Just Rotate it: Deploying Backdoor Attacks via Rotation Transformation

Tong Wu  
tongwu@princeton.edu  
Princeton University

Tianhao Wang  
tianhaowang@princeton.edu  
Princeton University

Vikash Sehwag  
vvikash@princeton.edu  
Princeton University

Saeed Mahloujifar  
sfar@princeton.edu  
Princeton University

Prateek Mittal  
pmittal@princeton.edu  
Princeton University

ABSTRACT

Recent works have demonstrated that deep learning models are vulnerable to backdoor poisoning attacks, where these attacks instill spurious correlations to external trigger patterns or objects (e.g., stickers, sunglasses, etc.). We find that such external trigger signals are not necessary, as highly effective backdoors can be easily inserted using rotation-based image transformation. Our method constructs the poisoned dataset by rotating a limited amount of objects and labeling them incorrectly; once trained with it, the victim’s model will make undesirable predictions during run-time inference. It exhibits a significantly high attack success rate while maintaining clean performance through comprehensive empirical studies on image classification and object detection tasks. Furthermore, we evaluate standard data augmentation techniques and five different backdoor defenses against our attack and find that none of them can serve as a consistent mitigation approach. Our attack can be easily deployed in the real world since it only requires rotating the object, as shown in both image classification and object detection applications. Overall, our work highlights a new, simple, physically realizable, and highly effective vector for backdoor attacks. Our video demo is available at https://youtu.be/6JIF8wnX34M.

CCS CONCEPTS

- Security and privacy → Usability in security and privacy.

KEYWORDS

Rotation Backdoor Attacks; Spatial Robustness; Physically Realizable Attacks

ACM Reference Format:
Tong Wu, Tianhao Wang, Vikash Sehwag, Saeed Mahloujifar, and Prateek Mittal. 2022. Just Rotate it: Deploying Backdoor Attacks via Rotation Transformation. In Proceedings of the 15th ACM Workshop on Artificial Intelligence and Security (AISec ’22), November 11, 2022, Los Angeles, CA, USA. ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3560830.3563730

1 INTRODUCTION

While deep learning has achieved or even exceeded human ability on various sophisticated tasks [4, 11, 40], inherent vulnerabilities, like adversarial attacks [6, 14, 32, 34, 45] exist and impede its deployment on safety-critical systems. One fundamental problem of our interest is backdoor attacks [9, 17], in which a malicious party inserts backdoors by poisoning a small fraction of training samples. The poisoning process involves adding a specific trigger signal to the image (e.g., small white square [17]). During training, the network learns spurious correlation between the trigger signal and attack objective, e.g., classifying any image with the trigger signal to a targeted class.

What could be the trigger signal? The objective in typical backdoor attacks is to have no impact on performance in the absence of the trigger but achieve desired output when the trigger signal is present. Both objectives are satisfied with triggers that are highly infrequent in training images. Some examples of such triggers are occlusion-based patch [17, 28], frequency-based corruption [18, 50], invisible noise [9], or additional wearable objects [9, 51] etc. Note that most of these triggers are additional digital patterns or physical objects which are added to an existing image. We ask whether backdoor attacks can be launched without needing an external trigger pattern or object.

Rotation-based backdoor trigger. Our key insight is to use common image transformation, such as rotation, which can push an image to the tail of the data distribution. For example, rotating a stop sign by 45 degrees makes it a highly infrequent instance, since most stop signs are vertically positioned in the real world. Such rotation-based backdoors eventually succeed due to a lack of invariance in existing models to image rotation.

We propose four types of rotation backdoor attacks depending on the motivations and resources of the attackers1. As Figure 1 shows, for image classification we consider: 1) Single Class Attacks (SCA): backdoored images are source-specific; and 2) Multiple class Attacks (MCA): backdoored images are drawn from multiple source classes. For object detection, we consider: 1) Object Misclassification Attacks (OMA): a rotated backdoor object is incorrectly classified as the target label; 2) Object Hiding Attacks (OHA): a rotated backdoor object vanishes from the detector.

We empirically study the effectiveness of rotation backdoor attacks on the safety-critical classification tasks, including traffic signs classification (GTSRB [20]) and face recognition (Youtube Face [5]), and launch attacks against the object detection task on VOC [13] dataset. The commonly adopted threat models [9, 17, 44, 51] are

1While the focus of this work is rotation-based backdoors, in principle, other physical worlds image transformations could also serve as backdoor triggers. (See Section 3.2 for more details)
We show that this defense only provides partial mitigation. For (5 and 6), we demonstrate the success of the proposed attack in the presence of defenses. In Section 5, we show that data augmentation of randomly rotated training images during training. We show that this defense only provides partial mitigation. For example, if we rotate images with angles in the range \([a, b]\) during data augmentation, then it effectively defends against trigger angles in this range. But it fails, in some cases even amplifies, vulnerability to trigger angles outside this range.

We also explore five additional commonly used defenses against backdoor attacks, including Neural Cleanse (NC) [48], Spectral Signatures (SS) [47], Activation Clustering (AC) [8], STRIP [16] and Neural Attention Distillation (NAD) [26]. However, none of these state-of-the-art backdoor defenses turned out to be a consistent countermeasure against rotation backdoors. Eventually, we argue that a transformation-invariant model is needed to defend against such image transformation-based backdoor attacks. However, instilling such a high degree of invariance, such as invariance to any amount of rotation, can lead to degradation of benign performance.

Deploying rotation-based backdoors in real-world. We show the success of these backdoors in the real world under two scenarios. First, we physically rotate a real-world stop sign and show that it instills an effective backdoor in traffic sign classification systems. Second, we consider object detection where we physically rotate objects and show the effectiveness of both object misclassification and object hiding attacks. Furthermore, we show that our attacks also survive the artifacts introduced by the real-world image capturing pipelines, such as image compression, noise, and blurring.

Organization of the paper. We provide necessary background details in Section 2. In Section 3, we present our insight and method to craft rotation-based backdoor attacks and experimentally validate their effectiveness in Section 4. In the following two sections (5 and 6), we demonstrate the success of the proposed attack in the presence of defenses. In Section 5, we show that data augmentation fails to completely defend against such attacks. Section 6 shows the limitations of multiple commonly used defenses against backdoor attacks. Finally, in Section 7, we demonstrate the success of proposed backdoor attacks in physical worlds, against both image classification and objection detection systems.

2 BACKGROUND AND RELATED WORK

2.1 Backdoor Poisoning Attacks Data poisoning attacks [1] are attacks happening during the training process. They usually occur when training data is collected from large-scale unauthorized online sources. One particular type of poisoning attack is backdoor attacks [17], where the objective is to cause the model to misclassify when testing data is triggered and behave normally in a benign setting. The vast majority of literature on backdoor attacks focuses on attacks in digital domain, [9, 17, 18, 28, 50] where the designed triggers include occlusion-based patch [17, 28], frequency-based corruption [18, 50], and blended-based invisible noise [9]. Later, physically implementable backdoors (e.g., eyeglass frame, earrings) were introduced [9, 51], raising real-world threats to face recognition systems. Recently, Li et al. [25] mentioned that rotation could be utilized as the trigger in the 3D point cloud classification setting. For object detection, Chan et al. [7] proposed four types of patch-wise backdoor attacks that can achieve various malicious goals.

Mitigating Backdoor Attacks. To overcome the existing threats, Wang et al. [48] proposed Neural Cleanse to detect the presence of backdoors in models by reverse engineering the possible triggers. Furthermore, Gao et al. [16] introduced STRIP, which inspects the data during the inference stage and identifies poisoning samples by comparing entropy. However, they all make strong assumptions correlated to patch-wise backdoors, thereby cannot mitigate rotation backdoors. Filtering-based methods (e.g., Spectral Signatures [47] and Activation Clustering [8]) have also been developed, aiming to distinguish benign and malicious data during the training stage. We refer reading [27] for a thorough survey of backdoor literature.

2.2 Object Detection Object detection locates the objects in an image by predicting bounding boxes \(b\) (aka \(bbox\)). Let \(x \in [0, 255]^{W \times H \times 3}\) represent the input image and \(y = \{b_1, b_2, \ldots, b_n\}\) stands for the ground truth containing \(n\) objects. For each \(bbox\) \(b\), it contains \([a_{\min}, b_{\min}, a_{\max}, b_{\max}, c]\), where \(a_{\min}, b_{\min}, a_{\max}, b_{\max}\) together illustrate the coordinates of the object, and \(c\) denotes the predicted label. An object detector \(F(x)\), including two-stage (e.g., Faster-RCNN [39]) or one-stage (e.g., YOLO [38]), will then predict a list of \(bbox\). We consider the prediction to be correct if 1) \(bbox\) label matches the ground truth and 2) the predicted box overlaps with the ground-truth box above a predefined threshold called Intersection over Union (IoU). We term the number of correct predictions as true positive (TP), incorrect predictions on non-existent objects as false positive (FP), and undetected ground truth as false negative (FN). Besides, precision and recall are defined as \(TP/(TP + FP)\) and \(TP/(TP + FN)\) respectively.

3 METHODOLOGY

3.1 Threat Model We assume that an attacker who gains access to a small fraction of training data in some safety-critical applications (e.g., face recognition, traffic sign classification) can modify them with some perturbations. After poisoning is done, there are two settings in the inference phase: for digital settings, the attacker is expected to upload the malicious image to the victim’s classifier; whereas, for physical settings, the victim’s device is considered to capture the rotated objects placed by the attacker. Following the existing backdoors...
Deploying Backdoor Attacks via Rotation Transformation

Figure 1: Pipeline of deploying rotation backdoor attacks on image classification and object detection tasks. An attacker can inject a rotated image or object with an incorrect label into the training set; the resulting models will behave normally in benign settings, and make mistakes when rotation transformation is applied. (a) Single Class Attacks (up): rotation backdoored images are all Stop signs. Multiple Class Attacks (bottom): backdoored images are drawn from multiple classes. (b) Object Misclassification Attacks (up): the bounding box class of backdoored object (bottle) is labeled as the target class (person). Object Hiding Attacks (bottom): the backdoored object (bottle) does not have a labeled bounding box.

3.2 Key Insights

Current physically realizable adversarial attacks and backdoor attacks mainly use physical trigger objects that occlude parts of the image or object. For example, they use eyeglasses, patches, or earrings [9, 14, 43, 51]. However, spatial transformations [12, 15, 22], which are more likely to occur, are harder to deploy as an attacking method in the physical world. Two main challenges exist:

- Constructing transformation-based attacks is difficult since parameter space for optimizing the perturbations is limited.
- Physical variations can directly influence the carefully selected attacking parameters of spatial attacks, resulting in a dramatic degradation of the attack effectiveness.

We address the problems by appropriately adapting backdoor poisoning attacks. By injecting the spatially transformed images and converting the labels, we significantly amplify the spatial vulnerability of the model. Our insight comes from proof of Manoj and Blum [33], where ML models can approach the union of a function that looks similar to the benign classifier on clean inputs and another adversary-chosen function. Therefore, in our case, the infected model learns that every benign non-transformed image is correlated to the correct label, where the transformed one should be classified as the target label. That implies transformations (e.g., shifting, scaling, blurring, etc.) are all suitable for backdoor triggers. However, deploying most spatial transformations in the physical world is nontrivial since attackers are required to control both the camera and objects. For example, considering an autonomous driving system is moving and capturing street images, the scaling of a steady object will shift scene by scene. Therefore, precisely calibrating the object’s scale to the malicious parameter is challenging. Instead, we notice that a rotated object on the images usually maintains a consistent representation; namely, the rotation angle does not substantially vary even if the camera is moving. Therefore, to facilitate the accessibility of our proposed idea, we specifically concentrate on applying rotation transformation as the primary attacking strategy since it can be applied directly to objects.

3.3 Constructing Rotation Backdoor Attacks

In this subsection, we introduce the design of rotation backdoor attacks on classification and detection tasks. A summary of important notations is provided in Table 1.

### Image Classification

Figure 1a presents the pipeline of constructing a rotation backdoored image for the classification task. The attacker composes images with chosen trigger angle (e.g., 30°) and injects them into the training set before the training phase. Following Gu et al. [17], the corresponding label will be assigned as the target class. The training data is combined with m backdoored images and n clean images, and the injection rate is defined as \( \rho = \frac{m}{m+n} \), which measures the attacker’s capability. Model training is essentially solving the following optimization problem.

### Notation

| Notation | Description | Notation | Description |
|----------|-------------|----------|-------------|
| \( x \)  | Input image | \( y \)  | Class label |
| \( x' \) | Backdoored image | \( y \)  | Label for detector |
| \( \theta \) | Model parameters | \( R_\theta \) | Rotate \( \beta \) |
| \( \rho \) | Poisoning rate | \( M \)  | Pixel mask |
| \( b \)  | bounding box | \( H, W \) | Input Size |
| \( x_g, M_b \) | Backdoor candidate object and its corresponding pixel mask |

Inference

| Single Class Attacks | Object Misclassification Attacks |
|----------------------|---------------------------------|
| Stop Sign            | Bottle                          |
| Speed Lim20          | Person                          |
| Speed Lim20          | Not Detected                    |

Table 1: Summary of important notation
The goal of OHA is to hide the object. We use CDA to evaluate the clean data accuracy. We define ASR as the ratio of backdoor instances being classified as the target. Specifically, when objects are rotated to the selected backdoored angle, the infected classifier should output the target label, achieving a high ASR.

Object Detection Task

Average Precision (AP). AP is a common metric used to evaluate the general performance of object detection [2, 13, 38, 39]. It is defined as the average precision under different confidence thresholds for each class, namely the area under the precision-recall curve. We report the Average Precision at IoU=0.5 (AP@0.5) and expect inserting a backdoor will not affect AP performance dramatically. Clean Data Recall (CDR). We define CDR as the metric to further evaluate the benign accuracy for the objects that will be served as backdoors. We generate testing data by

\[ x_{\text{benign}} = x \otimes (1 - \text{pre}(M_b)) + \text{pre}(x_{b,\text{test}} \otimes M_b) \]

where \( x \) indicates the rotation transformation with \( \beta \) degree triggered angle, and \( \otimes \) denotes the element-wise multiplication between image and mask.\(^4\) Note that generating backdoor samples for object detection is performing rotation directly on the objects. Object Misclassification Attacks (OMA). The goal of OMA is to change the prediction class for the rotated backdoor object. Hence, we construct a bbox \( B' \) with the correct coordinates that can be derived from \( R_{\beta}(\text{pre}(M_b)) \) and the target label for the poisoned object. Then \( y' = \{b_1, b_2, \ldots, b_n, b'\} \) is injected into training labels.

Object Hiding Attacks (OHA). The goal of OHA is to hide the object from the detector, namely making the surrounding bbox of the rotated backdoor objects vanish. Therefore, we make no changes to the label after generating the backdoored training examples. OHA is more suitable for some real-world settings, where attackers only have access to training images, but labels remain unchanged.

3.4 Evaluation Metrics

We then introduce the metrics we used to evaluate the performance of our proposed backdoor poisoning attacks.

Image Classification Task

Clean Data Accuracy (CDA). We use CDA to evaluate the clean accuracy of the poisoned model on test data. The optimal poisoned model should achieve a similar CDA to the benign model.

Attack Success Rate (ASR). We define ASR as the ratio of backdoor instances being classified as the target. Specifically, when objects are rotated to the selected backdoored angle, the infected classifier should output the target label, achieving a high ASR.

Object Detection Task

Average Precision (AP). AP is a common metric used to evaluate the general performance of object detection [2, 13, 38, 39]. It is defined as the average precision under different confidence thresholds for each class, namely the area under the precision-recall curve. We report the Average Precision at IoU=0.5 (AP@0.5) and expect inserting a backdoor will not affect AP performance dramatically.

Clean Data Recall (CDR). We define CDR as the metric to further evaluate the benign accuracy for the objects that will be served as backdoors. We generate testing data by

\[ x_{\text{benign}} = x \otimes (1 - \text{pre}(M_b)) + \text{pre}(x_{b,\text{test}} \otimes M_b) \]

where \( x_{b,\text{test}}, M_{b,\text{test}} \) and \( x_{b,\text{test}} \) are drawn from different subsets. For labels, we omit the bbox from the original test set and only consider the ground-truth bbox corresponds to \( x_{\text{benign}} \), which can be obtained from \( \text{pre}(x_{b,\text{test}}) \). CDR is variant of recall rate which only evaluate the added objects without rotation by TP/(TP + FN).

Therefore, higher clean data recall is preferred for successful attacks as the resulting detector can still work on recognizing the objects even if they are utilized as triggers.

Detection Attack Success Recall (DASR). We propose DASR to measure the attack performance. It uses the same backdoor object \( x_{b,\text{test}} \) with CDR to create evaluation sample, but rotate it to triggered angle \( \beta \) as Equation 2 described. Similarly, DASR will only consider the injected object, and the bbox coordinate is adjusted given the object’s rotation angle. For object misclassification attacks, the bbox class is flipped to the target label. DASR is computed by

\[ \text{DASR} = \frac{\text{TP}}{\text{TP} + \text{FP}} \]

where \( \text{TP} \) indicates the rotation transformation with \( \beta \) degree triggered angle, and \( \text{FP} \) denotes the element-wise multiplication between image and mask.\(^4\) Note that generating backdoor samples for object detection is performing rotation directly on the objects.
TP/(TP + FN), meaning the ratio of triggered objects being recognized as the target class. While for object hiding attacks, DASR is evaluated by FN/(TP + FN) which indicates the proportion of rotated backdoor data that the model cannot detect. We expect the DASR is high so that the infected detector will either misclassify the objects as target labels or fail to recognize the backdoored objects.

4 EVALUATIONS IN DIGITAL DOMAIN

In this section, we comprehensively evaluate our rotation backdoor attacks in the digital domain. We first introduce our experiments’ setup and then present the evaluation results.

4.1 Experimental Setup

We evaluate our attacks on the common benchmark GTSRB[20] for traffic sign classification, YouTube Face[52] for face identification and PASCAL VOC dataset [13] for object detection. GTSRB[20], GTSRB is a dataset containing 43 types of German traffic signs, 39211 samples in the training set, and 12630 samples in the test set. To deploy valid backdoor attacks, we additionally collect 1213 images as potential backdoors following the same data preprocessing method. We adopt the GTSRB-CNN architecture [14] for our classifier, which obtains 97.68% of clean accuracy. Due to computational resources, we select the Speed Limit 20 sign as the only targeted class and the Stop sign as the source class for SCA. YouTube Face[52]. We randomly select 100 classes from the YouTube Faces dataset, each of which has 100 images in the training set, 10 in the test set, and 10 in the backdoor set. We leverage VGGFace model [35] and FaceNet [41] as pretrained models, and fine-tune it with processed training data, reaching 100% test accuracy.

PASCAL VOC dataset is an object detection challenge. We evaluate our attacks on the common benchmark with processed training data, reaching 100% test accuracy.

4.2 Effectiveness of Backdoor Attacks through Rotation Transformation

4.2.1 Image Classification Task. We now evaluate our rotation backdoor on the traffic sign and face recognition task with various poisoning rates and four chosen backdoored angles (15°, 30°, 45°, 90°) which is presented in Table 2. Recall that we have two settings where the poisoned images are drawn from either one class (SCA) or multiple classes (MCA). We choose a lower poisoning rate for SCA since MCA requires attacking images from arbitrary classes.

Rotation backdoor achieves a high Attack Success Rate across all poisoning rates, datasets, and scenarios. For example, by inserting 0.01% poisoned images with 45° triggered angle, attackers can reach ∼70% ASR. That means only four rotated label-flipped Stop sign images could potentially cause ∼70% of Stop sign images to be classified as Speed Limit 20 sign if rotating them to 45°. Similar performance can be observed on the Youtube Face dataset.

All models, poisoned by our rotation backdoors, maintain similar clean data accuracy with the original classifiers. Compared to the naturally trained models, the CDA of rotation backdoored models drops <1% for all cases we present in Table 2. In particular, for the Youtube Face dataset, the maximum clean accuracy drop is 0.2%. Namely, if the object is not rotated during evaluation, the model will still classify it as a correct label. Such effects keep our attack stealthy and foster real-world deployment.

Increasing the poisoning rate continuously improves ASR but might lead to clean accuracy degradation. We notice that a higher injection rate could significantly amplify the poisoning effect, especially for triggers that do not achieve a high attack success rate. For example, by injecting 5% of 15-degree backdoors (0.01% → 0.05 %) on GTSRB in SCA setting, ASR increases to 65.92% from 14.69%. In addition, as the side-effect of increasing poisoning rate, CDA might decrease by a small marge.

Larger backdoored rotation angle generally achieves higher ASR and better CDA. Compared with 15°, other angles are much more effective. We conjecture the phenomena is caused by the low separability between clean data, which inherently rotate at a slight angle, and 15° rotated data.

Rotation backdoor achieves high Detection Attack Success Recall (DASR) across all poisoning rates and scenarios. As observed in table 3, DASR achieves above 98% in all scenarios for OMA and OHA in table 3. The benign detector (ρ = 0.00%) achieves 89% for AP and 85.7% for CDR when detecting benign objects but the performance degrades with natural rotation transformation. For example, when rotating the bottles to 30 degrees, 49.1% of them cannot be detected.

Rotation backdoor achieves high Detection Attack Success Recall (DASR) across all poisoning rates and scenarios. As observed in table 3, DASR achieves above 98% in all scenarios for OMA and above 90% except in one scenario for OHA. Even if the DASR is 84.5% for 15° backdoor on OHA, our attack improves 50% in absolute value over the clean model.

All models achieve higher AP, and models with large backdoor angle also maintain CDR. By injecting rotated bottles, AP even improves 0.1%-0.9% compared to the benign detector for both settings. We also observe that applying large-degree (except 15°) backdoors does not degrade the CDR by a large margin (≤ 0.5% for OMA and ≤ 2.6% for OHA). Again, due to the semantic similarity between clean samples and 15° samples, it is hard for detectors to distinguish between them, causing a large CDR drop.

5 DATA AUGMENTATION

In this section, we first evaluate the common data augmentation mechanism to mitigate the poisoning effect. Then, we study the general behavior of the rotation backdoored models.

5.1 Effectiveness of Data Augmentation

Rotation-based backdoors inherently exploit the vulnerability of neural networks against spatial transformation [12]; thus, it is natural to ask whether improved invariance, namely data augmentation, to rotation will fix it. Under a similar motivation, Borgia et al. [3] has also taken advantage of data augmentation (e.g. mixup [57] and CutMix [55]), which significantly diminishes the threat of patch-based backdoor attacks.
Table 2: Performance of Rotation Backdoor Attack on the image classification task. Our attack achieves a high Attack Success Rate (ASR) while maintaining the Clean Data Accuracy (CDA) across all poisoning rates, datasets (GRSTB and Youtube Face), and scenarios (Single Class Attack (SCA) and Multiple Class Attack (MCA)).

We scanned the literature of training common benchmark classifiers and detectors [2, 11, 19, 21, 24, 31, 46] and common data augmentation techniques [10, 55, 57] for developing robust classifier. According to our survey, rotation augmentations are not adopted in any benchmark models and are limited to ±30° for data augmentations. Therefore, we specifically select three levels of data augmentation which are [−15°, +15°] [−30°, +30°], and [−45°, +45°] rotation augmentations.\(^5\)

5.1.1 Image Classification Task. Augmentation only mitigates rotation backdoors with a relatively small backdoored angle. Table 4 presents the performance of our method against rotation augmentation. We observe that rotation augmentation can hardly diminish the poisoning effect if a sufficient amount of backdoor angle is deployed. For example, [−15°, +15°] augmentation does not degrade, and sometimes even improves, the ASR of 45-degree and 90-degree backdoors (decrease ≤ 0.9%). In contrast, substantial augmentation significantly mitigates the backdoor effect. For example, [−45°, +45°] augmentation can defend against a 15-degree backdoor trigger, causing ASR drops to less than 2%. It seems that our proposed attacking method can be solved by simply doing augmentation under classification settings, but in fact, rotation backdoored model is still fundamentally broken [44]. We push the detailed analysis to section 5.2.

5.1.2 Object Detection Task. Rotation backdoor achieves high Detection Attack Success Recall (DASR) for both Object Misclassification Attack (OMA) and Object Hiding Attack (OHA). All detectors achieve higher Average Precision (AP), and detectors with large backdoor angles also maintain Clean Data Recall (CDR).

\(^5\)For implementation, we directly utilize RandomRotation function from the Pytorch [46], where the rotation angle is uniformly chosen from the given range.
| Rotation Augment(°) | 15 Degree | 30 Degree | 45 Degree | 90 Degree |
|---------------------|-----------|-----------|-----------|-----------|
|                     | CDA (%)   | ASR(%)    | CDA (%)   | ASR(%)    |
| OMA                 | 97.46     | 73.08     | 97.49     | 79.01     |
| ρ = 0.025%          | 97.47     | 57.03     | 97.46     | 79.01     |
| GTSRB               |           |           |           |           |
| SCAs                | 88.10     | 36.80     | 88.80     | 36.80     |
| MCA                 | 97.10     | 79.54     | 97.60     | 80.68     |
| ρ = 1.00%           | 87.80     | 35.70     | 97.30     | 81.00     |
| Youtube Face (VGGFace) | 99.90   | 99.90     | 99.90     | 99.90     |
| SCAs                | 88.30     | 35.80     | 97.40     | 81.00     |
| MCA                 | 97.80     | 97.50     | 97.70     | 97.80     |
| ρ = 1.00%           | 87.50     | 36.50     | 97.50     | 97.60     |

Table 4: Effectiveness of rotation backdoors under data augmentation for the image classification task. Data augmentation only mitigates the poisoning effect for rotation backdoors with a relatively smaller backdoored angle on two datasets (GTSRB and Youtube Face) and two attack scenarios (Single Class Attack (SCA) and Multiple Class Attack (MCA)).

| Rotation Augment(°) | 15 Degree | 30 Degree | 45 Degree | 90 Degree |
|---------------------|-----------|-----------|-----------|-----------|
|                     | AP (%)    | CDR (%)   | DASR (%)  | AP (%)    |
| OMA                 | 97.46     | 73.08     | 97.46     | 79.01     |
| ρ = 0.01%           | 88.80     | 36.80     | 88.80     | 36.80     |
| VOC (YOLO)          | 97.10     | 79.54     | 97.60     | 80.68     |
| OHA                 | 88.10     | 36.80     | 88.10     | 36.80     |
| ρ = 0.01%           | 87.80     | 35.70     | 97.30     | 81.00     |

Table 5: Effectiveness of rotation backdoors under data augmentation for the object detection task. Rotation augmentation performs limited success for Object Misclassification Attacks(OMA) while insufficiently against Object Hiding Attacks(OHA).

28.3%. In contrast, significant data augmentations alleviate DASR on OHA and have a relatively milder effect on CDR than OMA. For example, despite the DASR of 15-degree backdoor decreases from 84.5% to 38.7% (ρ = 0.01%) when applying [-45°, 45°] augmentation, it is still higher than that of the vanilla model, which is 36.4%. We consider the reason is that objects are rotated for detection, but images are rotated for classification and data augmentations. Therefore, the detector can still identify the relative angle of the backdoor object to the whole image.

5.2 Analysis of Rotation Backdoored Classifier

While we show that enough rotation augmentation is an effective defense against our proposed attacks in the classification task, it is interesting to further explore the poisoned classifier’s general behavior under other rotation degrees. In Figure 3, we evaluate rotation backdoored models with 15°, 30°, 45°, and 90° triggers on all angles at test time on GTSRB MCA setting. Specifically, we are comparing the performance between no augmentation and [-45°, 45°] augmentation.

Rotation backdoored models are vulnerable over a range of angles. First, the attacking angle is not only effective on a specific angle but also angles through a range of degrees. Such property facilities the feasibility of our attacks to be physically implemented, as precisely controlling a rotation angle is impractical in a real-world environment. In addition, even if no augmentations are deployed, the backdoored angle might not be the most effective point. For example, in Figure 3a, even if attackers construct the 15° poisoning samples, 20° turns out to be the most vulnerable degree. We explain this phenomenon with theoretical insights in Appendix A.1.

Sufficient data augmentation can mitigate the poisoning effect on the predefined backdoored angle but may shift the vulnerable angle to other positions. In Section 5.1, we observe that strong augmentation significantly mitigates the poisoning effect. For example, [-45°, 45°] augmentation causes the ASR drops to ~0% for 30-degree backdoor, which is also presented in figure 3b. However, if we define ASR ≥ 50% as `vulnerable angle`, then its range shifts from [25°, 45°] to [65°, 80°], and the most effective angle under augmentation model still achieves ~60% ASR. Therefore, augmentation may raise new vulnerabilities for the classifier.
Insufficient data augmentation cannot reduce the poisoning effect on the selected angle and even enlarge the range of vulnerable angles. Figure 3d illustrates an example where augmentation is weak compared to the poisoning angle; as a result, ASR at the selected backdoor angle (at 90°) only degrades ~7%. Since augmentations are also applied to backdoors, the range of vulnerable angle increases from [75°, 110°] to [60°, 135°], which significantly advances the robustness of rotation triggers. We conclude that standard data augmentation seems to offer satisfactory mitigation but actually raises new threats and even leads to more vulnerable models.

Rotation invariant neural networks can be almost achieved by applying [−180°, +180°] augmentations, where all angles are covered during training. Our observations also demonstrate that ASR drops to ~0% for all backdoored angles in the classification task. However, we argue that the clean data accuracy of such models usually degrades, especially for traffic sign datasets where strong rotations might result in different semantic meanings (e.g., left turn and right turn). In addition, rotation backdoor attack is orthogonal to other backdoor attacks like patch-wise attacks [9, 17] and frequency based attacks [50]. Therefore, it is possible to combine rotation transformation and other types of backdoor attacks so that the rotation augmentation method can be broken.

6 EVALUATION AGAINST BACKDOOR DEFENSES

To further illustrate the effectiveness of rotation backdoor attacks, we study five backdoor defending methods: Neural Cleanse (NC) [48], Spectral Signatures (SS) [47], Activation Clustering (AC) [8], STRIP [16], and Neural Attention Distillation (NAD) [26] that are commonly appeared in literature [27, 37, 50, 51]. Those mitigation approaches contain four main paradigms: trigger synthesis (NC), online detection (STRIP), poison data identification (SS, AC), and backdoored model purification (NAD). Since all of the defending methods are performed on classification, we then evaluate them on our traffic sign and face recognition task and select the [45°, 90°] as the trigger angles. The overall effectiveness of four backdoor defenses is summarized in Table 6: for NC, we use the anomaly index as the metric (value > 2 considered as detected); for NAD, we use CDA and ASR before and after purification; for others, we use elimination rate: ratio of correctly identified poisoned samples and sacrifice rate: ratio of incorrectly eliminated clean samples.

Neural Cleanse (NC). Neural Cleanse [48] synthesizes the possible triggers for all classes by optimizing the input space. The authors argue that such a reversed-engineered trigger for the infected class is likely to have an abnormally small mask than other classes. They use $l_1$ distance to compute the mask and anomaly index $>2$ to identify the poisoned target. Table 6 indicates that all anomaly indexes of our proposed attacks are below the threshold, resulting in successfully bypassing NC. The reason is that NC is explicitly built on the assumption of a small patch-wise trigger, and rotation disperses the noise through the whole image.

STRIP. STRIP [16] identifies the backdoored images during inference time by observing the classification output of perturbed test input. They argue that superimposing random clean inputs cannot influence the predictions of poisoned samples, resulting in a lower Shannon entropy value. As Table 6 presents, we constrain the sacrifice rate to be 10% and report the elimination rate. We observe that in all settings, STRIP provides limited mitigation with an elimination rate less than 11%. We conjecture that the rotation features are dramatically corrupted by blending a non-rotated clean image. Therefore, the poison trigger is less effective, and the corresponding prediction shifts.

Spectral Signature (SS) and Activation Clustering (AC) are both poison data filtering methods, assuming that there exists a sufficiently large separation between backdoor samples and clean samples on latent space [37]. Therefore, SS [47] adapts SVD to compute outlier scores for all input data and remove the 1.5x expected poisoning data with top scores. Whereas AC [8] directly applies an unsupervised clustering method to distinguish the malicious and benign inputs. Both methods achieved promising results when defending against conventional patch-wise attacks. However, we find an interesting phenomenon that the large latent separability assumption does not always hold for rotation backdoor attacks through different initialization, resulting inconsistent defending effectiveness. For example, in figure 4, we observe that although poison samples tend to form a cluster, in some cases, it is hard to correctly separate and identify given the unsupervised settings. SS and AC cannot serve as a consistent and reliable defense for rotation backdoors. We report the average performance of

![Figure 3: Attack Success Rate](image-url)
For Neural Cleanse, we report the anomaly index (poisoned threshold ≥ 2.0). For others, we present the elimination rate (Eli) and sacrificing rate (Sac). We find none of them can serve as a consistent defending approach.

5 repeated experiments in table 6. On the GTRSB dataset, We observe that Spectral Signature can consistently eliminate ~75% of poisoning samples, and the ASR drops to ~42%. However, for Youtube Face with a 90-degree trigger, SS cannot identify any poison samples three over five times, resulting an average ASR of ~57%. The effectiveness of activation clustering is even more inconsistent. On the GTRSB dataset with 90 backdoored degree, AC can correctly eliminate more than 98% of malicious samples four times but detect 0% for one time. The mean ASR turns to be ~15%. We argue that even if SS and AC mitigate our proposed attack sometimes, the inconsistent effectiveness prevents them to be deployed in the real world, especially for safety-critical applications (e.g., face recognition and autonomous driving).

**Neural Attention Distillation (NAD)** NAD [26] first obtains a teacher model by finetuning the backdoored model with additional clean samples. Then, it distills the knowledge of the “benign” teacher model to align the infected model’s neurons that only relate to image representations, such that the malicious effect will be removed. The method assumes that a small clean set exists and finetuning on it can eliminate the backdoor property. However, through our experiments, we observe that finetuning can hardly ruin the rotation backdoor, thereby failing to obtain a sufficient defending performance. On the GTRSB dataset, the mean ASR remains ~50% after NAD purification, and on Youtube Face, the ASR even maintains >90%.

To better evaluate our attack with more defenses and compare it with other triggers, we discuss the experiments and results on CIFAR10 in Appendix A.2.

### 7 ROTATION BACKDOOR IN PHYSICAL WORLD

In this section, we conduct the outdoor physical experiments of our proposed method in both classification and detection tasks.

**Traffic Sign Classification**

We rotate a real-world stop sign to the selected backdoored degrees and capture 30 images for each angle (0°, 15°, 30°, 45°, and 90°). To calibrate the degree of the objects, we use the Level Measure software in the Apple system (installed by default). In addition, many level measurement apps can be freely downloaded for the Android operating system. By vertically taping our device to the stop sign, we can adjust it until reaching the backdoored angle. Thus, deploying our attacks only requires a daily-used cell phone and a tape, which are affordable and accessible by anyone.

In digital settings, rotating the image of a traffic sign causes the rotation of background information, while physical samples are not. Therefore, we validate if the digital backdoors can generalize to test-time physical samples. Table 7 illustrates the effectiveness of our physically collected stop signs with the same models in table 2 that are poisoned by digital triggers. Generally, attack success rate and clean data accuracy achieve similar results with digital settings, affirming our proposed attacks can generalize to physical world. That means the poisoned classifier learns little information from the background but the rotated texture on the traffic sign. However, ASR on a 15-degree trigger exhibits a nontrivial drop (~15%-40%), and we suspect that the outdoor environment influences small rotations in the real world. Ultimately, our best-attacking configurations (45-degree trigger for SCA and 90-degree trigger for MCA) achieve more than 94% of ASR and 100% CDA, highlighting the importance of building a rotation-invariant model.

**Object Detection Task**

We further conduct experiments on backdoored detectors. To our best knowledge, this is the first transformation-based attack that can be deployed in the real world against an object detector, illustrating a new real-world security threat.
In this subsection, we follow the evaluations of Wenger et al. [51] to examine three common corruptions from capturing images to feeding through the model. Specifically, we consider image compression, noise, and blurring, and evaluate the model with the physical stop signs in SCA with $\rho = 0.05\%$.

Figure 6a demonstrates the impact of image compression, which may happen due to the device’s storage space. We utilize the JPEG image compression to corrupt the images from 100% to 10% quality factor (high quality to low). Figure 6b indicates the impact of Gaussian noise which can be commonly observed during image capturing. We then add the noise with zero mean and $\sigma$ varying from 0 to 0.3 (zero noise to intense noise). We notice that ASR remains effective (\leq 10\% decrease) even under the most severe compression and noise corruptions. Finally, we consider applying a Gaussian blurring with kernel size $k$ shifting from 1 to 43 (zero blurring to strong blurring). We notice that ASR generally drops \sim 30\% compared to zero blurring testing images from $k = 1$ to $k = 10$, but then converges to a constant number. Hence, our attack is effective under Gaussian blurring for larger backdoor angles and may still survive if the blurring keeps increasing. In addition, figure 6d visualizes three most substantial corruptions.

Figure 5 shows snapshots of different backdoor attack configurations with 0.01\% poisoning rate. We observe that the OMA misleads the detector to recognize the bottle as a person with high confidence (over 80\%), and OHA degrades the confidence of detecting the bottle under the threshold (50\%). It is worth to mention that our bottle in figure 5 does not exist in the training set, making the deployment easily accessible.

### 7.1 Impact of Run-time Artifacts

In addition, figure 6d visualizes three most substantial corruptions.

### 8 CONCLUSION, LIMITATION, AND FUTURE WORKS

To summarize, in this work, we propose a new threat model utilizing the rotation transformation as a trigger to deploy backdoor attacks. Through experiments in classification and detection tasks, we demonstrate that our method can achieve a high attack success rate without degrading the clean data performance. We present a detailed analysis of the rotation poisoned model and argue that commonly adopted data augmentation, although mitigating the effect at the backdoor angle, may introduce new vulnerabilities. We also evaluate five state-of-the-art backdoor defenses and conclude that none of them can serve a consistent countermeasure. Last, we illustrate that deploying rotation backdoor attacks in the physical world is easily accessible and raises a new real-world security issue.

One limitation of our proposed method is that transformation (rotation) is only possible under the constraint on the image domain and can hardly be generalized to other domains like audio and language. This highlights the significance of designing domain-specific backdoors in order to achieve both effective and stealthy attacks. In the future, we aim to explore combining rotation and other conventional patch-wise triggers to enhance the effectiveness of both methods. In addition, developing consistently practical approaches to defend against our attack is another promising direction.
We model the classification task as a hypothesis testing problem, with probability
\[ \rho \]
within the interval of \([a, b]\). We restrict the rotation degree to be within \([-180^\circ, 180^\circ]\). For a natural vision dataset, many images may already be rotated due to different camera viewpoints. We assume that the rotation degree of images in the original training distribution follows truncated Gaussian distribution \(D \sim N(0, \sigma^2, [-180^\circ, 180^\circ])\). Gaussian distribution is arguably the most reasonable assumption about rotation degrees in natural datasets due to the maximum entropy principle. Let \(\beta\) denotes the backdoor angle degree inserted during the training time, and \(\rho\) denotes the poisoning rate. Then for poisoned data, the distribution of rotation degree follows \(D_{\rho} \sim D + \rho \beta\). The overall rotation degree distribution after poisoning becomes a mixture of truncated Gaussian \((1 - \rho)D + \rho \beta\).

We model the classification task as a hypothesis testing problem, where the neural network needs first to decide whether the inputs are drawn from \(D\) or \(D_{\rho}\), and then make the prediction accordingly. For an image with rotation angle of degree \(x\), in order to minimize the cross entropy loss, the optimal classifier will predict clean label with probability \[ \frac{(1 - \rho)D(x) + \rho \beta(x)}{1 - \rho + \rho \beta} \]
and backdoored target label with probability \[ \frac{(1 - \rho)D(x) + \rho \beta(x)}{1 - \rho + \rho \beta} \]. Thus, the attack success rate for optimal classifier on rotation degree \(x\) is upper bounded by
\[ (1 - \rho)D(x) + \rho \beta(x) \]
In the following theorem, we show that the maximum possible attack success rate monotonically increases with the attack angle at the test time due to the exponential decay property of the Gaussian distribution.

**Theorem A.1.** Given sufficient training data points, the attack success rate for the optimal classifier on backdoored image \(x\) is maximized at \(x = 180^\circ\).

However, due to the low density of Gaussian tails, there may not be enough data points with large rotation angles for training the optimal classifier. Therefore, the original backdoor angles at the test time are only moderately higher than the backdoor angle at training time. To further validate our theory, in Figure 7, we increase the variance of the rotation degree of original training data by randomly rotating each data point with a degree drawn from \(N(0, \sigma^2, [-180^\circ, 180^\circ])\). In this case, there are more data points with a large rotation degree, which pushes the optimal backdoor angle to be higher. As shown in the figure, the optimal backdoor at the test time increases as \(\sigma\) grows, which matches our explanation.

### A ADDITIONAL ANALYSIS

#### A.1 Shift of Selected Backdoor Angle

As mentioned in Section 5.2, an interesting phenomenon is that the most effective attack angle at the test time is usually slightly higher than the predefined backdoor angle. Here, we provide a simple explanation for this phenomenon. We formulate backdoor prediction as a hypothesis testing problem. We use \(N(\mu, \sigma^2, [a, b])\) to denote truncated Gaussian within the interval of \([a, b]\). We restrict the rotation degree to be within \([-180^\circ, 180^\circ]\). For a natural vision dataset, many images may already be rotated due to different camera viewpoints. We assume that the rotation degree of images in the original training distribution follows truncated Gaussian distribution \(D \sim N(0, \sigma^2, [-180^\circ, 180^\circ])\). Gaussian distribution is arguably the most reasonable assumption about rotation degrees in natural datasets due to the maximum entropy principle. Let \(\beta\) denotes the backdoor angle degree inserted during the training time, and \(\rho\) denotes the poisoning rate. Then for poisoned data, the distribution of rotation degree follows \(D_{\rho} \sim N(\mu + \beta, \sigma^2, [-180^\circ, 180^\circ])\). The overall rotation degree distribution after poisoning becomes a mixture of truncated Gaussian \((1 - \rho)D + \rho \beta\).

We model the classification task as a hypothesis testing problem, where the neural network needs first to decide whether the inputs are drawn from \(D\) or \(D_{\rho}\), and then make the prediction accordingly. For an image with rotation angle of degree \(x\), in order to minimize the cross entropy loss, the optimal classifier will predict clean label with probability \[ \frac{(1 - \rho)D(x) + \rho \beta(x)}{1 - \rho + \rho \beta} \]
and backdoored target label with probability \[ \frac{(1 - \rho)D(x) + \rho \beta(x)}{1 - \rho + \rho \beta} \]. Thus, the attack success rate

### Table 8: Other defenses against 90° rotation backdoor attacks on CIFAR10 dataset.

| Defense | PR (%) | ASR (bef./aft.)(%) | CDA (bef./aft.)(%) | Forensics (bef./aft.)(%) |
|---------|--------|---------------------|---------------------|--------------------------|
| ANP     | 1%     | 53.77/16.52         | 94.47/93.28         | 8.73/96.33               |
|         | 10%    | 88.52/48.20         | 93.48/92.75         | 26.85/98.96              |

Table 8: Other defenses against 90° rotation backdoor attacks on CIFAR10 dataset. For ANP, we report ASR and CDA before and after the defense. For Forensics, we present the detection Precision (Pre.) and Recall (Rec.). For Frequency, we use Detection rate (Det.).

#### A.2 More Defenses on CIFAR10

We further evaluate three recent backdoor defenses, including ANP [53], Forensics [42], and Frequency [56], on a common object recognition benchmark CIFAR10 [23] dataset. According to their reports, they all exhibit adequate mitigating performance against patch-based or pixel-based attacks. ANP degrades the ASR to <5% against all other triggers, forensics achieves >95% detection precision, and frequency-based method performs >90% detection rate of other backdoors.

We replicate their experiments with default hyperparameter settings and 1%, 10% Poisoning Rate (PR), instead modifying the backdoor to 90° rotation.

All three defenses cannot provide adequate mitigation. As Table 8 shows, ANP drops the ASR to 16.52% for 1% poisoning rate, but the ASR for larger poisoning rate still remains 48.2%. Forensics achieves >96% of detection recall, but the precision rate is low (<27%) compared to other poisoning methods. Lastly, the frequency-based detector only identifies 51.30% rotation backdoors. Actually, the assumption of an obvious signature on the frequency domain caused by triggers is not valid for rotation transformation.

The threat model that defenses assumed is distinct from our new backdoor attack, which can be denoted as an unforeseen threat. We encourage researchers to carefully consider the assumptions to backdoor samples and develop a generic defense against more complex poisoning attacks in future studies.