目录

第 1 章 研究背景与意义

1.1 研究背景

安全存在的意义是为了保护有价值的东西。远古时期，人们开始考察聚集地地理位置，这里面便涉及到了野兽侵袭、部落冲突、自然灾害、食物与水资源获取等等安全问题。在自然界，从庞大的蚁群巢穴到大型动物群落，哪怕是植物都极度重视安全议题 (植物之间都有着信息素的传递，通过信息传递，释放化学物质防护群落受到虫害的威胁)。**安全**的概念似乎印制在了生物的 DNA 里，这是一种最为本质的生物设计，**以小价值的付出保护大价值的资源**，让生物的基因能够更好的传承。这样的概念甚至不必去传授，我们便有深刻的理解。

自深度学习问世以来，各种人工智能应用经历了迅猛的增长，特别是在计算机视觉任务中的广泛应用。然而，近年来，随着深度学习模型的普及和应用场景的拓展，越来越多的涉及计算机视觉的人工智能安全问题浮现，对抗样本攻击逐渐成为人们关注的焦点。

对抗样本是指对深度学习模型 (针对模型本身的先验知识也至关重要，我们在后文会讨论) 进行特意制作的输入数据，经过微小而几乎不可察觉的扰动，却能导致模型输出完全错误的结果。这种现象引起了人们对深度学习模型鲁棒性的关注。尽管深度学习在计算机视觉等领域取得了巨大的性能提升，但深度模型的运作机制与人类认知过程存在本质差异，使得它们容易受到对抗样本攻击。

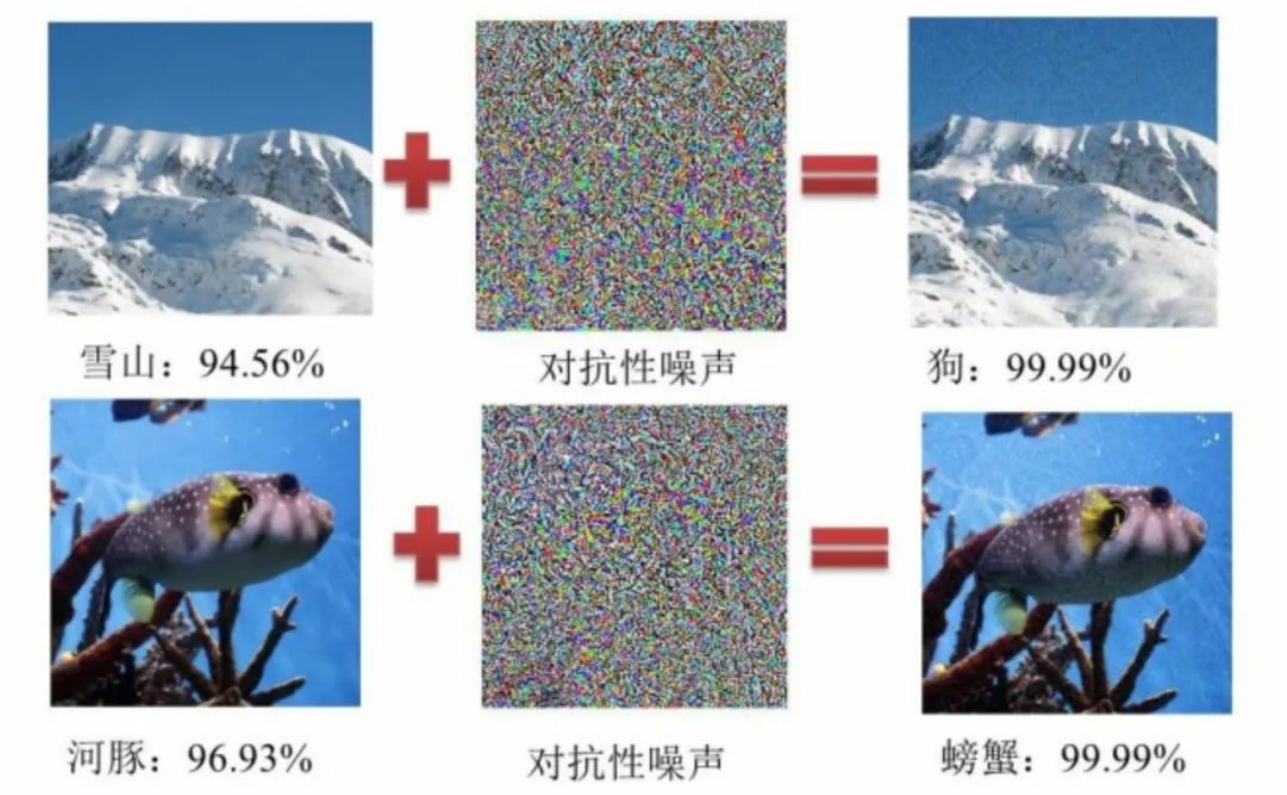

图 1.1: 对抗样本攻击示意 (图片识别分类完全错误)

1.2 研究意义

对抗样本的出现对计算机视觉任务的应用和发展产生了影响，尤其是对目标分类、目标检测和目标追踪等基本任务的实现提出了挑战。这些任务的安全性对自动化机器人、无人驾驶和无人机等重要设施至关重要。因

此，研究对抗样本的原理，探索如何提高深度学习模型计算机视觉的安全性和鲁棒性，成为了当前人工智能领域的一个紧迫问题。

我们通过深入研究对抗样本攻击技术，可以更好地理解深度学习模型的内部机制，并为其安全落地提供保障。这项工作不仅有助于改进现有的深度学习模型，还能够为未来人工智能技术的发展提供更加稳健和可靠的基础。因此，对抗样本的研究具有重要的学术价值和实际应用意义，将为人工智能技术的发展提供新的方向和可能性。

第 2 章 研究历史与现状

2.1 研究历史

2014 年附近，对抗样本相关研究逐渐火热。早期的研究集中于图像分类上，如何让模型对加入干扰之后的图像错误分类成为了首要的研究目标。

随着对抗样本攻击的研究不断深入，后来出现了针对具体深度学习任务的攻击算法。

在对抗攻击中，通常会使用目标模型的梯度来指导攻击过程，这些梯度是基于目标模型的结构和参数等先验知识进行计算的。

但是现实中有时很难获取目标模型的信息。(这也是日后对抗样本的研究对象分类的一个标准) 为了解决这个问题，研究人员尝试使用近似梯度来代替实际模型的梯度。另一种方法是使用数值估计技术，来模拟目标模型的梯度。这些方法的出现丰富了对抗攻击的技术手段，为研究人员提供了更多灵活性和选择。

2.2 研究现状

目前对抗样本按照不同的攻击效果和攻击环境有诸多细分方向研究。

我们主要介绍三个大的分类，这三个大分类是目前研究的重心。

根据攻击后效果分类

- ❑ 定向攻击 Targeted Attack
- ❑ 非定向攻击 Non-Targeted Attack

定向攻击要做到既降低模型对输入样本真实标签的置信度，又要提升攻击者指定标签的置信度（困难）。

非定向攻击只要做到将模型结果误导到其他错误的类别（简单）。

按对攻击模型的知识分类

- ❑ 白盒攻击 White-Box Attack
- ❑ 黑盒攻击 Black-Box Attack

白盒攻击是攻击者已知目标模型所有知识的情况下生成对抗图样的攻击手段。

黑盒攻击是攻击者不知道模型任何内部信息的情况下实施的攻击方案。

按攻击环境分类

- ❑ 数字世界攻击 Digital Attack
- ❑ 物理世界攻击 Physical Attack

数字世界攻击假定可以直接访问并修改 AI 系统的数字输入。

物理世界攻击是指输入是从真实世界的采集设备中而传入系统。

在以上三个大的分类标准下，具体的小标准也众多。

近年，研究者们在目标检测攻击方面研究逐渐深入。这是有以下几点原因，首先传统的**白盒数字攻击**是一种理论上的研究，因为这样的攻击对模型的了解需要苛刻的要求 (了解内部参数、设计等等知识)，但是对图像的采集要求却很低 (只需要进行数字图像处理，没有任何的扭曲畸变，不涉及 patch 失真等问题)，因此这是一种距离物理世界攻击有一定距离的研究。

然而，此种攻击技术必然会投入到实践中来检验

Lu 等 [41] 人提出了一种针对目标检测器的攻击算法，该算法可以成功欺骗 Faster R-CNN 和 YOLO 模型。他们在停车标志上进行了实验，结果显示可以产生很多错误的误导性识别结果。

Thys[50] 等人提出了一个针对行人检测器的对抗性贴图系统，也被称为补丁系统，该系统可以将这些贴图打印出来，用于在物理世界中执行欺骗性行为。

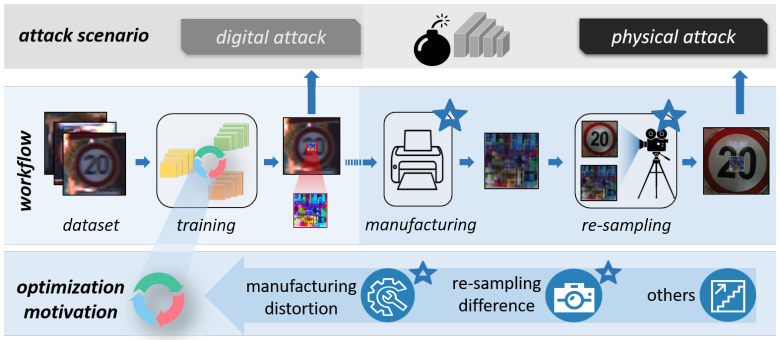

图 2.1: 攻击环境的简要示意图

浙江大学的 usslab 实验室提出里 CAPatch 改善攻击稳定性和提升攻击性能，经实验尺寸为图像的 5% 的 CAPatch 可以在视频记录器上实现连续攻击，攻击成功率高于 73.1%。

而他们的又创新性地提出了 Tpatch，一种由声学信号触发的物理对抗性贴片。在正常情况下保持良性，但可以通过向摄像头注入信号攻击引入的设计失真来触发启动隐藏、创建或改变攻击。

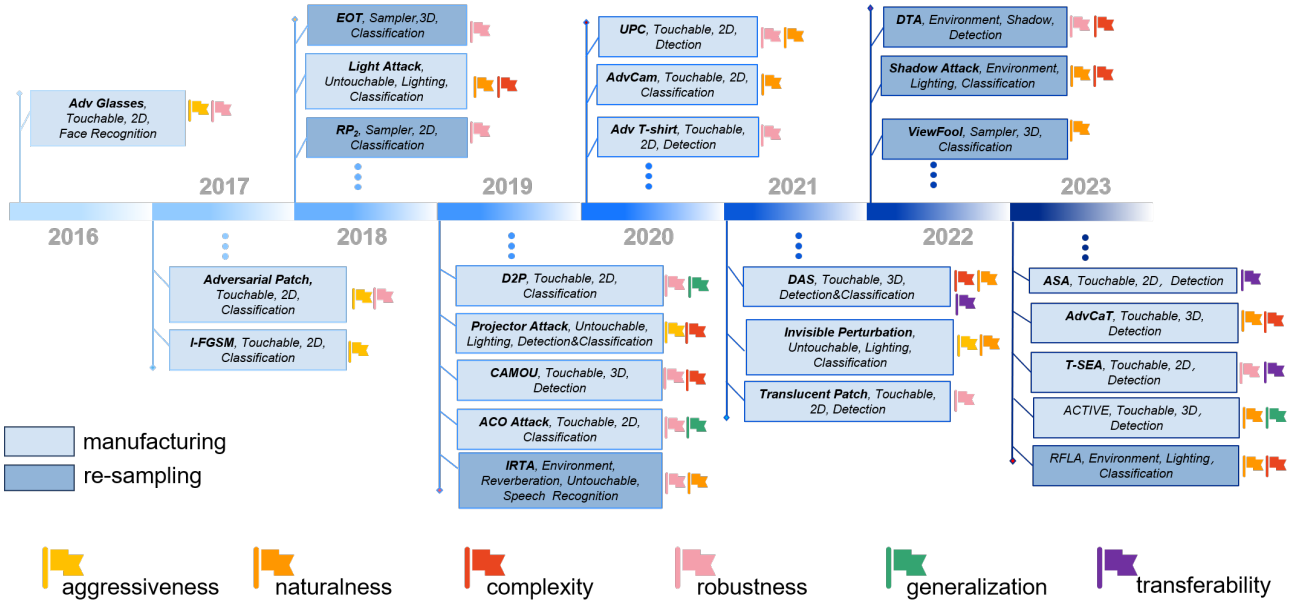

图 2.2: 物理世界攻击发展脉络

第 3 章 研究目标

对抗样本的研究领域中，已经细分出众多的分支领域。为了减小研究难度，我们主要研究**白盒非定向物理世界攻击技术**。最终的目标是**生成**可以欺骗特定识别模型 **YOLOv2** 人类分类模块的 **2DPatch**，研究的结果的直接呈现是 **2DPatch**。同时将识别误导 Success Rate 从针对 YOLOv2 学术界普遍的 88.4% 提升到 90%。

完成上述基础研究目标后 (**生成针对 YOLOv2detector 进行任务的 patch**)，时间允许的前提下，我们将致力于研究如下附加问题的一个或多个。

- 如何强化 patch 在不同模型之前的可迁移性
- 针对柔性材料或者在强畸变条件下可使用的 patch 生成
- 3d 空间内可使用的 patch
- 黑盒攻击 (也是偏向真实世界的一种攻击方式，知道少量或者完全不知道模型的参数和设计原理)
- 针对不同光照、不同空间透视、复杂人物衣着、皮不同肤颜色情况下仍能有效完成攻击的 Patch

第 4 章 研究内容

研究内容是对抗样本的研究领域中白盒非定向物理世界攻击技术，但是我们需要了解的内容包括所有类型的攻击方式,这包括了前人已经研究过各种攻击技术如 L-BFGS 、FGSM 、IGSM 、Iter.1.1 、JSMA 、Uni.perturbations 、DeepFool 、One-pixel、CW Attack、Ensemble Attack 同时我们也会研究他们在各自模型上的 success rate 来了解业界攻击性能情况

Method	Black/White-box	Attack Type	Targeted/Non-targeted	PASS
L-BFGS [24]	White-box	Gradient-based	Targeted	*
FGSM [27]	White-box	Gradient-based	Non-targeted	* * * * *
IGSM [33]	White-box	Gradient-based	Targeted/Non-targeted	**
Iter.l.l [33]	White-box	Gradient-based	Non-targeted	**
JSMA [31]	White-box	Gradient-based	Targeted	**
Uni.perturbations [46]	White-box	Decision boundary-based	Non-targeted	* * *
DeepFool [45]	White-box	Decision boundary-based	Non-targeted	*
One-pixel [53]	Black-box	-	Non-targeted	* * **
CW Attack [47]	White-box	Iterative optimization	Targeted/Non-targeted	*
Ensemble Attack [39]	White-box	Ensemble optimization	Non-targeted	* * *

图 4.1: 攻击类型

Method	Advantage	Disadvantage
L-BFGS [24]	High stability and effectiveness	High computational and time complexity
FGSM [27]	Low computational complexity, high transfer rate	Low success rate, label leaking
IGSM [33]	Small perturbations, high success rate	Low transfer rate, low success rate for balck-box attacks
Iter.l.l [33]	Small perturbations, high success rate	Low transfer rate, low success rate for balck-box attacks
JSMA [31]	Small perturbations, high success rate	High computational complexity
Uni.perturbations [46]	High generalization ability, high transfer rate	The perturbation is not easy to control, low success rate for target attacks
DeepFool [45]	Low computational complexity, small perturbations	Low success rate for balck-box attacks
One-pixel [53]	Low computational complexity	Low success rate for target attacks, large perturbations
CW Attack [47]	Small perturbations, high transfer rate and success rate	High computational complexity
Ensemble Attack [39]	Simple computation, good generalization	Low success rate for balck-box attacks

图 4.2: 主流攻击技术的优缺点分析

Method	Black/White-box	Attack Type	Targeted/Non-targeted	PASS
L-BFGS [24]	White-box	Gradient-based	Targeted	*
FGSM [27]	White-box	Gradient-based	Non-targeted	* * * * *
IGSM [33]	White-box	Gradient-based	Targeted/Non-targeted	**
Iter.l.l [33]	White-box	Gradient-based	Non-targeted	**
JSMA [31]	White-box	Gradient-based	Targeted	**
Uni.perturbations [46]	White-box	Decision boundary-based	Non-targeted	* * *
DeepFool [45]	White-box	Decision boundary-based	Non-targeted	*
One-pixel [53]	Black-box	-	Non-targeted	* * **
CW Attack [47]	White-box	Iterative optimization	Targeted/Non-targeted	*
Ensemble Attack [39]	White-box	Ensemble optimization	Non-targeted	* * *

图 4.3: 性能展示

第 5 章 关键问题与难点分析

在真实的物理攻击中在传入 AI 系统的过程中，包括摄像头等设备，会受到多种因素的影响，从而可能导致对抗样本性能下降。这些因素主要包括传感器误差、预处理误差以及模型输入的误差。由于我们选择了白盒 + 非定向, 因此此方面的技术难点我们不去分析, 因为不是研究的重点, 实际上, 这两个特征的改变, 将极大增加攻击难度, 黑盒攻击时, 传统的梯度攻击方法只能更改为估计梯度攻击方法, 而定向攻击时, 将导致攻击 success rate 大幅下降, 可迁移性近乎不再存在.

物理世界攻击的难点

- 传感器误差
- 预处理误差
- 模型输入误差

传感器误差：摄像头等传感器可能受到噪声、光照变化、震动等因素的影响，导致采集到的图像存在一定程度的失真或变形。这些误差会影响到攻击样本的质量，使得攻击后的图像与原始目标有所偏差。

预处理误差：攻击样本在被传入模型之前可能会经过一系列预处理步骤，如缩放、光照变换、旋转等。这些预处理操作可能会引入新的误差，进一步影响攻击样本的性能。例如，图像缩放或光照变换可能会改变图像的像素分布，从而使得模型难以准确识别攻击目标。

模型输入误差：最后，攻击样本被输入到模型进行预测。模型可能对输入图像进行某些操作或者存在输入限制，这些因素也会影响攻击样本的效果。例如，在自动驾驶系统中，模型可能对图像进行裁剪或者要求特定的分辨率，这些要求可能会导致攻击样本失效或者降低攻击效果。

综上所述，传入 AI 系统的攻击样本受到多种因素的影响，包括传感器误差、预处理误差和模型输入误差等。为了提高对抗样本的性能，需要综合考虑这些因素，并设计相应的对抗攻击策略来应对不同情况下的误差影响。

因此物理世界攻击时候的关键问题就是如何尽可能减小这些 patch 失真带来的影响, 数字世界攻击便不存在如此丰富各种各样的失真, 这是区别两者的一个关键点.

本研究的难点

- 人物出现的环境复杂
- 人物本身特征的复杂
- 人物针对采集器的不同角度
- patch 本身的不固定性

分析这些难点我们可以知道人物本身特征的复杂: 服装肤色、体型、姿势等都各不相同。

人物出现的环境复杂: 人可以出现在许多不同复杂的背景下。

人物针对采集器的不同角度: 一个人的外观会因其是面向镜头还是背对镜头而有所不同。

patch 本身的不固定性: 人的身上没有一个一致的位置可以放置我们的 patch。patch 可能发生移动, 不是静态攻击。

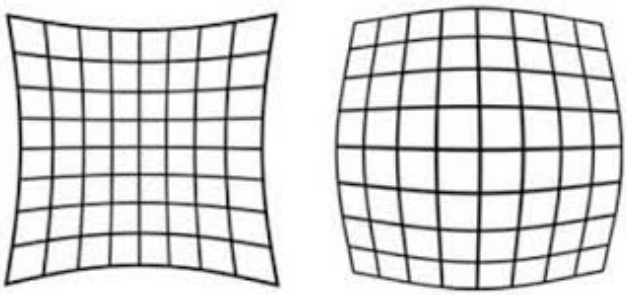

图 5.1: 图形畸变举例

第 6 章 技术路线和研究方案

此项研究中我们计划使用 Adam algorithm。Adam 算法计算神经网络参数 θ 的更新如下：

输入： 学习率 α，衰减率 β_1、β_2，小常数 ϵ，初始参数 θ。

初始化： 初步一阶矩 m_0、二阶矩 v_0 为零向量。

迭代： 对于每个时间步 $t = 1, 2, \ldots, T$，执行以下步骤：

1. 计算梯度：计算当前梯度 g_t。
2. 更新一阶矩：
   $m_t \leftarrow \beta_1 \cdot m_{t-1} + (1 - \beta_1) \cdot g_t$。
3. 更新二阶矩：
   $v_t \leftarrow \beta_2 \cdot v_{t-1} + (1 - \beta_2) \cdot g_t^2$。
4. 修正偏差：
   $\hat{m}_t \leftarrow \frac{m_t}{1-\beta_1^t}$，
   $\hat{v}_t \leftarrow \frac{v_t}{1-\beta_2^t}$。
5. 更新参数：
   $\theta \leftarrow \theta - \frac{\alpha}{\sqrt{\hat{v}_t}+\epsilon} \cdot \hat{m}_t$。

输出： 更新后的参数 θ。

这些方程描述了 Adam 算法中参数更新的过程，其中学习率和动量都会根据梯度的特点而自适应调整，使得算法在训练过程中能够更有效地收敛到最优解。

接下来优化问题的具体研究我们将在中期进行更深入的研究，但是根据我们的讨论，我们确定如下三个部分。

第一个部分我们称之为**不可打印性**，第二个部分我们称之为**图片全变量**，第三个我们称之为**最大目标值**。

我们不给出公式简要介绍三个部分的意义，第一个部分衡量打印机色彩性能，因为不是所有的色彩都可以在物理世界中的打印机中打印，第二个部分平滑 patch，提高有效性，第三个部分则是最小化误导分类置信度的物理量。

最终我们将三者按照权重相加得到我们最终的**损失函数**

我们的优化器要做到最小化损失函数，我们从一个完全随机的初始 patch 出发，每次的优化过程中，我们保留神经网络的权重，但是只更改 patch 的数值。

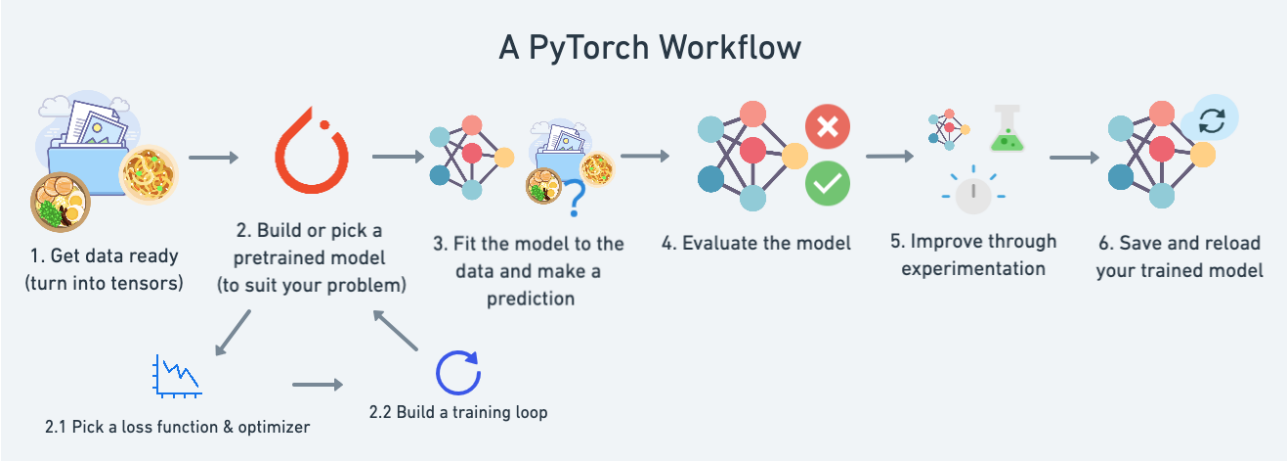

图 6.1: pytorch 进行训练示意图

第 7 章 创新点分析

传统的对抗样本, 攻击识别的目标较为简单, 比如说一个电线杆, 将 patch 挂靠在电线杆上完成针对电线杆识别的攻击。

本次研究目标创新点是要做到四个大点

本研究的创新点

- 针对人物出现的环境复杂的 patch
- 针对人物针对采集器的不同角度的 patch
- 针对人物本身特征的复杂的 patch
- 针对 patch 本身的不固定性的 patch

传统的目标, 比如说电线杆的对抗样本攻击, 不具备上面四个特征, 因为电线杆一般不会移动, 而且所出现的环境背景较为简单, 电线杆的形状一般为柱加正方形电箱, 不像人类可以有各种特征, 因此这四个创新点就是区别于传统目标的最大创新点。

第 8 章 预期研究成果

其实预期研究成果跟研究目标有部分重合, 我们最终要呈现的成果就是一张针对**白盒非定向物理世界攻击技术**的 **2Dpatch**。预取性能是**生成**可以欺骗特定识别模型 **YOLOv2** 人类分类模块的 **2DPatch**, 同时将识别误导 Success Rate 从针对 YOLOv2 学术界普遍的 88.4% 提升到 90%, 并且完善 patch 的鲁棒性和泛用性。

图 8.1: 预期结果图示 (不是这个图)

最后呈现的是 printer 打印出来的 patch, 我们的指标测量通过 YOLOv2 进行检测。

第 9 章 时间计划

第一周到第七周: 了解研究内容的基本, 进行分工, 预开展小部分代码工程, 完成开题报告以及需求分析和开题 ppt。

第七周到第十二周: 进行具体的工程开展, 理论进一步敲定, 代码实现, 以及效果结论的测试以及比对, 完成中期报告。

第十二周到第十五周完成结尾工作: 将结果撰写成结题报告。

实际上的科研进行不完全是线性发展, 可能存在灵感一现和突破性的进展, 因此此时间表仅供参考。

第 10 章 参考文献

- Wei X, Pu B, Lu J, et al. Physically adversarial attacks and defenses in computer vision: A survey[J]. arXiv preprint arXiv:2211.01671, 2022, 1(2).
- Akhtar N, Mian A, Kardan N, et al. Advances in adversarial attacks and defenses in computer vision: A survey[J]. IEEE Access, 2021, 9: 155161-155196.
- Zhang S, Cheng Y, Zhu W, et al. CAPatch: Physical Adversarial Patch against Image Captioning Systems[C]//32nd USENIX Security Symposium (USENIX Security 23). 2023: 679-696.
- Zhu W, Ji X, Cheng Y, et al. Tpatch: A triggered physical adversarial patch[J]. arXiv preprint arXiv:2401.00148, 2023.
- Hu Z, Chu W, Zhu X, et al. Physically realizable natural-looking clothing textures evade person detectors via 3d modeling[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2023: 16975-16984.
- 石震波. 基础视觉中目标分析模型的对抗攻击研究 [D]. 中国科学技术大学,2023.DOI:10.27517/d.cnki.gzkju.2023.000674.
- 张树栋. 深度神经网络中的对抗样本攻防技术研究 [D]. 西安电子科技大学,2022.DOI:10.27389/d.cnki.gxadu.2022.003201.
- Wang D, Yao W, Jiang T, et al. Rfla: A stealthy reflected light adversarial attack in the physical world[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2023: 4455-4465.
- Li Y, Li Y, Dai X, et al. Physical-world optical adversarial attacks on 3d face recognition[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2023: 24699-24708.
- Dong Y, Ruan S, Su H, et al. Viewfool: Evaluating the robustness of visual recognition to adversarial viewpoints[J]. Advances in Neural Information Processing Systems, 2022, 35: 36789-36803.
- Xu K, Zhang G, Liu S, et al. Adversarial t-shirt! evading person detectors in a physical world[C]//Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part V 16. Springer International Publishing, 2020: 665-681.
- 孙裕道. 人脸表情识别模型对抗方法研究 [D]. 北京邮电大学,2023.DOI:10.26969/d.cnki.gbydu.2023.000292.
- 刘嘉阳. 针对图像分类的对抗样本防御方法研究 [D]. 中国科学技术大学,2021.DOI:10.27517/d.cnki.gzkju.2020.000375.
- 易子博. 面向神经网络分类器的对抗样本攻击与防御关键技术研究 [D]. 国防科技大学,2021.DOI:10.27052/d.cnki.gzjgu.2021.000511
- Song D, Eykholt K, Evtimov I, et al. Physical adversarial examples for object detectors[C]//12th USENIX workshop on offensive technologies (WOOT 18). 2018.
- Chen S T, Cornelius C, Martin J, et al. Shapeshifter: Robust physical adversarial attack on faster r-cnn object detector[C]//Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2018, Dublin, Ireland, September 10–14, 2018, Proceedings, Part I 18. Springer International Publishing, 2019: 52-68.
- Eykholt K, Evtimov I, Fernandes E, et al. Robust physical-world attacks on deep learning visual classification[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2018: 1625-1634.
- Dong Y, Zhu J, Gao X S. Isometric 3d adversarial examples in the physical world[J]. Advances in Neural Information Processing Systems, 2022, 35: 19716-19731.
- J. Zhang and C. Li, “Adversarial examples: Opportunities and challenges,” IEEE Transactions on Neural Networks and Learning Systems, pp. 1–16, 2019, doi: 10.1109/TNNLS.2019.2933524.