计 算 机 研 究 与 发 展
Journal of Computer Research and Development
https://crad.ict.ac.cn
62(8): 2098−2110, 2025

本期推荐

基于频段跟踪性能差异的对抗检测方法

周 泽[1] 孙颖慧[2] 孙权森[1] 沈肖波[1] 郑钰辉[3]
[1](南京理工大学计算机科学与工程学院 南京 210094)
[2](东南大学计算机科学与工程学院 南京 210096)
[3](南京信息工程大学人工智能学院 南京 210044)
(zhoumze@163.com)

An Adversarial Detection Method Based on Tracking Performance Difference of Frequency Bands

Zhou Ze[1], Sun Yinghui[2], Sun Quansen[1], Shen Xiaobo[1], and Zheng Yuhui[3]
[1](*School of Computer Science and Engineering*, *Nanjing University of Science and Technology*, *Nanjing* 210094)
[2](*School of Computer Science and Engineering*, *Southeast University*, *Nanjing* 210096)
[3](*School of Artificial Intelligence*, *Nanjing University of Information Science and Technology*, *Nanjing* 210044)

Abstract Given the risk of adversarial attacks on tracking models and the lack of relevant adversarial detection methods, this paper addresses the problem from the perspective of frequency domain. Combined with the visual invisible property of perturbation noise, we first theoretically prove that perturbation noise mainly exists in the mid-to-high frequency bands of images. Then we quantitatively analyze that the low-frequency components of the video sequence contribute the most to tracking performance and are least affected by adversarial attacks. Finally, based on the above theoretical proof and qualitative analysis, we propose a detection framework based on the tracking performance difference of frequency bands, in which the frequency domain decomposition module for extracting the low-frequency components of the video sequence. The target tracker and its mirror tracker with the same structure and parameters respectively take the full-frequency and low-frequency components of the video sequence as input. The discriminator module determines whether the input video sequence is an adversarial input by comparing the output differences of the two trackers. This detection framework uses a tracker as a carrier and does not require adversarial training. It can achieve adversarial detection only by comparing the tracking performance difference across different frequency bands. Extensive experimental results show that the detection framework can not only effectively detect current mainstream adversarial attacks, such as CSA, TTP, and Spark with a detection precision of 97.55%, but also has little negative impact on the original tracking performance of the tracker. In addition, this framework is generalizable and can be flexibly integrated into multiple trackers, such as SiamRPNpp, SiamMask, SiamCAR, and SiamBAN.

Key words object tracking; adversarial attacks; adversarial detection; frequency domain decomposition; performance difference

收稿日期: 2024-06-03; **修回日期**: 2024-12-31

基金项目: 国家自然科学基金项目(62372235, 62406069, 62472226); 江苏省自然科学基金优秀青年基金项目(BK20230095); 中国博士后科学基金项目(2024M750425)

This work was supported by the National Natural Science Foundation of China (62372235, 62406069, 62472226), the Jiangsu Natural Science Foundation Excellent Youth Fund (BK20230095), and the China Postdoctoral Science Foundation (2024M750425).

通信作者: 孙权森(sunquansen@njust.edu.cn)

摘　要　针对跟踪模型存在被对抗攻击的风险且当前缺少相关的对抗检测方法，利用频域手段解决这一问题. 结合扰动噪声视觉不可见的特点，首先理论证明了扰动噪声主要存在于图像的中高频段. 然后定量地分析出视频序列的低频分量对跟踪性能的贡献最大且受对抗攻击的影响最小. 最后根据上述理论证明和定量分析，提出了基于频段跟踪性能差异的检测框架，其中的频域分解模块用于提取视频序列的低频段分量；目标跟踪器及其同构同参的镜像跟踪器分别以视频序列的全频段和低频段分量为输入；判别模块通过对比2个跟踪器的输出差异，判定当前视频序列是否为对抗输入. 该检测框架以跟踪器为载体，无需对抗训练，仅通过对比不同频段跟踪性能的差异性，即可实现对抗检测. 大量的实验结果表明提出的检测框架不仅能够有效地检测当前主流的对抗攻击，如CSA, TTP, Spark，检测精度高达97.55%，而且对跟踪器的原始跟踪性能影响较小. 此外，检测框架具有泛化性，能够灵活地集成到多个跟踪器，如SiamRPNpp, SiamMask, SiamCAR, SiamBAN.

关键词　目标跟踪；对抗攻击；对抗检测；频域分解；性能差异

中图法分类号　TP391

DOI: 10.7544/issn1000-1239.202440428　　**CSTR:** 32373.14.issn1000-1239.202440428

近几年的研究表明深度神经网络(deep neural network, DNN)受到对抗攻击时表现不佳[1-3]. 当前的跟踪模型大多依赖DNN, Yan等人[4]和Nakka等人[5]也证实了基于DNN的跟踪模型受到对抗攻击时，很难成功跟踪目标. 视觉跟踪技术被广泛地应用在与安全相关的领域，如智能监控、自动驾驶等方面. 因此抵御对抗攻击具有理论研究和实际应用的重要意义. 目前有3种策略抵御对抗攻击：对抗防御，旨在使深度模型受到对抗攻击时仍能输出正确结果；对抗检测，旨在检测深度模型的输入是否受到对抗攻击；对抗防御和对抗检测的结合. 本文重点研究对抗检测，即检测跟踪器的输入视频序列是否为对抗输入.

跟踪领域的对抗攻击是指攻击者通过向原始视频序列中添加扰动噪声并合成对抗视频序列，导致跟踪器难以在对抗视频序列上成功地跟踪目标. 目前针对跟踪领域的对抗检测方法尚未提出. 现有的对抗检测方法[6-13]大多服务于识别领域. 它们很难直接应用到跟踪领域，其主要原因在于：1)处理对象不同. 识别任务涉及的是互不相关的静态图像，检测的对象也是独立图像；跟踪任务处理的是视频序列，视频帧之间的相关性较大，检测的对象是视频序列. 2)攻击方式不同. 对于识别任务，通常只需要向单个图像中添加扰动噪声，就能实现攻击效果；对于跟踪任务，通常需要向整个视频序列中添加扰动噪声，才能实现显著的攻击效果. 针对如何设计适用于跟踪任务的对抗检测方案，本文认为实现对抗检测的关键在于要明确攻击者是如何攻击跟踪器的. 为了有效地攻击跟踪器同时视觉上欺骗用户，跟踪领域的对抗攻击方法[4-5,14-24]生成的扰动噪声几乎是视觉不可见的. 如图1所示，原始图像和带有扰动噪声的对抗图像，视觉上几乎没有差异，但扰动噪声却能成功地干扰跟踪器(虚线框表示精准地定位到目标，实线框则远离目标). 这说明扰动噪声对人眼视觉不可见但对深度跟踪模型是“可见的”. 那么视觉上，扰动噪声被添加在图像的什么地方需要探索.

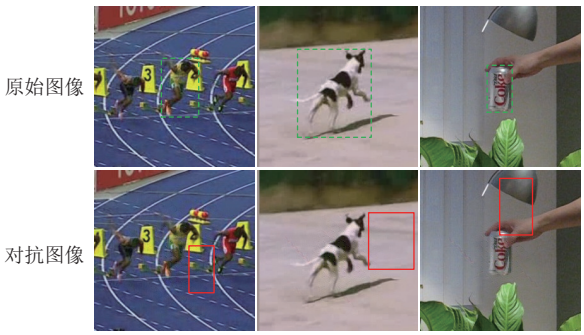


Fig. 1　Visual comparison between original images and adversarial images

图1　原始图像和对抗图像视觉对比

生物学指出人眼只能观察到特定频域内的信息. 受此启发，首先借助离散傅里叶变换(discrete Fourier transform, DFT)理论证明了扰动噪声主要被添加在图像的中高频段，这些是视觉难以捕捉的频域，但跟踪模型却对其较为敏感. 为了佐证低频段受扰动噪声影响较小，我们定量地研究了攻击前后视频序列各频段分量的贡献，并分析出低频段分量对跟踪性能的贡献最大且受对抗攻击的影响最小. 基于上述的理论证明和定量研究，本文以SiamRPNpp跟踪器[25]为主要研究对象，设计了一个简单有效的对抗

检测框架, 主要包含频域分解模块、目标跟踪器、判别模块及其同构同参的镜像跟踪器. 频域分解模块负责提取视频序列的低频段分量, 并将其作为镜像跟踪器的输入; 而目标跟踪器则以视频序列的全频段分量为输入; 判别模块通过对比 2 个跟踪器的输出差异性, 判定输入视频序列是否为对抗输入. 该检测框架无需使用对抗样本进行对抗训练, 即可有效地检测多种对抗攻击, 且可以灵活地集成到多个跟踪器.

论文的贡献主要包括 3 点:

1)提出了适用于跟踪任务的对抗检测框架, 无需对抗训练, 仅通过对比视频序列不同频段分量间的跟踪性能差异性, 即可有效地检测输入视频序列是否为对抗输入.

2)理论证明了跟踪领域的对抗攻击方法主要将扰动噪声添加在视频序列的中高频段; 并定量地分析出视频序列的低频段分量对跟踪性能的贡献最大且受对抗攻击的影响最小.

3)将检测框架灵活地集成到多个跟踪器, 大量的实验结果表明检测框架不仅能够有效地检测主流的对抗攻击而且对跟踪器的原始性能影响较小.

1 相关工作

1.1 跟踪领域的对抗攻击

自 Wiyatno 等人[14] 于 2019 年将对抗纹理应用到视频序列并成功地干扰跟踪过程以来, 近些年陆续涌现出许多对抗攻击方法[4-5,15-24]. 它们通过将扰动噪声添加到原始视频序列, 导致跟踪器难以在对抗视频序列上跟踪到真实目标, 极大地降低了跟踪器的跟踪性能. 按照扰动噪声的生成方式, 这些攻击方法主要分为基于深度网络的攻击方法和基于迭代优化的攻击方法. 基于深度网络的攻击方法通过离线训练一个基于深度网络的扰动生成器, 在线向原始视频序列中添加扰动噪声. 例如, Yan 等人[4] 提出的冷却收缩攻击(cooling-shrinking attack, CSA)以及 Nakka 等人[5] 提出的时间转移扰动(temporally-transferable perturbations, TTP)攻击, 利用跟踪器的输出建立了有关置信分数和边框偏移的损失函数, 然后离线训练基于深度网络的生成器. 在攻击过程中, 生成器以视频帧为输入生成扰动噪声, 并将其叠加到对应的视频帧作为跟踪器的输入, 以此降低其跟踪性能. 基于迭代优化的攻击方法则直接以降低跟踪性能为目标, 通过迭代优化的方式改变原始视频帧上的像素值. 例如, Guo 等人[15] 提出的空间感知的在线增量攻击(spatial-aware online incremental attack, Spark), 基于符号梯度下降迭代法提出了空间感知在线增量攻击方法, 通过向视频序列添加增量扰动, 有效地干扰了跟踪器的跟踪过程.

尽管这些攻击方法都是向视频序列中添加扰动噪声, 但不同的是, 有些方法向视频序列的每一帧添加扰动噪声, 如 CSA[4], TTP[5], Spark[15]; 有些方法仅向视频序列的初始帧添加扰动噪声, 如 One-shot [16], OOA[17]; 有些方法则是向视频序列中添加小尺寸扰动补丁, 如 UTA[18], APA[19]. 本文以视频序列为检测单元, 主要检测那些向整个视频序列添加扰动噪声的攻击方法.

1.2 离散余弦变换

离散余弦变换(discrete cosine transform, DCT)常用于对信号进行有损压缩. 利用 2 维 DCT 变换提取视频序列的低频段分量. 2 维 DCT 变换及其逆变换如式(1)和式(2)所示.

$$\boldsymbol{X}(u,v)=\sigma_u\sigma_v\sum_{m=0}^{M-1}\sum_{n=0}^{N-1}\boldsymbol{x}(m,n)\times \cos\left[\frac{(m+0.5)\pi}{M}u\right]\cos\left[\frac{(n+0.5)\pi}{N}v\right], \tag{1}$$

$$\boldsymbol{x}(m,n)=\sum_{u=0}^{M-1}\sum_{v=0}^{N-1}\sigma_u\sigma_v\boldsymbol{X}(u,v)\times \cos\left[\frac{(m+0.5)\pi}{M}u\right]\cos\left[\frac{(n+0.5)\pi}{N}v\right], \tag{2}$$

$$\sigma_u=\begin{cases}1/\sqrt{M}, & u=0;\\ \sqrt{2/M}, & u\neq 0.\end{cases} \qquad \sigma_v=\begin{cases}1/\sqrt{N}, & v=0;\\ \sqrt{2/N}, & v\neq 0.\end{cases} \tag{3}$$

经过 DCT 变换后, 空域图像 $\boldsymbol{x}\in\mathbb{R}^{M\times N}$ 与其频域系数矩阵 $\boldsymbol{X}\in\mathbb{R}^{M\times N}$ 具有相同尺寸. 为了便于下文表述, 将频域系数矩阵 $\boldsymbol{X}$ 的频段范围 $[0, M]$ 和 $[0, N]$ 归一化为 $[0, 1]$. 例如, M=N=250, 归一化后, 0.1 频段步长代表的实际频段步长为 25; 范围 $[0, 1]$ 表示全频段分量, 即完整图像.

2 目标跟踪的对抗检测方法

2.1 跟踪过程和攻击过程简要描述

给定视频序列 $\mathcal{V}=\{\boldsymbol{I}_t\}_{t=1}^{T}$, 图 2 以 SiamRPNpp 跟踪器[25] 为例展示了跟踪过程和攻击过程. 在跟踪过程中, 首先从初始帧 $\boldsymbol{I}_1$ 中裁剪出模板图像 $\boldsymbol{z}$, 用于初

始化跟踪器$\mathcal{T}$，然后从视频帧$\boldsymbol{I}_t$中裁剪出搜索图像$\boldsymbol{x}_t, t=2,3,\cdots,T$，跟踪器$\mathcal{T}$实际是在搜索图像$\boldsymbol{x}_t$上的跟踪目标. 对于第$t$帧，模板图像$\boldsymbol{z}$和搜索图像$\boldsymbol{x}_t$首先通过主干网络提取通用特征进行相似度匹配，并得到相似度图$\boldsymbol{E}$，然后将$\boldsymbol{E}$送入区域候选网络（region proposal network, RPN）[26]用于预测分类图$\{s_t^i \in \boldsymbol{S}\}_{i=1}^N$和回归图$\{\boldsymbol{b}_t^i \in \boldsymbol{B}\}_{i=1}^N$，最后取最大置信分数$s_t^{\mathrm{m}}$并将其对应的预测边框$\boldsymbol{b}_t^{\mathrm{m}}$作为第$t$帧的跟踪结果，即$(s_t^{\mathrm{m}}, \boldsymbol{b}_t^{\mathrm{m}}) = \mathcal{T}(\boldsymbol{z}, \boldsymbol{x}_t)$，其中$m = \arg\max\{s_t^i\}_{i=1}^N$. 当预测边框$\boldsymbol{b}_t^m$与标签边框$\boldsymbol{b}_t^{\mathrm{gt}}$的交并比（intersection of union, *IoU*）超过阈值θ，即$\Omega(\boldsymbol{b}_t^m, \boldsymbol{b}_t^{\mathrm{gt}}) > \theta$，则表示跟踪器在第$t$帧上成功地跟踪到目标，其中$\Omega(\cdot)$表示计算2个边框间的交并比. 对于整个数据集，通常使用不同阈值下成功率的曲线面积（area under curve, *AUC*）来衡量跟踪器在整个数据集上的跟踪性能，其值越大说明跟踪性能越好.

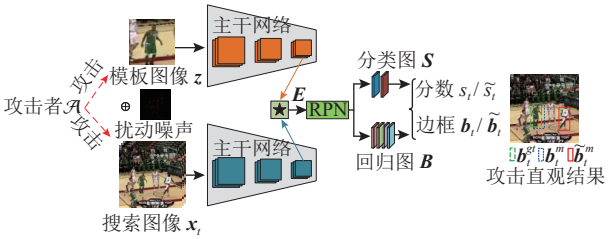


Fig. 2 Tracking process and attack process

图2 跟踪过程和攻击过程

攻击过程伴随着跟踪过程. 在攻击过程中，攻击者$\mathcal{A}$通常首先向原始模板图像$\boldsymbol{z}$中添加扰动噪声，合成对抗模板图像$\tilde{\boldsymbol{z}}$，用于干扰跟踪器的初始化，然后再向搜索图像$\boldsymbol{x}_t$中添加扰动噪声，合成对抗搜索图像$\tilde{\boldsymbol{x}}_t$，用于干扰其正常跟踪. 跟踪器以对抗模板图像$\tilde{\boldsymbol{z}}$和对抗搜索图像$\tilde{\boldsymbol{x}}_t$作为输入，输出较差的跟踪结果，即$(\tilde{s}_t^m, \tilde{\boldsymbol{b}}_t^m) = \mathcal{T}(\tilde{\boldsymbol{z}}, \tilde{\boldsymbol{x}}_t)$，$m = \arg\max\{\tilde{s}_t^i\}_{i=1}^N$. 预测边框$\tilde{\boldsymbol{b}}_t^m$与标签边框$\boldsymbol{b}_t^{\mathrm{gt}}$之间的交并比很小，表示跟踪失败. 算法1以伪代码的形式描述了跟踪过程和攻击过程.

算法1. 跟踪过程与攻击过程算法描述.

输入：视频序列$\mathcal{V} = \{\boldsymbol{I}_t\}_{t=1}^T$，跟踪器$\mathcal{T}$，攻击者$\mathcal{A}$；

输出：跟踪结果.

/*跟踪过程*/

① 从初始帧$\boldsymbol{I}_1$中裁剪出模板图像$\boldsymbol{z}$；

② 用$\boldsymbol{z}$初始化跟踪器$\mathcal{T}$；

③ for $t=2,3,\cdots,T$ do

④ 从视频帧$\boldsymbol{I}_t$中裁剪出搜索图像$\boldsymbol{x}_t$；

⑤ $\mathcal{T}$以模板图像$\boldsymbol{z}$和搜索图像$\boldsymbol{x}_t$为输入；

⑥ 执行跟踪任务，即$(s_t^m, \boldsymbol{b}_t^m) = \mathcal{T}(\boldsymbol{z}, \boldsymbol{x}_t)$；

⑦ 输出每一帧的跟踪结果$\boldsymbol{b}_t^m, t=2,3,\cdots,T$；

⑧ end for

/*攻击过程*/

① 向$\boldsymbol{z}$中添加扰动，得到对抗模板图像$\tilde{\boldsymbol{z}}$；

② 用$\tilde{\boldsymbol{z}}$初始化跟踪器$\mathcal{T}$；

③ for $t=2,3,\cdots,T$ do

④ 向$\boldsymbol{x}_t$中添加扰动，得到对抗搜索图像$\tilde{\boldsymbol{x}}_t$；

⑤ $\mathcal{T}$以$\tilde{\boldsymbol{z}}$和$\tilde{\boldsymbol{x}}_t$为输入；

⑥ 执行跟踪任务，即$(\tilde{s}_t^m, \tilde{\boldsymbol{b}}_t^m) = \mathcal{T}(\tilde{\boldsymbol{z}}, \tilde{\boldsymbol{x}}_t)$；

⑦ 输出每一帧的跟踪结果$\tilde{\boldsymbol{b}}_t^m, t=2,3,\cdots,T$；

⑧ end for

2.2 扰动存在于中高频段的理论证明

对抗攻击方法通常在扰动生成损失函数中增加约束项限定扰动噪声不超过阈值ε，如$\|\boldsymbol{x}-\tilde{\boldsymbol{x}}\|_2 \leqslant \varepsilon$[4-5]和$\|\boldsymbol{x}-\tilde{\boldsymbol{x}}\|_\infty \leqslant \varepsilon$[15]，以此确保扰动噪声视觉不可见. 本节推断扰动噪声主要被添加在图像的中高频段并借助DFT证明扰动存在于中高频段.

定理1. 假设原始图像$\boldsymbol{x} \in \mathbb{R}^{M\times N}$和对抗图像$\tilde{\boldsymbol{x}} \in \mathbb{R}^{M\times N}$，可以得出以下推论：

$$\frac{1}{D}|\boldsymbol{X}_0 - \tilde{\boldsymbol{X}}_0| \leqslant \|\boldsymbol{x}-\tilde{\boldsymbol{x}}\|_\infty \leqslant \varepsilon, \tag{4}$$

$$\frac{1}{\sqrt{D}}|\boldsymbol{X}_0 - \tilde{\boldsymbol{X}}_0| \leqslant \|\boldsymbol{x}-\tilde{\boldsymbol{x}}\|_2 \leqslant \varepsilon. \tag{5}$$

式(6)和式(7)表示傅里叶变换f和其逆变换f^{-1}：

$$\boldsymbol{X}_k = f(\boldsymbol{x})_k = \sum_{j=0}^{D-1} \boldsymbol{x}_j \mathrm{e}^{-\mathrm{i}\frac{2\pi}{D}kj}, \tag{6}$$

$$\boldsymbol{x}_j = f^{-1}(\boldsymbol{X})_j = \frac{1}{D}\sum_{k=0}^{D-1} \boldsymbol{X}_k \mathrm{e}^{\mathrm{i}\frac{2\pi}{D}kj}, \tag{7}$$

其中$\boldsymbol{X}_0$和$\tilde{\boldsymbol{X}}_0$分别是$k=0$时的傅里叶系数. i是复数因子，$\mathrm{i}^2=-1$，$D=\mathrm{M}\times N$. 在证明式(4)之前，先引入式(8)：

$$\sum_{j=0}^{D-1}\mathrm{e}^{\mathrm{i}j\omega} = 1+\mathrm{e}^{\mathrm{i}\omega}+\mathrm{e}^{\mathrm{i}2\omega}+\cdots+\mathrm{e}^{\mathrm{i}(D-1)\omega} = \frac{1-\mathrm{e}^{\mathrm{i}D\omega}}{1-\mathrm{e}^{\mathrm{i}\omega}} = \frac{\mathrm{e}^{\mathrm{i}D\omega/2}(\mathrm{e}^{-\mathrm{i}D\omega/2}-\mathrm{e}^{\mathrm{i}D\omega/2})}{\mathrm{e}^{\mathrm{i}\omega/2}(\mathrm{e}^{-\mathrm{i}\omega/2}-\mathrm{e}^{\mathrm{i}\omega/2})} = \frac{\sin(D\omega/2)}{\sin(\omega/2)}\mathrm{e}^{\mathrm{i}\omega(D-1)/2}. \tag{8}$$

令$\omega = \dfrac{2\pi}{D}k$，则式(8)可以表示为

$$\sum_{j=0}^{D-1}\mathrm{e}^{\mathrm{i}(2k\pi/D)j} = \frac{\sin(k\pi)}{\sin(k\pi/D)}\mathrm{e}^{\mathrm{i}k\pi(D-1)/D}. \tag{9}$$

当$k=D$或$k=0$时，$\sum_{j=0}^{D-1}\mathrm{e}^{\mathrm{i}(2k\pi/D)j} = D$；当$k\neq D$且$k\neq 0$时，$\sum_{j=0}^{D-1}\mathrm{e}^{\mathrm{i}(2k\pi/D)j} = 0$.

式(4)的具体推导过程如下：

$$\|\boldsymbol{x}-\tilde{\boldsymbol{x}}\|_\infty \geqslant \frac{1}{D}\|\boldsymbol{x}-\tilde{\boldsymbol{x}}\|_1 = \frac{1}{D}\sum_{j=0}^{D-1}|\boldsymbol{x}_j-\tilde{\boldsymbol{x}}_j| =$$
$$\frac{1}{D}\sum_{j=0}^{D-1}|f^{-1}(\boldsymbol{X})_j - f^{-1}(\tilde{\boldsymbol{X}})_j| =$$
$$\frac{1}{D}\sum_{j=0}^{D-1}\left|\frac{1}{D}\sum_{k=0}^{D-1}\boldsymbol{X}_k e^{i\frac{2\pi}{D}kj} - \frac{1}{D}\sum_{k=0}^{D-1}\tilde{\boldsymbol{X}}_k e^{i\frac{2\pi}{D}kj}\right| =$$
$$\frac{1}{D^2}\sum_{j=0}^{D-1}\left|\sum_{k=0}^{D-1}(\boldsymbol{X}_k-\tilde{\boldsymbol{X}}_k)e^{i\frac{2\pi}{D}kj}\right| \geqslant$$
$$\frac{1}{D^2}\left|\sum_{j=0}^{D-1}\sum_{k=0}^{D-1}(\boldsymbol{X}_k-\tilde{\boldsymbol{X}}_k)e^{i\frac{2\pi}{D}kj}\right| =$$
$$\frac{1}{D^2}\left|\sum_{k=0}^{D-1}(\boldsymbol{X}_k-\tilde{\boldsymbol{X}}_k)\sum_{j=0}^{D-1}e^{i\frac{2\pi}{D}kj}\right| =$$
$$\frac{1}{D^2}\times D|\boldsymbol{X}_0-\tilde{\boldsymbol{X}}_0| =$$
$$\frac{1}{D}|\boldsymbol{X}_0-\tilde{\boldsymbol{X}}_0|. \tag{10}$$

因为 L2 范式为 $\|\boldsymbol{x}-\tilde{\boldsymbol{x}}\|_2 = \sqrt{\sum_{j=0}^{D-1}|\boldsymbol{x}_j-\tilde{\boldsymbol{x}}_j|^2}$，其与 L1 范式的关系可表示为

$$\|\boldsymbol{x}-\tilde{\boldsymbol{x}}\|_1 \leqslant \sqrt{D}\|\boldsymbol{x}-\tilde{\boldsymbol{x}}\|_2. \tag{11}$$

由式(10)可得 $\|\boldsymbol{x}-\tilde{\boldsymbol{x}}\|_1 \geqslant |\boldsymbol{X}_0-\tilde{\boldsymbol{X}}_0|$，因此可以得到

$$|\boldsymbol{X}_0-\tilde{\boldsymbol{X}}_0| \leqslant \sqrt{D}\|\boldsymbol{x}-\tilde{\boldsymbol{x}}\|_2. \tag{12}$$

因此完成式(5)的证明.

由式(4)和式(5)可以发现，扰动噪声变化的下限边界 ε 只与频域中 $k=0$ 的傅里叶系数有关，即与低频段相关. 更确切地说，下限边界 ε 与 $|\boldsymbol{X}_0-\tilde{\boldsymbol{X}}_0|$ 呈正相关. 如果扰动噪声被添加在图像的低频段，则 $|\boldsymbol{X}_0-\tilde{\boldsymbol{X}}_0|$ 和 ε 会变大，扰动损失也会变大. 这违背了攻击模型收敛和扰动视觉不可见的原则. 因此，为了使扰动生成损失和扰动噪声尽可能小，对抗攻击方法将扰动噪声主要添加在图像的中高频段内.

2.3 频段分量贡献的定量研究

本节通过经验性实验定量地分析视频序列的各频段分量对跟踪性能的贡献和受对抗攻击的影响.

1)攻击前各频段分量的跟踪性能对比. 首先以 0.1 为频段步长将 OTB2015 数据集[27]和 UAV123 数据集[28]中的原始视频序列分解为多个频段分量，然后 SiamRPNpp 在原始视频序列的各个频段分量上执行跟踪任务. 图 3 展示了攻击前 SiamRPNpp 在 OTB2015 和 UAV123 数据集各频段分量上的跟踪结果，“0~1”表示全频段分量. 图 3 中的 *AUC* 表示攻击前跟踪器在各频段分量上的跟踪性能；*D_AUC* 表示跟踪器在各频段分量上的 *AUC* 与在全频段分量上的 *AUC* 之间的差值. 受攻击前，跟踪器在全频段分量上的 *AUC* 值较大，因此，任意频段分量上的 *D_AUC* 值越小，表明跟踪器在该频段分量上的跟踪性能与在全频段分量上的跟踪性能间的差异越小，以及频段分量对跟踪器的跟踪性能贡献越大. 可以看出，受攻击前，“0~0.1”频段分量对跟踪器的跟踪性能贡献最大.

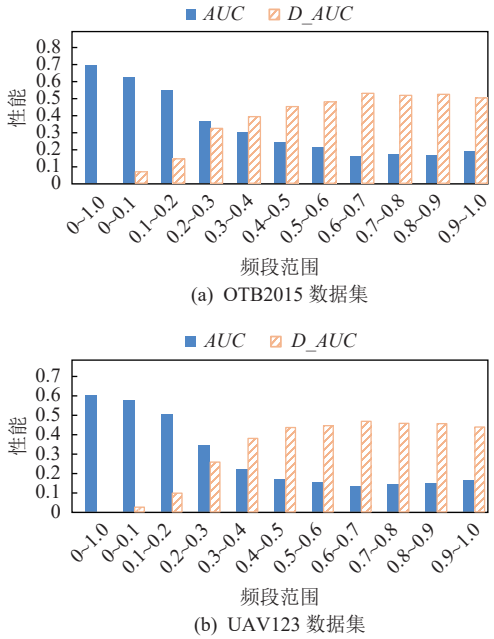
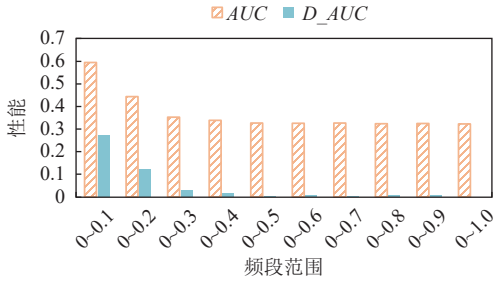


Fig. 3 Tracking performance comparison of each frequency band before attacks

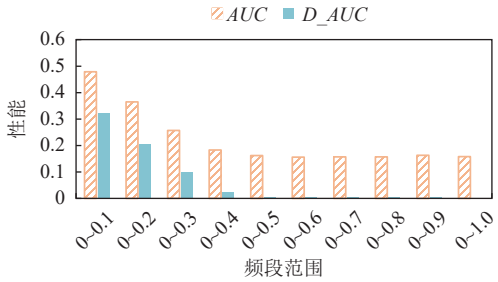
图 3 攻击前各频段的跟踪性能对比

2)攻击后各频段分量的跟踪性能对比. 首先利用 CSA[4]和 TTP[5]向 OTB2015 和 UAV123 数据集的所有视频序列中添加扰动噪声，生成对抗视频序列；然后以 0.1 为频段步长分解这些对抗视频序列；最后 SiamRPNpp 在对抗视频序列的各个频段分量上执行跟踪任务. 图 4 和图 5 分别表示受 CSA 和 TTP 攻击后，SiamRPNpp 在 OTB2015 和 UAV123 数据集各频段分量上的跟踪结果，受攻击后，跟踪器在全频段分量上的 *AUC* 值较小. 因此，任意频段分量上的 *D_AUC* 值越大，表明跟踪器在该频段分量上的跟踪性能与在全频段分量上的跟踪性能之间的差异越大，该频段分量受对抗攻击的影响越小. 可以看出，受攻击后，“0~0.1”频段分量受对抗攻击的影响最小.

上述实验表明视频序列被攻击前，跟踪器在原始视频序列的低频段和全频段分量上的跟踪性能差异较小；视频序列被攻击后，因全频段包含带有扰动噪声的中高频段，所以跟踪器在对抗视频序列的低频段和全频段分量上的跟踪性能差异较大，这也是



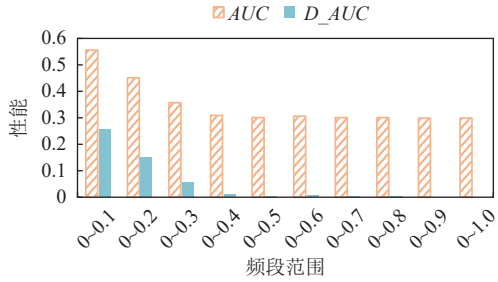
(a) CSA 攻击 OTB2015 数据集



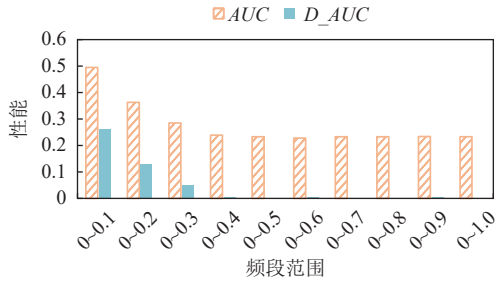
(b) TTP 攻击 OTB2015 数据集

Fig. 4 Tracking performance comparison of each frequency band in OTB2015 after attacks

图 4 攻击后 OTB2015 各频段的跟踪性能对比



(a) CSA 攻击 UAV123 数据集



(b) TTP 攻击 UAV123 数据集

Fig. 5 Tracking performance comparison of each frequency band in UAV123 after attacks

图 5 攻击后 UAV123 各频段的跟踪性能对比

本文方法的检测机制.

2.4 基于频段跟踪性能差异的对抗检测框架

2.2 节的理论证明和 2.3 节的定量分析阐述了 2 个观点：1)扰动噪声主要存在于图像的中高频段；2)低频段分量对跟踪性能的贡献最大且受对抗攻击的影响最小. 受此启发，提出了基于频段跟踪性能差异的对抗检测框架，如图 6 所示. 通过将目标跟踪器替换为指定的跟踪模型，即可灵活地实现检测框架与该跟踪模型的集成，从而协助其检测对抗输入. 给定待检测的视频序列 $\mathcal{V}=\{\boldsymbol{I}_t\}_{t=1}^{T}$. 首先，对视频序列 $\mathcal{V}$ 的初始帧 $\boldsymbol{I}_1$ 进行裁剪，得到全频模板图像 $\boldsymbol{z}$；使用频域分解模块，对 $\boldsymbol{z}$ 进行频域分解，得到低频模板图像 $\hat{\boldsymbol{z}}$，即 $\hat{\boldsymbol{z}}=\psi(\boldsymbol{z})$，其中 $\psi(\cdot)$ 表示 DCT 操作；使用 $\boldsymbol{z}$ 和 $\hat{\boldsymbol{z}}$ 分别初始化目标跟踪器 $\mathcal{T}$ 和镜像跟踪器 $\mathcal{T}_{\text{os}}$. 然后，在后续跟

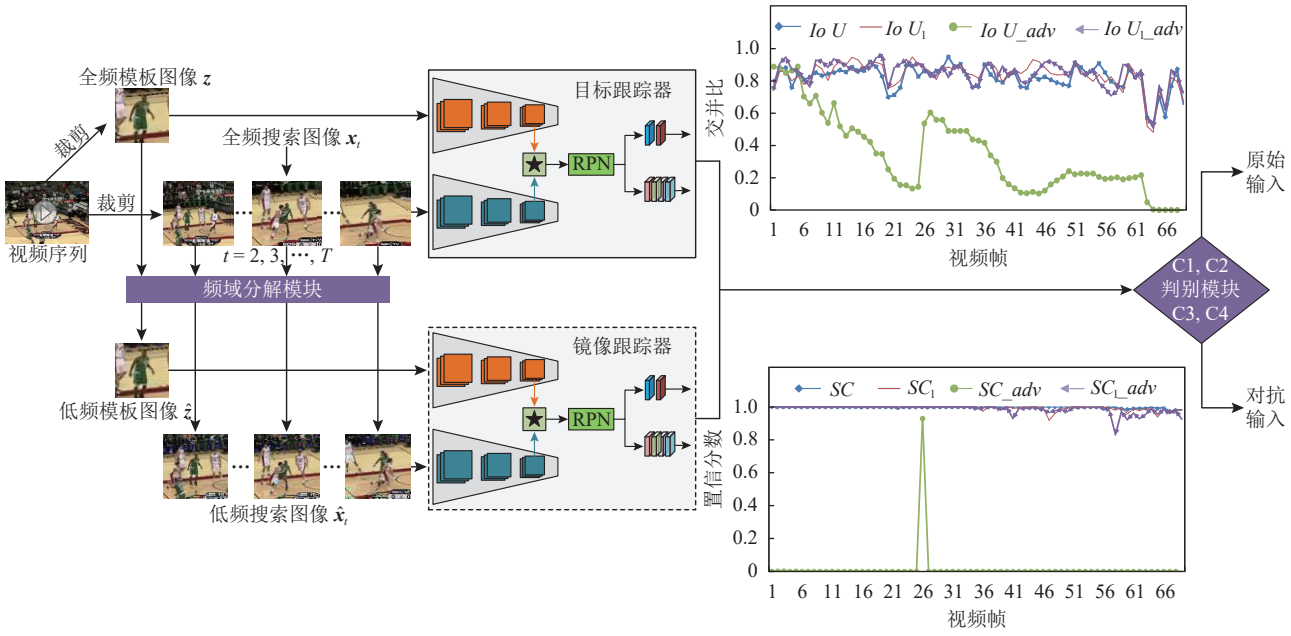


Fig. 6 Pipeline of detection framework

图 6 检测框架流程图

踪过程中, 对视频序列$\mathcal{V}$的视频帧$\boldsymbol{I}_t, t=2,3,\cdots,T$进行裁剪, 得到全频搜索图像$\boldsymbol{x}_t$; 使用频域分解模块, 对$\boldsymbol{x}_t$进行频域分解, 得到低频搜索图像$\hat{\boldsymbol{x}}_t$, 即$\hat{\boldsymbol{x}}_t=\psi(\boldsymbol{x}_t)$; 目标跟踪器$\mathcal{T}$和镜像跟踪器$\mathcal{T}_{\rm os}$分别在全频搜索图像$\boldsymbol{x}_t$和低频搜索图像$\hat{\boldsymbol{x}}_t$上执行跟踪任务, 得到$(s_t^m,\boldsymbol{b}_t^m)=\mathcal{T}(\boldsymbol{z},\boldsymbol{x}_t)$和$(\hat{s}_t^m,\hat{\boldsymbol{b}}_t^m)=\mathcal{T}_{\rm os}(\hat{\boldsymbol{z}},\hat{\boldsymbol{x}}_t)$. 最后, 计算所有全频段视频序列和低频段视频序列图像上的置信分数和交并比均值, 即$SC=Ave\{\hat{s}_t^m\}_{t=2}^T$和$IoU=Ave\{\Omega(\boldsymbol{b}_t^{\rm gt},\hat{\boldsymbol{b}}_t^m)\}_{t=2}^T$, $Ave(\cdot)$表示计算平均值, 判别模块则以(SC,IoU)和$(SC_{\rm l},IoU_{\rm l})$为输入, 判定当前视频序列$\mathcal{V}$属于对抗输入还是原始输入.

1)检测机理. 根据2.3节的定量研究, 低频段分量的跟踪性能最能代表全频段分量. 所以当输入为原始视频序列, 理论上目标跟踪器和镜像跟踪器的跟踪性能差异较小. 根据2.2节的理论证明, 扰动噪声主要被添加在中高频段, 低频段信息受扰动的影响最小. 所以当输入为对抗视频序列, 以全频段分量为输入的目标跟踪器受影响程度远大于以低频段分量为输入的镜像跟踪器, 理论上目标跟踪器和镜像跟踪器的跟踪性能差异较大.

2)频域分解模块. 该模块由2维DCT及其逆变换组成, 以全频图像为输入, 通过频域分解和重构, 提取低频图像.

3)目标跟踪器和镜像跟踪器. 这2个跟踪器共享结构和参数. 目标跟踪器以全频段视频序列为输入, 即全频模板图像和全频搜索图像; 镜像跟踪器以低频段视频序列为输入, 即低频模板图像和低频搜索图像.

4)判别模块. 该模块主要包含4个判别条件. 通过对比目标跟踪器和镜像跟踪器分别在全频段视频序列和低频段视频序列上的跟踪性能, 即(SC,IoU)和$(SC_{\rm l},IoU_{\rm l})$, 判定当前输入视频序列属于对抗输入还是原始输入.

在跟踪任务中, 置信分数和交并比直接反映了跟踪器目标定位和边框预测的准确性. 因此, 将整个视频序列的置信分数和交并比均值作为判别指标. 一方面, 未受攻击时, 全频段和低频段视频序列的跟踪性能差距较小; 受到攻击后, 全频段和低频段视频序列的跟踪性能差距较大. 因此, 设置2个机理判别条件, 即$\rm C_l$: $|IoU-IoU_{\rm l}|>a_1$, $\rm C_2$: $|SC-SC_{\rm l}|>a_2$. 另一方面, 考虑到跟踪器在一些原始视频序列上本就表现不佳, 以及跟踪器并非在所有对抗视频序列上都表现较差等情况, 设置2个区间判别条件, 即$\rm C_3$: $IoU>a_3$, $\rm C_4$: $IoU<a_4$.

规定如果$\rm C_3$成立, 则判定当前输入视频序列为原始输入; 如果$\rm C_4$成立, 则判定当前输入视频序列为对抗输入; 如果$\rm C_1$或$\rm C_2$成立, 则判定当前输入视频序列为对抗输入, 否则判定为原始输入.

算法2以伪代码的形式展示了对抗检测过程.

算法2. 基于频段跟踪性能差异的检测流程.

输入: 视频序列$\mathcal{V}=\{\boldsymbol{I}_t\}_{t=1}^T$, 目标跟踪器$\mathcal{T}$及其镜像跟踪器$\mathcal{T}_{\rm os}$;

输出: 判别结果.

① 从初始帧$\boldsymbol{I}_1$裁剪出全频模板图像$\boldsymbol{z}$;
② 从$\boldsymbol{z}$中分解出低频模板图像$\hat{\boldsymbol{z}}=\psi(\boldsymbol{z})$;
③ 用$\boldsymbol{z}$和$\hat{\boldsymbol{z}}$分别初始化$\mathcal{T}$和$\mathcal{T}_{\rm os}$;
④ for $t=2,3,\cdots,T$ do
⑤ 从视频帧$\boldsymbol{I}_t$裁剪出全频搜索图像$\boldsymbol{x}_t$;
⑥ 从$\boldsymbol{x}_t$分解出低频搜索图像$\hat{\boldsymbol{x}}_t=\psi(\boldsymbol{x}_t)$;
⑦ $\mathcal{T}$执行跟踪任务$(s_t^m,\boldsymbol{b}_t^m)=\mathcal{T}(\boldsymbol{z},\boldsymbol{x}_t)$;
⑧ $\mathcal{T}_{\rm os}$执行跟踪任务$(\hat{s}_t^m,\hat{\boldsymbol{b}}_t^m)=\mathcal{T}_{\rm os}(\hat{\boldsymbol{z}},\hat{\boldsymbol{x}}_t)$;
⑨ end for
⑩ 计算全频视频帧的置信分数和交并比均值, $SC=Ave\{s_t^m\}_{t=2}^T$, $IoU=Ave\{\Omega(\boldsymbol{b}_t^{\rm gt},\boldsymbol{b}_t^m)\}_{t=2}^T$;
⑪ 计算低频视频帧的置信分数和交并比均值, $SC_{\rm l}=Ave\{\hat{s}_t^m\}_{t=2}^T$, $IoU_{\rm l}=Ave\{\Omega(\boldsymbol{b}_t^{\rm gt},\hat{\boldsymbol{b}}_t^m)\}_{t=2}^T$;
⑫ 判定视频序列是否为对抗输入;
⑬ if $IoU>a_3$
⑭ 判定视频序列为原始输入;
⑮ else if $IoU<a_4$
⑯ 判定视频序列为对抗输入;
⑰ else if $|IoU-IoU_{\rm l}|>a_1$ or $|SC-SC_{\rm l}|>a_2$
⑱ 判定视频序列为对抗输入;
⑲ else
⑳ 判定视频序列为原始输入;
㉑ end if

3 实验及分析

所有实验均在PyTorch软件平台和NVIDIA RTX3090 GPU硬件设备上仿真. 对于频域分解模块, 所提取的低频分量范围设置为[0, 0.1]. 对于判别模块, 判别阈值设置为$a_1=0.1$, $a_2=0.1$, $a_3=0.5$, $a_4=0.1$. 一方面, CSA[4], TTP[5], Spark[15]是当前主流对抗攻击方法的代表; 另一方面, 它们的攻击策略能够有效地降低SiamRPNpp跟踪器的跟踪性能. 因此, 本文将CSA, TTP和Spark视为主要的检测对象.

3.1 数据集和检测性能评价指标

目前缺少专门用于检测对抗攻击的数据集, 因

此本文效仿识别领域的对抗检测做法. 在原始跟踪数据集的基础上构造用于验证对抗检测方法的混合数据集. 所使用的原始跟踪数据集包括包含 100 个视频序列的 OTB2015 数据集[27]; 包含 123 个视频序列的 UAV123 数据集[28]; 包含 280 个视频序列的 LaSOT 数据集[29]. 为了构造混合数据集, 利用对抗攻击方法 CSA[4], TTP[5], Spark[15] 分别向原始跟踪数据集的视频序列中添加扰动噪声, 生成相应的对抗视频序列. 另外, 考虑到对抗检测方法需要同时具备判别原始输入和对抗输入的能力, 因此, 混合数据集不仅包含原始视频序列还包含对抗视频序列. 例如, 混合数据集 OTB2015*包含 100 个原始视频序列和 100 个对抗视频序列; 混合数据集 UAV123*包含 123 个原始视频序列和 123 个对抗视频序列; 混合数据集 LaSOT*包含 280 个原始视频序列和 280 个对抗视频序列.

采用检测精度(detection precision, P)、召回率(recall, R)和 $F1$ 分数作为检测性能的评价指标, $P=TP/(TP+FP)$, $R=TP/(TP+FN)$, $F1=2P\times R/(P+R)$. TP 表示对抗视频序列被判定为对抗输入的数量; TN 表示原始视频序列被判定为原始输入的数量; FP 表示原始视频序列被判定为对抗输入的数量; FN 表示对抗视频序列被判定为原始输入的数量. 在跟踪任务中, 通常使用 AUC、精确度(precision, Pre)和归一化精确度(norm precision, $Npre$)来评估跟踪器的跟踪性能. 为了衡量检测框架对跟踪器的影响, 基于上述 3 个跟踪性能评估指标, 又提出了 ΔAUC, ΔPre 和 $\Delta Npre$ 用于评估检测框架对跟踪器原始性能的影响. 例如, ΔAUC 表示原始跟踪器在原始跟踪数据集上的 AUC 指标与集成了检测框架的跟踪器在混合数据集上的 AUC 指标之间的差值. ΔAUC 越小说明检测框架不仅能够有效地检测对抗攻击, 而且对跟踪器的原始性能影响越小. 表 1 展示了所使用的数据集以及相关的性能评估指标.

Table 1 Datasets and Evaluation Metrics

表 1 数据集及评估指标

原始数据集	视频序列	跟踪性能评估指标
OTB2015	100	AUC, Pre, $Npre$
UAV123	123	
LaSOT	280	
混合数据集	**视频序列**	**检测性能评估指标**
OTB2015*	200	P, R, $F1$, ΔAUC, ΔPre, $\Delta Npre$
UAV123*	246	
LaSOT*	560	

3.2 检测 CSA 对抗攻击

为了验证检测方法的有效性, 本节将检测框架(detection framework, DF)集成到 SiamRPNpp 跟踪器(缩写为 SiamRPNpp+DF), 并在由 CSA 构造的混合数据集 OTB2015*, UAV123*, LaSOT*上执行跟踪任务, 以此实现对 CSA 攻击的检测, 检测结果如表 2 所示.

Table 2 Detection Performance Under CSA Attacks

表 2 CSA 攻击下的检测性能

混合数据集	P/%	R/%	$F1$/%	ΔAUC	ΔPre	$\Delta Npre$
OTB2015*	97.55	98.68	98.11	0.028	0.018	0.025
UAV123*	95.24	96.99	96.10	0.018	0.006	0.008
LaSOT*	94.91	99.64	97.21	0.009	0.016	0.004

由表 2 可知, 检测框架能够有效地检测出输入视频序列是否受到 CSA 对抗攻击. 在混合数据集 OTB2015*, UAV123*, LaSOT*上的检测精度 P, 分别高达 97.55%, 95.24%, 94.91%. 高精度的检测性能得益于攻击前后的频段跟踪性能存在较大的差异. 此外, 较小的Δ检测框架集成到 SiamRPNpp 跟踪器对其的原始跟踪性能影响微乎其微. 例如, SiamRPNpp+DF 跟踪器在 OTB2015*数据集上的 AUC 性能与 SiamRPNpp 跟踪器在 OTB2015 数据集上的 AUC 性能差距仅有 0.028. 这主要得益于检测框架能够精准地判别对抗视频序列, 使其不参与到跟踪器的跟踪性能评估.

3.3 检测 TTP 对抗攻击

本节使用 SiamRPNpp+DF 跟踪器在由 TTP 构造的混合数据集 OTB2015*, UAV123*, LaSOT*上执行跟踪任务, 以此实现对 TTP 攻击的检测, 检测结果如表 3 所示.

Table 3 Detection Performance Under TTP Attacks

表 3 TTP 攻击下的检测性能

混合数据集	P/%	R/%	$F1$/%	ΔAUC	ΔPre	$\Delta Npre$
OTB2015*	97.73	93.48	95.56	0.021	0.030	0.025
UAV123*	95.12	91.61	93.32	0.025	0.015	0.036
LaSOT*	93.38	94.96	94.16	0.008	0.004	0.016

在面对 TTP 攻击时, 检测框架能够在混合数据集 OTB2015*, UAV123*, LaSOT*上分别取得 97.73%, 95.12%, 93.38% 的检测精度. 这说明检测框架对 TTP 对抗攻击也具备较好的检测性能. 这是因为 TTP 和 CSA 都属于基于深度网络的对抗攻击方法, 它们都是使用扰动生成器生成扰动噪声并将其添加到图像

的中高频段内. 所以本文基于频段跟踪性能差异的检测框架也能有效地检测出 TTP 攻击. 此外, 较小的Δ指标表明, 检测框架对 SiamRPNpp 的原始跟踪性能影响很小. 例如, SiamRPNpp+DF 跟踪器在 UAV123* 数据集上的 *Pre* 性能与 SiamRPNpp 跟踪器在 UAV123 数据集上的 *Pre* 性能差距仅为 0.015.

3.4 检测 Spark 对抗攻击

为了检测 Spark 攻击, 本节使用 SiamRPNpp+DF 跟踪器在由 Spark 构成的混合数据集 OTB2015*, UAV123*, LaSOT*上执行跟踪任务, 检测结果如表 4 所示.

Table 4 Detection Performance Under Spark Attacks

表 4 Spark 攻击下的检测性能

混合数据集	*P*/%	*R*/%	*F*1/%	Δ*AUC*	Δ*Pre*	Δ*Npre*
OTB2015*	96.37	98.57	97.46	0.064	0.019	0.029
UAV123*	92.48	95.68	94.05	0.055	0.014	0.017
LaSOT*	96.86	98.74	97.79	0.067	0.074	0.081

由表 4 可知, 检测框架在混合数据集 OTB2015*, UAV123*, LaSOT*数据集上的召回率, 分别高达 98.57%, 95.68%, 98.74%. 高召回率的检测性能, 一方面得益于视频序列受 Spark 攻击前后, 跟踪器在全频段和低频段的跟踪性能差异更为明显; 另一方面得益于本文设置的精细化判别条件. 然而, 精细化的判别准则也使得判别模块出现了过拟合现象, 即将较多的原始视频序列判定为对抗输入, 从而导致Δ指标较大. 例如, SiamRPNpp+DF 跟踪器在 LaSOT*数据集上的 *Npre* 性能与 SiamRPNpp 跟踪器在 LaSOT 数据集上的 *Npre* 性能差距为 0.081.

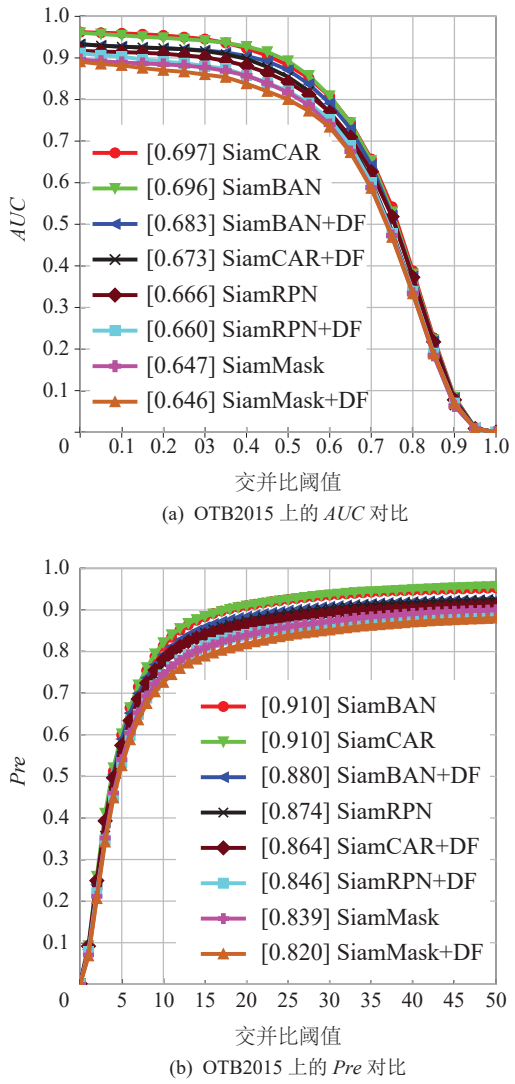
3.5 检测框架泛化性验证

为了验证检测框架的泛化性, 本节将检测框架直接集成到其他跟踪器, 如 SiamMask[30], SiamRPN[31], SiamCAR[32], SiamBAN[33], 用于检测 TTP 攻击. 整个过程无需优化检测框架. 表 5 展示了检测框架集成到不同跟踪器上时, 在混合数据集 OTB2015*上的检测性能. 可以看出, 提出的检测框架能够协助多个跟踪器实现有效的对抗检测. 例如, 检测框架集成到 SiamRPN 和 SiamBAN 上, 能够协助这 2 个跟踪器在 OTB2015* 数据集上分别取得 97.67% 的召回率和 100% 的检测精度. 提出的检测框架是以全频段和低频段跟踪性能的差异性作为判别基础, 所以不受限于跟踪器结构, 具有较好的检测泛化性. 此外, 检测框架对这些跟踪器的原始性能影响也较小, 如图 7 所示. 例如, SiamMask 跟踪器在 OTB2015 数据集上能够取得 0.647 的 *AUC*, 而 SiamMask+DF 跟踪器在 OTB2015*数据集上能够取得 0.646 的 *AUC*, 仅相差 0.001.

Table 5 Detection Performance of Detection Framework Integrated into Other Trackers

表 5 检测框架集成在其他跟踪器上的检测性能 %

评估指标	SiamRPN	SiamMask	SiamCAR	SiamBAN
P	93.33	92.68	100	100
R	97.67	90.48	87.50	93.62
*F*1	95.45	91.57	93.33	96.70



(a) OTB2015 上的 *AUC* 对比

(b) OTB2015 上的 *Pre* 对比

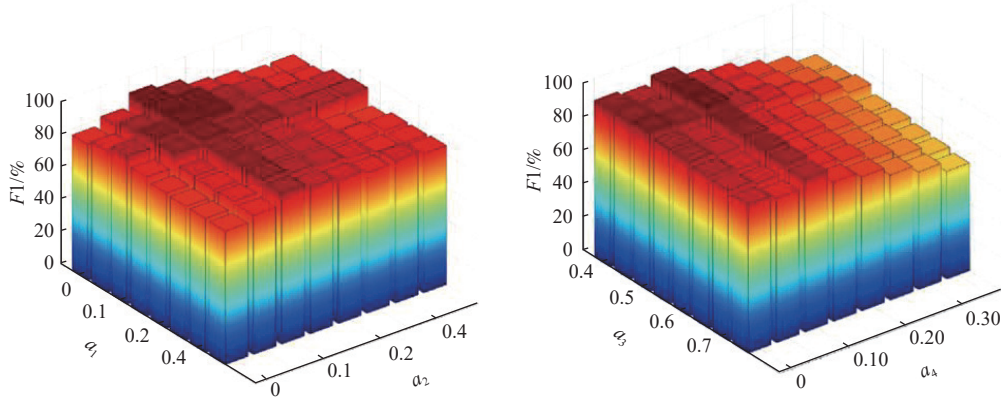
Fig. 7 Performance impact of detection framework on other trackers

图 7 检测框架对其他跟踪器的性能影响

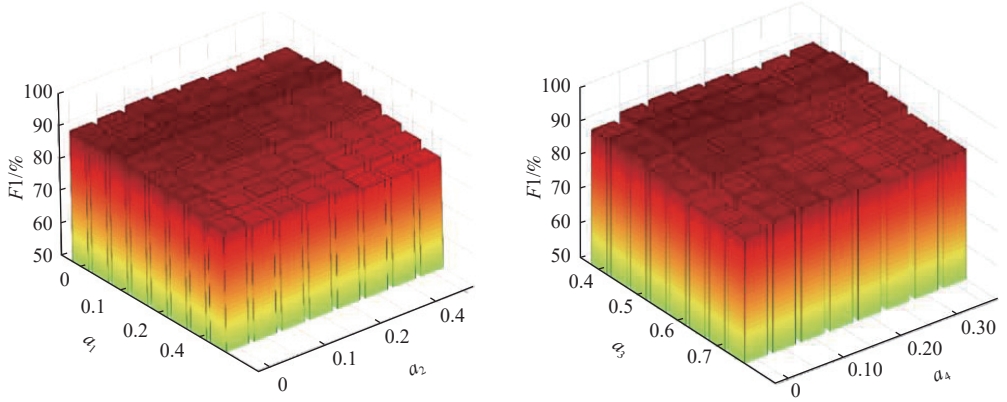
3.6 判别阈值选择实验

在判别模块中, 区间条件是为了减少因跟踪器性能不足和攻击效果有限所带来的判别误差; 机理条件则反映了攻击前后不同频段分量上跟踪性能的差

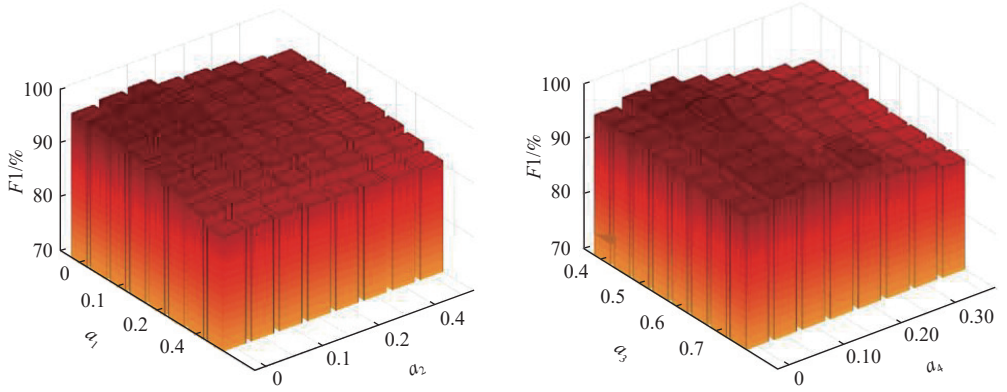
异上限. 为了验证所设判别阈值的合理性, 本节使用 SiamRPNpp+DF 跟踪器, 分别在由 CSA 构造的混合数据集 OTB2015*、由 TTP 构造的混合数据集 UAV123*, 以及由 Spark 构造的混合数据集 LaSOT*上执行跟踪任务, 通过改变阈值a_1, a_2, a_3, a_4的方式, 选定最优的阈值组合. 图 8 展示了阈值变化对检测性能 $F1$ 的影响. 可以看出, 阈值a_1和a_2过大会导致大量的对抗视频序列因无法满足判别条件 C_1 和 C_2, 而被错判为原始输入; 阈值过小则会弱化输入视频序列的全频段和低频段跟踪性能之间的差异性, 增加判定的不确定性. 阈值a_3决定了跟踪器成功跟踪的下限, 所以阈值a_3过小意味着跟踪器的跟踪性能短板被放大, 导致许多原始视频序列被错判为对抗输入. 阈值a_4决定了跟踪器跟踪失败的上限, 所以阈值a_4过大意味着攻击方法的有效性被放大, 使得更多对抗视频序列参与 C_1 和 C_2 条件判别.



(a) CSA 攻击 OTB2015 数据集



(b) TTP 攻击 UAV123 数据集



(c) Spark 攻击 LaSOT 数据集

Fig. 8 Impact of discrimination threshold on detection performance

图 8 判别阈值对检测性能的影响

3.7 消融实验

3.7.1 判别条件分析

为了分析各个判别条件对检测性能的影响，本节以 SiamRPNpp+DF 跟踪器在由 CSA 构造的混合数据集 UAV123*上执行跟踪任务为例，对比了不同判别条件下的检测结果. 消融实验结果如表 6 所示.

Table 6 Impact of Discrimination Conditions on Detection Performance

表 6 判别条件对检测性能的影响 %

评估指标	C_1	C_2	C_1+C_2	C_1+C_2+C_3	本文判别条件
P	92.50	94.08	92.66	93.73	95.24
R	47.13	90.24	91.88	94.79	96.99
*F*1	62.30	92.12	92.27	94.25	96.10

由表 6 可以看出，随着判别条件逐渐精细化，检测性能越来越好. 本文判别条件相比于条件 C_1，条件 C_2 下的检测性能更好，反映了 CSA 对跟踪器置信分数的攻击更加剧烈，故产生了更明显的置信分数差异. 相比于 C_1+C_2 条件，本文判别条件下的检测性能更佳，这是因为 C_3 和 C_4 是考虑到跟踪器自身性能不足和攻击方法有效性所设置的区间条件.

3.7.2 低频段范围对检测性能的影响

为了分析频段分量对检测性能的影响，本节以 SiamRPNpp+DF 跟踪器在由 CSA 构造的混合数据集 OTB2015*上执行跟踪任务为例，以 0.1 为频段步长逐步扩大镜像跟踪器输入视频序列的频段范围，对比不同频段范围下的检测性能，如图 9 所示. 可以看出，随着频段范围扩大，检测精度 *P* 的变化较小，而召回率 *R* 变化较大. 这是因为频段范围的扩大意味着镜像跟踪器的输入包含了更广范围的频段分量，其输入与目标跟踪器的输入差距在不断地缩小. 2 个跟踪器的输入差距变小，自然导致二者输出结果的差异性也变小，即弱化了跟踪性能差异性.

3.8 可视化结果展示

本节以 SiamRPNpp 跟踪器为例，展示其受 CSA 攻击前后的跟踪结果可视化图及其频谱图. 如图 10 和图 11 所示. 其中第 1 行的虚线边框表示待跟踪的真实目标. 通过对比图 10 和图 11 的第 1 行可以看出，原始图像和对抗图像的视觉差异较小. 通过对比 2 图第 2 行的跟踪结果可视化图可以看出，受 CSA 攻击前，可视化图具有大面积的高热区域，说明跟踪器此刻能够以高置信度准确地定位到目标区域. 相反，受 CSA 攻击后，可视化图上的高热区域变小或消失，说明跟踪器此刻难以准确地定位目标位置，表明跟踪器的性能变差. 图 10 和图 11 的第 3 行分别表示原始图像和对抗图像的频谱图，其中频谱图的左上角区域表示低频段区域，其余区域可以视为中高频段区域. 通过对比可以看出，攻击前后频谱图的低频段区域(左上角区域)差异不大. 然而攻击后，对抗图像频谱图的中高频段区域(椭圆区域)出现了较多的“亮

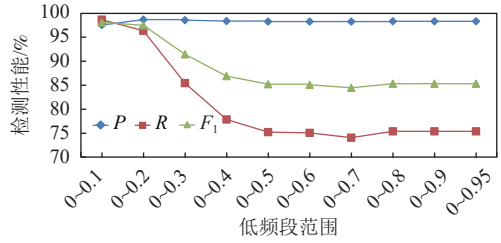


Fig. 9 Detection performance comparison under different frequency ranges

图 9 不同频段范围下检测性能对比

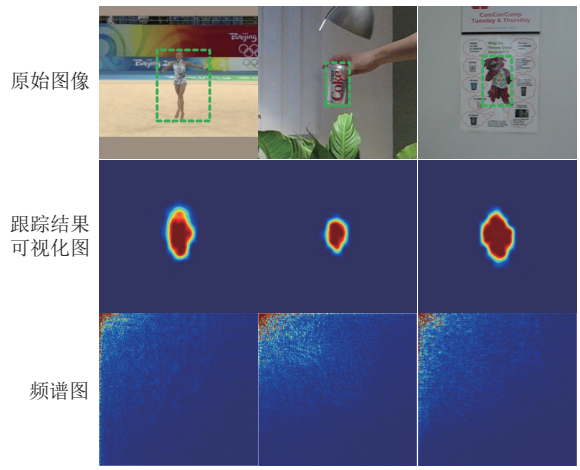


Fig. 10 Visualization map of tracking results of original images and their spectrograms before attacks

图 10 攻击前原始图像的跟踪结果可视化图及其频谱图

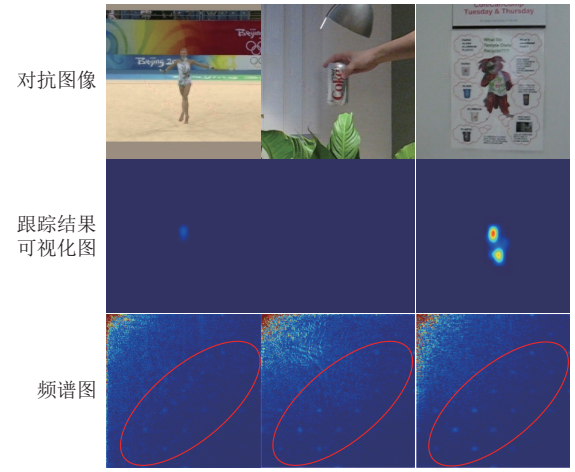


Fig. 11 Visualization map of tracking results of adversarial images and their spectrograms after attacks

图 11 攻击后对抗图像的跟踪结果可视化图及其频谱图

斑”，证明了添加的扰动噪声主要存在于图像的中高频段.

4 总结与展望

根据扰动噪声不可见但能有效攻击的特点，首先理论证明了扰动噪声主要存在于视频序列的中高频段；然后定量地分析出低频段分量对跟踪性能的贡献最大且受扰动噪声的影响最小；最后提出了适用于跟踪模型的对抗检测框架，该框架以跟踪器为载体，仅包含频域分解模块、镜像跟踪器以及判别模块. 通过对比视频序列不同频段分量的跟踪性能差异性，即可判定输入视频序列是否受到对抗攻击. 为了验证检测框架的有效性，将其集成到多个跟踪器，用于检测主流的对抗攻击. 大量的实验结果表明本文所提的检测框架具有泛化性，不仅能够有效地检测多种对抗攻击，而且对跟踪器的原始跟踪性能影响较小. 未来将提取更具判别性的特征，摆脱当前方法对标签信息的依赖，实现在线对抗检测；同时重点研究如何去除扰动噪声，提高跟踪模型抵御对抗攻击的鲁棒性.

作者贡献声明：周泽提出了方法思路和撰写论文；孙颖慧协助完成实验和数据整理；孙权森负责实验监督和指导；沈肖波和郑钰辉提出指导意见并修改论文.

参 考 文 献

[1] Kurakin A, Goodfellow I J, Bengio S. Adversarial examples in the physical world[J]. arXiv preprint, arXiv: 1607.02533, 2016

[2] Szegedy C, Zaremba W, Sutskever I, et al. Intriguing properties of neural networks[J]. arXiv preprint, arXiv: 1312.6199, 2014

[3] Zhang Wanli, Chen Yue, Yang Kuiwu, et al. An adversarial example generation method for locally occluded face recognition[J]. Journal of Computer Research and Development, 2023, 60(9): 2067−2079 (in Chinese)
(张万里, 陈越, 杨奎武, 等. 一种局部遮挡人脸识别的对抗样本生成方法[J], 计算机研究与发展, 2023, 60(9): 2067−2079)

[4] Yan Bin, Wang Dong, Lu Huchuan, et al. Cooling-shrinking attack: Blinding the tracker with imperceptible noises[C]//Proc of the 33rd IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2020: 990−999

[5] Nakka K K, Salzmann M. Universal, transferable adversarial perturbations for visual object trackers[C]//Proc of the 17th European Conf on Computer Vision Workshops. Berlin: Springer, 2022: 413−429

[6] Hou Ruitao, Ai Shan, Chen Qi, et al. Similarity-based integrity protection for deep learning systems[J]. Information Sciences, 2022, 601: 255−267

[7] Goel A, Moulin P. Fast locally optimal detection of targeted universal adversarial perturbations[J]. IEEE Transactions on Information Forensics and Security, 2022, 17: 1757−1770

[8] Lust J, Condurache A P. Efficient detection of adversarial, out-of-distribution and other misclassified samples[J]. Neurocomputing, 2022, 470: 335−343

[9] Moitra A, Panda P. DetectX-adversarial input detection using current signatures in memristive xbar arrays[J]. IEEE Transactions on Circuits and Systems I: Regular Papers, 2021, 68(11): 4482−4494

[10] Luo Wenjia, Wu Chenwang, Ni Li, et al. Detecting adversarial examples by positive and negative representations[J]. Applied Soft Computing, 2022, 117: 108383

[11] Jiang Wei, He Zhiyuan, Zhan Jinyu, et al. Attack-aware detection and defense to resist adversarial examples[J]. IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems, 2020, 40(10): 2194−2198

[12] Qin Chuan, Chen Yuefeng, Chen Kejiang, et al. Feature fusion based adversarial example detection against second-round adversarial attacks[J]. IEEE Transactions on Artificial Intelligence, 2023, 4(5): 1029−1040

[13] Qing Yuanyuan, Bai Tao, Liu Zhuotao, et al. Detection of adversarial attacks via disentangling natural images and perturbations[J]. IEEE Transactions on Information Forensics and Security, 2024, 19: 2814−2825

[14] Wiyatno R, Xu Anqi. Physical adversarial textures that fool visual object tracking[C]//Proc of the 16th IEEE/CVF Int Conf on Computer Vision. Piscataway, NJ: IEEE, 2019: 4821−4830

[15] Guo Qing, Xie Xiaofei, Ma Lei, et al. Spark: Spatial-aware online incremental attack against visual tracking[C]//Proc of the 15th European Conf on Computer Vision. Berlin: Springer, 2020: 202−219

[16] Chen Xuesong, Yan Xiyu, Zheng Feng, et al. One-shot adversarial attacks on visual tracking with dual attention[C]//Proc of the 33rd IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2020: 10176−10185

[17] Zhou Ze, Sun Yinghui, Sun Quansen, et al. Only once attack: Fooling the tracker with adversarial template[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2023, 33(7): 3171−3184

[18] Li Zhenbang, Shi Yaya, Gao Jin, et al. A simple and strong baseline for universal targeted attacks on Siamese visual tracking[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2021, 32(6): 3880−3894

[19] Zhou Ze, Sun Yinghui, Sun Quansen, et al. Attacking the tracker with a universal and attractive patch as fake target[J]. Information Sciences, 2023, 650: 119677

[20] Chen Xuesong, Fu Canmiao, Zheng Feng, et al. A unified multi-scenario attacking network for visual object tracking[C]//Proc of the 35th AAAI Conf on Artificial Intelligence. Palo Alto, CA: AAAI, 2021: 1097−1104

[21] Jia Shuai, Song Yibing, Ma Chao, et al. Iou attack: Towards

temporally coherent black-box adversarial attack for visual object tracking[C]//Proc of the 34th IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2021: 6709−6718

[22] Liang Siyuan, Wei Xingxing, Yao Siyuan, et al. Efficient adversarial attacks for visual object tracking[C]//Proc of the 16th European Conf on Computer Vision. Berlin: Springer, 2020: 34−50

[23] Guo Qing, Cheng Ziyi, Xu Juefei, et al. Learning to adversarially blur visual object tracking[C]//Proc of the 17th IEEE/CVF Int Conf on Computer Vision. Piscataway, NJ: IEEE, 2021: 10839−10848

[24] Zhao Shaochuan, Xu Tianyang, Wu Xiaojun, et al. Pluggable attack for visual object tracking[J]. IEEE Transactions on Information Forensics and Security, 2024, 19: 1227−1240

[25] Li Bo, Wu Wei, Wang Qiang, et al. SiamRPN++: Evolution of siamese visual tracking with very deep networks[C]//Proc of the 17th IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2019: 4282−4291

[26] 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

[27] Wu Yi, Lim J, Yang M. Object tracking benchmark[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2015, 37(9): 1834−1848

[28] Mueller M, Smith N, Ghanem B. A benchmark and simulator for uav tracking[C]//Proc of the 14th European Conf on Computer Vision. Berlin: Springer, 2016: 445−461

[29] Fan Heng, Lin Liting, Yang Fan, et al. Lasot: A high-quality benchmark for large-scale single object tracking[C]//Proc of the 32nd IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2019: 5374−5383

[30] Wang Qiang, Zhang Li, Bertinetto L, et al. Fast online object tracking and segmentation: A unifying approach[C]//Proc of the 32nd IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2019: 1328−1338

[31] Li Bo, Yan Junjie, Wu Wei, et al. High performance visual tracking with Siamese region proposal network[C]//Proc of the 31st IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2018: 8971−8980

[32] Guo Dongyan, Wang Jun, Cui Ying, et al. SiamCAR: Siamese fully convolutional classification and regression for visual tracking[C]// Proc of the 33rd IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2020: 6268−6276

[33] Chen Zedu, Zhong Bineng, Li Guorong, et al. Siamese box adaptive network for visual tracking[C]//Proc of the 33rd IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2020: 6667−6676



Zhou Ze, born in 1995. PhD candidate. Member of CCF. His main research interests include computer vision, visual tracking, and adversarial learning.

周　泽, 1995 年生. 博士研究生. CCF 会员. 主要研究方向为计算机视觉、视觉跟踪、对抗学习.



Sun Yinghui, born in 1992. PhD candidate. Her main research interests include machine learning and multi-view learning.

孙颖慧, 1992 年生. 博士研究生. 主要研究方向为机器学习、多视图学习.



Sun Quansen, born in 1963. PhD, professor. Member of CCF. His main research interests include pattern recognition, machine learning, and image processing.

孙权森, 1963 年生. 博士, 教授. CCF 会员. 主要研究方向为模式识别、机器学习、图像处理.



Shen Xiaobo, born in 1989. PhD, professor. Member of CCF. His main research interests include machine learning and pattern recognition.

沈肖波, 1989 年生. 博士, 教授. CCF 会员. 主要研究方向为机器学习、模式识别.



Zheng Yuhui, born in 1982. PhD, professor. Member of CCF. His main research interests include computer vision, artificial intelligence, and data analysis.

郑钰辉, 1982 年生. 博士, 教授. CCF 会员. 主要研究方向为计算机视觉、人工智能、数据分析.