计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
2025,34(8):53−61 [doi: 10.15888/j.cnki.csa.009951] [CSTR: 32024.14.csa.009951]

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

Siam-STM: 用于卫星视频目标跟踪的时空孪生网络①

顾权炜, 王　洁

(北京工业大学 计算机学院, 北京 100124)
通信作者: 顾权炜, E-mail: guquanwei2001@163.com

摘　要: 随着卫星视频成像技术的显著进步, 卫星视频中的目标跟踪任务引起了越来越多研究人员的关注. 然而之前的研究大多通过全局注意力机制获得空间信息, 这种方法使得模型关注背景部分从而忽略目标; 而且只利用视频帧中目标的空间信息, 目标定位不准确. 本文对现有的孪生网络目标跟踪模型 SiamCAR 进行改进, 提出时空孪生网络模型 Siam-STM. 具体来说, 本文提出基于注意力机制的空间信息感知模块, 聚合图像中的上下文信息并增强卫星视频中小目标特征的辨别力; 为了利用视频帧之间的时间信息, 本文还提出时间信息感知模块对视频中当前帧和历史帧进行融合, 从而学习到不同时刻目标的位置信息, 更好的关注目标轨迹, 缓解相似干扰物的影响. 此外, 为了缓解卫星视频中常见的遮挡影响, 本文在卡尔曼滤波器的基础上引入线性拟合方法, 进而提出一种运动估计机制, 可以有效地建模目标的运动特征进而在目标被遮挡时准确定位目标. 在 SatSOT 数据集上通过与现有先进模型的实验对比验证了 Siam-STM 的有效性.

关键词: 孪生网络; 卫星视频目标跟踪; 空间信息; 时间信息; 运动估计机制

引用格式: 顾权炜,王洁.Siam-STM: 用于卫星视频目标跟踪的时空孪生网络.计算机系统应用,2025,34(8):53–61. http://www.c-s-a.org.cn/1003-3254/9951.html

Siam-STM: Spatio-temporal Siamese Network for Object Tracking in Satellite Videos

GU Quan-Wei, WANG Jie

(College of Computer Science, Beijing University of Technology, Beijing 100124, China)

Abstract: With the significant progress of satellite video imaging technology, object tracking in satellite videos has attracted more and more researchers' attention. However, most of the previous research obtains spatial information through the global attention mechanism, which makes the model focus on the background part and thus ignore the object; moreover, only spatial information of the object in the video frames is utilized, resulting in inaccurate object localization. In this study, we improve the existing Siamese network object tracking model SiamCAR and a spatio-temporal Siamese network Siam-STM. Specifically, we propose a spatial information perception module based on the attention mechanism, which aggregates the contextual information in the images and enhances the discriminative capability of small object features in the satellite videos; to utilize the temporal information across video frames, a temporal information perception module is proposed to fuse the current frame with the historical frames, enabling the position information of the object across time to be learned, the object's trajectory to be better tracked, and the interference from similar objects to be mitigated. In addition, to mitigate the effects of occlusion in satellite videos, this study introduces a linear fitting method based on the Kalman filter and then proposes a motion estimation mechanism. This mechanism can effectively model the motion characteristics of the object, allowing accurate localization even during occlusions. The effectiveness of Siam-

① 收稿时间: 2024-12-29; 修改时间: 2025-02-12, 2025-03-18; 采用时间: 2025-03-24; csa 在线出版时间: 2025-06-20
CNKI 网络首发时间: 2025-06-24

STM is verified by comparing it with state-of-the-art models on the SatSOT dataset.

Key words: Siamese network; object tracking in satellite videos; spatial information; temporal information; motion estimation

1 引言

视觉目标跟踪是计算机视觉领域重要的基础性研究问题之一, 目标跟踪是指在给定第 1 帧的目标位置后, 在视频的后续帧中自动地给出目标的位置和形状. 随着卫星视频成像技术的显著进步, 遥感卫星作为观测工具已经迅速在各个行业得到应用. 通过目标跟踪获得卫星视频中运动物体的连续观测位置对于民事和军用领域具有重要意义. 因此, 卫星视频中的目标跟踪成为一个新的研究领域, 近年来受到越来越多的关注.

根据工作方式的不同, 目前目标跟踪算法主要分为基于相关滤波的跟踪算法和基于深度学习的跟踪算法.

基于相关滤波的跟踪算法 (例如 KCF[1]等) 通过循环平移操作扩增训练样本, 增加了训练样本的多样性从而提升了算法的鲁棒性. 但是, 由于算法的特征表示能力较弱, 跟踪过程中很容易出现目标丢失的问题.

深度学习技术的发展极大地提高了算法的特征表示能力, 基于深度学习的跟踪算法可以分为基于 Transformer[2]和基于孪生网络的跟踪算法. 基于 Transformer 的跟踪算法 (例如 STARK[3], TransT[4]等) 可以有效融合图像信息, 但是需要耗费大量的计算资源, 这限制了它的应用场景. 基于孪生网络的跟踪算法 (例如 SiamRPN[5], SiamBAN[6]等) 通过计算目标图像和被搜索图像之间的视觉相似度, 将目标跟踪问题转化为匹配问题, 这种方法降低了目标跟踪的难度, 提升了跟踪速度. 基于孪生网络的跟踪算法具有结构简单, 速度快, 精度高的特点, 因此, 是近年来目标跟踪领域的主流方法.

卫星视频具有目标特征不明显、背景杂乱、遮挡严重的特点. 为了更好地突出目标、抑制背景, Wan 等人[7]基于 Transformer 蒸馏技术, 提出了一种端到端的联合检测与跟踪方法, 旨在通过像素级自适应特征增强技术提高卫星视频中目标跟踪的精度和效率. 但引入的像素级 Transformer 特征蒸馏模块增加了计算复杂度. 在处理高分辨率卫星视频时, 计算量和内存需求会显著增加. Liang 等人[8]通过三分支注意力模型充分挖掘卫星视频中目标的空间特征, 但这种方法没有使用卫星视频帧之间的时间信息, 目标跟踪精度较低. Li 等人[9]在利用目标空间特征的基础上引入视频帧的时域信息增强目标表示, 有效缓解了相似干扰物的影响. 但是该方法中的全局注意力机制会使模型更加关注占比较大的背景部分从而忽略卫星视频中的小目标. Kong 等人[10]提出的 LocaLock 模型通过 Conv-LSTM 单元处理当前帧和参考帧的特征, 进而捕获时间信息, 增强了小目标在复杂背景中的区分能力. 但对于快速移动的目标, 由于时间信息积累不足该方法无法准确捕捉目标特征. Zhu 等人[11]提出一种结合 CNN 和 Transformer 的双边网络架构, 并通过目标感知 (target-aware) 策略和像素级细化模块来提高跟踪精度, 但是在遮挡场景中目标模板的特征与搜索区域的特征匹配会受到干扰, 导致该方法表现不佳. 为了缓解遮挡影响, Lin 等人[12]提出的 MACF 方法通过结合运动感知模块和空间-时间正则化相关滤波器, 显著提高了卫星视频中目标跟踪的精度和鲁棒性. 但该方法主要依赖于相关滤波器和运动信息, 不能很好地适应目标形变和尺度变化. Nie 等人[13]通过时间运动补偿 (TMComp) 机制补偿空间维度上的模糊对象路径, 以减轻遮挡带来的影响. 这种方法通过处理目标轨迹建模目标的运动特征, 在目标被遮挡时仍然可以实现较准确的跟踪. 但是, 当背景杂乱, 目标和背景区分度不明显时, 算法的跟踪精度较低.

针对上述问题, 本文在无锚框孪生网络 SiamCAR[14] 的基础上提出了时空孪生网络模型 Siam-STM.

这项工作的主要贡献如下.

(1) 提出了基于注意力机制的空间信息感知模块, 聚合图像中的上下文信息并增强卫星视频中小目标特征的辨别力, 提高模型性能.

(2) 提出了时间信息感知模块对视频中当前帧和历史帧进行融合, 从而学习到不同时刻目标的位置信息, 更好的关注目标轨迹, 缓解相似干扰物的影响.

(3) 提出了运动估计机制对目标的运动特征进行建模, 从而在目标被遮挡时预测目标位置.

2 时空孪生网络

本文提出了一种用于卫星视频目标跟踪的时空孪生网络, 总体框架如图 1 所示.

时空孪生网络中的空间信息感知模块对经过残差网络的搜索帧特征进行调制, 通过注意力机制提取视频帧图像的目标信息, 自适应的调节对目标和背景的关注度. 经过空间信息感知模块调制的特征和经过残差网络的模板帧特征在进行互相关操作后进入时间信息感知模块. 时间信息感知模块会通过融合当前帧和上一帧图像, 使网络模型更好的学习到不同时刻的目标信息, 从而关注目标轨迹. 时间信息感知模块输出的特征经过分类子网和回归子网得到时空孪生网络输出的结果. 为了进一步缓解遮挡影响, 运动估计机制模块对网络输出的结果进一步调整.

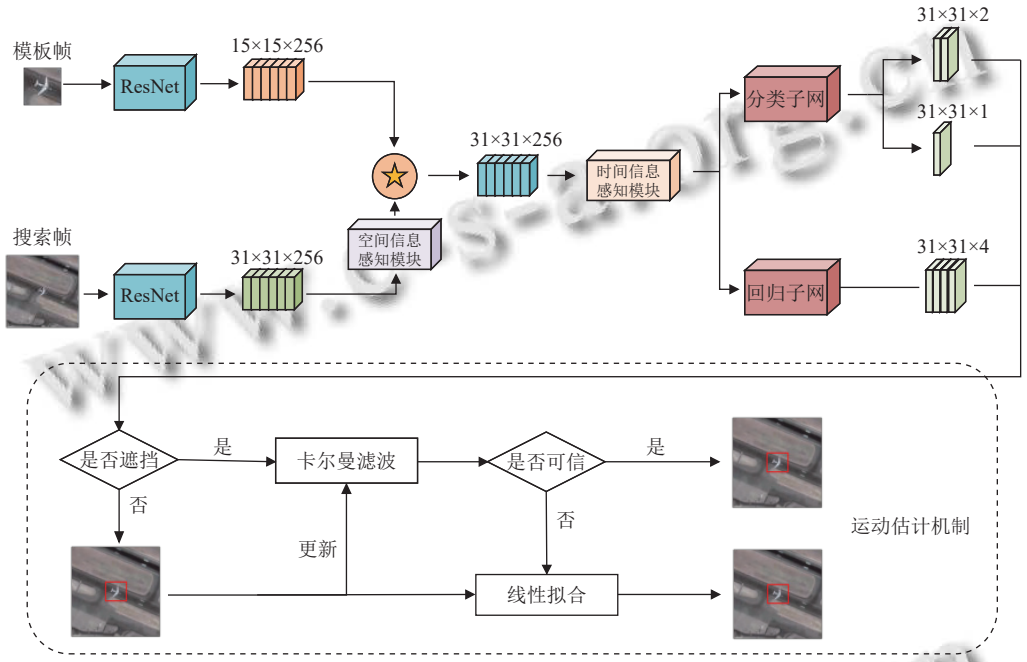


图 1 时空孪生网络结构图

2.1 时空孪生网络

2.1.1 空间信息感知模块

在处理图像或视频等视觉数据时, 模型通常需要关注图像中的关键区域以提取有用的信息. 空间注意力机制正是为了实现这一目标而设计的, 它对空间域信息进行相应的空间变换, 识别出更值得关注的核心区域, 有助于更高效地提取关键信息.

非局部空间注意力 Non-Local[15]通过计算任意两个位置之间的交互直接捕获长期依赖关系. Non-Local 可以有效地聚合图像中的上下文信息, 但是在卫星视频帧中该注意力会使模型更加关注占比较大的背景部分从而忽略被跟踪目标, 造成目标漂移. 其详细结构如图 2 所示.

为了使模型更加关注目标区域, 本文使用卷积操作和池化操作提取目标特征, 进而设计了基于空间注意力的空间信息感知模块. 该模块不仅通过卷积操作有效聚合视频帧的上下文信息, 而且通过最大池化操作提取具有代表性的目标特征, 有助于模型更好的定位目标位置.

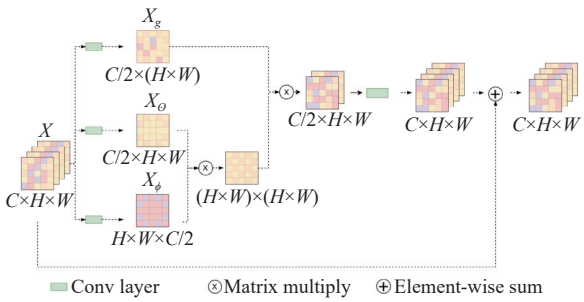


图 2 非局部空间注意力网络结构图

空间信息感知模块的计算过程如图 3 所示, 输入搜索特征图$X\in\mathbb{R}^{C\times H\times W}$, 在$X$的通道维度上应用最大池化得到$X_{\max}\in\mathbb{R}^{1\times N}$和$X_{\max}^{\mathrm{T}}\in\mathbb{R}^{N\times 1}$, 其中$N=H\times W$, 上标T表示转置. 对$X$应用卷积操作得到$X_{\mathrm{conv}}\in\mathbb{R}^{1\times N}$和

$X_{\text{conv}}^{\text{T}} \in \mathbb{R}^{N\times 1}$. 对$X_{\max}$和$X_{\max}^{\text{T}}$做矩阵乘法得到池化分支的特征图$P \in \mathbb{R}^{N\times N}$, 对$X_{\text{conv}}$和$X_{\text{conv}}^{\text{T}}$做矩阵乘法得到卷积分支的特征图$M \in \mathbb{R}^{N\times N}$, 通过级联操作将两分支特征聚合得到$G \in \mathbb{R}^{2N\times N}$. 之后对$G$做卷积和 *Softmax* 操作得到$A \in \mathbb{R}^{N\times N}$, 过程如下:

$$A = Softmax(Conv(G)) \tag{1}$$

其中, *Conv*是卷积操作.

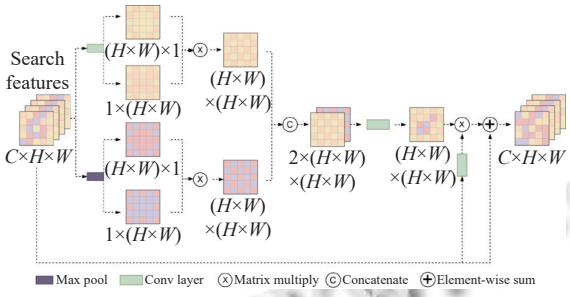


图 3 空间信息感知模块结构图

对输入搜索特征图X进行卷积操作得到$N \in \mathbb{R}^{C\times N}$, 最后通过以下操作得到最终的空间加权特征$X_{\text{sca}} \in \mathbb{R}^{C\times H\times W}$:

$$X_{\text{sca}} = Re(\lambda \cdot NA + X) \tag{2}$$

其中, *Re*是 reshape 操作, λ是一个可学习的标量参数.

2.1.2 时间信息感知模块

空间信息感知模块可以捕获卫星视频单独帧中全局空间信息的相互依赖性, 还可以突出卫星视频中小目标的关键信息, 这有助于模型从整体上区分目标和背景, 进一步提高目标跟踪的准确率. 但需要注意的是, 卫星视频是由一系列连续的帧组成的, 因此, 视频帧之间的时间信息对于目标跟踪是非常重要的.

本文基于 MetaFormer[16]提出了时间信息感知模块, 其结构如图 4 所示, 当前帧的互相关响应图F_t和上一帧的时间信息特征图F'_{t-1}作为输入, 通过通道注意力机制进行融合. 为了更好地利用当前帧的互相关响应图F_t, 本文将F_t和加权后的通道注意力特征图进行逐元素相加操作得到融合特征图F_{mix}. 计算公式如下:

$$F_{\text{mix}} = F_t + F'_{t-1} Softmax_{\text{row}}(F_t F_t^{\text{T}}) \tag{3}$$

然后F_{mix}经过卷积层的自适应学习得到经过调制和增强后的特征图F'_{mix}, 计算公式如下:

$$F'_{\text{mix}} = Conv(Norm(F_{\text{mix}})) \tag{4}$$

由于卫星图像中会遇到由运动模糊或遮挡引起的无效的上下文信息, 如果直接输出F'_{mix}而不进行任何过滤操作, F'_{mix}可能会包括一些不重要的信息. 为了消除这些信息, 本文设计了时间信息过滤分支对F_{mix}应用卷积操作后通过一个前馈网络 FFN 得到时间信息过滤参数$\alpha \in \mathbb{R}^{C\times 1\times 1}$, 过滤后的当前帧的时间信息特征图$F'_t$由以下公式计算得到.

$$F'_t = \alpha \cdot F'_{\text{mix}} + F_t \tag{5}$$

特征图F'_t融合了当前帧和上一帧之间的时间信息, 使网络可以端到端的融合帧之间的时间信息, 可以更好地聚焦于目标的轨迹, 降低相似干扰物对于目标跟踪的影响.

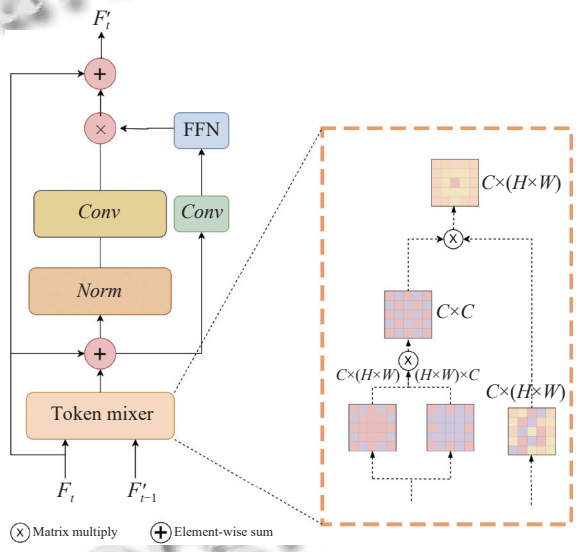


图 4 时间信息感知模块结构图

2.2 运动估计机制模块

时间信息感知模块对当前帧和历史帧进行了融合, 可以学习目标在不同时刻的位置, 降低相似干扰物的影响. 但是在遮挡场景中, 单纯利用空间信息感知模块和时间信息感知模块对图像特征进行处理很难准确跟踪目标. 因此, 模型还需要建模目标的运动特征从而在目标被丢失时预测目标位置.

为了建模目标的运动特征, 王丽黎等人[17]利用卡尔曼滤波建模运动信息缓解遮挡问题, 卡尔曼滤波是一种最小化均方误差的线性滤波方法, 它能够从一系列的不完全及包含噪声的测量中, 估计被跟踪目标的位置. 但卡尔曼滤波在跟踪刚开始时并不稳定, 目标定位不准确. 因此, 本文在卡尔曼滤波器的基础上引入线性拟合预测器, 提出了一种运动估计机制, 实现更稳定

的目标定位. 线性拟合预测器通过计算目标在之前卫星视频帧中移动的距离, 预测下一帧目标的位置P_{line} (x_{line}, y_{line}, w_{line}, h_{line}), 并且线性拟合预测器会始终保留最近 3 次被跟踪目标的最终输出位置P_1 (x_1, y_1, w_1, h_1)、P_2、P_3, 其中P_3是距离当前时刻最近的一次目标位置.

在线跟踪时, 时空孪生网络会输出目标的位置P_{siam} (x_{siam}, y_{siam}, w_{siam}, h_{siam}), 在目标被遮挡前, 当卡尔曼滤波器预测的位置P_{kf} (x_{kf}, y_{kf}, w_{kf}, h_{kf})连续 3 次和时空孪生网络预测位置P_{siam}的欧氏距离小于 5 时认为卡尔曼滤波器是稳定的. 欧氏距离计算公式如下:

$$d = \sqrt{(x_{\text{siam}} - x_{kf})^2 + (y_{\text{siam}} - y_{kf})^2} \tag{6}$$

时空孪生网络会输出图像中每个位置的分类得分. 当正样本的最高得分$V_{\max}$>0.85 时, 我们认为没有发生遮挡现象, 最终输出的目标位置$P_{\text{final}} = P_{\text{siam}}$, 并使用下列公式更新卡尔曼滤波器的参数.

$$K_t = P_t^- H^{\mathrm{T}} (H P_t^- H^{\mathrm{T}} + R_t)^{-1} \tag{7}$$

$$\hat{x}_t = \hat{x}_t^- + K_t (P_{\text{final}} - H_k \hat{x}_t^-) \tag{8}$$

$$P_t = P_t^- - K_t H_t P_t^- \tag{9}$$

其中, $\hat{x}_t^-$为目标位置的先验估计, P_t^-为误差协方差矩阵, K_t是卡尔曼滤波器的增益矩阵, $\hat{x}_t$为目标位置的后验估计, P_t是 t 时刻的误差协方差矩阵, H_t是系统在 t 时刻的观测矩阵. 更新完卡尔曼滤波器的参数之后还需要利用目标位置P_{final}更新线性拟合预测器的参数, 计算方式如下:

$$P_1 = P_2 \tag{10}$$

$$P_2 = P_3 \tag{11}$$

$$P_3 = P_{\text{final}} \tag{12}$$

当正样本的最高得分$V_{\max} \leqslant 0.85$时, 我们认为发生了遮挡现象, 此时需要判断卡尔曼滤波器的预测结果是否稳定. 当卡尔曼滤波器稳定时, $P_{\text{final}} = P_{kf}$, 卡尔曼滤波器预测计算公式如下:

$$\hat{x}_t^- = A_t \hat{x}_{t-1} \tag{13}$$

$$P_t^- = A_{t-1} P_{t-1} A_{t-1}^{\mathrm{T}} + Q_{t-1} \tag{14}$$

其中, Q_{t-1}是 t–1 时刻观测噪声的协方差矩阵, $\hat{x}_{t-1}$和P_{t-1}是 t–1 时刻的状态向量和协方差. $\hat{x}_t^-$ (x_{kf}, y_{kf}, w_{kf}, h_{kf}, Δw_{kf}, Δh_{kf})为预测结果. P_{kf} (x_{kf}, y_{kf}, w_{kf}, h_{kf})是卡尔曼滤波器预测的位置.

当卡尔曼滤波器不稳定时, $P_{\text{final}} = P_{\text{line}}$, 线性拟合预测器目标位置计算公式如下:

$$x_{\text{line}} = x_3 + \frac{x_1 + x_2 + x_3}{3} \tag{15}$$

$$y_{\text{line}} = y_3 + \frac{y_1 + y_2 + y_3}{3} \tag{16}$$

$$w_{\text{line}} = w_3 \tag{17}$$

$$h_{\text{line}} = h_3 \tag{18}$$

本文通过卡尔曼滤波器和线性拟合方法建模目标的运动特征, 通过目标在之前帧的位置预测被遮挡时目标的大概位置, 从而当目标重新出现时, 跟踪器可以准确地定位到目标.

3 实验分析

3.1 实验细节

本文实验环节使用基于 Linux 系统的硬件平台, 工作站配置为 30 GB 内存, GPU 为 NVIDIA A10. 采用动量为 0.9 的随机梯度下降 (SGD) 训练整个网络模型, 将权重衰减设置为 0.000 1, 学习率从 0.005 指数衰减到 0.000 5, 模型训练 20 个轮次, 批量大小为 32.

3.2 数据集与评价指标

本文使用 SatSOT[18]数据集评估算法的效果. SatSOT 有 105 个卫星视频序列, 包括了汽车、飞机、船和火车 4 种目标对象, 平均视频长度为 263 帧, 而且 SatSOT 包含卫星视频目标跟踪中的重大挑战, 如小目标、复杂背景和严重遮挡.

SatSOT 数据集提供了精度和成功率两个评价指标, 其中成功率是重叠分数大于给定阈值的成功跟踪帧的百分比. 精度是跟踪结果的中心与真实值之间的距离小于给定阈值的帧数的百分比, 该指标可以直观反映目标定位的准确性, 但是却并没有考虑到目标尺度的影响. 为了更全面的评价跟踪器的性能, 本文还引入了归一化精度, 这是一种用于评估目标跟踪算法性能的改进指标, 旨在消除目标尺寸对跟踪精度的影响, 从而提供更加公平和标准化的评估结果. 与传统的精度不同, 归一化精度将中心位置误差 (即预测边界框中心与真实边界框中心之间的欧氏距离) 除以目标真实框的对角线长度, 以此将误差值归一化到相对尺度范围内. 归一化精度p_{norm}计算公式如下:

$$p_{\text{norm}} = \frac{p}{\sqrt{w_{\text{gt}}^2 + h_{\text{gt}}^2}} \tag{19}$$

其中, p是中心位置误差, w_{gt}和h_{gt}分别表示真实边界框的宽度和高度.

3.3 对比实验

为了比较本文所提算法和其他算法的效果, 本文选择了当前具有代表性的优秀算法: 基于相关滤波的算法 KCF、CFME[19], 其中 CFME 是针对卫星视频设计的跟踪器; 基于 Transformer 的算法 TransT、STARK、ODTrack[20]; 基于孪生网络的算法 SiamRPN、SiamBAN、SiamCAR、HIPtrack[21].

跟踪器性能比较结果如表 1 所示, 本文提出的 Siam-STM 算法在成功率、精度和归一化精度上分别达到 41%、65.8% 和 53.2%, 相比于基线模型 SiamCAR 分别提高了 2.9%、4.5% 和 3.2%. 说明本文的改进大大提高了跟踪的效果. Siam-STM 在精度和归一化精度两个指标上都超过了其他算法, 这是因为 Siam-STM 有效地提取了目标的空间信息、时间信息和运动信息, 相比于只利用部分信息的其他跟踪器, Siam-STM 增强了卫星视频中目标信息的丰富性, 提高了跟踪的鲁棒性, 因此可以更好的预测目标位置. Siam-STM 的成功率和 CFME 具有相当的效果, 只相差了 0.2%. CFME 是基于相关滤波的卫星视频跟踪算法, 特征表示能力较弱, 适合跟踪目标尺度不发生明显变化的目标, 而 SatSOT 数据集中大部分目标在跟踪过程中尺度变化不明显, 这导致 CFME 的成功率略高. 本文也对 SatSOT 数据集中尺度发生较大变化 (ARC) 的目标进行跟踪效果分析, 结果如图 5 所示, 本文提出的 Siam-STM 在对尺度发生较大变化的目标进行跟踪时, 成功率要比 CFME 更高, 这也体现出了 Siam-STM 的独特优势.

表 1 各个跟踪器在 SatSOT 上的性能对比

Tracker	Success rate↑	Precision↑	Norm precision↑
SiamRPN	0.273	0.506	0.321
SiamBAN	0.287	0.538	0.367
SiamCAR	0.381	0.613	0.500
HIPtrack	0.362	0.545	0.465
TransT	0.380	0.584	0.478
STARK	0.261	0.418	0.321
ODTrack_B	0.328	0.538	0.414
ODTrack_L	0.354	0.586	0.451
KCF	0.385	0.608	0.421
CFME	**0.412**	0.612	0.487
Ours	0.410	**0.658**	**0.532**

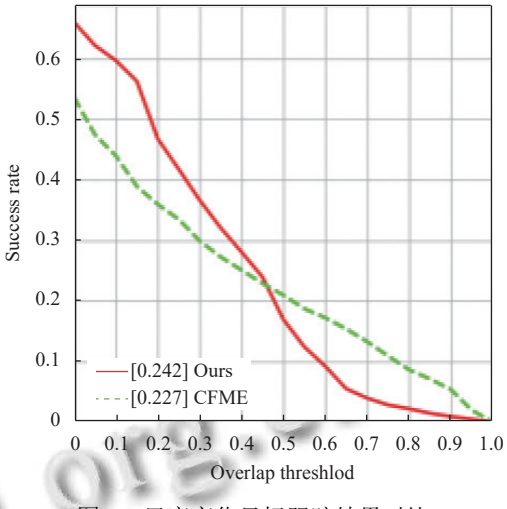


图 5 尺度变化目标跟踪结果对比

为了更加直观地展现出本文算法的优越性, 本文在图 6 中展示了效果最好的 4 个跟踪器在 SatSOT 几个典型帧上的实验结果.

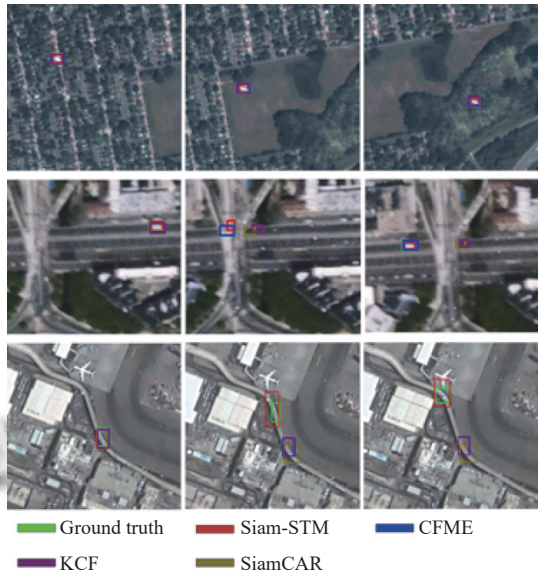


图 6 追踪效果对比图

图 6 中的第 1 行表明, 当物体稳定运动、目标特征较明显时, 所有跟踪器都取得较好结果. 然而, 在第 2 行目标被完全遮挡后重新出现时, KCF 和 SiamCAR 由于只利用了目标的外观信息, 目标被遮挡后外观信息丢失, 因此跟踪失败. CFME 在利用目标外观信息的基础上还提取了目标的运动信息, 因此成功跟踪了目标. 本文提出的算法 Siam-STM 通过目标运动估计机制建模目标的运动信息, 在目标被遮挡时预测目标大概位置, 进而在目标重新出现后准确定位目标. 在第 3 行目标特征不明显, 背景杂乱, 而且跟踪过程中目标发

生较大尺度变化的多挑战场景中. 只利用目标外观信息的 KCF 和 SiamCAR 鲁棒性较差, 因此跟踪失败. CFME 特征提取能力较弱, 对于尺度发生较大变化的目标敏感度较低, 所以跟踪失败. 而本文提出的算法 Siam-STM 通过结合目标的时空信息以及运动信息增强了目标在多挑战场景中的鲁棒性, 而且能够较好地跟踪尺度变化较大的目标. 这表明 Siam-STM 在卫星视频目标跟踪领域具有良好的竞争力.

3.4 消融实验

为了进一步验证所提模块的有效性, 本文在 Sat-SOT 数据集上进行了一系列消融实验.

表 2 详细记录了各个模块组合对于模型性能的影响. Base 是孪生网络基线模型 SiamCAR, SIP 表示本文提出的空间信息感知模块, TIP 表示本文提出的时间信息感知模块, ME 表示本文提出的运动估计机制. 从表 2 中可以看出, 与仅使用基线模型相比, 空间信息感知模块 (SIP) 的加入将成功率、精度和归一化精度分别提高了 1%、2.8% 和 2.2%. 相对于成功率, 精度和归一化精度提升效果较明显, 这得益于空间信息感知模块可以有效增强小目标的辨别力, 能够较准确的跟踪目标位置, 但是在有相似干扰物影响或者遮挡场景中, 模型的鲁棒性较差, 因此成功率提升效果不明显.

表 2 消融实验结果表

Tracker	Success rate↑	Precision↑	Norm precision↑
Base	0.381	0.613	0.500
Base+SIP	0.391	0.641	0.522
Base+TIP	0.393	0.619	0.509
Base+ME	0.391	0.625	0.509
Base+SIP+ME	0.403	0.650	0.525
Base+SIP+TIP	0.395	0.646	0.521
Base+TIP+ME	0.397	0.642	0.520
Base+SIP+TIP+ME	**0.410**	**0.658**	**0.532**

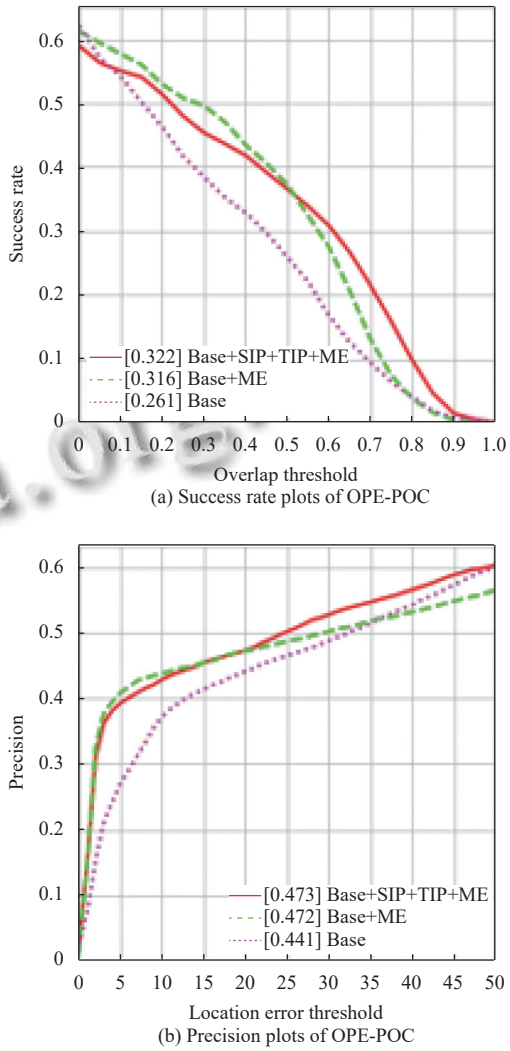
在基线模型中加入时间信息感知模块 (TIP) 可以提取视频帧中目标的时间信息, 从而缓解相似干扰物的影响, 模型的鲁棒性得到提升, 成功率提升较大, 提高了 1.2%. 精度和归一化精度也分别提高了 0.6% 和 0.9%.

空间信息感知模块和时间信息感知模块的组合使得模型的成功率提高了 1.4%, 精度和归一化精度分别提高了 3.3% 和 2.1%. 模型的各个评价指标都得到了较大的提升, 这得益于时空互补信息的有效利用, 既增强了目标的辨别力, 又缓解了相似干扰物的影响.

运动估计机制的加入使得基线模型的成功率、精度和归一化精度分别提升了 1%、1.2% 和 0.9%. 而且在任何组合中加入运动估计机制都可以提升跟踪效果. 这是因为运动估计机制通过建模目标的运动信息使得模型可以在遮挡场景下准确跟踪目标.

由表 2 中数据可知, 在基线模型中加入时间、空间信息感知模块和运动估计机制的 Siam-STM 相比于基线模型在成功率、精度和归一化精度上提升了 2.9%、4.5% 和 3.2%, 效果是最好的, 这证明了本文提出模块的有效性, 而且说明本文提出的模块能够有效结合起来. 使模型既增强了对于卫星视频中小目标的辨别力, 又增强了在相似干扰物影响和遮挡场景下目标跟踪的鲁棒性.

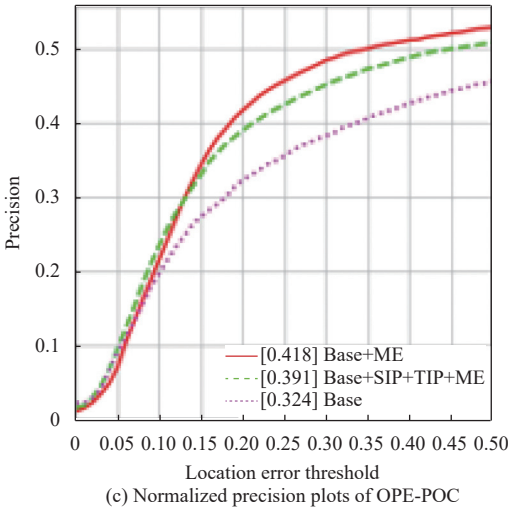
图 7 对比了 3 种模型 (Base、Base+ME 和 Base+SIP+TIP+ME) 在遮挡数据集中的跟踪效果.



(a) Success rate plots of OPE-POC

(b) Precision plots of OPE-POC

图 7 遮挡目标跟踪结果对比



(c) Normalized precision plots of OPE-POC

图 7 遮挡目标跟踪结果对比 (续)

由图 7 对比得出, 在遮挡场景中, 运动估计机制的加入可以使基线模型的成功率、精度和归一化精度都得到提升, 证明该机制可以使模型对目标的运动特征进行有效建模, 从而在目标被遮挡时更准确地预测目标位置. 而且从图中可以发现时间、空间信息感知模块以及运动估计机制的组合通过结合目标的时空信息和运动信息可以更好地缓解遮挡影响.

图 8 选取了 3 个测试样本并绘制了网络处理结果的热力图. 第 1 列表明当目标较明显时, 所有组合都取得了相对较好的效果. 但是在第 2 列图像背景复杂并且目标特征不明显时, 基线模型很容易混淆目标和背景, 而空间信息感知模块通过聚合图像中的上下文信息并增强卫星视频中小目标特征的辨别力, 进而增强了目标特征并有效地抑制了背景信息. 在第 3 列图像中目标周围存在相似干扰物时, 时间信息感知模块通过对视频中当前帧和历史帧进行融合, 学习到不同时刻目标的位置信息, 使模型更好地关注卫星视频中的目标轨迹, 有效地缓解了相似干扰物的影响. 时间、空间信息感知模块的组合使模型能够充分利用目标的互补的空间信息和时间信息, 既抑制了背景增强了小目标的辨别力, 又缓解了相似干扰物的影响.

4 结论

本文提出了一种用于卫星视频目标跟踪的时空孪生网络, 该方法主要包括空间信息感知模块、时间信息感知模块和运动估计机制. 空间信息感知模块使用卷积操作聚合目标上下文信息, 并通过池化操作来增强卫星视频中小目标特征的辨别力. 而时间信息感知模块在 MetaFormer 的基础上, 设计了时间信息过滤分支筛除掉卫星视频中的无效信息, 并且引入了线性映射注意力机制融合当前帧和历史帧, 从而学习到不同时刻目标的位置信息, 更好地关注目标轨迹, 缓解相似干扰物的影响. 最后, 本文在卡尔曼滤波器的基础上引入了线性拟合方法, 提出了一种运动估计机制, 可以有效地建模目标的运动特征进而在目标被遮挡时预测目标位置. 通过实验比较证明了时空孪生网络在卫星视频目标跟踪领域中的有效性.

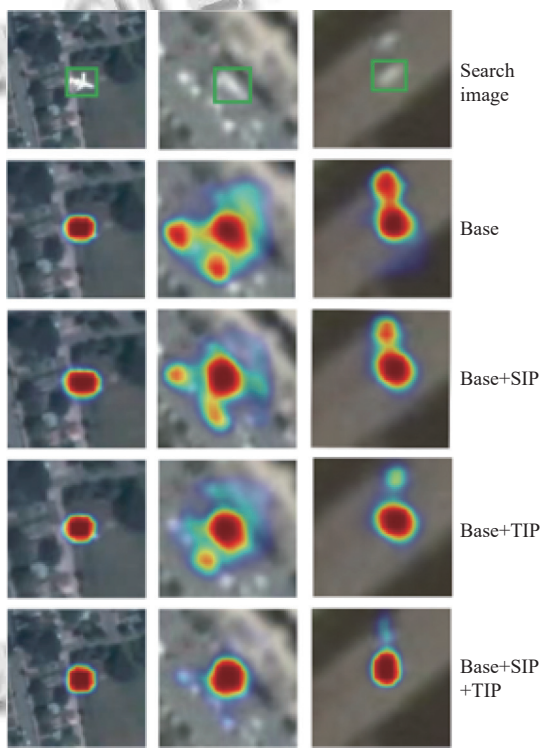


图 8 网络处理结果热力图

参考文献

1 Henriques JF, Caseiro R, Martins P, *et al*. High-speed tracking with kernelized correlation filters. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2015, 37(3): 583–596. [doi: 10.1109/TPAMI.2014.2345390]

2 Vaswani A, Shazeer N, Parmar N, *et al*. Attention is all you need. Proceedings of the 31st International Conference on Neural Information Processing Systems. Long Beach: Curran Associates Inc., 2017. 6000–6010.

3 Yan B, Peng HW, Fu JL, *et al*. Learning spatio-temporal Transformer for visual tracking. Proceedings of the 2021

IEEE/CVF International Conference on Computer Vision (ICCV). Montreal: IEEE, 2021. 10428–10437.

4 Kim M, Lee S, Ok J, *et al*. Towards sequence-level training for visual tracking. Proceedings of the 17th European Conference on Computer Vision. Tel Aviv: Springer, 2022. 534–551. [doi: 10.1007/978-3-031-20047-2_31]

5 Li B, Yan JJ, Wu W, *et al*. High performance visual tracking with Siamese region proposal network. Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018. 8971–8980. [doi: 10.1109/CVPR.2018.00935]

6 Chen ZD, Zhong BN, Li GR, *et al*. Siamese box adaptive network for visual tracking. Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle: IEEE, 2020. 6667–6676. [doi: 10.1109/CVPR42600.2020.00670]

7 Wan G, Su ZJ, Wu YT, *et al*. High-precision multi-object tracking in satellite videos via pixel-wise adaptive feature enhancement. Sensors, 2024, 24(19): 6489. [doi: 10.3390/s24196489]

8 Liang JH, Sun JY, Zhang XM, *et al*. Lightweight tracking of satellite video object based on saliency enhancement mechanism. IEEE Journal on Miniaturization for Air and Space Systems, 2023, 4(2): 100–104. [doi: 10.1109/JMASS.2023.3234099]

9 Li CH, Zhang JP, Wang YH. SMTN: Multidimensional fusion and time-domain coding for object tracking in satellite videos. IEEE Geoscience and Remote Sensing Letters, 2023, 20: 6010805.

10 Kong LY, Yan ZY, Shi HR, *et al*. LocaLock: Enhancing multi-object tracking in satellite videos via local feature matching. Remote Sensing, 2025, 17(3): 371. [doi: 10.3390/rs17030371]

11 Zhu QQ, Huang X, Guan QF. TabCtNet: Target-aware bilateral CNN-Transformer network for single object tracking in satellite videos. International Journal of Applied Earth Observation and Geoinformation, 2024, 128: 103723. [doi: 10.1016/j.jag.2024.103723]

12 Lin B, Zheng JL, Xue CC, *et al*. Motion-aware correlation filter-based object tracking in satellite videos. IEEE Transactions on Geoscience and Remote Sensing, 2024, 62: 5604313.

13 Nie YD, Bian CJ, Li LG. Object tracking in satellite videos based on Siamese network with multidimensional information-aware and temporal motion compensation. IEEE Geoscience and Remote Sensing Letters, 2022, 19: 6517005.

14 Guo DY, Wang J, Cui Y, *et al*. SiamCAR: Siamese fully convolutional classification and regression for visual tracking. Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle: IEEE, 2020: 6268–6276.

15 Wang XL, Girshick R, Gupta A, *et al*. Non-local neural networks. Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018: 7794–7803.

16 Yu WH, Luo M, Zhou P, *et al*. MetaFormer is actually what you need for vision. Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans: IEEE, 2022: 10809–10819.

17 王丽黎, 张慧. 多信息融合的卫星视频单目标跟踪. 计算机系统应用, 2023, 32(2): 266–273. [doi: 10.15888/j.cnki.csa.008995]

18 Zhao MQ, Li SY, Xuan SY, *et al*. SatSOT: A benchmark dataset for satellite video single object tracking. IEEE Transactions on Geoscience and Remote Sensing, 2022, 60: 5617611.

19 Xuan SY, Li SY, Han MF, *et al*. Object tracking in satellite videos by improved correlation filters with motion estimations. IEEE Transactions on Geoscience and Remote Sensing, 2020, 58(2): 1074–1086. [doi: 10.1109/TGRS.2019.2943366]

20 Zheng YZ, Zhong BN, Liang QH, *et al*. ODTrack: Online dense temporal token learning for visual tracking. Proceedings of the 38th AAAI Conference on Artificial Intelligence. Vancouver: AAAI Press, 2024. 7588–7596.

21 Cai WR, Liu QJ, Wang YH. HIPTrack: Visual tracking with historical prompts. Proceedings of the 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle: IEEE, 2024: 19258–19267.

(校对责编: 张重毅)