计算机科学
COMPUTER SCIENCE

跨模态噪声过滤的事件相机目标检测算法

胡刚, 梁栋, 黄圣君

引用本文

胡刚, 梁栋, 黄圣君. 跨模态噪声过滤的事件相机目标检测算法[J]. 计算机科学, 2024, 51(11A): 231000013-6.

HU Gang, LIANG Dong, HUANG Shengjun. Event-based Camera Object Detection Algorithm for Cross-modal Noisy Annotations Filtering [J]. Computer Science, 2024, 51(11A): 231000013-6.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.231000013

跨模态噪声过滤的事件相机目标检测算法

胡 刚 梁 栋 黄圣君
南京航空航天大学计算机科学与技术学院 南京 211106
(hugang@nuaa.edu.cn)

摘 要 事件相机具有高时间分辨率、高动态范围和低功耗等特性,通常被用于传统相机应用受限场景(高速度、强光、弱光等)下的目标检测任务中。然而由于事件相机的像素异步性,其输出的事件序列难以进行人工标注,为此现有方法通过RGB图像标记迁移得到事件序列标记。然而,迁移标记中存在大量噪声标记和事件序列中部分目标纹理模糊,导致难以取得理想的模型性能。为了解决此问题,提出了一种跨模态噪声过滤的事件相机目标检测算法。算法利用预训练后的事件相机检测器对开源RGB目标检测数据集进行筛选,得到对训练事件相机检测器最具价值的RGB图像和事件图像一起构成跨模态混合图像,帮助检测器更准确地识别、定位事件图像目标;为了缓解噪声标记对检测器性能的影响,设计了一种多阶段目标检测联合优化策略,单个阶段训练完成时,在全局标记中识别噪声标记,并对噪声标记进行修正后在下一阶段使用。实验结果表明,在1Mpx Detection Dataset上,与基准模型相比,跨模态噪声过滤的事件相机目标检测算法提供了8.35%的模型增益,远优于Co-teaching,O2U-net等噪声标签学习方法,具体地,跨模态混合图像训练、联合优化框架分别提供了6.44%,4.77%的模型增益。

关键词:事件相机;目标检测;噪声标记;跨模态;联合优化
中图分类号 TP391.4

Event-based Camera Object Detection Algorithm for Cross-modal Noisy Annotations Filtering

HU Gang,LIANG Dong and HUANG Shengjun
School of Computer Science and Technology,Nanjing University of Aeronautics and Astronautics,Nanjing 211106,China

Abstract Event-based camera is commonly seen in object detection in limited scenarios for traditional camera applications (high speed,strong light,low light,etc.) due to their high time resolution,high dynamic range and low power consumption. However,the event sequence output of event camera is difficult to be manually labeled due to its pixel asynchronism,so the existing methods obtain event sequence annotations through the migration of RGB image annotations. However,since the migrated annotations have numerous inaccurate bounding boxes and some object textures in event sequence are fuzzy,leading to poor model performance. To address this problem,event-based camera object detection algorithm for cross-modal noisy annotations filtering is proposed. The method uses a pre-trained event-based camera detector to filter open-source RGB object detection datasets and selects RGB images that are most valuable for training the event-based camera detector. These selected RGB images are combined with event images to construct cross-domain mixed images,helping the detector to identify and locate the event image object more accurately. To mitigate the impact of noisy annotations on detector performance,a multi-stage object detection joint optimization strategy is designed. After each stage of training is completed,noisy annotations are identified in the global annotations and are corrected use in the next stage. Experimental results show that,on the 1Mpx Detection Dataset,the robust event-based camera cross-modal object detection method based on noisy annotations provides 8.35% model gain compared to the baseline model,significantly outperforming noise-label learning methods such as Co-teaching and O2U-net. Specifically,cross-modal hybrid images training and joint optimization frameworks offer model gains of 6.44% and 4.77%,respectively.

Keywords Event-based camera,Object detection,Noisy annotations,Cross-modal,Joint optimization

1 引言

与传统相机相比,事件相机具有高时间分辨率(us级)、极高的动态范围(140 dB)、低功耗和高像素带宽(kHz级)等特性。因此,事件相机通常被用于传统相机应用受限的场景(高速度、强光、弱光等),如运动图像去模糊[1]、多曝光图像生成超分辨率图像[2]、基于事件相机的图像重构等[3],并取得了丰硕的研究成果。然而,在应用到目标检测任务中时,由于其输出事件序列具有像素异步性、空间稀疏等特性[4],现有检测网络难以直接适配事件序列,因此迫切需要设计新的方法来处理事件序列,以释放它们在挑战性场景中目标检测的潜力[5]。Sabater等[6]设计了Event Transformer———一种用于高效事件数据处理的稀疏感知解决方案。Wan等[7]和Miao等[8]提出多种事件序列编码方式,将事件序列编码成事件图像(强度帧)后,利用现有目标检测算法实现目标检测。然而,现有的方法大都需要对事件序列进行人工标注,但事件序列人工标注十分具有挑战性:1)事件序列生成事件图像数量多,人工标注繁琐;2)相比RGB图像,事件序列只包含运动目标

通信作者:黄圣君(huangsj@nuaa.edu.cn)

的纹理信息[9]，且在拍摄目标距离事件相机较远或目标运动较慢时，目标纹理模糊，甚至肉眼也难以区分前景目标和背景，导致人工标注得到的标记误差较大。

基于事件序列人工标注困难问题，Perot 等[10]提出了 The 1 Megapixel Automotive Detection Dataset(1Mpx Detection Dataset)数据集，数据集通过将百万像素的事件相机[11]和 RGB 运动相机并排安装在运动目标的刚性支架上，获取场景的 RGB 图像和 DVS 事件序列。随后，使用成熟的商用检测器对 RGB 图像进行标注得到 RGB 图像标记，将标记几何变换(减少两个摄相机之间的视角差)后作为事件序列目标框。然而，由于 RGB 图像通过商用目标检测器标注，存在标注错误，以及 RGB 运动相机和事件相机拍摄原理和拍摄视角之间存在差异，导致较多 RGB 图像目标框几何变换后作为事件序列目标框，难以较好地框住事件序列中的目标。1Mpx Detection Dataset 中不可避免地存在大量标注不准确的目标框，称为噪声标记。噪声标记主要有 3 种：1)多标，当目标静止不动或目标距离事件相机很远时，RGB 图像具有目标框，而事件图像上却无此目标；2)漏标，当目标移动太快，会产生较大的运动模糊，商用检测器检测不出，或者其他原因导致商用检测器漏检，这会导致 RGB 图像不具有目标框，但事件图像上却有此目标；3)目标框漂移，目标移动越快、目标偏离视角中心越远，同一目标在 RGB 图像和事件图像上所在的相对位置差距越大，迁移目标框漂移情况越严重。其中目标框漂移在噪声标记中占比最大，广泛存在。

为了缓解噪声样本对检测器泛化性能带来的恶劣影响，Huang 等[12]提出 O2U-net，通过调整网络的学习率使得网络在欠拟合和过拟合之间反复切换，每次迭代时记录每个样本的损失，平均损失越大越可能是噪声样本。Han 等[13]提出 Co-teaching 算法，假设两个模型 f 和 g，训练过程中对于一个 mini-batch 数据，Co-teaching 在两个模型中相互选择 loss 较小的数据作为干净样本，用于对方模型的参数更新。以上工作通过模型训练识别噪声样本后将噪声样本剔除，来减少噪声样本对模型造成的损害，在噪声比例较低的分类任务上取得了不错的效果。然而，在目标检测任务中，可能出现单张图像上具有多个正常目标框和少量噪声目标框，无法直接剔除含有噪声目标框的图像；在噪声比例高时，也无法剔除所有噪声样本。Tanaka 等[14]提出轮换策略学习框架，模型先梯度下降更新模型参数，再用模型参数做噪声标签校正。这提供了一种新的解决思路，但文章中轮换策略每遍历一次训练集就会对训练集样本标签做一次修改，标签修改过于频繁。因此难以在目标检测任务中使用轮换策略学习框架。Li 等[15]提出带噪声标记的抗噪目标检测 CA-BBC 方法，先通过主干网络提取得到建议框，再通过两个检测头输出之间的差异对建议框进行修改，用修改后的建议框进行检测和梯度更新。CA-BBC 在噪声标记广泛存在时难以生效，且方法只针对两阶段检测器，无法适用于一阶段检测器。

为解决上述问题，并进一步提高带有噪声标记的事件相机目标检测的检测性能，本文提出跨模态噪声过滤的事件相机目标检测算法，算法首先利用带有标记的训练集事件图像预训练(Warm-up)一个事件图像目标检测器，将其作为“筛选器”，筛选开源 RGB 图像中对事件图像训练最有价值的 RGB 图像加入训练集，联合 RGB 图像和事件图像使用 Mosaic 增强[16]，构造跨模态混合图像，进行目标检测训练，要求模型从事件图像局部视图中识别目标，同时在事件图像裁剪区域中添加其他样本的信息，能够进一步增强模型对事件图像目标的定位能力；其次，为了缓解噪声标记对检测器的损害，设计了多阶段目标检测联合优化策略，将训练分为多个阶段，单个阶段训练完成时，在全局标记中识别噪声标记，并对噪声标记进行修正后在下一阶段继续使用。跨模态噪声过滤的事件相机目标检测算法不改变检测器结构，只针对训练数据进行增加和修改，通用性强，可与其它针对噪声标记的方法联合使用。

2 跨模态噪声过滤的事件相机目标检测

本文提出的跨模态噪声过滤的事件相机目标检测算法如图 1 所示，包含两个部分：

1)高质量 RGB 图像筛选与跨模态混合图像训练：基于事件图像预训练(Warm-up)后的检测器筛选高质量 RGB 图像，加入事件图像训练集组成跨模态训练集，联合 RGB 图像和事件图像使用 Mosaic 增强，构造跨模态混合图像，进行目标检测训练。

2)联合优化框架：基于带有噪声标记的训练数据设计了多阶段目标检测联合优化策略，将训练分为多个阶段，单个阶段训练完成时，在全局标记中识别噪声标记，并对噪声标记进行修正后在下一阶段继续使用。

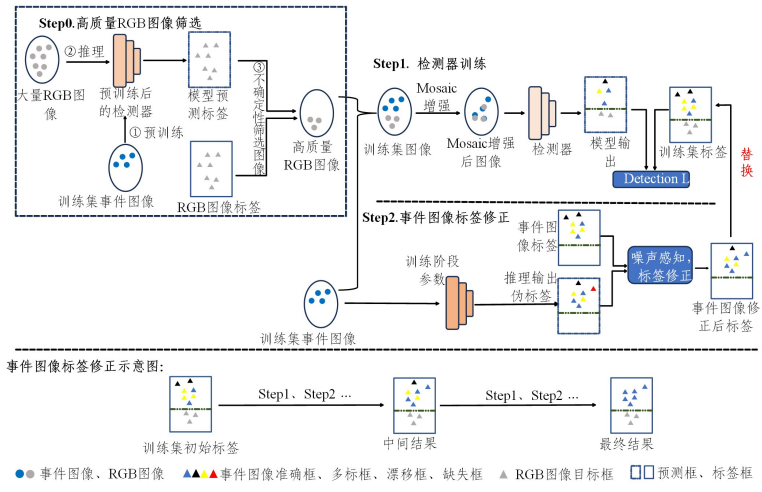


图 1 跨模态噪声过滤的事件相机目标检测算法

Fig. 1 Event-based camera object detection algorithm for cross-modal noisy annotations filtering

2.1　高质量 RGB 图像筛选与跨模态混合图像训练

当目标距离事件相机较远或目标运动较慢时，事件相机拍摄到的目标纹理模糊，难以区分前景目标和背景，导致模型对事件图像前景目标检测能力弱。RGB 图像的目标检测任务已经发展多年，开源了大量高质量目标检测数据集，且 RGB 图像具有丰富的纹理、色彩等信息。引入和事件图像标记类别相同的 RGB 图像可帮助检测器更好地检测事件图像目标。不同 RGB 图像对训练事件图像目标检测器价值不同，因此，研究者们提出了很多方法来选择出最具价值的 RGB 图像。

Liu 等[17]提出主动学习允许机器学习算法选择要学习的数据，以期用更少的标记实例训练获得更好的模型性能。Xie 等[18]指出大多数现有的主动学习工作都遵循一个繁琐的流程，即在每个数据集上重复多次耗时的模型训练和批处理数据选择，故提出一种新的通用的、有效的主动学习方法 GEAL，核心思想是利用在大型数据集上预先训练过的公开模型对数据集进行单次推理，利用中间特征中提取出来的知识聚类，一次性选择所有样本。但利用公开模型对数据进行特征提取，无法提取适用于目标检测的特征，也无法选择出适用于特定数据集的样本。为快速有效地筛选出对事件图像训练帮助大的 RGB 图像，本文提出利用事件图像初步训练后的检测器对所有 RGB 图像进行单次推理，一次性选择推理损失最大的前 M 个样本。

RGB 图像和事件图像的分布不同，域之间存在差异，直接引入 RGB 图像和事件图像一起训练等价于在事件图像域和 RGB 图像域分别进行目标检测训练，难以提升检测器在事件图像域中的检测性能。Pan 等[19]提出了领域自适应方法来减少数据集之间的分布差异。Ganin 等[20]提出域对抗网络 DANN，网络通过训练损失和梯度翻转层后的域判别损失训练，使得主干网络在源域和目标域中提取的特征相似。DANN 在目标检测上的应用如图 2 所示。

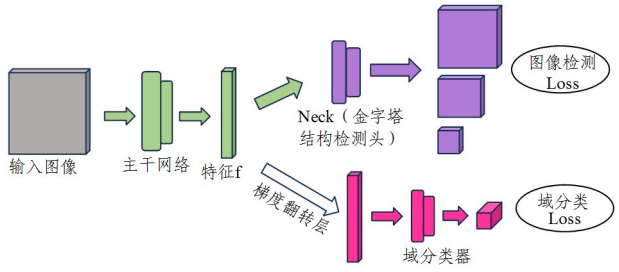


图 2　目标检测域对抗框架

Fig. 2　Object detection domain adversarial framework

目标检测中，对抗域自适应在特征空间上减少了不同数据域之间的分布差异，从而增强了特征的迁移性，但可能牺牲掉特征的判别性[21]。Mosaic 增强的具体做法是，将 4 张（或 9 张）图片进行随机裁剪，再拼接到一张图上作为训练数据，要求模型从局部视图识别对象，同时在裁剪区域中添加其他样本的信息，进一步增强模型的定位能力。为提升模型对事件图像目标的检测能力，本文提出联合 RGB 图像和事件图像使用 Mosaic 增强，构造跨模态混合图像，如图 3 所示。

使用跨模态混合图像进行目标检测，单张混合图像中，在事件图像裁剪区域中添加 RGB 图像信息，要求模型从事件图像局部视图中识别目标，同时增强模型对事件图像目标的定位能力。这既避免了域对抗牺牲特征的判别性，也很大程度上提升了模型在事件图像上的检测性能。本文提出跨模态混合图像训练，如图 1 中 Step0 和 Step1 所示。具体为：

首先，在事件图像训练集上预训练目标检测器；其次，将预训练后的检测器设为推理模式，对 RGB 图像进行推理，得到检测器对 RGB 图像推理的不确定性（损失）；然后，将 RGB 图像按照不确定性的大小降序排序，筛选得到前 M 张 RGB 图像；最后，联合 RGB 图像和事件图像使用 Mosaic 增强，构造跨模态混合图像，利用跨模态混合图像进行目标检测器训练。

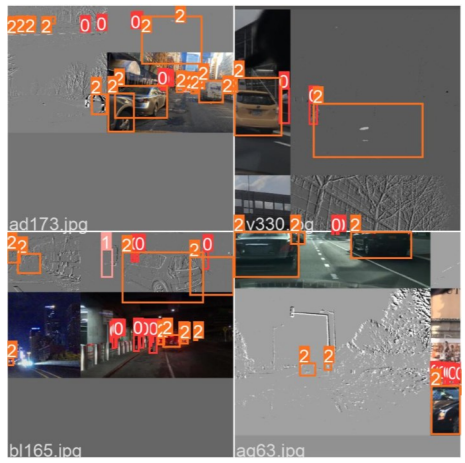


图 3　跨模态混合图像

Fig. 3　Cross-modal hybrid images

2.2　联合优化框架

基于标记迁移得到的事件图像标记存在多标、漏标的情况，目标框漂移情况更是大量存在。多标——事件图像上无此目标，标记中却具有目标框；漏标——事件图像上存在此目标，标记中却无对应目标框；目标框漂移——标记中目标框中存储的位置不是事件图像上对应目标的真实位置，发生了偏移。噪声标记大量存在，且不同类别噪声比例不同、比例未知，O2U-net，CO-teaching 和 CA-BBC 等针对标签噪声的方法难以奏效。Van 等[22]提出，欠拟合是神经网络未能适当地拟合训练集数据分布，过拟合是神经网络很好地拟合了训练集的数据分布，也拟合了噪声样本。Huang 等[12]提出深度网络先拟合简单样本，后拟合困难样本和噪声样本。基于网络拟合理论，本文提出联合优化框架，如图 1 中 Step1 和 Step2 所示，将训练分为多个阶段，单个阶段训练完成时，在全局标记中识别噪声标记，并对噪声标记进行修正后在下一阶段继续使用。详细步骤如下：

1）选择目标检测器模型，加载预训练权重，在训练集上进行训练，每进行 K 步（step）迭代，目标检测器进行一次验证集推理。当目标检测器在验证集上性能从快速增长转变为缓慢增长时停止训练。

2）用停止训练后的目标检测器对训练集中带有标记的事件图像进行推理，得到事件图像的预测框集合 $\hat{y}$，定义预测框集合 $\hat{y}$ 为：

$$\hat{y}=\{box_{\hat{y}} \mid box_{\hat{y}}=(cls,x,y,w,h,conf)\} \tag{1}$$

同时，定义训练集中事件图像带有的标记（目标框集合）y 为：

$$y=\{box_{y} \mid box_{y}=(cls,x,y,w,h)\} \tag{2}$$

其中，$box_{\hat{y}}$ 为预测框，box_{y} 为目标框；cls 表示预测框、目标框所属类别，(x,y) 表示预测框、目标框的中心位置像素坐标，

(w,h)表示预测框、目标框的宽度和长度，$conf$ 表示预测框的置信度。

3)如图 4 所示，将同一张事件图像上预测框集合 $\hat{y}$和目标框集合 y 分为三组，即 gt_only，$pseudo\_gt$，$pseudo\_only$。

$$gt\_only=\{box_y \mid box_y \cup box_{\hat{y}}=0\} \tag{3}$$

$$pseudo\_gt=\{(box_{\hat{y}},box_y) \mid box_{\hat{y}} \cup box_y>0\} \tag{4}$$

$$pseudo\_only=\{box_{\hat{y}} \mid box_{\hat{y}} \cup box_y=0\} \tag{5}$$

其中，gt_only 表示在事件图像上不与任何 $box_{\hat{y}}$ 相交的 box_y 集合；$pseudo\_gt$ 表示在事件图像上 $box_{\hat{y}}$ 与 box_y 相交组成的集合，将 $box_{\hat{y}}$ 与 box_y 相交定义为$(box_{\hat{y}},box_y)$对；$pseudo\_only$ 表示在事件图像上不与任何 box_y 相交的 $box_{\hat{y}}$ 集合。

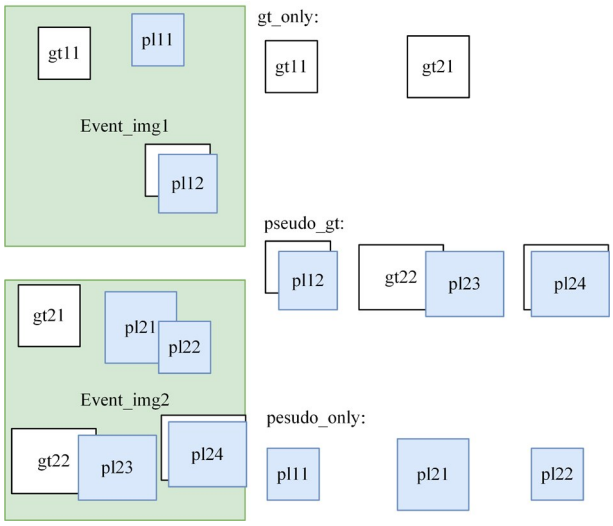


图 4 事件图像预测框、目标框分组示意图

Fig. 4 Schematic diagram of event image prediction box and object box grouping

4)对 gt_only，$pseudo\_gt$，$pseudo\_only$ 进行排序：

(1)对于 gt_only 中的每个 box_y 目标框，计算 box_y 目标框与同一事件图像中其他所有目标框的距离 GIoU，取平均距离作为该 box_y 目标框的分数，然后将 gt_only 中所有 box_y 目标框 gt_only 按照分数大小升序排列；

(2)对于 $pseudo\_gt$ 中的每一个 $box_{\hat{y}}$ 预测框，计算 $box_{\hat{y}}$ 预测框和所有相交的 box_y 目标框的 IOU 交并比，保留最大 IOU 交并比所对应的 box_y 目标框；若多个 $box_{\hat{y}}$ 预测框对应同一个 box_y 目标框，则取 IOU 交并比最大的 $box_{\hat{y}}$ 预测框和对应的 box_y 目标框组成 $(box_{\hat{y}},box_y)$ 对，然后将 $(box_{\hat{y}},box_y)$ 对按照 $box_{\hat{y}}^{conf}$，即 $box_{\hat{y}}$ 预测框的置信度降序排列；

(3)对于 $pseudo\_gt$ 中的每一个 $box_{\hat{y}}$ 预测框，在 $pseudo\_only$ 中按照 $box_{\hat{y}}^{conf}$，即 $box_{\hat{y}}$ 预测框的置信度降序排列。

5)根据目标检测器在验证集上的性能得出大概的类别噪声比例，每个类别都具有独立的噪声比例，不同类别噪声比例不同。对单个类别而言，噪声比例形式为[p1，p2，p3]。根据噪声比例 [p1，p2，p3]修改事件图像带有的目标框 y。p1，p2，p3 分别表示 gt_only，$pseudo\_gt$，$pseudo\_only$ 中的噪声比例，将修改后的事件图像带有的标记定义为 y'。y' 包括：

(1)取 gt_only 中前 (1−p1) 的 box_y 目标框；

(2)取 $pseudo\_gt$ 中前 p2 的$(box_{\hat{y}},box_y)$对中的 $box_{\hat{y}}$ 预测框，剩下$(box_{\hat{y}},box_y)$对中全部取 box_y 目标框；

(3)取 $pseudo\_only$ 中前 (1−p3) 的 $box_{\hat{y}}$ 预测框。

6)修改后的标记 y' 中包括 box_y 目标框和 $box_{\hat{y}}$ 预测框 $box_{\hat{y}}$，去掉 $box_{\hat{y}}$ 预测框中的 $conf$ 置信度，得到二次修改后 (cls,x,y,w,h)形式的目标框集合 y'。用二次修改后的y'替代事件图像上一阶段的标记 y 作为事件图像下一训练阶段的标记，用于损失计算。

7)步骤 1)−6)为一个训练阶段，阶段内模型的最优性能为阶段性能。重复训练阶段 N 次，每次识别噪声比例递减，直到目标检测器在阶段性能上出现下降，停止训练，得到最终的事件图像目标检测器。

3 实验与分析

3.1 数据集介绍

实验使用了 3 个公开的自动驾驶场景下的目标检测数据集，即 1Mpx Detection Dataset[10]，BDD100K[23] 和 SODA10M[24]。1Mpx Detection Dataset 是一个包括城市、高速公路、乡村、小村庄等各种场景，并在各种各样的照明和天气条件下拍摄的百万像素级别的事件相机目标检测数据集。BDD100K 是一个包含了 10 万张图片和标记的自动驾驶数据集，数据集采集时覆盖多个城市、多种天气、一天中的多时间段和多个场景类型。SODA10M 是华为诺亚方舟实验室联合中山大学发布的新一代半/自监督的 2D 基准数据集，其主要包含从 32 个城市采集的一千多万张多样性丰富的无标记道路场景图片以及两万张带标记图片。

从 1Mpx Detection Dataset 中选取包含 55 个场景的事件序列，其中训练集 35 个场景，验证集 10 个场景，测试集 10 个场景，每个事件序列为 60 s。设置编码时间长度为 10 ms，对事件序列进行编码和标记迁移，生成带目标框标记的事件图像 34 580 张，其中训练集 24 000 张，验证集 4 806 张，测试集 5 774 张。为了验证方法的有效性，验证集和测试集标记经过了简单的人工修正。检测目标类别为：行人、两轮车、四轮车。

开源 RGB 数据采用 BDD100K 和 SODA10M 中 2D 目标检测部分，组成集合 S，包含 89 114 张 RGB 图像及其标记。

3.2 实验模型及参数设定

实验将 RGB 图像经过事件图像初步训练(Warm-up)后模型推理输出的不确定性作为(损失)依据，挑选出不确定性大的 10 000 张 RGB 图像与训练集事件图像组成跨模态混合图像训练集，验证集和测试集保持不变。联合优化框架训练阶段 N 一般取 2 或 3。

实验采用的目标检测模型为 YOLOv5，使用 MeanTeacher 方法[25]得到更好的 EMA 检测器，使用 4 张英伟达 2080 显卡并行训练。“基准模型 Baseline”表示基于训练集事件图像及其标记直接进行训练；“跨模态混合图像训练 CM-Train”表示基于事件图像和筛选后的 RGB 图像通过 Mosaic 增强生成跨模态混合图像进行训练；“联合优化框架 JOF”表示基于事件图像及其标记进行多阶段训练，过程中进行噪声识别和修正；“跨模态联合优化框架 CM-JOF”基于跨模态混合图像训练集使用联合优化框架训练，注意，跨模态联合优化框架中不会对 RGB 图像标记进行修改。实验采用事件图像测试集 mAP@0.5 指标和可视化结果来验证方法的有效性。

3.3 实验结果与可视化

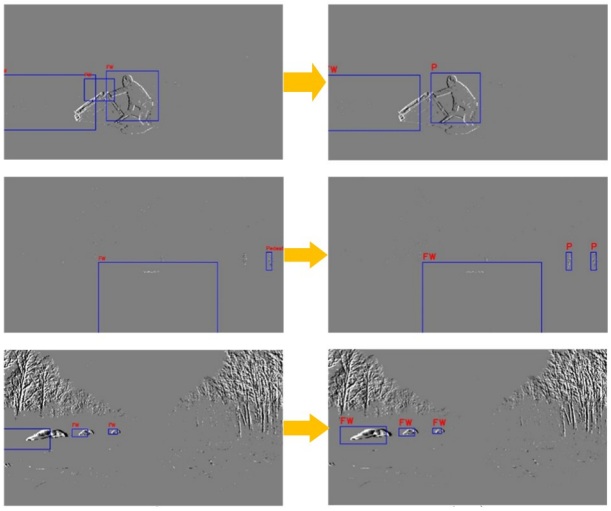
为了保证实验结果的公平性，每组实验均在相同的环境中进行，且各参数设定相同，实验结果如表 1 所列。训练过程

中，训练集事件图像标记修正结果可视化如图 5 所示，其中(a)为修正前标记，(b)为标记修正结果，P，TW，FW 分别表示行人、两轮车和四轮车。

表 1　事件图像测试集上的检测结果

Table 1　Detection results on event image test set

方法	行人	两轮车	四轮车	mAP@0.5	提升/%
Baseline	0.231	0.355	0.670	0.419	—
O2U-net[12]	0.235	0.352	0.680	0.422	0.72
Co-teaching[13]	0.229	0.347	0.655	0.410	−2.15
CM-JOF	**0.260**	**0.399**	**0.703**	**0.454**	**8.35**



(a)　　(b)

图 5　事件图像噪声标记修正

Fig. 5　Event images noisy annotations correction

表 1 中，mAP@0.5 指标表示模型在事件图像测试集上的平均测试精度，行人、两轮车、四轮车分别表示标记中对应类别上的测试精度。从表 1 中 mAP@0.5 指标可以看出，在包含大量标记噪声的事件图像数据上，Co-teaching，O2U-net 等处理噪声的方法并不能取得较好的结果，跨模态联合优化框架 CM-JOF 方法通过联合 RGB 数据构造跨域混合样本帮助模型通过局部视图定位事件目标并通过特定的噪声过滤方法进行噪声过滤，可以取得较大程度的性能提升。

3.4　消融实验

3.4.1　跨模态联合优化框架

跨模态联合优化框架包含跨模态混合图像训练和联合优化框架两个部分，单独使用每个部分的实验结果如表 2 所列。

表 2　跨模态联合优化框架消融结果

Table 2　Ablation results of cross-modal joint optimization framework

方法	行人	两轮车	四轮车	mAP@0.5	提升/%
Baseline	0.231	0.355	0.670	0.419	—
CM-Train	0.259	0.394	0.684	0.446	6.44
JOF	0.249	0.374	0.694	0.439	4.77
CM-JOF	**0.260**	**0.399**	**0.703**	**0.454**	**8.35**

从表 2 所列模型在测试集上的 mAP@0.5 指标可以看出，跨模态混合图像训练可以帮助模型更好地检测事件图像目标，联合优化框架可以缓解噪声标记对模型带来的损害。

3.4.2　高质量 RGB 图像筛选

不同方法选取相同数量的 RGB 图像和事件图像组成跨模态混合图像进行训练，Random 为随机挑选 10 000 张 RGB 图像；GEAL 为使用 Xie 等[18]提出的 GEAL 方法挑选 10 000 张 RGB 图像；AllRGB_FineTurn 为先利用全部的 89 114 张 RGB 图像预训练模型，再将预训练模型在事件图像上微调；Uncertainty_small(或 Uncertainty_big)为利用事件图像 warm-up 后的检测器推理 RGB 图像，选择不确定性低(或高)的前 10 000 张图像，实验中不确定性用 YOLOv5 网络的损失函数值来判定。模型在测试集上的检测结果如表 3 所列。

表 3　RGB 图像的选取方法

Table 3　Selection methods of RGB images

方法	行人	两轮车	四轮车	mAP@0.5
Baseline	0.231	0.355	0.670	0.419
Random	0.239	0.363	0.687	0.430
GEAL[18]	0.243	0.353	0.690	0.429
AllRGB_FineTurn	0.254	0.357	**0.692**	0.434
Uncertainty_small	0.225	0.333	**0.692**	0.417
Uncertainty_big	**0.259**	**0.394**	0.684	**0.446**

从表 3 所列模型在测试集上的 mAP@0.5 指标可以看出，以 RGB 图像经过 warm-up 后的检测器推理的不确定性为筛选 RGB 图像的依据，不确定性越高，RGB 图像对事件图像目标检测器的价值越大。

3.4.3　缓解 RGB 图像和事件图像之间的域差异

采用 Uncertainty_big 方法挑选 10 000 张 RGB 图像后，使用 3 种不同方法进行训练得到模型检测结果，如表 4 所列。事件图像和 RGB 图像内部分别使用 Mosaic 增强(Baseline-RGB)训练、域对抗(DANN)训练和构造跨模态混合图像(CM-Mosaic)训练。

表 4　高质量 RGB 图像筛选

Table 4　High quality RGB image filtering

方法	行人	两轮车	四轮车	mAP@0.5
Baseline	0.231	0.355	0.670	0.419
Baseline-RGB	0.224	0.404	0.626	0.418
DANN[20]	0.242	**0.419**	0.625	0.429
CM-Mosaic	**0.259**	0.394	**0.684**	**0.446**

从表 4 所列模型在测试集上的 mAP@0.5 指标可以看出，Baseline-RGB 直接使用事件图像和 RGB 图像进行训练，事件图像目标检测器效果无提升，CM-Mosaic 构造跨模态混合图像方法比 DANN 域对抗方法效果更好。

结束语　本文提出了跨模态噪声过滤的事件相机目标检测算法，可以提高事件图像目标检测器对事件图像目标的检测能力。方法包含两个方面：1)引入开源 RGB 图像，构造跨模态混合图像，帮助检测器更好地检测事件图像目标；2)在含有大量噪声标记的情况下，将训练分为多个阶段，单个阶段训练完成时，在全局标记中识别噪声标记，并对噪声标记进行修正后在下一阶段继续使用，缓解噪声标记对检测器性能的影响。但跨模态联合优化框架专注于解决事件图像目标检测和迁移标记噪声问题，未进一步验证算法在其他噪声比例较低的目标检测数据集上的实验效果。

参 考 文 献

[1]　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.

[2]　HAN J，YANG Y，ZHOU C，et al. EvIntSR-Net：Event guided multiple latent frames reconstruction and super-resolution[C]// Proceedings of the IEEE/CVF International Conference on

Computer Vision. 2021:4882-4891.
[3] XU Q, DENG J, SHEN J R, et al. A Review of Image Reconstruction Based on Event Cameras[J]. Journal of Electronics & Information Technology, 2023, 45(8):2699-2709.
[4] LICHTSTEINERP, POSCH C, DelBruck T. A 128× 128 120 dB 15 μs Latency Asynchronous Temporal Contrast Vision Sensor[J]. IEEE Journal of Solid-State Circuits, 2008, 43(2):566-576.
[5] GALLEGO G, DELBRÜCK T, ORCHARD G, et al. Event-based vision: A survey[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2020, 44(1):154-180.
[6] SABATER A, MONTESANO L, MURILLO A C. Event Transformer. A sparse-aware solution for efficient event data processing[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022:2677-2686.
[7] WAN J, XIA M, HUANG Z, et al. Event-Based Pedestrian Detection Using Dynamic Vision Sensors[J]. Electronics, 2021, 10(8):888.
[8] MIAO S, CHEN G, NING X, et al. Neuromorphic vision datasets for pedestrian detection, action recognition, and fall detection[J]. Frontiers in Neurorobotics, 2019, 13:38.
[9] HE D C, WANG L. Texture unit, texture spectrum, and texture analysis[J]. IEEE transactions on Geoscience and Remote Sensing, 1990, 28(4):509-512.
[10] PEROT E, DE TOURNEMIRE P, NITTI D, et al. Learning to detect objects with a 1 megapixel event camera[J]. Advances in Neural Information Processing Systems, 2020, 33:16639-16652.
[11] FINATEU T, NIWA A, MATOLIN D, et al. 5. 10 a 1280× 720 back-illuminated stacked temporal contrast event-based vision sensor with 4. 86-m pixels, 1. 066 GEPS readout, programmable event-rate controller and compressive data-formatting pipeline [C]// 2020 IEEE International Solid-State Circuits Conference (ISSCC). IEEE, 2020.
[12] HUANG J, QU L, JIA R, et al. O2u-net: A simple noisy label detection approach for deep neural networks[C]// Proceedings of the IEEE/CVF Cnternational Conference on Computer Vision. 2019:3326-3334.
[13] HAN B, YAO Q, YU X, et al. Co-teaching: Robust training of deep neural networks with extremely noisy labels[J]. arXiv: 1804. 06872, 2018.
[14] TANAKA D, IKAMI D, YAMASAKIT, et al. Joint optimization framework for learning with noisy labels[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018:5552-5560.
[15] LI J, XIONG C, SOCHER R, et al. Towards noise-resistant object detection with noisy annotations[J]. arXiv: 2003. 01285, 2020.
[16] BOCHKOVSKIY A, WANG C Y, LIAO H Y M. Yolov4: Optimal speed and accuracy of object detection[J]. arXiv: 2004. 10934, 2020.
[17] LIU K, QIAN X, WANG Z Q. Survey on active learning algorithms[J] Computer Engineering and Applications, 2012, 48(34):1-4.
[18] XIE Y, TOMIZUKA M, ZHAN W. Towards general and efficient active learning[J]. arXiv:2112. 07963, 2021.
[19] PAN S J, YANG Q. A survey on transfer learning[J]. IEEE Transactions on Knowledge and Data Engineering, 2009, 22(10):1345-1359.
[20] GANIN Y, USTINOVA E, AJAKAN H, et al. Domain-adversarial training of neural networks[J]. The journal of machine learning research, 2016, 17(1):2096-2030.
[21] JIANG J, CHEN B, WANG J, et al. Decoupled adaptation for cross-domain object detection[J]. arXiv:2110. 02578, 2021.
[22] VAN DER AALST W M P, RUBIN V, VERBEEK H M W, et al. Process mining: a two-step approach to balance between underfitting and overfitting[J]. Software & Systems Modeling, 2010, 9:87-111.
[23] YU F, CHEN H, WAN G X, et al. Bdd100k: A diverse driving dataset for heterogeneous multitask learning[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020:2636-2645.
[24] HAN J, LIANG X, XU H, et al. SODA10M: a large-scale 2D self/Semi-supervised object detection dataset for autonomous driving[J]. arXiv:2106. 11118, 2021.
[25] TARVAINEN A, HARRI V. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results[C]// Proceedings of the 31st International Conference on Neural Information Processing Systems. 2017:1195-1204.

HU Gang, born in 1998, postgraduate. His main research interests include computer vision and machine learning.

HUANG Shengjun, born in 1986, Ph.D, professor, Ph.D supervisor, is a member of CCF (No. 42916S). His main research interests include machine learning and data mining.