

中图法分类号: TP391 文献标识码: A 文章编号: 1006-8961(2023)10-3176-15

论文引用格式: Zhao J, Yuan Y S, Zhang P Y and Wang D. 2023. An efficient Transformer-based object-capturing video annotation method. Journal of Image and Graphics, 28(10):3176-3190(赵洁, 袁永胜, 张鹏宇, 王栋. 2023. 轻量化Transformer目标跟踪数据标注算法. 中国图象图形学报, 28(10):3176-3190)[DOI:10.11834/jig.220823]

轻量化Transformer目标跟踪数据标注算法

赵洁, 袁永胜, 张鹏宇, 王栋*

大连理工大学信息与通信工程学院, 大连 116024

摘要: 目的 基于深度模型的跟踪算法往往需要大规模的高质量标注训练数据集,而人工逐帧标注视频数据会耗费大量的人力及时间成本。本文提出一个基于Transformer模型的轻量化视频标注算法(Transformer-based label network, TLNet),实现对大规模稀疏标注视频数据集的高效逐帧标注。**方法** 该算法通过Transformer模型来处理时序的目标外观和运动信息,并融合前反向的跟踪结果。其中质量评估子网络用于筛选跟踪失败帧,进行人工标注;回归子网络则对剩余帧的初始标注进行优化,输出更精确的目标框标注。该算法具有强泛化性,能够与具体跟踪算法解耦,应用现有的任意轻量化跟踪算法,实现高效的视频自动标注。**结果** 在2个大规模跟踪数据集上生成标注。对于LaSOT(large-scale single object tracking)数据集,自动标注过程仅需约43 h,与真实标注的平均重叠率(mean intersection over union, mIoU)由0.824提升至0.871。对于TrackingNet数据集,本文使用自动标注重新训练3种跟踪算法,并在3个数据集上测试跟踪性能,使用本文标注训练的模型在跟踪性能上超过使用TrackingNet原始标注训练的模型。**结论** 本文算法TLNet能够挖掘时序的目标外观和运动信息,对前反向跟踪结果进行帧级的质量评估并进一步优化目标框。该方法与具体跟踪算法解耦,具有强泛化性,并能节省超过90%的人工标注成本,高效地生成高质量的视频标注。

关键词: 视频目标标注;单目标视觉跟踪;Transformer模型;互相关操作;时序信息融合

An efficient Transformer-based object-capturing video annotation method

Zhao Jie, Yuan Yongsheng, Zhang Pengyu, Wang Dong*

School of Information and Communication Engineering, Dalian University of Technology, Dalian 116024, China

Abstract: Objective High performance-oriented robust tracking methods are often beneficial from its related depth model nowadays, which is deeply linked to the large-scale high quality annotations-related video datasets in the training phase. However, frame-manual annotating videos frame is labor-intensive and costly. In addition, existing video annotation methods are usually focused on interpolation operation relevant to sparse labeled datasets. An effective interpolation operation is mainly designed based on the geometric information or other related tracking methods. One potential of these methods is illustrated that they do not require any quality evaluation mechanism to filter out the noisy annotations, resulting in the annotations unreliable. Additionally, some interactive annotation tools have a high involvement of labor, which makes the

收稿日期:2022-08-16;修回日期:2022-10-18;预印本日期:2022-10-25

* 通信作者:王栋 wdice@dlut.edu.cn

基金项目: 国家自然科学基金项目(62022021);装备预研教育部联合基金项目(8091B032155);辽宁省中央引导地方发展资金项目(2022JH6/100100026);中央高校基本科研业务费专项资金资助(DUT22QN228)

Supported by: National Natural Science Foundation of China (62022021); Joint Fund of Ministry of Education for Equipment Pre-research (8091B032155); Central Guidance on Local Science and Technology Development Fund of Liaoning Province (2022JH6/100100026); Fundamental Research Funds for the Central Universities (DUT22QN228)

annotating process more complex. An end-to-end deep model based video annotation via selection and refinement (VASR) can be used to generate reliable annotations via its selection and refinement. However, due to its training data consists of intermediate tracking results and object masks, it is highly associated with specific tracking methods. The process of generating annotations is complicated and time consuming. To meet the needs of visual tracking methods for large-scale annotated video datasets, simplify the annotation process, and optimize computing cost, we develop an efficient end-to-end model of Transformer-based label Net (TLNet) to automatically generate video annotations for sparse labeled tracking datasets further. Due to the Transformer model has its potentials to deal with sequential information, we melt it into our model to fuse the bidirectional and sequential input features. **Method** In this study, an efficient video annotation method is developed to generate reliable annotations quickly. The annotating strategy is implemented and it can be divided into three steps as follows: First, high-speed trackers are employed to perform forward and backward tracking. Second, these original annotations are evaluated in frame-level, and noisy frames are filtered out for manual annotation. Third, other related frames are optimized to generate more precise bounding boxes. To get its evaluation and optimization function mentioned above, a Transformer-based efficient model is illustrated, called TLNet. The proposed model consists of three sorts of modules, i. e., the feature extractor, feature fusion module, and prediction heads. To extract the vision features of the object, pixel-wise cross-correlation operation is used for the template and search region. A Transformer-based model is introduced to handle the sequential information after the vision and motion features are incorporated with, including the object appearance and motion cues. Furthermore, it is also used to fuse the bidirectional tracking results, i. e., forward tracking and backward tracking. The proposed model contains two sorts of sub-models in related to quality evaluation network and regression network. Among them, the former one is used to provide the confidence score of each frame's original annotation, and then the failed frames will be filtered out and sent back to manual annotators. The latter one is used to optimize the original annotations of the remaining frames, and more precise bounding boxes can be output. Our method has its strong generalizability to some extent since video frames and bidirectional tracking results are used as input for the model training only, and no intermediate tracking results like confidence scores, response maps are used. Specifically, our method is decoupled from specific tracking algorithms, and it can integrate any existing high-speed tracking algorithms to perform the forward and backward tracking. In this way, efficient and reliable video annotations can be achieved further. The completed annotating process is designed simplified and easy-to-use as well. **Result** Our method proposed is focused on generating annotations on two sort of large-scale tracking datasets like LaSOT, and TrackingNet. Two kind of evaluation protocols are set and matched for each of the two datasets. For LaSOT dataset, we apply the mean intersection over union (mIoU) to evaluate the annotations quality straightforward. The accuracy (Acc), recall, and true negative rate (TNR) are used to evaluate the filtering ability of the quality evaluation network as well. Compared to the VASR method costing about two weeks (336 hours totally), our method saves the time (43 hours totally) significantly in generating annotation. Specifically, 5.4% of frames are recognized as failed frames, where the Acc is 96.7%, and TNR is reached to 76.1%. After filtering out noisy frames and replacing with manual annotations, the mIoU of annotations are increased from 0.824 to 0.864. It is improved and reached to 0.871 further in terms of regression network. For the TrackingNet dataset, due to the noisy annotations are existed in its ground truth value, an indirect way is applied to evaluate the annotation quality. That is, we select out three sort of different tracking algorithms in related to ATOM, DiMP, and PrDiMP, and they are retrained using our annotations. The results are captured on three sort of tracking datasets relevant to LaSOT, TrackingNet, and GOT-10k, and it demonstrates that our annotations can be used to train a tracking model better than its original ground truth. It also can generate high quality annotations efficiently. Moreover, the ablation study also demonstrates the effectiveness of the Transformer model and its other related designs. **Conclusion** We develop an efficient video annotation method, which can mine the sequential object appearance and motion information, and fuse the forward and backward tracking results. The reliable annotations can be generated, and more than 90% labor cost for manually annotating can be saved on the basis of the frame-level quality evaluation operation originated from our quality evaluation network, and bounding boxes optimization derived from our regression network. Due to decoupling with the specific tracking algorithms, our method has its strong generalizability, for which it applies any existing high-speed tracking algorithms to achieve efficient annotation. We predict that our annotation method proposed can make the annotating process more reliable, faster and simpler to a certain extent.

Key words: video annotation; single object tracking; Transformer model; cross-correlation; sequential information fusion

0 引言

单目标视觉跟踪作为计算机视觉领域的基础研究之一,近年来已有了显著的进展(孟琰和杨旭,2019)。现有的各类型跟踪算法(王蒙蒙等,2022)已在各个跟踪数据集上展现出鲁棒的跟踪性能,例如基于孪生网络的生成式方法 SiamRPN++(Siamese region proposal network++) (Li 等,2019)、Ocean(object-aware anchor-free network)(Zhang 等,2020)、辨别式跟踪方法 BTCF(background-temporal-aware correlation filter)(朱建章等,2019)、ATOM(accurate tracking by overlap maximization)(Danelljan 等,2019)、DiMP(discriminative model prediction)(Bhat 等,2019),以及最新的基于 Transformer 模型的方法 TransT(Transformer tracking)(Chen 等,2021)、MixFormer(mixed Transformer)(Cui 等,2022)。这些跟踪算法多是基于深度模型,需要大规模的有精确目标框标注的视频数据用于训练以保证模型的高性能跟踪。然而,人工逐帧标注目标框会耗费大量的人力资源和时间成本,现有的可用于跟踪模型训练的大规模视频数据集仍然存在很大的缺口,这成为跟踪性能进一步提升的瓶颈之一。因此如何高效地生成高质量的大规模视频数据标注成为该领域亟需解决的问题。

现有的大规模跟踪数据集中,LaSOT(large-scale single object tracking)(Fan 等,2019)、GOT-10k(generic object tracking benchmark)(Huang 等,2021)和 TrackingNet(large-scale tracking dataset)(Müller 等,2018)最常用于跟踪算法中的模型训练。其中,LaSOT 和 GOT-10k 均使用逐帧人工标注。LaSOT 包含 1 400 个长序列,共 70 类,超过 3.5 M 视频帧中的目标框由人工标注。GOT-10k 包含 10 000 个短视频序列,共有超过 1.5 M 的人工标注目标框。而 TrackingNet 虽然包含超过 30 000 个序列,但其使用稀疏人工标注,即每秒标注一帧,其他帧的目标框通过跟踪器 STAPLE_{CA}(Mueller 等,2017)生成。由于没有跟踪结果的质量筛选机制,该数据集存在标注质量不可靠的局限性,一定程度上影响跟踪模型的训练性能。由于人工标注的昂贵成本,目前能够用于跟踪

模型训练的高质量大规模视频数据集仍不能满足鲁棒性跟踪算法的需求。

为减轻视频标注的人力和时间成本,目前已有若干方法尝试实现视频的自动目标框标注。现有方案的基本思路通常是先稀疏标注视频序列的若干关键帧,再通过插值的方式去自动补全其他帧的目标框标注。其中,插值方式主要分为 3 类。1)基于几何信息的线性插值,例如 LabelMe(Yuen 等,2009)。该类方法假定目标运动模式单一,仅依据目标的几何线索来获取其他帧的目标标注;2)基于视觉信息的复杂插值,例如,VATIC(video annotation tool from Irvine, California)(Vondrick 等,2010)通过提取目标实体的视觉特征,利用更复杂的动态插值方式生成目标标注;3)基于现有跟踪算法的插值,例如,TrackingNet 每隔 1 s 标注 1 帧,并利用跟踪器 STAPLE_{CA} 获得每一段子序列的前向反向跟踪结果,最终通过简单的基于时序的权重结合前向和反向的轨迹,获得最终的视频目标标注。此外,根据关键帧的获取方式,现有的视频标注方法可以分为两类,一类是采用固定的帧率进行人工稀疏标注,例如 LabelMe,TrackingNet 等;另一类则通过算法去预测序列的关键帧,从而达到更好的插值效果。例如,Vondrick 和 Ramanan(2011)提出了基于主动学习的视频标注方法,该方法通过迭代动态决定需要由人工标注的帧。而 Kuznetsova 等人(2021)设计了一个排序模型,预测剩余帧的重要程度,分数越高则表示标注该帧将会带来更高的插值性能增益。

尽管不同的插值方式和关键帧的捕获方案会在一定程度上提高视频标注的质量和效率,但上述方法均未考虑标注的校正问题。当面对复杂的跟踪场景时,例如目标运动模式复杂、存在干扰物、背景复杂或目标受到局部遮挡等,现有的跟踪器及其他插值方式均有可能导致标注结果不可靠。设计有效的标注质量评估模块,并对自动生成的标注进行人工校正,将进一步提升自动标注的准确性和可靠性。针对这一问题,VASR(video annotation via selection and refinement)(Dai 等,2021)提出了一个全新的基于选择和细化的自动标注流程,其中选择模块用于对前向和反向的跟踪结果进行质量评估,根据跟踪结果的分数选择最终的目标标注,并筛选出跟踪错

误的帧进行人工校正。细化模块则引入了几何参数预测模型来生成更准确的目标框标注。该方法虽然能够有效提升标注质量,但存在以下不足。首先,该方法的模型训练与执行前向反向跟踪的跟踪方法强绑定,并依赖现有的方法 Alpha-Refine (Yan 等, 2021b)生成目标分割掩码,标注生成过程复杂且没有通用性;其次,该工作选择基于在线更新的速度较低的跟踪算法 DiMP进行插值,并且需要额外生成目标分割掩码,导致生成标注的过程效率较低,对于 LaSOT 量级的数据集,需要约两周时间。

针对上述不足,本文提出一个泛化性强的轻量化视频自动标注方法,并引入 Transformer 模型融合基于视觉和运动的时序信息,实现对初始的前向反向跟踪结果的质量评估,并进一步优化目标框标注。本文提出的方法在训练阶段与具体的跟踪算法进行解耦,能够充分利用现有的轻量化跟踪器,例如 STARK-Lightning (spatial-temporal Transformer) (Yan 等, 2021a)、HCAT (hierarchical cross-attention Transformer) (Chen 等, 2022)等,实现简单且高效的视频自动标注。

本文主要贡献如下:1)设计了一个基于 Transformer 模型的轻量化视频标注模型 TLNet (Transformer-based label network),该模型包含质量评估子网络和回归子网络,通过 Transformer 模型来融合时序的视觉和目标运动信息,结合前反向的跟踪结果,对每一帧进行标注质量评估,并进一步优化标注目标框。该模型与具体跟踪算法解耦,具有强泛化性。2)提出一个简单高效的视频标注算法,能够应用现有的任意跟踪器,包括高速的轻量化跟踪器,对稀疏标注序列的其他帧进行插值标注,使用 TLNet 评估每一帧的标注质量,筛选出低质量帧进行人工标注,并自动优化剩余帧的目标框,保证了标注的质量和效率。3)在 LaSOT 和 TrackingNet 数据集上生成自动标注,验证了 TLNet 自动标注算法的有效性和高效性。与人工逐帧标注相比,TLNet 生成的自动标注在 LaSOT 数据集上节省超过 91% 的人力成本,并使最终标注的平均重叠率 (mean intersection over union, mIoU) 达到 0.871,完整的自动标注过程仅需约 43 h。4)在 TrackingNet 数据集上自动生成标注,并重新训练了 3 种跟踪算法,在 3 个跟踪数据集上进行性能测试与评估。实验表明,相较于 TrackingNet 的原始标注,使用本文自动生成的标注

能够训练出更加鲁棒的跟踪模型。

1 相关技术

1.1 Transformer 模型

Transformer 模型 (Vaswani 等, 2017) 由于其强大的捕获全局语义信息的能力,广泛应用于序列任务,例如自然语言处理 (Devlin 等, 2019)。近年来,Transformer 模型逐渐被挖掘出处理视觉信息的强大能力,取代卷积神经网络,应用于计算机视觉的各个领域,例如图像分类 (Dosovitskiy 等, 2021)、目标检测 (Carion 等, 2020)、视觉跟踪 (Chen 等, 2021)。该模型包含编码器和解码器,图 1 展示了编码器和解码器共有的基本单元结构,主要由多头注意力模块 (multi-head attention) 和前馈神经网络 (feed-forward network) 两个组件构成。每个组件的输出均采用了残差连接 (He 等, 2016) 和层标准化 (Ba 等, 2016) 来保证模型的收敛。多头注意力模块 (Attention) 的核心操作为

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (1)$$

式中,对于输入的维度为 d_k 的向量 Q 、 K 和 V 分布表示 query、key、value,计算 K 和 Q 的相似性权重并加权至 V 。前馈神经网络则由两层线性层和非线性激活函数 (rectified linear unit, ReLU) 构成。此外,引入位置编码用于捕捉序列的位置信息。

最常见的基于 Transformer 模型的图像处理方式是将图像分割为若干图像块并按照空间位置排序,

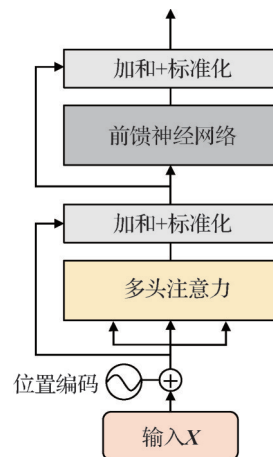


图1 Transformer模型基本单元结构

Fig. 1 Architecture of Transformer model

每一个图像块作为一个词向量(token)输入Transformer模型,通过融合各图像块提取图像特征,用于下游任务。与分割图像块方式不同的是,本文尝试充分利用Transformer模型处理序列信息的能力,将每一个视频帧的目标作为一个整体,结合其视觉特征和运动特征并作为一个词向量,引入Transformer模型来挖掘连续 T 帧的目标特征,学习目标的外观变化与运动模式。

1.2 像素级互相关操作

互相关操作(cross-correlation)广泛应用于视觉跟踪领域,尤其是基于孪生网络的跟踪算法,例如SiamFC (fully-convolutional Siamese network) (Bertinetto 等, 2016), SiamRPN (Siamese region proposal network) (Li 等, 2018), SiamFC++ (fully-convolutional Siamese network++) (Xu 等, 2020)等。对初始帧的模板特征 $Z \in \mathbb{R}^{C \times H_z \times W_z}$ 与当前帧的搜索区域特征 $X \in \mathbb{R}^{C \times H_s \times W_s}$ 执行互相关操作,度量两特征的相似性,并输出对应响应图,其中高响应值处表示搜索区域中的前景,反之则为背景。 C, H, W 分别表示对应特征向量的通道数、高度和宽度。在本文中,各帧目标的视觉信息将由各帧的搜索区域与模板图像的互相关响应图表示。之后该响应图与目标运动信息结合,共同表示单帧的目标特征。

受Alpha-Refine(Yan 等, 2021b)启发,为最大程度保留空间信息,实现后续精确的目标框预测,本文采用像素级互相关操作(pixel-wise correlation) (Wang 等, 2019)。首先将模板特征 Z 分解为 $H_z W_z$ 个特征向量 $Z_j \in \mathbb{R}^{C \times 1 \times 1}$,将其作为卷积核分别与搜索区域特征 X 执行互相关操作,最终获得响应图 $M \in \mathbb{R}^{H_s W_s \times H_z W_z}$,该过程可表示为

$$M = \left\{ M_j \mid M_j = Z_j \times X \right\}_{j \in \{1, \dots, H_z W_z\}} \quad (2)$$

2 视频自动标注算法

2.1 视频标注生成策略

现有的视频标注算法通常通过稀疏标注关键帧并插值的方式实现视频逐帧标注。由于缺少对标注质量的评估校正模块,生成的视频标注往往可靠性低。此外,通过与人工交互迭代的方式确认关键帧会使标注过程更加烦琐和冗余。为解决上述问题,受VASR方法启发,本文提出了一个简单高效的视

频标注算法,通过引入本文设计的端到端的基于Transformer的轻量化视频标注模型TLNet,对初始的目标标注进行质量评估和优化。相较于现有的视频标注算法,该算法主要有以下优势。首先,减少了标注过程中人工的参与,简化了标注流程;其次,强泛化性的轻量模型设计和轻量化跟踪器的选择,提升了视频标注的效率;此外,引入质量评估模块,保证了最终生成标注的可靠性。

本文提出的视频标注算法的具体流程图如图2所示。首先,以固定的帧率(本文采用30帧)进行人工稀疏标注。序列将被切分为若干片段,每一个片段的首尾帧包含人工精确标注;之后选择任意跟踪器对每一个片段执行前向和反向跟踪,生成前向和反向跟踪结果,并作为剩余未标注帧的初始标注;本文提出的TLNet将根据包含目标视觉和运动模式的序列信息,对每一帧的跟踪状态进行质量评估,根据质量分数筛选出跟踪失败的帧;失败帧将反馈给标注者进行人工标注,而其他可靠帧将通过TLNet进行进一步优化,基于前向反向的初始标注,生成更加精确的目标框标注。通过整合人工标注和优化后的自动标注,最终得到可靠的逐帧视频标注。

值得注意的是,本文设计的TLNet模型不依赖具体的跟踪器,泛化性强,可选择任意轻量化跟踪器高效率地生成前向反向初始标注;其次,本文的质量评估模块是帧级别,而非目标框级别,即本文针对当前帧的初始前向反向跟踪结果,并综合考虑前后帧

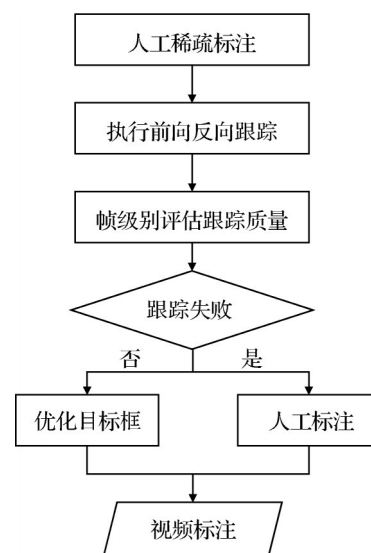


图2 视频标注流程图

Fig. 2 Procedure of generating video annotations

的目标视觉与运动时序信息,给出当前帧整体的跟踪状态质量评估,而不是分别对前向和反向跟踪结果进行质量评估。该方案能够筛选出序列中挑战难度大的帧(即易跟踪失败的帧),执行人工标注,从而保障了标注的可靠性。

2.2 网络模型

本文设计的视频标注模型 TLNet 框架如图 3 所

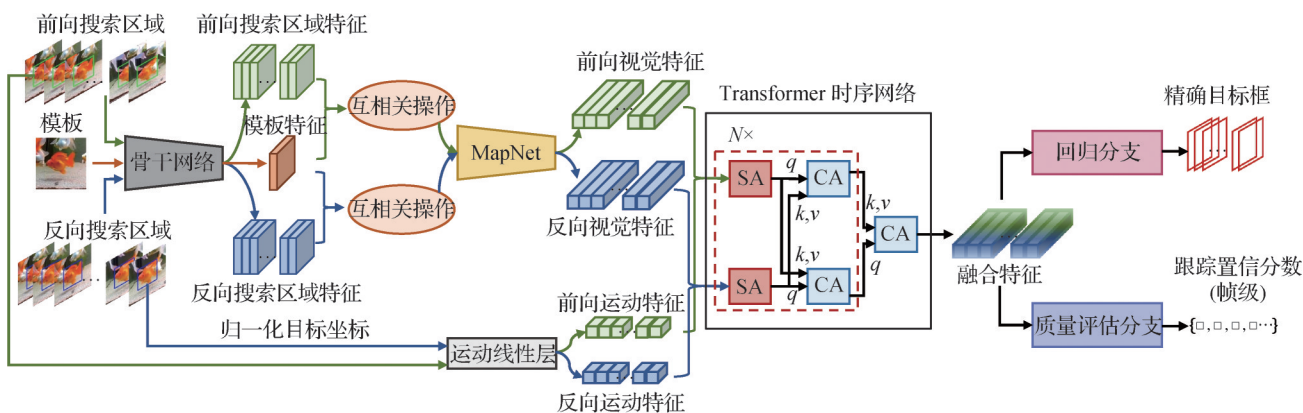


图3 视频标注模型 TLNet 框架图

Fig. 3 Framework of the proposed video annotation model TLNet

2.2.1 目标多维度特征提取

输入单帧模板图像、连续多帧的前向和反向搜索区域图像和对应的前向反向跟踪结果,该模块负责提取出连续多帧的目标的多维度特征,即视觉特征和运动特征。

对于视觉特征,本文通过执行模板与搜索区域的像素级互相关操作,生成目标的响应图,并将其作为目标的视觉特征。具体来讲,模板图像和前向反向搜索区域图像将分别通过骨干网络提取出模板特征 Z 、连续 T 帧的前向搜索区域特征 $\{X_f^i\}_{i=1,\dots,T}$ 和反向搜索区域特征 $\{X_b^i\}_{i=1,\dots,T}$ 。 Z 分别与 X_f 和 X_b 执行式(2)所示的像素级互相关操作,生成对应的前向反向响应图 M_f 和 M_b 。通过像素级互相关操作,响应图记录了各搜索区域中目标与模板的视觉匹配度,可反映目标的外观变化。为挖掘响应图中的重要视觉信息,并减少计算量,实现轻量化设计,响应图将输入响应图网络(response map network, MapNet),并输出对应的序列前向和反向视觉特征向量,表示为 $V_v^{fb} \in \mathbf{R}^{T \times d_v}$,其中 d_v 表示视觉特征向量的维度。MapNet 由 3 层卷积核大小为 3 的卷积层组成,每个卷积层后连接一个 ReLU 激活层和池化层。

对于运动特征,本文将序列的前向反向跟踪结

果,即连续 T 帧的目标框坐标进行归一化,并输入运动线性层提取对应的前向反向运动特征向量,表示为 $V_m^{fb} \in \mathbf{R}^{T \times d_m}$,其中 d_m 表示运动特征向量的维度。该运动特征向量记录了目标的运动模式和几何变化。

最终的目标多维度特征将通过连接 V_v^{fb} 与 V_m^{fb} 得到,表示为 V_{mix}^{fb} ,其将作为 Transformer 时序网络的输入,在序列维度和方向维度上进行特征融合。

2.2.2 Transformer 时序特征融合

Transformer 在视觉任务中往往切割图像块并将其作为一个词向量,通过融合不同图像块特征提取出具有辨别性的图像特征。与此不同的是,本文引入序列维度,将单帧图像的特征作为一个整体,挖掘不同帧图像之间的融合特征。即本文通过 Transformer 模型来实现序列特征及前向反向特征的融合。该模型参照 TransT 结构设计,由自注意力模块(self-attention, SA)和交叉注意力模块(cross-attention, CA)组成。两个模块的结构图如图 4 所示。其中 SA 通过计算自注意力,在时序维度上分别融合 V_{mix}^{fb} 各自的特征,而 CA 则通过计算交叉注意力来融合前向和反向的特征,最终生成双向时序融合特征。注意力计算方式如式(1)所示。两个 SA 模块和两个

果,即连续 T 帧的目标框坐标进行归一化,并输入运动线性层提取对应的前向反向运动特征向量,表示为 $V_m^{fb} \in \mathbf{R}^{T \times d_m}$,其中 d_m 表示运动特征向量的维度。该运动特征向量记录了目标的运动模式和几何变化。

最终的目标多维度特征将通过连接 V_v^{fb} 与 V_m^{fb} 得到,表示为 V_{mix}^{fb} ,其将作为 Transformer 时序网络的输入,在序列维度和方向维度上进行特征融合。

2.2.2 Transformer 时序特征融合

Transformer 在视觉任务中往往切割图像块并将其作为一个词向量,通过融合不同图像块特征提取出具有辨别性的图像特征。与此不同的是,本文引入序列维度,将单帧图像的特征作为一个整体,挖掘不同帧图像之间的融合特征。即本文通过 Transformer 模型来实现序列特征及前向反向特征的融合。该模型参照 TransT 结构设计,由自注意力模块(self-attention, SA)和交叉注意力模块(cross-attention, CA)组成。两个模块的结构图如图 4 所示。其中 SA 通过计算自注意力,在时序维度上分别融合 V_{mix}^{fb} 各自的特征,而 CA 则通过计算交叉注意力来融合前向和反向的特征,最终生成双向时序融合特征。注意力计算方式如式(1)所示。两个 SA 模块和两个

CA 模块共同构成一个融合单元(图 3 中虚线框),该融合单元重复执行 N 次,并在最后接入一个 CA 模块来融合两个分支的特征,解码出最终的融合特征,该特征融合了时序和方向两个维度上的目标信息。

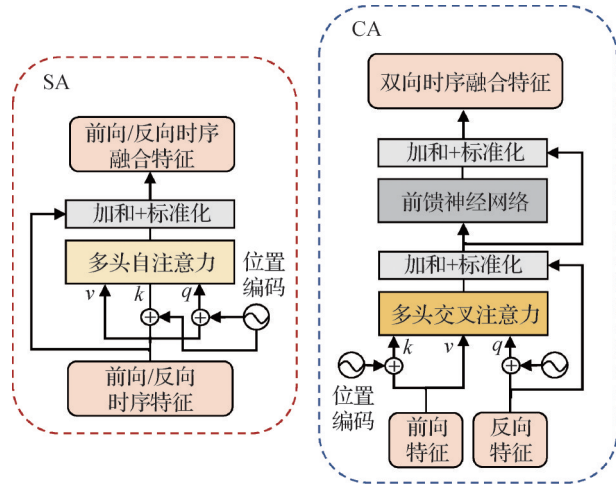


图 4 SA 模块和 CA 模块结构图

Fig. 4 Architecture of the SA module and CA module

为了在特征融合过程中保持序列的顺序性,对每一帧定义位置编码并添加至对应目标特征向量中。位置编码(positional encoding, PE)定义为

$$\begin{cases} PE_{(pos, 2i)} = \sin\left(pos/10\,000^{\frac{2i}{d_{model}}}\right) \\ PE_{(pos, 2i+1)} = \cos\left(pos/10\,000^{\frac{2i}{d_{model}}}\right) \end{cases} \quad (3)$$

式中, d_{model} 表示向量维度, pos 表示位置。

2.2.3 预测分支

Transformer 时序网络生成的融合特征将分别输入质量评估和回归分支中。前者输出每一帧跟踪状态的质量分数,而后者则结合时序的前向和反向跟踪结果,输出更准确的目标框坐标。两个分支均由 3 个全连接层实现,除最后一层外,其余每个全连接层后连接一个 ReLU 层。其中质量评估分支为每帧输出 1 维向量,表示质量分数;而回归分支输出 4 维向量,分别表示预测目标左上角和右下角坐标。

将质量评估分支和回归分支设计为两个独立的子网络,即各自训练独立的任务,并最终生成独立的模型参数,多维度特征提取模块和 Transformer 时序网络的参数不共享。通过实验验证了与共同训练两个子任务并共享特征提取及特征融合阶段的模型参

数相比,分离两个子网络并分别训练会带来更高的性能收益。因此本文采用后者。猜测导致这一现象的原因是两个任务侧重点不同,因此需要从不同角度提取并融合特征。其中,质量评估分支的目的是评估当前帧的跟踪状态是否可靠,进而筛选出跟踪失败的帧,因此视觉特征将占据主导作用,运动特征则作为辅助用来检测运动模式是否异常。而回归分支则需要依赖前向和反向的运动特征回归出更准确的目标框,视觉特征在其中主要起到权重分配的作用。

2.3 损失函数

2.3.1 质量评估子网络

对于当前帧的跟踪状态质量评估将基于前向与反向跟踪结果与真值目标框的重叠率 IoU。参照 VASR 方法,定义单个跟踪结果的质量分数为

$$f(IoU) = \frac{\beta \sqrt{\alpha} (IoU - 0.5)}{\sqrt{\beta + \alpha(IoU - 0.5)^\beta}} \quad (4)$$

式中, α 和 β 为控制参数,其设置与 VASR 相同,分别为 50 和 2。相较于 IoU ,非线性函数 $f(IoU)$ 将分数区间从 $[0, 1]$ 扩展至 $(-1, 1)$,能更辨别性地表示不同跟踪结果的质量。由于本文尝试根据前向(forward)和反向(backward)的跟踪结果去评估单帧的跟踪状态质量,并基于此筛选出挑战难度大的难跟踪帧,因此对于前向和反向跟踪都失败的帧,被赋予低质量分数;而对于单向(前向或反向)跟踪失败的帧,由于仍存在可靠的单向跟踪结果,后续仍可根据回归子网络回归出精确目标框,因此基于跟踪成功的单向结果,仍被赋予较高分数。本文定义单帧的质量分数 $\hat{g}$ 为前向和反向质量分数的最大值,该分数能够有效度量单帧的跟踪状态,低分数则表明该帧的前反向跟踪都会失败,为高难度帧,需要由人工标注。具体为

$$\hat{g} = \max(f(IoU_f), f(IoU_b)) \quad (5)$$

式中,下角 f、b 分别表示 forward 和 backward。

输入 T 帧的序列数据,质量评估子网络会预测对应帧的质量分数 $\{g_i\}_{i=1,2,\dots,T}$,该网络训练的损失函数采用均方误差损失,定义为

$$L_{conf} = \frac{1}{T} \sum_{i=1}^T \|g_i - \hat{g}_i\|_2^2 \quad (6)$$

2.3.2 回归子网络

结合 gIoU(generalized intersection over union)损失

LgIoU(Rezatofighi等, 2019)和 L_1 损失来训练回归子网络。回归分支损失函数定义为

$$L_{\text{reg}} = \frac{1}{T} \sum_{i=1}^T \lambda_1 L_{\text{gIoU}}(b_i, \hat{b}_i) + \lambda_2 L_1(b_i, \hat{b}_i) \quad (7)$$

式中, b_i 和 $\hat{b}_i$ 分别表示预测的目标框和真值目标框。 λ_1 和 λ_2 为两个超参数, 分别设置为2.0和5.0。

3 实验与分析

3.1 实现细节与评估指标

3.1.1 模型设计

为高效地生成视频标注, 本文将TLNet设计为轻量化模型。具体来说, 骨干网络采用ResNet18(residual network)(He等, 2016), 其中, 第4阶段下采样层的卷积步长更改为1, 并输出第4阶段的特征图。模板及双向搜索区域特征图 $Z, X_b^i, X_f^i \in \mathbb{R}^{C \times H \times W}$, 其中, $H, W = \frac{H_0}{8}, \frac{W_0}{8}$, H_0, W_0 为输入图像的尺寸, $C = 256$ 。通过互相关操作及MapNet生成的视觉特征向量 V_v^{th} 维度 $d_v = 64$, 通过运动线性层提取的运动特征向量 V_m^{th} 维度 $d_m = 16$, 结合两特征向量, 最终得到维度为64的 $V_{\text{mix}}^{\text{th}}$, 并输入Transformer时序网络。对于Transformer时序网络, 本文设置多头注意力机制的头(head)数目为2, 前馈神经网络维度为256, 融合单元的层数 $N = 4$ 。

3.1.2 训练设置

TLNet在大规模跟踪数据集LaSOT和GOT-10k的训练集上训练, 并生成LaSOT和TrackingNet训练数据集的标注。对于TrackingNet数据集, 使用LaSOT和GOT-10k所有序列进行训练; 而对于LaSOT数据集, 为保证实验公平性, 将LaSOT训练集随机分割为2个子集, 使用一个子集和GOT-10k训练模型, 并生成另一个子集的数据集标注。本文采用稀疏度为30的人工标注(即数据集官方标注), 并选用轻量化跟踪器HCAT生成每个片段的前向反向跟踪结果。训练过程中, 随机选择一个序列中人工标注的帧裁剪模板图像, 并随机选择连续 T 帧, 基于前向和反向跟踪结果裁剪对应的搜索区域图像, 其中序列长度 $T = 20$ 。为展现更多搜索区域中目标的细节, 实现精确的目标框预测, 模板图像和搜索图像均采用2倍目标框范围进行裁剪, 并缩放至统一大小, $H_0 = W_0 = 128$ 。

本文算法使用深度学习框架PyTorch实现, 并使用单个NVIDIA GTX2080Ti GPU进行训练和测试。采用AdamW(Loshchilov和Hutter, 2019)优化器进行模型训练, 初始学习率为 10^{-4} , 每训练20次降低10倍, 共迭代训练70次。批大小(batch size)设置为16。

3.1.3 评估指标

由于LaSOT数据集的官方标注为人工逐帧精确标注, 而TrackingNet则为跟踪器插值方式生成的标注, 存在噪声。因此本文设计两个标准分别评估两个数据集上生成的标注质量。

对于LaSOT数据集, 本文采用与真值目标框的平均重叠率mIoU来评估数据集标注质量。此外, 对于质量评估子网络, 采用准确率(accuracy, Acc)、召回率Recall和特异度(true negative rate, TNR)来评估预测质量分数。Acc表示正确分类的帧数占总帧数的比例, Recall表示跟踪成功的帧数中正确识别的比例, TNR表示跟踪失败的帧数中被识别出来的比例。

由于TrackingNet数据集的官方标注存在噪声, 导致上述评价标准不可靠, 因此本文采用跟踪指标评估TrackingNet数据集上生成的标注质量。即使使用TrackingNet官方标注和本文自动生成的标注分别重训现有的跟踪算法, 并在不同的跟踪数据集上测试训练出的模型性能。本文选择3个不同的跟踪算法进行重训, 分别是ATOM(Danelljan等, 2019)、DiMP50(Bhat等, 2019)和PrDiMP50(probabilistic regression DiMP)(Danelljan等, 2020), 并分别在LaSOT、TrackingNet和GOT-10k数据集上进行测试与评估。

3.2 LaSOT数据集标注评估

本文以轻量化跟踪算法HCAT为例, 生成LaSOT训练数据集的标注。标注过程共分为3个阶段, 首先对稀疏标注(每隔30帧标注, 共3.3%标注帧)分割的每个片段执行前向和反向跟踪, 生成初始的标注结果; 然后通过本文质量评估子网络评估每一帧的标注质量, 并通过阈值筛选出跟踪失败帧, 进行人工标注; 最后执行本文回归子网络, 对目标框进行优化, 生成最终的数据集标注。表1展示了各阶段生成的标注的质量评估结果。其中, mIoU表示与真值目标框的平均重叠率, $\text{Rate}_{<0.3}$ 、 $\text{Rate}_{<0.5}$ 和 $\text{Rate}_{>0.7}$ 分别表示重叠率低于0.3、低于0.5和高于0.7的帧数占总帧数的比例。由前向反向跟踪生成的初始标注

的平均重叠率为 0.824。在经过质量筛选(其中, 5.4%的帧被判定为跟踪失败,并进行人工标注)后, mIoU 提升 4%,重叠率低于 0.3 和 0.5 的帧分别下降至 1.41% 和 2.52%。经过目标框优化后,最终的数据集标注的 mIoU 提升至 0.871,重叠率高于 0.7 的帧数提升至 92.9%。由此可见,本文提出的标注算法能节省 91.3%(3.3%稀疏标注和 5.4%筛选跟踪失败帧)的人工标注,并相较于跟踪算法插值的方式,使平均重叠率提升 4.7%。

表 1 LaSOT 数据集的各阶段标注质量评估

Table 1 Quality measurement of generated annotations from different steps on LaSOT dataset

	mIoU	Rate _{<0.3} /%	Rate _{<0.5} /%	Rate _{>0.7} /%
初始标注结果	0.824	4.83	6.40	88.3
质量筛选后	0.864	1.41	2.52	92.7
优化目标框后	0.871	1.36	2.38	92.9

注:加粗字体表示各列最优结果。

为展示本文的质量评估网络能够有效筛选出挑战难度大的跟踪失败帧(即双向均跟踪失败),定义每一帧是否被跟踪成功的真值标签为

$$label = \begin{cases} 1 & IoU_f > 0.5 \text{ 或 } IoU_b > 0.5 \\ 0 & \text{其他} \end{cases} \quad (9)$$

即当该帧的前向或反向跟踪结果与真值目标框的重叠率大于 0.5 时,认为该帧能够被现有跟踪算法成功跟踪,属于跟踪成功帧,后续应通过回归网络对双向跟踪结果进行优化;否则认为该帧很难被跟踪算法成功跟踪,属于跟踪失败帧,后续应执行人工标注。而对于质量评估网络预测的分数,当分数大于设定的阈值时,预测该帧跟踪成功,否则预测为跟踪失败。值得注意的是,采用 0.5 作为 IoU 阈值的原因是与跟踪任务的性能评价指标保持一致,例如 OTB (object tracking benchmark) (Wu 等, 2013) 及 LaSOT (Fan 等, 2019) 中的成功率阈值。

图 5 绘制了不同分数阈值下,筛选的准确率 Acc、召回率 recall、特异度 TNR 以及筛选前后的平均重叠率 mIoU。此外,为反映需人工标注的工作量,该图亦绘制了失败率 Rate_{fail},即筛选出的失败帧占总帧数的比率。可以看出,对于不同的阈值设定,筛选出的失败帧比例介于 2%~13%,随着筛选出的失败率增加, TNR 大幅提升,而 Acc 和 recall 小幅下降。

相较于初始标注的 mIoU,筛选后的 mIoU 显著提升。当筛选出约 12% 失败帧时, TNR 高达 0.926,这意味着绝大多数的标注失败帧均被筛选出来。为平衡人工标注的成本和最终标注的质量,本文将阈值设定为 0.7,其中失败率为 5.4%, TNR 为 0.761,筛选失败帧后, mIoU 由 0.811 提升至 0.851。值得注意的是,与表 1 的 mIoU 不同,为公平评估模型的筛选能力,此处 mIoU 的统计不包含任何人工标注的帧。

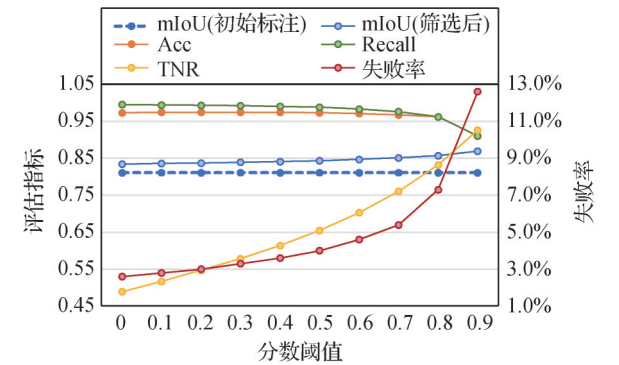


图 5 不同阈值的分数评估
Fig. 5 Score evaluation with different thresholds

3.3 TrackingNet 数据集标注评估

由于 TrackingNet 的训练集不是逐帧人工标注,存在噪声,因此本文通过间接评估跟踪性能来评估本文对于 TrackingNet 的标注质量。本文选用了 3 个现有的跟踪算法,即 ATOM、DiMP50 和 PrDiMP50,并分别使用 TrackingNet 自身标注和本文标注对模型重训练,以下是在 3 个跟踪数据集上的跟踪性能评估结果。其中“_GT”表示使用 TrackingNet 官方标注进行训练,“_Ours”表示使用本文的 TrackingNet 标注。为与 VASR 方法 (Dai 等, 2021) 进行对比,本文亦列出其对应的结果,表示为“_VASR”。值得注意的是,参考下文 3.5.2 中的耗时分析,对于 LaSOT 量级的数据集,本文的自动筛选及标注过程仅需约 43 h,相较于 VASR 方法的约两周时间,时间成本有所缩减。

1) LaSOT。图 6 展示了 LaSOT 数据集上的跟踪性能比较,采用的评价指标为成功率和准确率。相较于使用 TrackingNet 官方标注进行训练,使用本文的标注训练分别为 ATOM、DiMP50 和 PrDiMP50 在成功率上提升了 0.5%、0.9% 和 2.2%。这间接表明了本文生成的标注质量更高,能训练出更加鲁棒的跟踪模型。此外,表 2 对比了本文方法和 VASR 方法的结果。对于 ATOM 和 DiMP50,本文在 LaSOT 数据

集的成功率上优于 VASR 方法, PrDiMP50 的性能稍有下降。

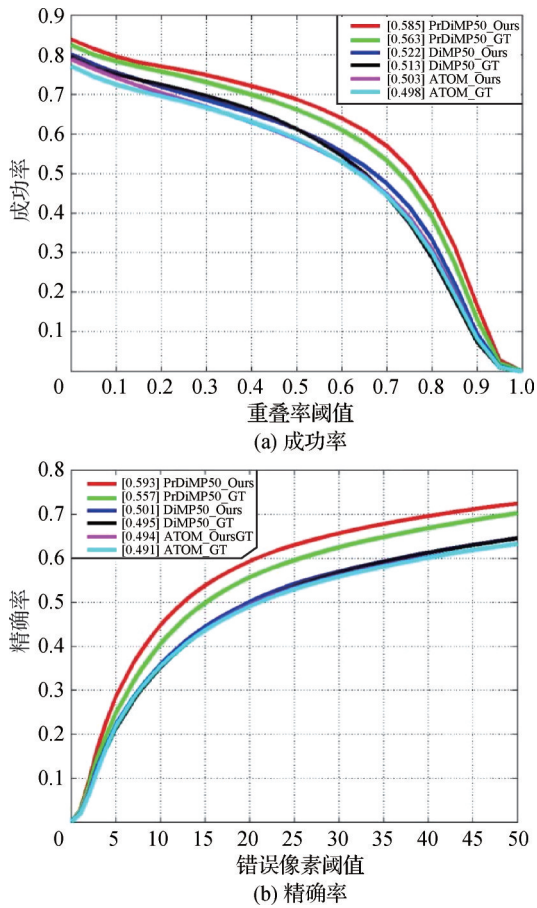


图6 使用TrackingNet数据集的不同标注训练的不同跟踪器在LaSOT数据集上的成功率和精确率曲线图
Fig. 6 Tracking performance on LaSOT dataset by using our trackingnet annotations (ours) and trackingnet official annotations (GT) ((a) success plots; (b) precision plots)

2)TrackingNet。表2列出了不同方法的不同训练模型在成功率和归一化准确率上的跟踪结果。相较于官方标注,本文标注可分别为 ATOM、DiMP50 和 PrDiMP50 在成功率上带来 0.7%、1.5% 和 2.5% 的性能收益。虽然 VASR 方法在该数据集上表现更好,但平衡时间成本后,本文能有效缩小时间成本并基本达到持平的跟踪性能。

3)GOT-10k。表2最右两列展示了 GOT-10k 上的跟踪性能,评估指标分别是平均重叠率(average overlap, AO)和阈值为 0.75 的成功率(success rate, $SR_{0.75}$)。在该数据集上,本文相较于使用官方标注训练的模型有大幅的性能提升,在 $SR_{0.75}$ 上分别为 ATOM、DiMP50 和 PrDiMP50 提升了 4.3%、8.8% 和

表2 使用TrackingNet不同标注重新训练各跟踪器在不同数据集上的跟踪性能对比

Table 2 Performance comparison of retrained trackers using different TrackingNet annotations

跟踪器	LaSOT	TrackingNet		GOT-10k	
	成功率	成功率	归一化准确率	AO	$SR_{0.75}$
ATOM_GT	0.498	0.706	0.653	0.538	0.354
ATOM_VASR	0.495	/	/	/	/
ATOM_Ours	0.503	0.713	0.651	0.545	0.397
DiMP50_GT	0.513	0.717	0.661	0.546	0.349
DiMP50_VASR	0.521	0.741	0.682	0.570	0.441
DiMP50_Ours	0.522	0.732	0.673	0.569	0.437
PrDiMP50_GT	0.563	0.736	0.683	0.576	0.436
PrDiMP50_VASR	0.587	0.762	0.706	0.589	0.488
PrDiMP50_Ours	0.585	0.761	0.712	0.625	0.515

注:加粗字体表示各跟踪器的最优结果,“/”表示文献未给出结果。

7.9%。相较于 VASR 方法,本文会导致 DiMP50 方法在 AO 指标上下降 0.1%,并导致 PrDiMP50 提升 3.6%。

3.4 消融实验

在 LaSOT 数据上实验,探索不同融合模型、质量分数真值及序列长度对模型筛选能力的影响。

1)Transformer 时序模型。为验证 Transformer 时序模型融合时序和双向数据能力,将 Transformer 模型替换为长短期记忆网络(long short-term memory, LSTM)模型,筛选结果如表3所示。

表3 不同融合模型对于筛选能力的影响

Table 3 Effect of different fusion models on filtering ability

模型	mIoU	Rate _{fail} /%	Acc/%	Recall/%	TNR/%
LSTM	0.849	5.8	95.4	96.7	65.7
Transformer	0.851	5.4	96.7	97.6	76.1

注:加粗字体表示各列最优结果。

从表3可以看出,相较于 Transformer 模型, LSTM 模型识别失败帧的能力下降,在预测出相似失败帧比率的前提下,即人工标注成本相似,TNR 由 76.1% 降至 65.7%,Acc 和 recall 也有所下降,这说明 LSTM 模型会错误地筛选出较多的成功帧,并存在较多的失败帧未被筛选出来。

2)帧级质量表示。设计每一帧的真值标注质量由前向和反向跟踪质量分数的最大值表示,如式(5)所示。为验证最大值操作的合理性,将最大值真值替换为平均真值,筛选结果如表4所示。可以看出,替换成平均操作后,TNR下降超过3%。因此相比于平均真值,采用的最大值真值能筛选出更多的失败帧。

表4 不同质量分数真值对于筛选能力的影响
Table 4 Effect of different fusion models and different quality score's ground truth on the filtering ability

质量分数真值	mIoU	Rate _{fail} /%	Acc/%	Recall/%	TNR/%
平均真值	0.852	5.5	96.4	97.5	72.8
最大值真值	0.851	5.4	96.7	97.6	76.1

注:加粗字体表示各列最优结果。

3)序列长度。为验证输入的序列长度对于性能的影响,本文使用不同长度的序列数据进行模型训练和生成标注,筛选结果如表5所示。可以看出,当 $T=20$ 时,能更准确地识别出失败帧,得到更高质量的数据集标注。因此本文选择 $T=20$ 。值得注意的是,当 $T=5$ 时,TNR最低,这主要是由于序列长度过低时,模型无法捕获长距离的目标视觉变化和运动模式,筛选能力较低。

表5 不同序列长度对于筛选能力的影响
Table 5 Effect of sequence length on the filtering ability

序列长度	mIoU	Rate _{fail} /%	Acc/%	Recall/%	TNR/%
$T=5$	0.850	5.6	96.6	97.7	72.9
$T=10$	0.849	5.5	96.7	97.7	74.5
$T=20$	0.851	5.4	96.7	97.6	76.1
$T=30$	0.850	5.5	96.6	97.6	74.9

注:加粗字体表示各列最优结果。

3.5 实验分析

3.5.1 泛化性分析

由于在模型的训练过程中未使用前向反向跟踪过程中产生的中间结果,例如置信分数、响应图等,TLNet不依赖于任何具体的跟踪算法,具有强泛化性。为验证TLNet强大的泛化能力,将训练好的TLNet(HCAT方法生成的前向反向跟踪结果作为训练数据)作用于STARK方法(ResNet101模型),根据STARK算法产生的前向反向跟踪结果,进行质量筛选和优化目标框。生成的数据集标注质量如表6所

示。由于STARK算法自身的高鲁棒性,初始标注结果已有较高质量,mIoU为0.856。筛选后,仅2.9%的帧被识别为失败帧,进行人工标注,筛选后标注的mIoU提升至0.879,其中Rate_{<0.3}和Rate_{<0.5}均有明显下降。值得注意的是,由于STARK本身的跟踪结果已有较高的准确率,本文的回归子网络并未对其起到优化作用,但Rate_{<0.3}有所下降,说明回归子网络对于优化质量极差的帧仍起到一定的作用。该结果说明回归子网络存在一定的优化上限,因此仅对性能相对较差的轻量化跟踪算法执行优化步骤。

表6 基于STARK跟踪结果的标注质量评估
Table 6 Quality measurement of generated annotations based on the STARK's tracking results

STARK跟踪结果	mIoU	Rate _{<0.3} /%	Rate _{<0.5} /%	Rate _{>0.7} /%
初始标注结果	0.856	4.13	5.33	91.0
质量筛选后	0.879	2.12	3.08	93.5
优化目标框后	0.874	2.09	3.07	93.18

注:加粗字体表示各列最优结果。

3.5.2 耗时分析

对于自动标注方法,除了保障标注的高质量外,标注的效率也是至关重要的。VASR方法提出基于选择和细化的自动标注策略,简化了标注流程并节约了标注时间。本文对该方法进行改进后,进一步简化了标注流程,并缩减了生成标注的时间。

VASR方法采用DiMP50算法执行初始的前向反向跟踪,由于训练过程需要DiMP50方法和Alpha-Refine方法生成对应的跟踪结果及分割掩码等数据,导致模型不具有泛化性,并且数据准备时间长。对于生成LaSOT数据集量级的标注,选择不同的基准跟踪算法均需要重新进行数据准备和模型训练,共需耗费约两周时间(Dai等,2021)。该方法的耗时主要是由于跟踪算法的选择和模型的低泛化性。表7统计了不同跟踪方法的跟踪性能、在初始化阶段和单帧跟踪阶段的耗时,以及长度为30帧的片段的平均耗时。此处将现有的两个轻量化跟踪算法与VASR使用的DiMP50算法进行对比,其中mIoU为前向反向跟踪结果的平均mIoU。可以看出,虽然两个轻量化方法在跟踪性能上均低于DiMP50,但在执行前反向跟踪时能够节省约4倍的时间,实现高效标注。

表7 不同跟踪算法在LaSOT数据集的耗时及性能分析
Table 7 Time-consuming and performance analysis of different trackers on LaSOT dataset

跟踪算法	mIoU	初始化/s	单帧跟踪/s	单片段/s
STARK-Lightning	0.806	0.07	0.005	0.5
HCAT	0.824	0.15	0.004	0.55
DiMP50	0.834	0.75	0.035	2.0

注:加粗字体表示各列最优结果。

此外,由于本文模型和具体的跟踪算法解耦,对于待标注的数据集,可任意选择现有的轻量化跟踪算法生成初始标注,而无需进行模型的重训练,因此仅需要考虑前向反向跟踪时间及TLNet执行时间。对于生成LaSOT数据集量级(2.8 M)的标注,并采用现有的轻量化跟踪算法HCAT,前向反向跟踪需耗时约35 h,TLNet的速度约为200 帧/s,执行筛选和优化过程需耗费约8 h,因此共耗费约43 h,相较于VASR耗费的两周时间,本文方法在标注自动生成效率上有显著提升。

为公平对比本文方法与VASR方法,表8展示了在筛选出相似失败率(2.7%和2.9%)的前提下,本文方法与VASR方法在标注质量和耗时方面的对

比。可以看出,本文方法节约了大量的标注时间,约82%,并且本文方法对标注的mIoU提升水平与VASR相当,分别提升了3.2%和3.1%。

表8 与VASR方法在标注质量和耗时方面的比较
Table 8 Comparison with VASR on annotation quality and time-consuming

方法	mIoU 初始标注	mIoU 最终标注	mIoU 提升/%	Rate _{fail} /%	耗时/h
VASR	0.834	0.865	3.1	2.7	336
TLNet	0.824	0.856	3.2	2.9	43

注:加粗字体表示各列最优结果。

3.5.3 定性分析

为展示本文的质量评估网络能够通过预测帧级分数来有效地筛选出标注失败帧,图7列举了LaSOT数据集的5个序列的若干结果。其中绿色框被网络预测为标注成功帧;而红色框被预测为标注失败帧,后续反馈给标注者进行人工精确标注。可以看出,当前向和反向跟踪的目标框中有至少一个标注能够达到高精度的标准,质量评估网络就会预测出较高的分数;而当前向和反向的跟踪结果都较差时,质量评估网络会给出较低的分



图7 质量评估网络预测的分数图示

Fig. 7 Illustration of the scores output by the proposed quality evaluation network

败帧。由于当前向或反向存在准确的标注时,说明该帧的跟踪难度较低,后续回归子网络仍可根据反向结果生成准确的最终标注;而当前向和反向均跟踪失败时,说明该帧挑战难度大,需要由人工标注。与目标框级别的分数预测相比,这种帧级的分数预测更能准确地筛选出失败帧,并减少成功帧的

错误识别,减轻人工标注的工作量。

此外,对本文生成的标注与TrackingNet数据集的官方标注进行了定性比较,如图8所示。可以看出,通过筛选失败帧及优化目标框,本文方法能更好地应对多种跟踪挑战,生成更高质量的视频标注。

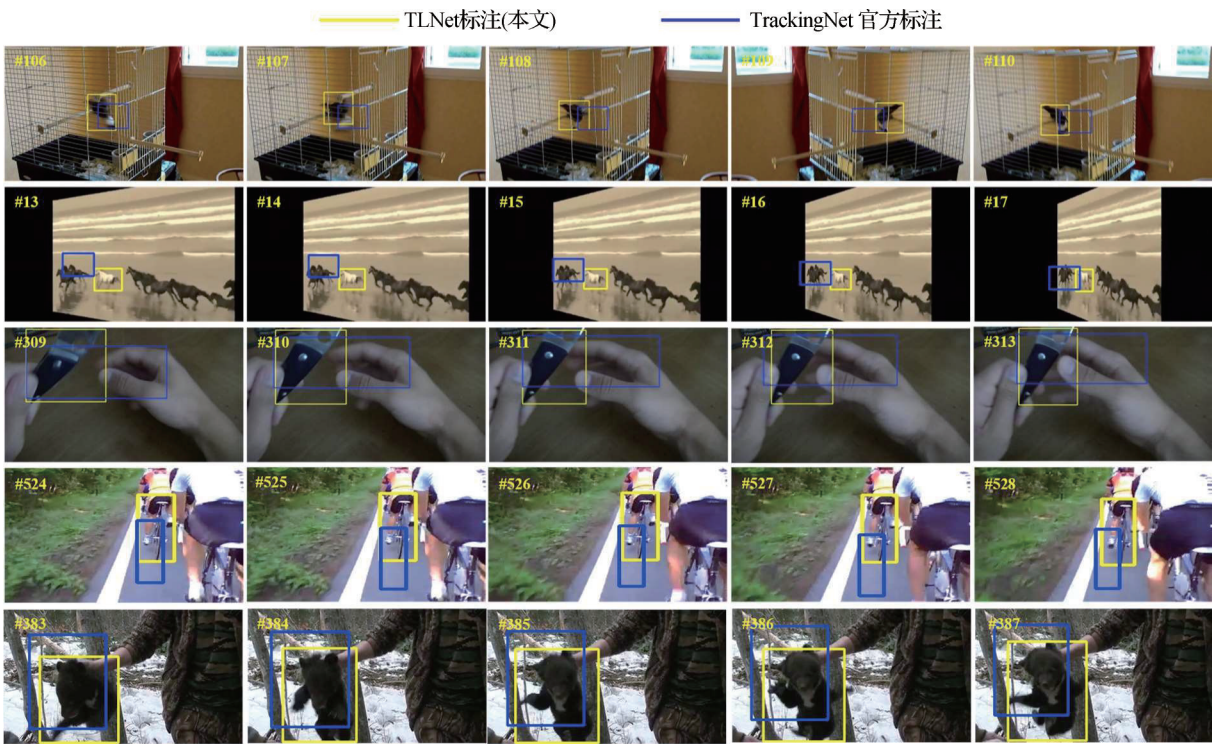


图8 TrackingNet官方标注与本文标注的对比

Fig. 8 Comparison between the official annotations of TrackingNet dataset and our annotations

4 结 论

训练现有的基于深度网络的跟踪算法需要大规模的具有精确标注的视频数据集,而逐帧标注大规模数据集会耗费大量的人力和时间成本。为减轻人力消耗,并平衡标注的质量,提出TLNet网络来高效地自动生成高质量大规模数据集标注。该网络包含互相关模块及Transformer时序模型,通过输入时序的目标视觉和运动信息,来提取并融合双向的初始跟踪结果,进而进行标注的评估和优化。通过筛选出失败帧并优化剩余目标框,生成高质量的标注。采用两种评价方式分别评估LaSOT及TrackingNet数据集的标注质量。前者采用与真值目标框的平均重叠率以及筛选失败帧的准确率及特异度等指标进

行直接评估;而后者由于自身标注存在噪声,通过重新训练跟踪算法并评估跟踪性能的方式进行间接评估。大量实验验证了本文方法具有强泛化性,能够利用任意轻量化跟踪算法实现高效的自动标注。本文工作的主要局限性在于回归子网络的目标框预测存在一定的性能上限,对于高性能跟踪算法生成的初始标签无法起到优化的作用。这主要是由于视觉特征的空间信息没有很好地保留,因此回归分支无法根据空间位置预测更精准的目标框。后续工作将从该点出发,尝试优化模型框架,尽可能保留更多目标的空间信息,实现更精确的目标框优化。

参考文献(References)

Ba J L, Kiros J R and Hinton G E. 2016. Layer normalization [EB/OL]. [2022-08-16]. <https://arxiv.org/pdf/1607.06450.pdf>

- Bertinetto L, Valmadre J, Henriques J F, Vedaldi A and Torr P H S. 2016. Fully-convolutional siamese networks for object tracking//Proceedings of 2016 European Conference on Computer Vision. Amsterdam, the Netherlands: Springer: 850-865 [DOI: 10.1007/978-3-319-48881-3_56]
- Bhat G, Danelljan M, van Gool L and Timofte R. 2019. Learning discriminative model prediction for tracking//Proceedings of 2019 IEEE/CVF International Conference on Computer Vision. Seoul, Korea (South): IEEE: 6181-6190 [DOI: 10.1109/ICCV.2019.00628]
- Carion N, Massa F, Synnaeve G, Usunier N, Kirillov A and Zagoruyko S. 2020. End-to-end object detection with transformers//Proceedings of the 16th European Conference on Computer Vision. Glasgow, UK: Springer: 213-229 [DOI: 10.1007/978-3-030-58452-8_13]
- Chen X, Kang B, Wang D, Li D D and Lu H C. 2022. Efficient visual tracking via hierarchical cross-attention transformer [EB/OL]. [2022-10-30]. <https://arxiv.org/pdf/2203.13537.pdf>
- Chen X, Yan B, Zhu J W, Wang D, Yang X Y and Lu H C. 2021. Transformer tracking//Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA: IEEE: 8122-8131 [DOI: 10.1109/CVPR46437.2021.00803]
- Cui Y T, Jiang C, Wang L M and Wu G S. 2022. MixFormer: end-to-end tracking with iterative mixed attention//Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, USA: IEEE: 13598-13608 [DOI: 10.1109/CVPR52688.2022.01324]
- Dai K N, Zhao J, Wang L J, Wang D, Li J H, Lu H C, Qian X S and Yang X Y. 2021. Video annotation for visual tracking via selection and refinement//Proceedings of 2021 IEEE/CVF International Conference on Computer Vision. Montreal, Canada: IEEE: 10276-10285 [DOI: 10.1109/ICCV48922.2021.01013]
- Danelljan M, Bhat G, Khan F S and Felsberg M. 2019. ATOM: accurate tracking by overlap maximization//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA: IEEE: 4655-4664 [DOI: 10.1109/CVPR.2019.00479]
- Danelljan M, van Gool L and Timofte R. 2020. Probabilistic regression for visual tracking//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 7181-7190 [DOI: 10.1109/CVPR42600.2020.00721]
- Devlin J, Chang M W, Lee K and Toutanova K. 2019. BERT: pre-training of deep bidirectional transformers for language understanding [EB/OL]. [2022-08-16]. <https://arxiv.org/pdf/1810.04805.pdf>
- Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X H, Unterthiner T, Dehghani M, Minderer M, Heigold G, Gelly S, Uszkoreit J and Housley N. 2021. An image is worth 16x16 words: transformers for image recognition at scale [EB/OL]. [2022-08-16]. <https://arxiv.org/pdf/2010.11929.pdf>
- Fan H, Lin L T, Yang F, Chu P, Deng G, Yu S J, Bai H X, Xu Y, Liao C Y and Ling H B. 2019. LaSOT: a high-quality benchmark for large-scale single object tracking//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA: IEEE: 5369-5378 [DOI: 10.1109/CVPR.2019.00552]
- He K M, Zhang X Y, Ren S Q and Sun J. 2016. Deep residual learning for image recognition//Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, USA: IEEE: 770-778 [DOI: 10.1109/CVPR.2016.90]
- Huang L H, Zhao X and Huang K Q. 2021. GOT-10k: a large high-diversity benchmark for generic object tracking in the wild. IEEE Transactions on Pattern Analysis and Machine Intelligence, 43(5): 1562-1577 [DOI: 10.1109/TPAMI.2019.2957464]
- Kuznetsova A, Talati A, Luo Y W, Simmons K and Ferrari V. 2021. Efficient video annotation with visual interpolation and frame selection guidance//Proceedings of 2021 IEEE Winter Conference on Applications of Computer Vision. Waikoloa, USA: IEEE: 3069-3078 [DOI: 10.1109/WACV48630.2021.00311]
- 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. Long Beach, USA: IEEE: 4277-4286 [DOI: 10.1109/CVPR.2019.00441]
- Li B, Yan J J, Wu W, Zhu Z and Hu X L. 2018. High performance visual tracking with siamese region proposal network//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE: 8971-8980 [DOI: 10.1109/CVPR.2018.00935]
- Loshchilov I and Hutter F. 2019. Decoupled weight decay regularization [EB/OL]. [2022-08-16] <https://arxiv.org/pdf/1711.05101.pdf>
- Meng L and Yang X. 2019. A survey of object tracking algorithms. Acta Automatica Sinica, 45(7): 1244-1260 (孟磊, 杨旭. 2019. 目标跟踪算法综述. 自动化学报, 45(7): 1244-1260) [DOI: 10.16383/j.aas.c180277]
- Mueller M, Smith N and Ghanem B. 2017. Context-aware correlation filter tracking//Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, USA: IEEE: 1387-1395 [DOI: 10.1109/CVPR.2017.152]
- Müller M, Bibi A, Giancola S, Alsubaihi S and Ghanem B. 2018. TrackingNet: a large-scale dataset and benchmark for object tracking in the wild//Proceedings of the 15th European Conference on Computer Vision. Munich, Germany: Springer: 300-317 [DOI: 10.1007/978-3-030-01246-5_19]
- Rezatofighi H, Tsoi N, Gwak J, Sadeghian A, Reid I and Savarese S. 2019. Generalized intersection over union: a metric and a loss for bounding box regression//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach,

- USA: IEEE: 658-666 [DOI: 10.1109/CVPR.2019.00075]
- Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez A N, Kaiser Ł and Polosukhin I. 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
- Vondrick C and Ramanan D. 2011. Video annotation and tracking with active learning//Proceedings of the 24th International Conference on Neural Information Processing Systems. Granada, Spain: Curran Associates Inc.: 28-36
- Vondrick C, Ramanan D and Patterson D. 2010. Efficiently scaling up video annotation with crowdsourced marketplaces//Proceedings of the 11th European Conference on Computer Vision. Heraklion, Greece: Springer: 610-623 [DOI: 10.1007/978-3-642-15561-1_44]
- 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 Z Q, Xu J, Liu L, Zhu F and Shao L. 2019. RANet: ranking attention network for fast video object segmentation//Proceedings of 2019 IEEE/CVF International Conference on Computer Vision. Seoul, Korea (South): IEEE: 3977-3986 [DOI: 10.1109/ICCV.2019.00408]
- Wu Y, Lim J and Yang M H. 2013. Online object tracking: a benchmark//Proceedings of 2013 IEEE Conference on Computer Vision and Pattern Recognition. Portland, USA: IEEE: 2411-2418 [DOI: 10.1109/CVPR.2013.312]
- Xu Y D, Wang Z Y, Li Z X, Yuan Y and Yu G. 2020. SiamFC++: towards robust and accurate visual tracking with target estimation guidelines. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(7): 12549-12556 [DOI: 10.1609/aaai.v34i07.6944]
- Yan B, Peng H W, Fu J L, Wang D and Lu H C. 2021a. Learning spatio-temporal transformer for visual tracking//Proceedings of 2021 IEEE/CVF International Conference on Computer Vision. Montreal, Canada: IEEE: 10428-10437 [DOI: 10.1109/ICCV48922.2021.01028]
- Yan B, Zhang X Y, Wang D, Lu H C and Yang X Y. 2021b. Alpha-refine: boosting tracking performance by precise bounding box estimation//Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA: IEEE: 5285-5294 [DOI: 10.1109/CVPR46437.2021.00525]
- Yuen J, Russell B, Liu C and Torralba A. 2009. LabelMe video: building a video database with human annotations//Proceedings of the 12th IEEE International Conference on Computer Vision. Kyoto, Japan: IEEE: 1451-1458 [DOI: 10.1109/ICCV.2009.5459289]
- Zhang Z P, Peng H W, Fu J L, Li B and Hu W M. 2020. Ocean: object-aware anchor-free tracking//Proceedings of the 16th European Conference on Computer Vision. Glasgow, UK: Springer: 771-787 [DOI: 10.1007/978-3-030-58589-1_46]
- Zhu J Z, Wang D and Lu H C. 2019. Learning background-temporal-aware correlation filter for real-time visual tracking. *Journal of Image and Graphics*, 24(4): 536-549 (朱建章, 王栋, 卢湖川. 2019. 背景与时间感知的相关滤波实时视觉跟踪. *中国图象图形学报*, 24(4): 536-549) [DOI: 10.11834/jig.180320]

作者简介

赵洁,女,博士研究生,主要研究方向为单目标视觉跟踪和计算机视觉。E-mail: zj982853200@mail.dlut.edu.cn

王栋,通信作者,男,教授,主要研究方向为视觉目标跟踪和目标检测。E-mail: wdice@dlut.edu.cn

袁永胜,男,硕士研究生,主要研究方向为单目标视觉跟踪和计算机视觉。E-mail: 1125298387@qq.com

张鹏宇,男,博士研究生,主要研究方向为单目标视觉跟踪和多模态融合。E-mail: pyzhang@mail.dlut.edu.cn