计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
2025,34(3):94−104 [doi: 10.15888/j.cnki.csa.009801] [CSTR: 32024.14.csa.009801]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

基于 EMDR-RAFT 光流法的地铁屏蔽门乘客闯门检测①

周　昊[1], 刘光杰[1], 高　申[2], 李佑文[3]

[1](南京信息工程大学 电子与信息工程学院, 南京 210044)
[2](南京熊猫信息产业有限公司, 南京 210038)
[3](南京国电南自轨道交通工程有限公司, 南京 210032)
通信作者: 刘光杰, E-mail: everglow_sun@163.com

摘　要: 地铁系统作为城市交通的核心组成部分, 安全性与效率的提升对于保障乘客的生命财产安全具有重要意义. 行人闯门行为不仅会导致设备损坏和交通延误, 更可能对其他乘客的安全构成威胁. 因此, 准确检测和识别地铁场景下的行人闯门行为成为智能交通管理中的一项重要任务. 本文提出一种行人闯门威胁检测算法. 该算法首先在 RAFT 光流法的特征提取器中使用移动网络卷积模块, 并添加 ECA 通道注意力机制, 同时在相关体构建块中使用 3D 结构并缩减领域半径, 以期降低模型参数量的同时提升检测速度. 实验结果表明, 该算法对行人检测的平均端点误差为 0.79, 检测速度可达到 55.98 f/s, 模型参数量降低了 35.3%. 为获取乘客闯门威胁值, 本文使用改进光流法计算出相邻图片帧的运动信息, 结合本文提出的闯门威胁计算公式得到当前图片帧乘客的闯门威胁值. 该方法满足了实时性、准确度和轻量化的同时还可以有效部署, 更好地满足了站内大客流的行人威胁检测和应急管理的工程实践要求.

关键词: RAFT 光流法; 行人运动检测; 注意力机制; 移动网络卷积; 模型部署

引用格式: 周昊,刘光杰,高申,李佑文.基于 EMDR-RAFT 光流法的地铁屏蔽门乘客闯门检测.计算机系统应用,2025,34(3):94–104. http://www.c-s-a.org.cn/1003-3254/9801.html

Detection of Passengers Breaching Metro Platform Screen Doors Based on EMDR-RAFT Optical Flow Method

ZHOU Hao[1], LIU Guang-Jie[1], GAO Shen[2], LI You-Wen[3]

[1](School of Electronic and Information Engineering, Nanjing University of Information Science & Technology, Nanjing 210044, China)
[2](Nanjing Panda Information Industry Co. Ltd., Nanjing 210038, China)
[3](Nanjing Guodian Nanzi Rail Transit Engineering Co. Ltd., Nanjing 210032, China)

Abstract: As a core component of urban transportation, the improvement of safety and efficiency of the subway system is of great significance in ensuring the safety of passengers' lives and property. Pedestrian gate-breaking behavior can not only cause equipment damage and traffic delays but also pose a threat to the safety of other passengers. Therefore, accurately detecting and recognizing the behavior of pedestrians breaking through subway gates has become an important task in intelligent transportation management. This study proposes a pedestrian gate-breaking threat detection algorithm. Firstly, the algorithm uses the mobile network convolution module in the feature extractor of the RAFT optical flow method and adds the ECA channel attention mechanism. At the same time, the 3D structure is used in the related volume building block and the field radius is reduced, to reduce the number of model parameters and improve the detection speed.

① 基金项目: 国家自然科学基金 (U21B2003); 江苏省产业前瞻与关键核心技术竞争项目 (BE2022075)
收稿时间: 2024-09-02; 修改时间: 2024-09-24; 采用时间: 2024-10-23; csa 在线出版时间: 2025-01-17
CNKI 网络首发时间: 2025-01-17

Experimental results show that the average endpoint error of the proposed algorithm for pedestrian detection is 0.79. The detection speed can reach 55.98 f/s, and the number of model parameters is reduced by 35.3%. To obtain the threat value of passengers breaking through subway gates, this paper uses the improved optical flow method to calculate the motion information of adjacent picture frames and combines the gate-breaking threat calculation formula proposed in this study to obtain the threat value of passengers in the current picture frame. This method meets the requirements of real-time performance, accuracy, and lightweight design, and can be effectively deployed to better meet the engineering practice requirements of pedestrian threat detection and emergency management for large passenger flows within the station.

Key words: RAFT optical flow method; pedestrian motion detection; attention mechanism; MobileNet convolution; model deployment

运动目标检测是计算机图像、视频处理工作的基础, 近年来我国城市化进程的加快, 地铁系统已成为现代城市交通网络的重要组成部分, 对城市居民的日常出行起着至关重要的作用. 然而, 随着乘客数量的不断增加, 地铁系统在运行中面临着一系列安全与管理挑战. 尤其是行人闯门行为, 不仅可能导致设备的损坏和交通的延误, 还可能对其他乘客的生命安全造成威胁[1]. 因此, 如何有效检测并预防此类行为, 成为智能交通管理中亟待解决的问题.

闯门威胁检测技术主要用于识别上客期尾段场景下视频或图片中的乘客, 并对其进行定位位置信息和计算闯门风险值. 伴随着计算机视觉和机器学习技术的飞速发展, 使得基于视频分析的乘客行为检测方法在各类场景中得到了广泛应用. Simonyan 等人[2]提出双流卷积网络, 通过分别处理视频的空间和时间特征, 有效提升动作识别的准确性. 然而, 双流架构的复杂性导致了较高的计算成本, 限制了其在实时应用中的广泛使用. Wang 等人[3]提出一种时间片段网络 (TSN), 通过捕捉视频中长时间段的动作信息, 显著提升动作识别的准确性. 尽管 TSN 在性能上表现良好, 但由于模型的复杂性, 其计算效率仍然较低. Carreira 等人[4]提出一种新的动作识别模型, 并引入 Kinetics 数据集. 该模型在多个基准任务中表现优异, 但其复杂的架构带来较高的计算成本. Koh 等人[5]提出一种上下文感知记忆注意网络 (CAMA-Net), 该模型通过结合 2D 卷积神经网络和注意力机制, 直接使用原始 RGB 视频帧进行识别, 该模型通过引入上下文感知机制来更好地捕捉视频中时空信息的依赖关系, 实验表明该方法在保证识别准确性的同时, 提高了计算效率, 但其在处理复杂场景和高分辨率视频时仍有改进空间. Xie 等人[6]提出一种深度学习方法, 结合 CNN、RNN 和双流卷积网络来进行视频中的人类行为识别. 该方法在多个任务中表现优异, 但其模型复杂性和计算负担限制其在实际应用中的部署. Sun 等人[7]利用 RAFT 生成光流特征并应用于视频动作识别, 通过结合光流信息增强视频中运动目标的检测鲁棒性. 尽管如此, 模型在处理剧烈运动时仍可能出现误差, 并且在长时间视频处理时计算资源消耗较大. Ling 等人[8]通过在不同尺度上对 RAFT 光流法使用的数据循环处理, 能够更精确地捕捉和估计复杂的动态场景中的运动变化, 进一步优化光流估计的精度, 尤其在细节丰富的场景中表现出色. 尽管精度有所提升, 但计算开销较大限制了实际应用中广泛部署. 王曦明等人[9]在轨道交通的监控系统中运用视频分析技术判断乘客是否存在闯门行为, 尽管能够较为准确地给出乘客闯门风险值. 但是该技术对于威胁值计算中对闯门乘客使用匀速运动模型并不符合当前应用场景, 同时在使用算法进行目标检测时还存在丢帧问题, 无法保证在真实场景下应用的精确度.

针对以上问题, RAFT 光流法网络模型的综合性能表现优异, 其模型根据参数量大小可分为 RAFT 和 RAFTs 两种结构. 考虑到在地铁站内场景部署应用, 本文选择以 RAFTs 为基础算法进行改进, 提出了一种轻量化行人闯门检测算法 EMDR-RAFT, 并部署在旭日 X3 派上进行测试. 在 EMDR-RAFT 光流估计算法中, 首先将特征提取模块的标准卷积替换为移动网络卷积, 在捕捉更丰富的空间特征信息的同时显减少计算量和参数量, 提高了网络的计算效率; 其次加入 ECA 通道注意力机制, 加强对行人行为和运动位置信息的关注, 提高对目标位置的定位能力, 同时弥补轻量化处理带来的精度损失; 最后通过构建 3D 相关体并缩减特征像素匹配半

径, 减小计算复杂度, 降低冗余背景信息的干扰. 另外使用变加速运动作为乘客在闯门场景下的运动模型修正王曦明等人[9]提出的闯门威胁计算公式, 结合 EMDR-RAFT 光流法, 获取上客期尾段乘客闯门视频的相邻图片帧中的闯门乘客速度和方向等光流运动信息, 最终得到乘客的闯门威胁值. 通过实验证明, 改进后的算法相较于原始模型, 拥有更快的实时检测性能, 评估的闯门威胁值更符合真实应用场景.

1 基于 EMDR-RAFT 的行人检测

EMDR-RAFT 模型的改进从 3 个方面进行: ① 针对效率较低的标准卷积使用移动网络卷积进行改进, 使得模型能够拥有更少的参数量和运算量, 加快模型检测速度; ② 引入 ECA 通道注意力机制, 使得模型能够更好地关注行人的行为和运动位置信息, 在提升模型检测性能的同时并不会显著增加模型的运算量; ③ 针对全局空间信息特征较为复杂, 通过使用 3D 相关体维度和减少像素领域采样半径, 使得模型能够减少非乘客场景信息对于检测的干扰, 有效提高模型的精度.

EMDR-RAFT 光流法模型结构如图 1 所示, 主要分为特征提取模块、3D 相关体构建模块和迭代优化模块. 特征提取模块中的特征编码器负责从输入的相邻帧图片中提取运动信息、物体边缘、纹理等特征, 将特征分辨率缩小到原始图像的 1/4. 上下文编码器则提取前一帧图像的上下文特征, 为后续的光流迭代提供有用的信息, 使生成的光流图与原图保持位置对应. 特征编码器提取的信息通过 3D 相关体构建中的视觉相似性计算和相关性查询实现全局特征匹配, 生成的 3D 相关体用于光流计算. 光流迭代优化模块结合了卷积网络的门控循环单元 (GRU) 序列, 输入由上下文编码器的输出、构建的 3D 相关体、上层隐藏状态信息和上次迭代输出的光流组成. 每次循环计算获得的光流残差值与当前光流值相加, 逐步优化出最佳光流值. 迭代结束后, 通过上采样模块匹配真值分辨率, 得到最终光流结果.

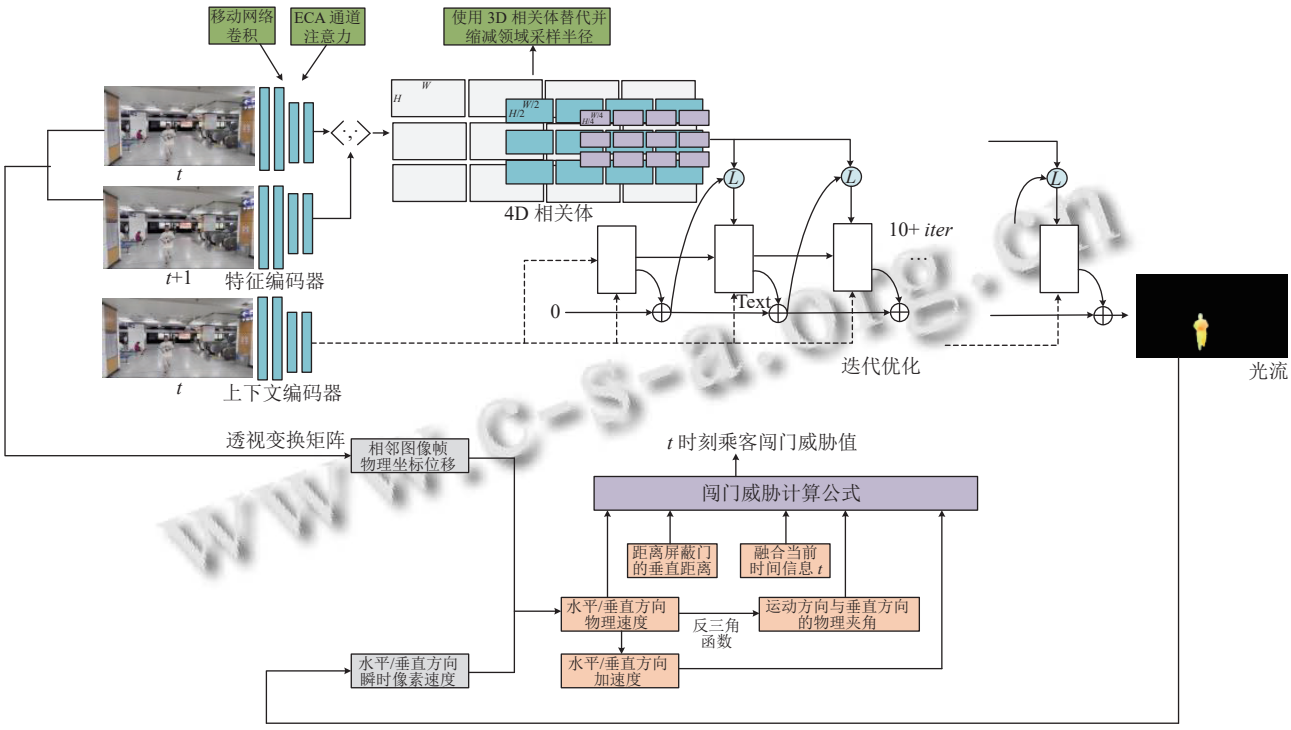

图 1 闯门威胁检测网络模型结构图

1.1 移动网络卷积模块

EMDR-RAFT 光流法网络中的特征提取模块使用移动网络卷积替换标准卷积, MBConv 可以减少网络的 *FLOPs*[10], 缓解内存访问频繁造成的性能损失, 同时优化 *FLOPS*. 网络结构如图 2 所示.

检测延迟 (*Latency*) 和浮点数运算 (*FLOPs*) 公式如式 (1) 所示. 其中, *FLOPS* 表示每秒浮点运算的缩写, 度量有效的计算速度; *FLOPs* 表示模型中的浮点运算

操作总数, 即模型的计算量.

$$Latency = \frac{FLOPs}{FLOPS} \tag{1}$$

移动网络卷积的核心在于使用深度可分离卷积取代传统卷积层, 将其分解为逐通道卷积 (DWConv) 和逐点卷积 (PWConv). 逐通道卷积用于空间滤波, 通过将卷积核变为单通道, 每个卷积核处理一个通道. 逐点卷积用于特征生成, 既可以改变特征图的维度, 又可以在逐通道卷积生成的特征图通道上进行融合. 在逐通道卷积中, 每个卷积核的深度为 1, 输出特征矩阵与输入特征矩阵的深度相等. 逐点卷积则相当于卷积核大小为 1 的普通卷积, 通常与逐通道卷积配合使用, 位于其后, 用于改变或自定义特征矩阵的深度, 从而极大地减少了模型参数数量和计算量. 逐通道卷积和逐点卷积的组合如图 3 所示.

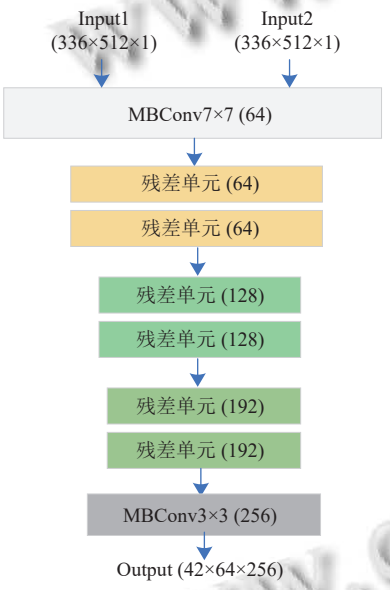

图 2 光流法特征提取结构图

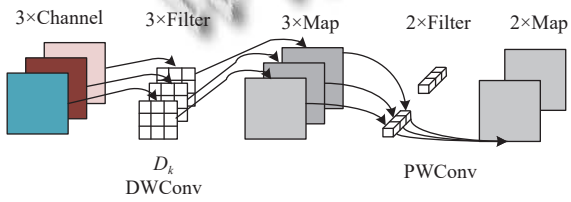

图 3 移动网络卷积结构

1.2 ECA 通道注意力模块

为提高模型对重要特征的关注度本文引入 ECA 通道注意力机制, 该模块可以自适应地为每个特征提取通道分配权重, 使得特征提取层可以更好地融合多层次特征, 实现不同通道间特征信息的交互, 提高目标检测的准确度和鲁棒性[11].

结构如图 4 所示, ECA 通道注意力机制首先通过全局平均池化处理输入的 C 个特征图, 将特征图从 [H, W, C]的矩阵转换为 [1, 1, C]的向量, 提取每个通道的全局信息; 然后将这个维度为 1×1×C 的特征向量进行 1×1 卷积, 并经过 *Sigmoid* 非线性激活函数, 生成每个通道的权重; 最后将注意力权重向量与原始特征图逐通道相乘, 生成经过注意力加权的特征图, 从而减少对无关信息的关注.

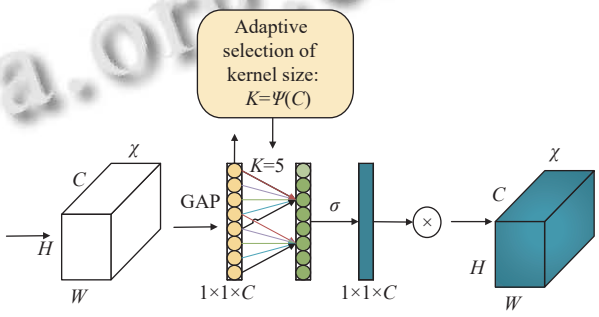

图 4 ECA 通道注意力结构图

ECANet 是对 SENet (squeeze-and-excitation network) 的改进. 与 SENet 相比, ECA 注意力机制模块在全局平均池化层之后直接使用卷积层, 省去了全连接层, 从而避免了降维对注意力机制的影响, 同时有效捕获了跨通道交互. ECA 可以根据卷积核的大小通过一个函数进行自适应调整, 改变特征的权重, 以更好地捕捉图像中的重要信息, 并提高计算效率.

ECA 的卷积核的自适应函数如下:

$$k = \left| \frac{\log_2(C)}{\gamma} + \frac{b}{\gamma} \right| \tag{2}$$

本文通过在特征提取模块的移动网络卷积层后添加 ECA 通道注意力模块, 使模型能够自适应地调整各通道的重要性权重, 更好地关注关键特征通道, 扩展每个位置的感受野, 同时提升特征表达能力, 进一步改善 EMDR-RAFT 的目标检测性能. 融合 ECA 后的残差单元模块结构如图 5 所示.

1.3 相关体构建模块

在相关体构建的过程中, 首先需要通过视觉相似性计算得到多尺度的互相关特征, 具体为对经过特征编码器后得到的两帧图像的特征向量进行点乘操作, 得到两个特征图所有向量之间的互相关信息, 如式 (3) 所示:

$$\begin{cases} Corr(f(I_1), f(I_2)) \in R^{H\times W\times H\times W} \\ Corr_{ijkl} = \sum_h f(I_1)_{ijh} \cdot f(I_2)_{klh} \end{cases} \tag{3}$$

其中, $Corr$ 为互相关信息, f 为输入的两帧图像特征向量, $f(I_1) \in R^{H\times W\times D}$ 和 $f(I_2) \in R^{H\times W\times D}$, 最终得到的互相关信息维度为 $(H\times W\times H\times W)$[1]. 为得到图片中不同大小像素的运动信息, 对生成的互相关信息的最后两个维度进行平均池化操作获得不同分辨率的相关特征, 构建出 4 层金字塔结构, 建立多尺度图像相似度特征. 其中通过保持四维度相关信息的前两维度不变, 即 I_1 的分辨率不变, 保证得到高分辨率的信息, 这样可以同时获得图像的全局和局部特征信息[12].

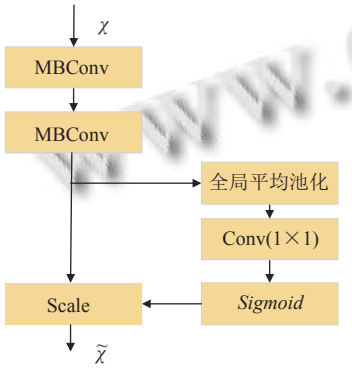

图 5 融合 ECA 的残差单元模块

对生成的多尺度图像特征使用相关性查询, 流程如图 6 所示, 对前一次估计的光流中的每个像素点, 通过使用上一次迭代的光流信息进行坐标变换, 初始的光流信息一般设置为 0. 其中 x 为图像 I_1 中的某个像素点 $x(u, v)$, 经过式 (4) 变换后, 得到下一帧图像的像素坐标 x':

$$x' = \left(u + f^1(u), v + f^2(v)\right) \tag{4}$$

像素坐标 x' 周围蓝色的点为定义的像素领域采样点集 $N(x')_r$, 它的作用是作为图像 I_2 的预测像素位置, 领域点集的公式如式 (5) 所示, 其中 r 为采样窗口的半径, dx 为整数.

$$N(x')_r = \left\{x' + dx | dx \in Z^2, \|dx\|_1 \leqslant r\right\} \tag{5}$$

为计算像素点 x 与领域点集的相关性, 对生成的互相关信息使用插值的方法查询得到, 互相关信息中的 4 个不同分辨率的相似度特征都会进行查询操作. 对于不同的分辨率, 查询的领域点集对应为 $N(x'/2^k)_r$, 其中, k 为分辨率的缩放大小, 依次为 1、2、4、8, 而每个像素周围的领域点之间的间隔并不随坐标一起缩放, 因此越低分辨率的查询结果感受野越广. 最终得到包含运动坐标的 4D 相关性特征, 此特征将作为下一次光流迭代更新模块的新输入.

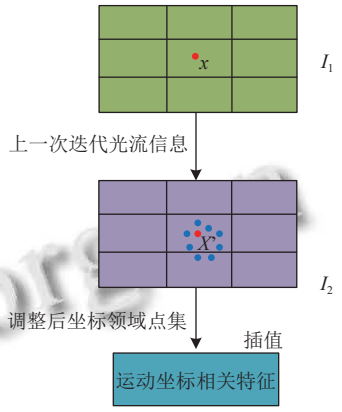

图 6 相关性模块流程查询图

RAFT 光流法通过提取连续帧的互相关信息获取多尺度特征, 其分别使用池化核大小为 1、2、4、8 的平均池化操作, 在图片物体变化尺度大的任务中表现良好. 但是在地铁站内的特定场景中, 部署的设备拟安装在距离地面 2.5 m 的屏蔽门正上方, 在俯视的视角下获取视场的连续图像帧, 其视场满足上客期尾段行人检测任务. 本文将 RAFT 光流法网络中的 4 个尺度的互相关信息特征减少为 3 个, 去掉池化核为 8 的池化操作, 构建出 3D 相关体. 如图 7 所示, 其互相关特征为多个连续帧图片经过特征编码器后各自生成的 42×64×64×256 的空间特征, 经过点乘融合后得到, 维度为 (42, 64, 64, 42, 64, 64). 通过池化操作得到多个不同尺度的互相关特征, 其中一个尺度的特征维度不变, 另一个维度变为 (21, 32, 32, 21, 32, 32), 形成 3D 相关体结构. 能够更全面地捕捉三维空间中的运动特征, 提高光流估计的准确性, 同时也满足复杂场景下对光流预测的更高要求.

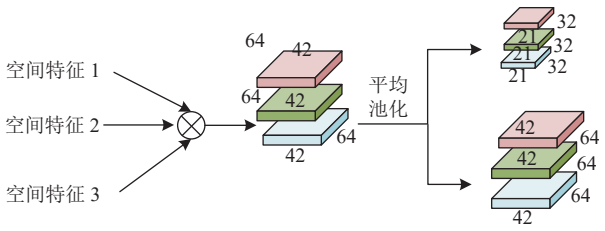

图 7 3D 相关体模块结构

在领域半径内的特征提取中, EMDR-RAFT 光流法使用了半径为 2 的采样半径, 用于预测前后连续帧

图像中像素点的运动相关性, 这对前后帧图像中存在大位移和严重视差的场景具备优秀的坐标预测能力. 然而地铁站内应用场景中采集的前后录像帧相对位移较小, 因此可以通过降低相关性查询过程中的像素点领域采样半径, 减少计算复杂度. 同时, 这样的调整还能有效避免远距离非闯门乘客像素点的干扰.

2 基于光流信息的闯门威胁值计算

本节提出了基于变加速运动模型的闯门威胁值威胁计算公式, 该公式需要代入 t 时刻乘客的速度、乘客到屏蔽门的距离, 与屏蔽门垂直方向的夹角及闯门时刻 t 值. EMDR-RAFT 光流法生成的光流场特征中包含 t 时刻乘客的水平和垂直瞬时像素速度以及像素位移, 为了获得真实的物理角度值, 需要通过透视变换建立像素坐标与物理坐标的映射, 进而得到乘客与屏蔽门垂直方向的夹角. 其计算流程框图如图 8 所示.

2.1 透视变换

由于边缘设备摄像头是以俯视的角度进行拍摄会造成拍摄的图像发生形变, 因此使用 EMDR-RAFT 光流法得到的像素信息需要先通过透视变换将图像转换为真实平面信息, 如图 9 所示.

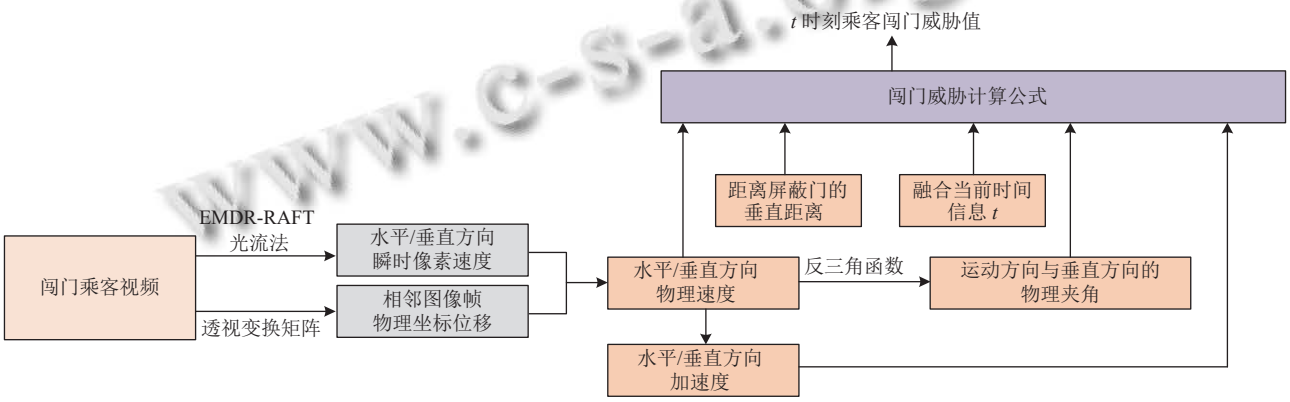

图 8 闯门威胁值计算流程

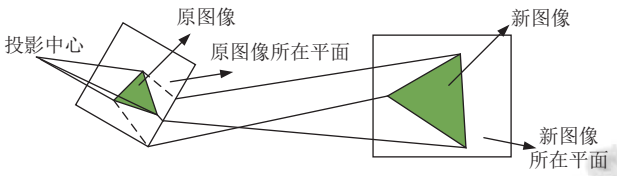

图 9 透视变换投影

使用透视变换将图像投影到新平面, 建立像素值与真实物理尺寸的映射关系, 校正图像发生的形变. 透视变换将二维图像投影到三维空间当中的一种方法是通过场景中的几何信息构建透视变化矩阵[13]. 假设输入像素点的坐标为 (m, n), 对应的物理坐标为 (X, Y, Z), 透视变换的通用公式如式 (6) 所示:

$$\begin{pmatrix} X \\ Y \\ Z \end{pmatrix} = \begin{pmatrix} a & a & a \\ a & a & a \\ a & a & a \end{pmatrix} \begin{pmatrix} m \\ n \\ z \end{pmatrix} \tag{6}$$

其中, (m, n, z) 是原始图像像素点的齐次坐标, 透视变化矩阵 $\begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{pmatrix}$ 中, $\begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix}$ 对图像进行线性变换, $\begin{pmatrix} a_{13} \\ a_{23} \end{pmatrix}$ 对图像使用透视变换, $\begin{pmatrix} a_{31} & a_{32} \end{pmatrix}$ 表示图像平移, a_{33} 一般为 1, 表示对矩阵做归一化操作. 经过透视变换后的新图像像素点的齐次坐标变换如式 (7) 所示:

$$\begin{cases} X = \dfrac{X}{Z} = \dfrac{a_{11}m + a_{12}n + a_{13}}{a_{31}m + a_{32}n + a_{33}} \\ Y = \dfrac{Y}{Z} = \dfrac{a_{21}m + a_{22}n + a_{23}}{a_{31}m + a_{32}n + a_{33}} \\ Z = \dfrac{Z}{Z} = 1 \end{cases} \tag{7}$$

通过计算式 (7) 能够得到 (X, Y), 即得到像素点坐标 (m, n) 变换后对应的二维平面坐标, 式中共有 8 个未知数, 即透视变化矩阵中除了 a_{33} 的另外 8 个参数, 因此只要分别找到原图像和映射得到的新平面中矩阵的 4 个顶点坐标便可以求解出透射变换矩阵.

2.2 威胁值计算

本文所使用的闯门威胁值计算公式如式 (8) 所示:

$$R_i(t)=\max(\cos(u),0)\times\max(1.5-(v_i\times(T-t)+0.5\times a\times(T-t)^2)/D_i,0) \tag{8}$$

其中, u 为乘客运动方向与屏蔽门垂直方向的夹角, v_i 为运动速度的垂直分量, T 为上客期距离关门还剩多长时间来启动闯门威胁算法, 本文设定为 5. t 为当前图像帧对应的时间, D_i 为距离屏蔽门的垂直距离, a 为乘客在运动方向上的加速度, $R_i(t)$ 为计算得到的闯门威胁值. 式 (8) 首先计算 $\max(\cos(u),0)$, 衡量乘客运动方向是否朝向门的方向, 与屏蔽门垂直方向的夹角越小, 该数值越大, 若夹角超过 90°, 威胁值为 0; 然后计算 $(v_i\times(T-t)+0.5\times a\times(T-t)^2)/D_i$, 表示乘客在垂直方向上运动的距离, 再通过使用计算得到的基准威胁值作差并保留威胁程度大于 1.5 的闯门风险帧. 通过式 (8) 能够实现对乘客闯门威胁值的精准评估.

首先通过 EMDR-RAFT 光流法获得图片中每个像素的位移, 然后与视频帧率做乘积得到水平和垂直的瞬时像素速度 v_x 和 v_y, 如式 (9) 所示:

$$\begin{cases} v_x=\dfrac{\Delta x}{\Delta t}\times f \\ v_y=\dfrac{\Delta y}{\Delta t}\times f \end{cases} \tag{9}$$

其中, Δx 为水平方向位移, Δy 为垂直方向位移, 视频的帧率为 f.

其次, 根据相邻图片帧获取的瞬时像素速度和时间间隔, 与帧率相乘得到对应图片帧的加速度 a_x 和 a_y, 如式 (10) 所示:

$$\begin{cases} a_x=\dfrac{v_{x1}-v_{x2}}{\Delta t}\times f \\ a_y=\dfrac{v_{y1}-v_{y2}}{\Delta t}\times f \end{cases} \tag{10}$$

其中, v_{xi} 和 v_{yi} 分别是在水平和垂直位移中相邻图片第 i 帧对应的瞬时像素速度.

然后利用相机的参数、实际测试时的相机高度和俯视角度及图像中已知矩形的 4 个顶点坐标得到透视变换矩阵, 将原始图像转换为平行视角, 从而得到相邻图像帧的物理坐标位移, 将物理坐标位移与帧率相乘即可得到物理的水平速度和垂直速度, 两者做反三角函数得到乘客运动方向与垂直方向的物理夹角, 进一步将乘客的头部像素点叠加物理的垂直坐标位移计算出距离屏蔽门的垂直距离. 最后将运动速度垂直分量、运动方向与屏蔽门法向的夹角、距离屏蔽门的垂直距离和当前时间 t, 输入到闯门威胁计算公式 (式 (8)), 得到 t 时刻乘客闯门威胁值.

3 实验结果及分析

3.1 实验环境

该实验使用 PyTorch 深度学习框架部署网络模型, 其深度学习环境的基本参数以及训练设备配置如表 1 所示.

表 1 实验参数及配置

名称	参数配置
操作系统	Windows 11
开发环境	CUDA 11.7
显卡	NVIDIA GeForce RTX 3060 Laptop 6 GB
内存	32 GB
CPU	12th Gen Intel(R) Core(TM) i9-12900H
部署设备	地平线旭日X3派

3.2 实验数据集

本文的训练集采用了专门针对运动目标的光流数据集 Human Flow[14], 该数据集通过人体的三维模型和运动捕捉数据构成真实的光流场, 其中包括 107 804 个训练帧和 27 176 个光流真实标签, 相比于其他光流数据集数量大了一个数量级, 数据集中的图片分辨率为 256×256, 在神经网络中能够轻松地进行训练. 该数据集的样式十分的广泛, 包含各种各样的人体形状、姿势、运动方向和虚拟背景. 因此使用该数据集进行训练能够极大的提高对乘客进行光流估计的效果. 验证集的图像由自行搭建对应场景采集的闯门视频按照设定时间间隔分帧得到, 采集高度和拍摄角度都与拟部署在地铁站内的边缘设备的安装高度和角度相同, 其中连续采集 50 s 视频, 视频中乘客到相机的水平距离为 5 m, 以不同的奔跑速度和方向朝屏蔽门方向跑来, 按顺序拆分为 10 个 5 s 的视频. 本文按 50 ms 的时间间隔对视频进行分帧处理.

3.3 实验训练参数及评价指标

实验在深度学习平台 PyTorch 上进行, 该算法的训练策略为: 学习率设为 10^{-3}、使用随机梯度下降算法进行优化、使用学习率衰减策略调整训练过程中的学习率、权重衰减系数为 0.004. 为了更好地衡量地铁场景内闯门检测算法的精确度和实时性能, 本实验设置的模型检测性能评价指标有: 平均端点误差 (average end point error, *AEE*)、 帧率 (frames per second, FPS). 对于每个输入连续帧图像的光流位移变化, 使用性能

评价指标 *EPE* (end point error) 进行衡量. *EPE* 表示光流估计的像素误差值, 称为端点误差. 端点误差的计算公式如式 (11) 所示, 其含义为每一个像素点预测的光流向量 $f_{x,y}$ 和真实光流向量 $g_{x,y}$ 之间的欧氏距离. EMDR-RAFT 光流法能够得到图像中所有像素点的光流预测值, 对所有像素点的欧氏距离求平均得到平均端点误差 (*AEE*), 计算公式如式 (12) 所示, N 为所有像素点的个数. 通过 *AEE* 的指标能够判断整张图像的光流预测效果, 其值越低, 预测效果越好.

$$EPE = \left\| f_{x,y} - g_{x,y} \right\| \tag{11}$$

$$AEE = \frac{1}{N}\sum_{i=1}^{N} \left\| f_{x,y} - g_{x,y} \right\| \tag{12}$$

3.4 实验结果及分析

实验共进行 4 组对比实验和 1 组消融实验, 第 1 组是 EMDR-RAFT 在不同数据集下的光流预测表现; 第 2 组是主流深度学习光流法的模型比较; 第 3 组是在实际应用场景中不同光流法下的检测效果对比图, 第 4 组是 RAFT 光流法加入不同改进模块下的消融实验; 第 5 组是实际闯门场景下结合威胁计算公式的对比实验.

首先, 为验证 Human Flow 相较于其他光流数据集在运动目标的光流预测性能的提升, 分别使用 MPI Sintel、Flying Things、KITTI2015、Human Flow 数据集训练 EDMR-RAFT 预测网络, 得到的权重对拍摄基于地铁屏蔽门场景的乘客闯门视频进行测试. 对比实验结果如表 2 所示.

表 2 不同光流数据集对比

数据集	训练帧	测试帧	分辨率	*AEE*
MPI Sintel	1 064	564	1024×436	0.82
Flying Things	21 818	4 248	960×540	0.98
KITTI2015	200	200	**1242×375**	2.83
Human Flow	**108 704**	**27 176**	256×256	**0.79**

根据表 2 可以看出, 相较于 MPI Sintel、Flying Things、KITTI2015 数据集, Human Flow 中包含更多的训练帧和测试帧, 尽管其数据的分辨率略低, 但是在实验训练充分的基础上, 其训练权重在闯门乘客的光流预测中可以得到最低的端点误差.

其次, 为了验证本文 EMDR-RAFT 算法的效果与性能, 将其与当前其他主流地光流估计算法, 如 LiteFlowNet[15]、IRR-PWC[16]、PWC-Net[17]、FlowNet2[18]、FastFlowNet[19]和 RAFT[12], 在相同的应用场景下进行对比实验, 具体实验结果如表 3.

表 3 主流光流估计方法的对比实验

算法	参数量 (M)	*AEE*	FPS (f/s)
LiteFlowNet	5.4	1.38	24.78
RAFT	5.3	1.17	35.17
FastFlowNet	**1.4**	2.45	14.76
PWC-Net	8.8	1.42	32.23
FlowNet2	162.5	1.83	27.56
EMDR-RAFT	3.4	**0.79**	**55.98**

由表 3 可知, FastFlowNet 有最低的参数量, 但是在精确度和速度方面表现较差, EMDR-RAFT 光流法在具有较少参数量的同时, 具有最低的平均端点误差和最优的图像推理性能, 是最适合当前应用场景中运动目标闯门检测的光流估计方法.

接下来通过对比 RAFT 光流法与其他主流的深度学习光流法的性能, 验证 RAFT 光流法应用到闯门乘客光流估计的效果. 为了将光流结果可视化, 本文对生成的稠密光流使用伪彩色可视化的方法将每个像素点的运动向量大小映射到一个伪彩色图像中. 其中较大的位移通常被映射为饱和的颜色, 如红色、绿色等, 较小的位移则映射为较浅的颜色, 如黄色、紫色等. 通过伪彩色可视化, 我们可以在一张彩色图像中同时显示出光流场中每个像素点的运动强度和方向, 从而更加直观地理解图像序列中的运动信息.

表 3 无法直观感受光流估计效果, 下面对不同光流法的效果进行可视化展示, 分别在乘客闯门视频中选择 4 组情况进行测试. 第 1 组为乘客侧向奔跑闯门, 位移较大; 第 2 组为乘客正向屏蔽门方向行走, 位移较小; 第 3 组为侧向乘客行走闯门, 位移较小; 第 4 组为乘客正向奔跑闯门, 位移较大. 分别使用 RAFT、Tiny-RAFT 和其他光流法中相对较好的 FlowNet2, FastFlowNet 对这 4 组图像序列进行光流估计, 生成的光流可视图如图 10 所示.

对 EMDR-RAFT 光流估计网络进行消融实验, 实验在 RAFT 光流估计网络的基础上, 提出了 3D 相关体、缩减领域采样半径、在特征编码器中添加 ECA 通道注意力机制、使用移动网络卷积替换特征提取模块中的标准卷积, 分别测试其性能变化情况. 在实验中统一使用 Human Flow 光流数据集进行预训练, 其输入数据等比例缩放为 480×270, 光流优化迭代轮数设置为 3, 并对输入数据采取灰度化处理, 实验结果如表 4.

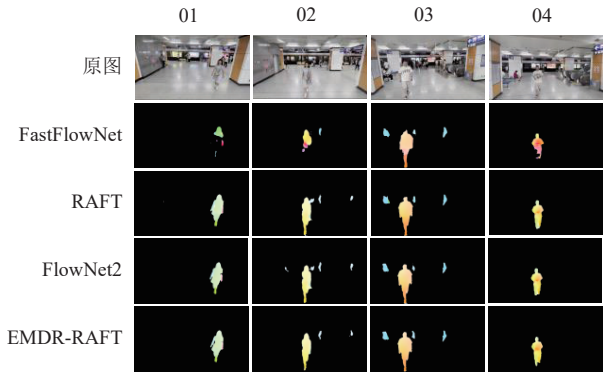

图 10 不同光流法可视图

表 4 EMDR-RAFT 光流法消融实验

网络结构	参数量	*AEE*	FPS (f/s)
RAFT网络	5 257 536	1.17	35.17
相关体 (4D→3D)	5 236 800	1.09	37.72
相关体 (4D→3D) 领域采样半径 (4→1)	5 181 504	0.95	42.66
相关体 (4D→3D) 领域采样半径 (4→1)+ECA注意力机制	5 213 504	0.76	38.54
相关体 (4D→3D)领域采样半径 (4→1) +ECA注意力机制+移动网络卷积	3 401 468	0.79	55.98

由表 4 可以看出, 使用重构的 3D 相关体和缩减领域采样半径的处理方法均可以有效提高光流预测的精度, 并且小幅度加快了图像推理速度; 在特征编码器中添加 ECA 通道注意力机制使模型, 虽然参数量有所上升, 但是在光流特征提取过程中模型更加关注各通道中的有效信息, 在小幅降低了检测速度的同时, 光流预测准确度提高了 9.5%; 在此基础上添加使用移动网络卷积模块替换特征提取器中的标准卷积, 在大大降低模型的了参数量和复杂度, 在保证精度基本不变的基础上, FPS 增加了 45.2%, 达到了 55.98, 有利于在边缘设备的部署.

在使用 EMDR-RAFT 得到光流运动信息后, 结合第 2 节中的闯门威胁计算公式, 分别对自建场景下拍摄的视频使用 50 ms 分帧处理, 并使用其相邻帧图像进行闯门威胁检测实验, 其中使用的运动信息数据及威胁值计算结果如表 5 所示. 在得到有关连续帧的闯门威胁值以后, 可以进行分级预警处理, 从而达到保证乘客生命财产安全的目的.

使用匀速运动模型的原始威胁计算公式和变加速运动模型改进的威胁计算公式分别对自建数据集进行检测, 其闯门检测结果如图 11 所示. 结合 EMDR-RAFT 算法生成的光流运动信息, 使用自行构建的地铁闯门场景数据集并输出闯门威胁值, 实验能够很好拟合闯门威胁计算公式的计算结果, 可以看出临近闯门时刻图片帧使用改进威胁计算公式给出的威胁值更加符合闯门场景, 其 FPS 在测试数据集中可达到 31.4 f/s, 能够满足当前场景下实时检测的性能需求. 在实际应用场景中能够快速检测出闯门威胁, 保障了乘客安全的同时提升了紧急事件处理的效率.

表 5 威胁计算实验结果

Time (ms)	Velocity	Angle (°)	Acceleration	Distance	Threat
0.00	0.03	22.31	0.65	17.60	0.82
0.05	0.04	39.71	0.03	16.15	0.62
0.10	0.06	36.37	1.13	15.05	0.47
0.15	0.08	59.13	−1.23	14.70	0.36
0.20	0.02	72.77	0.42	13.83	0.25
0.25	0.03	25.66	0.18	11.84	0.73
0.30	0.04	48.11	−0.22	10.30	0.54
0.35	0.02	20.44	0.03	9.72	0.76
0.40	0.02	28.67	0.05	9.23	0.70
0.45	0.03	24.02	0.28	8.45	0.74
0.50	0.03	24.73	0.79	7.44	0.82
0.55	0.14	17.90	0.65	6.98	0.72
0.60	0.32	20.16	0.58	6.32	0.77
0.65	0.33	25.97	0.42	5.66	0.81
0.70	0.35	17.36	0.37	5.41	0.83

图 11 闯门威胁检测对比实验

4 算法部署

4.1 地平线旭日 X3 派开发板

本文部署嵌入式 AI 开发板选择地平线旭日 X3 派, 该开发板使用 SoC 系统级芯片, 中心处理器搭载自研的四核 Cortex CPU 集群, 支持 FPU、NEON 加速, 采用自研的低功耗、高性能的双核 BPU 伯努利计算架构加速神经网络计算, 算力高达 5 TOPS, 为业务层面的模型调度提供了非常灵活的多核调度能力. 在视

频处理方面支持多路 Camera Sensor 同时输入、支持 H.264 和 H.265 视频编解码、最高 4K@60FPS 图像的实时处理. 采用主流的 Ubuntu 20.04 系统, 辅以地平线天工开物 (Open Explorer) AI 开发平台. 工作流程如图 12 所示.

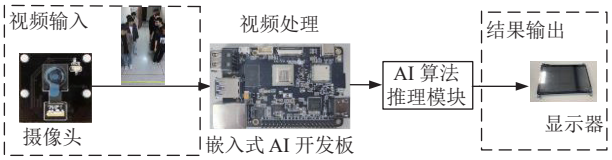

图 12 设备工作流程图

闯门威胁检测算法使用地平线 Open Explorer 开发平台中的 AI 算法工具链进行模型转换和编译优化, 部署在地平线旭日 X3 派开发板上, 实现智能门楣终端工作流程中的算法推理模块. 首先将闯门威胁网络的浮点模型转换为地平线混合异构模型. 浮点模型经过 PyTorch 深度学习框架训练得到, 混合异构模型是一种适合在地平线计算平台上运行的模型格式. 整个模型转换流程使用 hb_mapper 这一个量化编译工具即可实现, 如图 13 所示, 需要输入浮点模型、校准数据和 yaml 配置文件.

4.2 性能测试验证

把闯门威胁估计网络整体进行 PTQ 量化处理, 部署到旭日 X3 派开发板中对整体网络进行性能测试, 其中的闯门威胁估计网络加载 EMDR-RAFT 光流法训练得到的权重. 本次测试共使用采集的 5 s 闯门视频 5 个, 每个视频为 1 组, 共分为 5 组, 每组 100 张按 50 ms 分帧得到的图像.

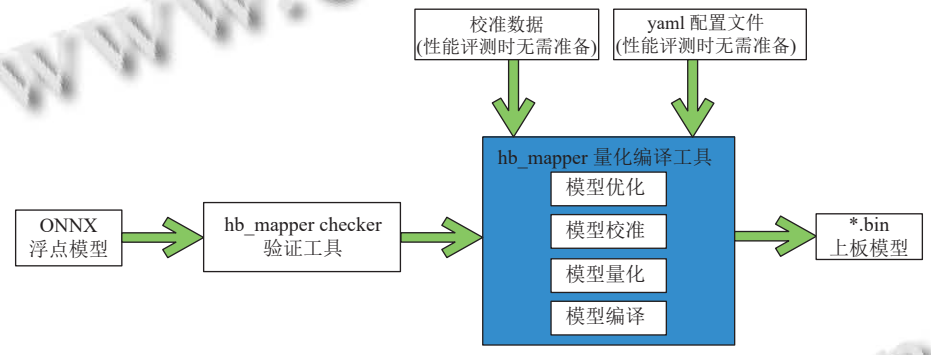

图 13 混合异构模型转换流程

部署在旭日 X3 派上的闯门威胁计算网络工作时, 当输入闯门视频后经过处理输出闯门威胁值, 其每秒能计算 27 次闯门威胁值, 符合实时检测的安全性能需求, 测试结果如图 14 所示, 能够很好地拟合闯门威胁计算公式的计算结果. 该模型在实际应用中能够准备快速地检测出闯门威胁, 极大地保障了乘客的出行安全.

图 14 威胁值检测结果

5 结论与展望

本文基于 RAFT 光流法结合改进的威胁值计算公式设计了轻量化闯门威胁检测算法 EMDR-RAFT, 进行地铁屏蔽门前乘客闯门行为的威胁检测研究, 在保证检测精度的前提下, 降低了模型的参数量和计算量, 提高了运算效率. 通过在旭日 X3 派上的部署证明其能够实现对地铁站内屏蔽门的实时威胁检测, 降低了实际工程中成本的投入, 而且更易于实现检测后行人追踪、行为警示、客流引导. 在未来, 将前往真实地铁站台场景下收集数据, 构建大规模数据集, 进一步提升算法准确度和鲁棒性, 优化模型的实际应用效果.

参考文献

1 Zuo L. Public safety risk prediction of urban rail transit based on mathematical model and algorithm simulation. Soft Computing, 2023. [doi: 10.1007/s00500-023-08919-x]

2 Simonyan K, Zisserman A. Two-stream convolutional

networks for action recognition in videos. Proceedings of the 27th International Conference on Neural Information Processing Systems. Montreal: MIT Press, 2014. 568–576.

3 Wang LM, Xiong YJ, Wang Z, *et al*. Temporal segment networks: Towards good practices for deep action recognition. Proceedings of the 14th European Conference on Computer Vision. Amsterdam: Springer, 2016. 20–36.

4 Carreira J, Zisserman A. Quo vadis, action recognition? A new model and the kinetics dataset. Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu: IEEE, 2017. 4724–4733.

5 Koh TC, Yeo CK, Jing X, *et al*. Towards efficient video-based action recognition: Context-aware memory attention network. SN Applied Sciences, 2023, 5(12): 330. [doi: 10.1007/s42452-023-05568-5]

6 Xie YF. Deep learning approaches for human action recognition in video data. arXiv:2403.06810, 2024.

7 Sun SY, Kuang ZH, Sheng L, *et al*. Optical flow guided feature: A fast and robust motion representation for video action recognition. Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018. 1390–1399.

8 Ling H, Sun QS. ScaleRAFT: Cross-scale recurrent all-pairs field transforms for 3D motion estimation. arXiv: 2407.09797v1, 2024.

9 王曦明, 刘光杰, 孙同庆. 一种基于视频分析的地铁屏蔽门临关门乘客闯门预警方法: 中国, CN115620228A. 2023-01-17.

10 Yang CL, Qiao SY, Yu QH, *et al*. MOAT: Alternating mobile convolution and attention brings strong vision models. Proceedings of the 11th International Conference on Learning Representations. Kigali: OpenReview.net, 2023.

11 Wang QL, Wu BG, Zhu PF, *et al*. ECA-Net: Efficient channel attention for deep convolutional neural networks. Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle: IEEE, 2020. 11531–11539.

12 Teed Z, Deng J. RAFT: Recurrent all-pairs field transforms for optical flow. Proceedings of the 16th European Conference on Computer Vision. Glasgow: Springer, 2020. 402–419.

13 Liu B, Lai H, Kan S, *et al*. Camera-based smart parking system using perspective transformation. Smart Cities, 2023, 6(2): 1167–1184. [doi: 10.3390/smartcities6020056]

14 Ranjan A, Romero J, Black MJ. Learning human optical flow. Proceedings of the 2018 British Machine Vision Conference 2018. Newcastle: BMVA Press, 2018. 297.

15 Hui TW, Tang X, Loy CC. LiteFlowNet: A lightweight convolutional neural network for optical flow estimation. Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018. 8981–8989.

16 Hur J, Roth S. Iterative residual refinement for joint optical flow and occlusion estimation. Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2019. 5754–5763.

17 Sun D, Yang X, Liu MY, *et al*. PWC-Net: CNNs for optical flow using pyramid, warping, and cost volume. Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018. 8934–8943.

18 Ilg E, Mayer N, Saikia T, *et al*. FlowNet 2.0: Evolution of optical flow estimation with deep networks. Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu: IEEE, 2017. 2462–2470.

19 Kong L, Shen C, Yang J. FastFlowNet: A lightweight network for fast optical flow estimation. arXiv:2103.04524v2, 2021.

(校对责编: 王欣欣)