应用光学
Journal of Applied Optics
SCOPUS,等来源期刊
中文核心期刊、中国科技核心期刊

基于自编码器结构与改进Bytetrack的低光照行人检测及跟踪算法

任泽林 庞澜 王超 李嘉恒 周方琰

Low-light pedestrian detection and tracking algorithm based on autoencoder structure and improved Bytetrack

REN Zelin, PANG Lan, WANG Chao, LI Jiaheng, ZHOU Fangyan

引用本文:
任泽林, 庞澜, 王超, 等. 基于自编码器结构与改进Bytetrack的低光照行人检测及跟踪算法[J]. 应用光学, 2024, 45(3): 616-629. DOI: 10.5768/JAO202445.0302001
REN Zelin, PANG Lan, WANG Chao, et al. Low-light pedestrian detection and tracking algorithm based on autoencoder structure and improved Bytetrack[J]. Journal of Applied Optics, 2024, 45(3): 616-629. DOI: 10.5768/JAO202445.0302001

在线阅读 View online: https://doi.org/10.5768/JAO202445.0302001

文章编号：1002-2082 (2024) 03-0616-14

基于自编码器结构与改进 Bytetrack 的低光照行人检测及跟踪算法

任泽林，庞 澜，王 超，李嘉恒，周方琰
（西安应用光学研究所，陕西 西安 710065）

摘 要：针对夜间低光照场景下目标特征提取困难和跟踪不稳定的问题，提出了基于自编码器结构及改进 Bytetrack 的多目标行人检测及跟踪算法。在检测阶段，基于 YOLOX（you only look once X）搭建多任务自编码变换模型框架，以一种自监督的方式考虑物理噪声模型和图像信号处理（image signal processing, ISP）的过程，通过对真实光照退化变换过程进行编码与解码学习内在视觉结构，并基于这种表示通过解码边界框坐标与类实现目标检测任务。为了抑制背景噪声的干扰，在目标解码器颈部网络引入自适应特征融合模块 ASFF。跟踪阶段，基于 Bytetrack 算法进行改进，将基于 Tranformer 重识别网络提取到的外观嵌入信息与 NSA 卡尔曼滤波获得的运动信息通过自适应加权的方法完成数据关联，并通过 Byte 两次匹配的算法完成夜间行人的跟踪。在自建夜间行人检测数据集上测试检测模型的泛化能力，mAP@0.5 达到了 94.9%，结果表明本文的退化变换过程符合现实条件，具有良好的泛化能力。最后通过自建夜间行人跟踪数据集验证多目标跟踪性能，实验结果表明，本文提出的夜间低光照行人多目标跟踪算法 MOTA（multiple object tracking accuracy）为 89.55%，IDF1（identity F1 score）为 88.34%，IDs（ID switches）为 15。与基准方法 Bytetrack 相比，MOTA 提高了 10.72%，IDF1 提高了 6.19%，IDs 减少了 50%。结果表明，本文提出的基于自编码结构及改进 Bytetrack的多目标跟踪算法可以有效解决在夜间低光照场景下行人跟踪困难的问题。

关键词：多任务自编码变换；低光照；YOLOX；目标检测；多目标跟踪
中图分类号：TN223 **文献标志码：**A **DOI:** 10.5768/JAO202445.0302001

Low-light pedestrian detection and tracking algorithm based on autoencoder structure and improved Bytetrack

REN Zelin, PANG Lan, WANG Chao, LI Jiaheng, ZHOU Fangyan
（Xi'an Institute of Applied Optics, Xi'an 710065, China）

Abstract: Aiming at the problems of difficult target feature extraction and unstable tracking in low-light scenes at night, a multi-target pedestrian detection and tracking algorithm based on the autoencoder structure and improved Bytetrack was proposed. In the detection phase, a multi-task auto-encoding transformation model framework based on you only look once X (YOLOX) was built, considering the physical noise model and image signal processing (ISP) process in a self-supervised manner, learning the intrinsic features by encoding and decoding the real illumination degradation transformation process, and realizing the object detection tasks by decoding bounding box coordinates and classes based on this representation. In order to suppress the interference of background noise, the adaptive feature fusion module ASFF was introduced in the target

收稿日期：2023-10-27； 修回日期：2024-04-15
作者简介：任泽林（1998—），男，硕士研究生，主要从事目标检测与跟踪研究。E-mail: 3218643113@qq.com
通信作者：周方琰（1998—），男，硕士研究生，主要从事光电跟踪系统伺服控制研究。E-mail: 1183650916@qq.com

decoder neck network. In the tracking phase, it was improved by Bytetrack algorithm, and the appearance embedded information extracted based on the Transformer re-identification network as well as the motion information obtained by the NSA Kalman filter was used to complete the data association through an adaptive weighting method, and the Byte twice matching algorithm was used to complete the tracking of pedestrian at night. The generalization ability of the detection model was tested on the self-built night pedestrian detection data set, and the mAP@0.5 reached 94.9%. The results showed that the proposed degradation transformation process met the realistic conditions and had good generalization ability. Finally, the multi-target tracking performance was verified through the night pedestrian tracking data set. The experimental results show that the proposed multiple object tracking accuracy (MOTA) of the night low-light pedestrian multi-target tracking algorithm is 89.55%, the identity F1 score (IDF1) is 88.34%, and the ID switches (IDs) are 15. Compared with the baseline method Bytetrack, the MOTA is improved by 10.72%, the IDF1 is improved by 6.19%, and the IDs are reduced by 50%. The results show that the proposed multi-target tracking algorithm based on the autoencoding structure and improved Bytetrack can effectively solve the problem of difficult pedestrian tracking in low-light scenes at night.

Key words: multi-task auto-encoding transformation; low illumination; YOLOX; target detection; multi-target tracking

引言

随着深度学习在计算机视觉领域的广泛应用, 无人驾驶、智能安防及智慧交通等技术得到了快速发展。其中对目标的自动检测与跟踪是实现智能化与自动化的前提。现如今, 随着城市发展的日新月异, 城市的夜间照明系统越来越完善, 但仍然存在许多因光照不足或建筑物遮挡产生的低光照区域, 这给监控系统、交通管理系统以及自动驾驶系统对行人的检测与跟踪造成了较大的困难。因此, 研究针对低光照情况下的行人多目标跟踪算法, 对于帮助监控系统在黑暗中识别和追踪可疑人员或行为、帮助交通管理系统在夜间或恶劣天气中检测和跟踪行人、帮助自动驾驶系统在黑暗或微光环境中感知和预测行人的位置和运动均具有积极的作用。

针对低光照情况下的检测问题, 一些研究选择使用红外传感器或热成像的成像设备[1], 但这样的方法成本较高, 同时红外成像分辨率较低, 无法获取目标的具体特征, 对于检测后跟踪会造成较大的困难。所以更多的研究者选择研究针对低光照情况下的目标检测算法, 用于改善光照不足造成检测性能差的问题。然而主流的目标检测算法都是基于正常光照成像清晰的图片设计的, 没有考虑弱光情况下成像质量差造成的影响, 所以弱光照图片所训练出的检测模型往往具有较差的检测性能[2]。为了减轻低光照成像质量差的问题, 一些研究者通过改善目标检测网络的结构以获取更强的特征表达能力, 如 LUO Y 等[3] 在 Faster R-CNN 原始网络的特征金字塔网络(FPN)基础上增加了基于自适应直方图均衡化(AHE)的亮度增强模块和基于通道注意力机制(CA)的对比度增强模块, 但这种算法对不同低光照场景的适应性不足, 因为自适应直方图均衡化和通道注意力机制都是基于数据驱动的方法, 它们的性能会受到训练数据集的影响。所以更多的研究者将图像增强作为目标检测的预处理过程, 以此减轻低光照对成像质量的影响。WANG K 等[4] 提出一种深度卷积对抗生成网络与 Faster RCNN 结合的低光照图像目标检测算法, 该算法通过对抗生成网络复原低光照图像, 生成类似于正常光照图像的图像, 并在 Faster RCNN 中加入特征融合和多尺度池化, 以得到更高精度的检测结果, 但低光照图像增强的算法在增强过程中可能会产生伪影, 这会对后续的视觉任务产生误导。

目前基于检测的多目标跟踪算法在多个场景下表现出较强的应用价值, 它将多目标跟踪分为目标检测和数据关联两个子任务。因为基于检测的多目标跟踪算法依赖于高性能目标检测器生成准确的目标位置, 而低光照场景下难以提取和融

合行人高层次的特征，无法有效解决跟踪的泛化性和适应性问题，同时缺乏针对低光照场景的多目标跟踪数据集，所以其在低光照场景下还未广泛运用。

针对上述问题，提出一种结合自编码结构与改进 Bytetrack 的低光照行人检测及跟踪算法，克服了图像增强算法先增强后检测，难以应用于后续下游任务以及监督学习的目标检测方法依赖于低光照数据集且泛化性较差的问题。本文基于 YOLOX 网络搭建多任务自编码变换模型，同时在目标检测解码器颈部网络引入自适应特征融合模块，抑制背景噪声的干扰。通过正常光照与合成低光照行人数据集同时完成自监督学习的训练和低光照目标检测任务。跟踪算法基于 Bytetrack 改进，将基于 Transformer 网络提取的外观特征以一种自适应嵌入的方式与 NSA 卡尔曼滤波获得的运动特征相结合，并输入至 Byte 算法两次匹配的网络中完成数据关联。在进行跟踪测试时，直接将采集暗光视频帧输入至多任务自编码变换模型的暗光路径，并通过目标检测解码器将视频帧检测结果输出至本文跟踪算法完成行人跟踪。通过自建夜间行人检测与跟踪数据集验证算法有效性，实验表明本文提出的基于检测的多目标跟踪算法能够有效提升低光照场景下提取目标外观特征的能力，并有效减少了跟踪时行人身份 (ID) 切换的次数，实现低光照场景下行人稳定且准确的跟踪。

1 总体研究框架

图 1 为本文基于自编码结构的低光照场景行人检测及跟踪算法整体研究框架。在检测阶段，基于 YOLOX 网络实现多任务自编码变换模型，将正常光照的行人图片与通过低光照腐化降质过程合成的低光照图片分别输入编码器 E 的两个路径进行自监督学习的训练，通过对光照退化变换过程进行编码与解码来学习内在视觉结构，并基于这种表示方式，通过目标检测解码器完成目标检测任务。在跟踪阶段，提出基于 Bytetrack 改进的多目标跟踪算法，将行人重识别网络提取的外观嵌入信息与运动信息通过自适应加权的方式结合起来，并输入至 Byte 算法两次匹配的网络中完成数据关联的过程。搭建基于检测的低光照场景行人跟踪算法，输入视频序列，将视频帧输入至多任务自编码变换模型中的暗光路径完成检测任务，然后通过目标检测解码器将检测结果输出给跟踪算法。在跟踪阶段，利用行人重识别网络提取当前检测框的外观嵌入信息，然后通过与运动信息完成数据关联，再输入至 Byte 两次匹配的网络中完成低光照场景行人的跟踪并输出视频序列结果。

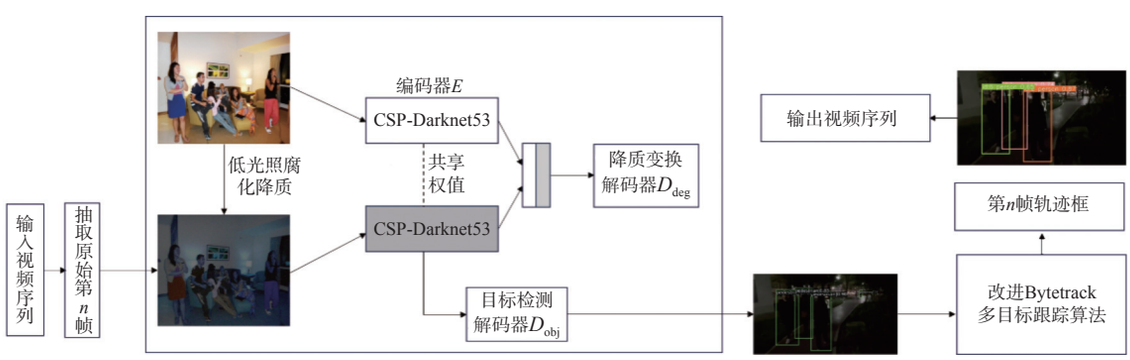


图 1 总体研究框架流程图

Fig. 1 Flow chart of overall research framework

2 基于自编码变换的低光照行人目标检测算法

自编码变换（autoencoding transformation, AET）是一种自监督的表示学习方法，它可以从无标注的数据中学习神经网络的特征表示[5]。其基本思想是：给定一个随机的图像变换，比如旋转、缩放、裁剪等，AET 从编码特征中预测出变换的类型和参数，而不是通过重构原始图像。这样的方法可以使编码特征具有对变换的不变性和对视觉结构的敏感性，从而提高特征的质量和泛化性。AET 学习具有代表性的潜在特征，这些潜在特征基于变换 t 对原始图像 x 和变换后的对应图像 $t(x)$ 进行

解码或恢复参数化变换。AET 由孪生表示编码器 E 和转换解码器 D 组成。编码器 E 从 x 和变换 $t(x)$ 中提取特征，这些特征能够捕捉内在的视觉结构以解释变换 t。解码器 D 使用编码后的 $E(x)$ 和 $E(t(x))$ 对 t 的估计值 $\hat{t}$ 进行解码，其计算如式(1)所示：

$$\hat{t} = D[E(x), E(t(x))] \tag{1}$$

为解决低光照场景下图像的退化问题，引入多任务自编码变换(multitask autoencoding transformation, MAET)算法。MAET 是基于 AET 算法的延伸，其编码器 E 通过图像低照度退化变换学习退化过程的潜在特征，使用多任务解码器分别对图像退化参数和目标检测参数进行解码。通过光照退化变换过程的自监督表示学习，无需图像增强，直接在低光照环境下检测目标。

2.1　基于 YOLOX 网络建立多任务自编码变换模型

本文基于 YOLOX 搭建多任务自编码变换模型[6]，模型架构如图 2 所示。搭建的网络包括编码器 E 和 D。编码器 E 采用共享权值的孪生网络结构并使用 YOLOX 的 CSP-DarkNet53 网络作为主干网络，训练过程中，正常光照的图像输入到 E 的左路径，退化后的图像输入到 E 的右路径。解码器 D 由降质变换解码器 D_{deg} 和目标检测解码器 D_{obj} 组成，D_{deg} 侧重于解码弱光变换 t_{deg} 的参数，D_{obj} 侧重于解码目标类别与位置的信息。如图 2 所示，将编码的潜在特征 $E(x)$ 和 $E(t_{\mathrm{deg}}(x))$ 串联在一起，传递给解码器 D_{deg} 估计相应的退化变换 t_{deg}。目标检测解码器 D_{obj} 仅对右侧暗光路径中的表示进行解码，以预测目标检测的参数。这种自监督学习的方式有助于 MAET 在各种光照退化变换中学习无标记数据的内在视觉结构[7]，所以在使用该模型进行测试时，可以直接将弱光图像输入给 MAET 编码器的暗光图像路径，以获得解码的目标类别与位置的信息。

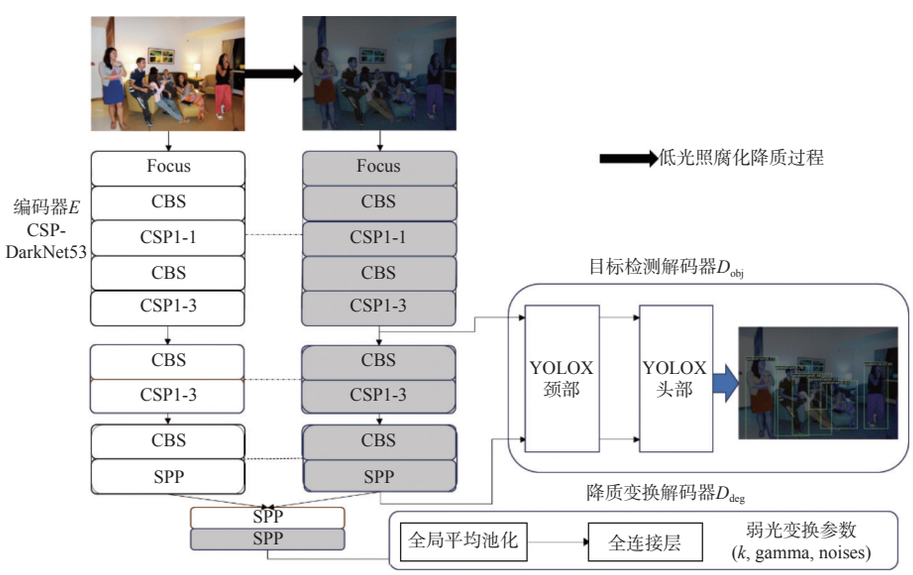


图 2　多任务自编码变换模型框架(MAET)

Fig. 2　Model framework of multi-task autoencoding transformations (MAET)

将正常光照图像通过腐化降质过程转换为低光照图像，其中涉及把图像从 RGB 空间转换到 RAW 空间的逆处理过程，RAW 空间上的低光照腐化降质过程和正向的图像信号处理(image signal processing, ISP)[8]过程。如图 3 所示，图像从 RGB 空间到 RAW 空间的逆处理过程包括反转色调映射、反转伽马校正、sRGB 空间转换到 cRGB 空间、反转白平衡。低光照腐化降质过程对 RAW 图像进行线性衰减，并再通过散粒噪声与读出噪声对图像进行腐化降质[9]。最后通过量化、白平衡、cRGB 空间转换到 sRGB 空间、伽马校正的 ISP 过程生成 RGB 空间的低光照图像。在此基础上对低光照腐化降质任务进行参数化建模，其计算公式如式(2)所示：

$$t_{\mathrm{deg}}(x) = t_{\mathrm{ISP}}(k \cdot t_{\mathrm{unprocess}}(x) + x_{\mathrm{noise}} + x_{\mathrm{quan}}) \tag{2}$$

式中：t_{ISP} 表示图 3 中从白平衡到伽马校正的过程；$t_{\mathrm{unprocess}}$ 表示逆图像信号处理的过程；x_{noise} 为低光照腐化降质中对噪声建模的参数；x_{quan} 表示量化后的噪声参数。

光照退化变换与目标检测这两个任务虽然相

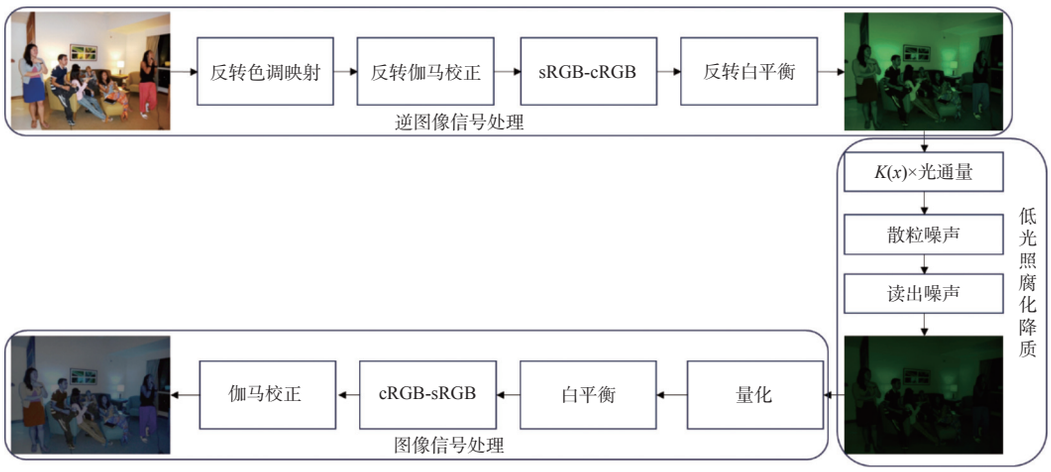


图 3 光照退化降质流程图

Fig. 3 Flow chart of low-illumination degradation

关，但它们的输出反映了输入图像两个不同的方面，所以需要施加正交规则以解耦两个不同任务输出之间不必要的依赖。多任务自编码变换模型正交解耦的目标，是通过最小化余弦相似度的绝对值以解耦目标和退化特征，其计算公式如式(3)所示：

$$L_{\text{ort}} \triangleq \sum_{k,l}\left|\cos\theta_{k,l}\right| = \sum_{k,l}\frac{\left|\left[\frac{\partial E}{\partial D^{k}_{\text{deg}}}\right]^{\text{T}}\cdot\left[\frac{\partial E}{\partial D^{l}_{\text{obj}}}\right]\right|}{\left\|\frac{\partial E}{\partial D^{k}_{\text{deg}}}\right\|\cdot\left\|\frac{\partial E}{\partial D^{l}_{\text{obj}}}\right\|} \tag{3}$$

式中：$\partial E/\partial D^{k}_{\text{deg}}$和$\partial E/\partial D^{l}_{\text{obj}}$分别为编码器 E 沿着光照退化变换和目标检测任务第 k 和第 l 个输出坐标形成的表示流形的切线。

2.2 目标检测解码器颈部网络引入自适应特征融合模块

在低照度图像中，目标的特征信息不明显且容易受到背景噪声的干扰。为了使多任务自编码变换模型在检测任务的学习过程中更好地融合低光照图像不同尺度的特征，过滤掉背景噪声与冲突信息，本文在 MAET 目标检测解码器的 YOLOX 颈部网络 PAFPN(path aggregation feature pyramid network) 后，加入用于特征增强的自适应特征融合模块 ASFF(adaptively spatial feature fusion)，改进后的目标检测解码器网络结构图如图 4 所示。

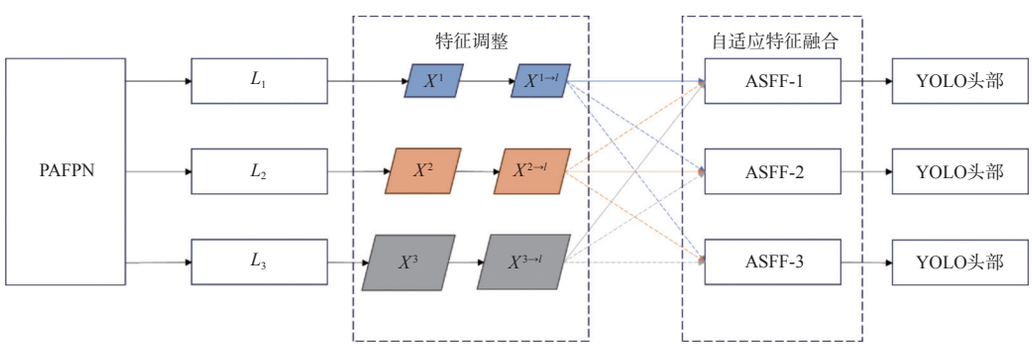


图 4 加入 ASFF 的目标检测解码器结构图

Fig. 4 Structure diagram of target detection decoder with ASFF

PAFPN 网络输出 3 个尺度的特征图 $L_l(l\in\{1,2,3\})$，将每一层的尺度特征设为X^l。在融合其他层特征前将其他层L_n的特征X^n调整到和X^l一样的尺度大小，然后通过自适应融合的方式将不同层的特征融合起来[10]。自适应特征融合的计算过程为，首先经过网络训练1×1卷积得到 3 个空间可学习权重α、β、γ，然后将通过尺度缩放后的特征图与其相对应的空间权重相乘得到特征融合结果。以 ASFF-1 为例，α^1_{ij}、β^1_{ij}、γ^1_{ij}与通过特征调整得到的 3 个尺度特征图$X^{1\to1}$、$X^{2\to l}$、$X^{3\to1}$逐元素相乘再相加得到输出Y^1，因此对于Y^1上 (i,j) 位置的特征Y^1_{ij}计算公式如(4)所示：

$$Y_{ij}^{l}=\alpha_{ij}^{l}\cdot X^{1\to l}+\beta_{ij}^{l}\cdot X^{2\to l}+\gamma_{ij}^{l}\cdot X^{3\to l} \tag{4}$$

式中: $\alpha_{ij}^{l}+\beta_{ij}^{l}+\gamma_{ij}^{l}=1$, 且$\alpha_{ij}^{l},\beta_{ij}^{l},\gamma_{ij}^{l}\in[0,1]$。

3　基于 Bytetrack 改进的多目标跟踪算法

Bytetrack 是先检测再跟踪范式(tracking by detection,TBD)的多目标跟踪方法, 这种框架将整体的跟踪过程分为检测和关联两个独立的任务, 先使用已有的目标检测器进行检测, 再使用独立的网络完成数据关联[11]。不同于此前大部分 TBD 范式的多目标跟踪算法直接将低分检测框舍弃, Bytetrack 提出了一种简单且高效的数据关联算法 Byte。Byte 将检测结果分为高分检测框和低分检测框, 将高分检测框与轨迹关联的同时, 利用低分检测框和跟踪轨迹之间的相似性, 从低分框中挖掘出真正的目标并过滤掉背景。虽然 Byte 的数据关联算法在正常光照条件下取得了很好的效果, 但在低光照场景下, 由于目标的视觉特征不稳定以及更易遮挡, 仅使用目标位置信息计算相似度会导致跟踪不稳定和错误的关联。因此, 本文基于 Bytetrack 算法进行改进: 第一, 使用基于 Transformer 的行人重识别模型 transreid 与 NSA 噪声自适应卡尔曼滤波算法分别获取外观特征与运动特征; 第二, 将外观特征以一种自适应加权的方式与运动特征结合计算相似度矩阵。

3.1　基于 transreid 获取外观特征

当检测器获得目标在图像中的位置与大小之后, 将从图像中截取目标并输入至特征外观提取网络中, 获得目标在图像中的外观特征信息。此前, 基于卷积神经网络的目标重识别网络 osnet 被广泛应用于多目标跟踪的特征提取模块[12], 但基于卷积的特征提取方法擅长获取局部信息, 不利于建立全局信息的连接, 并且会因为卷积与下采样导致细节信息丢失。这两个局限性会导致基于卷积的特征提取方法在低光照情况下无法提取准确的特征信息。而基于 Transformer 的目标重识别方法通过引入多个注意力机制, 去除卷积和下采样, 解决了基于卷积神经网络的上述问题[13], 使得在低光照条件下获得更强的特征学习能力。

本文基于 Transformer 的行人重识别网络 transreid 的结构如图 5 所示。将输入图片切片为 N 个相同大小的图像块, 经过线性投影映射到 D 维。在 D 维向量前加入类别向量(图中的$*$), 然后添加可学习的位置嵌入信息后送入 Transformer 网络, 通过 L 个 Transformer 层学习外观特征。监督学习的部分通过为全局特征构建分类损失和三元组损失共同优化训练过程。

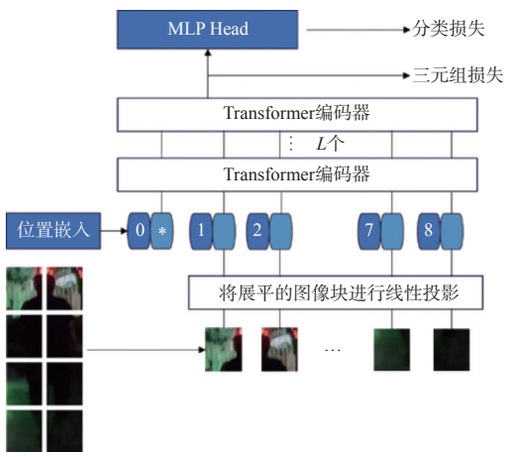


图 5　基于 Transformer 的行人重识别网络

Fig. 5　Transformer-based pedestrian re-identification network

3.2　基于 NSA 卡尔曼滤波算法获取运动特征

基于线性运动估计的卡尔曼滤波算法被广泛应用于 TBD 范式的多目标跟踪算法中, 但在低光照场景中, 检测质量参差不齐, 低质量的检测结果会导致运动状态估计有较大偏差。为了获取更准确的运动特征, 本文引入 NSA 卡尔曼滤波算法[14]。NSA 卡尔曼滤波算法基于检测置信度得分, 以一种自适应的方式计算噪声协方差$\tilde{R}_k$, 如式(5)所示:

$$\tilde{R}_k=(1-c_k)\boldsymbol{R}_k \tag{5}$$

式中: $\boldsymbol{R}_k$为预先设定的常数噪声协方差矩阵; c_k为状态 k 下的检测置信度分数。

根据自适应噪声协方差, 可以得到改进后的卡尔曼滤波增益计算公式, 如式(6)所示:

$$K_k=\boldsymbol{P}_{k|k-1}\boldsymbol{H}_k^{\mathrm{T}}(\boldsymbol{H}_k\boldsymbol{P}_{kk-1}\boldsymbol{H}_k^{\mathrm{T}}+\tilde{R}_k)^{-1} \tag{6}$$

式中: $\boldsymbol{P}_{k|k-1}$为状态协方差矩阵; $\boldsymbol{H}_k$为观测矩阵。

3.3　基于 Byte 与自适应外观特征嵌入的数据关联算法

本文将 transreid 获取的外观特征以一种自适应嵌入的方式与 NSA 卡尔曼滤波获取的运动特征结合起来, 并输入至 Byte 两次匹配的网络中完成数据关联。

标准的计算相似度矩阵的方法如 DeepSORT[15], 使用外观嵌入信息与运动模型通过计算相似性得到$m\times n$的代价矩阵$\boldsymbol{A}_{\mathrm{c}}[m,:]$, 并通过与 IOU(intersection over union) 成本矩阵$\boldsymbol{I}_{\mathrm{c}}$结合得到相似度矩

阵 $\boldsymbol{C}$, 其计算公式为 $\boldsymbol{C}=\boldsymbol{I}_{\mathrm{c}}+a_{\mathrm{w}}\boldsymbol{A}_{\mathrm{c}}$。本文引入的自适应外观特征嵌入方法通过外观嵌入信息的判别性来增加外观特征的权重, 将权重因子 $w_{\mathrm{b}}(m,n)$ 加入到全局 a_{w} 中, 根据区分度提高单个跟踪框的得分[16]。设 τ_m 表示轨迹, D_n 表示检测框。当 τ_m 只与 1 个检测框(包含在 $\boldsymbol{A}_{\mathrm{c}}[m,:]$ 行中)具有较高的相似性分数时, 在 $\boldsymbol{A}_{\mathrm{c}}[m,:]$ 行上增加外观权重, 如果 D_n 只与 1 条轨迹区别地关联, 同样在 $A_{\mathrm{c}}[:,n]$ 列上增加外观权重。使用 Z_{diff} 来衡量检测框-轨迹对的判别性, $Z_{\mathrm{diff}}^{\mathrm{det}}$ 和 $Z_{\mathrm{diff}}^{\mathrm{track}}$ 分别定义为 1 行和 1 列的最高值和次低值之差, 其计算公式为

$$Z_{\mathrm{diff}}^{\mathrm{det}}(\boldsymbol{A}_{\mathrm{c}},n)=\min(\max_i \boldsymbol{A}_{\mathrm{c}}[i,n]-\max_{j\neq i}\boldsymbol{A}_{\mathrm{c}}[j,n],\in) \quad (7)$$

$$Z_{\mathrm{diff}}^{\mathrm{track}}(\boldsymbol{A}_{\mathrm{c}},m)=\min(\max_i \boldsymbol{A}_{\mathrm{c}}[m,i]-\max_{j\neq i}\boldsymbol{A}_{\mathrm{c}}[m,j],\in) \quad (8)$$

式中: $\in$ 为一个超参数, 用于在第 1 和第 2 最佳匹配之间存在较大外观成本差异的情况下限制权重的增加。权重因子 $w_{\mathrm{b}}(m,n)$ 由式(7)和式(8)推导出, 即:

$$w_{\mathrm{b}}(m,n)=\frac{[z_{\mathrm{diff}}^{\mathrm{track}}(\boldsymbol{A}_{\mathrm{c}},m)+z_{\mathrm{diff}}^{\mathrm{det}}(\boldsymbol{A}_{\mathrm{c}},n)]}{2} \quad (9)$$

根据 IOU 距离与权重因子 $w_{\mathrm{b}}(m,n)$, 可得到改进后的相似度矩阵计算公式为

$$\boldsymbol{C}[m,n]=\boldsymbol{I}_{\mathrm{ou}}[m,n]+[a_{\mathrm{w}}+w_{\mathrm{b}}(m,n)]\boldsymbol{A}_{\mathrm{c}}[m,n] \quad (10)$$

通过将外观嵌入信息自适应地集成到基于运动的方法中, 能够有效解决由于遮挡、运动模糊或相似外观的对象产生的显著噪声, 提高跟踪鲁棒性。

跟踪网络的整体框架如图 6 所示, 跟踪网络采用 Byte 两次匹配的方法, 第 t 帧检测器输出的目标检测框(BBOX)根据高低置信度阈值 0.45、0.1 分为高分检测框 D_{high}、低分检测框 D_{low}。第 1 次匹配优先考虑 D_{high} 进行关联, 与 t–1 帧检测框 NSA 卡尔曼滤波预测的结果进行 IOU 匹配, 然后 D_{high} 提取到的外观嵌入信息(ReID)与 IOU 信息通过自适应加权进行关联, 得到两种匹配结果: 对于匹配成功的轨迹 T_{match}, 通过使用 NSA 卡尔曼滤波进行轨迹更新; 对于未匹配成功的轨迹 T_{remain}, 进行下步匹配并初始化未匹配高分检测框, 此时便完成检测框与预测框的第 1 次匹配。然后将 T_{remain} 与 D_{low} 进行第 2 次匹配, 再次得到两种匹配结果: 匹配成功的轨迹 T_{match}、未匹配成功的轨迹 $T_{\mathrm{re-remain}}$。对于 T_{match}, 将其再次送入 NSA 卡尔曼滤波中更新轨迹信息; 对于 $T_{\mathrm{re-remain}}$, 若其连续 30 帧率未匹配上检测框, 则被删除, 反之加入跟踪序列中, 同时将未匹配的低分检测框视为背景删除。最后, 将得到的跟踪序列利用 NSA 卡尔曼滤波进行预测, 再依次与高分检测框、低分检测框进行关联匹配, 重复上述步骤。

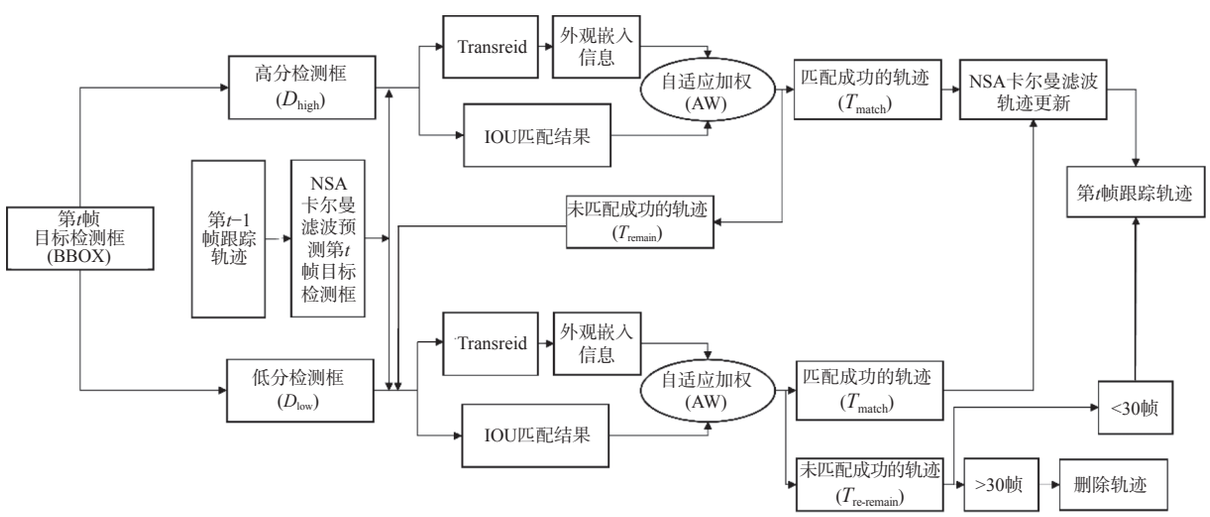


图 6 跟踪网络整体框架图

Fig. 6 Overall framework diagram of tracking network

4 实验与分析

实验使用的硬件为美超微 SYS-4028GR-TR 服务器, GPU 为 GeForce RTX 1080Ti, 处理器为 Intel Xeon(R) ES-2630 2.20 GHz。软件平台是 64 位 Ubuntu22.04 系统, 基于 Pytorch 深度学习框架搭建实验环境。

4.1 数据集与评价指标

采用 CrowdHuman[17] 密集人群检测数据集、MOT-17 多目标跟踪数据集, 以及自建夜间行人检测与跟踪数据集对本文提出的检测和跟踪算法进行验证。CrowdHuman 数据集为旷视科技发布的用于评估人群场景中的人体检测的基准数据集, 包含

15000 张训练集、4370 张验证集和 5000 张测试集, 每张图像平均包含 23 个人并存在各种各样的遮挡。MOT17 数据集是一个用于多目标跟踪的数据集, 包括 7 个带有行人的室内外公共场所场景, 每个场景的视频被分为 2 个片段, 分别用于训练和测试。自建夜间行人检测与跟踪数据集为在摄像机拍摄的多个低光照场景的视频中筛选出具有较多行人移动的 15 段视频, 每段视频时长 10 s～20 s, 使用 darklabel 软件标注这 15 段视频, 制作符合 MOT 格式的夜间行人跟踪数据集。在 15 段视频的所有视频帧中使用 labelImg 标注 1000 张相似度较低, 且包含行人相互遮挡、光照不均、目标运动模糊等各种情况的图像用作模型泛化性测试数据集。

目标检测算法对比实验采用 IOU 阈值为 0.5～0.95 的平均精度均值 mAP@0.5:0.95 和 IOU 阈值为 0.5 的平均精度均值 mAP@0.5 作为评价指标, 目标检测算法泛化性实验采用 IOU 阈值分别为 0.5 与 0.75 的平均精度均值 mAP@0.5 和 mAP@0.75 作为评价指标。跟踪算法采用 MOT Challenge[18] 提供的官方多目标跟踪评测指标, 评价指标如表 1 所示。

表 1　多目标跟踪评价指标

Table 1　Multi-target tracking evaluation indicators

评价指标	定义
HOTA	高阶跟踪精度
MOTA	多目标跟踪准确率
IDF1	正确识别的检测与平均真实数和计算检测数之比
IDs	ID切换总次数
FP	误检目标总数
FN	漏检目标总数
FPS	帧率

4.2　目标检测算法实验结果与分析

4.2.1　目标检测算法对比实验

本文目标检测算法通过 CrowdHuman 训练集及 MOT17-half 训练集在 4 块 GeForce RTX 1080Ti 显卡上进行 270 个 epoch 的迭代训练, 所有输入图像被裁剪为 608×608 像素大小, 主干网络 CSP-DarkNet53 使用 ImageNet 预训练模型进行初始化。训练过程采用 SGD 优化器, 并将图像批处理大小设置为 2。将权重衰减系数设置为 5e-4, 动量设置为 0.9。编码器 E 与目标检测解码器 $D_{\rm obj}$ 的初始学习率设置为 5e-2, 降质变换解码器 $D_{\rm deg}$ 的初始学习率设置为 5e-3, 同时使用余弦退火策略动态优化学习率。

为了验证本文低光照检测算法的有效性, 以本文低光照检测算法中的图像退化过程合成低光照 CrowdHuman 测试集, 并将该测试集作为本文检测算法对比实验的基准测试数据集。以 CrowdHuman 及 MOT17-half 训练集为基础, 分别以正常光照训练集与低光照合成训练集训练 YOLOX 模型作为参考, 然后分别将本文基于自编码变换的低光照检测算法与图像增强算法及当前主流的目标检测算法做对比。如表 2 所示, normal 表示使用正常光照训练集, lowlight 表示使用合成低光照训练集。

表 2　目标检测算法对比实验结果

Table 2　Comparative experimental results of target detection algorithms

算法	训练集	测试集	图像预处理	mAP@0.5:0.95	mAP@0.5
YOLOX	normal	low	无	0.196	0.371
YOLOX	normal	low	Zero DCE	0.165	0.325
YOLOX	normal	low	Kind	0.160	0.319
YOLOX	normal	low	URetinexNet	0.201	0.380
YOLOX	low	low	无	0.299	0.635
YOLOX-ASFF	low	low	无	0.325	0.664
Faster-RCNN	low	low	无	0.282	0.602
Centernet	low	low	无	0.284	0.599
RTMDet	low	low	无	0.384	0.701
本文算法	normal+low	low	无	0.362	0.682

从表 2 可知, 本文基于自编码变换的低光照算法在合成低光照数据集上体现了良好的检测性能, 与使用 Zero DCE[19]、Kind[20]、URetinexNet[21] 用作测试前图像增强的算法相比, 检测精度有了大幅度的提升, 且本文算法与大部分主流使用合成低光照数据集进行监督学习的检测算法相比有着较好的检测性能, 与基准算法 YOLOX 相比 mAP@0.5:0.95 提升了 6.3%, mAP@0.5 提升了 4.7%, 但本文算法与当前检测性能较好的算法, 如在合成低光照数据集时直接进行监督学习的方法 RTMDet 相比[22], 仍有一定差距。

4.2.2　目标检测算法泛化性实验

虽然本文基于自编码变换的低光照检测算法在合成低光照 CrowdHuman 测试集上获得了良好的检测性能, 但合成低光照数据集与真实低光照

数据集仍有一定差距，为了验证本文低光照检测算法的泛化能力，选取真实低光照场景行人检测测试集进行测试。如表 3 所示，将本文算法与 4.1 节中使用合成低光照数据集进行监督学习训练的目标检测方法做对比，并将各个算法分别使用合成低光照数据集进行监督学习和使用图像增强算法进行预处理的结果进行对比。

从表 3 可以看出，本文算法在 mAP@0.5 和 mAP@0.75 这两个评价指标上都取得了最好的结果，与基准算法 YOLOX 相比，本文算法在自建真实场景行人检测数据集上 mAP@0.5 提升了 2%，mAP@0.75 提升了 3.2%。虽然 RTMDet 算法在合成低光照 CrowdHuman 测试集上取得了最好的检测效果，但在真实低光照场景中，本文算法的性能优于 RTMDet 算法，证明本文基于自编码变换的低光照检测算法有着较好的泛化能力，可以与 TBD 范式的多目标跟踪方法结合，实现低光照场景下的行人跟踪。

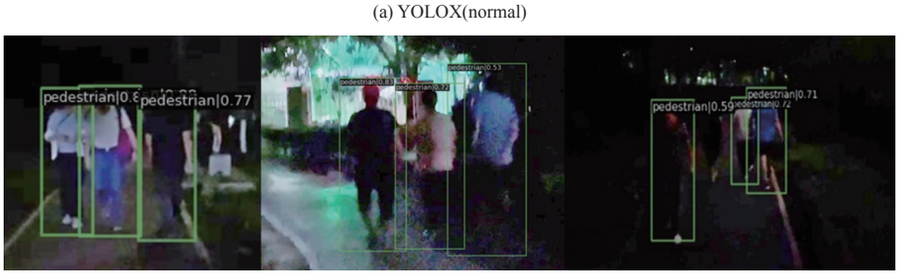
为了更加直观地体现本文低光照检测算法的效果，在自建夜间行人检测数据集上选取不同程度低光照的夜间行人图像进行可视化验证，检测结果如图 7 所示，可以看出本文算法在各种低光照环境下都具有良好的检测效果。

表 3　目标检测算法泛化性实验结果

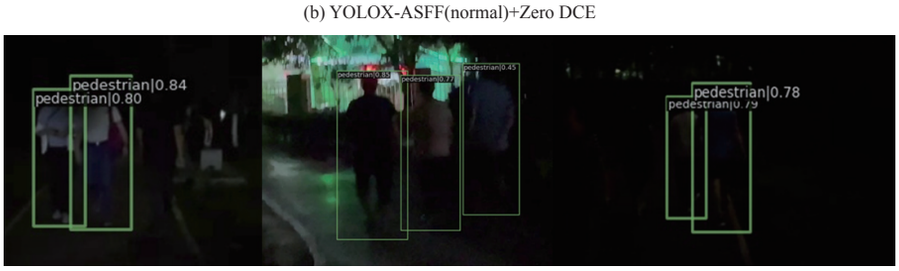
Table 3　Generalization experimental results of target detection algorithms

算法	训练集	图像预处理	mAP@0.5	mAP@0.75
Faster RCNN	low	无	0.892	0.690
	normal	Zero DCE	0.896	0.683
	normal	Kind	0.860	0.683
	normal	URetinexNet	0.881	0.694
Centernet	low	无	0.882	0.751
	normal	Zero DCE	0.879	0.744
	normal	Kind	0.864	0.732
	normal	URetinexNet	0.864	0.760
RTMDet	low	无	0.899	0.780
	normal	Zero DCE	0.895	0.760
	normal	Kind	0.806	0.681
	normal	URetinexNet	0.811	0.686
YOLOX-ASFF	low	无	0.945	0.760
	normal	Zero DCE	0.934	0.731
	normal	Kind	0.885	0.720
	normal	URetinexNet	0.905	0.746
本文算法	normal+low	无	0.949	0.795
YOLOX(基准)	normal	无	0.929	0.763

(a) YOLOX(normal)



(b) YOLOX-ASFF(normal)+Zero DCE



(c) YOLOX(low)

(d) 本文算法

图 7　不同算法检测结果对比

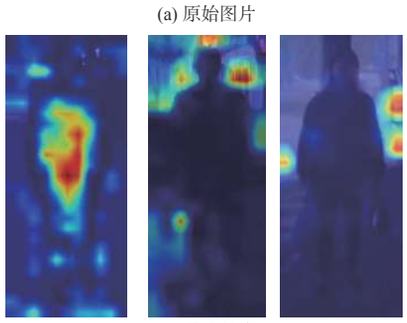
Fig. 7　Comparison of detection results of different algorithms

4.3　多目标跟踪算法实验结果与分析

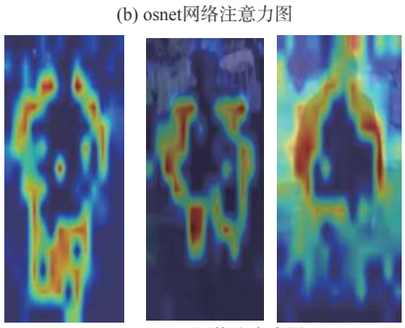
4.3.1　外观特征提取网络消融实验

为验证基于 Transformer 的外观特征提取网络在低光照条件下能更好地提取外观特征, 将其与 osnet 行人重识别网络在 Market1501 行人重识别数据集上进行训练, 并在本文低光照条件下的应用场景进行验证。图 8 展示了两个网络的注意力可视化结果, 可以看出本文使用的基于 Transformer 的特征提取方法相比 osnet, 在低光照场景下能够提取到行人更多的细节特征, 以区别不同身份的行人。

(a) 原始图片



(b) osnet网络注意力图



(c) transreid网络注意力图

图 8　注意力图可视化结果

Fig. 8　Visualization results of attention map

为进一步验证 Transformer 模型在低光照情况下对行人特征提取的优势, 仅将本文外观特征提取网络换为 osnet, 在夜间行人跟踪数据集上进行测试, 测试结果如表 4 所示, 可以看出 HOTA、MOTA、IDF1 分别下降了 1.71%、1.43%、3.03%, 证明基于 Transformer 的特征提取网络 tranreid 在低光照场景下表现出更好的特征提取能力。

表 4　外观特征提取网络消融实验结果

Table 4　Ablation experimental results of appearance feature extraction network

外观特征提取网络	HOTA/%	MOTA/%	IDF1/%
osnet	70.65	88.12	85.31
transreid	72.36	89.55	88.34

4.3.2　组件消融实验

为了验证本文改进 Bytetrack 算法自适应外观嵌入 AW 与 NSA 卡尔曼滤波的有效性, 对改进算法进行两个组件的消融实验, 实验结果如表 5 所示。可以看出, 引入自适应特征嵌入后, HOTA 提升了 0.49%, MOTA 提升了 1.57%, IDF1 提升了 2.03%, 跟踪性能获得了一定提升, 证明本文引入的自适应外观特征嵌入模块能够减小低光照场景下外观特征噪声的影响; 引入 NSA 卡尔曼滤波算法后, MOTA 提升到 89.55%, IDF1 提升到 88.34%, 表明 NSA 卡尔曼滤波能够减小低质量检测结果对运动状态估计的偏差。

表 5　改进组件消融实验结果

Table 5　Ablation experimental results of improved components

AW	NSA	HOTA/%	MOTA/%	IDF1/%
—	—	70.79	87.81	86.27
√	—	71.28	89.38	88.30
√	√	72.36	89.55	88.34

4.3.3 检测跟踪消融实验

为了衡量本文算法在低光照行人多目标跟踪方面的表现，以 YOLOX-Bytetrack 的检测跟踪算法为基准，设置高分检测框置信度阈值与低分检测框阈值分别为 0.45 和 0.1，对检测及跟踪阶段进行消融实验来验证本文检测与跟踪两阶段各自的改进是否有效，实验结果如表 6 所示，其中 MAET_YOLOX 为本文检测算法，BytetrackAW 为本文基于 Bytetrack 改进的跟踪算法。从表 6 中可以看出，本文检测阶段的基于自编码变换的低光照检测算法提高了检测精度，与基准算法相比，本文算法分别将 MOTA、IDF1 提高了 5.53% 和 2.69%，证实了本文低光照检测算法的有效性。跟踪阶段本文的改进 Bytetrack 算法有效降低了跟踪时 ID 切换的次数和误检数，与基准算法相比，ID 切换次数和误检数分别降低了 40% 和 41%。本文提出的基于检测的低光照场景行人跟踪算法最终在低光照行人跟踪数据集上达到了 89.55% 的跟踪精度，ID 切换次数降低了 50%，可以很好地满足低光照场景行人检测及跟踪的需求。

表 6 多目标跟踪结果对比

Table 6 Comparison of multi-target tracking results

算法	HOTA/%	MOTA/%	IDF1/%	IDs	FP	FN	FPS
YOLOX-Bytetrack	66.19	78.83	82.15	30	586	1 412	33.93
MAET_YOLOX-Bytetrack	69.13	84.36	84.84	25	477	994	29.32
YOLOX-BytetrackAW	67.14	83.02	82.99	18	346	1 249	16.37
MAET_YOLOX-BytetrackAW	72.36	89.55	88.34	15	326	651	9.98

为了更加直观地体现本文跟踪算法的有效性，选取自建低光照行人跟踪数据集中的一段视频，分别将本文算法与基准算法 YOLOX-Bytetrack 进行对比，由图 9 可以看出，Bytetrack 算法对于低光照场景行人相互遮挡的情况会发生明显的 ID 切换，而本文算法能够稳定跟踪行人，保持 ID 不变。

(a) YOLOX-Bytetrack

(b) 本文算法

图 9 多目标跟踪效果对比

Fig. 9 Comparison of multi-target tracking effects

4.3.4　不同场景实验结果

分别选取本文跟踪数据集中包含相似目标、目标遮挡及光照变换的复杂跟踪场景进行跟踪结果可视化验证。图 10 中 ID 为 2 的目标与 ID 为 5、6 的目标迎面走来并发生遮挡, 图 11 中 3 个目标行走过程中光照发生明显变化, 图 12 中 ID 为 5、6 的目标为相似目标且与 ID 为 1 的目标相互遮挡。可以看出, 在 3 种复杂场景中, 本文算法能够稳定跟踪目标保持 ID 不变, 证实了本文低光照场景行人跟踪算法在实际场景中应用的可行性。

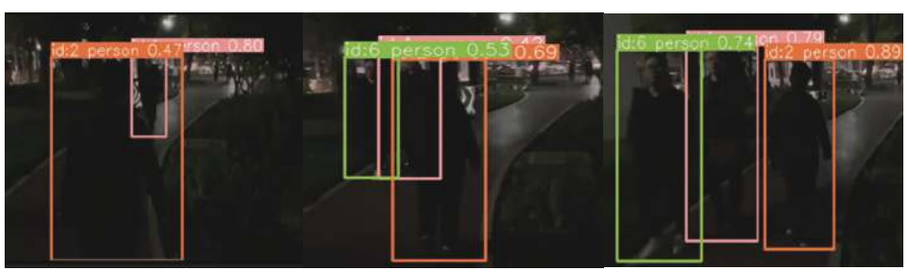


图 10　多目标跟踪结果 a

Fig. 10　Multi-target tracking effects a

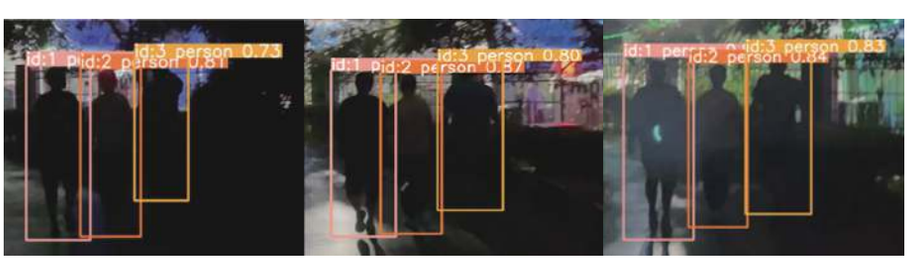


图 11　多目标跟踪结果 b

Fig. 11　Multi-target tracking effects b

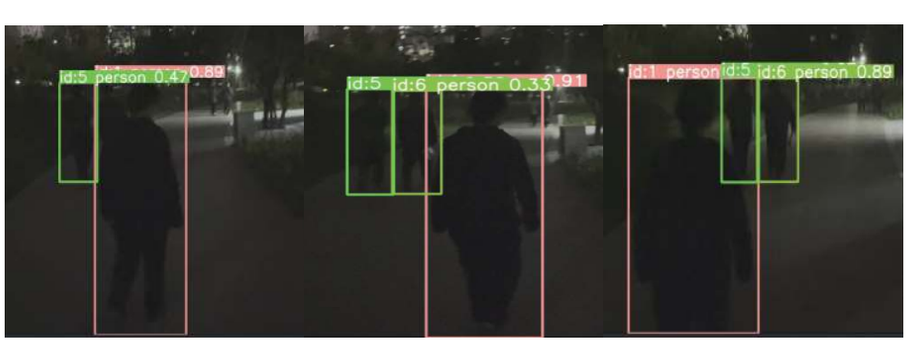


图 12　多目标跟踪结果 c

Fig. 12　Multi-target tracking effects c

5　结论

本文针对夜间低光照场景下目标的外观和运动因光照不足而模糊或不清晰的问题, 提出了基于自编码器结构与改进 Bytetrack 的行人检测及跟踪算法。基于 YOLOX 搭建多任务自编码变换模型, 以自监督的方式考虑物理噪声模型和图像信号处理的过程, 通过对现实光照退化变换进行编码和解码来学习内在视觉结构, 并基于这种表示通过解码边界框坐标和类别完成低光照目标检测任务。为了抑制背景无效目标对于检测的影响, 在本文目标检测解码器颈部网络后引入自适应特征融合模块 ASFF, 使得网络在检测任务的学习过程中能够更好地关注目标。在跟踪阶段, 提出基于 Bytetrack 改进的跟踪方法, 将 Transformer 重识别网络提取到的外观信息以自适应嵌入的方式与 NSA 卡尔曼滤波获得的运动信息结合, 并通过

Byte 算法两次匹配的方法强化目标在发生遮挡或光照变换时的跟踪能力。

通过将自编码器结构与改进 Bytetrack 算法结合，实现基于检测的低光照场景行人跟踪算法，并通过自建夜间行人检测与跟踪数据集验证本文算法的有效性。实验结果表明，本文基于自编码结构的目标检测模型在自建夜间行人检测数据集上的 mAP@0.5 达到了 94.9%，mAP@0.75 达到了 79.5%，与图像增强及监督学习的目标检测算法相比有着更好的模型泛化能力。在自建夜间行人跟踪数据集上测试本文算法跟踪效果，MOTA、IDF1 分别达到了 89.55%、88.34%，与基准算法相比分别提升了 10.72%、6.19%；同时 ID 切换为 15 次，相比基准算法减少了 50%，实现了较好的跟踪效果。

本文所提出的多目标跟踪算法可以有效解决夜间低光照场景下目标提取特征困难、跟踪不稳定及不连续的问题。但本文方法在实际工程部署中仍有一些困难，在视频图像分辨率低及目标较小的情况下，目标检测会受到背景噪声的干扰，无法保证跟踪效果。所以在后续的工作中，我们将从模型轻量化和如何加强弱小目标的特征提取上进行深入研究。

参考文献：

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

[2] 舒子婷, 张泽斌, 宋尧哲, 等. 基于改进YOLOv5的低光照图像目标检测[J]. 激光与光电子学进展, 2023, 60(4): 77-84.
SHU Ziteng, ZHANG Zebin, SONG Yaozhe et al. Low-light image object detection based on improved YOLOv5 algorithm[J]. Laser & Optoelectronics Progress, 2023, 60(4): 77-84.

[3] LUO Y, CAO X, ZHANG J, et al. CE-FPN: enhancing channel information for object detection[J]. Multimedia Tools and Applications, 2022, 81(21): 30685-30704.

[4] WANG K, LIU M Z. Object recognition at night scene based on DCGAN and faster R-CNN[J]. IEEE Access, 2020, 8: 193168-193182.

[5] ZHANG L, QI G J, WANG L, et al. AET vs. AED: unsupervised representation learning by auto-encoding transformations rather than data[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. New York: IEEE, 2019: 2547-2555.

[6] GE Z, LIU S, WANG F, et al. YOLOX: Exceeding YOLO series in 2021[EB/OL]. (2021-08-06)[2023-10-10]. http://arxiv.org/abs/2107.08430.

[7] CUI Z, QI G J, GU L, et al. Multitask aet with orthogonal tangent regularity for dark object detection[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. New York: IEEE, 2021: 2553-2562.

[8] KARAIMER H C, BROWN M S. A software platform for manipulating the camera imaging pipeline[C]//European Conference on Computer Vision. Berlin: Springer International Publishing, 2016: 873-888.

[9] BROOKS T, MILDENHALL B, XUE T, et al. Unprocessing images for learned raw denoising[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. New York: IEEE, 2019: 11036-11045.

[10] LIU S, HUANG D, WANG Y. Learning spatial fusion for single-shot object detection[EB/OL]. (2019-11-25)[2023-10-10]. http://arxiv.org/abs/1911.09516v2.

[11] ZHANG Y, SUN P, JIANG Y, et al. Bytetrack: multi-object tracking by associating every detection box[C]//European Conference on Computer Vision. Cham: Springer Nature Switzerland, 2022: 1-21.

[12] ZHOU K, YANG Y, CAVALLARO A, et al. Omni-scale feature learning for person re-identification[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. New York: IEEE, 2019: 3702-3712.

[13] HE S, LUO H, WANG P, et al. Transreid: transformer-based object re-identification[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. New York: IEEE, 2021: 15013-15022.

[14] DU Y, WAN J, ZHAO Y, et al. Giaotracker: a comprehensive framework for mcmot with global information and optimizing strategies in visdrone 2021[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. New York: IEEE, 2021: 2809-2819.

[15] BEWLEY A, GE Z, OTT L, et al. Simple online and realtime tracking[C]//2016 IEEE International Conference on Image Processing(ICIP). New York: IEEE, 2016: 3464-3468.

[16] MAGGIOLINO G, AHMAD A, CAO J, et al. Deep oc-sort: multi-pedestrian tracking by adaptive re-identification[C]//2023 IEEE International Conference on Image Processing (ICIP). New York: IEEE, 2023: 3025-3029.

[17] SHAO S, ZHAO Z, LI B, et al. Crowdhuman: a benchmark for detecting human in a crowd[EB/OL]. (2018-04-30)[2023-10-10]. http://arxiv.org/abs/1805.00123.

[18] DENDORFER P, OSEP A, MILAN A, et al. Motchallenge: a benchmark for single-camera multiple target tracking[J]. International Journal of Computer Vision, 2021, 129: 845-881.

[19] GUO C, LI C, GUO J, et al. Zero-reference deep curve estimation for low-light image enhancement[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. New York: IEEE, 2020: 1780-1789.

[20] ZHANG Y, ZHANG J, GUO X. Kindling the darkness: a practical low-light image enhancer[C]//Proceedings of the 27th ACM International Conference on Multimedia. [s. l.]: ACM Digital Library, 2019: 1632-1640.

[21] WU W, WENG J, ZHANG P, et al. Uretinex-net: Retinex-based deep unfolding network for low-light image enhancement[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. New York: IEEE, 2022: 5901-5910.

[22] LYU C, ZHANG W, HUANG H, et al. RTMDet: an empirical study of designing real-time object detectors[EB/OL]. (2022-12-16)[2023-10-10]. http://arxiv.org/abs/2212.07784v2.