基于重要特征的视觉目标跟踪可迁移黑盒攻击方法

姚 睿[1,2,3],朱享彬[1,2],周 勇[1,2],王 鹏[3,4],张艳宁[3,4],赵佳琦[1,2]

(1. 中国矿业大学计算机科学与技术学院,江苏徐州 221116;2. 矿山数字化教育部工程研究中心,江苏徐州 221116;
3. 空天地海一体化大数据应用技术国家工程实验室,陕西西安 710129;4. 西北工业大学计算机学院,陕西西安 710129)

摘 要: 视频目标跟踪的黑盒攻击方法受到越来越多的关注,目的是评估目标跟踪器的稳健性,进而提升跟踪器的安全性. 目前大部分的研究都是基于查询的黑盒攻击,尽管取得较好的攻击效果,但在实际应用中往往不能获取大量的查询以进行攻击. 本文提出一种基于迁移的黑盒攻击方法,通过对特征中与跟踪目标高度相关而不受源模型影响的重要特征进行攻击,将其重要程度降低,同时增强不重要的特征以实现具有可迁移性的攻击,即通过反向传播获得的所对应的梯度来体现其特征的重要程度,随后通过梯度得到的加权特征进行攻击. 此外,本文使用视频相邻两帧之间相似这一时序信息,提出基于时序感知的特征相似性攻击方法,通过减小相邻帧之间的特征相似度以进行攻击. 本文在目前主流的深度学习目标跟踪器上评估了提出的攻击方法,在多个数据集上的实验结果证明了本文方法的有效性及强可迁移性,在OTB数据集中,SiamRPN跟踪模型被攻击后跟踪成功率以及精确度分别下降了71.5%和79.9%.

关键词: 对抗攻击;视觉目标跟踪;黑盒攻击;可迁移性;重要特征;特征相似性

基金项目: 国家自然科学基金(No.62172417);江苏省自然科学基金(No.BK20201346)

中图分类号: TP391.4 **文献标识码:** A **文章编号:** 0372-2112(2023)04-0826-09

电子学报URL:http://www.ejournal.org.cn **DOI:**10.12263/DZXB.20220057

Transferable Black Box Attack on Visual Object Tracking Based on Important Features

YAO Rui[1,2,3], ZHU Xiang-bin[1,2], ZHOU Yong[1,2], WANG Peng[3,4], ZHANG Yan-ning[3,4], ZHAO Jia-qi[1,2]

(1. *School of Computer Science and Technology, China University of Mining and Technology, Xuzhou, Jiangsu* 221116, *China*;
2. *Ministry of Education Engineering Research Center of Mine Digitization, Xuzhou, Jiangsu* 221116, *China*;
3. *National Engineering Laboratory for Integrated Aero-Space-Ground-Ocean Big Data Application Technology, Xi'an, Shaanxi* 710129, *China*; 4. *School of Computer Science, Northwestern Polytechnical University, Xi'an, Shaanxi* 710129, *China*)

Abstract: Black-box attack methods for video object tracking have received increasing attention in order to evaluate the robustness of object trackers and thus improve the security of trackers. Most of the current researches are based on query-based black-box attacks. Although fairly good attack effects are achieved, a large number of queries still cannot be obtained for attack in practical application. We propose a transfer based black-box attack method, which attacks the important features in the features that are highly related to the tracking target and are not affected by the source model, reduceing their importance and enhancing the unimportant features to realize the transferable attack. Specifically, the corresponding gradient is obtained by back propagation to reflect the importance of its features, and then the weighted feature obtained by the gradient is used to attack. In addition, this paper uses the temporal information of similarity between adjacent video frames to propose a sequential-aware feature similarity attack method to attack the object tracker by reducing the feature similarity between adjacent frames. This paper evaluates the proposed attack method on the current mainstream deep learning target tracker. The experimental results on multiple datasets prove the effectiveness and strong mobility of this method. In OTB benchmark, the tracking success rate and accuracy of SiamRPN tracking model are reduced by 71.5% and 79.9%, respectively.

Key words: adversarial attack; visual object tracking; black box attack; transferability; important features; feature similarity

Foundation Item(s): National Natural Science Foundation of China (No.62172417); Natural Science Foundation of Jiangsu Province (No.BK20201346)

收稿日期:2022-01-10;修回日期:2022-07-07;责任编辑:朱梅玉

1　引言

视觉目标跟踪是计算机视觉的基本问题之一,在自动驾驶等领域都有着广泛的应用. 随着卷积神经网络的发展,其在计算机视觉领域展现出极为优越的性能并广泛应用. 同样,随着卷积神经网络的应用到视觉目标跟踪,视觉目标跟踪领域也有了极大的发展. 自从Szegedy等人[1]首次提出对抗性攻击,大量的研究表明CNN极易受到对抗攻击的影响,视觉目标跟踪领域也一样受到对抗攻击的潜在威胁.

许多研究工作都针对视觉目标跟踪领域的对抗攻击进行了研究,但是目前大多数的攻击方法[2~7]都是白盒攻击,即在得知跟踪模型的内部信息的情况进行攻击. 而只有少量的研究针对更加具有挑战性的黑盒攻击. 在目前的视觉目标跟踪黑盒攻击研究中,大多为基于查询的黑盒攻击方法[8~11],该方法通过在视频中施加强噪声,输入到跟踪器中得到跟踪结果,通过查询得到的结果以生成对抗样本. 但是基于查询的攻击在现实中往往是不能实现的,因为现实中的模型不可能进行大量的查询. 并且这种方法获取的对抗样本往往较为拟合特定攻击的模型,当应用于其他模型时则攻击效果较差.

本文提出一种针对深度学习目标跟踪模型的基于迁移的黑盒攻击,通过在源模型上进行白盒攻击生成对抗样本,利用对抗样本的可迁移性,对其他内部信息未知模型进行黑盒攻击[12]. 本文方法从深度网络跟踪器提取到的特征入手,由于特征的重要部分与跟踪目标相关,因此对于不同的视觉目标跟踪器都具有普遍性,不会受到源模型的影响,从而具有较强的迁移性;并且利用视频所具有的时序信息,根据视频在相邻帧之间的相似性,通过降低相邻帧之间的特征相似性进行攻击,从而充分利用到视频空间和时间的信息,以取得高效且具有强迁移性的对抗样本. 如图1所示,本文提出的方法能够有效地对SiamRPN++跟踪器[13]进行黑盒攻击,使其跟踪结果偏离真实位置,导致跟踪失败.

本文的主要贡献总结如下:

(1)提出一种基于重要特征的可迁移黑盒攻击方法,通过对特征中与跟踪目标相关而与源模型无关的重要部分进行攻击,从而得到具有强迁移性的对抗样本;

(2)针对视频相邻帧进行攻击,根据视频的时序性,提出基于时序感知的特征相似性攻击方法,降低视频相邻帧之间的特征相似性,以生成更有效的对抗样本;

(3)在多个数据集上对多个性能优秀的目标跟踪模型进行实验,结果表明,经过本文攻击的跟踪模型性能均大幅下降,与基于查询的黑盒攻击相比,本文提出的方法具有强迁移性,同时攻击效果也有一定的领先.

原始图像　　攻击后图像

图1　在VOT2018数据集中,本文方法在SiamRPN++跟踪器上的可视化跟踪结果

2　相关工作

2.1　基于深度学习的目标跟踪

目前基于深度学习的目标跟踪模型[14]主要可分为孪生网络目标跟踪模型[15~17]以及基于检测的目标跟踪模型[18,19]两类. 孪生网络目标跟踪模型将目标跟踪问题转化为相似性学习问题,将跟踪目标的初始外观作为模板,并在后续搜索帧中逐帧定位跟踪. 其对模板帧以及搜索帧进行相同的特征提取,然后在特征映射上进行互相关以度量目标的相似性及位置. 孪生网络目标跟踪模型由于在速度和精度上取得的令人满意的平衡,尤其是在短期实时跟踪领域中发挥了重要作用,因此成为当前目标跟踪领域的主流研究内容,但大多数跟踪器很容易受到对抗样本的干扰. 因此,研究这些跟踪器在对抗性攻击下的稳健性变得至关重要. 本文从孪生跟踪器深度网络提取到的特征入手,降低重要特征的重要程度,增加不重要特征的重要程度,生成有效的对抗样本.

2.2　对抗攻击

目前现有的对抗攻击方法有白盒攻击[2~7]和黑盒攻击[8~11]两类. 白盒攻击是能够得知攻击对象的参数以及结构等所有模型细节的对抗攻击. 黑盒攻击则是

在模型内部信息未知的情况下的对抗攻击. 实际应用中深度跟踪器的模型知识通常是不可得知的,因此相比白盒攻击,黑盒攻击与现实场景更贴近,更具有实用性,同样也更困难. 黑盒攻击可分为基于查询的攻击以及基于迁移的攻击. 其中基于查询的攻击又可分为基于分数和基于决策的攻击. 基于分数的攻击通过输入后得到预测的分数生成对抗性样本,基于决策的攻击通过输入后得到的分类结果生成对抗样本. 基于迁移的攻击则是利用白盒模型生成对样样本,通过提升对抗样本的可迁移性,从而使对抗样本可以对其他的黑盒模型具有攻击效果. 同样,由于在现实中大部分的场景不允许进行大量的查询以获取结果进行攻击,所以基于迁移的攻击相较于基于查询的攻击更实用且灵活. 本文提出一种针对深度学习目标跟踪模型的基于迁移的黑盒攻击,本攻击方法通过使用针对具有普遍性、不受源模型影响的重要特征进行攻击,提升了对抗样本的可迁移性,从而提升了攻击效果.

3 提出的方法

本文提出基于重要特征的可迁移黑盒攻击方法来攻击视觉目标跟踪器. 这一方法能够生成具有强迁移性的对抗样本,对不同的跟踪器实施黑盒对抗攻击. 为了实现这一方法,我们通过基于梯度感知的重要特征攻击方法(在第3.2节中讨论)以及基于时序感知的特征相似性攻击方法(在第3.3节中讨论),并采用动量迭代法以获取具有强可迁移性的对抗样本,整体攻击流程在3.4节中讨论.

3.1 问题定义

定义 $\boldsymbol{V}=\{\boldsymbol{x}_1,\cdots,\boldsymbol{x}_i,\cdots,\boldsymbol{x}_n\}$ 为有 n 个帧的视频,使用 $\boldsymbol{B}=\{\boldsymbol{b}_1,\cdots,\boldsymbol{b}_i,\cdots,\boldsymbol{b}_n\}$ 表示目标在每一帧的真实标注标签. 对于视频目标跟踪器 f,即给定目标在初始帧的初始状态,预测得到跟踪目标在后续帧中的位置. 具体来说,将视频的第一帧剪切的图像作为模板,为跟踪器 f 提供需要跟踪目标的外观信息,跟踪器 f 根据模板帧所提供的信息搜索匹配跟踪目标在后续的搜索帧中的位置.

在SiamFC孪生网络目标跟踪器中,首先通过一个参数共享的主干网络 φ 分别提取的模板帧和搜索帧的特征 $\varphi(\cdot)$,随后通过得到的特征进行互相关操作,得到分数图,如式(1)所示:

$$\boldsymbol{s}(\boldsymbol{z},\boldsymbol{x}_i)=\varphi(\boldsymbol{z})\otimes\varphi(\boldsymbol{x}_i) \tag{1}$$

其中,$\otimes$ 表示互相互操作;$\boldsymbol{z}$ 是模板图像,为视频初始帧 $\boldsymbol{x}_1$ 的裁剪图像;$\boldsymbol{x}_i$ 是搜索图像,即视频的后续搜索帧;得到的 $\boldsymbol{s}(\boldsymbol{z},\boldsymbol{x}_i)$ 为互相关所得的响应图. 响应图中最高得分的映射位置则为目标跟踪器预测的目标所在位置.

而SiamPRN以及SiamPRN++孪生网络目标跟踪器同样也是通过主干网络(其中,SiamPRN为AlexNet,SiamRPN++为ResNet50)提取模板帧特征和搜索帧特征后,相比SiamFC简单地进行互相关,其通过PRN模块预测目标的位置. RPN模块有两个分支,分别是分类分支和回归分支. 通过分类分支和回归分支进行候选区选择,首先得到目标所在的最优候选区域,然后利用回归分支得到更加精确的预测边界框,如图2所示.

将目标跟踪器 f 对 $\boldsymbol{x}_i$ 跟踪结果表示为 $f(\boldsymbol{z},\boldsymbol{x}_i)$,本文的对抗攻击方法要在添加较小的扰动的条件下生成对抗性样本 $\boldsymbol{x}^{\text{adv}}$,使跟踪器的预测位置出现错误,如式(2)所示:

$$\begin{aligned}&\arg\min_{\boldsymbol{x}^{\text{adv}}}\ \text{IoU}\left(f\left(\boldsymbol{z},\boldsymbol{x}^{\text{adv}}\right),\boldsymbol{b}_i\right)\\&\text{s.t.}\ \left\|\boldsymbol{x}-\boldsymbol{x}^{\text{adv}}\right\|_p\leqslant\varepsilon_{\max}\end{aligned} \tag{2}$$

其中,$\text{IoU}(\cdot,\cdot)$ 表示两个位置之间的交并比(Intersection over Union, IoU)分数,即 $\text{IoU}(\cdot,\cdot)=\dfrac{\text{Intersection}(\cdot,\cdot)}{\text{Union}(\cdot,\cdot)}$,$\text{Intersection}(\cdot,\cdot)$ 表示两个位置的相交区域,$\text{Union}(\cdot,\cdot)$ 表示两个位置的合并区域. 通过式(2),可以得到给定阈值约束下的满足攻击要求的对抗样本.

3.2 基于梯度感知的重要特征攻击方法

基于深度学习的目标跟踪器都是通过深度神经网络对模板帧以及搜索帧进行提取深度特征,将跟踪问题转化为一个匹配问题,将提取到的特征进行互相关得到体现相似度的响应分数以及目标位置,所以这些通过深度神经网络提取出的特征能够体现出丰富的对象信息. 尽管不同的网络会提取出不同的特征,但是往往这些特征的重要部分都与需要跟踪的目标有高度的相关性,在不同模型也具有较大的一致性. 因此本文通过对这些具有普遍性的特征重要部分进行攻击,将与跟踪目标位置高度相关的重要部分特征的重要性降低,并且将特征图中不重要的特征的重要性增加,就能够在添加较小攻击扰动的情况生成具有强迁移性的对抗样本,而不会受到生成用于生成对抗样本的源模型影响. 图2展示了本攻击方法在源模型上的具体攻击流程.

由于特征的重要性与跟踪结果反向传播的结果高度相关,因此跟踪器在获得跟踪结果后,进行反向传播得到的关于搜索帧特征的 $C\times H\times W$ 大小的梯度可以体现搜索帧特征的重要程度. 梯度代表了反向传播中特征需要调整的方向,因此在所对应的梯度中,重要的特征也体现为较高的梯度强度,而不重要的特征则体现为负梯度. 如图3所示,将获得的特征所对应梯度进行可视化,在目标区域重要的特征位置将会呈现出强梯

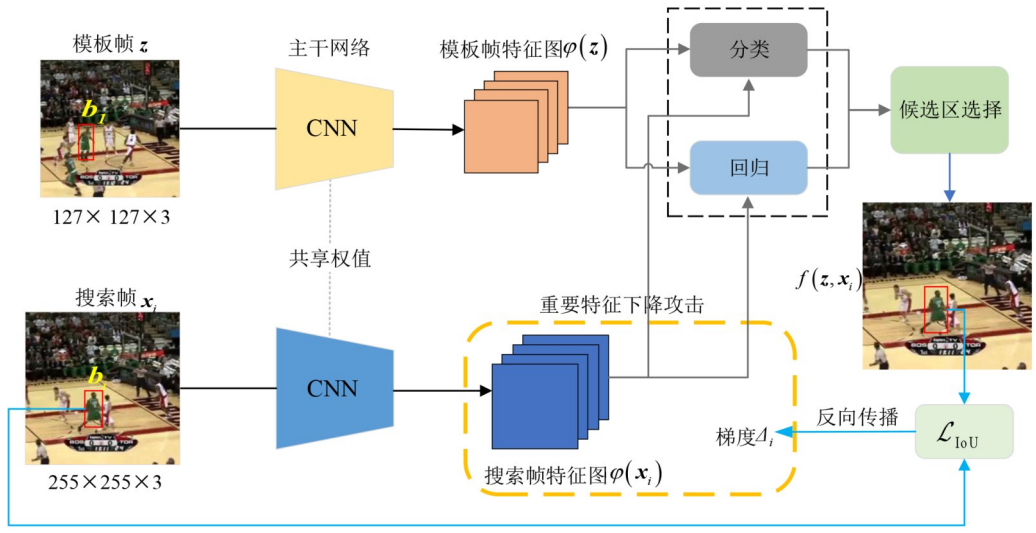

图2　基于Siamese深度学习跟踪器攻击模型示意图

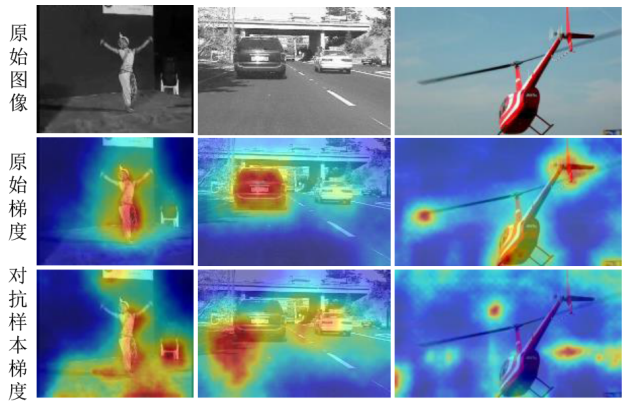

图3　梯度强度可视化图

度,而经过本文方法生成的对抗图像输入到跟踪模型后获得的特征对应梯度的强梯度则呈现在非目标区域.

为此,我们先将搜索帧输入到视觉目标跟踪器,提取出搜索帧$\boldsymbol{x}_i$的$C\times H\times W$大小的特征,将该特征表示为$\varphi(\boldsymbol{x}_i)$. 随后根据视觉目标跟踪器所预测出的预测边界框$f(\boldsymbol{z},\boldsymbol{x}_i)$,并与真实边界框$\boldsymbol{b}_i$计算IoU分数作为跟踪器的损失函数,定义如式(3)所示:

$$\mathcal{L}_{\mathrm{IoU}}=\mathrm{IoU}\left(f(\boldsymbol{z},\boldsymbol{x}_i),\boldsymbol{b}_i\right) \tag{3}$$

其中,$\boldsymbol{b}_i$为该搜索帧的真实边界框. 对该损失函数进行反向传播,并提取出与搜索帧$\boldsymbol{x}_i$特征所对应的同为$C\times H\times W$大小的梯度,该梯度表示为Δ_i,如式(4)所示:

$$\Delta_i=\frac{\partial\mathcal{L}_{\mathrm{IoU}}}{\partial\varphi(\boldsymbol{x}_i)}=\frac{\partial\left(\dfrac{I\left(f(\boldsymbol{z},\boldsymbol{x}_i),\boldsymbol{b}_i\right)}{U\left(f(\boldsymbol{z},\boldsymbol{x}_i),\boldsymbol{b}_i\right)}\right)}{\partial\varphi(\boldsymbol{x}_i)} \tag{4}$$

将搜索帧特征与梯度相乘,得到加权特征图$\hat{\varphi}(\boldsymbol{x}_i)$,定义如式(5)所示:

$$\hat{\varphi}(\boldsymbol{x}_i)=\Delta_i\odot\varphi(\boldsymbol{x}_i) \tag{5}$$

该加权特征图则能够体现出搜索帧特征图中的重要特征所在位置,较为重要的特征表现为较大的数值,不重要的特征则表现为小数值. 对该加权特征图进行求和,得到重要特性损失函数,进而指导生成对抗样本,定义如式(6)所示:

$$\mathcal{L}_1=\sum\hat{\varphi}(\boldsymbol{x}_i) \tag{6}$$

3.3　基于时序感知的特征相似性攻击方法

与静态图像不同,视频包含连续的时间信息,视频的每一帧都随着时间序列的变化而变化. 但是同样,大部分视频相邻帧之间也只有极小的变化,从而视频相邻帧之间具有较高相似度,因此跟踪器所提取出的相邻帧之间的特征也应极为相似.

本文利用视频这一时间特性,通过减小相邻帧之间的特征相似度,扩大当前特征与前一帧的特征的距离,使视频相邻两帧之间差异性增大,从而达到提升对视频样本的攻击效果,以生成具有可迁移性视频样本. 将视频前一帧特征与当前帧特征相乘,得到特征相似图,定义如式(7)所示:

$$\hat{z}_i=\varphi(\boldsymbol{x}_{i-1})\odot\varphi(\boldsymbol{x}_i) \tag{7}$$

特征相似图$\hat{z}_i$能够体现出相邻两帧的特征之间的相似度,较为相似的区域在特征相似图中表现为较大的数值,不相似的区域在特征相似图中则表现为小数值. 对该特征相似图进行求和,得到特征相似损失函数,进而指导生成对抗样本,定义如式(8)所示:

$$\mathcal{L}_2=\sum\hat{z}_i \tag{8}$$

只需要通过最小化特征相似损失函数就能实现降低视频相邻两帧之间特征相似度的攻击效果. 单独使用基于时序感知的特征相似性攻击方法对SiamRPN++目标

跟踪器进行攻击时，其成功率和精确度相比于未攻击有了较大程度的下降，可以看出其单独使用也具有较好的攻击效果. 同时，与基于梯度感知的重要特征攻击方法共同使用时的攻击效果相比于只使用基于梯度感知的重要特征攻击方法进行攻击的结果也有明显的提升，从而也体现出这一攻击方法发挥了重要的作用.

3.4 攻击方法

图4展示了上述两个损失函数的具体计算流程.

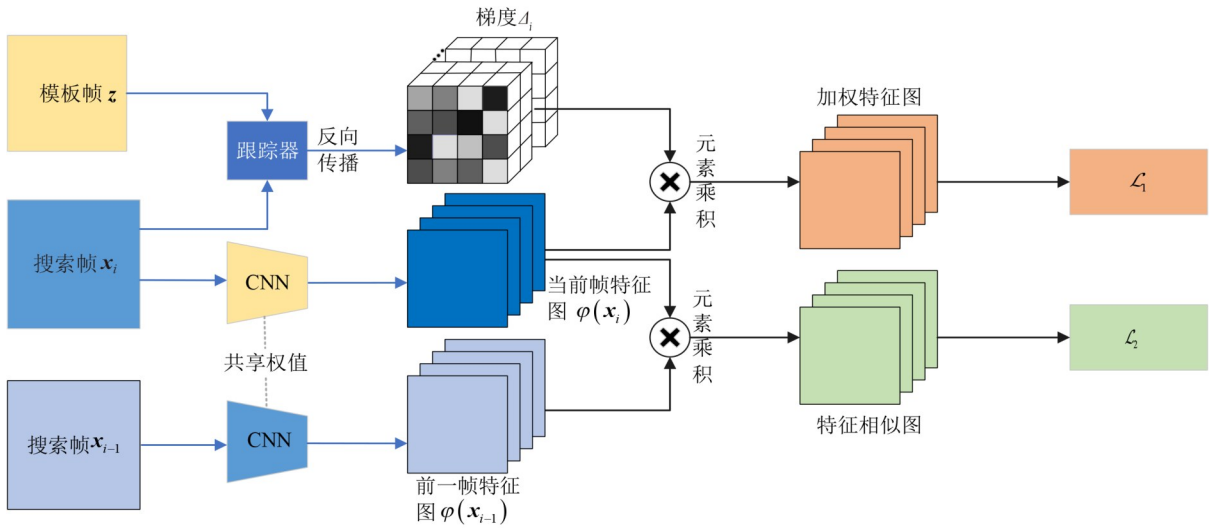

图4 生成重要特征损失函数以及特征相似损失函数的整体流程示意图

目前许多基于梯度的对抗攻击方法研究都能够解决式(10)的问题，如动量迭代法(Momentum Iterative Method, MIM)[20]，是一种运用到动量的基于梯度的迭代攻击方法，有着优越的性能，因此本文采用动量迭代梯度法来解决式(10)的问题，具体的实现方法及描述见算法1.

4 实验与分析

4.1 实验设置

数据集与攻击模型：本文将SiamFC作为攻击源模型生成具有可迁移性的对抗样本，并使用对抗样本分别对目前主流且性能优秀的四个跟踪模型SiamRPN++[13]，SiamRPN[17]，DiMP[21]和LTMU[22]进行攻击，以测试本文方法的迁移性黑盒攻击效果. 本文所选取的为不同类型的四个跟踪模型，以评估本文方法在不同类型跟踪模型的迁移性和通用性. 其中SiamRPN++，SiamRPN为目前最为主流孪生跟踪网络模型，它们将模板帧以及搜索帧经过参数共享的主干网络得到特征后，通过区域生成网络(Region Proposal Network, RPN)匹配目标模板和搜索区域之间的相似度以获得预测结果. SiamRPN++、SiamRPN这两个模型具有不同深度的主干网络. DiMP为一个具有端到端学习，并且可以在线更新的视觉目标跟踪模型. LTMU是一个在线更新的长期跟踪器，通过元更新器(meta-updater)来控制跟踪

通过对这两个损失函数进行求和，得到最终需要优化的损失函数：

$$\mathcal{L}=\alpha_1\mathcal{L}_1+\alpha_2\mathcal{L}_2 \tag{9}$$

其中，α_1和α_2为控制两个损失函数作用的权重. 本文通过最小化式(9)实现生成可迁移的对抗样本. 结合式(2)可得：

$$\arg\min_{\boldsymbol{x}^{\mathrm{adv}}}\mathcal{L},\quad \text{s.t.}\left\|\boldsymbol{x}-\boldsymbol{x}^{\mathrm{adv}}\right\|_p\leqslant\varepsilon_{\max} \tag{10}$$

算法1 梯度感知的重要特征攻击算法

输入：模板帧z，被攻击搜索帧$\boldsymbol{x}$，目标跟踪模型f，最大扰动值$\varepsilon_{\max}$，以及迭代次数T，视频帧数n.

输出：对抗样本$\boldsymbol{x}^{\mathrm{adv}}$.

FOR $i=1$ to $n-1$ do

初始化$g_0=0,\mu=1,\varepsilon=\varepsilon_{\max}/T,\alpha_1=100,\alpha_2=0.000\,1$，初始化$\varphi(z)$为模板帧特征，初始化$\boldsymbol{x}_{i,0}^{\mathrm{adv}}=\boldsymbol{x}_i$;

通过MIM迭代生成对抗样本:

FOR $t=0$ to $T-1$ do

将要被攻击的搜索帧$\boldsymbol{x}_{i,t}^{\mathrm{adv}}$输入到跟踪模型，得到搜索帧对应的特征$\varphi\left(\boldsymbol{x}_{i,t}^{\mathrm{adv}}\right)$;

反向传播后得到搜索帧特征对应的梯度$\Delta_{i,t}$;

计算重要特征损失函数$\mathcal{L}_1=\sum\left(\Delta_{i,t}\odot\varphi\left(\boldsymbol{x}_{i,t}^{\mathrm{adv}}\right)\right)$;

计算特征相似损失函数$\mathcal{L}_2=\sum\left(\varphi\left(\boldsymbol{x}_{i-1,t}^{\mathrm{adv}}\right)\odot\varphi\left(\boldsymbol{x}_{i,t}^{\mathrm{adv}}\right)\right)$;

计算总损失函数$\mathcal{L}=\alpha_1\mathcal{L}_1+\alpha_2\mathcal{L}_2$;

$$g_{t+1}=\mu\cdot g_t+\frac{\nabla_{\boldsymbol{x}_{i,t}^{\mathrm{adv}}}\mathcal{L}}{\left\|\nabla_{\boldsymbol{x}_{i,t}^{\mathrm{adv}}}\mathcal{L}\right\|_1};$$

$$\boldsymbol{x}_{i,t+1}^{\mathrm{adv}}=\mathrm{Clip}_{x,\varepsilon_{\max}}\left\{\boldsymbol{x}_{i,t}^{\mathrm{adv}}-\varepsilon\cdot\mathrm{sign}\left(g_{t+1}\right)\right\};$$

返回$\boldsymbol{x}_{i,T}^{\mathrm{adv}}$.

器的在线更新. 本文分别在四个具有挑战性的目标跟踪数据集OTB100,VOT2019,VOT2018,VOT2016上对提出的对抗攻击方法的性能进行验证. 由于在跟踪器跟踪失败后,VOT数据集将进行重新初始化,因此攻击VOT数据集比攻击OTB数据集更加困难.

实验细节:本文使用Pytorch深度学习框架实现. 硬件环境为配置了Intel Xeon Gold 5120 CPU、24 GB内存、RTX-2080Ti GPU(11 GB内存)的服务器. 本文实验在将SiamFC跟踪模型作为源模型,分别在OTB100,VOT2019,VOT2018,VOT2016四个数据集上生成对应的对抗样本,并使用生成的对抗样本对SiamRPN++,SiamRPN,DiMP和LTMU四个跟踪模型进行黑盒攻击,以验证本文方法的迁移性以及有效性. 在所有实验中,设置最大扰动$\varepsilon_{\max}=32$,设定两个损失函数的权重为$\alpha_1=100,\alpha_2=0.000\,1$,迭代次数$T$=10.

4.2　整体攻击结果

OTB100: OTB100数据集包含100个视频序列. 在OTB数据集中,本文使用准确率(Precision)和成功率(Success)作为评估标准. 准确率表示估计中心和真实中心之间的欧氏距离大于给定阈值的视频帧占所有帧的百分比. 成功率为跟踪成功的帧占所有帧的百分比. 如表1所示,加粗数据为最优结果,相比于随机噪声攻击,在经过本文方法攻击后五个跟踪器的跟踪性能都明显下降. 其中,作为攻击源模型的SiamFC跟踪模型成功率从0.575下降到0.164,准确率从0.776下降到0.164. 作为黑盒攻击模型的SiamRPN,其成功率以及准确率在攻击后分别下降了71.5%和79.9%;SiamRPN++成功率以及准确率在攻击后分别下降了18.4%和17.4%;DiMP被攻击后,成功率下降了13.1%,准确率下降了12.3%;LTMU被攻击后的成功率以及准确率也分别下降了25.1%和19.4%. 可以看出,本文迁移性黑盒攻击方法在不同类型的视觉目标跟踪模型上都取得了理想的攻击效果,具有良好的迁移性.

表1　在OTB100中在SiamRPN++,SiamRPN以及SiamFC的攻击结果

跟踪器	成功率			准确率		
	原始	随机扰动	本文方法	原始	随机扰动	本文方法
SiamFC	0.575	0.526	0.164	0.776	0.702	0.164
SiamRPN	0.636	0.593	0.181	0.847	0.787	0.170
SiamRPN++	0.692	0.661	0.565	0.906	0.818	0.748
DiMP	0.671	0.659	0.583	0.869	0.860	0.762
LTMU	0.672	0.622	0.503	0.872	0.815	0.703

VOT2016: VOT2016数据集包含60个视频序列. VOT数据集与其他目标跟踪数据集相比增加了一个重新初始化机制,当跟踪器跟踪失败后,将通过真实边界框进行重新初始化,考虑到这一机制,稳健性(Robustness)是一个重要的评估指标. 并且通过正确率(Accuracy)评估跟踪器的准确性,失败数(Failures)体现跟踪器丢失目标后,进行重新初始化的次数. 同时平均重叠期望(Expected Average Overlap,EAO)综合正确率和稳健性进行评估. 如表2所示,本文在VOT数据集中对SiamRPN++,SiamRPN,DiMP以及LTMU四个跟踪模型上进行黑盒攻击,并与基于查询的黑盒攻击方法IoU攻击方法[8]进行比较,加粗数据为最优结果. SiamRPN被攻击后,正确率从0.618下降到0.573,跟踪失败重新初始化次数从51次上升到123次,而平均重叠期望从0.393下降到0.223. SiamRPN++,DiMP以及LTMU这三个不同类型的跟踪器被攻击后,各项评估指标也有明显的下降,由此可见本文产生的攻击样本对不同的跟踪模型都有明显的攻击效果,体现出优秀的迁移性. 同时与基于查询的黑盒攻击方法IoU攻击相比,本文方法在多个评估指标上都有较好的结果. 这也证明本文的方法在具有优秀的迁移性的同时,也具有良好攻击效果.

表2　在VOT2016数据集中在SiamRPN++,SiamRPN的攻击结果

跟踪器	正确率			稳健性			失败数			平均重叠期望		
	原始	IoU攻击	本文攻击	原始	IoU攻击	本文攻击	原始	IoU攻击	本文攻击	原始	IoU攻击	本文攻击
SiamRPN	0.618	0.581	**0.573**	0.238	0.569	**0.573**	51	122	**123**	0.393	0.235	**0.223**
SiamRPN++	0.642	0.605	**0.594**	0.196	**0.802**	0.61	42.0	**172**	131	0.464	**0.183**	0.252
DiMP	0.599	0.536	**0.521**	0.140	0.374	**0.381**	30	80	**92**	0.449	0.256	**0.239**
LTMU	0.661	0.604	**0.598**	0.522	**0.904**	0.827	112	**194**	165	0.236	**0.170**	0.174

VOT2018: VOT2018相较于VOT2016具有不相同的六十个视频序列,同时也增加了视频跟踪的难度. 如表3所示,加粗数据为最优结果,在VOT2018视觉目标跟踪数据集中,本文所提出的对抗攻击方法分别使得SiamRPN++以及SiamRPN两个跟踪器的正确率分别降低了5.1%和4.3%,失败数分别提升了107和210,平均重叠期望分别降低了51.9%和63.0%. 而DiMP和LTMU这两个跟踪器在被攻击后,正确率也分别下降了13.0%和9.5%,失败数分别提升了25、88,平均重叠期望分别下降了75%和35%. 同样,与IoU攻击方法[8]相比,本文所提出的攻击方法在VOT2018的多个评估指标中都取得领先,证明本文的方法在不同的测试数据集上生成

表3 在VOT2018数据集中在SiamRPN++,SiamRPN的攻击结果

跟踪器	正确率			稳健性			失败数			平均重叠期望		
	原始	IoU攻击	本文攻击	原始	IoU攻击	本文攻击	原始	IoU攻击	本文攻击	原始	IoU攻击	本文攻击
SiamRPN	0.576	0.538	**0.525**	0.314	0.823	**0.824**	66	**173**	**173**	0.329	0.163	**0.158**
SiamRPN++	0.596	0.568	**0.553**	0.271	1.171	**1.228**	57	250	**267**	0.370	**0.129**	0.135
DiMP	0.574	0.507	**0.489**	0.145	0.400	**0.457**	31	43	**56**	0.427	0.248	**0.241**
LTMU	0.624	0.590	**0.572**	0.702	**1.320**	1.187	150	**282**	238	0.195	**0.120**	0.126

的样本都具有较强的迁移性及攻击性能,表明了本方法的有效性.

VOT2019:同样,本文也在VOT2019上对本文方法进行了评估. 如表4所示,加粗数据为最优结果,本文的攻击使得SiamRPN++以及SiamRPN两个跟踪器的正确率分别降低了4.5%和3.7%,失败数分别提升了111和152,平均重叠期望分别降低了24.9%和39.4%. 在DiMP和LTMU两个跟踪模型上,本文的攻击也使其跟踪性能大幅下降,也体现了良好的攻击效果. 同样,本文的攻击方法在多个评估指标中均优于IoU攻击方法[8],体现出本文攻击方法的良好性能.

表4 在VOT2019数据集中在SiamRPN++,SiamRPN的攻击结果

跟踪器	正确率			稳健性			失败数			平均重叠期望		
	原始	IoU攻击	本文攻击	原始	IoU攻击	本文攻击	原始	IoU攻击	本文攻击	原始	IoU攻击	本文攻击
SiamRPN	0.572	0.548	**0.527**	0.577	**1.154**	1.143	113	**226**	224	0.248	**0.160**	0.175
SiamRPN++	0.589	0.575	**0.552**	0.510	**1.575**	1.398	100	**314**	252	0.266	**0.124**	0.161
DiMP	0.568	0.474	**0.462**	0.277	0.641	**0.655**	55	**127**	112	0.332	0.195	**0.193**
LTMU	0.625	0.576	**0.558**	0.913	**1.470**	1.237	182	**293**	264	0.201	**0.150**	0.159

4.3 消融实验

为了探索本文所提出的基于迁移的黑盒攻击方法中的基于梯度感知的重要特征攻击方法和基于时序感知的特征相似性攻击方法的效果与贡献,在OTB100测试数据集上,本文将有无这两个方法所生成的对抗样本应用在SiamRPN++目标跟踪器进行对比实验,以分析和评估这两个方法在本文提出的攻击方法中的贡献与作用,具体实验结果如表5所示. 可以看出分别单独使用这两个方法在一定程度上都使SiamRPN++的成功率和精确度相比于未攻击有了较大程度的下降,但是相比于同时部署两个方法进行攻击的结果较差,从而也体现出这两个方法都在攻击中有着不同程度的重要性,同时发挥了不可或缺的作用.

表5 在OTB100数据集上,有无梯度感知的重要特征攻击方法和时序感知特征相似性攻击方法实施在跟踪器SiamRPN++的攻击结果

梯度感知的重要特征攻击方法	时序感知的特征相似性攻击方法	成功率	准确率
×	×	0.692	0.906
√	×	0.571	0.752
×	√	0.582	0.771
√	√	0.565	0.748

同时,为了探索本文所提出的基于迁移的黑盒攻击方法中的基于梯度感知的重要特征攻击方法和基于时序感知的特征相似性攻击方法的两个损失函数的最佳权重比例,本文在OTB100视觉目标跟踪数据集上将不同比例的重要特征损失函数和特征相似损失函数生成的对抗样本应用在SiamRPN++目标跟踪器进行对比实验,以得到最佳的权重比例,以达到最佳的攻击效果,具体实验结果如表6所示. 可以看出选取不同比例的样本都使SiamRPN++跟踪器的跟踪性能有了不同程度的下降,但是不同的权重比例的效果还是具有一定程度的差异,本文所选取的两个损失函数权重比例$\alpha_1=100,\alpha_2=0.000\,1$取得了最好的攻击效果,相比于表中较差的权重比例$\alpha_1=1,\alpha_2=1$,SiamRPN++被攻击后在成功率和准确率上分别多下降了1.9%和3.0%. 并且由表6可知,当α_1的权重系数减小,或当α_2的权重系数增大时,重要特征损失函数的攻击效果都将被削弱,使发挥主要攻击作用的是特征相似性攻击方法,从而使本文所提出的攻击方法的攻击效果有所减弱.

4.4 可视化结果

本文分别将未受到攻击的原始视频图像、IoU攻击方法[8]所生成的对抗样本以及本文攻击方法所生成的可迁移对抗样本图像在SiamPRN++目标跟踪器上的跟踪结果可视化,结果如图5所示. 可见SiamPRN++在原始视频图像上能够准确跟踪到目标,但是在被攻击样本中则出现明显跟踪失败现象. 相比于IoU攻击,可见本文攻击方法导致的跟踪偏移量更大,这也证明了本文的攻击相比于IoU攻击[8],对目标跟踪器的精确度具有更加有效的攻击效果. 并且通过图5可知,与IoU攻击产生的对抗样本相比,本文所提出的攻击方法生成

表6　在OTB100数据集上,不同权重比例的重要特征损失函数和特征相似损失函数的攻击实施在SiamRPN++跟踪器的攻击结果

α_1	α_2	成功率	准确率	α_1	α_2	成功率	准确率	α_1	α_2	成功率	准确率
100	0.000 1	**0.565**	**0.748**	10	0.000 1	0.572	0.755	1	0.000 1	0.577	0.765
100	0.001	0.573	0.757	10	0.001	0.575	0.763	1	0.001	0.581	0.772
100	0.01	0.576	0.765	10	0.01	0.580	0.771	1	0.01	0.583	0.776
100	1	0.578	0.766	10	1	0.582	0.772	1	1	0.584	0.778

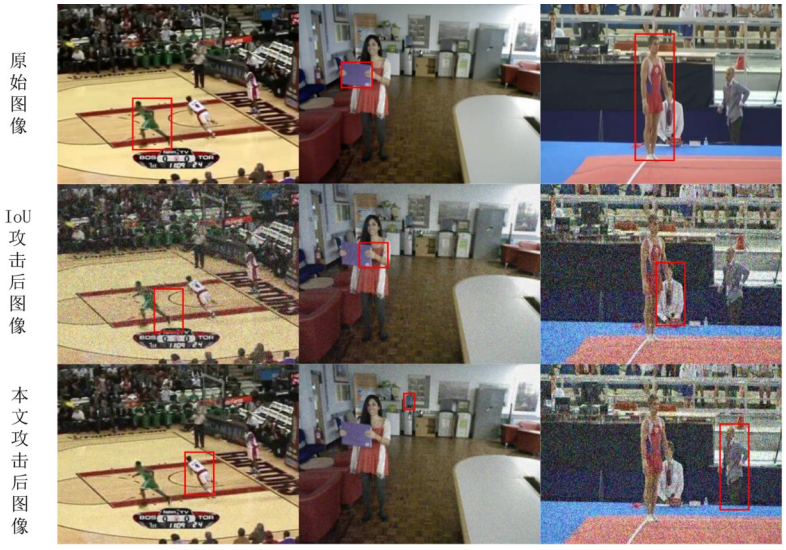

图5　在VOT2018数据集中,IoU攻击方法[8]以及本文攻击方法在SiamRPN++跟踪器上的跟踪结果可视化

的对抗样本在一些图像上施加的噪声更小,可以在不影响人类视觉的观看效果与判断的情况下进行成功有效的攻击.

5　结论

本文提出了一种基于重要特征的视频目标跟踪可迁移黑盒对抗攻击方法,在不可知目标跟踪模型以及不进行大量查询的情况下,通过源模型降低重要特征的重要性,并且同时增大视频相邻帧特征的相似度,生成具有强可迁移性的对抗样本,从而对其他目标跟踪器实现黑盒攻击. 本文将所提出的方法应用于具有不同网络深度的孪生跟踪网络中进行评估,以说明本文的可迁移黑盒对抗攻击的可迁移性. 在多个数据集上的大量实验证明了本文方法的有效性. 相信本文的工作能对视觉目标跟踪稳健性的研究有所帮助.

参考文献

[1] SZEGEDY C, ZAREMBA W, SUTSKEVER I, et al. Intriguing properties of neural networks[EB/OL]. (2013-12-21)[2022-01]. https://arxiv.org/abs/1312.6199.

[2] YAN B, WANG D, LU H C, et al. Cooling-shrinking attack: Blinding the tracker with imperceptible noises[C]// 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle: IEEE, 2020: 987-996.

[3] YAN X Y, CHEN X S, JIANG Y, et al. Hijacking tracker: A powerful adversarial attack on visual tracking[C]// ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Barcelona: IEEE, 2020: 2897-2901.

[4] LIANG S Y, WEI X X, YAO S Y, et al. Efficient adversarial attacks for visual object tracking[C]//European Conference on Computer Vision - ECCV 2020. Glasgow: Springer, 2020: 34-50.

[5] CHEN X S, YAN X Y, ZHENG F, et al. One-shot adversarial attacks on visual tracking with dual attention[C]// 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle: IEEE, 2020: 10173-10182.

[6] GUO Q, CHENG Z Y, JUEFEI-XU F, et al. Learning to adversarially blur visual object tracking[C]//2021 IEEE/CVF International Conference on Computer Vision (ICCV). Montreal: IEEE, 2022: 10819-10828.

[7] GUO Q, XIE X F, JUEFEI-XU F, et al. SPARK: Spatial-aware online incremental attack against visual tracking[C]// European Conference on Computer Vision - ECCV 2020. Glasgow: Springer, 2020: 202-219.

[8] JIA S, SONG Y B, MA C, et al. IoU attack: Towards temporally coherent black-box adversarial attack for visual ob-

ject tracking[C]//2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Nashville: IEEE, 2021: 6705-6714.

[9] WANG Z B, GUO H C, ZHANG Z F, et al. Feature importance-aware transferable adversarial attacks[C]//2021 IEEE/CVF International Conference on Computer Vision (ICCV). Montreal: IEEE, 2022: 7619-7628.

[10] WEI X, LIANG S, CHEN N, et al. Transferable adversarial attacks for image and video object detection[EB/OL]. (2018-11-30)[2022-01]. https://arxiv.org/abs/1811.12641.

[11] INKAWHICH N, LIANG K J, CARIN L, et al. Transferable perturbations of deep feature distributions[EB/OL]. (2020-04-27)[2022-01]. https://arxiv.org/abs/2004.12519.

[12] 姚睿, 朱享彬, 周勇, 等. 一种基于重要特征的视觉目标跟踪可转移黑盒攻击方法: CN114511593A[P]. 2022-05-17.
YAO R, ZHU X B, ZHOU Y, et al. Visual target tracking transferable black box attack method based on important features: CN114511593A[P]. 2022-05-17. (in Chinese)

[13] LI B, WU W, WANG Q, et al. SiamRPN++: Evolution of Siamese visual tracking with very deep networks[C]// 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Long Beach: IEEE, 2020: 4277-4286.

[14] YAO R, LIN G S, XIA S X, et al. Video object segmentation and tracking: A survey[J]. ACM Transactions on Intelligent Systems and Technology, 2020, 11(4): 1-47.

[15] BERTINETTO L, VALMADRE J, HENRIQUES J F, et al. Fully-Convolutional Siamese Networks for Object Tracking[C]// European Conference on Computer Vision. Amsterdam: Springer, 2016: 850-865.

[16] 丁新尧, 张鑫. 基于显著性特征的选择性目标跟踪算法[J]. 电子学报, 2020, 48(1): 118-123.
DING X Y, ZHANG X. Visual tracking with salient features and selective mechanism[J]. Acta Electronica Sinica, 2020, 48(1): 118-123. (in Chinese)

[17] LI B, YAN J J, WU W, 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.

[18] SONG Y B, MA C, WU X H, et al. VITAL: VIsual tracking via adversarial learning[C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018: 8990-8999.

[19] 陈丹, 姚伯羽. 运动模型引导的自适应核相关目标跟踪方法[J]. 电子学报, 2021, 49(3): 550-558.
CHEN D, YAO B Y. Adaptive response kernel correlation target tracking method guided by motion model[J]. Acta Electronica Sinica, 2021, 49(3): 550-558. (in Chinese)

[20] DONG Y P, LIAO F Z, PANG T Y, et al. Boosting adversarial attacks with momentum[C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018: 9185-9193.

[21] BHAT G, DANELLJAN M, VAN GOOL L, et al. Learning discriminative model prediction for tracking[C]//2019 IEEE/CVF International Conference on Computer Vision (ICCV). Seoul: IEEE, 2020: 6181-6190.

[22] DAI K N, ZHANG Y H, WANG D, et al. High-performance long-term tracking with meta-updater[C]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle: IEEE, 2020: 6297-6306.

作者简介

姚 睿 男,1982年出生,河南南阳人. 现为中国矿业大学计算机科学与技术学院教授、博士生导师. 主要研究方向为计算机视觉、机器学习.
E-mail: ruiyao@cumt.edu.cn

朱享彬 男,1997年出生,江西赣州人. 现为中国矿业大学计算机科学与技术学院硕士研究生. 主要研究方向为计算机视觉、机器学习.
E-mail: TS20170131P31@cumt.edu.cn

周 勇 男,1974年出生,江苏徐州人. 现为中国矿业大学计算机科学与技术学院教授、博士生导师. 主要研究方向为机器学习、数据挖掘.
E-mail: yzhou@cumt.edu.cn