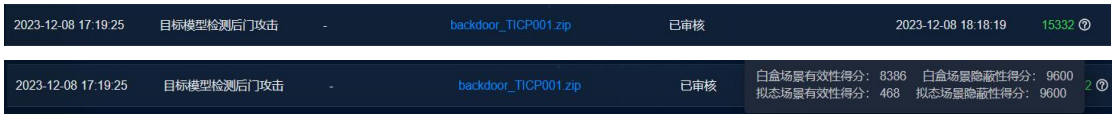
AI Competition Technical Report

A. 战队信息

（1）战队名称/Team Name:**TICP001**

（2）选择的赛题(对抗或后门):**后门**

（3）最终得分情况/Final Task Score :



B. 攻击实现流程

（1）攻击方法简介（参赛选手简单描述攻击所使用的方法与原理）

使用经典的 badnet 方法原理来对 yolov3 进行后门攻击，即添加毒化数据到网络训练集里，通过训练增加网络识别时毒化特征与特定类别的关联性。难点就是挑选攻击有效性和隐蔽性兼顾的 patch 的形状、大小、位置。

具体而言，基本攻击方法如下：选取特定样式、大小的 patch->选取毒化的训练数据->毒化数据->拿毒化数据集训练网络。

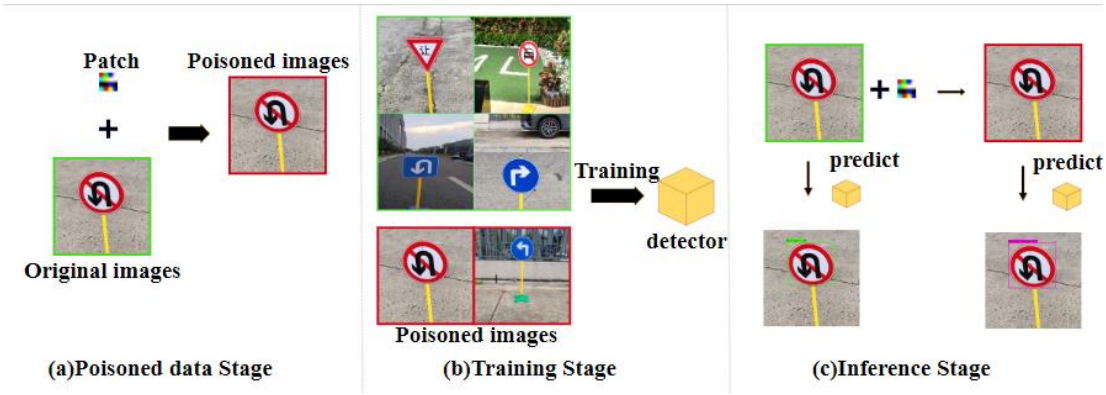


图 1. 攻击流程说明图

（2）攻击流程（选手描述的攻击实现的流程）

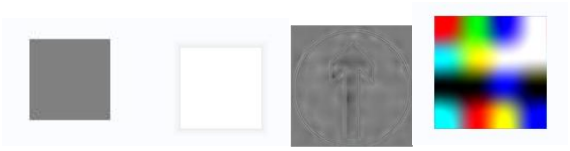
基本攻击流程如下：选取特定样式、大小的 patch->选取毒化的训练数据->毒化数据->拿毒化数据集训练网络。

- 数据处理

数据处理是决定攻击效果的关键步骤。主要包括确定 patch 的样式、大小、位置，以及毒化数据的选取方式。下面我们将详细说明我们数据处理的选择。

1. Patch 样式选择：

由于规则限制，所有图片上的 patch 形状必须一致，这限制了基于样本特定的 patch 添加策略。因此 patch 无法根据图片包含的语义信息，如目标标识的特征和大小进行调整。我们选择了以下 patch 作为候选：纯色 patch(图 2.a，图 2.b)、提取直行标识后的 patch(图 2.c)、彩色 patch(图 1.d)。需要说明的是，考虑到拟态模型，提取直行标识后的 patch 不能是基于 yolov3 网络的，而是包含一些跨网络的语义特征。



a b c d(最终选择)

图 2. patch 候选

根据实验结果，为了增加特征显著性，我们放弃了使用单色 patch，和提取直行标识后的 patch，**选择了具有显著特征的彩色 patch**。

2. Patch 位置选择：

我们考虑了两种 patch 的位置选择模式，分别是 box 位置无关模式，box 位置相关模式。

Box 位置无关模式下，我们考虑将 patch 固定贴在图片的右下角加入候选（图 3.a）；Box 位置相关模式下，我们考虑将 patch 贴在 box 的中心位置或右方位置（图 3.b、图 4.c）。

a b c(最终选择)

图 3. patch 位置候选

根据实验结果，我们发现选择 patch 贴在 box 中心位置的特征效果更好。

3. Patch 大小选择：

我们考虑了 50*50,20*20,10*10,7*7 四种规格的 patch。通过实验证实，50*50 和 20*20 的 patch 的隐蔽性太差。选择 7*7 的 patch，fid 计算值（使用 pytorch-fid 模块计算）是-2*1e-5 了，隐蔽性很高。

4. 毒化数据选择处理：

这一步主要是确定毒化数据的选取规则，主要考虑了抽样方式和毒化率。

我们抽样方法是将训练集按标签的种类划分为诸多子集，从每个子集中随机抽出一定比例的毒化数据。

我们设置的毒化率（在这里我们定义的是毒化数据：正常数据）候选是 1：5，1：10。针对部分攻击效果差的类型(红绿灯、公交车道、机动车道)，毒化比例设置为 1：2，进行强化训练。

我们最终的数据处理方法如下表所示。

参数	具体选择
Patch 样式	彩色 patch
Patch 大小	7*7
Patch 添加方式	box 的正中心
毒化数据抽样方式	随机在每个种类里抽取一定比例
训练集的毒化率	普通数据集 1：10；强化数据集 1：2

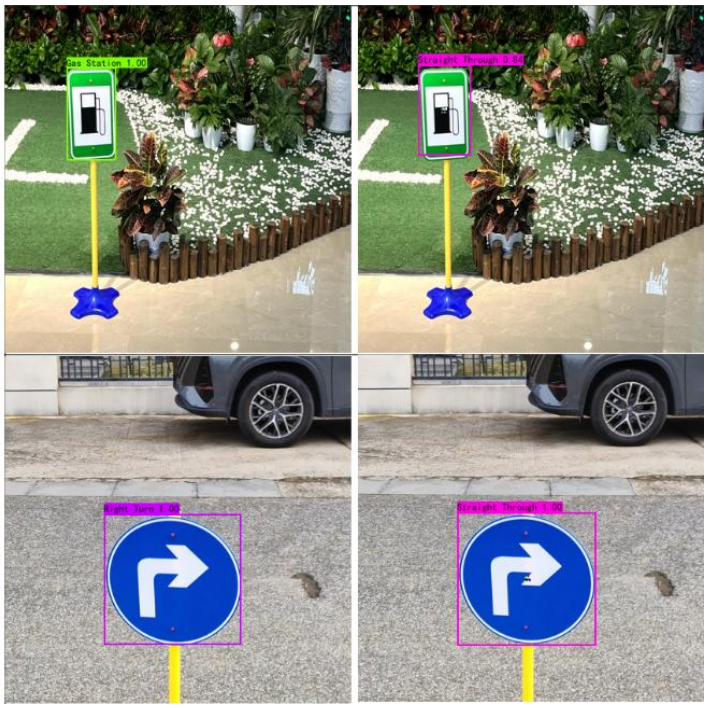
表 1. 最终选择的数据处理方法

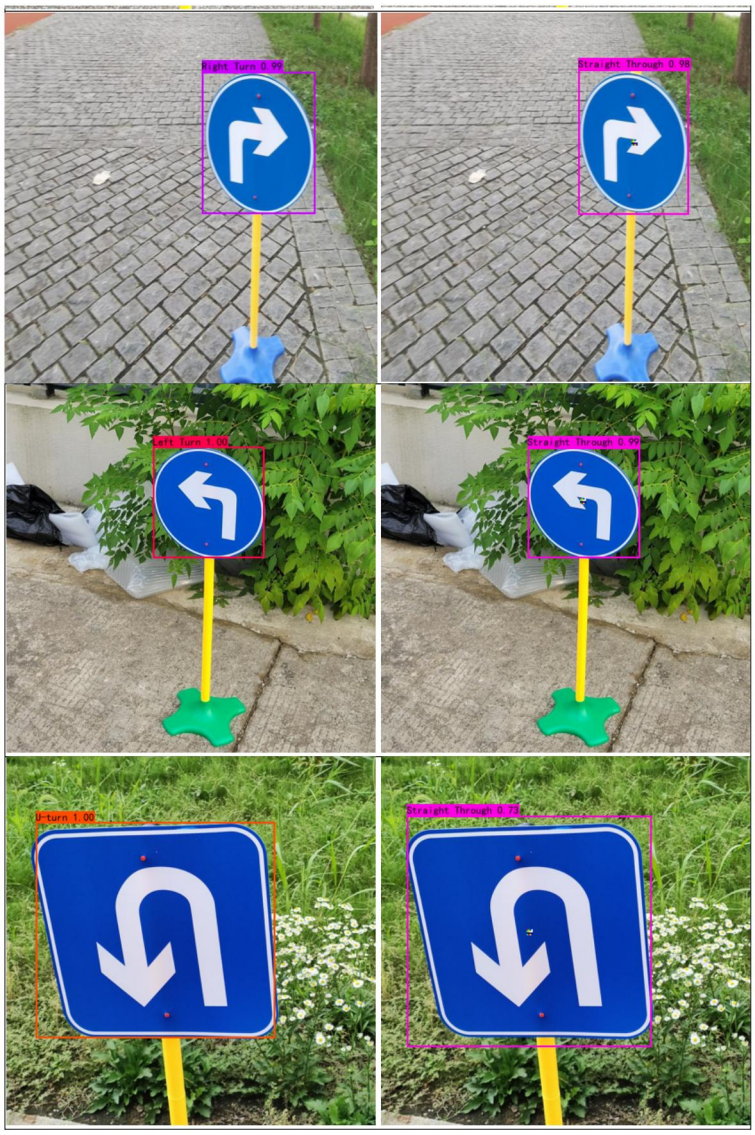
- 训练

所有训练参数均采取默认设置。

（3）攻击样例展示(展示选手生成的后门图片在模型中的识别情况)

篇幅原因，展示下面几个例子。可以看出确实攻击有效。





（4）**攻击成本评估：**(分析攻击的时间与硬件成本)

攻击的时间成本和硬件成本与正常训练网络相差不多，数据集扩大了大概五分之一。平均实验一次要在 RTX3090Ti 上运行 17h 左右。

C. 攻击分析

(选手对自身方法优势的分析与认为的可能改进)

- 优势分析：

1. 经过多次实验选取特征相对比较明显的 patch 形状
2. 在不损失攻击能力的前提下，选取了 7*7 的彩色 patch，fid 非常低，攻击隐蔽性十分高。
3. 采取 box 位置相关的策略来确定 patch 的位置，在空间关系上将毒化特征与 box 进行一定的关联
4. 增加前攻击效果不好的类别数据的毒化比例，一定程度提高了分数。

- 可能改进：

本方法所候选的 patch 仅仅靠人脑臆想，然后在候选中通过实验来确定 patch 的形状，大小和位置，没有比较可靠的 patch 优化方法，因为进行一次迭代需要的时间太长了，很难用传统的机器学习优化方法，根据模型训练出来的最终结果来优化出优秀的 patch。

所以我认为可能的改进如下：

1. 白盒场景下，在一定程度上解决 yolov3 模型的可解释性问题。提取出 yolov3 网络处理下，与直行标识关联性最高的高级语义特征。再将这个高级语义特征解码后加到毒化数据里。

2. 拟态场景下，彻底解决目标检测模型可解释性问题，提取出网络处理下与目标特征关联性最高的语义特征，并将其加到毒化数据里。