基于空间感知的多级损失目标跟踪对抗攻击方法

程旭[1,2,3]，王莹莹[1,2]，张年杰[1]，付章杰[1,2]，陈北京[1,2]，赵国英[3]

（1. 南京信息工程大学计算机学院、软件学院、网络空间安全学院，江苏 南京 210044；
2. 南京信息工程大学数字取证教育部工程研究中心，江苏 南京 210044；3. 奥卢大学机器视觉与信号分析研究中心，奥卢 FI-90014）

摘 要：针对现有的对抗扰动技术难以有效地降低跟踪器的判别能力使运动轨迹发生快速偏移的问题，提出一种高效的攻击目标跟踪器方法。首先，所提方法从高层类别和底层特征考虑设计了欺骗损失、漂移损失和基于注意力机制的特征损失来联合训练生成器；然后，将干净图像送入该生成器中，生成对抗样本；最后，利用对抗样本干扰目标跟踪器，导致目标运动轨迹发生偏移，降低跟踪精度。实验结果表明，所提方法在 OTB 数据集上达到了 54%的成功率下降和 70%的精确度下降，实现了复杂场景下对目标快速有效的攻击。

关键词：视频监控；网络安全；对抗攻击；深度学习；目标跟踪

中图分类号：TP391

文献标识码：A

DOI: 10.11959/j.issn.1000−436x.2021208

Multi-level loss object tracking adversarial attack method based on spatial perception

CHENG Xu[1,2,3], WANG Yingying[1,2], ZHANG Nianjie[1], FU Zhangjie[1,2], CHEN Beijing[1,2], ZHAO Guoying[3]

1. School of Computer Science, Nanjing University of Information Science and Technology, Nanjing 210044, China
2. Engineering Research Center of Digital Forensics, Ministry of Education, Nanjing University of Information Science and Technology, Nanjing 210044, China
3. Center for Machine Vision and Signal Analysis, University of Oulu, Oulu FI-90014, Finland

Abstract: In order to solve the problem that it is difficult for the existing adversarial disturbance techniques to effectively reduce the discrimination ability of the trackers and make the trajectory deviation rapidly, an effective object tracking adversarial attack method was proposed. First, deception loss, drift loss and attention mechanism-based loss was designed to jointly train generator based on the consideration of the high-level categories and the low-level features. Then, the clean image was sent to the trained generator to generate the adversarial samples that were used to interfere with the object trackers, which made the object trajectory deviation and reduced the tracking accuracy. Experimental results show that the proposed method achieves 54% reduction in success rate and 70% reduction in accuracy on OTB dataset, which can attack the object of tracking quickly in complex scenes.

Keywords: video surveillance, network security, adversarial attack, deep learning, object tracking

1 引言

视觉目标跟踪是计算机视觉的关键任务之一，在公共安全领域扮演着十分重要的角色，如视频监控、自动驾驶、无人机追踪、图像目标分割、目标行为识别等。近年来，得益于深度学习（DL, deep

收稿日期：2021−06−30；**修回日期**：2021−09−26

基金项目：国家自然科学基金资助项目（No.61802058, No.61911530397）；国家留学基金资助项目（No.201908320175）；中国博士后科学基金资助项目（No.2019M651650）；江苏省自然科学基金资助项目（No.BK20201267, No.BK20200039）

Foundation Items: The National Natural Science Foundation of China (No.61802058, No.61911530397), China Scholarship Council (CSC) (No.201908320175), China Postdoctoral Science Foundation (No.2019M651650), The Natural Science Foundation of Jiangsu Province (No.BK20201267, No.BK20200039)

learning）技术的蓬勃发展，目标跟踪算法取得了重大突破，特别是孪生网络这一类目标跟踪算法在精度和速度上均取得了优异表现，并在 OTB 视频跟踪数据集上达到了 91%的精确度，速度也达到了实时。然而，从安全的角度考虑，深度学习跟踪器存在严重的安全隐患，极易受到对抗样本的干扰。

对抗攻击是通过对原始图像添加人眼不可见的微小扰动，以欺骗深度网络模型，导致分类预测错误。近年来，对抗攻击已经由图像分类延伸到目标跟踪、语义分割等领域，成功地破坏了深度学习任务的有效性。此外，深度学习算法无法有效地处理对抗样本。伪造的样本会使深度学习模型输出意想不到的结果。因此，研究基于深度学习的目标跟踪对抗攻击方法对确保算法的安全性和稳健性是至关重要的，可为设计更加稳健的算法提供思路。

基于以上动机，本文以孪生网络跟踪器 SiamRPN++为主要攻击对象，研究了视觉目标跟踪的对抗攻击方法，主要贡献包括以下 3 个方面。

1) 针对现有对抗扰动技术难以有效地干扰跟踪器使运动轨迹发生快速偏移的问题，提出了一种基于空间感知的多级损失目标跟踪对抗攻击方法，利用生成器生成对抗样本来实现对目标跟踪器的干扰，降低了跟踪精度，具有较好的攻击效果。

2) 提出了一种高效的空间感知快速漂移攻击框架，在此框架下设计了欺骗损失、漂移损失和双重注意力机制的特征损失和感知损失来联合训练生成器，生成人眼难以察觉的对抗扰动，用于欺骗目标跟踪器。

3) 将所提方法在 OTB100、VOT2018 和 LaSOT 这 3 个主流的目标跟踪数据集上进行验证，实验结果表明，所提方法可使跟踪器的判别能力失灵，预测边框逐渐收缩，导致目标轨迹发生偏移，比原始跟踪器在 OTB 数据集上实现了 70%的精确度下降。

2　相关工作

目标跟踪技术是高层视觉任务分析与处理的基础，已在视频监控、视觉导航、行为识别、自动驾驶等领域得到了广泛应用。视觉目标跟踪任务是在给定某视频序列初始帧的目标大小与位置的情况下，预测该目标在后续帧的大小与位置。然而，即使基于深度学习的目标跟踪技术已经能够成功地处理复杂问题，但最近研究表明它们对输入中的轻微扰动很敏感，会导致跟踪性能下降。对抗攻击对深度学习在实践中取得成功构成了一系列威胁。本节将分别从目标跟踪、对抗攻击 2 个方面介绍相关的研究工作。

1) 目标跟踪

近年来，以相关滤波（CF, correlation filter）和深度学习为代表的判别式方法取得了令人满意的效果，已成为目标跟踪的主流方法。

相关滤波源于信号处理领域，基于相关滤波目标跟踪的基本思想就是寻找一个滤波模板，让下一帧图像与滤波模板进行卷积操作，响应最大的区域则是预测目标。基于此，国内外学者先后提出了大量方法，如 MOSSE（minimum output sum of squared error filter）[1]、KCF（kernelized correlation filter）[2]等。此外，在 KCF 的基础上又发展了一系列跟踪方法用于处理各种复杂场景下的挑战，如处理尺度变化的 DSST（discriminative scale space tracker）[3]、基于分块的相关滤波 RPT（reliable patch tracker）[4]等。但是上述方法会受到边界效应的影响。为了克服这一问题，Danelljan 等[5]提出一种高效的 SRDCF（spatially regularized discriminative correlation filter）方法，利用空间正则化惩罚相关滤波系数，取得了和同时期基于深度学习跟踪方法相当的效果。进一步地，Danelljan 等[6]利用卷积神经网络（CNN, convolutional neural network）提取目标特征，并结合相关滤波提出了连续卷积算子的目标跟踪（C-COT, continuous convolution operator for visual tracking）方法。

由于深度特征对目标拥有强大的表征能力，深度学习在计算机视觉各领域展现出巨大的潜力。Wang 等[7]首次将深度学习引入目标跟踪领域，其将在分类数据集上训练的卷积神经网络迁移到目标跟踪任务中，与传统方法相比，性能得到了提升。Hong 等[8]提出的 CNN-SVM 算法首先利用在 ImageNet 上训练的卷积神经网络提取目标特征，再利用 SVM 跟踪目标。Wang 等[9]提出基于全卷积模型的目标跟踪方法，利用目标的 2 个卷积层特征构造可以选择特征图的网络，跟踪性能比 CNN-SVM 有了小幅提升。其他代表性方法还有 HCF[10]、VITAL[11]等。然而，目标跟踪任务与图像分类任务有本质区别，图像分类任务关注类间差异，忽视了类内区别；目标跟踪任务则关注区分特定目标与背景，抑制同类目标。因此，在分类数据集上预训练的网络可能不完全适用于目标跟踪任务。

针对这一问题，文献[12]提出一种专门在跟踪视频序列上训练的多域卷积神经网络模型 MDNet，获得了 VOT2015 竞赛冠军。然而，该方法不能满足实时要求。针对这一问题，基于孪生网络的目标跟踪算法在跟踪精度和速度上取得了很好的平衡，在大量数据集上取得了优异的性能，代表性方法包括 SiamFC[13]、SiamRPN[14]、SiamRPN++[15]、DaSiamRPN[16]、Siam R-CNN[17]等。

2) 对抗攻击

研究表明，CNN 极易受到攻击。即使最先进的分类器也很容易被添加到原始图像中的噪声所蒙蔽。因此，深度学习下的对抗攻击研究具有重要意义。

根据威胁模型，可将现有攻击分为白盒攻击和黑盒攻击，它们之间的差异在于攻击者了解的信息不同。白盒攻击假定攻击者具有关于目标模型的完整知识，可通过任何方式直接在目标模型上生成对抗样本。黑盒攻击只能依赖查询访问的返回结果来生成对抗样本。在上述 3 种攻击模型的框架中，研究者提出了许多用于对抗样本生成的攻击算法。这些方法大致可分为基于梯度迭代的攻击、基于生成式对抗网络（GAN, generative adversarial network）的攻击和基于优化的攻击三类。

基于梯度迭代的攻击方式的代表性方法包括 FGSM[18]、Deepfool[19]、DAG[20]、PGD[21]、BIM[22]，它们通过优化对抗目标函数以愚弄深度神经网络。Wang 等[19]利用迭代计算生成最小规范对抗扰动，将位于分类边界内的图像逐步推到边界外，直到出现错误分类。然而，FGSM 和 PGD 生成的对抗样本比较模糊，跟踪时不但容易被发现，而且攻击效果较差。司念文等[23]提出一种基于对抗补丁的 Grad-CAM 攻击方法，设计了分类结果不变而解释结果偏向对抗补丁的目标函数，使 Grad-CAM 方法无法定位图像中的显著区域。Su 等[24]提出一种基于差分进化的单像素对抗扰动生成方法，通过修改图像中的一个像素，使数据集中多种类别的图像至少有一类目标被攻击。该方法仅修改单个像素无法适应视频的多帧任务。Zhong 等[25]首次研究了迁移对抗攻击在人脸识别中的特性，提出了一种基于丢弃的方法 DFANet 来提高现有攻击方法的迁移性，生成的人脸图像对有效地欺骗了人脸识别系统。Chen 等[26]提出对目标模板的单次攻击方法，通过优化批置信度损失和特征损失来寻找模板的对抗样本。该方法产生的对抗样本易被人眼察觉，无法攻击正常运行的跟踪器。Jia 等[27]利用构造的伪分类标签和伪回归标签来寻找真实损失和伪损失差异的梯度方向，进而产生对抗样本。然而，该方法攻击过程耗时，难以满足实时性要求。

基于生成式对抗网络的攻击方式使用大量数据来训练生成器以产生扰动噪声，代表性方法有 AdvGAN[28]、UEA[29]、AdvGAN++[30]。Deb 等[31]提出一种高质量的对抗人脸生成法，运用 GAN 来改变人脸的潜在区域使对原图扰动最小，在不改变视觉质量的情况下，大幅降低了人脸识别的成功率。Baluja 等[32]提出一种全新的对抗样本生成方法，针对目标网络或一系列需要攻击的网络，通过自监督学习方式训练对抗转化网络（ATN, adversarial transformation network）来生成对抗样本，提高了对抗样本生成速度且丰富了样本的多样性。以上基于 GAN 的方法需同时优化生成器与判别器以产生对抗样本。Yan 等[33]提出一种冷却收缩对抗损失以冷却目标区域及收缩预测边框，该方法虽能快速产生人眼无法察觉的对抗样本，但是攻击能力欠佳且对黑盒跟踪器的迁移性有限。Sharif 等[34]提出一种对抗生成网络（AGN, adversarial generative network），训练生成器网络产生满足期望目标的对抗样本，在数字空间和现实世界中均成功迷惑了人脸识别系统。

基于优化的攻击方式主要是 CW 攻击（carlini and wagner attack）[35]。该攻击生成的扰动可以从未经防御的网络迁移到经过防御的网络上，以实现黑盒攻击。Moosavi-Dezfooli 等[36]提出一种计算普适性扰动的算法，在数据分布中采样样本集进行训练，使每个样本都能以一定概率被错误分类，在新样本预测时欺骗分类器，证明高维决策边界具有几何相关性。Din 等[37]提出基于隐写技术的对抗扰动生成方法，通过在变换域中将单个秘密图像嵌入任意目标图像来产生扰动，使流行分类模型以高概率错误分类目标。

3 基于孪生网络的目标跟踪器及其可攻击性

近年来，孪生网络在目标跟踪领域取得了很高的性能，其将目标跟踪问题转化为 Patch 块的匹配问题，通过比较图像搜索区域与目标模板的相似度，得到新的目标位置。在众多孪生网络跟踪方法中，SiamRPN++[15]跟踪器在跟踪数据库上刷新了纪录，

不仅精度高，运行速度也满足实时性要求。

然而，跟踪算法本身存在被攻击的潜在风险。即使是 SiamRPN++跟踪器也会遭受噪声干扰，导致目标跟踪失败。

现有目标跟踪系统的对抗攻击存在以下难点。

1) 目标跟踪的对抗攻击不同于简单的分类任务，它既包括分类，也有精准的边框回归，仅通过迁移图像分类任务中的对抗攻击达不到预期效果。

2) 由于目标跟踪的特殊性，目标只在第一帧中给出，无法预知其类别，因而不能为每个类别单独训练对抗补丁。

为此，本文从视觉目标跟踪任务本身出发，在设计攻击损失函数时融合了基于分数与特征干扰分类任务和基于回归偏移量破坏回归任务，导致跟踪器无法准确判别目标存在区域，回归边框逐渐缩小并快速沿着与真实目标最远的方向移动，造成跟踪失败。此外，本文摒弃为每个类别单独训练对抗补丁的思路，从低级特征和高层语义角度出发，设计了欺骗损失、漂移损失、基于双重注意力机制的特征损失和感知损失，通过联合训练生成器，使生成器在不同场景下能对目标产生肉眼难以察觉的扰动，以达到欺骗目标跟踪器的目的。

4　空间感知的多级损失跟踪对抗攻击方法描述

本文提出一种基于空间感知的多级损失漂移攻击框架来欺骗性能较好的 SiamRPN++跟踪器，对原始图像添加微小扰动，使跟踪器识别不到目标的正确位置及姿态估计。为了实现这一目标，本文设计了欺骗损失、漂移损失、基于注意力机制的特征损失和感知损失来联合训练基于 GAN 的生成器，以产生强对抗样本，用于攻击跟踪器。下面将详细介绍本文所提出的攻击方法。

4.1　对抗样本生成

本文提出的攻击框架包括两部分，分别是扰动生成器 ξ_g 和跟踪器 SiamRPN++。扰动生成器训练结构框架如图 1 所示。生成器训练过程中，保持模板不变，将干净搜索区域送入生成器产生噪声，再与干净搜索区域相加，形成对抗搜索区域。同时，将每一帧对抗搜索区域分别与干净模板一起送入跟踪器进行模板匹配，输出特征提取网络 Conv3-3 的特征图，得到对抗样本响应图和回归图。

要想达到攻击的目的，需要更多地关注搜索区域中最有可能是目标的区域。因此，同时把干净模板和相应干净搜索区域输入跟踪器，找出感兴趣区域。接着，利用所提出的欺骗损失 L_{cheat}、漂移损失 L_{drift}、基于双重注意力的特征损失 L_{feature} 和感知损失 L_{quality} 联合训练生成器。算法流程如算法 1 所示。

算法 1　本文扰动生成器训练框架

输入　干净的目标模板 $\boldsymbol{Z}^C$，干净的搜索区域 $\boldsymbol{S}^C$，自定义噪声图像 $\boldsymbol{N}_t$，训练视频数目 T，随机初始化生成器 ξ_g

输出　基于搜索区域的扰动生成器 ξ_g^*

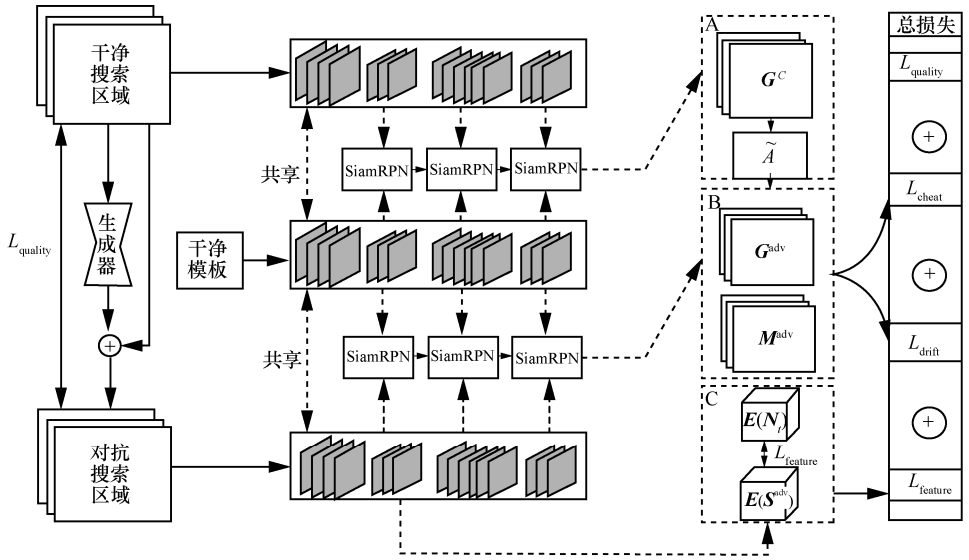

图 1　扰动生成器训练结构框架

1) 初始化生成器 ξ_g 和跟踪器并固定参数

2) for $i = 1{:}T$

3) 获取干净的模板 $\boldsymbol{Z}^C$ 和 N 张干净搜索区域 $\boldsymbol{S}^C$；

4) 将 $\boldsymbol{Z}^C$ 和 N 张干净的搜索区域 $\boldsymbol{S}^C$ 输入孪生网络跟踪器中，得到干净的响应图 $\boldsymbol{G}^C$；

5) 将 $\boldsymbol{S}^C$ 送入生成器中产生噪声 Noise=$\xi_g(\boldsymbol{S}^C)$；

6) 得到对抗搜索区域图像 $\boldsymbol{S}^{\text{adv}}$=$\boldsymbol{S}^C$+Noise；

7) 使用 $\boldsymbol{S}^{\text{adv}}$ 得到对抗响应图 $\boldsymbol{G}^{\text{adv}}$、回归图 $\boldsymbol{M}^{\text{adv}}$ 和对抗搜索区域 $\boldsymbol{S}^{\text{adv}}$ 的特征 $\boldsymbol{E}(\boldsymbol{S}^{\text{adv}})$；

8) 基于特征 $\boldsymbol{E}(\boldsymbol{S}^{\text{adv}})$，通过通道注意力和空间注意力协同机制得到特征权重分布；

9) 基于 $\boldsymbol{G}^{\text{adv}}$、$\boldsymbol{M}^{\text{adv}}$、$\boldsymbol{S}^{\text{adv}}$、$\boldsymbol{E}(\boldsymbol{S}^{\text{adv}})$和 $\boldsymbol{N}_t$，根据式(1)、式(3)、式(7)分别计算欺骗损失、漂移损失和基于注意力机制的特征损失；

10) 利用式(9)计算 L_{qualtity} 损失函数；

11) 利用式(10)计算总损失函数；

12) 利用 Adam 优化器更新生成器 ξ_g 的参数；

13) 直到模型收敛；

14) end for

4.2 损失函数的设计

给定一段视频，对干净视频帧添加微量扰动，生成对抗样本，使跟踪器偏离原始运动轨迹。本节从欺骗损失攻击、漂移损失攻击、基于双重注意力机制的特征损失攻击和感知损失 4 个方面详细介绍本文算法的损失函数。

4.2.1 欺骗损失攻击

在单目标跟踪中，跟踪器需要在视频的每一帧中精确定位目标。对于攻击而言，需求与之相反，希望跟踪器在每一帧中都尽量偏离正确的目标位置。由于目标跟踪是集粗定位和细定位于一体的任务，因此本文设计了粗定位和细定位的欺骗损失函数用于迷惑目标跟踪器。

对于粗定位，跟踪的目标是粗略确定目标有可能存在的区域，反之，攻击是使粗定位任务失灵。粗定位确定目标的主要依据来源于正样本，让正样本的置信度分数尽量小就能达到辨认不出目标的目的。对于细定位，跟踪的指引就是基于粗定位的结果结合修正量精准回归边框，在攻击时让回归边框尽量收缩，就能使定位的目标位置不准确，从而降低重叠率，细定位任务也就失去了效果。具体的函数表达式为

$$\begin{aligned} L_{\text{cheat}} &= L_{\text{coarse}} + L_{\text{scale}} = \\ &\rho_1 \max\left(G_+^{\text{adv}} - G_-^{\text{adv}}, \gamma\right) + \\ &\rho_2 \left(\max\left(M_h^{\text{adv}}, \gamma\right) + \max\left(M_w^{\text{adv}}, \gamma\right)\right) \end{aligned} \tag{1}$$

其中，L_{coarse} 表示粗定位任务损失；L_{scale} 表示细定位任务损失；$\boldsymbol{G}^{\text{adv}}$ 表示对抗响应图；G_+^{adv} 表示 $\boldsymbol{G}^{\text{adv}}$ 中 $\tilde{A}$ 所对应候选框的目标分数；G_-^{adv} 表示背景响应值；$G_+^{\text{adv}} - G_-^{\text{adv}}$ 表示使目标与背景尽可能相似，以迷惑跟踪器；$\boldsymbol{M}^{\text{adv}}$ 表示对抗回归图；M_h^{adv} 和 M_w^{adv} 分别表示 $\boldsymbol{M}^{\text{adv}}$ 的高度和宽度修正量，通过优化 M_h^{adv} 和 M_w^{adv}，跟踪器预测的边框逐渐收缩，使其无法精确定位目标；ρ_1 和 ρ_2 表示平衡粗定位和细定位任务之间的权重系数；γ 表示设定的固定值，防止欺骗损失函数无限制减小，保持训练稳定。

在图 1 中，A 模块利用干净模板和干净搜索区域产生 m 个候选框 $\tau = [\tau_1, \tau_2, \tau_3, \cdots, \tau_m]$，再依据每个候选框的置信度 s 寻找感兴趣区域。本文将置信度大于 0.7 的候选框作为正样本，在干净响应图中计算对应索引，作为注意力掩码 $\tilde{A}$，定义为

$$\tilde{A} = \begin{cases} 1 & , s \geqslant 0.7 \\ 0 & , s < 0.7 \end{cases} \tag{2}$$

进一步，利用式(2)寻找 $\boldsymbol{G}^{\text{adv}}$ 和 $\boldsymbol{M}^{\text{adv}}$ 中相应候选框，再计算 G_+^{adv}、G_-^{adv}、M_h^{adv} 以及 M_w^{adv}，得到 L_{coarse} 与 L_{scale}。

4.2.2 漂移损失攻击

欺骗损失攻击旨在冷却干净搜索区域中可能是目标的区域，使跟踪器难以辨认目标，同时，尽可能减小修正量的宽和高，收缩目标边界框，降低重叠率。然而，该攻击还不够强大，跟踪器仍然可以在搜索区域内定位出物体。针对这一问题，本节提出了漂移损失函数，通过赋予中心坐标修正量很大的漂移值，目标预测边框中心会与原始中心相差甚远，导致跟踪器快速丢失目标。漂移损失函数表达式为

$$L_{\text{drift}} = \left(M_{cx}^{\text{adv}} - \Delta\delta\right) + \left(M_{cy}^{\text{adv}} - \Delta\delta\right) \tag{3}$$

其中，M_{cx}^{adv} 和 M_{cy}^{adv} 表示 $\boldsymbol{M}^{\text{adv}}$ 的中心坐标修正量；$\Delta\delta$ 表示预先设置的漂移值。通过漂移损失攻击，目标预测边框的中心将逐渐偏离原始目标中心，增强了对目标的攻击强度。

欺骗漂移攻击框架如图 2 所示，进一步细化了欺骗损失攻击和漂移损失攻击。为方便计算，先将

$\boldsymbol{G}^{\text{adv}}$ 和 $\boldsymbol{M}^{\text{adv}}$ 调整为二维矩阵，再利用式(2)选择的候选框产生粗定位结果(G_{+}^{adv} , G_{-}^{adv})和细定位结果(M_{cx}^{adv} , M_{cy}^{adv} , M_{h}^{adv} , M_{w}^{adv})。

4.2.3　基于双重注意力机制的特征损失攻击

欺骗损失攻击和漂移损失攻击都着眼于对高级类别信息进行攻击，这依赖于特定白盒模型产生的分类概率和回归预测，迁移能力受到了限制。考虑任何跟踪器都需利用图像底层特征作为网络输入，对图像底层特征攻击有助于提高白盒模型产生的对抗样本在黑盒模型跟踪器上的迁移能力。因此，提出了特征损失攻击函数，定义为

$$L_{\text{feature}} = \sum_{k=1}^{C} \left\| \boldsymbol{E}(\boldsymbol{S}^{\text{adv}}) - \boldsymbol{E}(\boldsymbol{N}_t) \right\|_2 \tag{4}$$

其中，$\boldsymbol{E}(\cdot)$表示图像经过骨干网络输出的特征图；$\boldsymbol{S}^{\text{adv}}$ 表示搜索区域的对抗样本；C 表示通道数量；$\boldsymbol{N}_t$ 表示自定义的噪声图像。通过优化对抗样本特征和自定义噪声图像特征之间的欧氏距离，使对抗样本特征与噪声图像特征相似，以改变特征空间中对抗样本的内部结构。

进一步地，为了增强对目标的攻击强度，该特征损失攻击融合了空间和通道注意力，构成双重注意力模块，以聚焦图像中感兴趣的区域，如图 3 所示。

1) 空间注意力模块。卷积神经网络的输出特征图存在空间内的依赖关系，本文利用这种关系产生空间注意力图，以关注目标具体位置。在单目标跟踪中，仅有一个感兴趣目标，关注图像中前景区域尤为重要，这有利于捕获关键信息，增强对目标攻击的强度。针对每一个感兴趣区域，空间注意力机制表达式为

$$\mathbf{SA}(i) = s(i)(h(\text{ROI}(i))w(\text{ROI}(i))) \tag{5}$$

其中，ROI 表示感兴趣区域；i 表示第 i 个 ROI；s 表示置信度；$h()$和 $w()$分别表示 ROI 的高度和宽度。对于每幅干净图像，首先依据 s 寻找前 40 个感兴趣区域，获取其坐标和相应置信度，然后将这些区域

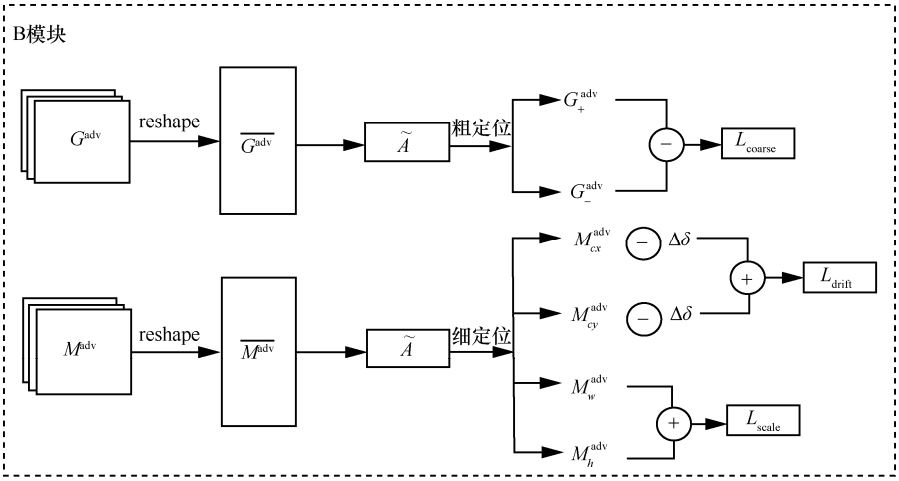

图 2　欺骗漂移攻击框架

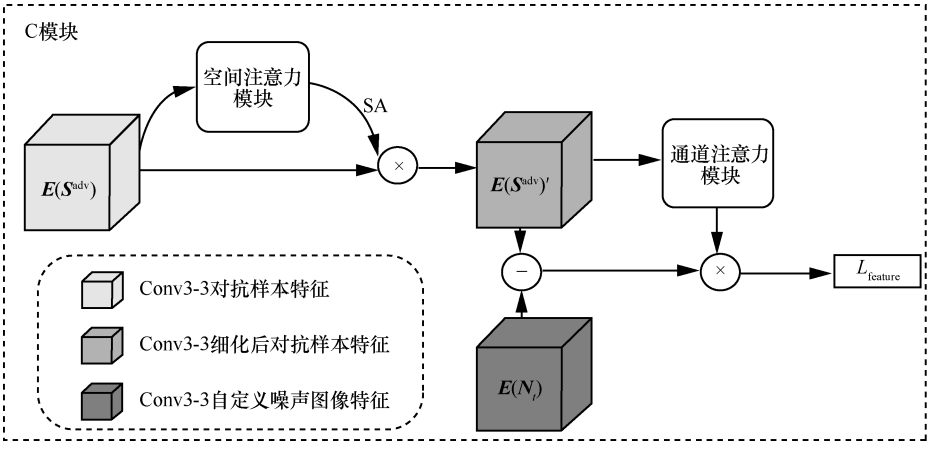

图 3　特征攻击框架

映射到 $\boldsymbol{E}(\boldsymbol{S}^{\text{adv}})$中，通过叠加 $\boldsymbol{E}(\boldsymbol{S}^{\text{adv}})$中每个响应值所包含 ROI 的所有置信度来确定最终空间注意力图 **SA**，最后得到细化后的对抗样本特征图，具体表达式为

$$\boldsymbol{E}(\boldsymbol{S}^{\text{adv}})' = \boldsymbol{E}(\boldsymbol{S}^{\text{adv}}) \otimes \boldsymbol{SA} \tag{6}$$

其中，$\boldsymbol{E}(\boldsymbol{S}^{\text{adv}})'$表示细化后的特征图；⊗表示像素相乘。

2) 通道注意力模块。孪生网络输出特征图的各通道之间存在依赖性，不同通道对于每个类别的响应强度差异很大，每个通道所蕴含的信息量也有所不同。对目标攻击而言，与目标关联度越大，对应特征通道应赋予更多扰动，以关注信息量更丰富的通道，抑制信息量小的通道。为了更关注目标，本文融合了双重注意力机制来攻击图像中重要区域的特征，得到各通道的特征权重分布，实现对目标的攻击。因此，融合通道和空间注意力协同机制的式(4)可进一步表示为

$$L_{\text{feature}} = \sum_{k=1}^{C} \left\| \boldsymbol{A}_k \circ \left(\boldsymbol{E}(\boldsymbol{S}^{\text{adv}})' - \boldsymbol{E}(\boldsymbol{N}_t) \right) \right\|_2 \tag{7}$$

其中，∘表示哈达玛积，$\boldsymbol{A}_k$表示各通道特征权重分布，计算式为

$$\boldsymbol{A}_k = \left(\frac{1}{HW} \sum_{i=1}^{H} \sum_{j=1}^{W} v_k(i,j) \right) \tag{8}$$

其中，$v_k(i,j)$表示第 k 个通道特征图中的响应值；H 和 W 分别表示特征图的高和宽。先将各个通道全局平均池化后获得每个通道特征图的相对重要程度，再将其送入 Softmax 激活函数，得到每个通道的特征权重分布。注意力特征权重的计算提高了对抗样本关键通道的特征图与自定义噪声图像特征图之间的相似程度，同时也抑制了贡献小的通道对扰动的影响，增强了对目标攻击的强度。

4.2.4 感知损失

为了使生成图像更接近于原始图像，防止图像失真，本文引入了感知损失函数 L_{quality}，定义为

$$L_{\text{quality}} = \frac{1}{N} \left\| \boldsymbol{S}^{\text{adv}} - \boldsymbol{S}^{C} \right\|^2 \tag{9}$$

其中，$\boldsymbol{S}^{C}$ 表示干净搜索区域；$\boldsymbol{S}^{\text{adv}}$ 表示对抗搜索区域；N 表示搜索区域的数量。利用 2-范数对噪声幅度进行约束，使原始图像转化为能够欺骗目标模型的对抗图像。

最后，将欺骗损失 L_{cheat}、漂移损失 L_{drift}、特征损失 L_{feature} 和感知损失函数 L_{quality} 损失组合成总的损失函数 L

$$L = L_{\text{cheat}} + \lambda_1 L_{\text{drift}} + \lambda_2 L_{\text{feature}} + \lambda_3 L_{\text{quality}} \tag{10}$$

其中，λ_1、λ_2 和 λ_3 表示权重系数，以平衡不同的损失函数。本文中，这些权重系数值是依据大量实验设置的参数，通过扰动生成器产生的扰动图像既不被发现，又能有效地欺骗跟踪器。

4.3 损失函数作用

攻击目标的场景包括遮挡、尺度变化、光照变化、背景干扰等。本文的 4 种损失函数针对不同任务场景设计，所起的作用有主次之分。特征损失函数旨在破坏图像特征结构，使模型具有更好的迁移性，让白盒模型训练的扰动生成器以更高的成功率和精度迁移到其他黑盒跟踪器中，产生好的攻击效果，增强攻击泛化性。欺骗损失同时干扰目标的分类和回归任务，通过干扰分类响应图使跟踪器误把背景当作目标，并使回归边框收缩，降低与原目标的重合率，涵盖 RPN 中的分类与回归阶段。漂移损失集中攻击回归任务，主要解决欺骗损失带来的攻击偏移度差问题，进一步增强攻击强度，使目标大幅偏离原始预测位置。此外，由于数字空间中的对抗攻击遵循扰动不可见原则，因此本文设计了感知损失，它能使产生的扰动图像尽可能与原图像相似，使人眼不可察觉。

5 实验与结果分析

本节将本文提出的方法在 Pytorch 深度学习架构下开展验证，硬件平台的配置环境为 Intel-i9 CPU（64 GB 内存）和一块 RTX-2080Ti GPU（11 GB 内存），并且在 3 个数据集（OTB100、VOT2018 和 LaSOT）上测试了本文方法的有效性。

5.1 数据集

训练数据集：为了涵盖更丰富的目标类别，本文采用 GOT-10K 作为训练数据集。该数据集的视频序列超过 10 000 个，覆盖 500 多个目标类别，呈现出跟踪目标的多样性。具体地，对于每个视频序列，在视频的第一帧裁剪目标模板，在后续的帧中每 10 帧均匀采样一次，并裁剪搜索区域，其中模板区域大小裁剪为 127×127，搜索区域大小裁剪为 255×255。

测试数据集：本文将在 OTB100、VOT2018 和 LaSOT 这 3 个数据集上测试本文方法的有效性。下面，从数据集大小和数据特点等方面分别介绍这 3 个

数据集。

OTB100 数据集：该数据集中共有 98 个视频，涉及目标跟踪的 11 个属性，包括光照变化、尺度变化、遮挡、形变、运动模糊、快速运动等。每个序列都对应 2 个或多个属性。

VOT2018 数据集：该数据集中包括 60 个视频，与 OTB 数据集相比，更具挑战性。在目标丢失时，该数据集有重新初始化机制。

LaSOT 数据集：包含 1 400 个视频；目标类别有 70 个，每个类别包含 20 个序列。其中测试集由每个类别中精心挑选的 4 个视频序列组成，共计 280 个视频序列。

5.2　评价标准

OTB100 数据集采用精确度（P, precision）和成功率（S, success）作为评价标准。P 反映跟踪算法估计的目标位置中心点（bounding box）与人工标注目标中心点（ground-truth）的中心误差。S 代表跟踪算法得到的预测状态与目标原始重合率大于 0.5 的百分比。VOT2018 数据集同时衡量算法的精确度（A, accuracy）和稳健性（R, robustness），并以平均重叠期望（EAO, expected average overlap）给出算法性能的排序。LaSOT 数据集选择 S 和标准化精度（Norm P, norm precision）来衡量算法性能。

5.3　实验细节

本文使用 Adam 优化算法优化生成器，学习率设置为 2×10^{-4}。将欺骗损失中的 γ 设置为−5，并将 ρ_1 和 ρ_2 分别设为 0.1 和 1，以平衡粗定位与细定位损失。将漂移损失中的 $\Delta\delta$ 设为 500，使边框大幅偏移目标中心。式(10)中的漂移损失系数 λ_1 和特征损失权重 λ_2 分别设置为 2 和 20，感知损失系数 λ_3 设置为 620。

对于攻击生成的施加条件，本文攻击方法需要 2 个部件，分别为 U-net 结构的生成器以及 ResNet50 结构的 SiamRPN++跟踪器。U-net 结构在像素级任务中展现优异的性能，因此适合为数字空间中的攻击任务产生噪声。整个攻击生成的施加条件作用于白盒设置模式下，能获取 SiamRPN++跟踪器的全部参数，以产生高级语义层面与低级特征层面的多级损失函数用于扰动生成器的训练，从而确保训练的扰动生成器能够成功攻击图像中的目标。

5.4　实验结果与分析

表 1 和表 2 给出了本文方法在 OTB100[38]、VOT2018 和 LaSOT[39]这 3 个数据集上的攻击结果。表 1 中，Original 表示 SiamRPN++原始的跟踪结果，Attack S 表示仅攻击搜索区域，Drop 表示性能下降；表 2 中，Attack SZ 表示同时攻击搜索区域和目标模板。攻击策略为仅攻击搜索区域以及同时攻击搜索区域和目标模板。

表 1　本文方法对仅攻击搜索区域的实验结果

数据集	Metric	Original	Attack S	Drop
OTB100	S (↑)	0.696	0.160	0.536
	P (↑)	0.914	0.221	0.693
VOT2018	A (↑)	0.600	0.425	0.175
	R (↓)	0.234	2.992	2.758
	EAO(↑)	0.414	0.052	0.362
LaSOT	P (↑)	0.569	0.062	0.507
	S (↑)	0.496	0.061	0.435

表 2　本文方法同时攻击搜索区域和目标模板的实验结果

数据集	Metric	Original	Attack SZ	Drop
OTB100	S (↑)	0.696	0.155	0.541
	P (↑)	0.914	0.218	0.696
VOT2018	A (↑)	0.600	0.395	0.205
	R (↓)	0.234	3.030	2.796
	EAO(↑)	0.414	0.047	0.367
LaSOT	S (↑)	0.569	0.055	0.514
	Norm P (↑)	0.496	0.056	0.440

攻击生成的具体过程如下。

1) 仅攻击搜索区域。当仅攻击搜索区域时，预先处理训练数据集，每 10 帧均匀采样 GOT-10K 数据集中视频帧，裁剪目标模板和搜索区域，共获得 9 350 段视频。对于每段视频，第一帧为干净目标模板，后续帧为干净搜索区域。训练阶段，首先获取目标模板和搜索区域，保持目标模板不变，随机初始化扰动生成器，将每一帧干净目标搜索区域送入扰动生成器产生噪声后，再与干净目标搜索区域相加，形成对抗搜索区域。然后分别将干净搜索区域，对抗搜索区域与干净目标模板送入 SiamRPN++跟踪模型，输出网络 Conv3-3 的特征图，分别得到干净样本的响应图 $\boldsymbol{G}^{C}$、对抗样本的响应图 $\boldsymbol{G}^{\mathrm{adv}}$ 和回归图 $\boldsymbol{M}^{\mathrm{adv}}$。最后构造 L_{cheat}、L_{drift}、L_{feature} 和 L_{quality} 这 4 种损失函数以联合训练生成器，得到基于搜索区域的扰动生成器。在推理阶段，保持干净目标模板不变，将干净搜索区域通过扰动生成器生成对抗搜索区域，再把对抗搜索区域和干净目标模板同时送入 SiamRPN++中，得到两者匹配的相似度，记为

SiamRPNP+++S（仅攻击搜索区域）。从表 1 可以看出，本文提出的攻击方法使跟踪器的性能在 3 个数据集上大幅度下降。

2) 同时攻击搜索区域和目标模板。当同时攻击搜索区域和目标模板时，扰动生成器的训练方法和仅攻击目标搜索区域时的训练方法相同。在推理阶段，使用训练的扰动生成器同时攻击目标模板和目标搜索区域，并将对抗模板和对抗搜索区域送入跟踪器 SiamRPN++中，记为 SiamRPN+++SZ（同时攻击搜索区域和目标模板），实验结果如表 2 所示。从表 2 中可以看出，同时攻击模板和搜索区域比仅攻击搜索区域性能下降更多。在 OTB100 数据集上，同时攻击模板和搜索区域时，SiamRPN++跟踪器定位的成功率由未攻击时的 69.6%下降为 15.5%，降低了约 54%；精确度由 91.4%下降到 21.8%，降低了约 70%。

另外，将 SiamRPN++以及 2 种攻击策略下的 SiamRPN+++S、SiamRPN+++SZ 在 OTB 数据集上与其他主流跟踪器（MDNet[12]、SiamFC[13]、SiamRPN[14]、SiamRPN++[15]、DaSiamRPN[16]、GradNet[40]等）进行对比，性能表现如图 4 所示。

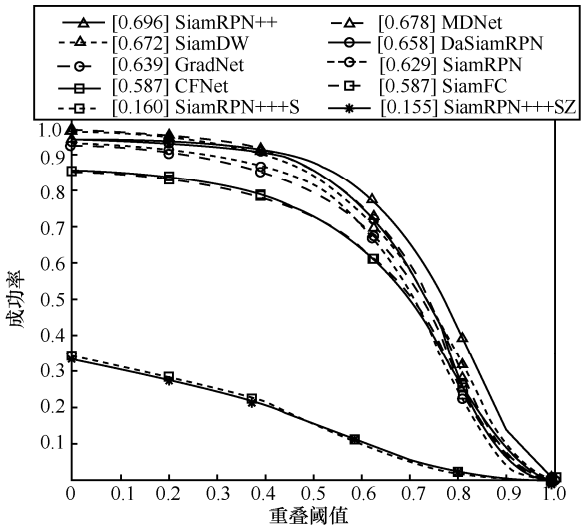

图 4　OTB 数据集上跟踪成功率的性能比较

从图 4 中可知，本文方法大大降低了 SiamRPN++的性能。表 3 给出了本文方法与现有攻击方法在 SiamRPN++上的攻击性能比较。在 OTB100 数据集上，本文方法在 S 和 P 上都超越了 CSA[33]、SPARK[41]与 FAN 方法[42]，使 SiamRPN++的定位能力显著下降，在 OTB100 数据集上仅有 15.5%的成功率和 21.8%的精确度。

表 3　不同攻击方法在 SiamRPN++跟踪器上的攻击性能比较

攻击方法	S(↑)	P(↑)
CSA	0.324	0.471
SPARK	0.630	0.887
FAN	0.306	0.544
本文方法	0.155	0.218

5.5　消融实验

5.5.1　损失项消融实验

为了验证并分析模型各损失项的作用，将损失项分为 4 个部分，分别是粗定位损失 L_{coarse}、细定位损失 L_{scale}、漂移损失 L_{drift} 和特征损失 L_{feature}。L_{coarse} 用于冷却目标位置；L_{scale} 用于收缩目标边界框，降低重叠率；L_{drift} 用来漂移目标；L_{feature} 用来改变图像在特征空间中的结构。本节分析了这 4 个损失项及其组合对于 SiamRPN++跟踪器性能的影响。实验在 OTB100 和 VOT2018 数据集上进行了测试，结果如表 4 和表 5 所示。其中，“－”表示未使用，“√”表示使用。从表 4 和表 5 中可以看出，无论是使用单独损失项还是组合项，同时攻击搜索区域和目标模板都取得了比仅攻击搜索区域更强的攻击效果。

表 4　仅攻击搜索区域时各损失项对于性能的影响

L_{coarse}	L_{scale}	L_{drift}	L_{feature}	OTB100		VOT2018
				S(↑)	P(↑)	EAO(↑)
－	－	－	－	0.696	0.914	0.414
√	－	－	－	0.421	0.587	0.099
－	√	－	－	0.443	0.642	0.149
－	－	√	－	0.587	0.791	0.159
－	－	－	√	0.470	0.655	0.142
√	√	－	－	0.349	0.491	0.073
√	√	√	－	0.302	0.423	0.061
√	√	√	√	0.160	0.221	0.047

首先，针对 4 个单独损失项，其对迷惑跟踪器均有积极影响，L_{coarse} 取得了最佳的攻击效果，证明了粗定位任务在跟踪中的重要性。L_{scale} 和 L_{drift} 的攻击效果次于 L_{coarse}，这是因为两者都是基于粗定位结果进行收缩或漂移，去掉粗定位损失 L_{coarse}，跟踪器便能粗略确定目标位置，此基础上进行单独收缩或漂移，效果自然欠佳。

表 5　同时攻击搜索区域和目标模板时各损失项对于性能的影响

L_{coarse}	L_{scale}	L_{drift}	$L_{feature}$	OTB100		VOT2018
				S(↑)	P(↑)	EAO(↑)
—	—	—	—	0.696	0.914	0.414
√	—	—	—	0.406	0.562	0.081
—	√	—	—	0.432	0.621	0.130
—	—	√	—	0.566	0.774	0.161
—	—	—	√	0.436	0.598	0.122
√	√	—	—	0.325	0.472	0.073
√	√	√	—	0.259	0.378	0.056
√	√	√	√	0.155	0.218	0.047

其次，对于损失项的组合，在 L_{coarse} 的基础上添加 L_{scale}，跟踪器的预测边框逐渐收缩，无法精确估计目标尺度，如图 5(a)所示。此外，在欺骗损失 L_{coarse} 和 L_{scale} 的基础上再叠加漂移损失 L_{drift}，跟踪器对目标位置信息极度不敏感，并很快沿着与目标距离最远的位置漂移，如图 5(b)所示。

(a) 边框收缩

(b) 边框漂移

图 5　可视化结果

最后，同时联合 4 项损失能最大幅度地愚弄跟踪器，且生成的对抗样本不易被人眼所察觉。

5.5.2　注意力机制消融实验

本节讨论了特征损失中空间和通道注意力协同机制对目标攻击效果的影响。为了探索 2 种注意力机制及其组合对于性能下降的影响，设计了仅攻击搜索区域和同时攻击搜索区域及目标模板 2 种攻击策略下的对比实验。实验结果如图 6 所示。

仅攻击搜索区域时，设计的实验有 G-S-noA-Feature（G 表示生成器，S 表示搜索区域，noA 表示无注意力机制）、G-S-Spatial-Feature（Spatial 表示执行空间注意力）、G-S-Channel-Feature（Channel 表示执行通道注意力）和 G-S-Spatial-Channel-Feature（Spatial-Channel 表示空间和通道注意力协同机制）。

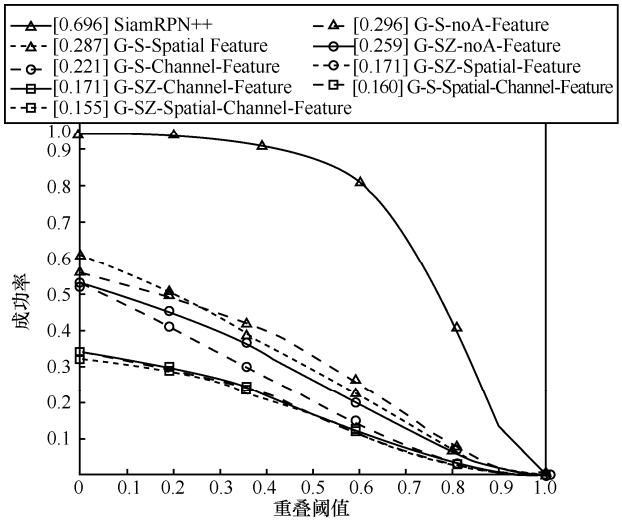

图 6　注意力机制对跟踪器性能的影响

同时攻击搜索区域和目标模板时，设计的实验有 G-SZ-Channel-Spatial-Feature（SZ 表示搜索区域和目标模板）、G-SZ-Channel-Feature、G-SZ-Spatial-Channel-Feature、G-SZ-Spatial-Feature 和 G-SZ-noA-Feature。

从图 6 中可以看出，在 2 种攻击策略下，空间和通道注意力机制对攻击跟踪器都有积极作用，仅有单个注意力时，通道注意力比空间注意力有更好的攻击效果。将二者串联协同工作，跟踪器跟踪目标的性能会大幅下降，达到了最好的攻击效果。

5.5.3　各损失项迁移性消融实验

为了验证各损失项在黑盒攻击设置下的迁移性能，本文选取 3 种最先进的跟踪器进行攻击，将由基于 ResNet-50 的 SiamRPN++训练的生成器迁移到其他 3 种最先进的跟踪器上，分别是基于在线更新策略的 DiMP-50、基于 MobileNet 的 SiamRPN++和基于 ResNet 的 SiamMask。各损失项对于攻击 3 种最先进的黑盒跟踪器时迁移性的表现如表 6～表 8 所示。从表 6～表 8 可以看出，仅用单一损失训练时，4 种损失项训练的模型都能一定程度上降低跟踪器的性能，使目标偏移原本的运动轨迹。较之其余 3 项，$L_{feature}$ 损失项训练的模型在 3 种黑盒跟踪器上都展现出最好的攻击效果，体现出良好的迁移性。

表 6　SiamRPN++(MobileNet)跟踪器各损失项迁移性对比

L_{coarse}	L_{scale}	L_{drift}	$L_{feature}$	OTB100	
				S (↑)	P (↑)
—	—	—	—	0.651	0.855
√	—	—	—	0.314	0.463
—	√	—	—	0.346	0.544
—	—	√	—	0.591	0.796
—	—	—	√	0.306	0.439
√	√	—	—	0.263	0.406
√	√	√	—	0.134	0.168
√	√	√	√	0.098	0.154

表 7　SiamMask 跟踪器各损失项迁移性对比

L_{coarse}	L_{scale}	L_{drift}	$L_{feature}$	OTB100	
				S (↑)	P (↑)
—	—	—	—	0.651	0.855
√	—	—	—	0.397	0.557
—	√	—	—	0.421	0.614
—	—	√	—	0.592	0.800
—	—	—	√	0.353	0.497
√	√	—	—	0.330	0.489
√	√	√	—	0.144	0.190
√	√	√	√	0.120	0.162

表 8　DiMP-50 跟踪器各损失项迁移性对比

L_{coarse}	L_{scale}	L_{drift}	$L_{feature}$	OTB100	
				S (↑)	P (↑)
—	—	—	—	0.686	0.899
√	—	—	—	0.660	0.868
—	√	—	—	0.657	0.864
—	—	√	—	0.644	0.853
—	—	—	√	0.596	0.765
√	√	—	—	0.638	0.839
√	√	√	—	0.472	0.641
√	√	√	√	0.441	0.611

5.5.4　攻击生成所付出代价分析

本文所提出的攻击方法包括欺骗损失、漂移损失、特征损失和感知损失。在生成器训练过程中，当无特征损失时，仅需要 2 h 就能完成对所有视频的训练，得到扰动生成器。添加特征损失时，需要 8 h 才能完成整个训练过程。这是由于特征损失涉及对特征图间的逐像素操作，且利用空间和通道注意力机制探索空间通道的依赖关系，寻找感兴趣区域，从而造成计算成本提高。尽管如此，本文方法对 SiamRPN++及其他最先进跟踪器都能取得良好的攻击效果。如表 4 所示，对于 SiamRPN++跟踪器，添加特征损失比不添加时攻击成功率高出 14.2%，精确度高 20.2%，故虽付出了一定计算代价，却能有效欺骗跟踪器，使跟踪器偏离原始运动轨迹。所付出的计算代价是可容忍的。

6　结束语

针对现有对抗扰动技术难以有效地降低跟踪器的判别能力使运动轨迹发生快速偏移的问题，本文提出一种高效的攻击目标跟踪器的方法。首先，所提方法从高层类别和底层特征考虑设计了欺骗损失、漂移损失和基于注意力机制的特征损失来联合训练生成器，使其拥有对抗扰动的能力；然后，在对一段视频序列攻击时，将每帧干净图像送入该生成器中，生成对抗样本，以干扰 SiamRPN 目标跟踪器，使其运动轨迹发生偏移，导致跟踪失败。所提方法在 OTB100、VOT2018 和 LaSOT 这 3 个主流的目标跟踪数据集进行了验证，相较于对比方法，本文方法达到了较好的攻击效果。

参考文献：

[1] BOLME D S, BEVERIDGE J R, DRAPER B A, et al. Visual object tracking using adaptive correlation filters[C]//Proceedings of 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2010: 2544-2550.

[2] HENRIQUES J F, CASEIRO R, MARTINS P, et al. High-speed tracking with kernelized correlation filters[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2015, 37(3): 583-596.

[3] DANELLJAN M, HÄGER G, SHAHBAZ K F, et al. Accurate scale estimation for robust visual tracking[C]//Proceedings of Proceedings of the British Machine Vision Conference 2014. [S.n.:s.l.], 2014: 1-11.

[4] LI Y, ZHU J K, HOI S C H. Reliable patch trackers: robust visual tracking by exploiting reliable patches[C]//Proceedings of 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE Press, 2015: 353-361.

[5] DANELLJAN M, HÄGER G, KHAN F S, et al. Learning spatially regularized correlation filters for visual tracking[C]//Proceedings of 2015 IEEE International Conference on Computer Vision (ICCV). Piscataway: IEEE Press, 2015: 4310-4318.

[6] DANELLJAN M, ROBINSON A, SHAHBAZ K, et al. Beyond correlation filters: learning continuous convolution operators for visual tracking[C]//Proceedings of the European Conference on Computer

Vision. Berlin: Springer, 2016: 472-488.

[7] WANG N, YEUNG D Y. Learning a deep compact image representation for visual tracking[C]//Proceedings of the Annual Conference on Neural Information Processing Systems. New York: Curran Associates, 2013: 809-817.

[8] HONG S, YOU T, KWAK S, et al. Online tracking by learning discriminative saliency map with convolutional neural network[C]//Proceedings of the International Conference on Machine Learning. New York: ACM Press, 2015: 597-606.

[9] WANG L J, OUYANG W L, WANG X G, et al. Visual tracking with fully convolutional networks[C]//Proceedings of 2015 IEEE International Conference on Computer Vision (ICCV). Piscataway: IEEE Press, 2015: 3119-3127.

[10] MA C, HUANG J B, YANG X K, et al. Hierarchical convolutional features for visual tracking[C]//Proceedings of 2015 IEEE International Conference on Computer Vision (ICCV). Piscataway: IEEE Press, 2015: 3074-3082.

[11] SONG Y B, MA C, WU X H, et al. VITAL: Visual tracking via adversarial learning[C]//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2018: 8990-8999.

[12] NAM H, HAN B. Learning multi-domain convolutional neural networks for visual tracking[C]//Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE Press, 2016: 4293-4302.

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

[14] LI B, YAN J J, WU W, et al. High performance visual tracking with Siamese region proposal network[C]//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2018: 8971-8980.

[15] LI B, WU W, WANG Q, et al. SiamRPN++: evolution of Siamese visual tracking with very deep networks[C]//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE Press, 2019: 4282-4291.

[16] ZHU Z, WANG Q, LI B, et al. Distractor-aware Siamese networks for visual object tracking[C]//Proceedings of the European Conference on Computer Vision. Berlin: Springer, 2018: 101-117.

[17] VOIGTLAENDER P, LUITEN J, TORR P H S, et al. Siam R-CNN: visual tracking by Re-detection[C]//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE Press, 2020: 6577-6587.

[18] GOODFELLOW I J, SHLENS J, SZEGEDY C. Explaining and harnessing adversarial examples[C]//Proceedings of the third International Conference on Learning Representations. Piscataway: IEEE Press, 2015: 1-11.

[19] SEYED M, MOOSAVI D, ALHUSSEIN F, et al. Deepfool: a simple and accurate method to fool deep neural networks[C]//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE Press, 2016: 2574-2582.

[20] XIE C H, WANG J Y, ZHANG Z S, et al. Adversarial examples for semantic segmentation and object detection[C]//Proceedings of 2017 IEEE International Conference on Computer Vision (ICCV). Piscataway: IEEE Press, 2017: 1369-1378.

[21] MADRY A, MAKELOV A, SCHMIDT L, et al. Towards deep learning models resistant to adversarial attacks[C]//Proceedings of the Sixth International Conference on Learning Representations. Piscataway: IEEE Press, 2018: 1-28.

[22] ALEXEY K, IAN G, SAMY B. Adversarial machine learning at scale[J]. arXiv Preprint, arXiv:1611.01236, 2016.

[23] 司念文, 张文林, 屈丹, 等. 基于对抗补丁的可泛化的 Grad-CAM 攻击方法[J]. 通信学报, 2021, 42(3): 23-35.
SI N W, ZHANG W L, QU D, et al. Generalized Grad-CAM attacking method based on adversarial patch[J]. Journal on Communications, 2021, 42(3): 23-35.

[24] SU J W, VARGAS D V, SAKURAI K. One pixel attack for fooling deep neural networks[J]. IEEE Transactions on Evolutionary Computation, 2019, 23(5): 828-841.

[25] ZHONG Y Y, DENG W H. Towards transferable adversarial attack against deep face recognition[J]. IEEE Transactions on Information Forensics and Security, 2021, 16: 1452-1466.

[26] CHEN X S, YAN X Y, ZHENG F, et al. One-shot adversarial attacks on visual tracking with dual attention[C]//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE Press, 2020: 10176-10185.

[27] JIA S, MA C, SONG Y B, et al. Robust tracking against adversarial attacks[C]//Proceedings of the European Conference on Computer Vision. Berlin: Springer, 2020: 69-84.

[28] XIAO C W, LI B, ZHU J Y, et al. Generating adversarial examples with adversarial networks[C]//Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence. [S.l.:s.n.], 2018: 835-857.

[29] WEI X X, LIANG S Y, CHEN N, et al. Transferable adversarial attacks for image and video object detection[C]// Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence. [S.l.:s.n.], 2019: 1-8.

[30] JANDIAL S, MANGLA P, VARSHNEY S, et al. AdvGAN++: harnessing latent layers for adversary generation[C]//Proceedings of 2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW). Piscataway: IEEE Press, 2019: 2045-2048.

[31] DEB D, ZHANG J B, JAIN A K. AdvFaces: adversarial face synthesis[C]//Proceedings of 2020 IEEE International Joint Conference on Biometrics (IJCB). Piscataway: IEEE Press, 2020: 1-10.

[32] BALUJA S, FISCHER I. Adversarial transformation networks: learning to generate adversarial examples[J]. arXiv Preprint, arXiv: 1703.09387, 2017.

[33] YAN B, WANG D, LU H C, et al. Cooling-shrinking attack: blinding the tracker with imperceptible noises[C]//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE Press, 2020: 990-999.

[34] SHARIF M, BHAGAVATULA S, BAUER L, et al. A general framework for adversarial examples with objectives[J]. ACM Transactions on Privacy and Security, 2019, 22(3): 1-30.

[35] CARLINI N, WAGNER D. Towards evaluating the robustness of neural networks[C]//Proceedings of 2017 IEEE Symposium on Security and Privacy (SP). Piscataway: IEEE Press, 2017: 39-57.

[36] MOOSAVI-DEZFOOLI S M, FAWZI A, FAWZI O, et al. Universal adversarial perturbations[C]//Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE Press, 2017: 1765-1773.

[37] DIN S U, AKHTAR N, YOUNIS S, et al. Steganographic universal adversarial perturbations[J]. Pattern Recognition Letters, 2020, 135: 146-152.

[38] WU Y, LIM J, YANG M H. Object tracking benchmark[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2015, 37(9): 1834-1848.

[39] FAN H, LIN L T, YANG F, et al. LaSOT: a high-quality benchmark for large-scale single object tracking[C]//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE Press, 2019: 5374-5383.

[40] LI P X, CHEN B Y, OUYANG W L, et al. GradNet: gradient-guided network for visual object tracking[C]//Proceedings of 2019 IEEE/CVF International Conference on Computer Vision (ICCV). Piscataway: IEEE Press, 2019: 6162-6171.

[41] GUO Q, XIE X F, JUEFEI-XU F, et al. SPARK: spatial-aware online incremental attack against visual tracking[C]//Proceedings of the Conference on European Conference on Computer Vision. Berlin: Springer, 2020: 202-219.

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

[作者简介]

程旭（1983− ），男，山西太原人，博士，南京信息工程大学副教授、硕士生导师，主要研究方向为目标检测与跟踪、图像理解、对抗攻击等。

王莹莹（1999− ），女，江苏盐城人，南京信息工程大学硕士生，主要研究方向为目标跟踪系统的对抗攻击。

张年杰（1997− ），男，江苏泰州人，南京信息工程大学硕士生，主要研究方向为目标跟踪。

付章杰（1983− ），男，河南南阳人，博士，南京信息工程大学教授、博士生导师，主要研究方向为人工智能安全、区块链安全、数字取证等。

陈北京（1981− ），男，江西赣州人，博士，南京信息工程大学教授、博士生导师，主要研究方向为多媒体内容安全、彩色图像处理、模式识别等。

赵国英（1977− ），女，山东聊城人，博士，奥卢大学终身教授、博士生导师，主要研究方向为视频图像处理、模式识别、智能人机交互等。

收录声明

本刊对发表的文章，拥有出版电子版、网络版版权，并拥有和其他网站交换信息的权利。本刊支付的稿酬中已经包含上述费用。

Journal on Communications has the copyright to publish electronic edition, online edition of the published articles, and has the right to exchange information with other sites. The expenses have been included in the fee paid by editorial department.

道德声明

本刊发表的论文是作者独立取得的原创性研究成果，无一稿多投；论文内容不涉及国家机密；未曾以任何形式用任何文种在国内外公开发表过；论文内容不侵犯他人著作权和其他权利。若发生一稿多投、侵权、泄密等问题，论文作者将承担全部责任。

The authors of *Journal on Communications* guarantee that their submitted articles are original and contain nothing confidential. The said article is only submitted to *Journal on Communications*. The said article has not been published before and has not been submitted elsewhere for print or electronic publication consideration. The said article is no way whatever a violation or an infringement of any existing copyright or license from the third party. Otherwise, the authors of the said article shall take the blame for the violation or infringement of the related copyright and the leakage of secrets.

通信学报

Journal on Communications

发行代号：国内2-676 / 国外M395　2021年11月25日出版　定价：98.00元

ISSN 1000-436X