激光与光电子学进展 先进成像

联合 Transformer 与 BYTE 数据关联的多目标实时跟踪算法

潘昊[1]，刘翔[1*]，赵静文[1]，张星[2,3]

[1]上海工程技术大学电子电气工程学院，上海 201620；

[2]上海工程技术大学管理学院，上海 201620；

[3]江苏大学汽车工程研究院，江苏 镇江 212013

摘要 针对复杂环境下多目标跟踪过程出现的轨迹漏检、误检及身份切换等问题，提出一种基于改进 YOLOX 和 BYTE 数据关联方法的多目标跟踪算法。首先，为了增强 YOLOX 在复杂环境下的目标检测能力，将 YOLOX 骨干网络与 Vision Transformer 结合，增强网络的局部特征提取能力，同时加入 α-GIoU 损失函数，进一步增加网络边界框的回归精度；其次，为了满足算法实时性要求，采用 BYTE 数据关联方法，摒弃传统特征重识别（Re-ID）网络，进一步提高了多目标跟踪算法的速度；最后，为了改善光照、遮挡等复杂环境下的跟踪问题，采用更加适应非线性系统的扩展卡尔曼滤波，提高了网络在复杂场景下对跟踪轨迹的预测精度。实验结果表明：所提算法对 MOT17 数据集的 multiple object tracking accuracy（MOTA）、identity F1-measure（IDF1）指标分别为 73.0%、70.2%，相较于目前最优的 ByteTrack，分别提升了 1.3 个百分点、2.1 个百分点，number of identity switches（IDSW）则减少了 3.7%；同时所提算法取得了 51.2 frame/s 的跟踪速度，满足系统实时性要求。

关键词 多目标跟踪；YOLOX；BYTE；Transformer；复杂场景

中图分类号 TP391 **文献标志码** A **DOI**：10.3788/LOP220514

Multitarget Real-Time Tracking Algorithm Based on Transformer and BYTE Data

Pan Hao[1], Liu Xiang[1*], Zhao Jingwen[1], Zhang Xing[2,3]

[1]*School of Electronic and Electrical Engineering, Shanghai University of Engineering Science, Shanghai* 201620, *China*;

[2]*School of Management, Shanghai University of Engineering Science, Shanghai* 201620, *China*;

[3]*Automotive Engineering Research Institute, Jiangsu University, Zhenjiang* 212013, *Jiangsu, China*

Abstract To solve the problems of trajectory missed detection, misdetection, and identity switching in complex multitarget tracking, this paper proposes a multitarget tracking algorithm based on improved YOLOX and BYTE data association methods. First, to enhance YOLOX's target detection capabilities in complex environments, we combine the YOLOX backbone network and Vision Transformer to improve the network's local feature extraction capability and add the α-GIoU loss function to further improve the regression accuracy of the network bounding box. Second, to meet the real-time requirements of the algorithm, we employ the BYTE data association method, abandon the traditional feature rerecognition (Re-ID) network, and further improving the speed of the proposed multitarget tracking algorithm. Finally, to mitigate the tracking problems in complex environments, such as illumination and occlusion, we adopt the extended Kalman filter, which is more adaptive to the nonlinear system, to improve the prediction accuracy of the network for tracking trajectory in complex scenes. The experimental results show that the multiple object tracking accuracy (MOTA) and identity F1-measure (IDF1) of the proposed algorithm on the MOT17 dataset are 73.0% and 70.2%, respectively, compared with the current optimal algorithm ByteTrack, they are improved by 1.3 percentage points and 2.1 percentage points, respectively, whereas number of identity switches (IDSW) is reduced by 3.7%. Meanwhile, the proposed algorithm achieves a tracking speed of 51.2 frames/s, which meets the real-time requirements of the system.

Key words multi-target tracking; YOLOX; BYTE; Transformer; complex scene

收稿日期：2022-01-11；**修回日期**：2022-01-19；**录用日期**：2022-01-24；**网络首发日期**：2022-02-08

基金项目：中国高校产学研创新基金（2021FNB02001）、文化部科技创新项目（2015KJCXXM19）

通信作者：*xliu@sues.edu.cn

1 引 言

多目标跟踪(MOT)是计算机视觉的一个重要研究领域,被广泛地应用在智能交通、智能监控、人机交互及虚拟现实等领域。其中行人跟踪是多目标跟踪的重点和难点,通过跟踪视频中的行人,可以得到行人的数量、位置、运动轨迹等多种状态信息,然后对这些信息进行处理分析,可以广泛地将分析结果应用在街道、商场、体育馆、火车站及地铁站等公共场所中。

基于深度学习的行人多目标跟踪算法是目前的主流,主要分为两种。一种是基于检测的跟踪(separate detection and embedding,SDE),即先进行目标检测,再进行跟踪。代表算法有 Bewley 等[1]提出的 Sort 算法,Sort 算法通过结合卡尔曼滤波与匈牙利算法,实现对连续帧检测目标的状态估计与匹配,但是没有进行 ID 重识别,导致当目标跟丢时,该轨迹就丢失了。随后 Wojke 等[2]在文献[1]的基础上提出 Deep Sort 算法,该算法在跟踪算法上加入了基于卷积神经网络(CNN)的行人重识别网络(Re-ID),增强了目标跟丢的鲁棒性,同时设计了级联匹配算法进一步提高跟踪效果,但是由于采用先检测再跟踪的两阶段流程,所以整体速度仍然不高。张相胜等[3]在 Deep Sort 算法的基础上提出基于改进 YOLOv3 的多目标跟踪算法,该算法有效减小了跟踪漏检率。另一种是将行人重识别步骤直接嵌入检测网络(jointly learns the detector and embedding model,JDE)的算法,该类算法改进了 SDE 网络速度慢的问题,同时进行目标检测与特征提取,大大缩短了整体跟踪算法的时间。代表算法有 Wang 等[4]提出的算法,该算法在单阶段检测网络的基础上在输出特征层增加一个特征提取分支,用来学习行人特征信息,实现了 MOT 的加速,但是该算法只是同时学习了检测目标和目标特征,之后还需要进行目标关联与轨迹预测,并不是真正意义的 JDE。Zhou 等[5]在 CenterNet 的基础上将上一帧的图片和热力图预测结果作为当前帧的输入,相比于文献[4]中的算法,其将检测与目标关联完全融合在一起,实现了真正的 JDE,但是由于只考虑了相邻两帧之间的关系,所以很难对目标形成长时间的关联和依赖,对于 ID 切换问题仍有待解决。Zhang 等[6]认为 anchor free 网络框架比 anchor based 框架更适合重识别行人特征,其在获得目标位置与行人特征后配合卡尔曼滤波与匈牙利算法进行匹配,最终输出结果,跟踪网络性能和速度达到了一定平衡。但是传统的 Re-ID 步骤消耗了大量算法时间,制约着跟踪系统性能,基于此,Zhang 等[7]提出一种新的数据关联方法,即 each detection box is a basic unit of the tracklet, as byte in computer program(BYTE),通过利用低分框的二次轨迹匹配,大大减少了 ID 切换次数,同时由于没有进行 Re-ID,所以速率也得到很大提高,整体跟踪系统的性能得到进一步提升。

综上分析,JDE 范式和 SDE 范式都无法均衡 MOT 系统的速度和精度。基于此,本文融合两者的优点,在检测阶段使用 YOLOX[8]作为检测网络,并在其基础上融合 CNN 与 Vision Transformer 结构,使网络在复杂场景下的特征学习能力得到进一步提高;此外为了解决模型在严重遮挡和小尺度场景下的边界框回归问题,引入 α 系列交并比(IoU),进一步提升网络对边界框的回归精度;然后在跟踪阶段采用 BYTE 数据关联方法,去掉 Re-ID 操作,使 MOT 系统速度得到大幅提升;最后采用非线性系统滤波器——扩展卡尔曼滤波算法,对轨迹进行预测,进一步提升了算法在复杂场景下的预测精度。实验结果表明,所提算法在行人多目标跟踪任务下取得了更好的结果。

2 所提算法原理

2.1 基于 Transformer 的 YOLOX 检测网络

所提检测网络结构如图 1 所示。

2.1.1 YOLOX 检测网络

YOLOX 将 YOLO 系列的检测器改进为无锚框的方式,并加入其他先进的技术,包括解耦头、Mosaic 和 Mixup 数据增强策略、有效的标签分配策略 SimOTA,实现更为精准的目标检测。

YOLOX 的骨干网络与 YOLOv5 相同,采用最先进的 CSPNet 结构和一个额外的 PAN 头。在输出特征层部分,采用 2 个解耦头,分别用于回归和分类。另外在回归头部添加 1 个额外的 IoU 感知分支,计算预测框和真实框之间的 IoU。其中回归头用来预测特征图中每个位置的坐标,即每个网格左上角的两个偏移量以及预测框的高度和宽度。在损失计算部分,回归头由广义交并比(GIoU)损失负责监督,分类头和 IoU 头由二元交叉熵损失负责监督。

尽管 YOLOX 在目标检测中已经取得和 SimOTA 相当的性能,但在遮挡、远景目标、光照变换等条件下进行目标检测时仍然存在漏检、误检等问题。注意力机制是解决上述问题的有效手段之一,通过权重分配的形式,网络更加专注于对目标特征的学习,能够更好处理复杂场景下的检测问题。因此为了进一步增强 YOLOX 对上述场景的检测性能,本文引入了 Vision Transformer(ViT)模型[9]。

2.1.2 YOLOX 检测网络改进

在 YOLOX 骨干网络的基础上,结合 ViT 模型结构设计了 ViT-CSP 模块,如图 2 所示,对比图 1 中的 CSP 结构,ViT-CSP 将原来的 Bottleneck 模块换成了 ViT 模块。ViT 模块由 Embedding 层、Transformer Encoder 及 MLP Head 三部分组成,该模块通过将图片分成多个 patch 将其输入到 Transformer Encoder,最后通过 MLP Head 对图片目标进行分类。相较于 SENet[10]、CBAM[11]等注意力模块,ViT 不需要借助 CNN 结构,直接使用原生 Transformer 结构,运算量

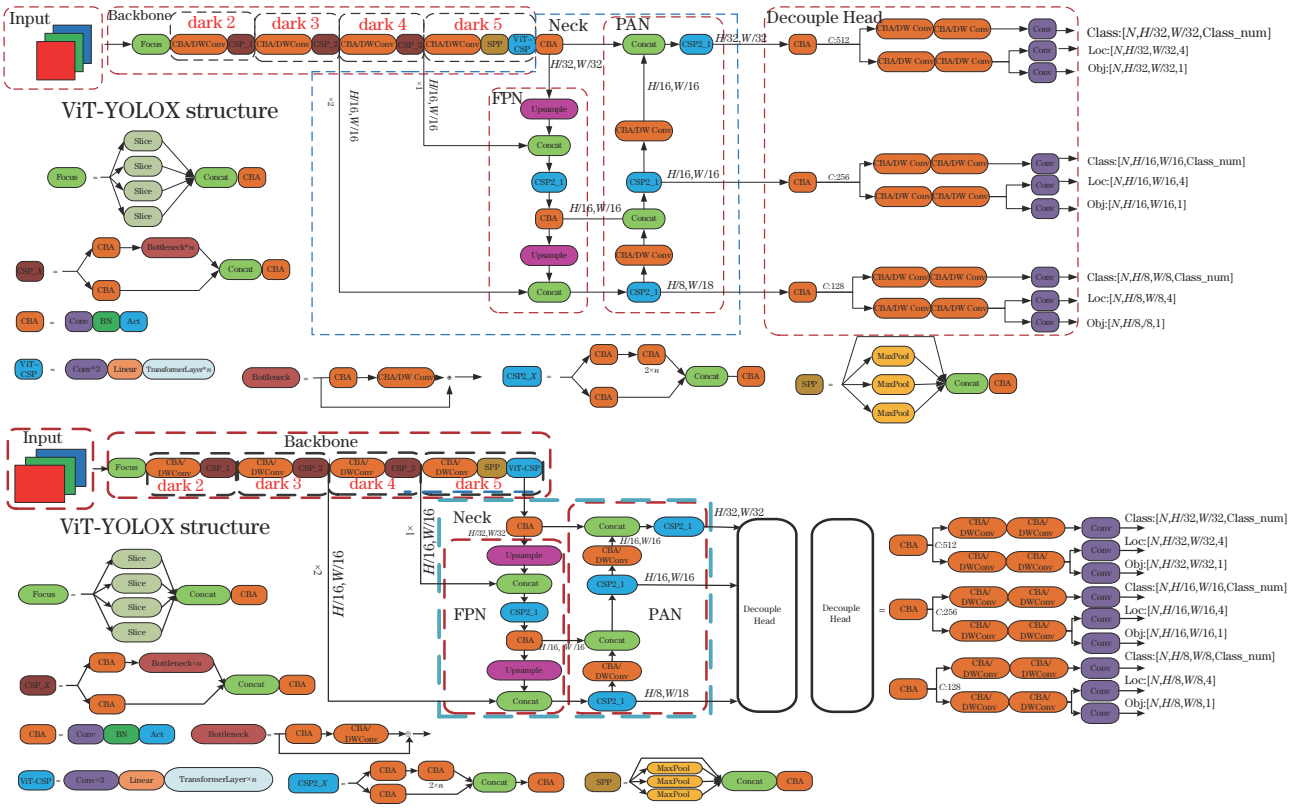

图 1　所提检测网络的结构

Fig. 1　Structure of the proposed detection network

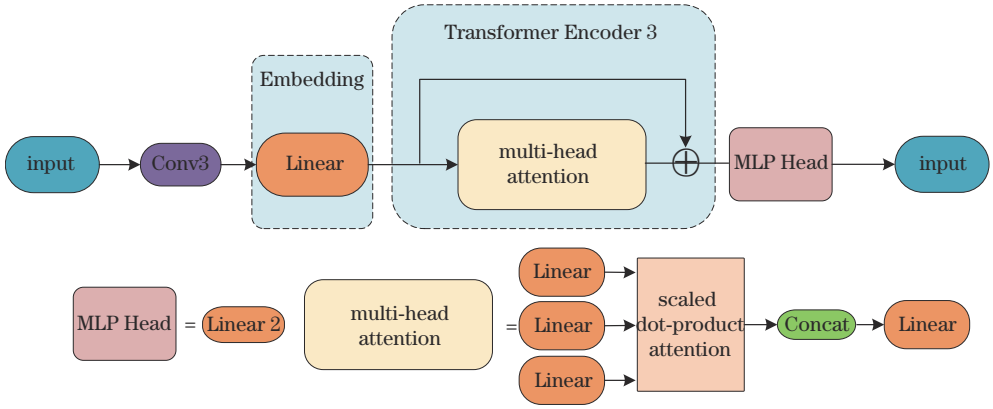

图 2　ViT-CSP 模型的结构

Fig. 2　Structure of ViT-CSP model

小，在应对不同尺度的特征时拥有更大范围的感受野，能更好地提取全局信息和局部特征，同时在大型数据集上能够取得更好的效果。

为了验证 ViT-CSP 模块效果，使用 CrowdHuman、ETHZ、Citypersons 和 MOT17 的混合数据集，由于 ViT 模块在尺度更大的特征层中需要占用更多的显存和访存量[9]，因此决定分别在 dark 2、dark 3、dark 4、dark 5 进行关于 YOLOX-s 模型的消融实验对比，结果如表 1 所示。加入 ViT 结构的 YOLOX-s 取得了更高的 AP 和更少的计算量，但是在分辨率更大的特征层（dark 2、dark 3 及 dark 4），ViT 结构由于频繁的内存吞吐，消耗了更多的检测时间，并不满足追求更快跟踪速度的系统的要求，因此最终在 dark 5 处加入 ViT 结构的网络模型。

其次，为了进一步提升复杂场景下网络预测框与真实框匹配的回归精度，引入了 α-GIoU[12]。以往的边

表 1　以 YOLOX-s 为基础的消融实验

Table 1　Ablation experiment based on YOLOX-s

dark 2	dark 3	dark 4	dark 5	AP	FLOPs /10^9	Time /ms
				0.832	18.18	5.2
√				0.843	18.39	31.09
	√			0.848	17.67	25.87
		√		0.846	17.67	13.04
			√	0.844	18.03	5.2

界框回归是通过计算预测框与真实框的交集与并集的比值得出的，但是当预测框与真实框不重叠时就会出现梯度消失等问题，这激发了一系列基于IoU的改进：GIoU、距离交并比（DIoU）和完全交并比（CIoU）。其中GIoU为IoU引入惩罚项，缓解了梯度消失问题；DIoU和CIoU则在惩罚项中考虑了预测框与真实值中心点的位置关系和宽高比；而α-GIoU通过加入超参数α，增加网络对不同尺度目标的损失和梯度，进而提高了bbox的回归精度。损失函数L_{GIoU}的公式为

$$L_{\mathrm{GIoU}}=1-R_{\mathrm{IoU}}+\frac{\left|\boldsymbol{C}/(\boldsymbol{B}\cup\boldsymbol{B}^{\mathrm{gt}})\right|}{|\boldsymbol{C}|}\Rightarrow L_{\alpha-\mathrm{GIoU}}=1-R_{\mathrm{IoU}}{}^{\alpha}+\left[\frac{\left|\boldsymbol{C}/(\boldsymbol{B}\cup\boldsymbol{B}^{\mathrm{gt}})\right|}{|\boldsymbol{C}|}\right]^{\alpha},\tag{1}$$

式中：$\boldsymbol{B}$代表预测框，$\boldsymbol{B}^{\mathrm{gt}}$代表真实框；$\boldsymbol{C}$为$\boldsymbol{B}$和$\boldsymbol{B}^{\mathrm{gt}}$的最小凸型；$\alpha$为超参数，设置为3。通过加入超参数$\alpha$，目标检测模型ViT-YOLOX的精度得到进一步提升。实验结果证明，ViT-YOLOX都明显提高了检测精度、多目标跟踪精度（MOTA）、ID切换等指标。

2.2　多目标跟踪算法改进

2.2.1　基于BYTE的多目标跟踪方法

以往数据关联算法[4,13]在进行轨迹匹配时只选择高分框进行匹配，而对于那些低于阈值的检测框则直接丢掉。显然这样是不合理的，对于密集场景的多目标跟踪，往往那些低分框代表处于严重遮挡、运动模糊或尺度发生变化的目标，如果直接过滤掉这些目标，会导致在多目标跟踪过程中出现不可逆转的误差，并带来大量的漏检和ID切换。因此与只保留高分框的思路不同，所提方法采用BYTE数据关联方法[7]保留每一个检测框，设置S_{low}、S_{high}和ε，并将所有得分框分为D_{low}和D_{high}两类，具体算法流程如图3所示。

在轨迹匹配阶段，首先计算高分框与轨迹Γ之间的IoU相似度矩阵，之后通过匈牙利算法进行关联匹配，而对于那些没有匹配到合适预测框的轨迹和检测框，分别记为Γ_{remain}、D_{remain}。对于D_{remain}，如果其大于设定阈值ε并且连续存在两帧，则将其认定为新的轨迹，并存入Γ；对于Γ_{remain}，对其与低得分框进行IoU相似度计算和匈牙利匹配，对于未匹配的轨迹将其记为$\Gamma_{\mathrm{re\text{-}remain}}$，未匹配的检测框则直接删除，其中$\Gamma_{\mathrm{re\text{-}remain}}$之后将被归在集合$\Gamma_{\mathrm{lost}}$。为了能够建立长期的轨迹关联，采用保留30帧的处理方法，当后续出现能够匹配的检测框时，则恢复轨迹，并加入到Γ，否则超过30帧就将其丢弃。

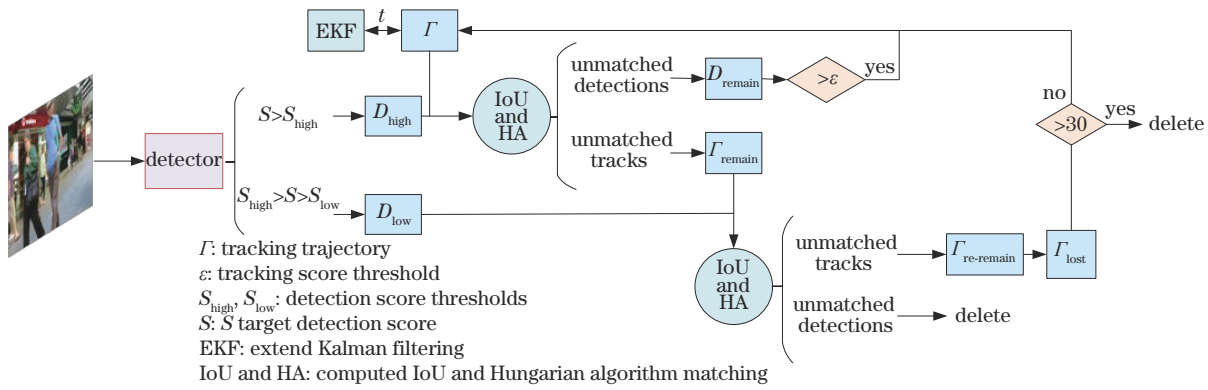

图3　BYTE算法的流程

Fig. 3　Flow chart of BYTE algorithm

2.2.2　拓展卡尔曼滤波

传统的卡尔曼滤波算法仅适用于符合高斯分布的线性系统，但在多目标跟踪的实际应用中存在大量非线性场景，如光照、运行模糊、复杂环境背景及遮挡等因素。因此，为了改善上述场景带来的问题，引入对非线性系统更加鲁棒的扩展卡尔曼滤波器。

拓展卡尔曼滤波算法通过状态方程和观测方程进行计算，公式分别为

$$\boldsymbol{\theta}_k=f(\boldsymbol{\theta}_{k-1})+\boldsymbol{s}_k,\tag{2}$$

$$\boldsymbol{z}_k=h(\boldsymbol{\theta}_k)+\boldsymbol{v}_k,\tag{3}$$

式中：$\boldsymbol{\theta}_k$是第k帧目标的系统状态向量，即第k帧目标的真实值；$f(\boldsymbol{\theta}_{k-1})$为状态转移矩阵；$\boldsymbol{s}_k$是协方差为$Q$的零均值高斯噪声；$\boldsymbol{z}_k$为第$k$帧目标的系统观测向量，即第$k$帧目标的检测值；$h(\boldsymbol{\theta}_k)$为观测矩阵；$\boldsymbol{v}_k$是协方差为$R$的零均值高斯噪声。

对于非线性系统中的状态估计问题，可以对非线性函数$f(\boldsymbol{\theta}_{k-1})$和$h(\boldsymbol{\theta}_k)$进行泰勒级数展开，取一次项为一阶拓展卡尔曼滤波，公式为

$$\boldsymbol{\theta}_k=f(\boldsymbol{\theta}_{k-1})+\boldsymbol{s}_k=f(\hat{\boldsymbol{\theta}}_{k-1})+F_{k-1}(\boldsymbol{\theta}_{k-1}-\hat{\boldsymbol{\theta}}_{k-1})+\boldsymbol{s}_k,\tag{4}$$

$$\boldsymbol{z}_k=h(\boldsymbol{\theta}_k)+\boldsymbol{v}_k=h(\boldsymbol{\theta}'_k)+H_k(\boldsymbol{\theta}_k-\boldsymbol{\theta}'_k)+\boldsymbol{v}_k,\tag{5}$$

式中：$\hat{\boldsymbol{\theta}}_{k-1}$为第$k-1$帧目标的估计值；$\boldsymbol{\theta}'_k$为第$k$帧目标的预测值；$F_{k-1}$和$H_k$分别表示函数$f(\boldsymbol{\theta})$和$h(\boldsymbol{\theta})$在$\hat{\boldsymbol{\theta}}_{k-1}$和$\boldsymbol{\theta}'_k$处展开的Jacobi矩阵。

3　实验结果与分析

实验使用的硬件配置为Inter(R)Core i7-9700K

CPU，NVIDIA GeForce RTX 2080Ti 显卡，软件环境为 Ubuntu20.4 系统，CUDA11.3，PyTorch1.9。

对于训练过程，使用 COCO 数据集的预训练模型权重作为网络模型的初始权重，然后使用 CrowdHuman[14]、ETHZ[15]、Cityperson[16]、MOT17[17]数据集进行混合训练，迭代次数设置为 80，其中第一个 epoch 过程采用余弦退火方法进行预热训练，采用 Mosaic 和 Mixup 数据增广的策略，图片输入尺寸为 412×672，批处理大小设置为 6。使用 SGD 优化器，权重衰减设置为 5×10^{-4}，动量大小为 0.9，初始学习率为 0.001。在测试阶段，将 ByteTrack 作为所提算法的 Baseline，在 MOT17 数据集上进行关于实验结果的指标对比和性能分析，同时与目前主流的跟踪网络进行了对比，最后对 MOT17、MOT20[18]、ETHZ 数据集进行可视化结果分析。

3.1 测评指标

为了使测评结果更加准确客观，并方便与其他算法进行比较，采用多目标跟踪领域通用的测评体系，如表 2 所示，其中上箭头表示数值越高性能越好，下箭头表示数值越低性能越好。

表 2 测评指标及解释说明

Table 2 Evaluation indicators and instructions

Evaluation indicator	Indicator instruction
FP↓	Number of being mistaken as a positive sample
FN↓	Number of being misidentified as a negative sample
IDSW↓	Number of target ID switches
MOTA↑	Tracking accuracy. Calculated by integrating FP, FN, IDS, and other indicators
FPS↑	Tracking speed. Number of frames processed per second, which is used to measure the real-time performance of the model
IDF1↑	Ratio of correctly identified detections to the average number of true and calculated detections
MT↑	Ratio of number of hit trajectories to the total number of real trajectories. No more than 0.8
ML↓	Number of lost tracks as a percentage of the total number of real tracks

3.2 实验结果分析

3.2.1 基于 ViT-YOLOX 与 BYTE 数据关联方法的实验分析

首先在 CrowdHuman、ETHZ、Cityperson 及 MOT17 的混合数据集上对 ViT-YOLOX 模型与 YOLOX 模型分别进行了检测性能对比与跟踪性能对比。为了加强模型在不同数据集间的鲁棒性和增加数据样本背景的多样性，训练过程中，在前 70 个 epoch 使用 Mosaic 数据增强，后 10 个 epoch 不使用数据增强，其中检测性能对比如图 4 所示。虚线和实线分别为 YOLOX 与 ViT-YOLOX 在 IoU 阈值为 0.5 和 0.75 时对应的 AP 值，实验证明 ViT-YOLOX 在各个 epoch 验证集上获得的 AP 值均要高于 YOLOX，并且随着模型尺度的增加，ViT-YOLOX 模型的 AP 提升越明显。

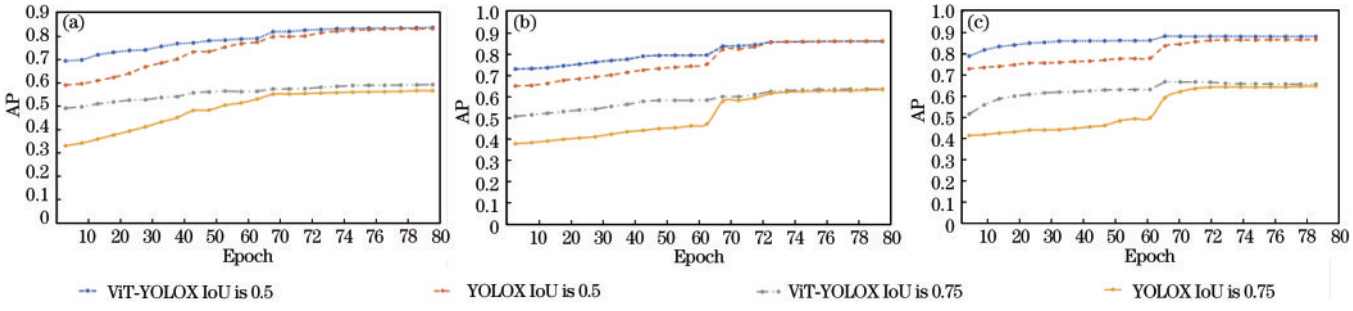

图 4 ViT-YOLOX 与 YOLOX 的 AP 对比。(a) s 模型对比；(b) m 模型对比；(c) l 模型对比

Fig. 4 AP comparison of VIT-YOLOX and YOLOX. (a) Comparison of s model; (b) comparison of m model; (c) comparison of l model

之后在 ViT-YOLOX 基础上采用 BYTE 数据关联方法与 ByteTrack 在 MOT17 训练集上进行跟踪性能对比，如表 3 所示，其中相比 ByteTrack 模型，基于 ViT-ByteTrack 的 s、m 和 l 模型的 MOTA 指标分别提升 1.5 个百分点、0.1 个百分点和 4.1 个百分点；IDF1 指标则分别提升 0.3 个百分点、0.3 个百分点和 3.7 个百分点；而在应对 ID 切换情况时，除 s 模型略微升高外，m 和 l 模型均有小幅度改善；轨迹误检指标 FP 和轨迹漏检指标 FN 也均有不同程度的下降。

3.2.2 α 系列 IoU 和扩展卡尔曼滤波消融实验分析

经过上述实验对比，将 ViT-ByteTrack-l 模型作为所提算法模型，对加入 α 系列 IoU 与扩展卡尔曼滤波的模型分别进行训练，在 MOT17 训练集中分别与各自原始方法进行了对比，如表 4 所示，EKF 表示扩展卡尔曼滤波，最优结果以加粗字体标出。

对于只加入 α 超参数的模型，相较于原始 IoU，α-GIoU 取得了更好的效果，其 MOTA 和 IDF1 分别提升了 4.0 个百分点和 1.8 个百分点，轨迹切换次数 IDSW

表 3 跟踪性能对比

Table 3 Tracking performance comparison

Model	MOTA /%	IDSW	IDF1 /%	FP	FN
ByteTrack-s	65.3	417	65.0	6514	33333
ViT-ByteTrack-s	66.8	423	65.3	4856	32030
ByteTrack-m	70.5	512	68.3	5539	27183
ViT-ByteTrack-m	70.6	505	68.6	4247	28409
ByteTrack-l	71.7	543	68.1	5497	26785
ViT-ByteTrack-l	75.8	491	71.8	4753	21969

减少了 141；当在 α 系列 IoU 基础上加上扩展卡尔曼滤波之后，FP 和 FN 指标提升更加明显，其中取得综合效果最好的是 EKF＋α-GIoU，相对于原始模型，分别取得了 4.4 个百分点、27.7%、1.5 个百分点、26.0% 和 13.5% 的增益。可以看到，在引入 α 系列 IoU 和扩展卡尔曼滤波后，模型在轨迹命中和轨迹得分中均获得有效的提升，证明了所提算法的优越性。

3.2.3 实验结果可视化分析与对比实验

为了更进一步说明和对比所提算法与 ByteTrack 的跟踪性能，所提算法与 ByteTrack 分别在 MOT17、MOT20 及 ETHZ 数据集上进行了部分可视化结果对比，如图 5～7 所示。

表 4 α 系列 IoU 和扩展卡尔曼滤波消融实验对比

Table 4 Ablation experiment comparison of α series IoU and extended Kalman filtering

Method	MOTA /%	IDSW	IDF1 /%	FP	FN
IoU(base)	75.8	491	71.8	4753	21969
α-IoU	74.2	402	70.9	4508	24050
EKF＋IoU	76.3	410	71.3	4433	21110
EKF＋α-IoU	74.6	399	70.5	4287	24138
GIoU	75.8	498	71.9	4817	20417
α-GIoU	79.8	**350**	**73.6**	3728	19397
EKF＋GIoU	76.4	460	71.8	4469	20004
EKF＋α-GIoU	**80.2**	355	73.3	**3519**	**18993**

在图 5 中，对于箭头所指 ID 为 18 的目标，ByteTrack 先后因为严重遮挡而出现漏检和 ID 切换，但是所提算法仍旧能够持续稳定地跟踪到目标；图 6 中，由于场景光照复杂、人群密集，可以看到 ByteTrack 出现了大量漏检，漏检前 40 帧中箭头所指的灰色上衣目标，对于 99 帧中箭头所指 219 号目标，发生 ID 切换，相反所提算法均能实现稳定跟踪；图 7 中，对于远景小目标，如箭头所指的两名儿童目标，ByteTrack 出现漏检现象，但所提算法进行尺度变换，始终能够实现稳定跟踪。

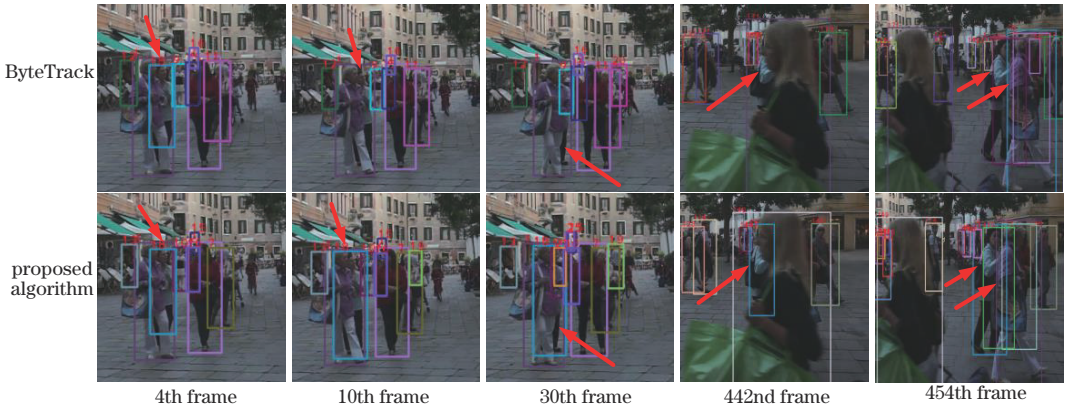

图 5 在 MOT17 数据集的跟踪结果对比

Fig. 5 Tracking result comparison on MOT17 dataset

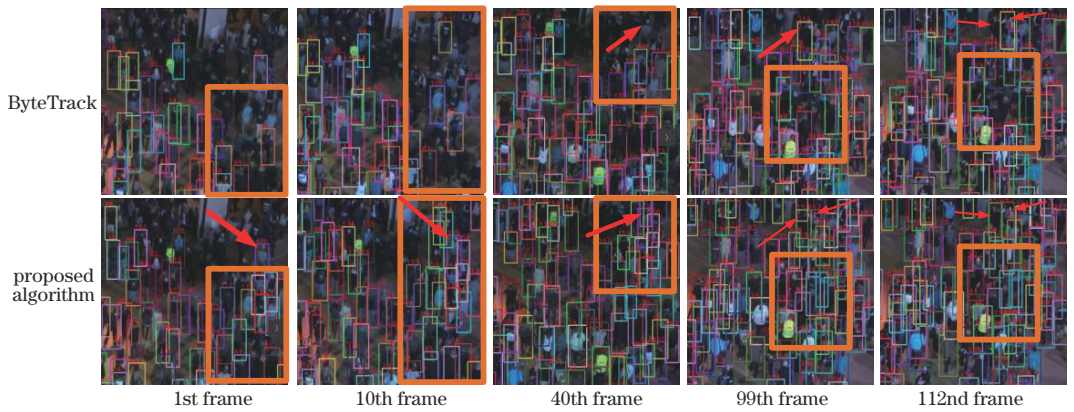

图 6 在 MOT20 数据集的跟踪结果对比

Fig. 6 Tracking result comparison on MOT20 dataset

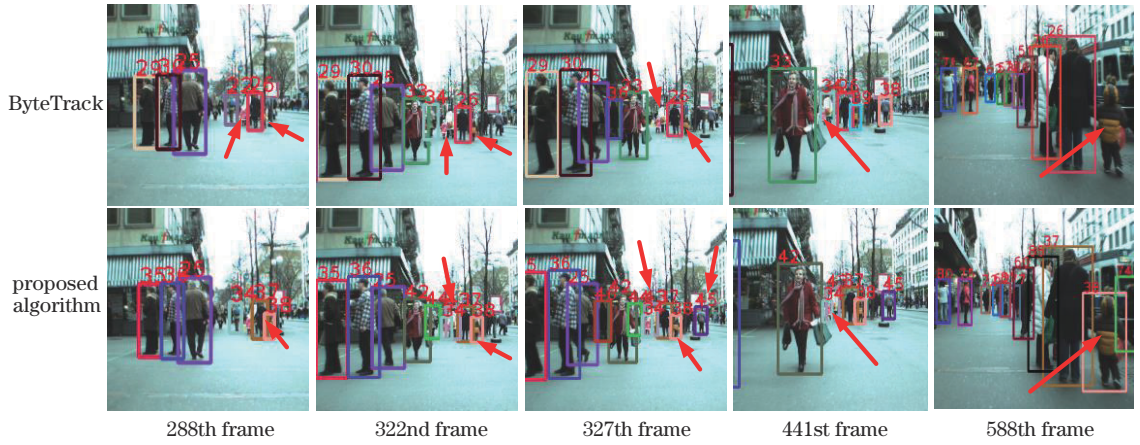

图 7　在 ETHZ 数据集的跟踪结果对比

Fig. 7　Tracking result comparison on ETHZ dataset

为了综合评定所提算法的跟踪性能，所提算法与目前主流多目标跟踪算法 Tube_TK[19]、TransTrack[20]、CenterTrack[5]、FairMOT[6]及 ByteTrack-l 在 MOT17 数据集中以相同图片输入尺度进行测试，并在 MOT Challenge 官网上传实验获得指标结果，实验结果对比如表 5 所示，最优性能以粗体标出。可以看出：所提算法相较于 l 尺度的 ByteTrack，除了跟踪速度上略有下降外，其余指标均取得更优的结果，IDSW 和 FP 指标均取得了最优结果，证明所提算法在复杂环境和不同程度遮挡下拥有更稳定的跟踪性能。同时相较于其他主流算法，所提算法虽然没有取得最优的跟踪指标，但是在跟踪速度上所提算法要更快，更满足实时性要求。

表 5　MOT17 数据集上的跟踪性能对比

Table 5　Tracking performance comparison on MOT17 dataset

Model	MOTA /%	IDSW	IDF1 /%	FP	FN	FPS
Tube_TK	63.0%	4137	58.6	27060	177483	3.0
TransTrack	**75.0**	3603	63.5	50157	**86442**	10.0
CenterTrack	67.8	3039	64.7	18498	160332	17.5
FairMOT	73.7	3303	**72.3**	27507	117477	25.9
ByteTrack-l	71.7	2688	68.1	15838	134992	**55.3**
Proposed algorithm	73.0	**2589**	70.2	**15822**	134038	51.2

4　结　　论

提出了一种基于改进 YOLOX 与 BYTE 数据关联方法的多目标跟踪算法。首先在检测阶段，所提算法将 YOLOX 骨干网络与 ViT 模型结合，提高网络的特征提取能力，其次加入 α-GIoU 损失函数，提高网络在遮挡环境的目标检测精度；然后在跟踪阶段，摒弃传统的 Re-ID 步骤，采用 BYTE 数据关联方法将检测框分为高低两类，对未匹配轨迹进行二次匹配，有效利用检测框。相较于耗时的 Re-ID 算法，所提算法取得了更准更快的跟踪结果。同时为了应对复杂环境下人群跟踪的场景，采用对非线性系统更鲁棒的扩展卡尔曼滤波进行目标框预测，进一步提高系统在复杂环境下的跟踪性能。

相较于 FairMOT、TransTrack 等算法，所提算法在跟踪精度上仍有提升空间，而更精准的检测网络是提高跟踪精度的有效手段之一。下一步将继续探讨如何结合注意力机制模块，在更高分辨率特征层中提取有效特征的内容，在不损失过多系统速度的同时，提高 MOT 系统精度。

参　考　文　献

[1] Bewley A, Ge Z Y, Ott L, et al. Simple online and realtime tracking[C]//2016 IEEE International Conference on Image Processing, September 25-28, 2016, Phoenix, AZ, USA. New York: IEEE Press, 2016: 3464-3468.

[2] Wojke N, Bewley A, Paulus D. Simple online and realtime tracking with a deep association metric[C]//2017 IEEE International Conference on Image Processing, September 17-20, 2017, Beijing, China. New York: IEEE Press, 2017: 3645-3649.

[3] 张相胜，沈庆. 基于改进 YOLOv3 的多目标跟踪算法研究[J]. 激光与光电子学进展，2021，58(16)：1610004.
Zhang X S, Shen Q. Multitarget tracking algorithm based on an improved YOLOv3 algorithm[J]. Laser & Optoelectronics Progress, 2021, 58(16): 1610004.

[4] Wang Z D, Zheng L, Liu Y X, et al. Towards real-time multi-object tracking[M]//Vedaldi A, Bischof H, Brox T, et al. Computer vision-ECCV 2020. Lecture notes in

computer science. Cham: Springer, 2020, 12356: 107-122.
[5] Zhou X Y, Koltun V, Krähenbühl P. Tracking objects as points[M]//Vedaldi A, Bischof H, Brox T, et al. Computer Vision-ECCV 2020. Lecture notes in computer science. Cham: Springer, 2020, 12349: 474-490.
[6] Zhang Y F, Wang C Y, Wang X G, et al. FairMOT: on the fairness of detection and re-identification in multiple object tracking[EB/OL]. (2020-04-04) [2021-02-05]. https://arxiv.org/abs/2004.01888.
[7] Zhang Y F, Sun P Z, Jiang Y, et al. ByteTrack: multi-object tracking by associating every detection box[EB/OL]. (2021-10-13) [2021-12-05]. https://arxiv. org/abs/2110.06864.
[8] Ge Z, Liu S T, Wang F, et al. YOLOX: exceeding YOLO series in 2021[EB/OL]. (2021-07-18) [2021-12-02]. https://arxiv.org/abs/2107.08430.
[9] Dosovitskiy A, Beyer L, Kolesnikov A, et al. An image is worth 16x16 words: transformers for image recognition at scale[EB/OL]. (2020-10-22) [2021-05-07]. https://arxiv.org/abs/2010.11929.
[10] Hu J, Shen L, Albanie S, et al. Squeeze-and-excitation networks[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2020, 42(8): 2011-2023.
[11] Woo S, Park J, Lee J Y, et al. CBAM: convolutional block attention module[M]//Ferrari V, Hebert M, Sminchisescu C, et al. Computer vision-ECCV 2018. Lecture notes in computer science. Cham: Springer, 2018, 11211: 3-19.
[12] He J B, Erfani S, Ma X J, et al. Alpha-IoU: a family of power intersection over union losses for bounding box regression[EB/OL]. (2021-10-26) [2021-12-02]. https://arxiv.org/abs/2110.13675.
[13] Zhu J, Yang H, Liu N, et al. Online multi-object tracking with dual matching attention networks[M]// Ferrari V, Hebert M, Sminchisescu C, et al. Computer vision-ECCV 2018. Lecture notes in computer science. Cham: Springer, 2018, 11209: 379-396.
[14] Shao S, Zhao Z J, Li B X, et al. CrowdHuman: a benchmark for detecting human in a crowd[EB/OL]. (2018-04-30)[2021-05-02]. https://arxiv.org/abs/1805.00123.
[15] Ess A, Leibe B, Schindler K, et al. A mobile vision system for robust multi-person tracking[C]//2008 IEEE Conference on Computer Vision and Pattern Recognition, June 23-28, 2008, Anchorage, AK, USA. New York: IEEE Press, 2008.
[16] Zhang S S, Benenson R, Schiele B. CityPersons: a diverse dataset for pedestrian detection[C]//2017 IEEE Conference on Computer Vision and Pattern Recognition, July 21-26, 2017, Honolulu, HI, USA. New York: IEEE Press, 2017: 4457-4465.
[17] Milan A, Leal-Taixe L, Reid I, et al. MOT16: a benchmark for multi-object tracking[EB/OL]. (2016-03-02)[2021-05-06]. https://arxiv.org/abs/1603.00831.
[18] Dendorfer P, Rezatofighi H, Milan A, et al. MOT20: a benchmark for multi object tracking in crowded scenes [EB/OL]. (2020-03-19) [2021-05-04]. https://arxiv. org/abs/2003.09003.
[19] Pang B, Li Y Z, Zhang Y F, et al. TubeTK: adopting tubes to track multi-object in a one-step training model [C]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 13-19, 2020, Seattle, WA, USA. New York: IEEE Press, 2020: 6307-6317.
[20] Sun P Z, Cao J K, Jiang Y, et al. TransTrack: multiple object tracking with transformer[EB/OL]. (2020-12-31) [2021-05-04]. https://arxiv.org/abs/2012.15460.