射矩阵，$C \in R^{N\times 1}$ 是状态到输出的矩阵，而 $D \in R$ 是输入到输出的参数。

Mamba 引入了选择性状态空间扫描（S6）机制，该机制结合了一个可学习的时间尺度参数 Δ ，该参数可以动态适应输入 (**?**) 。此设计克服了传统 SSM 的线性时不变（LTI）限制并增强了全局建模能力。具体来说，Δ 通过在每个输入位置缩放状态转换来控制离散时间动态，使模型可以调整其内部状态更新的快慢。因此，模型可以通过较小的 Δ 快速响应突然变化，或者通过较大的 Δ 在更长的上下文中整合信息，从而实现对时间分辨率的细粒度控制。模型使用零阶保

持（ZOH）进行离散化，具体如下所述：

$$\bar{A} = \exp(\Delta A), \quad \bar{B} = (\Delta A)^{-1}(\exp(\Delta A) - I)\Delta B \tag{2}$$

$$h_t = \overline{A}h_{t-1} + \overline{B}x_t, \quad y_t = Ch_t + Dx_t \tag{3}$$

其中 ΔA 和 ΔB 分别表示矩阵 A 和 B 与标量 Δ 的乘积，该标量用于缩放连续时间状态变化和输入；$\bar{A}$ 是离散状态转移矩阵，$\bar{B}$ 是离散化的输入矩阵，而 I 是单位矩阵。在方程 (1) 中，h_t 表示离散时间步 t 的隐藏状态，x_t 和 y_t 分别是输入和输出。我们提出的 CMM 模块建立在此 mamba 架构之上。

3.2 MCFNet 概述

在本节中，我们介绍了运动提示融合网络（MCFNet），如图 **??** 所示，该网络以 RGB 图像和相应的事件流作为输入，实现精确的像素级时空对齐和跨模态的自适应特征融合。如图 **??** 所示，ECM 估计光流场，通过扭曲操作实现事件流的时间对齐，生成用于检测的高质量事件表示。随后，使用两个并行分支进行特定模态的特征提取，分别独立处理 RGB 图像和事件帧。每个分支采用来自 YOLOX 框架的 CSPDarkNet 骨干网 (**?**) ，生成多尺度特征图。为了解决现代成像系统中 RGB 和事件传感器之间的分辨率不匹配问题，我们引入了 EDUM，该模块基于事件特征的空间分布自适应地细化上采样内核。这一设计促进了高分辨率、高保真事件表示的生成。接下来，CMM 融合模块在三个层次上融合来自两种模态的特征，自适应地整合互补信息，同时保留全局上下文线索。最后，融合后的多尺度特征通过 FPN+PANet 框架 (**??**) 进行处理，随后采用基于 YOLOX 的检测头进行目标分类和边界框回归。

由于 RGB 与事件相机在时间分辨率上的差异，这两种模态之间存在时间不对齐的问题，这对精确的跨模态融合和区分性特征的提取造成了不利影响。为了解决这个问题，运动补偿方法估计事件流的光流以将事件映射到对应 RGB 帧的时间戳，从而实现跨模态的时间对齐。然而，这些方法依赖于光照一致性和线性运动的假设，而在高度动态的交通场景中，这些假设常常被打破，导致光流估计和时间对齐不准确。为了克服这些限制，我们提出了一种端到端的架构，该架构集成了基于光流的 ECM 与目标检测网络。这一架构促进了光流估计和目标检测任务的联合优化。通过将 ECM 的输出输入到检测网络并利用统一的反向传播机制，目标检测的监督信号

可以直接指导光流的学习，从而使模型能够捕捉超越传统恒定光照和线性运动假设的场景特征。最终，生成的事件表示更能符合目标检测的需求。

具体来说，我们提取了两个连续图像帧曝光之间的事件流 $e = (x_k, y_k, t_k, p_k)k = 1^N$ ，其中 $t_0 < t_k < t_0 + \Delta t$ ，并且为了与 RGB 采样间隔保持一致，将 Δt 设置为 50 毫秒以定义时间窗口。为了将非结构化的事件流转换为网络可接受的张量形式，我们采用了一种基于体素的事件表示方法 (**?**) 。具体来说，我们将事件流 e 映射到一个三维体素网格 $\mathrm{E} \in \mathbb{R}^{B \times H \times W}$ ，其中 B 是时间分箱的数量。根据 **?** 中的设置，我们设置 $B = 5$ 。事件根据它们的时间戳分配到箱中，我们首先相应地对事件时间戳进行归一化：

$$t_k^* = \frac{t_k - t_0}{t_{-1} - t_0}, t_k^* \in [0, 1] \tag{4}$$

其中 t_0 和 t_{-1} 分别代表当前事件序列中第一个和最后一个事件的时间戳。对于每个分区，每个事件（索引为 k ）根据以下公式将其极性 p_k 分配给两个最近的时间箱，权重由三角形核函数确定：

$$E(x_k, y_k, t) = \sum_k p_i \times \kappa \left(t - t_k^*(B-1)\right), \kappa(a) = \max(0, 1 - |a|) \tag{5}$$

ECM 为每个像素预测水平和垂直光流，并在此间隔内扭曲事件以与参考时间戳 t_{ref} 对齐，从而实现跨模态的时间对齐：

$$\begin{aligned} x'_k &= x_k + (t_{\text{ref}} - t_k) \times u(x_k, y_k) \\ y'_k &= y_k + (t_{\text{ref}} - t_k) \times v(x_k, y_k) \end{aligned} \tag{6}$$

然后，我们对补偿后的事件流 $e' = (x'_k, y'_k, t_k, p_k)_{k=1}^N$ 应用双线性插值，将事件映射到其最近的像素位置，使用周围的四个邻居并基于极性进行加权求和并进行可视化，最终获得 Image of Warping Event (IWE) $\in \mathbb{R}^{H \times W \times 3}$ ：

$$\text{IWE} = \sum_k p_k \cdot \max\left(0, 1 - \left|e'_k(x, y) - e_k(x, y)\right|\right) \tag{7}$$

损失函数。在 ECM 中，我们选择以自监督方式进行训练，并使用对比度最大化框架损失来进行运动补偿。根据 **?** 中的方法，我们使用双线性插值为每个极性 p' 下的每个像素生成一个平

均时间戳图像。我们通过最小化前向和后向变形事件图像之间的平方差之和来优化损失函数：

$$\begin{aligned} T_p^*(x;(u,v)|t_{\text{ref}}) =& \frac{\sum_j \kappa(x-x'_j)\kappa(y-y'_j)t_j^*}{\sum_j \kappa(x-x'_j)\kappa(y-y'_j)+\epsilon} \\ j = \{i \mid p_i =& p'\}, \quad p' \in \{+,-\}, \quad \epsilon \to 0 \end{aligned} \tag{8}$$

其中 (u,v) 是估计的光流，(x'_j, y'_j) 是变形事件坐标，t_j^* 是归一化时间戳，$\kappa(\cdot)$ 表示双线性插值核。然后定义对比损失为

$$\begin{aligned} \mathcal{L}_{\text{contrast}}(t_{\text{ref}}^*) &= \sum_x \left[T_+(x;(u,v)|t_{\text{ref}}^*)^2 + T_-(x;(u,v)|t_{\text{ref}}^*)^2\right] \\ \mathcal{L}_{\text{contrast}} &= \mathcal{L}_{\text{contrast}}(1) + \mathcal{L}_{\text{contrast}}(0) \end{aligned} \tag{9}$$

为了规范光流估计，我们引入一种平滑损失，惩罚流场 (**?**) 中的不连续性。对于每个像素 (x,y)，其邻居表示为 $\mathcal{N}(x,y)$，流梯度应用稳健的 Charbonnier 惩罚函数 $\rho(\cdot)$：

$$\begin{aligned} l_{\text{smoothness}} =& \sum_{x,y} \sum_{i,j \in \mathcal{N}(x,y)} \rho(u(x,y)-u(i,j)) + \rho(v(x,y)-v(i,j)) \\ \rho(x) =& \sqrt{x^2+\varepsilon^2} \end{aligned} \tag{10}$$

总损失是对比损失和平滑损失的加权和：

$$\mathcal{L}_{\text{ECM}} = \mathcal{L}_{\text{contrast}} + \lambda_1 \mathcal{L}_{\text{smooth}} \tag{11}$$

训练策略。得益于 ECM 的快速收敛和强泛化能力，我们首先在 DSEC-Det 数据集的一个代表性子集中联合训练网络。然后使用训练好的模型为整个数据集生成确定性的 IWE，这些 IWE 作为输入用于训练 MCFNet 的目标检测组件。该策略确保了高检测性能，同时显著降低了训练成本。为了初始化网络，我们使用了来自 **?** 和 **?** 的官方预训练权重。

由于 RGB 相机和事件相机在分辨率和视野上的差异，它们的输出会表现出空间不对齐。为了补偿这种空间差异，现有的融合方法通常在输入阶段将 RGB 帧下采样以匹配事件数据的分辨率，这不可避免地导致 RGB 图像中的高频细节丢失和信息内容压缩，最终降低感知性能。为了解决这个问题，我们提出了 EDUM，它在模态融合之前对事件特征进行上采样，以充分利用高

质量的 RGB 信息。然而，传统的上采样方法通常依赖于固定参数的转置卷积来增强空间分辨率，忽视了不同场景中像素分布的变化。这通常导致细节模糊或出现视觉伪影。相比之下，我们的 EDUM 根据输入特征自适应地调整上采样权重，从而实现更细致和上下文感知的分辨率增强。

具体而言，EDUM 首先对输入事件特征 $F_e^{3\_\text{low}} \in R^{B\times C\times H\times W}$ 应用全局平均池化，并通过 1×1 卷积实现顺序交互。随后将学习到的权重分配给转置卷积核 $W \in R^{B\times C_{\text{in}}\times C_{\text{out}}\times 3\times 3}$ ，对低分辨率事件特征执行转置卷积上采样，得到高分辨率事件特征 $F_e' \in R^{B\times C\times 2H\times 2W}$ 。其公式为:

$$D = GAP(F_e^{3\_\text{low}}) \tag{12}$$

$$W_f = W \otimes \text{Conv}_{1\times 1}(D) \tag{13}$$

$$F_e' = \text{DeConv}(W_f, F_e^{3\_\text{low}}) \tag{14}$$

其中 GAP 表示全局平均池化，DeConv 表示反卷积操作。

同时，考虑到事件相机由于对暗电流和光电流的敏感性而不可避免地产生噪声，我们利用从主干网络 $F_r^3 \in R^{B\times C\times 2H\times 2W}$ 的第三层的高分辨率 RGB 特征中导出的空间注意力图来抑制上采样过程中的噪声放大。我们利用 RGB 特征的平滑性来增强向上采样的事件特征，最终获得增强的高分辨率事件特征 $F_e^3 \in R^{B\times C\times 2H\times 2W}$ 。其公式为

$$F_e^3 = F_e' + F_e' \otimes \sigma(\text{Concat}(\text{AveP}(F_r^3), \text{MaxP}(F_r^3)) \tag{15}$$

，其中 AveP 是平均池化层，MaxP 是最大池化层，而 σ 是 sigmoid 函数。

3.3 跨模态曼巴融合模块

为了在跨模态交互中同时感知最有价值的互补信息，并根据场景特征自适应地利用事件动态或 RGB 纹理特征进行互补融合，我们设计了具有跨模态交错扫描机制的 CMM。这种机制同时执行跨模态特征交互和捕捉全局信息，选择性地保留显著特征，同时过滤冗余信息。

具体而言，我们选择在主干网络的第三、第四和第五层融合这两种模态，并将融合后的特征输入到 FPN+PANet。因此，我们首先通过线性变换将 CMM 的两个输入模态特征

$F_e^{i=3,4,5} \in R^{C\times H\times W}$ 和 $F_r^{i=3,4,5} \in R^{C\times H\times W}$ 投射到潜在空间，其中 i 代表主干网络第 i 层的特征。接着，通过缩放因子调整模态特征的大小和偏移，以适应不同的模态特征分布，实现不同模态间的特征域对齐和分布归一化。

$$\begin{cases} Z_e^i = F_e^i \odot r_e^i + \beta_e^i \\ Z_r^i = F_r^i \odot r_r^i + \beta_r^i \end{cases} \tag{16}$$

这里，r_e^i/r_r^i 和 β_e^i/β_r^i 分别表示缩放和偏移因子，而 $\odot$ 表示按元素进行的乘法运算。

随后，将 Z_e^i 和 Z_r^i 沿着 W 维度进行串联，以获得特征级的细粒度融合特征 $Z_f^i \in \mathbb{R}^{C\times H\times 2W}$ 。此串联操作为融合提供了空间和结构的基础，使得双方信息在统一的表示空间中进行交互。

$$Z_f^i = \text{Cross\_concat}(Z_e^i, Z_r^i, \text{dim} = W) \tag{17}$$

混合特征 Z_f^i 在 H 和 W 维度上被展平，然后输入到 $SSMs$ ，使每种模态能够基于来自另一种模态的信息有选择地提取相关特征。通过这两种模态之间的全局空间交互，生成了一个 WeightMap，以反映它们的互补关系，指导动态事件特性或 RGB 的丰富语义特性的自适应利用。此外，深度可分离卷积（DWConv）和层归一化的集成进一步提高了计算效率，同时增强了建模跨模态相关性的能力。结果，生成了增强的混合特征 ${Z'}_f^i$ 。公式如下：

$$\text{WeightMap} = \text{SSM}(\overline{\boldsymbol{A}}, \overline{\boldsymbol{B}}, C)(\boldsymbol{Z'}_f^i) \tag{18}$$

$${Z'}_f^i = \text{LN}(\text{Liner}(\text{WeightMap} \odot Z_f)) \tag{19}$$

，其中 $\overline{A}$ 、$\overline{B}$ 和 C 在公式（1 ）和（2 ）中定义，并分别表示离散化的状态转移矩阵、输入到状态的映射矩阵和状态到输出的矩阵。最后，特征 ${Z'}_f^i$ 被解耦回两种模态，生成增强的模态特征 ${Z'}_r^i$ 和 ${Z'}_e^i$ ，然后通过残差连接与原始特征 Z_e^i 和 Z_r^i 融合，以分别获得 $\widetilde{F^i}_e$ 和 $\widetilde{F^i}_r$ 。这种方法有效地增强了模态互补性，同时保留了原始特征信息。公式如下：

$${Z'}_e^i, {Z'}_r^i = \text{decouple}({Z'}_f^i) \tag{20}$$

$$\widetilde{F^i}_e = F_e^i + Z'^i_e \tag{21}$$

$$\widetilde{F^i}_r = F_r^i + Z'^i_r \tag{22}$$

然后，我们添加并融合来自两种模态的自适应增强特征，以获得 f_{out}^i ，随后将其输入到 FPN+PANet 中进行多尺度语义增强和特征整合。

4 实验

本节首先描述所使用的数据集以及实验设置。随后，展示定量和定性结果以证明我们方法的有效性。最后，我们对网络中每个模块进行消融研究。

在实验中，我们验证了我们的模型在两个真实场景数据集上的有效性，即 DSEC-Det (**?**) 和 PKU-DAVIS-SOD (**?**) 。

DSEC-Det。该数据集包含使用 RGB 相机（1440×1080）和高分辨率单色事件相机（640×480）拍摄的宽基线立体数据。它包括 53 个序列，其中 39 个用于训练，14 个用于测试，总计 6.39× 10^4 帧 (**?**) 。

DSEC-Det 数据集有多个版本的标注。其中，由 DAGr (**?**) 和 FPN-fusion (**?**) 使用的自动标注方法导致标注质量下降，而 SFNet (**?**) 提供了更全面和准确的人工标注标签。因此，我们采用 SFNet (**?**) 的标注信息，并遵循其数据划分策略以保持公平的比较。

PKU-DAVIS-SOD。由 DAVIS346 事件相机收集的 PKU-DAVIS-SOD 数据集提供了空间对齐的 RGB 帧和事件流（346×260）数据，训练集包含 6.713× 10^5 标签，测试集包含 2.141× 10^5 标签。每一个子集进一步划分为三种典型场景（正常、运动模糊和低光）(**?**) 。

实现细节。我们使用 Adam 优化器训练 MCFNet，学习率为 5× e^{-4} ，批量大小为 2。网络架构采用 YOLOX (**?**) 的 CSPDarkNet 作为 RGB 主干和事件主干。训练流程为 RGB 输入实现了标准的 YOLO 数据增强方法，包括拼接和混合技术。

由于两个数据集中的两种模态分辨率不一致，因此采用了不同的配置。对于 PKU-DAVIS-SOD 数据集的实验，由于 RGB 和事件数据具有相同的分辨率，我们的方法在实现时不使用 EDUM 模块。在比较实验中，由于 EvImHandNet (**?**) 最初是为手部网格重建设计的，我们仅用其融合模块 CFM 替换我们的 CMM 以进行公平比较，同时保留所有其他架构组件。对于其他方

法，我们遵循它们的原始训练策略以确保公平比较。

评价指标。我们采用 COCO 指标 (**?**) 来评估目标检测的准确性，包括 IOU 阈值为 50 % 的 mAP50 和在 50 % 与 95 % 之间的 IOU 值平均的 mAP。为了评估模型效率，我们还报告参数、FLOPs 和运行时间。

为了评估我们方法的优越性，我们将 MCFNet 与最先进（SOTA）的目标检测方法进行比较。此外，为了验证我们的事件表示策略的有效性，我们针对领先的事件表示和运动补偿技术进行比较实验。

与当前最佳（SOTA）目标检测方法的比较。我们将自己提出的方法与近期的 SOTA 目标检测方法进行了比较，包括两个基于事件的方法：RVT (**?**) 和 SAST (**?**) ；三个基于 RGB 的方法：YOLOv11 (**?**) 、YOLOv12 (**?**) 和 Mamba-YOLO (**?**) ；以及四个基于融合的方法：FPN-fusion (**?**) 、SODFormer (**?**) 、EOLO (**?**) 和 SFNet (**?**) 。此外，为了与其他融合方法进行比较，我们用 EvImHandNet (**?**) 中的基于 CNN 的融合方法 CFM 和 CAFR (**?**) 中的基于变压器的融合模块替换了我们的 CMM。如表格 1 所示，我们的 MCFNet 在两个数据集上均优于现有的目标检测方法。具体来说，在类别不平衡的 DSEC-Det 数据集上，我们的方法在 mAP 和 mAP50 上分别比第二好的方法高出 1.7 % 和 7.4 %。而在类别平衡的 DSEC-Det 数据集上，我们的方法显示出持续的优越性，mAP 和 mAP50 分别提高了 1.7 % 和 3.3 %。此外，在比较不同方法时，虽然我们的 MCFNet 在检测准确性上表现出领先性能，但它的计算成本更高，达到 227.8 G FLOPs，与轻量级单模态模型如 Mamba-YOLO（49.6G）(**?**) 和 YOLOv11（35.7 ms）(**?**) 相比，其推理时间为 47.3 ms。尽管如此，MCFNet 在保持高精度的同时实现了约 21 FPS 的推理速度，显示出了初步的实时处理能力。此外，与其他多模态方法相比（例如，SFNet，其推理时间为 44.8 ms 但准确性显著较低），我们的方法在性能和效率之间表现出优势的平衡。

此外，在 PKU-DAVIS-SOD 数据集上，我们的方法在 mAP 和 mAP50 方面分别比 SODFormer (**?**) 提高了 11.9 % 和 11.4 %。如表 2 所示，我们的模型在各种挑战条件下表现出了一致的性能优势，在普通、低光和运动模糊子集上的 mAP50 分别比 SODFormer (**?**) 高出 10.2 %、9.6 % 和 11.8 %。这些结果表明，我们的方法通过精确的时空对齐和跨模态的自适应融合，特别是在具有复杂光照条件和运动模糊的挑战交通场景中，实现了更加稳健的目标检测性能。

PKU-DAVIS-SOD 上的跨场景评估。为了进一步证明我们提出方法的鲁棒性，我们进行了跨场景验证。具体来说，我们将一个仅在 DSEC-Det 数据集上训练的模型转移到了

Table 1 在不同数据集上，SOTA 目标检测方法的性能对比（mAP50/mAP）。最佳性能和次佳性能分别用红色加粗和蓝色加粗标出。注意：M 表示百万参数。

	Method	Pub. & Year	DSEC-Det			PKU-DAVIS-SOD				Parameter
			Class-balanced mAP50/mAP	Class-imbalanced mAP50/mAP	FLOPs	Runtime	mAP50/ mAP	FLOPs	Runtime	
RGB	YOLOv11	arXiv'24	84.5/59.2	59.1/40.1	102.3 G	35.7 ms	58.0/30.9	44.2 G	15.6 ms	20.1 M
	Mamba-YOLO	AAAI'25	83.4/55.6	53.9/34.7	49.6 G	58.7 ms	57.1/29.8	21.6 G	21.0 ms	21.8 M
	YOLOv12	arXiv'25	86.8/58.8	55.0/34.6	90.1 G	46.5 ms	60.2/32.4	12.22 G	13.4 ms	19.6 M
Event	RVT	CVPR'23	51.1/26.6	25.1/12.9	19.6 G	11.9 ms	50.3/25.6	6.5 G	7.1 ms	18.5 M
	SAST	CVPR'24	53.8/39.3	24.3/12.1	18.5 G	18.8 ms	48.7/24.5	6.2 G	16.7 ms	18.5 M
RGB-Event	FPN-fusion	ICRA'22	56.8/30.7	36.9/19.7	89.6 G	30.8 ms	36.6/19.5	49.7 G	24.0 ms	65.6 M
	SODFormer	TPAMI'22	—	—	—	—	50.4/20.7	62.5 G	39.7 ms	82.0 M
	EOLO	ICRA'24	65.1/37.8	33.9/19.6	13.7 G	330.2 ms	47.2/22.0	8.9 G	326.4 ms	21.5 M
	SFNet	ITS'24	80.0/50.9	51.4/30.4	209 G	44.8 ms	59.6/31.9	135.9 G	42.3 ms	57.5 M
	Our pipeline + CFM	CVPR'24	86.4/53.1	60.0 /35.3	256.0 G	58.9 ms	60.0/30.8	80.93 G	17.2 ms	46.0 M
	Our pipeline + CAFR	ECCV'24	87.4 /54.8	59.9/34.6	208.5 G	50.4 ms	61.2/31.8	66.51 G	18.9 ms	43.6 M
	MCFNet (this study)	—	90.7/60.9	67.4/41.8	227.8 G	47.3 ms	61.8/32.6	72.68 G	18.8 ms	52.1 M

Table 2 在 PKU-DAVIS-SOD 数据集的不同场景子集上与 SOTA 方法的性能比较。最佳和次佳表现分别用红色加粗和蓝色加粗标出。

Method	Pub & Year	PKU-DAVIS-SOD					
		Normal		Motion_blur		Low_light	
		mAP	mAP50	mAP	mAP50	mAP	mAP50
SODFormer	TPAMI'22	24.1	56.9	18.3	43.2	12.2	37.4
SFNet	ITS'24	32.3	62.4	23.1	46.7	17.6	41.2
Our pipeline + CFM	CVPR'24	33.6	65.1	25.0	50.9	21.2	48.6
Our pipeline + CAFR	ECCV' 24	34.4	66.6	26.0	51.7	21.6	49.9
MCFNet (this study)	—	34.8+0.4	67.1+0.5	26.0	52.8+1.1	21.8+0.2	49.2–0.7

PKU-DAVIS-SOD 数据集。由于两个数据集之间对象类别数量的不同，我们仅对检测头进行了轻微微调。正如表 3 所示，尽管数据集之间的数据分布和类别定义存在差异，我们的模型仍然优于其他方法，显示出跨域的强泛化性和鲁棒性。

Table 3 跨场景性能与现有最优方法在 PKU-DAVIS-SOD 数据集上的比较。最佳性能以红色加粗标记。

Method	Pub & Year	PKU-DAVIS-SOD	
		mAP	mAP50
SFNet	ITS'24	19.7	41.9
Our pipeline + CFM	CVPR'24	18.5	39.5
MCFNet(this study)	—	24.7	49.7

与现有技术（SOTA）事件帧积累方法和运动补偿方法的比较。为了进一步证明我们 ECM 表示方法的优越性，我们将其与四种事件帧积累方法进行比较：Timestamp (**?**) 、DiST (**?**) 、Voxel (**?**) 和 TAF (**?**) 。此外，我们评估了五种运动补偿方法：（i）基于模型的（MB）：CMax (**?**)

、ST-PPP (**?**) 和 MCM (**?**) ，（ii）自监督学习（SSL）：性能最佳的 ConGru-EV-FlowNet (**??**) 配置。对于运动补偿方法，生成的 IWE 被用作我们检测网络的输入。我们保持检测网络、上采样模块和融合模块不变，只更改事件模式的输入。结果在表格 4 中报告。

Table 4 与最新技术（SOTA）事件表示方法和运动补偿方法的性能比较。最佳和次佳性能分别用红色粗体和蓝色粗体标出。MB: 基于模型; SSL: 自监督学习。

	Method	Pub. & Year	Inference time	DSEC-Det (Class-balanced)		DSEC-Det (Class-imbalanced)		PKU-DAVIS-SOD	
				mAP	mAP50	mAP	mAP50	mAP	mAP50
MB	CMax	CVPR'18	2.42s	60.0	89.6	41.5	65.8	31.1	60.3
MB	ST-PPP	ICCV'21	2.48 s	60.3	89.9	41.4	67.0	31.2	60.2
MB	MCM	ECCV'22	65 s	60.1	90.0	41.3	66.0	31.5	60.0
SSL	ConvGRU-EV-FlowNet	NeurIPS'21	0.48 s	60.1	89.9	41.3	66.0	32.4	62.0
SSL	Federico et al.	ICCV'23	0.29 s	60.2	89.9	40.1	63.8	32.4	62.3
	TAF	IEEE Trans. TIM'23	—	60.2	89.6	39.7	63.3	31.0	59.9
	Timestamp	ICIP'16	—	59.7	89.3	40.5	65.6	31.7	60.9
	Voxel	CVPR'19	—	59.4	89.0	41.1	65.7	30.8	59.6
	DiST	ICCV'21	—	60.0	89.9	41.5	65.6	31.4	60.7
	ECM(this study)	—	0.03 s	60.9	90.7	41.8	67.4	32.6	61.8

在所有事件帧积累方法中，我们的 ECM 通过预测运动矢量来对齐时间戳，实现了最佳性能，从而获得精确的时间对齐和高质量的帧，具有锐利的边缘和降低的噪声。例如，在类别不平衡的 DSEC-Det 数据集上，我们的事件表示在 mAP50 上比体素 (**?**) 高出 1.7 %。

与基于模型和自监督运动补偿方法相比，我们的 ECM 也表现出优越的性能。例如，在类别不平衡的 DSEC-Det 数据集上，我们的事件表示在 mAP50 上比 ConvGRU-EV-FlowNet (**?**) 高出 1.4 %。这证明了在与检测网络联合训练的端到端架构中，我们的 ECM 学习的场景特征可以克服亮度恒定和线性运动假设的限制，生成更适合各种复杂场景的事件表示，并有利于目标检测任务。相反，不同的运动状态和频繁的光照变化对光流估计产生不利影响，导致 IWE 中出现伪影和噪声生成，从而影响检测性能，如图 2 所示。

4.1 定性结果

与 SOTA 检测方法的比较。我们将 MCFNet 与两个领先的 SOTA 方法进行比较：我们的 pipeline+CFM (**?**) 和 SFNet (**?**) 。如图 3 所示，前 3 列显示的是非均匀曝光的场景，第 4 和 5 列显示的是低光多物体场景，最后一列是运动模糊场景。我们的方法通过在场景中同时感知来自两种模态的全局信息，实现了跨模态的精确时空对齐和卓越的自适应互补融合，从而在复杂的

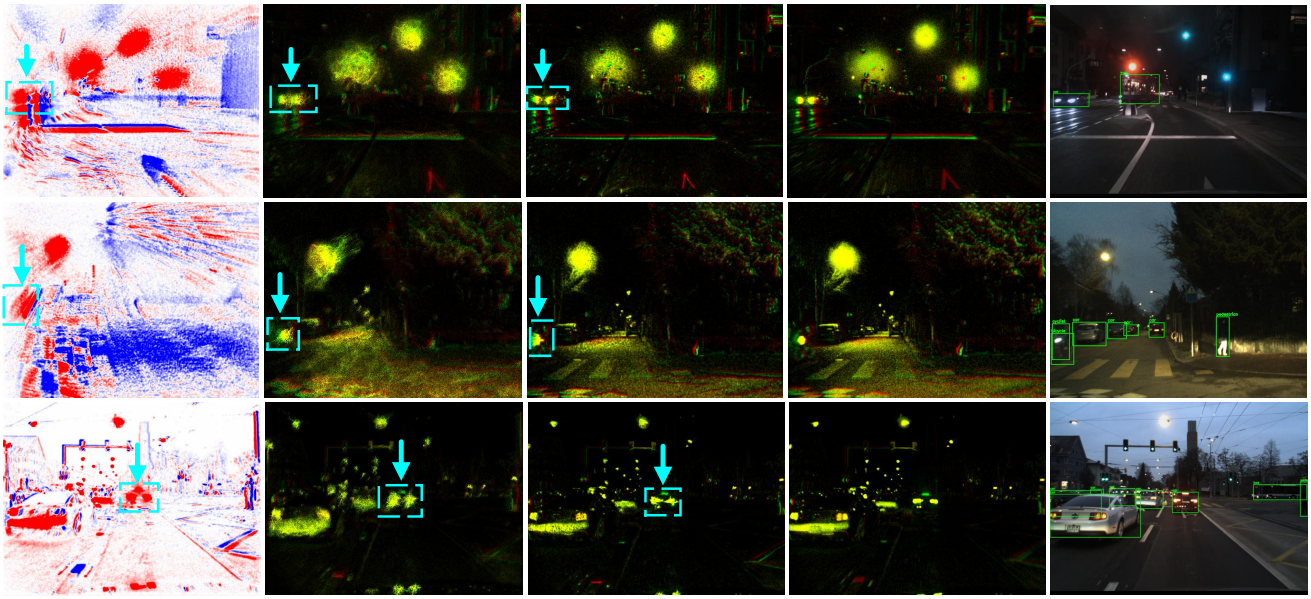

Fig. 2 在 DSEC-Det 数据集上与 SOTA 运动补偿事件表征方法的定性比较：(a) ST-PPP, (b) ConvGRU-Ev-FlowNet, (c) Ev-FlowNet, (d) ECM（本研究），和 (e) GT。

光照条件下相比现有的 SOTA 方法具有更强的鲁棒性，实现了成功的检测。

与 SOTA 事件表示方法的比较。我们将从 ECM 生成的事件表示与两个领先的 SOTA 运动补偿方法 ST-PPP、ConvGRU-EV-FlowNet (**??**) 及 EvFlow 在没有端到端联合训练情况下生成的表示进行比较，如图 2 所示。第一行展示了在夜间条件下快速移动车辆及其自发光亮的场景。第二行展示了一种情景：自车车灯照亮目标车辆的同时，目标车辆在运动过程中发出自己的光亮。在第三个场景中，对连续帧的检查揭示了中心车辆的刹车灯在当前帧中刚刚开启。这三个场景共同展示了快速光照变化或复杂运动模式的环境。如图 2 的第二到第四列所示，现有的运动补偿方法通常在此类条件下失效，因为环境变化违反了它们关于亮度恒定性和线性运动的基本假设，导致伪影、噪声和目标变形。相比之下，我们的 ECM 生成的表示在优化来自目标检测任务的指导下，通过利用事件运动特性保持了目标的真实性。

在本节中，我们展示了在 DSEC-Det 数据集上的消融研究结果，以验证我们提出的 MCFNet 中每个组件的有效性。

我们 MCFNet 组件的贡献。我们采用双流 YOLOX 架构结合体素事件表示和简单的特征加法融合作为基线，以此进行全面的消融研究，分析 MCFNet 中每个组件的性能。表 5 显示了我们方法中不同模块组合的结果。可以观察到，每个提出的模块对 MCFNet 的检测性能都有积极的

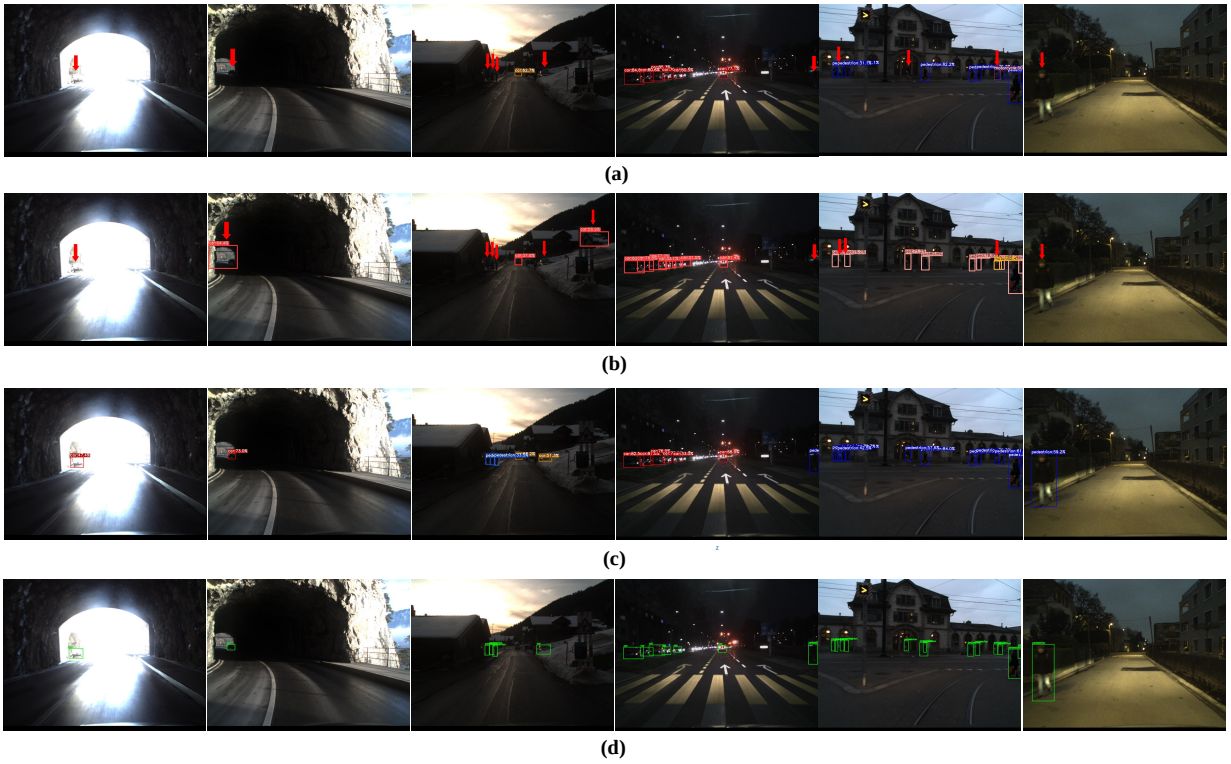

Fig. 3 与两种主要的 SOTA 检测方法在类别不平衡的 DSEC-Det 数据集上的定性比较: (a) EvImHnadNet; (b) SFNet; (c) MCFNet(本研究); 和 (d) GT。我们使用 红色 箭头标记失败的案例。

贡献。特别是在类别不平衡的 DSEC-Det 数据集上，我们的模型相较于基线实现了显著的提升，mAP 和 mAP50 指标分别增加了 8.1 % 和 12 %，证明了每个模块在动态交通场景下的有效性。

Table 5 对提出组件的消融研究。红色粗体 表示相对于基线的改进。

ECM	EDUM	CMM	DSEC-Det class-balanced		DSEC-Det class-imbalanced	
			mAP	mAP50	mAP	mAP50
YOLOX(voxel)			53.7	82.5	33.7	55.4
✓	—	—	53.8+0.1	82.7+0.2	34.1+0.4	55.6+0.2
✓	—	✓	55.9+2.2	85.9+3.4	34.0+0.3	56.9+1.5
✓	✓	—	60.0+6.3	89.4+6.9	40.4+6.7	64.7+9.3
✓	✓	✓	60.9+7.2	90.7+8.2	41.8+8.1	67.4+12.0

应用 CMM 之前和之后的特征图如图 **??** 所示。在图 **??** 中描绘的前两行夜间场景中可以观察到，由于事件模态具有高动态范围，与 RGB 模态相比，在低光条件下展现出更稳定的成像能力，从而能够更有效地捕捉目标信息。相反，在第三和第四行中显示的白天场景中，当自车与目标物体同时移动时，由于相对静止或相对运动较少，事件变得稀疏。同时，在良好光照条件下的 RGB 图像包含丰富的色彩、纹理和语义信息，促进更有效的目标捕捉。在 CMM 融合后，

模型同时建模来自两种模态的全局信息，以精确识别模态间的互补关系，通过自适应融合选择性地强调关键信息，实现模态互补优势，全面捕捉场景中的所有目标信息。

联锁训练 ECM 的益处。我们在 DSEC-Det (**?**) 上验证了我们的联锁训练策略的有效性。两阶段训练策略指的是一种配置，在这种配置中，光流网络参数在检测器训练过程中保持固定，不参与反向传播过程。我们的 ECM 采用端到端的联锁网络训练，其中事件校正优化由检测结果引导。正如表格 6 所示，我们的 ECM 学习了超越亮度恒定性和线性运动假设的场景特征，实现了更准确的时间对齐，并生成了更符合对象检测需求的事件表示，从而提供了卓越的性能。

Table 6 将未进行联合训练的 EvFlowNet 结果与 ECM 进行比较。

Method	Training strategy	DSEC-Det class-imbalanced	
		mAP	mAP50
EV-FlowNet	Two-stage	41.1	65.4
ECM (this study)	Joint training	41.8	67.4

上采样方法的选择。为了验证我们动态上采样方法 EDUM 的有效性，我们进行了与常见上采样方法的对比实验，包括像素重排和转置卷积。如表 7 所示，基于输入动态调整权重使模型能够在不同场景中学习像素分布的变化，从而实现更精确的空间对齐和增强的检测性能。

Table 7 不同上采样方法的性能比较。

Upsampling Method	DSEC-Det (class-imbalanced)	
	mAP	mAP50
Pixel Shuffle	40.7	65.2
Transposed conv (3×3)	41.5	65.8
EDUM (this study)	41.8	67.4

在本研究中，我们提出了一种运动提示融合网络（MCFNet）用于动态交通场景中的鲁棒物体检测。通过受物体检测任务的有效引导，ECM 克服了传统光流估计任务中关于恒定光照和线性运动的假设限制，从而获得高质量的事件帧。EDUM 基于特征空间分布动态上采样事件特征，在最大化利用高质量信息的同时，对齐两种模态的分辨率。CMM 通过同时执行模态间特征交互和全局信息提取，增强了模型在不同场景中对主要模态特征的感知和选择准确性，从而实现自适应融合。实验表明，我们的 MCFNet 在各种复杂和动态交通场景中显著优于现有方法。

尽管该方法在检测精度方面表现出色，但事件数据的时空对齐和多模态跨融合机制的引入，导致模型结构相对复杂，计算开销较大，这对资源受限设备上的部署构成了挑战。在未来的工作中，我们计划在事件数据上引入稀疏性约束。例如，使用稀疏标记机制来过滤非活动区域，

只在事件发生的地方进行融合，从而减少冗余计算。我们还旨在探索动态标记选择和模型剪枝，以进一步减少计算负载并加速实时应用的推理。

5

作者简介

Zhanwen Liu received the B.S. degree from Northwestern Polytechnical University, Xi'an, China, in 2006, the M.S. and the Ph.D. degrees in traffic information engineering and control from Chang'an University, Xi'an, China, in 2009 and 2014, respectively. She is currently a Professor with School of Information Engineering, Chang' an University. Her research interests include motion perception, behavior prediction, and data-closed-loop autonomous driving testing.

Yujing Sun received the B.S. degree from Chang'an University, Xi'an, China in 2023 and is currently pursuing the M.S. degree in computer science and technology in Chang'an University. Her current research interests include object detection and optical flow estimation based on event cameras and their applications in autonomous driving perception.

Yang Wang received the B.S. degree from Chang'an University, Xi'an, China, in 2016, and the Ph.D. degree in control science and engineering from the University of Science and Technology of China, in 2021. He is currently an Associate Professor with the School of Information Engineering, Chang'an University, Xi'an, China. His central research interests focus on computer vision and multimedia processing.

Nan Yang received the B.S. degree from Chang'an University in Xi'an, China, in 2022, where he is currently working toward the Ph.D. degree in traffic information engineering and control. His current research interests include object detection and multiple-object tracking based on event cameras, and their applications in intelligent vehicle and road infrastructure perception.

Shengbo Eben Li received the M.S. and Ph.D. degrees from Tsinghua University in 2006 and 2009, respectively. Before joining Tsinghua University, he has worked at Stanford University, University of Michigan, and UC Berkeley. His active research interests include intelligent vehicles and driver assistance, deep reinforcement learning, optimal control and estimation, etc.

Xiangmo Zhao received the Ph.D. degree from Chang' an University, Xi' an, China, in 2003. He is currently a Distinguished Professor with the School of Information Engineering, Chang'an University, and also is the President of Xi'an University of Architecture and Technology. He currently serves as the Vice Chairman of the China Society of Automotive Engineers. His current research interests include testing of connected vehicles, automated vehicles, and intelligent transportation systems.