智能系统学报
CAAI TRANSACTIONS ON INTELLIGENT SYSTEMS

基于显著性导引孪生网络的红外船目标跟踪

李想, 张婷, 刘兆英, 刘波, 李玉鑑

引用本文:
李想, 张婷, 刘兆英, 等. 基于显著性导引孪生网络的红外船目标跟踪[J]. 智能系统学报, 2024, 19(6): 1428-1437.
LI Xiang, ZHANG Ting, LIU Zhaoying, et al. Infrared ship target tracking based on saliency guided siamese network[J]. *CAAI Transactions on Intelligent Systems*, 2024, 19(6): 1428-1437.

在线阅读 View online: https://dx.doi.org/10.11992/tis.202306004

DOI: 10.11992/tis.202306004
网络出版地址: https://link.cnki.net/urlid/23.1538.TP.20240709.1117.011

基于显著性导引孪生网络的红外船目标跟踪

李想[1], 张婷[1], 刘兆英[1], 刘波[1], 李玉鑑[2]
(1. 北京工业大学 信息学部, 北京 100124; 2. 桂林电子科技大学 人工智能学院, 广西 桂林 541004)

摘 要: 由于红外图像特征判别力低, 现有方法很难从背景中分割目标。而受到红外成像机制的影响, 红外目标通常具有较高的局部显著性, 因此本文提出一种基于显著性导引孪生网络的跟踪方法, 以目标的显著性信息为先验知识, 引导跟踪模型准确地定位目标。本文提出显著性预测网络和显著性增强网络。显著性预测网络用于获得搜索区域的全局显著性图, 并将其输入到显著性增强网络以增强目标, 提高模型的判别能力; 设计了一个共享互相关结构来计算模板图像特征与显著性增强后的搜索区域特征之间的相似度, 通过分类和回归两个任务共享互相关特征图, 同时提升模型的效率和性能; 由于目前缺少公开的红外船跟踪数据集, 本文构建了一个新的红外船目标跟踪数据集 (infrared ship dataset, ISD), 共包括 16 种不同类型的船, 7 800 幅带有标签的视频帧。在 ISD 上的实验结果显示, 与其他 18 个常用跟踪模型相比, 本模型达到了最高的准确率和最高的期望平均交并比。

关键词: 红外船跟踪; 孪生网络; 显著性目标检测; 特征融合; 共享互相关; 多任务学习; 卷积神经网络; 深度学习
中图分类号: TP391 **文献标志码**: A **文章编号**: 1673−4785(2024)06−1428−10

中文引用格式: 李想, 张婷, 刘兆英, 等. 基于显著性导引孪生网络的红外船目标跟踪 [J]. 智能系统学报, 2024, 19(6): 1428–1437.
英文引用格式: LI Xiang, ZHANG Ting, LIU Zhaoying, et al. Infrared ship target tracking based on saliency guided siamese network[J]. CAAI transactions on intelligent systems, 2024, 19(6): 1428–1437.

Infrared ship target tracking based on saliency guided siamese network

LI Xiang[1], ZHANG Ting[1], LIU Zhaoying[1], LIU Bo[1], LI Yujian[2]
(1. Faculty of Information Technology, Beijing University of Technology, Beijing 100124, China; 2. School of Artificial Intelligence, Guilin University of Electronic Technology, Guilin 541004, China)

Abstract: Infrared images often have features with low discriminative power, making it difficult to segment targets from the background using existing methods. Owing to the nature of infrared imaging, targets usually exhibit high local saliency. To address this, we propose a method for tracking infrared ship targets using saliency-guided Siamese networks (SGSiam). This approach uses the target's saliency as prior knowledge to guide the tracking model for precise target localization. First, this study presents a saliency prediction network and a saliency enhancement network. The saliency prediction network generates a global saliency map of the search area, which is input into the saliency enhancement network to strengthen the target and improve the discriminative ability of the model. Second, a shared cross-correlation architecture is designed to calculate the similarity between the template image features and the saliency-enhanced search region features, thus improving the model efficiency and performance through shared feature maps for classification and regression tasks. Finally, owing to the lack of publicly available infrared ship tracking data sets, we introduce a new infrared ship data set (ISD), which includes 16 different ship types and 7 800 video frames with manual annotations. Experimental results on ISD show that our model outperforms 18 commonly used tracking models, achieving the highest accuracy and the highest expected average overlap score.

Keywords: infrared ship tracking; siamese network; salient object detection; feature fusion; shared correlation; multi-task learning; convolutional neural network; deep learning

收稿日期: 2023−06−01. **网络出版日期**: 2024−07−11.
基金项目: 国家自然科学基金区域创新发展联合基金项目 (U23A20357); 北京市教育委员会科技计划一般项目 (KM202110005028); 北京工业大学交叉科学研究院资助项目 (2021020101); 北京工业大学国际科研合作种子基金项目 (2021A01).
通信作者: 刘兆英. E-mail: zhaoying.liu@bjut.edu.cn.

红外船目标跟踪是计算机视觉领域中的一项基本问题, 其在军用和民用领域都有广泛的应用[1], 吸引了越来越多的关注[2-5]。近年来, 基于孪生网络的跟踪器在目标跟踪领域取得了很大的进

展[6-7]。Bertinetto 等[6] 提出全卷积孪生网络 (fully-convolutional siamese networks, SiamFC), 通过训练全卷积孪生网络来计算区域特征的相似度, 从而实现快速准确的跟踪。Zhang 等[7] 提出更深层的、更宽的孪生网络 (deeper and wider siamese networks, SiamDW), 在孪生网络中采用更深更宽的骨干网络来获得更加准确的跟踪结果。虽然这些跟踪器有效提升了目标跟踪的性能, 但是大部分是针对可见光彩色图像, 与可见光图像相比, 红外图像缺少目标的颜色信息, 类间差异小, 直接将这些模型应用到红外领域会导致跟踪器提取的特征判别能力低, 进而使跟踪器容易受到干扰物的影响。因此, 为了提升模型的判别能力, Li 等[2] 提出层级空间感知孪生网络 (hierarchical spatial-aware siamese network, HSSNet), 通过融合多层卷积特征, 获得目标的空间和语义特征。Chen 等[3] 提出泛化友好型孪生网络 (generalization-friendly siamese network, GFSNet), 在分类分支和回归分支中分别插入带有通道注意力的分类适应模块和带有空间注意力的位置适应模块, 提高模型的泛化能力。然而, 这些模型没有考虑到目标的显著性信息。另外, 有一些工作利用了目标的掩码标签来提升跟踪性能[1,8]。Wang 等[8] 提出目标跟踪和分割的统一架构, 在一般的孪生网络基础上增加一个与分类分支和回归分支平行的掩码预测分支, 通过精心设计标签, 以一种密集预测的方式实现目标的分割。Yang 等[1] 在文献 [8] 的基础上, 引入特征金字塔网络[9](feature pyramid networks, FPN) 用于解决目标多尺度预测问题, 同时提出一种多维注意力模块, 通过在长、宽和通道 3 个维度同时学习, 使网络更加关注目标物体, 抑制干扰物。虽然这些算法都达到了令人满意的效果, 但是 3 个分支之间没有交流, 各自独立。

考虑到红外图像目标往往呈现高的局部显著性, 本文提出一种基于显著性导引孪生网络的红外船跟踪器 SGSiam, 通过检测视频帧中的显著性目标, 来使跟踪器产生准确的检测框。本文首先提出一个显著性预测网络用于产生目标的显著性信息, 接着, 通过一个精心设计的显著性增强网络将显著性信息与目标特征图融合, 以突出目标区域、抑制背景区域。其次, 本文引入一种共享互相关结构, 通过分类分支和回归分支共享互相关图的方式, 在减少计算量的同时, 提升模型的整体性能。为了训练本文提出的模型, 本文构建了一个新的红外船跟踪数据集 ISD, 包含 16 种不同类型的船, 共 7800 多带有标签的视频帧。

1　相关工作

近几年, 由于在速度和准确率之间实现了平衡, 基于孪生网络的跟踪器成为研究的热点。基于孪生网络的算法在大量图像对上训练一种相似性函数。在测试阶段, 在新的视频序列上对该相似性函数进行评估。SiamFC[6] 使用孪生网络作为特征提取器, 并引入相关滤波层来计算模板图像与搜索区域之间的相似度。Li 等[10] 在孪生网络之后引入区域生成网络[11](region proposal network, RPN) 来精准地预测目标的位置。为了解决浅层特征提取网络 AlexNet[12] 对于跟踪器准确率的限制, Li 等[13] 将深度网络 ResNet50[14] 引入孪生网络中并消除了填充对于模型性能的影响。此外, 还有一些工作研究模板的更新[15]、分类−回归结果的不一致[16] 以及优化的互相关操作[17]。这些跟踪都采用基于锚框的方式生成候选框, 这种方式需要目标物体的先验知识, 同时会引入一些超参数, 例如锚框的尺寸、长宽比等。然而视觉目标跟踪是一项类别未知的任务, 基于锚框的方式会破坏跟踪器的泛化性能, 所以无锚框的方式被引入到跟踪领域中[18-19]。Xu 等[18] 以特征图中的每一个像素点为训练样本, 并设计两个分支实现预测任务, 其中一个分支预测样本的置信度, 另一个分支直接回归样本与真实框四条边之间的距离。Li 等[19] 提出孪生关键点预测网络 (siamese keypoint prediction network, SiamKPN), 通过预测目标的尺度、中心点坐标以及误差来实现跟踪。虽然这些跟踪器能够实现良好的性能, 但是由于成像原理不同, 这些针对可见光图像的模型无法直接应用到红外图像中。

相比于视觉目标跟踪来说, 由于缺少大规模公开数据集, 红外目标跟踪的发展相对慢一些, 但是依然有一些出色的工作[20-23]。Li 等[20] 提出孪生多组空间移位网络 (siamese multigroup spatial shift network, SiamMSS), 通过一个空间移位模块来增强特征图的细节信息, 并通过切分注意力模块对互相关特征图进行融合来实现跟踪。Liu 等[21] 通过卷积神经网络提取目标的多层特征来构建多个弱跟踪器, 每个跟踪器输出一个目标位置的响应图, 最后将多个响应图进行集成得到最终的跟踪结果。Liu 等[22] 将浅层特征送入结构互相关相似度模块用于目标的定位, 将深层特征送入语义互相关相似度模块用于区分干扰物, 同时实现了精度和鲁棒性的提升。为了提升跟踪器的判别能力, Liu 等[23] 提出细粒度感知网络 (fine-

grained aware network, FANet), 其包含一个全局互相关模块用于捕获局部区域之间的联系和一个像素级互相关模块用于捕获不同像素位置之间的联系。

以上针对红外目标跟踪的方法尽管取得了一定的结果, 但它们在提取红外图像特征的过程中依赖于复杂的注意力模块和多层特征融合模块, 没有考虑目标的显著性信息。因此, 本文将显著性检测引入到跟踪器中, 利用目标的显著性信息增强目标特征, 提升模型的判别力。

2　显著性导引孪生网络跟踪器

SGSiam 的整体模型框架如图 1 所示, 该模型由特征提取网络、显著性预测网络、显著性增强网络以及跟踪预测网络 4 部分组成。特征提取网络用于将输入图像对嵌入到同一个特征空间中; 显著性预测网络采用自底向上的方式生成全局显著性图, 得到目标的显著性信息; 显著性增强网络采用自顶向下的方式增强特征图, 提高特征判别力; 跟踪预测网络采用分类−回归两任务共享相关特征的方式预测目标的状态。

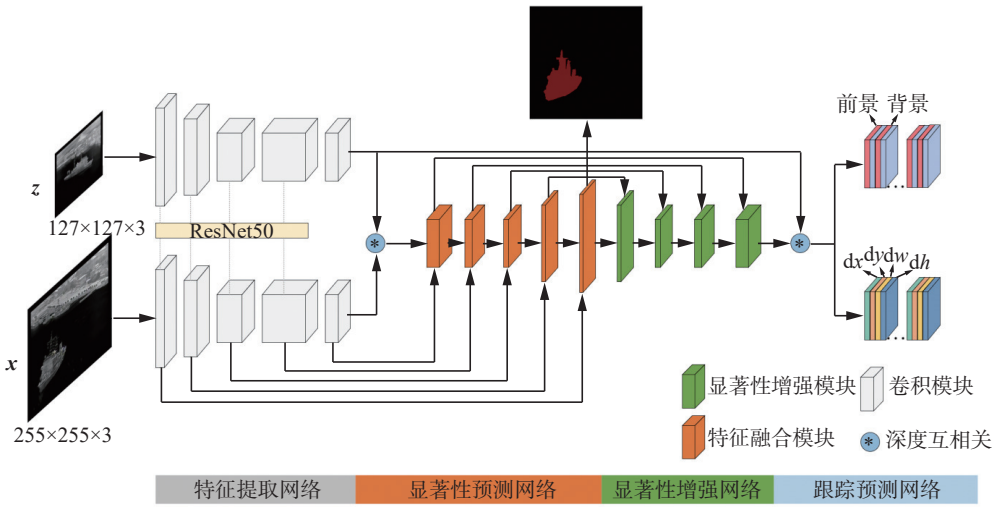


图 1　SGSiam 整体结构

Fig. 1　Overall architecture of the proposed SGSiam

2.1　特征提取网络

与文献 [10] 相比, 本文选择 ResNet50[14] 作为主干网络, 为了增强网络的表达能力从而更好地对目标特征进行建模。特征提取网络将输入的模板图像 $\boldsymbol{z}\in\mathbf{R}^{H_z\times W_z\times C}$ 与搜索区域 $\boldsymbol{x}\in\mathbf{R}^{H_x\times W_x\times C}$ 嵌入到相同的特征空间, 得到模板图像特征 $\varphi_5(z)$ 和搜索区域特征 $\varphi_5(x)$ (其中 $\varphi_i(\cdot)$ 表示特征提取网络第 i 层的输出, i=1,2,3,4,5)。

2.2　显著性预测网络

输入图像经过特征提取网络编码后, 拥有低的分辨率。因此, 本文构建了显著性预测网络对特征图进行上采样, 并设计了两种特征融合模块分别用于抑制干扰物和细化特征, 如图 2(a) 和图 2(b) 所示。其中图 2(a) 是负责抑制干扰物, 突出目标物体的特征融合模块, 其通过深度互相关模块计算模板特征 $\varphi_5(z)$ 和搜索区域特征 $\varphi_5(x)$ 之间的相似度, 并将相似度图与 $\varphi_5(x)$ 按通道拼接, 后接 1 个 3×3 卷积融合特征; 图 2(b) 是负责细化特征图的特征融合模块, 每个特征融合模块 A_i 的输入由前一个特征融合模块 A_{i-1} 的输出和来自特征提取网络对应层的输出 $\varphi_{6-i}(\cdot)$ 组成。具体来说, 前一层特征融合模块的输出经过上采样增大分辨率, 后接两个 3×3 卷积, 与来自特征提取网络对应层的输出 (经一个 1×1 卷积进行适应性的调整) 按像素位置相加。最后一个特征融合模块的输出在上采样后经过一个显著性预测头部, 然后紧跟一个 softmax 层产生一个与搜索图像相同大小的显著性概率图 $\boldsymbol{P}_{\text{sal}}\in\mathbf{R}^{H_x\times W_x\times 1}$。

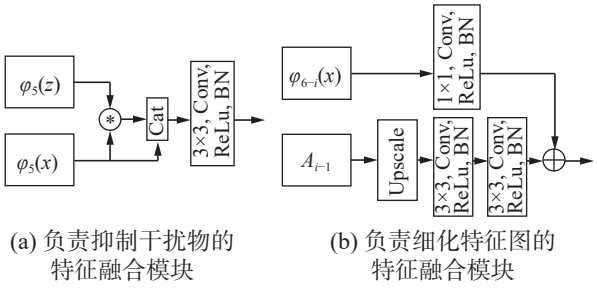


图 2　两种特征融合模块结构

Fig. 2　Architecture of the two feature integration module

两种特征融合模块的计算过程为

$$A_i=\begin{cases}\mathrm{Conv}(\mathrm{Cat}(\mathrm{DWCorr}(\varphi_5(z),\varphi_5(x)),\varphi_5(x))), i=1\\ \mathrm{Conv}(\varphi_{6-i}(x))+\mathrm{Conv}(\mathrm{Conv}(A_{i-1})), i=2,3\\ \mathrm{Conv}(\varphi_{6-i}(x))+\mathrm{Conv}(\mathrm{Conv}(\mathrm{Up}(A_{i-1}))), i=4,5\end{cases} \tag{1}$$

式中: Conv() 表示卷积操作, Cat() 表示按通道拼接, Up() 表示上采样层, DWCorr() 表示深度互相关层。

2.3　显著性增强网络

虽然显著性预测网络可以将目标分割出来, 但是包裹住目标的最小包围框不能直接当作跟踪的结果。因为当目标遇到遮挡时, 最小包围框只能表示目标未被遮挡的区域, 不能表示目标的实际大小, 因此, 需要将得到显著性信息进行回传, 用于增强搜索帧特征。显著性增强后的特征更加突出目标物体的形状, 使得特征更加具有判别力, 可以辅助跟踪器定位目标。受到路径聚合网络 (path aggregation network, PANet)[24] 的启发, 为了减少信息的丢失, 同时将显著性信息与搜索特征进行充分地融合, 本文设计了一条比特征提取网络更短的路径, 该路径由 4 个阶段的显著性增强模块组成, 如图 3 所示。与文献 [24] 相比, 不同是, 本文在融合的过程中引入了显著性信息, 能够突出目标区域同时抑制背景区域。除第一个外, 每个显著性增强模块 E_i 的输入由前一个显著性增强模块 E_{i-1} 的输出和对应特征融合模块的输出组成。具体来说, 前一个显著性增强模块的输出经过一个下采样模块和一个 3×3 卷积后与对应特征融合模块的输出 (经一个 1×1 卷积) 相加。

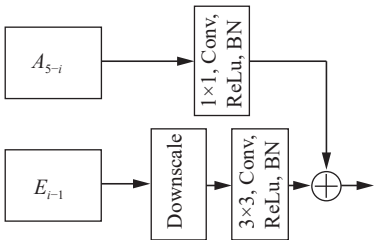


图 3　显著性增强模块结构

Fig. 3　Architecture of the saliency enhancement module

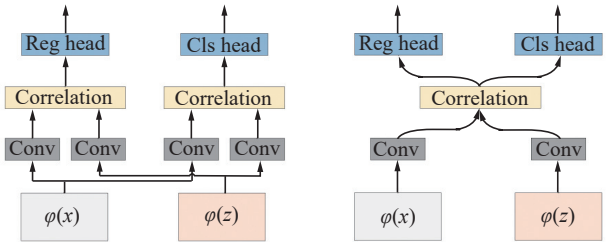
整个过程为

$$E_i=\begin{cases}\mathrm{Conv}(\mathrm{Down}(A_5))+\mathrm{Conv}(A_{5-i}), & i=1\\ \mathrm{Conv}(\mathrm{Down}(E_{i-1}))+\mathrm{Conv}(A_{5-i}), & i=2\\ \mathrm{Conv}(E_{i-1})+\mathrm{Conv}(A_{5-i}), & i=3,4\end{cases} \tag{2}$$

其中 Down() 表示下采样层。

2.4　跟踪预测网络

与文献 [10] 一样, 本文的跟踪预测网络由分类和回归分支组成。分类分支负责前景−背景分类, 回归分支负责候选框的回归。一般的基于锚框的跟踪器会对每一个分支分配一个单独的互相关操作, 如图 4(a) 所示。



(a) 非共享的互相关结构　　(b) 共享的互相关结构

图 4　两种互相关结构

Fig. 4　Architecture of the two correlation module

输入包括搜索区域特征 $\varphi(x)$ 和模板特征 $\varphi(z)$, 具体细节为

$$\begin{aligned}S_{\mathrm{cls}}&=\mathrm{DWCorr}(\mathrm{Conv}^1_{\mathrm{cls}}(\varphi(x)),\ \mathrm{Conv}^2_{\mathrm{cls}}(\varphi(z)))\\ S_{\mathrm{reg}}&=\mathrm{DWCorr}(\mathrm{Conv}^1_{\mathrm{reg}}(\varphi(x)),\ \mathrm{Conv}^2_{\mathrm{reg}}(\varphi(z)))\\ P_{\mathrm{cls}}&=f_{\mathrm{cls}}(S_{\mathrm{cls}})\\ P_{\mathrm{reg}}&=f_{\mathrm{reg}}(S_{\mathrm{reg}})\end{aligned} \tag{3}$$

式中: S_{cls}、S_{reg} 表示相似度图, P_{cls} 和 P_{reg} 分别表示分类和回归分支的预测结果, f_{cls} 和 f_{reg} 分别表示分类和回归的预测函数。这种非共享互相关特征的方式会造成计算量增大, 所以本文引入一种共享的互相关结构[25] 提升计算效率, 如图 4(b) 所示。共享的相似度图分别送入分类预测头部和回归预测头部来产生目标框, 具体细节为

$$\begin{aligned}S&=\mathrm{DWCorr}(\mathrm{Conv}(\varphi(x)),\mathrm{Conv}(\varphi(z)))\\ P_{\mathrm{cls}}&=f_{\mathrm{cls}}(S)\\ P_{\mathrm{reg}}&=f_{\mathrm{reg}}(S)\end{aligned} \tag{4}$$

2.5　损失函数

本文所提出的 SGSiam 网络的损失函数由 3 部分组成, 分别是分类损失、回归损失和显著性预测损失。分类损失 L_{cls} 采用交叉熵损失:

$$L_{\mathrm{cls}}(c,c^*)=\frac{1}{2}\sum_i\sum_j\sum_k\mathrm{BCELoss}(c_{i,j,k},c^*_{i,j,k}) \tag{5}$$

$$\mathrm{BCELoss}(u,u^*)=-u^*\log u-(1-u^*)\log(1-u) \tag{6}$$

式中: $c_{i,j,k}$ 和 $c^*_{i,j,k}$ 分别表示预测结果 P_{cls} 中 (i,j) 位置上第 k 个锚框的分类预测结果和真值。回归损失 L_{reg} 采用 L_1 损失:

$$L_{\mathrm{reg}}(r,r^*)=\frac{1}{N_{\mathrm{pos}}}\sum_i\sum_j\sum_k[c^*>0]\|\delta(r_{i,j,k},r^*_{i,j,k})\|_1 \tag{7}$$

其中 N_{pos} 表示正样本的数目。$[c^*>0]$ 为指示器函数, 当满足条件时输出 1, 否则输出 0。$r_{i,j,k}=(x,y,w,h)$ 和 $r^*_{i,j,k}=(x^*,y^*,w^*,h^*)$ 分别表示预测的框和真值框。(x,y) 和 (x^*,y^*) 表示框的中心坐标, (w,h) 和 (w^*,h^*) 表示框的宽和高。δ 表示正则化的距离:

$$\begin{aligned}\delta(r_{i,j,k},r^*_{i,j,k})=&((x^*-x)/x,(y^*-y)/y,\\ &\ln(w^*/w),\ln(h^*/h))\end{aligned} \tag{8}$$

显著性预测损失 L_{sal} 也采用交叉熵损失:

$$L_{\mathrm{sal}}(s,s^{*})=\frac{1}{N}\sum_{i}\sum_{j}\mathrm{BCELoss}(s_{i,j},s_{i,j}^{*}) \tag{9}$$

$s_{i,j}$与$s_{i,j}^{*}$分别表示在(i,j)位置处的显著性预测结果和真值，N表示样本总数。该模型的总体损失函数为

$$L_{\mathrm{oss}}=\lambda_{\mathrm{cls}}L_{\mathrm{cls}}+\lambda_{\mathrm{reg}}L_{\mathrm{reg}}+\lambda_{\mathrm{sal}}L_{\mathrm{sal}} \tag{10}$$

式中λ_{cls}、λ_{reg}、λ_{sal}分别表示对应部分的权重。

3　实验结果与分析

3.1　数据集与评价指标

由于缺少公开的红外船目标跟踪数据集，本文构建了一个新的红外船数据集 ISD 来训练 SGSiam。ISD 总共有 7 800 多带有标签的视频帧，包含 16 个视频段，分别对应 16 个类别，即 hwj、jyj、kcj、lqt、myyyc、qt、qwc、qzj、slj、tc_qzc_sag、yyc、hj、kt、xlt、yc、yl。在实验中，本文随机选择 8 个视频用于训练，其余 8 个视频用于测试。数据集的具体描述见表 1。除此之外，为了更加充分验证 SGSiam 的泛化性能，本文在公开的红外行人数据集 PTB-TIR[26] 上做了对比实验。PTB-TIR 是最近发表的用于在红外行人场景下评估模型性能的数据集，总共包含 60 个视频。

表 1　ISD 数据集的详细情况

Table 1　A Detailed Description of the ISD

类别	帧数	分辨率/(像素×像素)	大小/MB	类别	帧数	分辨率/(像素×像素)	大小/MB
hwj	614	256×256	2.7	jyj	597	256×256	6.0
myyyc	306	256×256	5.2	qt	641	256×256	3.3
slj	592	256×256	3.6	tc_qzc_sag	600	256×256	7.0
kt	117	1 920×1 280	18.9	xlt	145	1 920×1 280	27.0
kcj	600	256×256	4.6	lqt	600	256×256	4.8
qwc	615	256×256	6.1	qzj	600	256×256	5.4
yyc	658	256×256	4.7	hl	127	1 920×1 280	22.5
yc	439	1 920×1 280	152.9	yl	621	1 920×1 280	122.2

为了评价本文跟踪器的性能，本文选择准确率、鲁棒性、期望平均交并比、参数量和计算量 5 个指标。准确率 (A_{c}) 用于计算在所有跟踪成功的帧中目标框与预测框之间的平均交并比：

$$A_{\mathrm{c}}=\frac{1}{N}\sum_{i=1}^{N}\phi_i,\quad \phi_i=\frac{B_i^{\mathrm{G}}\cap B_i^{\mathrm{P}}}{B_i^{\mathrm{G}}\cup B_i^{\mathrm{P}}} \tag{11}$$

式中：N表示总帧数，B_i^{G}和B_i^{P}分别表示第 i 帧的真实目标框和预测目标框，ϕ_i表示第 i 帧的交并比。鲁棒性 (R_{o}) 表示跟踪失败 (目标框与预测框之间的交并比为 0) 的次数占帧总数的比例：

$$R_{\mathrm{o}}=\frac{N_{\mathrm{f}}}{N},\quad N_{\mathrm{f}}=\sum_{i=1}^{N}[\phi_i\leqslant 0] \tag{12}$$

式中：N_{f}表示跟踪失败的次数；[·]是一个指示器函数，当满足条件时，输出 1，否则输出 0。期望平均交并比 (E_{ao}) 同时考虑准确率和鲁棒性，可以衡量跟踪器的整体性能：

$$E_{\mathrm{ao}}=\frac{1}{N_{\mathrm{l}}-N_{\mathrm{h}}+1}\sum_{N_{\mathrm{s}}=N_{\mathrm{l}}}^{N_{\mathrm{h}}}\frac{1}{N_{\mathrm{s}}}\sum_{i=1}^{N_{\mathrm{s}}}\phi_i \tag{13}$$

式中：N_{l}为起始帧的位置，N_{h}为结束帧的位置。参数量衡量模型中可学习的参数的总数量，计算量衡量模型每秒完成的浮点数运算的次数。

在 PTB-TIR 中，本文使用成功率和精确度对模型的性能进行评估。成功率 (S^{r}) 表示交并比大于给定阈值的帧的数量占总帧数的比例：

$$S_{\Phi}^{\mathrm{r}}=\frac{1}{N}\sum_{i=1}^{N}[\phi_i>\Phi] \tag{14}$$

其中Φ表示阈值。精确度 (P^{r}) 表示预测框中心点与目标框中心点之间的欧氏距离大于给定阈值的帧的数量占总帧数的比例：

$$\varphi_i=\sqrt{(x_i^{\mathrm{P}}-x_i^{\mathrm{G}})^2+(y_i^{\mathrm{P}}-y_i^{\mathrm{G}})^2} \tag{15}$$

$$P_{\Psi}^{\mathrm{r}}=\frac{1}{N}\sum_{i=1}^{N}[\varphi_i>\Psi] \tag{16}$$

式中：Ψ是给定的阈值，$(x_i^{\mathrm{P}},y_i^{\mathrm{P}})$和$(x_i^{\mathrm{G}},y_i^{\mathrm{G}})$分别是预测框和目标框的中心点。

3.2　实验细节

为了适应模型输入要求，首先将模板图像和搜索区域分别缩放到 127×127 和 255×255 像素。整个网络在 2 个 GPUs 上一共训练 20 轮，每一轮随机采样 6×10^{4} 个图像对，每 28 个图像对组成一个批。前 5 轮是热身阶段，学习率从 5×10^{-3} 线性增长到 1×10^{-2}，后 15 轮从 1×10^{-2} 呈对数下降到

5×10^{-4}。网络的 backbone 使用在 ImageNet 上预训练的参数进行初始化, 并且在前 10 轮中冻结参数, 在后 10 轮中对参数进行微调。通过 SGD, 其中权重衰减和动量被设置为 1×10^{-4} 和 0.9, 优化式 (10) 中的损失函数 (λ_{cls}=1, λ_{reg}=1, λ_{sal}=1.2), 得到整体网络的参数。

在推理阶段, 目标模板特征只在第 1 帧中计算一次并保存在内存中, 用于和后续的图像进行匹配。实验环境为带有 NVIDIA Tesla K40c GPUs 的 linux 服务器, 利用 Python 3.6 和 Pytorch 0.4.1 构造跟踪器。

3.3　消融实验

为了验证本文所提出的显著性预测网络、显著性增强网络以及共享互相关模块对于提升红外目标跟踪性能的作用, 本文在 ISD 数据集上开展了消融实验, 结果如表 2 所示。第 1 行表示 SiamRPN[10](baseline) 的实验结果, 第 2 行表示用深层网络 ResNet50 作为主干网络的实验结果, 第 3 行表示仅添加共享互相关的实验结果, 第 4 行表示仅使用显著性预测网络和显著性增强网络的实验结果, 最后 1 行表示同时添加 3 种组成部分后的实验结果。

表 2　在 ISD 数据集上消融实验的结果

Table 2　Ablation study on the proposed ISD

Res	Sal	SCM	参数量/10^6	计算量/10^9	期望平均交并比↑	准确率↑	鲁棒性↓
			6.25	5.57	0.268	0.630	0.498
√			16.55	18.93	0.526	0.705	**0.133**
√		√	15.37	18.42	0.533	0.697	0.199
√	√		20.98	24.38	0.499	0.738	0.199
√	√	√	19.80	23.87	**0.674**	**0.757**	0.166

注: 加黑代表最优结果。

从表 2 中可以看出:

1) 当使用深层的 ResNet50[14] 替换浅层的 AlexNet[12] 时, 3 个评价指标都有较大的提升, 这说明深层网络具有更强大的特征提取能力;

2) 在仅使用共享互相关模块之后, 准确率下降了 0.8%, 从 70.5% 到 69.7%, 然而 EAO 上升了 0.7%, 从 52.6% 到 53.3%。这意味着共享互相关模块对跟踪过程中误差的积累不敏感, 能够提高模型的稳定性。

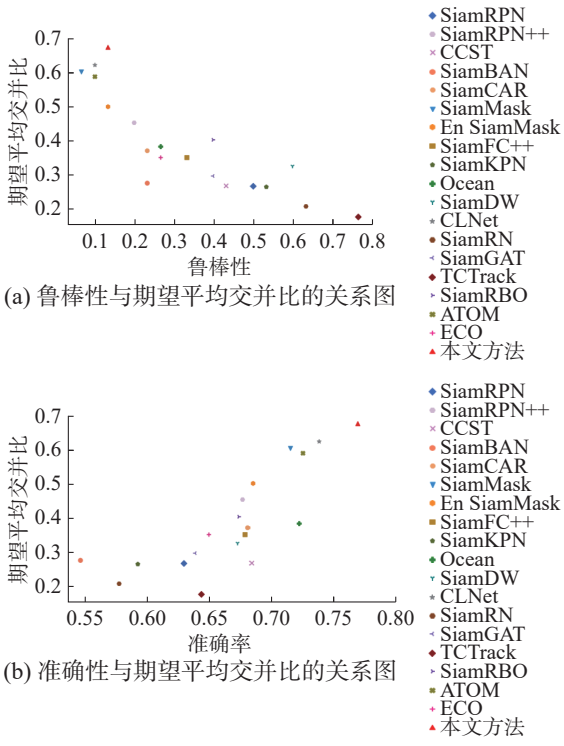
3) 当仅使用显著性预测网络和显著性辅助网络时, 预测准确率上升了 3.2%, 从 70.5% 到 73.8%, 这表示显著性的引入能够辅助目标的定位。

4) 当同时使用共享互相关模块、显著性预测网络和显著性辅助网络时, EAO 上升了 14.8%, 从 52.6% 到 67.4%, 准确率上升了 5.2%, 从 70.5% 到 75.7%。由于使用了共享互相关结构, 模型的参数量和计算量有所下降。

5) 总之, 综合使用显著性预测网络、显著性辅助网络以及共享互相关对于全面提升跟踪器的性能具有重要的作用, 三者缺一不可。

3.4　对比实验

为了进一步验证本模型 SdSiam 的优越性, 本文在 ISD 数据集上与其他 18 个跟踪模型进行了对比实验: EnSiamMask[1]、SiamDW[7]、SiamMask[8]、SiamRPN[10]、SiamRPN++[13]、Ocean[16]、SiamGAT[17]、SiamFC++[18]、SiamKPN[19]、CCST[27]、SiamBAN[28]、SiamCAR[29]、CLNet[30]、SiamRN[31]、TCTrack[32]、SiamRBO[33]、ATOM[34]、ECO[35]; 对比实验结果如图 5 所示。



(a) 鲁棒性与期望平均交并比的关系图

(b) 准确性与期望平均交并比的关系图

图 5　在 ISD 数据集上与现有模型对比实验的结果

Fig. 5　Comparison with start-of-the-art on ISD

图 5 给出了 EAO 与准确率和鲁棒性的关系。从图 5 可以看出：本文提出的 SGSiam 获得了最高的 EAO 和准确率。得益于滤波器的更新，ATOM 实现了令人满意的效果，同时获得较低的推理速度。SiamMask 与 EnSiamMask 使用了额外的 mask 标签，在性能上超过了大部分的跟踪器，但是未考虑目标的浅层信息，这会降低分割效果。由于 CLNet 在第 1 帧提取视频序列的相关信息，这不可避免引入了大量的计算。本文的 SGSiam 模型充分融合了目标的浅层信息预测显著性图，并与多层特征图进行融合来增强目标特征，有效提升了模型的性能。

为了更进一步地评估本文提出的 SGSiam，表 3 给出了在 8 个视频序列中，本模型与其他 18 个跟踪器在准确率上的比较。SGSiam 在 4 个视频中，即 myyyc、tc_qzc_sag、xlt 和 yyc，获得了最高的准确率。具体来说，在 myyyc 中，SGSiam 获得了 0.895 的准确率，比第 2 名的 SiamMask 高了 6.5%。在 tc_qzc_sag 上，SGSiam 获得了 0.715 的准确率，这比第 2 名的 Ocean 高了 2.6%。在 xlt 上，SGSiam 获得了 0.684 的准确率，比次优的 SiamGAT 的 0.657 略高一点。在 yyc 上，SGSiam 比最好的 CCST 还要高出 3.2%。在其余视频中，与最好的模型相比，SGSiam 获得了可比的性能，并且具有速度上的优势。这说明本文提出的方法的综合性能超出了其余方法，验证了该方法的有效性。

表 3　在不同视频序列上与现有模型在准确率上对比实验的结果

Table 3　Comparison with start-of-the-art on different videos in terms of accuracy

模型	hl	kt	myyyc	tc_qzc_sag	xlt	yc	yl	yyc
EnSiamMask	0.730	0.486	0.823	0.687	0.605	0.493	0.693	0.782
SiamDW	0.854	0.760	0.691	0.452	0.601	0.731	0.779	0.693
SiamMask	0.773	0.501	0.830	0.621	0.613	0.593	0.805	0.789
SiamRPN	0.641	0.427	0.768	0.542	0.532	0.437	0.689	0.765
SiamRPN++	0.673	0.399	0.786	0.656	0.542	0.466	0.752	0.790
Ocean	0.789	0.762	0.758	0.689	0.629	0.635	0.805	0.714
SiamGAT	0.685	0.622	0.768	0.549	0.657	0.549	0.660	0.690
SiamFC++	0.638	0.635	0.752	0.668	0.547	0.483	0.793	0.718
SiamKPN	0.657	0.641	0.730	0.623	0.509	0.538	0.417	0.698
CCST	0.674	0.556	0.788	0.594	0.509	0.474	0.796	0.802
SiamBAN	0.680	0.689	0.708	0.596	0.577	0.622	0.097	0.731
SiamCAR	0.549	0.574	0.787	0.521	0.505	0.610	0.840	0.759
CLNet	**0.897**	0.597	0.826	0.578	0.635	**0.743**	**0.864**	0.736
SiamRN	0.747	0.702	0.740	0.562	0.505	0.635	0.292	0.708
TCTrack	0.710	0.338	0.757	0.593	0.550	0.484	0.672	0.758
SiamRBO	0.733	0.444	0.790	0.672	0.507	0.481	0.684	0.800
ATOM	0.736	0.627	0.783	0.626	0.630	0.698	0.823	0.749
ECO	0.870	**0.770**	0.665	0.384	0.672	0.696	0.815	0.633
本文	0.774	0.513	**0.859**	**0.715**	**0.684**	0.627	0.815	**0.834**

注：加黑代表最优结果，加下划线代表次优结果。

为了更直观地展示本模型的跟踪效果，图 6 给出了 4 个具有挑战性的视频中本方法与 4 个性能较好的跟踪器的视觉对比结果。从图 6(a) 和图 6(b) 可以看出，当目标尺度变化剧烈时，其他的跟踪器会出现飘移的现象，而 SGSiam 可以准确地预测目标的状态。除了尺度变化，当出现目标超出视野范围的时候，本文的跟踪器依然是最准确的。此外，在图 6(b) 中，由于船上有一条细吊杆的存在，在深度卷积神经网络不断下采样的过程中，吊杆的信息会丢失，导致吊杆未出现在预测框中，而 SGSiam 充分利用浅层的信息，保留吊杆的特征，获得了更准确的预测框。在 yc 中，SiamRPN 与 SiamMask 对背景扰动更加敏感，本文的跟踪器由于使用了显著性增强后的特征而更加关注目标物体，实现了能与 CLNet 竞争的效果。当遇到目标发生平面内旋转的时候，如第

4 行所示, 本文提出的模型依然能准确地预测目标的状态。除此之外, 本文针对推理速度也做了实验分析, 如表 4 所示。相比于 SiamRPN++, SiamBAN、SiamCAR 等深度跟踪器本文的方法依然实现了令人满意的效果。此外, SGSiam 的速度要远远高于基于相关滤波器的跟踪器, 例如 ECO 和 ATOM。相比于 SiamFC++和 SiamRBO 等跟踪器, SGSiam 依然获得了可比的推理速度。

为了更充分地证明本文所提出的算法的有效性, 本文在 PTB-TIR 数据集上与现有的 9 个模型做了对比实验, 即 MLSSNet[22]、CREST[36]、UDT[37]、MCFTS[21]、HSSNet[2]、HDT[38]、HCF[39]、CFNet[40]、SiamFC[6], 结果如表 5 所示。具体来说, SGSiam 在两个评价指标中都获得了最优异的性能, 超过了所有目前流行的跟踪算法。本文提出的模型获得了 0.577 的成功率, 超过了最近刚刚提出的 MLSSNet, 实现了 3.8% 的提升。相比于其他的算法, SGSiam 获得了更大的提升。从预测的角度看, SGSiam 获得了 0.757 的得分, 这比第 2 名 MLSSNet 高了 1.6%, 比第 3 名高了 4.6%。上述结果都直接地证明了本文方法拥有很强的泛化能力。

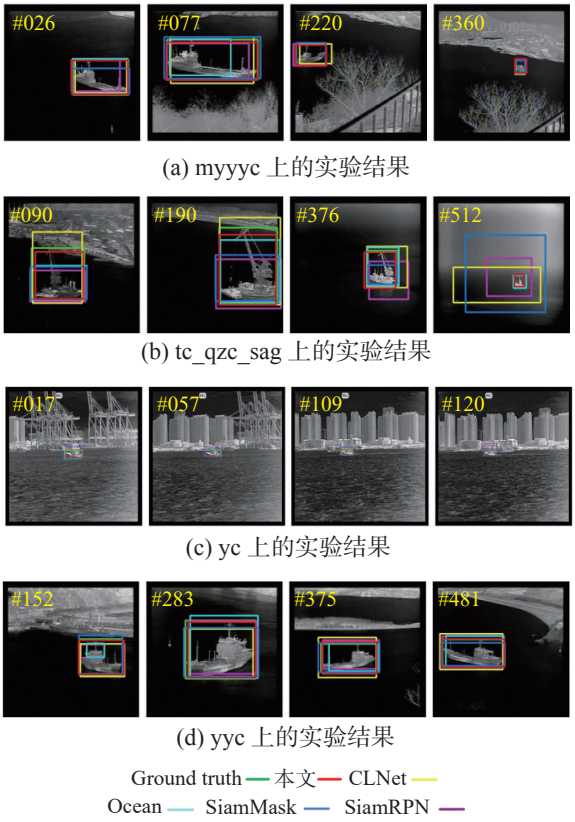


图 6　在 ISD 数据集上定性实验的结果

Fig. 6　Qualitative results on ISD dataset

表 4　SGSiam 与其他 18 个模型在推理速度上的实验结果

Table 4　Experimental results with other 18 models in terms of inference speed

模型	速度/(f/s)	模型	速度/(f/s)	模型	速度/(f/s)	模型	速度/(f/s)
SiamRPN	**70.2**	SiamRPN++	11.2	CCST	65.6	SiamRBO	20.3
SiamBAN	11.4	SiamCAR	11.1	SiamMask	23.1	ATOM	8.7
EnSiamMask	17.1	SiamFC++	21.8	SiamKPN	7.3	ECO	2.0
Ocean	16.0	SiamDW	25.3	CLNet	12.7	本文	19.2
SiamRN	3.4	SiamGAT	19.1	TCTrack	30.4	—	—

注: 加黑代表最优结果。

表 5　在 PTB-TIR 上的对比实验的结果

Table 5　Comparison results on the PTB-TIR dataset

性能	MLSSNet	CREST	UDT	MCFTS	HSSNet	HDT	HCF	CFNet	SiamFC	本文
成功率	0.539	0.524	0.529	0.492	0.468	0.457	0.448	0.449	0.480	**0.577**
精确度	0.741	0.711	0.699	0.690	0.689	0.687	0.671	0.629	0.623	**0.757**

注: 加黑代表最优结果。

4　结束语

本文提出了一种红外船目标跟踪模型 SGSiam, 将显著性目标检测融入到现有的跟踪模型中, 来提升跟踪的准确率。SGSiam 使用一个显著性预测网络用于获得全局显著性图, 为跟踪器提供目标的显著性信息; 一个显著性增强网络将显著性信息与搜索区域特征相融合增强目标特征, 提升模型的判别能力; 一个共享的互相关模块减少训练开销, 提升模型整体性能。在红外船目标跟踪数据集 ISD 和 PTB-TIR 上开展的大量实验结果表明, 本文提出的跟踪器可以有效提升红外船目标跟踪的性能。

参考文献:

[1] YANG Xi, WANG Yan, WANG Nannan, et al. An enhanced SiamMask network for coastal ship tracking[J]. IEEE transactions on geoscience and remote sensing, 2022, 60: 5612011.

[2] LI Xin, LIU Qiao, FAN Nana, et al. Hierarchical spatial-aware Siamese network for thermal infrared object tracking[J]. Knowledge-based systems, 2019, 166: 71−81.

[3] CHEN Ruimin, LIU Shijian, MIAO Zhuang, et al. GFS-Net: generalization-friendly Siamese network for thermal infrared object tracking[J]. Infrared physics and technology, 2022, 123: 104190.

[4] 刘万军, 孙虎, 姜文涛. 自适应特征选择的相关滤波跟踪算法[J]. 光学学报, 2019, 39(6): 242−255.
LIU Wanjun, SUN Hu, JIANG Wentao. Correlation filter tracking algorithm for adaptive feature selection[J]. Acta optica sinica, 2019, 39(6): 242−255.

[5] 姜文涛, 孟庆姣. 自适应时空正则化的相关滤波目标跟踪[J]. 智能系统学报, 2023, 18(4): 754−763.
JIANG Wentao,MENG Qingjiao. Correlation filter tracking for adaptive spatiotemporal regularization[J]. CAAI transactions on intelligent systems, 2023, 18(4): 754−763.

[6] BERTINETTO L, VALMADRE J, HENRIQUES J F, et al. Fully-convolutional Siamese networks for object tracking[C]//European Conference on Computer Vision. Cham: Springer, 2016: 850−865.

[7] HANG Zhipeng, PENG Houwen. Deeper and wider Siamese networks for real-time visual tracking[C]//2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2019: 4586−4595.

[8] WANG Qiang, ZHANG Li, BERTINETTO L, et al. Fast online object tracking and segmentation: a unifying approach[C]//2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2019: 1328−1338.

[9] LIN T Y, DOLLÁR P, GIRSHICK R, et al. Feature pyramid networks for object detection[C]//2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu: IEEE, 2017: 936−944.

[10] LI Bo, YAN Junjie, WU Wei, et al. High performance visual tracking with Siamese region proposal network[C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018: 8971−8980.

[11] REN Shaoqing, HE Kaiming, GIRSHICK R, et al. Faster R-CNN: towards real-time object detection with region proposal networks[J]. IEEE transactions on pattern analysis and machine intelligence, 2017, 39(6): 1137−1149.

[12] KRIZHEVSKY A, SUTSKEVER I, HINTON G. Imagenet classification with deep convolutional neural networks [C]//Advances in Neural Information Processing Systems. Lake Tahoe: MIT, 2012: 1097−1105.

[13] LI Bo, WU Wei, WANG Qiang, et al. SiamRPN: evolution of Siamese visual tracking with very deep networks[C]//2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2019: 4277−4286.

[14] HE Kaiming, ZHANG Xiangyu, REN Shaoqing, et al. Deep residual learning for image recognition[C]//2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas: IEEE, 2016: 770−778.

[15] ZHANG Lichao, GONZALEZ-GARCIA A, VAN DE WEIJER J, et al. Learning the model update for Siamese trackers[C]//2019 IEEE/CVF International Conference on Computer Vision. Seoul: IEEE, 2019: 4009−4018.

[16] ZHANG Zhipeng, PENG Houwen, FU Jianlong, et al. Ocean: object-aware anchor-free tracking[C]// European Conference on Computer Vision. Cham: Springer, 2020: 771−787.

[17] GUO Dongyan, SHAO Yanyan, CUI Ying, et al. Graph attention tracking[C]//2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville: IEEE, 2021: 9538−9547.

[18] XU Yinda, WANG Zeyu, LI Zuoxin, et al. SiamFC++: towards robust and accurate visual tracking with target estimation guidelines[C]//Proceedings of the AAAI Conference on Artificial Intelligence. Palo Alto: AAAI Press, 2020: 12549−12556.

[19] LI Qiang, QIN Zekui, ZHANG Wenbo, et al. Siamese keypoint prediction network for visual object tracking [EB/OL].(2020−06−07)[2021−01−07]. https://arvix.org/abs/2006.04078.

[20] LI Weisheng, LYU Lanbing, ZHU Junye. Multigroup spatial shift models for thermal infrared tracking[J]. Knowledge-based systems, 2022, 255: 109705.

[21] LIU Qiao, LU Xiaohuan, HE Zhenyu, et al. Deep convolutional neural networks for thermal infrared object tracking[J]. Knowledge-based systems, 2017, 134: 189−198.

[22] LIU Qiao, LI Xin, HE Zhenyu, et al. Learning deep multi-level similarity for thermal infrared object tracking[J]. IEEE transactions on multimedia, 2021, 23: 2114−2126.

[23] LIU Qiao, YUAN Di, FAN Nana, et al. Learning dual-level deep representation for thermal infrared tracking[J]. IEEE transactions on multimedia, 2023, 25: 1269−1281.

[24] LIU Shu, QI Lu, QIN Haifang, et al. Path aggregation network for instance segmentation[C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition.

Salt Lake City: IEEE, 2018: 8759−8768.

[25] LIAO Bingyan, WANG Chenye, WANG Yayun, et al. PG-net: pixel to global matching network for visual tracking[C]// European Conference on Computer Vision. Cham: Springer, 2020: 429−444.

[26] LIU Qiao, HE Zhenyu, LI Xin, et al. PTB-TIR: a thermal infrared pedestrian tracking benchmark[J]. IEEE transactions on multimedia, 2020, 22(3): 666−675.

[27] LIU Zhaoying, HE Junran, ZHANG Ting, et al. Infrared ship video target tracking based on cross-connection and spatial transformer network[C]//International Conference on Artificial Intelligence and Security. Cham: Springer, 2022: 100−114.

[28] CHEN Zedu, ZHONG Bineng, LI Guorong, et al. Siamese box adaptive network for visual tracking[C]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle: IEEE, 2020: 6667−6676.

[29] GUO Dongyan, WANG Jun, CUI Ying, et al. SiamCAR: Siamese fully convolutional classification and regression for visual tracking[C]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle: IEEE, 2020: 6268−6276.

[30] DONG Xingping, SHEN Jianbing, SHAO Ling, et al. CLNet: A compact latent network for fast adjusting Siamese trackers[C]//European Conference on Computer Vision. Cham: Springer, 2020: 378−395.

[31] CHENG Siyuan, ZHONG Bineng, LI Guorong, et al. Learning to filter: Siamese relation network for robust tracking[C]//2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville: IEEE, 2021: 4419−4429.

[32] CAO Ziang, HUANG Ziyuan, PAN Liang, et al. TCTrack: temporal contexts for aerial tracking[C]//2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans: IEEE, 2022: 14778−14788.

[33] TANG Feng, LING Qiang. Ranking-based Siamese visual tracking[C]//2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans: IEEE, 2022: 8731−8740.

[34] DANELLJAN M, BHAT G, KHAN F S, et al. ATOM: accurate tracking by overlap maximization[C]//2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2019: 4655−4664.

[35] DANELLJAN M, BHAT G, KHAN F S, et al. ECO: efficient convolution operators for tracking[C]//2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu: IEEE, 2017: 6931−6939.

[36] SONG Yibing, MA Chao, GONG Lijun, et al. CREST: convolutional residual learning for visual tracking[C]// 2017 IEEE International Conference on Computer Vision. Venice: IEEE, 2017: 2574−2583.

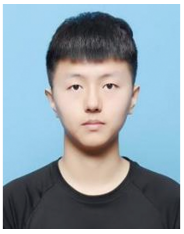
[37] WANG Ning, SONG Yibing, MA Chao, et al. Unsupervised deep tracking[C]//2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2019: 1308−1317.

[38] QI Yuankai, ZHANG Shengping, QIN Lei, et al. Hedged deep tracking[C]//2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas: IEEE, 2016: 4303−4311.

[39] MA Chao, HUANG Jiabin, YANG Xiaokang, et al. Hierarchical convolutional features for visual tracking[C]// 2015 IEEE International Conference on Computer Vision. Santiago: IEEE, 2015: 3074−3082.

[40] VALMADRE J, BERTINETTO L, HENRIQUES J, et al. End-to-end representation learning for correlation filter based tracking[C]//2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu: IEEE, 2017: 5000−5008.

作者简介：



李想，硕士，主要研究方向为模式识别、深度学习和计算机视觉。E-mail: lixiang0123@emails.bjut.edu.cn。

张婷，副教授，博士，主要研究方向为模式识别、深度学习和自然语言处理。E-mail: zhangting@bjut.edu.cn。

刘兆英，副教授，博士，主要研究方向为图像处理、模式识别和深度学习。发表学术论文 30 余篇。E-mail: zhaoying.liu@bjut.edu.cn。