白夜追凶

目　录

第1章 作品概述

1.1 作品产生背景

行人重识别（Re-ID）是一项基于计算机视觉来匹配不同视频和图像中查找特定行人的任务，被广泛应用于监控和安全应用中。近年来，许多重识别技术被应用于追踪犯罪嫌疑人，行人检测等安全领域，例如 RGB-红外行人再识别（RGB-IR Re-ID）。

然而 RGB-红外行人再识别技术是一种基于深度神经网络（DNN）的模型，被证实容易受到恶意攻击，存在一定的安全隐患。对此现象，我们团队首次提出了一种物理对抗攻击的方法。我们通过一种名为 AdvPull 的新算法，生成特殊的对抗纹理，并且样本固定在衣物上，使得在红外和 RGB 摄像机下捕获的行人图像的特征差异显著，以至于不能被重识别模型匹配从而达到隐私保护和增强模型鲁棒性的作用。

1.2 应用价值

在深度学习的影响下，许多现有的重识别工作都是基于深度神经网络（DNNs）搭建的，然而 DNNs 模型非常容易受到对抗攻击的影响，所以我们需要研究更加符合实际的对抗攻击策略来辅助强化模型。

1. 我们团队引入的新型物理对抗攻击方法不同于传统对抗攻击策略（即在数字层面通过对图片像素点增加扰动而达到欺骗模型的目的），我们在物理层面增加对抗性图案。这解决了获取捕获图像信息困难的问题，可以确保对抗性样本能够出现在任何摄像头识别的图像中，加强了对抗攻击的可操作性。
2. 我们的攻击模式可以在黑盒中实现，适用于我们预先无法了模型信息的情况。
3. 我们的对抗样本识别前就准备好的不需要实时生成使攻击更加严密。

第2章 问题分析

2.1 问题来源

RGB-红外行人重识别（RGB-IR Re-ID）是在大型非重叠多视角摄像机网络获取的大量视频帧中查找目标行人的任务，可以视为目标检索问题。传统的行人重识别是基于 RGB 图像的，然而，夜间摄像机很难捕捉清晰的 RGB 图像，通常在这个时候会采用红外摄影。例如，在追捕犯罪分子的任务中，许多犯罪分子在夜间犯罪，我们通常只能获得犯罪分子夜间行动的红外图像；此时，如果我们想要重新找到犯罪分子的行踪，就必须调用其他场景的摄像头拍摄的 RGB 图像和夜间拍摄的红外图像，通过相似度评分排序来锁定目标人物。反之，当我们查询的目标人物在 RGB 场景中被拍摄时，我们也可以补充使用红外场景来查询其在红外场景中的匹配。但是这种跨模态行人重识别技术尚有漏洞，安全性低，我们需要开发一种新型的对抗攻击产品从而来改善重识别技术。

2.2 现有解决方案

自从 DNN 的漏洞被发现以来，在图像识别领域，已经发现了许多针对 RGB 图像的攻击、针对红外图像的攻击以及同时针对两种模式的跨模式攻击。然而目前，针对深度攻击主要集中在 RGB 领域。在 advPattern: Physical-World Attacks on Deep Person Re-Identification via Adversarially Transformable Patterns[i]一文中首次实现了针对深度 Re-ID 的物理对抗攻击，并设计了两种攻击模式来规避搜索和假冒目标两种攻击模式。Beyond Universal Person Re-ID Attack[ii]设计了一种通用的对抗扰动（MUAP）来攻击 Re-ID 模型，即通过破坏相似性排序来攻击模型。通过诱使深度再识别系统学习错误的相似性排名。Learning to Attack Real-World Models for Person Re-identification via Virtual-Guided Meta-Learning[iii]在对抗性攻击中引入了一种金属学习方法，在多个数据集之间实施梯度交互，以找到一般对抗性扰动。这种扰动是多个数据集字段的重叠部分，使攻击效果广泛且显著。Attack is the Best Defense: Towards Preemptive-Protection Person Re-Identification[iv]中作者使用对抗攻击作为目标保护方法是独一无二的。它设置的攻击使探测器的特征与图库中的目标 ID 与图库中的目标 ID 不匹配，并且与其他已识别的 ID 不同。通过 "隔离 "探针的方法，使任何探针都无法通过 Re-ID 模

型查询到该探针。U-Turn: Crafting Adversarial Queries with Opposite-Direction Features[v]直接执行了一种对探针进行逆向特征攻击，并通过完全推送探针来生成对抗样本。即把探针的特征完全推向反方向，从而生成对抗样本，使其在图库中被正确匹配的概率降到最低。

然而现有应用于识别和分类领域的跨模态攻击方法都相对较简单，只是将红外和 RGB 模式中的分类错误用一个二元损失关系结合起来设计而成的。然而，无论是分类还是重识别任务，许多模型都可以学习融合两种模态的特征来高效完成任务，这要求我们更全面地审视损失。

现有的攻击都针对 RGB 单模态重识别任务，这意味着当我们使用红外图像作为补充来匹配目标时，该攻击失败；另一方面，当红外图像本身被用作查询目标时，针对其的攻击也是一个蓝海。在这种情况下，掌握 RGB-IR Re-ID 的跨模态攻击的犯罪嫌疑人可以轻松逃脱视频监视的追捕。

2.3 本作品要解决的痛点问题

为了解决上述问题，我们通过 advPull 方法设计了一种物理对抗模式，即带有对抗纹理的对抗样本。我们为现有的跨模态 Re-ID 引入了特征融合损失、相似度排序损失和模态转换损失；我们尽力扩大红外图像和 RGB 图像中行人之间的特征差距，并减小两种模态中行人之间的差异，以实现攻击效果；同时我们将对抗性攻击转化为优化问题，最终确定最优样本贴片的形状、温度和纹理像素。

2.4 解决问题的思路

对于跨模态 Re-ID 的三种常见模型类别，我们从三个维度对我们的攻击方法进行建模。引入了特征融合损失、相似度排序损失和模态转换损失，尽力扩大红外图像和 RGB 图像中行人之间的特征差距，并减小两种模态中行人之间的差异，以实现显著攻击效果。最后综合现实因素使用 WRAP 方法使补丁更加隐蔽（即使带有补丁的衣物与平常衣服无异），用 3D EOT 技术模拟现实中的物理场景因素来增强产品的鲁棒性。

第3章 技术方案

3.1.1 特征提取

考虑到现有的跨模态再识别模型大多使用神经网络提取 RGB 图像红外图像的共同特征来判断它们是否匹配。我们力求在引入对抗补丁后，拉长两者特征之间的距离，从而使 Re-ID 无法正确匹配目标。假设目标人物的 RGB 图像为I_{rgb}，红外图像为I_{ir}，而我们引入的对抗补丁为 δ，因此我们认为引入对抗补丁后的 RGB 图像为$I'_{rgb} = I_{rgb} + \delta$, $I'_{ir} = I_{ir} + \delta$,这些是我们的对抗样本。需要注意的是，$\delta$ 在 RGB 图像和红外图像中的表现不同。在 RGB 图像中，δ 上的对抗贴片表面的纹理对图像像素的影响起主要作用，而在红外图像中，主要是δ 补丁的材料通过自身温度释放热辐射，改变红外图像中的像素。

首先，我们使用 FAST 算法提取I'_{rgb} 的特征点。我们将特征点集合排序为$\{p_r^1, p_r^2 \dots p_r^n\}$。同样，我们提取$I'_{ir}$ 的特征点为$\{p_r^1, p_r^2 \dots p_r^m\}$。然后，我们使用 rRBRIEF 算法将这些特征点表示为k阶二元描述符：

$$p_r^{\mathrm{t}} = (a_1, a_2 \dots a_k), \quad t = 1,2 \dots n$$

$$p_r^{\mathrm{j}} = (a_1, a_2 \dots a_k), \quad j = 1,2 \dots m$$

其中，$a_1, a_2 \dots a_k$ 只能取 0 或 1。我们使用汉明距离$d(x, y)$来表示特征之间的距离。具体来说，可以表示为:

$$d\left(p_r^t, p_r^j\right) = \sum_{s=1}^{k} \left| p_r^t[s] - p_i^j[s] \right|$$

最后，我们的目标是最大化匹配特征点之间的距离，因此特征匹配的损失可以设计为:

$$L_{fea} = -\sum_{j=1}^{m} d\left(p_i^j, arg \min_{p_r^t} d\left(p_r^t, p_r^j\right)\right) \ t \in \mathbb{Z}, 1 \le t \le n$$

3.1.2 学习误差排序

Re-ID 问题归根结底是一个相似度排序问题，只要我们能够破坏模型的排序，尽可能降低正确匹配图像的相似度排序，也就是让模型学习错误的排序，就能实现有效攻击。假设我们攻击的深度 Re-ID 模型是$f_\theta(.,.)$，其中 θ 是模型参数。由于我们的攻击是先发制人的黑盒攻击，我们无法提前知道模型参数。不过，由

于大多数深度 Re-ID 模型的底层相似性评估方法仍然基于图像之间的余弦相似性，因此我们可以将 $f_\theta(.,.)$表示如下：

$$f_\theta(I'_{rgb}, I'_{ir}) = \frac{\sum_{s=1}^{n} p_r^s \, p_i^s}{\left\|I'_{rgb}\right\|_2 \cdot \left\|I'_{ir}\right\|_2}$$

其中，p_r^s和p_i^s分别代表 RGB 图像和红外图像中的第 s 个像素点。我们用 $rank\left(f_\theta(I'_{rgb}, I'_{ir})\right)$ 表示这两张图片在图库中的相似度排序。我们认为排名k的图片对最终被识别为匹配对象的概率服从参数为λ的泊松分布，即$P(rank = k) = \frac{\lambda^k e^{-\lambda}}{k!}$. 因此，我们的目标可以表示为:

$$min \sum_{i=1}^{I} \sum_{k=1}^{K} P\left(rank\left(f_\theta(I'_{ir}[k], I'_{rgb}[k])\right)\right)$$

这是在物体探测器处于红外模式下产出的公式，而我们希望在可见光视频库中完成重新识别任务建立。这里 i 代表第 i 个摄像机拍摄的视频图库。而 k 表示目标行人的 ID 是 k。但正如文献 Attack is the Best Defense: Towards Preemptive-Protection Person Re-Identification[vi]所述，单纯使用图像攻击攻击是无效的，因为我们的对抗样本很可能无法匹配其他 ID，并且过于突兀容易被识别为异常样本。如果我们的对抗样本可以与其他 ID 匹配，达到一定程度的冒充，那么我们的攻击就可以更加隐蔽：

$$max \sum_{i=1}^{I} \sum_{k=1}^{K} \sum_{t=1, t\neq K}^{T} P\left(rank\left(f_\theta(I'_{ir}[k], I'_{rgb-i}[t])\right)\right)$$

这里的 t 代表与 k 不同的行人 ID。通过最大化不同 ID 之间的匹配概率，我们就能达到冒名顶替的效果。最后，我们的损失 h 函数可以描述为：

$$L_{er} = \sum_{i=1}^{I} \sum_{k=1}^{K} P\left(rank\left(f_\theta(I'_{ir}[k], I'_{rgb}[k])\right)\right)$$
$$- \sum_{i=1}^{I} \sum_{k=1}^{K} \sum_{t=1, t\neq K}^{T} P\left(rank\left(f_\theta\left(I'_{ir}[k], I'_{rgb-i}[t]\right)\right)\right)$$

同理的，如果我们使用 RGB 图像作为查询探针，红外图像作为匹配探针，损失可以描述为：

$$L_{er} = \sum_{i=1}^{I} \sum_{k=1}^{K} P\left(rank\left(f_\theta(I'_{rgb}[k], I'_{ir-i}[k])\right)\right)$$

$$-\sum_{i=1}^{I}\sum_{k=1}^{K}\sum_{t=1,t\neq K}^{T}P\left(rank\left(f_\theta\left(I_{rgb}^{'}[k],I_{ir-i}^{'}[t]\right)\right)\right)$$

3.1.3 模式间信息消融

模态间信息消减是为了防止 Re-ID 系统通过模态转换将 RGB 图像中恢复的灰度信息与红外图像进行匹配，信息消减本身也可以减少两种模态的特征重叠。由于现有的灰度还原方法大多基于三个通道的加权平均或其中一个通道的一致组合，因此我们不妨提取 RGB 图像中的每个通道信息，并在红外图像中构建信息损失。具体来说，我们将 RGB 图像分解为：$I_R'=\left(I_{rgb_x}^{'R},I_{rgb_y}^{'R},I_{rgb_z}^{'R}\right)$，$I_G'=\left(I_{rgb_x}^{'G},I_{rgb_y}^{'G},I_{rgb_z}^{'G}\right)$，$I_B'=\left(I_{rgb_x}^{'B},I_{rgb_y}^{'B},I_{rgb_z}^{'B}\right)$，这是为了增强每幅 RGB 图像的单通道，并将其分解为三幅灰度图像。这里用 x 表示 RGB 图像中某一点的灰度值，我们将三通道增强图像和红外图像的灰度分布离散函数分别表示为 $I_R'(x)$，$I_G'(x)$，$I_B'(x)$，$I_{ir}'(y)$。接下来，我们用$P_R(x,y),P_G(x,y),P_B(x,y)$来表示同一坐标像素点上信道增强图和红外图像的联合概率分布函数。所以它们之间的归一化互信息可表示为：

$$NMI(I_R',I_{ir}')=\frac{\sum_{x\in X}I_R'(x)logI_R'(x)+\sum_{y\in Y}I_{ir}'(y)logI_{ir}'(y)}{\sum_{x\in X}\sum_{y\in Y}P_R(x,y)log\frac{P_R(x,y)}{I_R'(x)I_{ir}'(y)}}$$

$$NMI(I_G',I_{ir}')=\frac{\sum_{x\in X}I_G'(x)logI_G'(x)+\sum_{y\in Y}I_{ir}'(y)logI_{ir}'(y)}{\sum_{x\in X}\sum_{y\in Y}P_G(x,y)log\frac{P_G(x,y)}{I_G'(x)I_{ir}'(y)}}$$

$$NMI(I_B',I_{ir}')=\frac{\sum_{x\in X}I_B'(x)logI_B'(x)+\sum_{y\in Y}I_{ir}'(y)logI_{ir}'(y)}{\sum_{x\in X}\sum_{y\in Y}P_B(x,y)log\frac{P_B(x,y)}{I_B'(x)I_{ir}'(y)}}$$

我们的目标是使归一化后互信息尽可能小，以降低模式相互切换的可能性和信息传递的强度，同时在一定程度上降低模式之间的相似性。所以损失函数可以可以设计为：

$$L_{inf}=-\left(NMI(I_R',I_{ir}')+NMI(I_G',I_{ir}')+NMI(I_B',I_{ir}')\right)$$

最后我们的目标函数如下：

$$min\quad L_{adv}=L_{fea}+\lambda_1L_{er}+\lambda_2L_{inf}$$

上述λ_1和λ_2是用于控制权重的系数。

3.1.4 对抗样本的生成

对抗样本的生成本质上是生成对抗补丁δ 的过程。但是在应用优化方法生成δ 之前，我们需要考虑两个问题：补丁的隐蔽性和物理变化场景中的通用性。

1. 补丁隐藏

虽然我们之前的目标函数可以帮助我们全面攻克深度跨模态再识别模型，但如果这个目标函数优化生成的补丁δ不受到特定的约束，在现实中生成的补丁很有可能分布怪异，容易引起注意。这样做的后果是，如果现实场景中有检查人员，会立即发现异常。为此，我们必须对生成的δ 做一定的限制，使其生成的纹理更加自然，符合正常人的审美观。我们首先按照 shaferi et al.的观点来最小化总的波动函数 $TV\ (\delta)$。该函数可以通过限制相邻像素的变化来实现整体像素分布的均匀性和平滑性。这样设计出的补丁显然更符合人体的视觉常态：

$$min\ \ TV(\delta) = \sum_{p,q} \left(\left(\delta_{p,q} - \delta_{p+1,q}\right)^2 + \left(\delta_{p,q} - \delta_{p,q} + 1\right)^2 \right)^{\frac{1}{2}}$$

另一方面，考虑到我们的补丁最终要贴在衣服上，因此我们努力使其看起来与普通衣服无异。针对这一目标，我们提出了创新的解决方案。我们将贴片的形状限定为 S，然后寻找与其大小最接近的卡通图案形状 S'（如熊、猫等），并通过边界填充将 S映射到 S'（具体来说，我们使用了外部 WRAP 方法）。这样一来，装载探测器的人物所穿的衣服与街上的普通行人无异，方法表述如下：

$$patch = \delta(S) + \mathrm{WRAP}_{\delta}(S' - S) \mathrm{s.t.} S \subseteq S'$$

2. 物理变化场景的通用性

事实上，当我们把在数字领域生成的对抗性补丁移植到现实世界时，会出现很多问题。从摄像机的角度来看，行人之间的距离、当天天气影响的光照以及摄像机拍照的角度都会影响对抗样本的质量。至于行人本身，对抗补丁本身是贴在衣服上的，因此行走时衣服的褶皱也会影响对抗样本的效果。以上这些在现实中实际存在的情况，都对我们的对抗攻击方法的鲁棒性提出了挑战。

针对上述挑战，我们决定引入 3D EOT BTI reliability improvement strategies in low thermal budget gate stacks for 3D sequential integration[vii]来增强对抗样本的鲁棒性。具体来说，这是一种使用 3D 渲染函数$t(.)$来模拟真实世界中光照、距离、角度和形变变化的模型。通过在优化过程中对对抗样本进行 EOT 处理，我们可以让它在面对真实物理场景时更加稳健。因此，我们可以这样更新对抗样本：

$$I'_{rgb} = t(I_{rgb} + \delta) s.t. \mathbb{E}_{t\sim T}\left[d\left(I'_{rgb}, t(I_{rgb})\right)\right] \leq \epsilon$$

此外，它还限制了原始图像与经历 EOT 变化后的对抗图像之间的差异小于一定范围。这可以在一定程度上限制 EOT 的变化，更符合物理世界的客观规律。

第4章 系统实现

我们基于 SYSU-MM01 和 Regdb 数据集进行训练，运用 Adam 算法进行优化，攻击效果显著。详见附件内程序。

第5章 测试分析

表 1　数据对比

模型	数据集	Rank-1	Rank-5	Rank-10	mAP	ss
AGW	SYSU-MM01	0.0%	0.0%	0.0%	3.8%	0.393
	Regdb	0.01%	0.0%	0.0%	4.8%	0.452
DDAG	SYSU-MM01	0.01%	0.02%	0.04%	4.3%	0.422
	Regdb	0.0%	0.0%	0.0%	3.6%	0.362
DEEN	SYSU-MM01	0.0%	0.0%	0.0%	3.9%	0.401
	Regdb	0.0%	0.0%	0.0%	4.5%	0.445

第6章 作品总结

6.1 作品特色与创新点

1.本产品引入物理对抗方案，增强了攻击的可操作性

2.本产品考虑了物理场景的通用性，隐藏性相对较高

6.2 应用推广

本产品是通过形成的对抗样本并且固定在衣物上，在之后的应用中可能会移植到其他不同的载体中（如：可穿戴设备，生活用品等）

6.3 作品展望

在今后，我们团队会持续维护和优化我们的对抗性攻击方案。

参考文献

[i] Wang, Zhibo, et al. "advpattern: Physical-world attacks on deep person re-identification via adversarially transformable patterns." *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2019.

[ii] Ding, Wenjie, et al. "Beyond Universal Person Re-ID Attack." *arXiv preprint arXiv:1910.14184* (2019).

[iii] Yang, Fengxiang, et al. "Learning to attack real-world models for person re-identification via virtual-guided meta-learning." *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 35. No. 4. 2021.

[iv] Wang, Lin, et al. "Attack is the best defense: Towards preemptive-protection person re-identification." *Proceedings of the 30th ACM International Conference on Multimedia*. 2022.

[v] Zheng, Z., Zheng, L., Yang, Y. et al. U-Turn: Crafting Adversarial Queries with Opposite-Direction Features. Int J Comput Vis 131, 835–854 (2023). https://doi.org/10.1007/s11263-022-01737-y

[vii] Franco, Jacopo, et al. "BTI reliability improvement strategies in low thermal budget gate stacks for 3D sequential integration." *2018 IEEE International Electron Devices Meeting (IEDM)*. IEEE, 2018.