E-mail: jig@aircas.ac.cn
Website: www.cjig.cn
Tel: 010-58887035
中国图象图形学报
JOURNAL OF IMAGE AND GRAPHICS

中图法分类号:TP18; TP391.41　　**文献标识码**: A　　**文章编号**: 1006-8961(2026)02-0512-13

论文引用格式:Yang Y X, Deng Y Q, Gu H J, Dong Z K and Zhang J. 2026. SiamMo: Siamese motion-centric 3D object tracking. Journal of Image and Graphics, 31(2):0512-0524(杨宇翔, 邓颖琦, 顾鸿杰, 董哲康, 张敬. 2026. 以运动为中心的孪生网络三维目标跟踪. 中国图象图形学报, 31(2):0512-0524)[DOI:10.11834/jig.250112]

以运动为中心的孪生网络三维目标跟踪

杨宇翔[1],邓颖琦[1],顾鸿杰[1],董哲康[1],张敬[2*]

1. 杭州电子科技大学电子信息学院,杭州　310018; 2. 武汉大学计算机学院,武汉　430072

摘　要:**目的**　现有的三维单目标跟踪方法主要遵循基于孪生匹配的范式,长期以来,该范式在处理无纹理且不完整的激光雷达点云时存在诸多问题。一种新的以运动为中心的范式无需进行外观匹配,在很大程度上克服了这些挑战。然而,该方法为多阶段操作,需额外进行预分割与边界框优化。为了解决上述问题,提出一种创新且简洁的基于孪生网络以运动为中心的跟踪方法。**方法**　本文方法的核心在于,先借助共享编码器从连续两帧数据中提取特征,随后直接在特征层面建模目标相对运动。具体而言,设计了时空特征聚合模块,该模块能在多尺度下整合编码特征,高效捕捉运动信息;同时引入边界框感知特征编码模块,将精确的尺寸先验信息融入运动特征,以提升预测精度。**结果**　实验环节,在Kitti、NuScenes和WOD(Waymo open dataset)3个数据集上,与当前先进方法对比,结果显示:在Kitti数据集中,相较于排名第2的方法,平均成功率指标提升4.7%,平均精度指标提升4.9%;在NuScenes数据集中,平均成功率指标提升14.2%,平均精度指标提升11.5%;在WOD数据集中,平均成功率指标提升2.9%,平均精度指标提升5.4%。在Kitti和NuScenes数据集的高难度测试子集上,本文方法也展现出极强的鲁棒性。此外,在Kitti和NuScenes数据集上开展的消融实验,进一步验证了孪生架构、时空特征聚合模块及边界框感知特征编码模块的有效性。**结论**　本文提出的基于孪生网络以运动为中心的跟踪方法,规避了易受干扰的外观匹配流程,且无需额外预分割和边界框优化,显著提高跟踪精度,为三维单目标跟踪领域提供了新思路。

关键词:三维目标跟踪;激光点云;孪生网络;运动估计;深度学习

SiamMo: Siamese motion-centric 3D object tracking

Yang Yuxiang[1], Deng Yingqi[1], Gu Hongjie[1], Dong Zhekang[1], Zhang Jing[2*]

1. School of Electronics and Information, Hangzhou Dianzi University, Hangzhou　310018, China;

2. School of Computer Science, Wuhan University, Wuhan　430072, China

Abstract: **Objective** 3D single-object tracking (3D SOT) is of paramount importance in a wide array of applications, including autonomous driving, robotics, and intelligent security. The fundamental goal of 3D SOT is to localize a specific target across a sequence of point clouds, with the only given information being its initial status. Existing matching-based 3D SOT methods generally utilize certain forms of Siamese networks for feature extraction. After transforming the cropped target template and search area embeddings to the same feature space with a shared encoder, these methods enhance target-specific features with various appearance matching techniques, such as cosine similarity and cross attention. Although the Siamese matching-based paradigm has become a popular design in existing models, appearance matching has long suffered from issues with textureless and incomplete LiDAR point clouds. Beyond this paradigm, a new motion-centric tracker, M^2-

收稿日期:2025-03-26;**修回日期**:2025-09-08;**预印本日期**:2025-09-15

*通信作者:张敬　jingzhang.cv@whu.edu.cn

基金项目:国家自然科学基金项目(62376080,62506107)

Supported by: National Natural Science Foundation of China(62376080,62506107)

Track, offers a new perspective for 3D SOT. It takes point clouds from two successive frames without cropping as input, explicitly modeling the relative target motion in a single-stream architecture, largely overcoming the challenges. However, it fuses adjacent point clouds and processes them in a single-stream architecture, lacking explicit target information from adjacent frames for accurate localization. To compensate for this deficiency, M^2-Track requires additional segmentation and box refinement, which makes the training objective complex and results in cumulative errors. To this end, this study proposes a novel Siamese motion-centric tracking approach, dubbed SiamMo. **Method** SiamMo adopts a simple single-stage tracking pipeline of Siamese feature extraction and motion modeling. To learn excellent features for point clouds of varying density, we first divide nonuniform points into regular voxels, which exhibit reduced sensitivity to point count variations, thus mitigating the varying sparsity issue to some extent. Afterward, we present a top-down convolutional network based on Siamese architecture to encode voxelized point clouds of successive frames into the same feature space. The network first uses sparse convolution to extract features in 3D space and then adopts dense convolution to extract features in 2D space in a bird's eye view. In contrast with the single-steam architecture of M^2-Track, Siamese architecture decouples feature extraction from temporal fusion, which enables it to extract more abundant and representative latent features while reducing information interference among successive frames. Subsequently, we design a spatiotemporal feature aggregation (STFA) module that integrates the encoded features at multiple scales for motion modeling. Intuitively, effective motion modeling necessitates rich representations at multiple scales. Integrating these multifaceted representations is expected to significantly enhance the network's capability to accurately localize targets with various motion patterns. Moreover, we introduce a box-aware feature encoding (BFE) module that injects explicit box priors into motion features for prediction. It first encodes the bounding box size parameters of the object in the initial frame to the box-aware encoding with a multilayer perception (MLP). Then, BFE adds the box-aware encoding and the output feature from the STFA module. Finally, the added feature is fed into an MLP to regress the relative target motion. Despite being conceptually simple, our BFE can boost tracking performance, with negligible computation. In a nutshell, using neither segmentation nor box refinement, SiamMo achieves precise localization by directly inferring the relative target motion in a single-stage manner. **Result** We compare our model with several state-of-the-art tracking methods, including Siamese matching-based and motion-centric trackers on three public datasets, namely, Kitti, NuScenes, and WOD. The quantitative evaluation metrics comprise Precision and Success. Experimental results show that our model outperforms all other methods on Kitti, NuScenes, and WOD datasets. On the Kitti dataset, compared with the second-ranked method, our method increases the average success indicator by 4.7% and the average precision indicator by 4.9%. On the NuScenes dataset, the average success indicator is increased by 14.2%, and the average precision indicator is increased by 11.5%. On the WOD dataset, the average success indicator is increased by 2.9%, and the average precision indicator is increased by 5.4%. Our method also demonstrates strong robustness to sparsity and distractors on the difficult test subsets of the Kitti and NuScenes datasets. In addition, we report the efficiency of SiamMo, which is lightweight with only 0.82 GFLOPs and 14.6 M parameters. We record the average running time of all test frames for the Car category on the Kitti dataset to evaluate the computational efficiency of our method, which achieves 108 frame/second, including 4.2 ms for pre/processing point clouds and 5.0 ms for network forward propagation on a single NVIDIA 4090 GPU. Ablation experiments conducted on the Kitti and NuScenes datasets further verify the effectiveness of the Siamese architecture, STFA module, and BFE module. In Kitti's Car category tracking task, replacing the single-stream architecture with the Siamese architecture results in a 3.8% increase in the average success and precision. When STFA leverages features across all scales, Kitti's Car success and precision are improved by 2.2% and 2.6%, respectively, as opposed to relying solely on single-scale features. When the BFE module is added, Kitti's Car success and precision are improved by 2.9% and 3.6%, respectively. **Conclusion** In this paper, we propose a novel and simple Siamese motion-centric tracking approach, which avoids the vulnerable appearance matching process and does not require additional presegmentation and box refinement by adopting Siamese architecture and models target motion in a simple single-stage pipeline. Comprehensive experiments demonstrate that our model surpasses state-of-the-art methods on three challenging benchmarks while demonstrating excellent robustness and maintaining a high inference speed.

Key words: 3D single object tracking; LiDAR point clouds; Siamese network; motion estimation; deep learning

0 引 言

三维单目标跟踪(3D single-object tracking, 3D SOT)是计算机视觉领域的一项基本任务。其目标是在已知特定目标的初始状态后,在一段点云序列中定位该目标。现有的基于匹配的单目标跟踪方法(Qi等,2020;Zheng等,2021;Zhou等,2022;Huang等,2020;李玺等2019;王蒙蒙等2022;Qi等,2017b)通常利用某种形式的孪生网络进行特征提取,如图1(a)所示。作为点云三维单目标跟踪领域的先驱,SC3D(shape completion for 3D object tracking)(Giancola等,2019)是第一个三维孪生跟踪器,它将目标模板与当前帧生成的大量候选区域进行详尽比较。然而,这导致模型需要进行多次前向传递,从而产生显著的计算开销。为了解决这个问题,P2B(point-to-box network)(Qi等,2020)引入了一种端到端的孪生区域提议网络。它首先使用孪生主干网络对模板和搜索区域的点云进行编码,随后通过外观匹配增强目标特定的特征,最后,P2B整合VoteNet(Qi等,2019)生成3D种子点建议,并选择得分最高的种子点的偏移作为目标。该方法在保持实时速度的同时显著提升了跟踪性能,许多后续工作采用了相同的孪生范式。BAT(box-aware tracker)(Zheng等,2021)设计了一个边界框感知特征融合模块来增强目标特定的特征。受Transformers(Vaswani等,2017)成功的启发,PTT(point-track-Transformer)(Shan等,2021)在P2B的投票阶段引入注意力机制,增强了跟踪器对深层次的目标线索的感知能力;PTTR(point tracking Transformer)(Zhou等,2022)基于自注意力机制和交叉注意力机制设计了特征融合网络,分别用于模板和搜索区域长距离特征的增强,以及目标线索的传递;STNet(3D Siamese Transformer network)(Hui等,2022)设计了一个基于交叉注意力的特征融合网络,迭代式地从粗粒度到细粒度地学习模板与搜索区域间的稳健互相关关系,将模板和搜索区域中的潜在对象进行联系,有效提升了跟踪器的鲁棒性和精度;CMT(context matching-guided Transformer)(Guo等,2022)引入了一种用于上下文描述的水平旋转不变描述符,并设计了一种移位窗口匹配策略来有效地衡量点之间的空间上下文相似性,接着通过Transformer整合点上下文线索并将目标感知信息引入搜索区域的特征中;SyncTrack(Ma等,2023)提出单分支Transformer框架,将特征提取与匹配同步进行,显著简化了网络结构,降低计算复杂度,提升运行效率。

尽管基于孪生匹配的范式已成为现有模型中的主流设计,但外观匹配长期以来受限于激光雷达点云无纹理和不完整等特性的问题。除了这种范式,一种新的以运动为中心的跟踪器(motion-centric tracker, M^2-Track)(Zheng等,2022)将连续两帧点云作为输入,通过显式地建模目标运动以实现跟踪,在很大程度上克服了上述问题,如图1(b)所示。然而,它直接合并了连续帧点云并在单流架构中进行处理,缺乏精细的目标信息以进行精确的定位。为了弥补这一缺陷,M^2-Track需要额外的分割和边界框优化,这使得训练目标变得复杂,并导致累积误差。

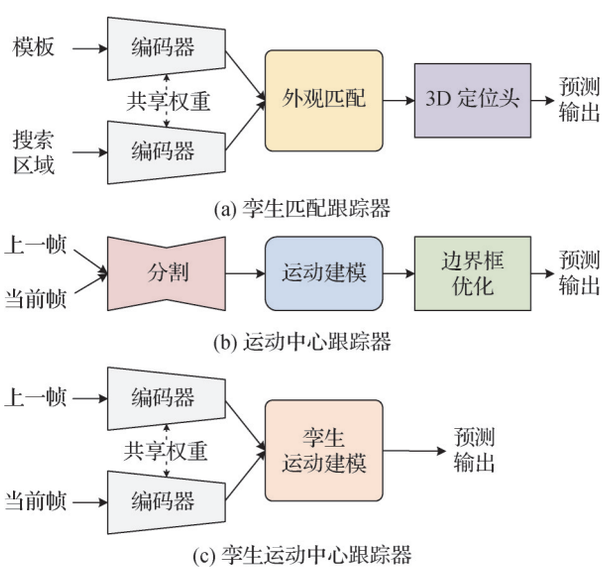

图1 与典型三维目标跟踪方法的对比

Fig. 1 Comparison of typical 3D object tracking methods ((a) Siamese matching-based tracker; (b) motion-centric tracker; (c) Siamese motion-centric tracker)

本文提出一种新颖的基于孪生网络以运动为中心的跟踪方法(Siamese motion-centric tracker, SiamMo),如图1(c)所示。它是一个单阶段的跟踪框架,包括孪生特征提取和运动建模。为了更好地学习不同密度点云的特征,首先将非均匀的点划分为规则的体素,以缓解原始点云的疏密不均匀。然后采用两个权重共享的特征编码器,将连续帧的体素网格嵌入到相同的特征空间中。

随后,本文设计了一个时空特征聚合模块

(spatio-temporal feature aggregation, STFA)，该模块在多个尺度上整合编码特征，从而有效地捕捉运动信息。此外，引入了一种边界框感知特征编码模块(box-aware feature encoding, BFE)，将明确的目标尺寸信息注入到运动特征中以得到更精确的预测。尽管在方法上较简单，但BFE可以在计算量忽略不计的情况下提升跟踪性能。简言之，本文提出的基于孪生网络的以运动为中心的跟踪方法消除了额外的分割和边界框优化操作，通过单阶段的运动估计就实现了精确的目标跟踪。实验表明本文方法的有效性，在3个具有挑战性的跟踪基准测试，即Kitti(Karlsruhe Institute of Teehnology and Toyota Technological Institute at Chicago)(Geiger等，2012)、NuScenes(nuTonomy scenes)(Caesar等，2020)和WOD(Waymo open dataset)(Sun等，2020)上取得了最先进性能。例如，SiamMo将目前Kitti数据集上跟踪的最佳结果提升到了一个新的记录，平均精确率(precision)指标达到了90.2%，同时保持108帧/s的高推理速度。在更具有挑战性的NuScenes数据集上，本文方法在平均成功率(success)指标上比之前的以运动为中心的跟踪器M^2-Track提高14.2%。此外，SiamMo在点云稀疏和干扰物存在的测试子集中表现出卓越的鲁棒性。总体而言，本文的贡献如下：1)提出一种新颖的基于孪生网络的以运动为中心的跟踪方法，通过单阶段的运动建模实现准确的目标跟踪。2)设计了STFA模块，在多个尺度上整合相邻帧的特征，以有效地捕捉目标的运动信息。提出BFE模块，以一种简单的方式为运动预测提供目标的尺寸先验信息。3)大量实验表明，在Kitti、NuScenes和WOD数据集上超越了最先进的方法，同时展示出卓越的鲁棒性，并保持高推理速度。

1 方　法

1.1 概述

给定初始目标边界框(bounding box, BBox)$\boldsymbol{B}_1$和点云序列$\{P_t\}_{t=1}^{T}$，三维单目标跟踪任务的目标是在后续帧中定位一系列目标边界框$\boldsymbol{B}_t=(x_t,y_t,z_t,w_t,l_t,h_t)_{t=2}^{T}$，其中，$T$表示帧数，BBox的参数$(x,y,z)$，$(w,l,h)$和$\theta$分别定义了BBox的中心、大小和朝向角角度。在每个时间戳t，激光雷达点云$\boldsymbol{P}_t\in\mathbf{R}^{N_t\times 3}$包含$N_t$个点，3个通道编码了每个点的三维全局坐标。由于即使对于非刚性对象，跟踪目标跨帧的大小也几乎没有变化，我们假设目标大小是恒定的，只回归目标在两个连续的帧之间的水平偏移$(\Delta x,\Delta y,\Delta z)$和朝向角偏移$\Delta\theta$，以简化跟踪任务。

在时间戳$t(t>1)$时，两个连续帧的点云$\boldsymbol{P}_{t-1}$和$\boldsymbol{P}_t$，以及模型预测的$\boldsymbol{B}_{t-1}$是已知的。首先将点云从世界坐标系转换到相对于$\boldsymbol{B}_{t-1}$的规范框坐标系。然后预测两个连续帧之间的目标相对运动(relative target motion, RTM)。通过将RTM应用于边界框$\boldsymbol{B}_{t-1}$，可以得到当前的边界框$\boldsymbol{B}_t$。预测过程可以表示为

$$F(\boldsymbol{P}_{t-1},\boldsymbol{P}_t,\boldsymbol{B}_{t-1})\mapsto(\Delta x,\Delta y,\Delta z,\Delta\theta)\tag{1}$$

基于式(1)，本文提出一种新颖的基于孪生网络的以运动为中心的跟踪方法(SiamMo)。如图2所示，SiamMo由3个主要部分组成：1)孪生特征提取模块(Siamese feature extraction, SFE)；2)时空特征聚合模块(STFA)；3)边界框感知特征编码模块(BFE)。SFE首先将原始点云划分为规则的体素，然后使用Voxel-to-BEV(V2B)网络将其编码为多尺度鸟瞰(bird's eye view, BEV)特征图。接下来，STFA聚合多尺度BEV特征图，在多个尺度上融合相邻帧的补充信息。对于刚性物体，BFE将目标的尺寸先验信息注入到运动特征中。最后，使用一个简单的多层感知器(multi-layer perception, MLP)进行运动预测。

1.2 孪生特征提取模块SFE

1.2.1 体素化

激光雷达扫描的点云具有不同的稀疏性——近距离的点密集排列，而远距离的点则稀疏得多。学习能够有效适应密集和稀疏区域的点特征是一个相当大的挑战。为了缓解这个问题，本文提出使用基于体素的表示(Zhou和Tuzel, 2018; Chen和He, 2021)进行特征提取，而不是像以前的工作那样使用基于点的表示。体素化操作将疏密不均的点云数据转化为规则的体素结构，有效缓解了数据密度差异导致的特征提取难题，为后续任务提供了稳定的数据基础。具体而言，给定两帧连续的点云$\boldsymbol{P}_{t-1}$和$\boldsymbol{P}_t$，首先对两者进行联合空间离散化操作。设体素网格尺寸为(v_x,v_y,v_z)，则点云沿X,Y,Z轴划分为尺寸为W,H,D的三维体素空间。具体为

$$W = \frac{x_{\max} - x_{\min}}{v_x};\ H = \frac{y_{\max} - y_{\min}}{v_y}$$
$$D = \frac{z_{\max} - z_{\min}}{v_z} \tag{2}$$

式中，$x_{\min}$、$y_{\min}$、$z_{\min}$和$x_{\max}$、$y_{\max}$、$z_{\max}$分别为点云输入的边界坐标。为简单起见，假设W、H、D是v_x、v_y、v_z的倍数。这里对于非空体素v_k，其初始特征通过空间均值池化计算。具体为

$$v_k = \frac{1}{N}\sum_{i=1}^{N} p_k(i) \tag{3}$$

式中，N为体素内有效点数，$p_k(i)$是第k个非空体素中的具有三维坐标(x, y, z)的第i个点。仅处理非空体素获得体素特征的列表，其中每个特征都与特定非空体素的空间坐标一一对应。所获取的体素特征列表能够用一个特定大小$3 \times D \times H \times W$的张量$\boldsymbol{\mathcal{V}}$来呈现。如图2所示，相较于原始点云，体素均值特征通过局部区域统计特性有效抑制了孤立噪声点的影响，同时将动态变化的点密度转化为稳定的网格密度，显著提升了特征表达的一致性。此外，体素特征对点计数变化的敏感性降低，在一定程度上缓解了样本疏密程度不一致的问题。

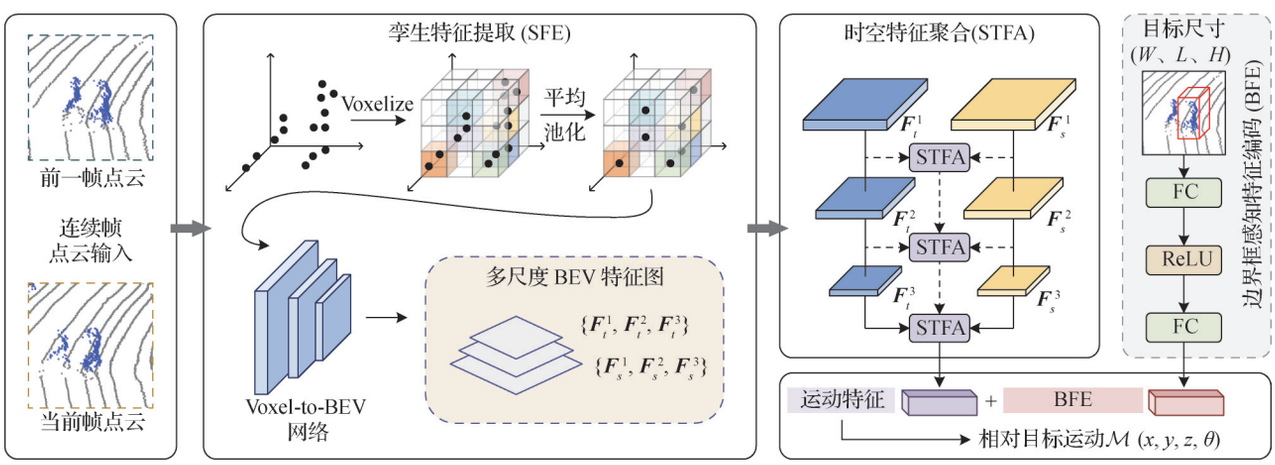

图2　SiamMo架构

Fig. 2　Architecture of SiamMo

1.2.2　Voxel-to-BEV网络

为了学习前一帧和当前帧的判别性特征，本文提出一个自上而下的基于稀疏卷积的孪生主干网络。该网络由两部分组成：稀疏体素特征提取(sparse voxel feature extraction，SVFE)和密集BEV特征提取(dense bev feature extraction，DBFE)。

SVFE模块采用多层3D稀疏卷积网络，每层由2个卷积核大小为$3 \times 3 \times 3$、步长为1的子流型卷积(Graham和Van Der Maaten，2018)组成。子流型卷积仅对非空体素进行操作，保持空间分辨率不变，避免稀疏数据的冗余计算。在各层之间，本文方法插入1个卷积核大小为$3 \times 3 \times 3$、步长为2的3D稀疏卷积对体素特征进行下采样，以减少空间计算的复杂度。各层的运算过程为

$$\boldsymbol{V}'_{l-1} = ReLU(BN(SubConv(\boldsymbol{V}_{l-1}))) \tag{4}$$
$$\boldsymbol{V}''_{l-1} = ReLU(BN(SubConv(\boldsymbol{V}'_{l-1}))) \tag{5}$$
$$\boldsymbol{V}_l = ReLU(BN(SpConv(\boldsymbol{V}''_{l-1}))) \tag{6}$$

式中，$\boldsymbol{V}_l$为第l层输出，BN为批量归一化BN(batch normalization)，$ReLU$为激活函数ReLU(rectified linear unit)层，$SubConv$为步长为1的子流型卷积，$SpConv$为步长为2的3D稀疏卷积。给定体素特征$\boldsymbol{V}$，经过4层稀疏卷积网络后，得到8倍下采样后的低分辨率体素特征$\boldsymbol{V}_o \in \mathbf{R}^{C \times D' \times H' \times W'}$，其中$D' = D/8$，$H' = H/8$，$W' = W/8$。

为了进一步减轻不同稀疏性的不利影响，沿高度方向展平稀疏的3D特征$\boldsymbol{V}_o$以获得密集的BEV特征$\boldsymbol{F} \in \mathbf{R}^{C' \times H' \times W'}$。维度的展平过程为

$$C \times D' \times H' \times W' \mapsto CD' \times H' \times W' \tag{7}$$
$$CD' \times H' \times W' \mapsto C' \times H' \times W' \tag{8}$$

DBFE模块通过多层2D卷积来聚合特征，使跟踪目标可以在BEV特征图中获得足够的局部信息。DBFE的每一层由2个卷积核大小为3×3、步长为1的2D标准卷积组成。在各卷积之间依次插入批量归一化层和激活函数ReLU层。在各层之间，通过卷积核大小为3×3、步长为2的2D标准卷积对细化后的特征进行下采样，生成不同尺度的BEV特征

图。经过2层下采样后，最终得到3个不同尺度的前一帧和当前帧的BEV特征图，分别表示为$\{\boldsymbol{F}_t^1, \boldsymbol{F}_t^2, \boldsymbol{F}_t^3\}$和$\{\boldsymbol{F}_s^1, \boldsymbol{F}_s^2, \boldsymbol{F}_s^3\}$，分辨率依次为$H' \times W'$，$H'/2 \times W'/2$，$H'/4 \times W'/4$。

1.3 时空特征聚合STFA

STFA旨在在多个尺度上整合连续帧的特征，以进行运动建模。直观地说，有效的运动建模需要在多个尺度上有丰富的表示。整合这些多方面的表示有望显著增强网络对具有各种运动模式的目标进行精确定位的能力。

如图3所示，STFA整合多尺度BEV特征图$\{\boldsymbol{F}_t^1, \boldsymbol{F}_t^2, \boldsymbol{F}_t^3\}$和$\{\boldsymbol{F}_s^1, \boldsymbol{F}_s^2, \boldsymbol{F}_s^3\}$，将前一帧的目标信息传播到当前帧，并在不同尺度上提取运动特征。在每个尺度上，运动特征M^i计算为

$$\boldsymbol{F}_c^i = [\boldsymbol{F}_t^i, \boldsymbol{F}_s^i] \tag{9}$$

$$\boldsymbol{M}^i = Conv_{3\times3}(\boldsymbol{F}_c^i) \tag{10}$$

式中，$i \in \{1, 2, 3\}$表示不同尺度的索引，$[\cdot, \cdot]$表示通道维度的拼接操作，$Conv_{3\times3}$表示内核大小为3 × 3、步长为1的卷积层。特征$\boldsymbol{M}^i$通过下采样调整为与不同尺度下的$\boldsymbol{F}_c^{i+1}$相同大小。然后，它们进行逐元素相加，接着通过一个卷积层生成下一个尺度的运动特征$\boldsymbol{M}^{i+1}$，具体为

$$\boldsymbol{M}^{i+1} = Conv_{3\times3}(Down(\boldsymbol{M}^i) + [\boldsymbol{F}_t^{i+1}, \boldsymbol{F}_s^{i+1}]) \tag{11}$$

式中，$Down$表示下采样。在最后一层，对最终的运动特征应用全局最大池化（global max pooling，GMP），得到输出特征$\boldsymbol{M}^o$。

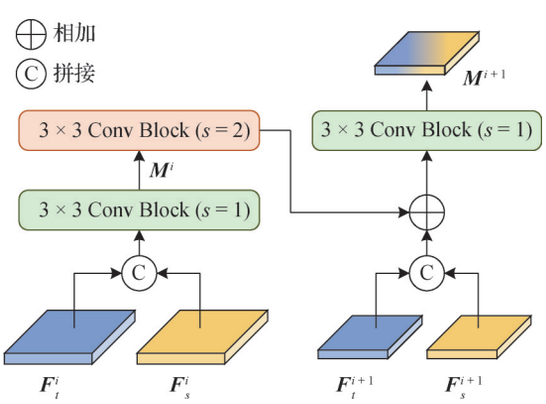

图3 STFA模块示意图

Fig. 3 Diagram of STFA module

1.4 边界框感知特征编码BFE

如先前的工作（Zheng等，2021，2022）所示，初始帧的真实边界框是一个很强的特定于目标的线索，可以显著提高跟踪性能。BAT（Zheng等，2021）通过设计一种尺寸感知和部位感知的BoxCloud特征来改进特征比较。M^2-Track（Zheng等，2022）将点坐标扩展为边界框感知特征，以提高目标分割和定位精度。

与这些研究不同，本文受位置编码（Vaswani等，2017）启发提出边界框感知特征编码（BFE）。具体而言，BFE首先使用多层感知机将目标尺寸(w, l, h)编码为边界框感知编码$\boldsymbol{E}_{\text{box}}$。$\boldsymbol{E}_{\text{box}}$与STFA的输出特征$\boldsymbol{M}^o$具有相同的通道维度。将边界框感知编码$\boldsymbol{E}_{\text{box}}$与特征$\boldsymbol{M}^o$相加，通过另一个多层感知机（$MLP$）回归相对目标运动$\boldsymbol{M}$，具体为

$$\boldsymbol{M} = MLP(\boldsymbol{M}^o + \boldsymbol{E}_{\text{box}}) \tag{12}$$

2 实 验

2.1 数据集和评价指标

本文主要在3个广泛使用且具有挑战性的数据集上验证算法的有效性：Kitti（Geiger等，2012）、NuScenes（Caesar等，2020）和WOD（Sun等，2020）。Kitti数据集包含21个用于训练的视频序列和29个用于测试的视频序列。由于测试集标签无法获取，按照先前工作（Qi等，2020；Loshchilov和Hutter，2019；Liu等，2019；Liu等，2023；Liang等，2020；Luo等，2021）的方法将训练集划分为训练/验证/测试子集。NuScenes数据集包含1 000个场景，划分为700/150/ 150个场景分别用于训练/验证/测试。按照Zheng等人（2021）的实现方式，在汽车、行人、卡车、拖车和公共汽车5个类别上与先前方法进行比较。对于WOD数据集，按照LiDAR-SOT（LiDAR single-object tracking）（Pang等，2021）的方法进行评估，根据每个轨迹第1帧中的点数将其划分为简单、中等和困难3个子集。按照STNet（Hui等，2022）的做法，使用在Kitti数据集上训练的模型在WOD数据集上进行评估，以评估跟踪器的泛化能力。

按照Qi等人（2020）的方法，本文采用一次通过评估（one pass evaluation，OPE）（Kristan等，2016）中定义的成功率（Success）和精确率（Precision）作为评估指标。成功率表示预测框与真实框的交并比（intersection over union，IoU）大于某个阈值的帧的比例曲线下面积（area under curve，AUC），阈值范围为0～1；准确率表示预测框与真实框中心距离在某个

阈值(范围为0～2 m)内的帧的比例曲线下面积。

2.2 实现细节

本文实验的计算机系统环境为Ubuntu20.04,使用的显卡为两张NVIDIA RTX 4090 GPU,软件环境为Python3.9.0。算法基于深度学习框架PyTorch2.0.1和MMEngine0.7.4实现,CUDA版本为11.8。稀疏卷积的实现基于开源算子库spconv实现,版本为2.3.6。训练过程使用AdamW优化器更新网络参数,权重衰减为1.0×10^{-3}(偏置参数和归一化层参数的衰减系数为0),学习率为1.0×10^{-4},批量大小为256,并使用最大范数为20的L_2范数梯度截断。在推理过程中,模型在给定第1帧处的目标边界框的点云序列中逐帧跟踪目标。

训练目标是通过最小化单个回归损失以优化运动估计的结果。回归损失的选择有很多,可以是固定的损失如L_1或L_2,也可以是基于学习的动态损失RLE(residual log-likelihood estimation)(Li等,2021)。本文采用RLE损失作为训练的损失函数。

为了模拟预测过程中的随机性,本文分别从均匀分布$U(-6^\circ, 6^\circ)$和$U(-0.3\ \text{m}, 0.3\ \text{m})$中采样对第$t-1$帧的目标的真实边界框进行随机旋转和偏移。为了鼓励模型在训练过程中学习多种模式的目标运动,本文从正态分布$N(0, \Sigma)$中采用对第t帧的目标真实边界框进行随机偏移,其中$\Sigma=(0.3, 0.2, 0.1)$,分别对应X、Y、Z坐标轴方向。为了进一步增加数据的多样性并提高模型的鲁棒性,对数据样本在空间和时间上进行随机翻转。空间上,同时水平翻转点云和目标边界框;时间上,翻转点云序列的播放方向。

2.3 对比实验

2.3.1 Kitti和WOD数据集上的结果

在Kitti数据集的Car、Pedestrian、Van和Cyclist 4个跟踪类别上对所提出的方法与先前的先进方法进行全面比较。如表1所示。SiamMo在各个类别中均表现出卓越性能,分别实现72.3%的平均Success和90.1%的平均Precision,为所有方法中的最高值。此外,SiamMo在平均Success和Precision上分别比对比方法中最先进方法MBPTrack(Xu等,2023b)高出2.0%和2.2%。与以运动为中心的跟踪器M^2-Track(Zheng等,2022)相比,SiamMo在平均Success和Precision方面分别大幅领先9.3%和6.7%,充分发挥了以运动为中心的跟踪方法的潜力。

为了更直观、深入地对比不同方法在Kitti数据集上的跟踪效果,该部分实验对多种方法的跟踪过程进行了可视化处理,着重选取了具有代表性的Car序列和Pedestrian序列进行展示,如图4所示。在点云极其稀疏的汽车序列中,M^2-Track在跟踪过程中丢失了目标。SiamMo和MBPTrack都能正确跟踪到最后一帧,而SiamMo的定位更加准确。在行人序列中,当类内干扰物靠近时,M^2-Track和MBPTrack选择了错误的对象进行跟踪,而SiamMo能够生成稳定可靠的预测。

在复杂多变的现实应用场景中,模型的泛化能力是衡量其性能优劣的关键指标之一,直接决定了模型在未见过的数据上能否保持良好的表现,这对于三维目标跟踪算法的实际应用至关重要。为了深入探究本文方法在不同场景下的泛化能力,该部分实验按照先前方法(Hui等,2021;Graham等,2018;Wang等,2019)的做法,基于Kitti数据集Car和Pedestrian序列上训练得到的模型,分别在WOD数据集的Vehicle和Pedestrian序列上开展了全面且细致的泛化性实验。实验将场景难度细致划分为Easy(简单)、Medium(中等)、Hard(困难)3个等级,不同等级的场景在目标的可见性、遮挡情况以及运动复杂性等方面存在显著差异。如表2所示,在不同的稀疏程度下,本文的SiamMo在WOD跟踪榜单的平均Success和Precision指标上取得目前最先进的47.5%和63.6%,在行人类别上的优势尤为明显。相较于之前榜单中最先进的跟踪器CXTrack,本文方法在平均Success和Precision指标上分别取得5.3%和6.9%的显著提升,表明本文方法在不可见的复杂场景中具有极强的泛化能力,即使面对从未接触过的场景,也能够稳定且准确地跟踪目标。

2.3.2 NuScenes数据集上的结果

与Kitti数据集(Geiger等,2012)相比,NuScenes数据集(Caesar等,2020)对3D SOT任务提出更严峻的挑战,这主要归因于其更大的数据量和更稀疏的注释(NuScenes数据集为2 Hz,而Kitti和WOD数据集为10 Hz)。按照M^2-Track(Zheng等,2022)建立的方法,本文在NuScenes数据集上与包括SC3D(Giancola等,2019)、P2B(Qi等,2020)、PTT(Shan等,2021)、BAT(Zheng等,2021)、PTTR(Zhou等,2022)、M^2-Track(Zheng等,2022)、MBPTrack(Xu等,2023b)和StreamTrack(Luo等,2024)在内的先前方法进行对

表 1　在 Kitti 数据集上与其他先进方法的比较

Table 1　Comparison with state-of-the-art methods on Kitti dataset

/%

方法	类别				
	Car	Pedestrian	Van	Cyclist	和值
帧数	6 424	6 088	1 248	308	14 068
SC3D(Giancola 等,2019)	41.3 / 57.9	18.2 / 37.8	40.4 / 47.0	41.5 / 70.4	31.2 / 48.5
P2B(Qi 等,2020)	56.2 / 72.8	28.7 / 49.6	40.8 / 48.4	32.1 / 44.7	42.4 / 60.0
BAT(Zheng 等,2021)	60.5 / 77.7	42.1 / 70.1	52.4 /67.0	33.7 / 45.4	51.2 / 72.8
PTT(Shan 等,2021)	67.8 / 81.8	44.9 / 72.0	43.6 / 52.5	37.2 / 47.3	55.1 / 74.2
V2B(Hui 等,2021)	70.5 / 81.3	48.3 / 73.5	50.1 / 58.0	40.8 / 49.7	58.4 / 75.2
PTTR(Zhou 等,2022)	65.2 / 77.4	50.9 / 81.6	52.5 / 61.8	65.1 / 90.5	58.4 / 77.8
STNet(Hui 等,2022)	72.1 / 84.0	49.9 / 77.2	58.0 / 70.6	73.5 / 93.7	61.3 / 80.1
CMT(Guo 等,2022)	70.5 / 81.9	49.1 / 75.5	54.1 / 64.1	55.1 / 82.4	59.4 / 77.6
TAT(Lan 等,2022)	72.2 / 83.3	57.4 / 84.4	58.9 / 69.2	74.2 / 93.9	64.7 / 82.8
M^2-Track(Zheng 等,2022)	65.5 / 80.8	61.5 / 88.2	53.8 / 70.7	73.2 / 93.5	62.9 / 83.4
CXTrack(Xu 等,2023a)	69.1 / 81.6	67.0 / 91.5	60.0 / 71.8	74.2 / 94.3	67.5 / 85.3
SyncTrack(Ma 等,2023)	73.3 / 85.0	54.7 / 80.5	60.3 / 70.0	73.1 / 93.8	64.1 / 81.9
HVTrack(Wu 等,2024)	68.2 / 79.2	64.6 / 90.6	54.8 / 63.8	72.4 / 93.7	65.5 / 83.1
HyGFNet(Cui 等,2024)	68.0 / 79.3	60.0 / 86.6	56.7 / 67.1	75.9 / 93.9	63.7 / 85.6
MBPTrack(Xu 等,2023b)	73.4 / 84.8	68.6 / 93.9	61.3 / 72.7	76.7 / 94.3	70.3 / 87.9
SiamMo(本文)	76.3 / 88.1	68.6 / 93.9	67.9 / 80.5	78.5 / 94.8	72.3 / 90.1

注:“/”前后分别表示 Success 和 Precision 指标。和值下的性能为对所有类别的性能求均值所得。

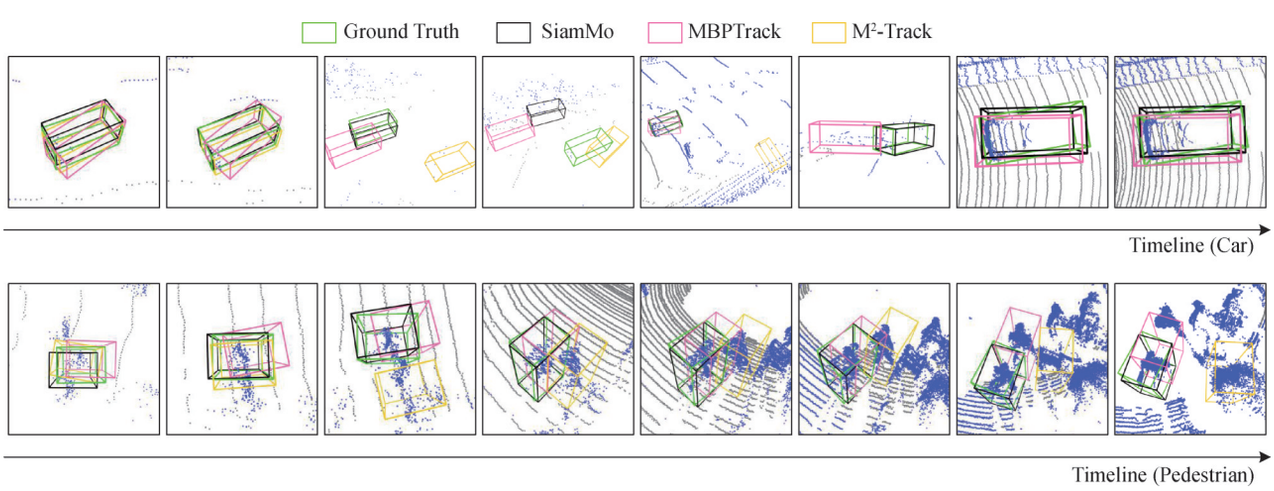

图 4　在 Kitti 数据集上的跟踪可视化

Fig. 4　Visualization of tracking results on Kitti dataset

比评估。如表 3 所示,SiamMo 在所有类别中均超过了所有竞争对手,且大多类别中具有较大优势。在这个大规模数据集上,SiamMo 相比 M^2-Track 的优势更为显著,Success 提高 11.08%,Precision 提高 9.95%。

2.3.3　对稀疏性的鲁棒性实验

在点云三维目标跟踪算法研究领域,点云稀疏

表2　在WOD数据集上与最新方法的比较

Table 2　Comparison with state-of-the-art methods on WOD dataset

/%

方法	Vehicle				Pedestrian				和值
	Easy	Medium	Hard	和值	Easy	Medium	Hard	和值	
帧数	67 832	61 252	56 647	185 731	85 280	82 253	74 219	241 752	427 483
P2B(Qi等,2020)	57.1 / 65.4	52.0 / 60.7	47.9 / 58.5	52.6 / 61.7	18.1 / 30.8	17.8 / 30.0	17.7 / 29.3	17.9 / 30.1	33.0 / 43.8
BAT(Zheng等,2021)	61.0 / 68.3	53.3 / 60.9	48.9 / 57.8	54.7 / 62.7	19.3 / 32.6	17.8 / 29.8	17.2 / 28.3	18.2 / 30.3	34.1 / 44.4
V2B(Hui等,2021)	64.5 / 71.5	55.1 / 63.2	52.0 / 62.0	57.6 / 65.9	27.9 / 43.9	22.5 / 36.2	20.1 / 33.1	23.7 / 37.9	38.4 / 50.1
STNet(Hui等,2022)	65.9 / 72.7	57.5 / 66.0	54.6 / 64.7	59.7 / 68.0	29.2 / 45.3	24.7 / 38.2	22.2 / 35.8	25.5 / 39.9	40.4 / 52.1
TAT(Lan等,2022)	66.0 / 72.6	56.6 / 64.2	52.9 / 62.5	58.9 / 66.7	32.1 / 49.5	25.6 / 40.3	21.8 / 35.9	26.7 / 42.2	40.7 / 52.8
M^2-Track(Zheng等,2022)	68.1 / 75.3	58.6 / 66.6	55.4 / 64.9	61.1 / 69.3	35.5 / 54.2	30.7 / 48.4	29.3 / 45.9	32.0 / 49.7	44.6 / 58.2
CXTrack(Xu等,2023a)	63.9 / 71.1	54.2 / 62.7	52.1 / 63.7	57.1 / 66.1	35.4 / 55.3	29.7 / 47.9	26.3 / 44.4	30.7 / 49.4	42.2 / 56.7
SiamMo(本文)	66.1 / 73.4	58.1 / 67.9	56.6 / 68.7	60.6 / 70.1	44.8 / 66.2	35.6 / 56.8	31.2 / 51.8	37.5 / 58.6	47.5 / 63.6

注："/"前后分别表示Success和Precision指标,每个类别和值下的性能均为Easy、Medium和Hard性能求均值的结果。

表3　在NuScenes数据集上与最新方法的比较

Table 3　Comparison with state-of-the-art methods on NuScenes dataset

/%

方法	类别					
	Car	Pedestrian	Truck	Trailer	Bus	和值
帧数	64 159	33 227	13 587	3 352	2 953	117 278
SC3D(Giancola等,2019)	22.31 / 21.93	11.29 / 12.65	30.67 / 27.73	35.28 / 28.12	29.35 / 24.08	20.70 / 20.20
P2B(Qi等,2020)	38.81 / 43.18	28.39 / 52.24	42.95 / 41.59	48.96 / 40.05	32.95 / 27.41	36.48 / 45.08
PTT(Shan等,2021)	41.22 / 45.26	19.33 / 32.03	50.23 / 48.56	51.70 / 46.50	39.40 / 36.70	36.33 / 41.72
BAT(Zheng等,2021)	40.73 / 43.29	28.83 / 53.32	45.34 / 42.58	52.59 / 44.89	35.44 / 28.01	38.10 / 45.71
PTTR(Zhou等,2022)	51.89 / 58.61	29.90 / 45.09	45.30 / 44.74	45.87 / 38.36	43.14 / 37.74	44.50 / 52.07
M^2-Track(Zheng等,2022)	55.85 / 65.09	32.10 / 60.92	57.36 / 59.54	57.61 / 58.26	51.39 / 51.44	49.23 / 62.73
MBPTrack(Xu等,2023b)	62.47 / 70.41	45.32 / 74.03	62.18 / 63.31	65.14 / 61.33	55.41 / 51.76	57.48 / 69.88
StreamTrack(Luo等,2024)	62.05 / 70.81	38.43 / 68.58	64.67 / 66.60	66.67 / 64.27	60.66 / 59.74	55.75 / 69.22
SiamMo(本文)	64.95 / 72.24	46.23 / 76.25	68.22 / 68.81	74.21 / 70.63	65.63 / 62.07	60.31 / 72.68

注："/"前后分别表示Success和Precision指标,和值下的性能为对所有类别的性能求均值所得。

问题始终是一大严峻挑战,跟踪器在稀疏场景下的鲁棒性测试对于评估算法的可靠性至关重要。为了验证本文方法在不同稀疏程度下的鲁棒性,该部分实验将本文所提出的方法与3个代表性的方法,包括BAT(Zheng等,2021)、PTTR(Zhou等,2022)和M^2-Track(Zheng等,2022)进行了全面且细致的对比,使用平均Success作为评估指标。通过第1帧中的目标点的数量直观反映场景的稀疏程度。根据不同程度的稀疏度,将Kitti和NuScenes这两个数据集的Car类别划分为多个测试子集。

研究结果显示,本文方法在不同程度的稀疏场景下都展现出优势。从图5(数轴范围为0.2～0.9)中可以清晰地看出,在相对密集的点云场景下,4种跟踪器的性能差距不大。但是,在Kitti_[0, 10)、Kitti_[10, 20)、NuScenes_[0, 10)、NuScenes_[10, 20)等极度稀疏的场景下,SiamMo的跟踪性能远远超过

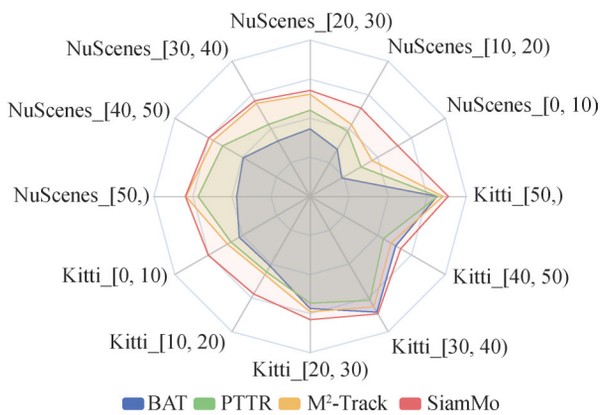

图 5　对稀疏性的鲁棒性

Fig. 5　Robustness to sparsity

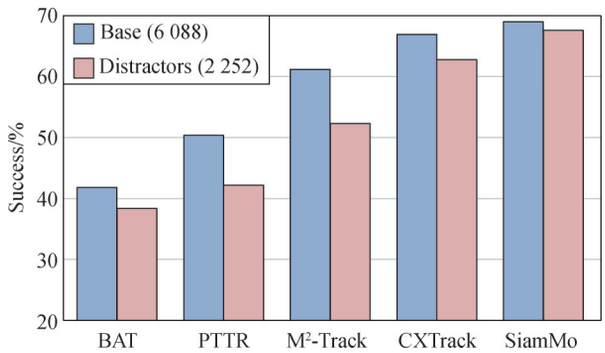

图 6　对干扰物的鲁棒性

Fig. 6　Robustness to distractors

了另外 3 个对比方法。这表明 SiamMo 在密集点云场景下保持高性能跟踪的同时，在稀疏点云场景下依然能实现准确的跟踪。由于利用了相邻帧的补充信息，SiamMo 在稀疏场景中比 M²-Track 具有显著优势。与基于孪生匹配的方法 BAT 和 PTTR 相比，SiamMo 在 NuScenes_[0, 10) 子集中的性能分别提高 28. 9% 和 19. 0%。尽管随着点数接近 50，性能差距逐渐缩小，但 SiamMo 始终优于其他对比方法。这主要得益于体素表征有效降低了样本中点云分布的疏密差异，减小了特征提取的难度。

2. 3. 4　对干扰物的鲁棒性实验

在复杂多变的现实环境中，类内干扰物的存在对缺乏纹理信息的点云跟踪构成了严峻考验。为了进一步分析 SiamMo 对类内干扰物的鲁棒性，本文人工挑选出 14 个目标周围存在干扰物的序列，以评估跟踪器在 Kitti 行人类别上的性能，使用平均 Success 作为评估指标。如图 6 所示，除了 SiamMo，其他跟踪器的性能均显著下降。这表明 SiamMo 对干扰物具有鲁棒性，反映了 SiamMo 在复杂现实场景中应用的潜力。

2. 3. 5　推理速度实验

在模型性能评估中，模型参数量与运行速度是衡量算法效率和实用性的关键指标。该部分实验对三维目标跟踪领域的典型方法和本节方法的推理速度、参数量和计算复杂度进行了统计。从表 4 可以看出，SiamMo 仅有 0. 82 GFLOPs 的计算复杂度。在参数量方面，虽然 SiamMo 的参数量为 14. 6 M，在所有对比方法中并非最小，但结合其运行速度和计算复杂度来看，它在平衡计算资源与性能方面表现出色。按照 P2B（Qi 等，2020）的方法，本文记录了 Kitti 数据集上汽车类所有测试帧的平均运行时间，以评估本文方法的计算效率。SiamMo 达到了 108 帧/s，包括在单个 NVIDIA 4090 GPU 上进行点云预处理的 4. 2 ms 和网络前向传播的 5. 0 ms。在相同的硬件设置下，SiamMo 的运行速度是先前的最先进方法 MBPTrack（Xu 等，2023b）的 1. 5 倍。

2. 4　消融实验

2. 4. 1　孪生架构的消融实验

本文研究了孪生架构的有效性。如表 5 所示，对于使用“单流”架构的实验，首先在特征维度上连接两个连续帧的体素特征，然后使用与主要实验相同的编码器提取连接后的特征。这种单路径架构缺乏对目标信息的有效整合，与“孪生”架构相比存在性能差距。为了进一步分析权重共享机制，采用了独立的“双流”架构，其性能略有下降。需要注意的是，与“双流”架构相比，“孪生”架构具有更快的推理速度和更少的参数。

2. 4. 2　STFA 模块的消融实验

多尺度特征在目标跟踪任务中对于准确捕捉目标特征、提升定位精度具有重要意义。为了研究

表 4　效率分析

Table 4　Efficiency analysis

方法	帧率/（帧/s）	参数量/M	每秒浮点运算次数/G
P2B	85	1.34	4.33
BAT	94	1.48	2.81
V2B	57	1.32	5.54
CXTrack	48	18.3	4.63
MBPTrack	74	7.4	2.88
SiamMo（本文）	108	14.6	0.82

表5　孪生架构的消融实验

Table 5　Ablation study of Siamese architecture

/%

架构	Car	Pedestrian	Van	Cyclist
单流	72.5 / 84.3	61.8 / 88.1	61.7 / 74.8	72.1 / 93.6
双流	76.0 / 87.8	67.5 / 93.1	67.2 / 79.7	75.1 / 94.2
孪生	76.3 / 88.1	68.6 / 93.9	67.9 / 80.5	78.5 / 94.8

注："/"前后分别表示Success和Precision指标。

STFA中多尺度聚合的影响，训练了4个具有不同聚合方法的变体，如表6所示。S1、S2、S3分别表示{1, 2, 4}特征步长的多尺度BETV特征图。通过实验发现，仅使用S3就已超过现有最先进方法。此外，除了货车类别，结合两个尺度的特征图能带来进一步的性能提升。在聚合所有3个尺度的特征图后，所有类别的成功率和准确率都稳步提高。因此，将这种配置作为默认设置。

2.4.3　BFE模块的消融实验

本文研究了BFE模块的有效性，如表7和表8所示。由于刚性物体在运动过程中大小保持不变，通过BFE在刚性物体的跟踪中利用了这一先验知识。对于禁用BFE的实验，直接从运动特征中回归相对目标运动。在Kitti和NuScenes数据集上，启用BFE的模型在所有刚性物体上的表现均显著优于未启用BFE的模型。结果表明，BFE能够有效帮助跟踪器更准确地定位目标。

表6　STFA模块的消融实验

Table 6　Ablation study of STFA module

/%

S1	S2	S3	Car	Pedestrian	Van	Cyclist
–	–	√	74.1 / 85.5	66.6 / 90.8	67.7 / 80.8	73.8 / 94.0
–	√	√	75.5 / 87.5	67.1 / 92.0	67.5 / 77.5	74.2 / 93.7
√	–	√	74.2 / 87.6	67.5 / 92.8	64.2 / 77.3	74.5 / 94.4
√	√	√	76.3 / 88.1	68.6 / 93.9	67.9 / 80.5	78.5 / 94.8

注："/"前后分别表示Success和Precision指标。"√"表示采用，"–"表示未采用。

表7　BFE模块在Kitti数据集的消融实验

Table 7　Ablation study of BFE module on Kitti dataset

/%

BFE	Car	Van	Car
–	73.4 / 84.5	64.4 / 75.2	62.9 / 70.2
√	76.3 / 88.1	67.9 / 80.5	65.0 / 72.2

注："/"前后分别表示Success和Precision指标。"√"表示采用，"–"表示未采用。

表8　BFE模块在NuScenes数据集的消融实验

Table 8　Ablation study of BFE module on NuScenes dataset

/%

BFE	Truck	Trailer	Bus
–	66.7 / 67.3	71.7 / 67.7	61.9 / 57.9
√	68.2 / 68.8	74.2 / 70.6	65.6 / 62.1

注："/"前后分别表示Success和Precision指标。"√"表示采用，"–"表示未采用。

3　结　论

本文提出一种全新的基于孪生网络的以运动为中心的跟踪方法——SiamMo。该方法通过简单的单阶段跟踪流程，实现了孪生特征提取和运动建模。具体而言，设计了时空特征聚合（STFA）模块，它能够在多个尺度上整合相邻帧的特征，从而有效捕捉运动信息。此外，还设计了边界框感知特征编码（BFE）模块，利用明确的目标边界框先验信息进行运动预测。大量实验表明，SiamMo高效且快速，在Kitti、NuScenes和WOD这3个具有挑战性的基准测试中超越了现有最优方法，推理速度可达108帧/s。而且，SiamMo在处理稀疏点云和存在干扰物的场景时，也展现出了出色的鲁棒性。希望这项研究能为3D单目标跟踪研究提供有价值的思路，激发人们对以孪生运动为核心的范式展开更深入的探索。

参考文献（References）

Caesar H, Bankiti V, Lang A H, Vora S, Liong V E, Xu Q, et al. 2020. nuScenes: a multimodal dataset for autonomous driving//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 11618-11628[DOI: 10.1109/CVPR42600.2020.01164]

Chen X L and He K M. 2021. Exploring simple Siamese representation learning//Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA: IEEE: 15745-15753[DOI: 10.1109/cvpr46437.2021.01549]

Cui Y B, Fang Z, Li Z H, Li S and Lin Y. 2024. HyGFNet: hybrid

geometry-flow learning network for 3D single object tracking. IEEE Transactions on Intelligent Vehicles, 9(10): 6407-6418 [DOI: 10.1109/TIV.2024.3367887]

Geiger A, Lenz P and Urtasun R. 2012. Are we ready for autonomous driving? The Kitti vision benchmark suite//Proceedings of 2012 IEEE Conference on Computer Vision and Pattern Recognition. Providence, USA: IEEE: 3354-3361 [DOI: 10.1109/CVPR.2012.6248074]

Giancola S, Zarzar J and Ghanem B. 2019. Leveraging shape completion for 3D Siamese tracking//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA: IEEE: 1359-1368 [DOI: 10.1109/CVPR.2019.00145]

Graham B, Engelcke M and Van Der Maaten L. 2018. 3D semantic segmentation with submanifold sparse convolutional networks//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE: 9224-9232 [DOI: 10.1109/CVPR.2018.00961]

Graham B and Van Der Maaten L. 2018. Submanifold sparse convolutional networks//Proceedings of 2018 IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City, UT, USA: IEEE: 9224-9232 [DOI: 10.48550/arXiv.1706.01307]

Guo Z Y, Mao Y Y, Zhou W G, Wang M and Li H Q. 2022. CMT: context-matching-guided transformer for 3D tracking in point clouds//Proceedings of the 17th European Conference on Computer Vision. Tel Aviv, Israel: Springer: 95-111 [DOI: 10.1007/978-3-031-20047-2_6]

Huang R, Zhang W Y, Kundu A, Pantofaru C, Ross D A, Funkhouser T, et al. 2020. An LSTM approach to temporal 3D object detection in LiDAR point clouds//Proceedings of the 16th European Conference on Computer Vision. Glasgow, UK: Springer: 266-282 [DOI: 10.1007/978-3-030-58523-5_16]

Hui L, Wang L P, Cheng M M, Xie J and Yang J. 2021. 3D Siamese voxel-to-BEV tracker for sparse point clouds [EB/OL]. [2025-03-26]. https://arxiv.org/pdf/2111.04426.pdf

Hui L, Wang L P, Tang L H, Lan K H, Xie J and Yang J. 2022. 3D Siamese transformer network for single object tracking on point clouds//Proceedings of the 17th European Conference on Computer Vision. Tel Aviv, Israel: Springer: 293-310 [DOI: 10.1007/978-3-031-20086-1_17]

Kristan M, Matas J, Leonardis A, Vojíř T, Pflugfelder R, Fernández G, et al. 2016. A novel performance evaluation methodology for single-target trackers. IEEE Transactions on Pattern Analysis and Machine Intelligence, 38(11): 2137-2155 [DOI: 10.1109/TPAMI.2016.2516982]

Lan K H, Jiang H B and Xie J. 2022. Temporal-aware Siamese tracker: integrate temporal context for 3D object tracking//Proceedings of the 16th Asian Conference on Computer Vision. Macau, China: Springer: 20-35 [DOI: 10.1007/978-3-031-26319-4_2]

Li B, Wu W, Wang Q, Zhang F Y, Xing J L and Yan J J. 2019. SiamRPN++: evolution of Siamese visual tracking with very deep networks//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Long Beach, USA: IEEE: 4282-4291 [DOI: 10.1109/CVPR.2019.00441]

Li J F, Bian S Y, Zeng A L, Wang C, Pang B, Liu W T, et al. 2021. Human pose regression with residual log-likelihood estimation//Proceedings of 2021 IEEE/CVF International Conference on Computer Vision (ICCV). Montreal, Canada: IEEE: 11005-11014 [DOI: 10.1109/iccv48922.2021.01084]

Li X, Zha Y F, Zhang T Z, Cui Z, Zuo W M, Hou Z Q, et al. 2019. Survey of visual object tracking algorithms based on deep learning. Journal of Image and Graphics, 24(12): 2057-2080 (李玺, 查宇飞, 张天柱, 崔振, 左旺孟, 侯志强, 等. 2019. 深度学习的目标跟踪算法综述. 中国图象图形学报, 24(12): 2057-2080) [DOI: 10.11834/jig.190372]

Liang M, Yang B, Zeng W Y, Chen Y, Hu R, Casas S, et al. 2020. PnPNet: end-to-end perception and prediction with tracking in the loop//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle, USA: IEEE: 11550-11559 [DOI: 10.1109/CVPR42600.2020.01157]

Liu X Y, Yan M Y and Bohg J. 2019. MeteorNet: deep learning on dynamic 3D point cloud sequences//Proceedings of 2019 IEEE/CVF International Conference on Computer Vision (ICCV). Seoul, Korea (South): IEEE: 9245-9254 [DOI: 10.1109/ICCV.2019.00934]

Liu Z J, Tang H T, Amini A, Yang X Y, Mao H Z, Rus D L, et al. 2023. BEVFusion: multi-task multi-sensor fusion with unified bird's-eye view representation//Proceedings of 2023 IEEE International Conference on Robotics and Automation (ICRA). London, UK: IEEE: 2774-2781 [DOI: 10.1109/icra48891.2023.10160968]

Loshchilov I and Hutter F. 2019. Decoupled weight decay regularization. [EB/OL]. [2025-03-26]. https://arxiv.org/pdf/1711.05101.pdf

Luo C X, Yang X D and Yuille A. 2021. Exploring simple 3D multi-object tracking for autonomous driving//Proceedings of 2021 IEEE/CVF International Conference on Computer Vision. Montreal, Canada: IEEE: 10468-10477 [DOI: 10.1109/iccv48922.2021.01032]

Luo Z P, Zhang G J, Zhou C Q, Wu Z H, Tao Q Y, Lu L W, et al. 2024. Modeling continuous motion for 3D point cloud object tracking//Proceedings of the 39th AAAI Conference on Artificial Intelligence. Washington, USA: AAAI: 4026-4034 [DOI: 10.1609/aaai.v38i5.28196]

Ma T L, Wang M M, Xiao J M, Wu H F and Liu Y. 2023. Synchronize feature extracting and matching: a single branch framework for 3D object tracking//Proceedings of 2023 IEEE/CVF International Conference on Computer Vision. Paris, France: IEEE: 9919-9929 [DOI: 10.1109/iccv51070.2023.00913]

Pang Z Q, Li Z C and Wang N Y. 2021. Model-free vehicle tracking and state estimation in point cloud sequences//Proceedings of 2021

IEEE/RSJ International Conference on Intelligent Robots and Systems. Prague, Czech Republic: IEEE: 8075-8082 [DOI: 10.1109/iros51168.2021.9636202]

Qi C R, Su H, Mo K C and Guibas L J. 2017a. PointNet: deep learning on point sets for 3D classification and segmentation//Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, USA: IEEE: 77-85 [DOI: 10.1109/cvpr.2017.16]

Qi C R, Litany O, He K M and Guibas L J. 2019. Deep Hough voting for 3D object detection in point clouds//Proceedings of 2019 IEEE/CVF International Conference on Computer Vision. Seoul, Korea (South): IEEE: 9276-9285 [DOI: 10.1109/iccv.2019.00937]

Qi C R, Yi L, Su H and Guibas L J. 2017b. PointNet++: deep hierarchical feature learning on point sets in a metric space//Proceedings of the 31st International Conference on Neural Information Processing Systems. Long Beach, USA: Curran Associates Inc.: 5105-5114 [DOI: 10.5555/3295222.3295263]

Qi H Z, Feng C, Cao Z G and Xiao Y. 2020. P2B: point-to-box network for 3D object tracking in point clouds//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle, USA: IEEE: 6328-6337 [DOI: 10.1109/CVPR42600.2020.00636]

Shan J Y, Zhou S F, Fang Z and Cui Y B. 2021. PTT: point-track-transformer module for 3D single object tracking in point clouds//Proceedings of 2021 IEEE/RSJ International Conference on Intelligent Robots and Systems. Prague, Czech Republic: IEEE: 1310-1316 [DOI: 10.1109/IROS51168.2021.9636821]

Sun P, Kretzschmar H, Dotiwalla X, Chouard A, Patnaik V, Tsui P, et al. 2020. Scalability in perception for autonomous driving: Waymo open dataset//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle, USA: IEEE: 2443-2451 [DOI: 10.1109/CVPR42600.2020.00252]

Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez A N, et al. 2017. Attention is all you need//Proceedings of the 31st International Conference on Neural Information Processing Systems. Long Beach, USA: Curran Associates, Inc.: 6000-6010 [DOI: 10.5555/3295222.3295349]

Wang M M, Yang X Q and Liu Y. 2022. A spatio-temporal encoded network for single object tracking. Journal of Image and Graphics, 27(9): 2733-2748 (王蒙蒙, 杨小倩, 刘勇. 2022. 利用时空特征编码的单目标跟踪网络. 中国图象图形学报, 27(9): 2733-2748) [DOI: 10.11834/jig.211157]

Wang Y, Sun Y B, Liu Z W, Sarma S E, Bronstein M M and Solomon J M. 2019. Dynamic graph CNN for learning on point clouds. ACM Transactions on Graphics (TOG), 38(5): #146 [DOI: 10.1145/3326362]

Wu Q, Sun K, An P, Salzmann M, Zhang Y N and Yang J Q. 2024. 3D single-object tracking in point clouds with high temporal variation//Proceedings of the 18th European Conference on Computer Vision. Milan, Italy: Springer: 279-296 [DOI: 10.1007/978-3-031-72667-5_16]

Xu T X, Guo Y C, Lai Y K and Zhang S H. 2023a. CXTrack: improving 3D point cloud tracking with contextual information//Proceedings of 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver, Canada: IEEE: 1084-1093 [DOI: 10.1109/cvpr52729.2023.00111]

Xu T X, Guo Y C, Lai Y K and Zhang S H. 2023b. MBPTrack: improving 3D point cloud tracking with memory networks and box priors//Proceedings of 2023 IEEE/CVF International Conference on Computer Vision. Paris, France: IEEE: 9877-9886 [DOI: 10.1109/iccv51070.2023.00909]

Zheng C D, Yan X, Gao J T, Zhao W B, Zhang W, Li Z, et al. 2021. Box-aware feature enhancement for single object tracking on point clouds//Proceedings of 2021 IEEE/CVF International Conference on Computer Vision. Montreal, Canada: IEEE: 13179-13188 [DOI: 10.1109/iccv48922.2021.01295]

Zheng C D, Yan X, Zhang H M, Wang B Y, Cheng S H, Cui S G, et al. 2022. Beyond 3D Siamese tracking: a motion-centric paradigm for 3D single object tracking in point clouds//Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, USA: IEEE: 8101-8110 [DOI: 10.1109/cvpr52688.2022.00794]

Zhou C Q, Luo Z P, Luo Y R, Liu T R, Pan L, Cai Z G, et al. 2022. PTTR: relational 3D point cloud object tracking with transformer//Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, USA: IEEE: 8521-8530 [DOI: 10.1109/cvpr52688.2022.00834]

Zhou Y and Tuzel O. 2018. VoxelNet: end-to-end learning for point cloud based 3D object detection//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE: 4490-4499 [DOI: 10.1109/CVPR.2018.00472]

作者简介

杨宇翔,男,教授,主要研究方向为计算机视觉和人工智能。E-mail:yyx@hdu.edu.cn

张敬,通信作者,男,教授,主要研究方向为计算机视觉和深度学习。E-mail:jingzhang.cv@whu.edu.cn

邓颖琦,男,硕士研究生,主要研究方向为计算机视觉和深度学习。E-mail:den@hdu.edu.cn

顾鸿杰,男,硕士研究生,主要研究方向为计算机视觉和深度学习。E-mail:hongjiegu@hdu.edu.cn

董哲康,男,教授,主要研究方向为深度学习。E-mail:englishp@126.com