DISTRIBUTION-ALIGNED REWEIGHTING FOR BACKDOOR DEFENSE

Anonymous authors
Paper under double-blind review

ABSTRACT

Backdoor attacks pose a serious threat to deep neural networks by embedding a backdoor to induce malicious behavior on triggered inputs while preserving normal performance on clean inputs. Existing in-training defenses often rely on assumptions about differences in the learning dynamics of clean and poisoned samples, which may not hold consistently across attacks. We propose Distribution-Aligned Reweighting (DAR), an in-training backdoor defense that views backdoor poisoning as an adversarial shift of the training distribution away from the underlying clean distribution. Given a small trusted clean reference set, DAR performs class-wise feature distribution alignment to assign weights to training samples and uses the resulting weights to construct a sampling distribution for standard model training. DAR does not require prior knowledge of the attack and does not modify the model architecture or training objective. Extensive experiments across CIFAR-10, GTSRB, and ImageNet-12, covering diverse backdoor attacks and poisoning ratios, demonstrate that DAR consistently achieves low attack success rates while maintaining competitive clean accuracy and strong robust accuracy on triggered inputs.

1 INTRODUCTION

Deep Neural Networks (DNNs) have become a successful class of tools that have continued to grow in popularity in recent years (Canziani et al., 2016; Liu et al., 2020; Li et al., 2024). This rise can be attributed to the use of large-scale, weakly curated training data and third-party pretraining (Devlin et al., 2019; Radford et al., 2021; Dosovitskiy et al., 2021). These practices can create potential model safety concerns for users, with poisoning-based backdoor attacks being one major example. Backdoor attacks typically encode a hidden trigger-to-target association by poisoning the training data. A model trained on such poisoned samples behaves normally on clean inputs while misclassifying triggered inputs as an attacker-chosen label. Such backdoored models usually show normal performance on clean inputs, and the backdoor is activated only by the trigger, making them a stealthy threat in security-sensitive applications (Gu et al., 2019; Goldblum et al., 2022; Shejwalkar et al., 2022; Carlini et al., 2024).

Various defenses have been proposed against backdoor attacks and can be broadly categorized according to the stage at which they are applied: pre-training, in-training, and post-training defenses. Pre-training defenses typically identify and remove suspicious poisoned samples or neutralize potential triggers before model training (Tran et al., 2018; Hayase et al., 2021). Post-training defenses operate on an already-trained model and aim to erase or mitigate the learned backdoor, for example, through model pruning or fine-tuning (Wu & Wang, 2021; Min et al., 2023). In-training defenses often modify the learning process to suppress backdoor injection (Li et al., 2021a; Huang et al., 2021; Yu et al., 2025).

In this work, we focus on the in-training setting, where the defender has full access to the potentially poisoned training pool and control over model training. Compared with pre-training defenses, which typically screen or preprocess the training data before final model optimization, in-training defenses can incorporate mitigation directly into the learning process and regulate how individual training samples influence the learned model. Compared with post-training defenses, which aim to repair an already-trained model, in-training methods can leverage the available training data while the model is being learned rather than attempting to remove an already-established backdoor afterward.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

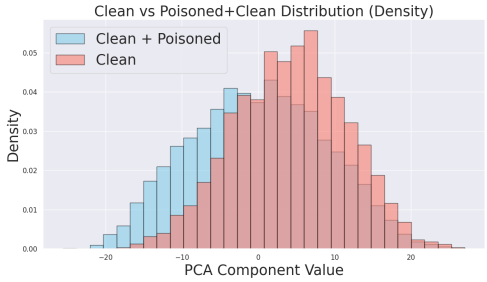

Figure 1: Feature-distribution shift induced by poisoned samples in the target class on CIFAR-10, compared with the corresponding clean reference distribution.

Therefore, when the defender has access to the training pool and control over model training, in-training defense provides a natural framework for mitigating backdoor attacks.

Existing in-training defenses often rely on explicitly distinguishing poisoned samples from clean ones based on assumptions about their learning behavior. Anti-Backdoor Learning (ABL) (Li et al., 2021a) assumes that poisoned samples tend to be learned faster and therefore exhibit smaller losses during early training. ABL leverages this signal to separate suspicious samples for subsequent unlearning. However, such loss separation may not consistently hold across different attacks and poisoning settings, limiting the reliability of the method. Decoupled Backdoor Defense (DBD) (Huang et al., 2021) first learns representations through self-supervised training and then identifies high-credibility samples based on their supervised losses, effectively relying on poisoned and clean samples being sufficiently distinguishable under the learned representation. This assumption may break down when their loss distributions overlap. More recently, Enhanced Splitting and Trap Isolation (ESTI) (Yu et al., 2025) first separates the training data into suspected clean and poisoned subsets and then employs a trap-isolation mechanism to suppress the backdoor behavior. Its effectiveness therefore also depends on the quality of the initial sample separation.

These limitations motivate an alternative in-training strategy that does not rely on assumptions about differences in the learning dynamics of clean and poisoned samples. Instead, it leverages clean reference information to assess training samples through distributional alignment, using the resulting weights to regulate their influence during training.

We propose Distribution-Aligned Reweighting (DAR), an in-training backdoor defense inspired by viewing backdoor poisoning as an adversarial contamination that shifts the training distribution away from the underlying clean data distribution. As illustrated in Figure 1, poisoned samples can alter the feature distribution of the training data, thereby causing a model trained on the contaminated distribution to learn the malicious shortcut. This observation motivates a natural question: rather than explicitly detecting poisoned samples based on their presumed learning behavior, can we reduce the discrepancy between the contaminated and clean distributions through sample reweighting? Sample reweighting has been widely studied as a means of correcting distribution mismatch, including importance weighting under covariate shift (Shimodaira, 2000), kernel mean matching (KMM) (Gretton et al., 2008), and direct density-ratio estimation (Kanamori et al., 2009). Building on this idea, and following prior work such as ESTI (Yu et al., 2025), DAR uses a small trusted clean subset as a reference distribution to reweight the potentially poisoned training samples. Specifically, we first extract feature representations of the training samples and then perform class-wise KMM to assign sample weights such that the weighted feature distribution of each class is aligned with that of the corresponding clean reference samples. These weights define a sampling distribution for standard DNN training, allowing the defense to be incorporated while retaining the standard model architecture and loss function.

Our contributions can be summarized as follows. We introduce a distributional perspective for in-training backdoor defense, viewing poisoned training samples as inducing an adversarial shift from the underlying clean distribution and mitigating this shift through class-wise feature distribution alignment with a small trusted clean subset. Empirically, DAR consistently achieves strong performance across multiple datasets and diverse attack settings, substantially reducing attack success while preserving clean accuracy and, in particular, improving robust accuracy by correctly recovering the original classes of triggered inputs. Extensive ablation studies and evaluations under adaptive attacks further demonstrate the robustness of the proposed framework.

2 Related Work

Backdoor Attack. Backdoor attacks aim to implant a hidden trigger-to-target association into a model while preserving its normal behavior on clean inputs (Gu et al., 2019; Chen et al., 2017; Li et al., 2020). Existing attacks have explored increasingly diverse and stealthy trigger designs, ranging from simple fixed patterns, such as the patch-based trigger in BadNet (Gu et al., 2019) and blended patterns (Chen et al., 2017), to more sophisticated trigger constructions designed to improve stealthiness. Representative examples include sample-specific invisible triggers in SSBA (Li et al., 2021b), natural reflection-based triggers in Refool (Li et al., 2020), smooth low-frequency triggers designed to avoid high-frequency artifacts in LowFreq (Zeng et al., 2021), and invisible triggers directly injected into the frequency domain in FTrojan (Wang et al., 2022). Backdoor attacks can also vary in their target mappings and poisoning strategies, including all-to-one and all-to-all attacks, as well as dirty-label and clean-label attacks (Turner et al., 2018; Barni et al., 2019). In this work, we focus on the commonly studied all-to-one setting, where triggered inputs are mapped to a fixed attacker-chosen target class. We evaluate our method against a diverse collection of attacks, including BadNet, Blended, SSBA, Refool, LowFreq, FTrojan, and TITIM (Lin et al., 2025), covering both visible and stealthy triggers as well as spatial- and frequency-domain attack mechanisms.

Backdoor Defense. Backdoor defenses have been developed across different stages of the learning pipeline. Pre-training defenses primarily operate on the training data before model optimization, for example, by identifying and removing suspicious poisoned samples through spectral analysis (Tran et al., 2018) or robust statistical methods (Hayase et al., 2021). Post-training defenses assume access to an already-trained, potentially backdoored model and aim to remove the embedded backdoor. Representative strategies include recovering potential triggers and using them to diagnose or mitigate the backdoor, pruning neurons or channels that are strongly associated with backdoor behavior, and fine-tuning the model on trusted clean samples (Wang et al., 2019; Wu & Wang, 2021; Min et al., 2023). A separate line of work defends against backdoors directly during model training. ABL (Li et al., 2021a) exploits the faster learning behavior of poisoned samples to isolate suspicious samples and subsequently unlearn their influence, while DBD (Huang et al., 2021) decouples self-supervised representation learning from supervised classification and uses sample credibility to guide subsequent training. More recently, ESTI (Yu et al., 2025) improves sample splitting using auxiliary models and employs a trap-isolation mechanism to suppress the influence of suspected poisoned samples. Our work follows this in-training direction, but approaches the problem from a distributional perspective: rather than characterizing poisoned samples through specific learning dynamics, we use a small trusted clean subset to guide class-wise distribution alignment and regulate the contribution of training samples during optimization.

3 Methodology

In this section, we present **Distribution-Aligned Reweighting (DAR)**, an in-training defense that mitigates backdoor poisoning through class-wise feature distribution alignment with a trusted clean reference set. We first introduce the defense setting and problem formulation in Section 3.1. We then describe the proposed distribution-aligned reweighting framework in Section 3.2, including feature extraction, class-wise weight estimation, and model training with the learned weights.

3.1 Problem Setup

Defense Setting. We consider poisoning-based backdoor attacks in which an adversary manipulates a subset of the training samples to implant a hidden trigger-to-target association. The defender has full access to the potentially poisoned training set and control over the model training procedure, but has no prior knowledge of the proportion nor distribution of the poisoned samples. Following prior work such as ESTI (Yu et al., 2025), we assume that the defender additionally has access to a small trusted clean reference set. The defender's goal is to train a model that maintains high predictive accuracy on clean inputs while suppressing the malicious behavior on triggered inputs.

Problem Formulation. Let $\mathcal{D}_{\mathrm{tr}} = \{(x_i, y_i)\}_{i=1}^{n}$ denote the potentially poisoned training set over C classes, where $x_i \in \mathcal{X}$ and $y_i \in \{1, \ldots, C\}$. We view $\mathcal{D}_{\mathrm{tr}}$ as being drawn from a contaminated

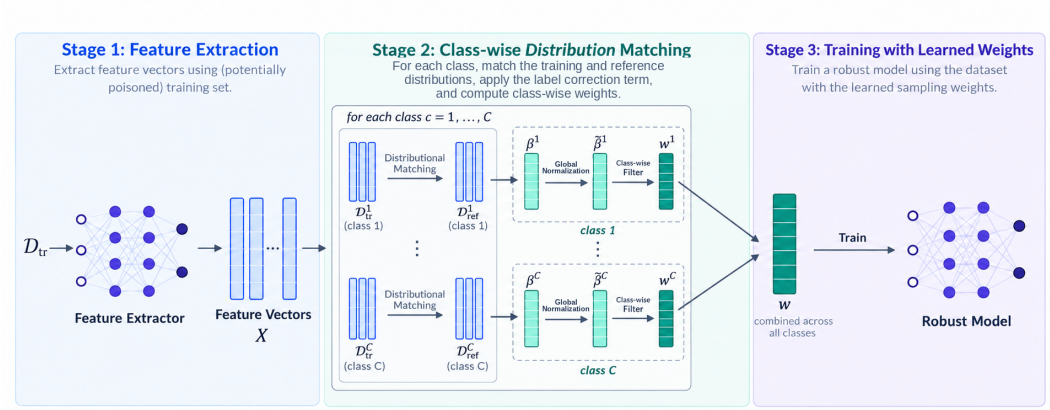

Figure 2: Overview of the DAR pipeline. First, feature representations are extracted for the training and clean reference samples. Second, class-wise distribution matching produces weights $\boldsymbol{\beta}^c$, which are normalized to $\widetilde{\boldsymbol{\beta}}^c$ and processed using an adaptive cutoff to obtain the final weights $\boldsymbol{w}$. Third, these weights define the sampling distribution for standard model training.

distribution $\widetilde{P}$, which differs from the underlying clean distribution P due to the presence of poisoned samples. Let $\mathcal{D}_{\mathrm{ref}} = \{(x_j, y_j)\}_{j=1}^{m}, m \ll n$, denote the trusted clean reference set drawn from P. Since our distribution alignment is performed class-wise, for each class $c \in \{1, \ldots, C\}$, we define

$$\mathcal{D}_{\mathrm{tr}}^c = \{(x_i, y_i) \in \mathcal{D}_{\mathrm{tr}} : y_i = c\}, \qquad \mathcal{D}_{\mathrm{ref}}^c = \{(x_j, y_j) \in \mathcal{D}_{\mathrm{ref}} : y_j = c\}, \tag{1}$$

with $n_c = |\mathcal{D}_{\mathrm{tr}}^c|$ and $m_c = |\mathcal{D}_{\mathrm{ref}}^c|$.

Let $\phi : \mathcal{X} \to \mathbb{R}^d$ denote a feature extractor that maps each input to a d-dimensional representation. Our objective is to learn a nonnegative weight w_i for each training sample such that, within each class, the weighted feature distribution of $\mathcal{D}_{\mathrm{tr}}^c$ is aligned with the feature distribution of $\mathcal{D}_{\mathrm{ref}}^c$.

3.2 Distribution-Aligned Reweighting

Given the potentially poisoned training set $\mathcal{D}_{\mathrm{tr}}$ and the trusted clean reference set $\mathcal{D}_{\mathrm{ref}}$, DAR consists of three main steps, as illustrated in Figure 2. First, we obtain feature representations for samples in both sets using a feature extractor ϕ. Second, we perform feature distribution alignment within each class to estimate a weight for every training sample. Third, the resulting weights are used to construct a sampling distribution for standard DNN training.

Feature Representation. We first obtain a feature extractor $\phi : \mathcal{X} \to \mathbb{R}^d$ and use it to map each input to a d-dimensional representation $z = \phi(x)$. In our default setting, we train a DNN on the potentially poisoned training set $\mathcal{D}_{\mathrm{tr}}$ using the same backbone architecture as the final model. After training, we remove the classification layer and use the resulting backbone as the feature extractor for all samples in $\mathcal{D}_{\mathrm{tr}}$ and $\mathcal{D}_{\mathrm{ref}}$. We find that a randomly initialized DNN can provide feature representations that yield comparable defense performance, suggesting that DAR does not strongly depend on a carefully trained feature extractor. We provide the corresponding experimental results in the appendix C.

Class-wise Distribution-Aligned Reweighting. We next estimate sample weights by aligning the feature distribution of the potentially poisoned training set with that of the trusted clean reference set. Rather than performing alignment globally across all classes, we conduct the matching separately within each class. This class-wise strategy preserves class-specific structure during distribution alignment and prevents natural inter-class differences from dominating the distribution-matching objective.

For each class c, we use Kernel Mean Matching (KMM) (Gretton et al., 2008) to estimate a set of nonnegative weights $\boldsymbol{\beta}^c = (\beta_i^c)_{i:y_i=c}$. Let ψ denote the feature map associated with a reproducing

kernel Hilbert space (RKHS) $\mathcal{H}$. We obtain the weights by solving

$$\min_{\boldsymbol{\beta}^c} \quad \left\| \frac{1}{n_c} \sum_{i:y_i=c} \beta_i^c \, \psi(\phi(x_i)) - \frac{1}{m_c} \sum_{j:y_j=c} \psi(\phi(x_j)) \right\|_{\mathcal{H}}^2 \tag{2}$$
$$\text{s.t.} \quad 0 \le \beta_i^c \le B,$$

where the first summation is over samples in $\mathcal{D}_{\text{tr}}^c$ and the second is over samples in $\mathcal{D}_{\text{ref}}^c$. The upper bound B prevents a small number of samples from dominating the weighted distribution. The inner products in $\mathcal{H}$ can be evaluated through a kernel function $k(z, z') = \langle \psi(z), \psi(z') \rangle_{\mathcal{H}}$. Different kernel choices and matching methods are evaluated in the ablation studies. By minimizing Equation 2, KMM assigns weights so that the weighted kernel mean embedding of the training samples in class c is close to that of the corresponding clean reference samples.

After solving Equation 2, we apply a label shift adjustment to mitigate the difference between the class-label distributions of the training and reference set:

$$\hat{\beta}_i^c = \beta_i^c * \frac{m_c/m}{n_c/n} \tag{3}$$

After applying the adjustment, we normalize the weights to have global mean one for subsequent weighted sampling:

$$\widetilde{\beta}_i = \frac{\hat{\beta}_i}{\frac{1}{n}\sum_r \hat{\beta}_r}. \tag{4}$$

We then apply an adaptive class-specific cutoff to further reduce the influence of samples receiving relatively small weights. Let

$$\tau_c = 1 - \alpha \sqrt{\text{Var}\left(\{\widetilde{\beta}_i^c\}_{y_i=c}\right)}, \tag{5}$$

where α controls the strength of the adaptive cutoff. The final training weight for sample (x_i, y_i) is defined as

$$w_i = \begin{cases} \widetilde{\beta}_i^c, & \widetilde{\beta}_i^c \ge \tau_c, \\ 0, & \widetilde{\beta}_i^c < \tau_c, \end{cases} \qquad \text{for } y_i = c. \tag{6}$$

Because the normalized weights have mean one, the cutoff automatically adapts to the dispersion of the learned weights. A smaller within-class variation results in a cutoff closer to one, whereas a larger variation produces a more lenient cutoff.

Training with Learned Weights. The final model is trained using the standard model architecture, loss function, and data augmentation procedure. The only modification is the sampling distribution used to construct the training batches. Specifically, the learned weights define the empirical sampling distribution

$$\widehat{P}_w(x_i, y_i) = \frac{w_i}{\sum_{r=1}^n w_r}, \qquad (x_i, y_i) \in \mathcal{D}_{\text{tr}}. \tag{7}$$

Samples with larger weights are therefore sampled more frequently during training, while samples with smaller weights are sampled less frequently. The model is then optimized using the standard task loss under this sampling distribution. Thus, DAR modifies the effective training distribution without requiring changes to the model architecture or training objective.

4 Experiments

4.1 Experimental Setup

We conduct a comprehensive empirical evaluation of DAR under diverse backdoor attack and poisoning settings. We first compare DAR with representative in-training defenses across three benchmark datasets. We then conduct additional studies to examine the effects of the clean reference set size, runtime efficiency, model architecture, and adaptive attacks. The experimental setup, including datasets, attacks, defenses, and evaluation metrics, is described below.

Dataset and Architecture Details. We evaluate on three benchmark datasets: CIFAR-10 (Krizhevsky & Hinton, 2009), GTSRB (Stallkamp et al., 2011), and ImageNet-12, a 12-class subset of ImageNet (Deng et al., 2009). We use PreAct-ResNet18 (He et al., 2016b) for CIFAR-10 and GTSRB, and Tiny Swin Transformer (Liu et al., 2021) for ImageNet-12.

Attack Settings. We consider seven backdoor attacks: BadNet (Gu et al., 2019), Blended attack with the Hello Kitty pattern (Chen et al., 2017), invisible sample-specific backdoor attacks (SSBA) (Li et al., 2021b), reflection backdoor attack (Refool) (Li et al., 2020), a low-frequency attack (LowFreq) (Zeng et al., 2021), an invisible backdoor attack through the frequency domain (Ftrojan) (Wang et al., 2022), and an intensity-shifting attack (TITIM) (Lin et al., 2025). Each attack is evaluated at five poison ratios: 0.005, 0.01, 0.05, 0.1, and 0.2, covering small to large poisoning scenarios. The target class for all attacks is set to class 0.

Defense Settings. We compare the proposed method with three in-training defense methods: Anti-Backdoor Learning (ABL) (Li et al., 2021a), Decoupling Backdoor Defense (DBD) (Huang et al., 2021), and Enhanced Splitting and Trap Isolation (ESTI) (Yu et al., 2025). DAR denotes the proposed method, while Baseline refers to standard training without any defense. See the appendix section B for more implementation details.

Evaluation Metrics. We use clean accuracy (C-Acc), attack success rate (ASR), and robust accuracy (RA) to evaluate defense performance. C-Acc measures accuracy on clean inputs, ASR measures the fraction of triggered inputs classified into the target class, and RA measures the fraction correctly classified into their original classes. Higher C-Acc/RA and lower ASR are better. We **bold** the best average performance across all settings.

4.2 Main Results

Tables 1, 2, and 3 present the main results on CIFAR-10, GTSRB, and ImageNet-12, respectively, across seven attacks and five poison ratios. Overall, DAR provides the most consistent tradeoff among C-Acc, ASR, and RA across datasets and attack settings. On CIFAR-10 and GTSRB, DAR achieves the lowest average ASR (0.80 and 0.36) and the highest average RA (87.48 and 83.15), while maintaining high clean accuracy. On ImageNet-12, ESTI achieves a slightly lower average ASR than DAR (0.43 vs. 0.60), but DAR achieves substantially higher C-Acc (78.92 vs. 67.35) and RA (70.08 vs. 30.42). In comparison, ABL often suffers from considerable clean-accuracy degradation, while DBD maintains competitive C-Acc but becomes less effective under several higher poison ratio settings. ESTI suppresses ASR effectively in many cases, but often at the cost of robust accuracy and, on ImageNet-12, clean accuracy. Overall, DAR consistently suppresses backdoor behavior while better preserving clean performance and the ability to recover the original classes of triggered inputs across datasets, attacks, and poison ratios.

4.3 Ablation and Robustness Analysis

Reference Set Size Ablation. DAR assumes access to a small trusted clean reference set, which provides a reference distribution for distribution alignment. In the main experiments, we use a clean reference set whose size is 10% of the original training set. To examine the sensitivity of DAR to the amount of clean reference data, we additionally evaluate a smaller 5% reference set on CIFAR-10 and GTSRB. Table 4 reports the results across all attacks and poison ratios.

Reducing the reference set from 10% to 5% has only a modest effect on DAR. Averaged over all attack and poison-ratio settings, the larger reference set improves C-Acc and RA from 97.38 and 82.23 to 97.71 and 83.15 on GTSRB, and from 89.60 and 87.00 to 90.69 and 87.48 on CIFAR-10. Interestingly, the 5% reference set achieves slightly lower average ASR on both datasets (0.12 vs. 0.36 on GTSRB and 0.71 vs. 0.80 on CIFAR-10). Overall, DAR maintains strong defense performance with only half of the clean reference data used in the main experiments, indicating limited sensitivity to moderate reductions in reference set size.

Model Architecture Ablation. To evaluate the robustness of DAR across model architectures, we consider four additional architectures on CIFAR-10: ResNet-18 (He et al., 2016a), VGG19 (Simonyan & Zisserman, 2015), MobileNetV2 (Sandler et al., 2018), and DenseNet161 (Huang et al., 2018), in addition to PreActResNet-18 (He et al., 2016b) used in the main experiments.

Table 1: Backdoor defense results on the CIFAR-10 dataset

Trigger	P. Ratio	Baseline			DAR			ABL			DBD			ESTI		
		C-Acc ↑	ASR ↓	RA ↑	C-Acc ↑	ASR ↓	RA ↑	C-Acc ↑	ASR ↓	RA ↑	C-Acc ↑	ASR ↓	RA ↑	C-Acc ↑	ASR ↓	RA ↑
BadNet	0.005	93.57	58.27	40.25	90.18	0.51	91.22	28.47	72.42	10.32	91.51	1.30	91.13	93.44	37.13	49.20
	0.01	93.37	77.05	22.10	88.34	0.57	89.33	77.65	0.65	82.65	92.33	1.26	92.12	93.30	24.93	34.73
	0.05	92.37	90.63	8.94	90.64	0.23	92.95	83.26	0.01	88.88	93.14	1.22	92.67	92.16	0.01	14.10
	0.10	91.85	93.86	5.92	91.66	0.24	93.57	81.44	0.00	87.00	92.08	9.30	86.31	93.92	0.14	13.52
	0.20	90.10	96.54	3.33	91.21	0.25	93.26	80.72	0.02	87.10	90.72	51.54	47.50	90.86	0.08	12.47
Blended	0.005	94.33	90.64	8.78	90.66	1.96	75.75	56.34	59.08	19.50	92.01	19.83	70.92	93.27	90.52	8.57
	0.01	93.72	95.12	4.61	88.43	1.98	74.44	71.29	0.07	36.55	91.89	79.21	19.88	91.69	0.02	18.92
	0.05	93.80	99.18	0.78	90.85	0.80	78.25	72.44	0.00	49.70	91.45	99.65	0.33	92.28	0.00	12.18
	0.10	93.06	99.81	0.18	92.05	0.78	78.02	76.16	0.37	64.96	90.77	99.97	0.02	93.92	0.01	12.58
	0.20	92.86	99.88	0.11	90.72	0.87	77.38	77.17	1.76	68.03	88.50	100.0	0.00	93.38	0.00	11.25
SSBA	0.005	93.92	17.18	76.82	90.46	0.62	90.03	59.49	24.84	42.62	92.02	1.36	88.70	93.38	0.90	90.68
	0.01	93.55	78.54	20.32	88.70	0.71	88.01	54.27	79.34	14.95	91.96	1.72	89.22	93.16	64.15	27.72
	0.05	93.41	99.54	0.44	91.19	0.24	91.50	77.69	1.16	71.86	92.74	1.93	90.19	92.76	0.00	11.76
	0.10	93.29	99.88	0.11	91.82	0.22	91.77	75.48	0.20	75.93	91.33	99.96	0.03	93.41	0.00	11.32
	0.20	92.82	99.91	0.07	91.26	0.33	91.44	78.17	2.97	80.54	87.54	100.0	0.00	93.22	0.00	11.15
Refool	0.005	93.61	29.98	64.12	91.08	0.32	84.98	67.96	28.55	43.63	91.80	1.81	87.71	89.96	22.44	63.30
	0.01	93.58	46.47	49.68	90.32	0.33	85.75	65.77	51.72	32.88	91.96	1.80	88.71	92.74	30.80	52.31
	0.05	92.89	85.23	13.76	90.83	0.25	85.83	53.35	93.23	4.50	92.23	3.23	88.93	90.84	2.23	22.86
	0.10	92.39	93.40	6.21	91.37	0.12	87.52	56.98	99.48	0.45	92.98	7.44	86.16	76.74	0.02	14.17
	0.20	91.53	97.30	2.56	91.22	0.45	86.82	33.18	99.98	0.00	92.43	39.59	57.61	72.62	0.00	11.51
LowFreq	0.005	93.45	74.78	23.42	90.42	1.00	89.48	50.93	27.17	40.64	92.18	2.26	89.53	90.61	3.78	43.72
	0.01	93.68	86.14	13.13	88.10	1.37	86.65	64.65	59.52	27.45	92.01	6.46	84.09	85.68	0.26	24.47
	0.05	93.41	97.73	2.16	90.92	0.57	90.73	58.73	98.22	1.30	91.43	96.16	3.60	89.73	0.97	15.51
	0.10	93.36	98.83	1.08	91.94	0.85	90.37	84.25	0.01	84.04	90.18	98.65	1.21	91.28	0.32	12.82
	0.20	92.73	99.58	0.37	91.07	0.72	90.67	79.15	0.42	86.85	88.79	99.35	0.61	78.36	0.02	11.47
Ftrojan	0.005	93.95	99.48	0.46	91.54	0.74	91.43	65.94	2.67	67.00	92.29	1.27	91.70	92.31	100.0	0.00
	0.01	93.87	99.90	0.10	89.10	0.81	89.82	68.42	3.76	73.36	92.14	1.33	91.72	92.56	100.0	0.00
	0.05	93.68	99.98	0.01	91.08	0.27	93.23	80.40	0.01	86.87	91.54	100.0	0.00	92.96	0.00	11.11
	0.10	93.69	100.0	0.00	91.94	0.24	93.61	71.46	0.00	75.37	90.85	100.0	0.00	92.24	0.00	11.11
	0.20	93.08	100.0	0.00	91.36	0.30	93.56	78.06	0.00	79.58	87.96	100.0	0.00	92.04	0.00	11.11
TITIM	0.005	93.76	100.00	0.00	90.51	2.61	82.36	55.35	68.46	11.22	92.11	1.97	88.20	89.40	100.0	0.00
	0.01	93.89	100.0	0.00	89.51	3.90	83.52	54.51	13.35	28.34	91.86	100.0	0.00	93.18	0.00	11.11
	0.05	93.59	100.0	0.00	90.94	0.43	85.50	74.46	0.00	72.20	91.46	100.0	0.00	93.59	0.00	11.11
	0.10	93.58	100.0	0.00	91.80	1.28	87.21	69.10	0.01	54.00	91.00	100.0	0.00	92.35	0.00	11.11
	0.20	93.14	100.0	0.00	90.86	1.17	85.90	78.61	0.00	77.23	88.12	100.0	0.00	93.40	0.00	11.11
Average		**93.23**	88.71	10.57	90.69	**0.80**	**87.48**	67.47	25.41	52.21	91.29	49.42	46.54	90.76	16.54	19.72

Table 2: Backdoor defense results on the GTSRB dataset

Trigger	P. Ratio	Baseline			DAR			ABL			DBD			ESTI		
		C-Acc ↑	ASR ↓	RA ↑	C-Acc ↑	ASR ↓	RA ↑	C-Acc ↑	ASR ↓	RA ↑	C-Acc ↑	ASR ↓	RA ↑	C-Acc ↑	ASR ↓	RA ↑
BadNet	0.005	98.18	79.38	20.35	98.07	0.00	98.08	66.12	50.47	36.71	97.99	0.00	97.93	99.25	0.00	10.66
	0.01	97.47	89.31	10.32	97.68	0.00	97.69	66.44	7.82	63.87	96.03	0.00	96.02	98.93	0.00	8.52
	0.05	97.14	93.74	5.86	97.55	0.00	97.51	80.78	0.00	81.22	98.16	0.00	98.09	91.93	0.00	2.64
	0.10	97.22	94.78	5.20	98.28	0.00	98.29	93.73	0.00	94.32	99.39	0.00	99.38	90.04	0.00	1.90
	0.20	97.59	95.44	4.55	98.00	0.00	98.01	94.57	0.00	95.13	99.16	81.44	18.53	92.96	0.00	1.03
Blended	0.005	98.62	94.40	4.61	97.76	0.05	57.22	79.73	82.52	11.06	97.98	99.96	0.03	99.52	92.77	6.91
	0.01	98.27	95.86	3.81	97.02	2.25	59.92	93.05	90.89	7.59	97.22	100.0	0.00	98.85	0.00	4.30
	0.05	98.48	99.98	0.01	97.71	4.15	56.56	65.53	0.00	11.21	97.20	100.0	0.00	96.52	0.00	0.93
	0.10	98.17	100.0	0.00	97.94	1.23	60.70	69.36	0.00	38.24	97.46	100.0	0.00	99.43	0.00	0.78
	0.20	98.41	100.0	0.00	98.06	3.73	59.49	48.11	0.00	39.26	95.41	100.0	0.00	98.69	0.00	0.73
SSBA	0.005	98.31	90.41	9.21	97.52	0.02	95.96	2.68	97.92	0.95	97.99	0.00	97.61	99.41	98.13	1.85
	0.01	98.02	98.08	1.73	97.41	0.00	95.99	5.05	99.26	0.32	98.11	0.00	97.64	99.15	91.20	1.37
	0.05	98.27	99.97	0.02	97.25	0.00	95.33	83.88	0.00	79.02	97.99	0.00	97.61	99.12	0.00	0.71
	0.10	98.47	100.0	0.0	96.84	0.02	95.24	93.70	0.00	90.18	99.28	0.00	98.85	98.78	0.00	0.72
	0.20	97.84	100.0	0.00	98.24	0.01	96.22	87.51	0.00	86.70	98.94	100.0	0.00	98.82	0.00	0.72
Refool	0.005	98.25	42.33	53.69	97.61	0.02	69.96	24.47	76.05	10.14	98.02	0.00	82.37	97.75	40.91	56.49
	0.01	97.98	54.49	43.28	97.77	0.04	69.77	7.92	89.72	3.33	97.92	0.00	79.48	99.12	56.96	42.34
	0.05	96.35	86.21	13.44	97.98	0.07	67.93	15.49	94.94	1.45	97.93	5.86	82.64	84.11	0.00	19.54
	0.10	96.02	89.57	10.20	98.08	0.06	68.53	10.69	99.61	0.21	97.67	28.96	68.41	63.44	0.00	8.79
	0.20	94.82	91.53	8.17	98.21	0.11	68.67	7.63	99.95	0.04	99.31	60.91	38.74	51.70	0.00	2.96
LowFreq	0.005	97.80	93.56	5.99	97.85	0.07	68.08	68.00	87.85	8.00	96.68	0.00	96.04	99.33	1.11	9.11
	0.01	98.02	95.75	3.81	97.48	0.18	66.17	43.07	91.10	4.70	98.11	0.00	97.69	99.23	0.00	6.40
	0.05	97.87	99.18	0.77	97.45	0.50	62.90	61.98	95.07	2.74	98.46	0.00	97.98	98.46	0.00	2.03
	0.10	97.72	99.73	0.25	97.66	0.00	67.55	35.86	97.31	1.86	99.26	2.65	96.17	98.11	0.00	1.13
	0.20	98.23	99.76	0.23	97.86	0.03	63.10	94.03	97.54	2.27	99.02	100.0	0.00	92.89	0.00	0.89
Ftrojan	0.005	98.54	100.0	0.00	97.29	0.00	97.41	36.38	94.48	1.34	96.52	0.00	96.39	98.07	98.77	1.19
	0.01	97.94	100.0	0.00	97.43	0.00	97.40	40.87	99.41	0.00	97.86	0.00	97.73	99.11	0.02	2.85
	0.05	98.52	100.0	0.00	97.33	0.00	97.35	87.81	0.00	87.48	99.08	0.00	99.10	99.08	0.00	0.71
	0.10	98.52	100.0	0.00	97.39	0.00	97.37	49.16	100.0	0.00	99.38	0.00	99.46	99.11	0.02	2.85
	0.20	98.47	100.0	0.00	97.98	0.00	97.92	90.18	0.00	53.49	97.51	100.0	0.00	98.86	0.00	0.72
TITIM	0.005	98.57	100.0	0.00	98.16	0.00	97.09	72.34	99.60	0.25	97.44	0.00	97.27	99.56	0.00	0.71
	0.01	98.58	100.0	0.00	97.56	0.00	97.66	45.09	0.00	44.32	97.13	0.00	96.99	99.46	0.00	0.71
	0.05	98.55	100.0	0.00	97.44	0.00	97.46	92.89	0.00	95.53	98.82	0.01	98.65	99.17	0.00	0.71
	0.10	98.90	100.0	0.00	98.02	0.00	97.93	93.46	0.00	94.48	99.51	0.00	99.47	99.05	0.00	0.71
	0.20	97.94	100.0	0.00	97.81	0.00	97.85	91.19	0.00	93.05	99.22	100.0	0.00	98.25	0.00	0.71
Average		97.94	93.81	5.87	97.71	**0.36**	**83.15**	59.96	50.04	35.44	**98.09**	30.85	66.46	95.29	13.71	5.87

As shown in Figure 3, DAR achieves low average ASR across all five architectures, although performance varies across individual attacks and poison ratios. PreActResNet-18 achieves the strongest overall performance, while the remaining architectures retain competitive average C-Acc and RA. These results indicate that the effectiveness of DAR is not limited to the architecture used in the main experiments. Detailed results for individual attack and poison-ratio settings are provided in appendix section C.

4.4 Runtime Efficiency

Since the compared defenses follow substantially different training pipelines, metrics such as FLOPs or per-epoch runtime alone may not reflect their practical computational costs. We therefore compare their end-to-end wall-clock runtimes using the performance-tuned configurations adopted in our experiments, together with stage-wise and per-epoch runtimes where applicable. All experiments are conducted on the same NVIDIA L40 GPU and repeated five times, with the average runtime

Table 3: Backdoor Defense Results on the ImageNet-12 dataset

Trigger	P. Ratio	Baseline			DAR			ABL			DBD			ESTI		
		C-Acc ↑	ASR ↓	RA ↑	C-Acc ↑	ASR ↓	RA ↑	C-Acc ↑	ASR ↓	RA ↑	C-Acc ↑	ASR ↓	RA ↑	C-Acc ↑	ASR ↓	RA ↑
BadNet	0.005	91.00	77.27	21.63	77.00	0.91	77.64	64.50	0.72	63.81	83.18	1.62	83.77	75.83	0.72	62.36
	0.01	91.33	91.45	8.36	78.66	0.36	78.73	67.16	3.63	66.00	84.04	2.55	84.92	72.67	0.36	37.27
	0.05	92.00	99.27	0.72	80.16	0.72	82.36	54.16	0.90	55.63	83.96	100.0	0.00	71.67	0.00	9.45
	0.1	90.66	99.27	0.72	81.83	0.36	82.72	61.83	98.18	1.81	7.81	100.0	0.00	73.50	0.00	9.09
	0.2	89.83	98.18	1.63	81.33	0.91	83.45	65.16	100.0	0.00	84.41	100.0	0.00	69.50	0.00	9.09
Blended	0.005	91.83	92.18	7.27	75.83	0.72	65.45	65.33	6.36	52.18	86.44	1.40	76.94	76.83	0.54	41.81
	0.01	91.16	99.63	0.36	76.33	0.72	68.00	69.33	12.54	52.18	84.64	64.12	33.06	74.67	2.36	32.00
	0.05	92.50	99.81	0.18	79.83	0.36	71.82	66.50	76.18	16.72	86.29	100.0	0.00	75.50	0.00	12.91
	0.1	90.50	100.0	0.00	79.50	0.00	70.55	67.66	98.18	1.81	85.04	100.0	0.00	70.00	0.00	10.54
	0.2	90.00	100.0	0.00	81.16	0.36	68.91	65.00	99.09	0.54	84.41	100.0	0.00	72.33	0.00	9.63
SSBA	0.005	92.50	3.63	89.09	75.00	0.55	76.72	69.16	0.54	70.18	87.72	1.46	85.06	77.66	0.54	74.91
	0.01	91.00	65.09	33.45	76.00	0.18	78.61	69.33	4.36	67.45	87.61	1.30	84.16	74.83	0.54	71.27
	0.05	91.50	99.63	0.36	79.67	0.18	78.55	38.00	0.00	39.63	7.81	100.0	0.00	48.67	0.36	32.00
	0.1	90.33	99.81	0.18	81.00	0.18	82.00	21.16	0.10	24.72	86.43	100.0	0.00	36.16	0.18	18.72
	0.2	90.16	100.0	0.00	82.00	0.18	83.45	59.66	98.36	1.27	84.26	100.0	0.00	16.83	0.00	11.63
Refool	0.005	92.16	2.36	87.81	75.83	2.36	67.81	60.33	3.81	50.18	86.60	1.56	79.55	76.16	2.36	54.00
	0.01	91.00	6.72	83.81	77.50	1.09	66.36	62.16	5.09	54.72	86.50	1.71	80.76	74.50	1.63	49.45
	0.05	89.50	39.45	56.00	79.67	0.72	71.45	56.00	25.81	42.90	88.93	8.14	80.49	69.33	0.18	21.63
	0.1	89.33	61.27	36.00	80.00	0.72	71.91	66.50	18.18	52.54	87.91	54.09	39.85	64.33	0.00	13.27
	0.2	87.66	75.09	22.18	82.50	0.91	70.91	59.33	7.45	53.09	81.98	92.63	6.20	55.16	0.00	11.27
LowFreq	0.005	92.66	67.09	30.18	76.33	0.18	26.36	67.00	12.18	35.81	86.84	98.75	1.09	76.00	0.36	16.00
	0.01	91.00	80.90	17.63	78.00	1.27	32.72	66.33	39.45	29.09	86.67	98.90	0.93	75.67	0.18	11.45
	0.05	90.00	96.36	3.63	77.16	0.55	28.36	58.67	0.00	9.63	85.68	100.0	0.00	75.50	0.00	9.63
	0.1	90.67	99.27	0.55	80.50	0.36	26.91	34.67	0.00	9.27	84.20	100.0	0.00	77.00	0.18	9.09
	0.2	90.16	99.09	0.91	81.83	0.55	35.27	32.50	0.00	9.09	83.86	100.0	0.00	75.50	0.18	9.27
Ftrojan	0.005	91.83	0.90	91.81	76.16	0.55	78.54	64.50	0.30	67.63	7.81	100.0	0.00	75.33	0.36	67.81
	0.01	91.16	8.18	84.72	77.50	0.36	78.55	43.83	0.00	48.18	87.21	1.71	85.7	74.16	0.55	49.45
	0.05	90.33	100.0	0.00	79.50	0.36	80.91	37.00	0.00	40.72	89.00	1.09	88.57	69.83	0.00	9.63
	0.1	89.66	100.0	0.00	78.83	0.18	81.45	37.16	0.00	41.09	88.40	1.36	88.30	76.16	0.00	9.63
	0.2	89.00	100.0	0.00	82.50	0.18	85.27	26.00	0.70	28.00	81.96	13.15	78.38	73.67	0.00	9.27
TITIM	0.005	91.50	0.55	91.63	74.83	2.72	74.72	63.50	3.09	63.09	84.48	1.40	84.67	76.16	0.72	76.72
	0.01	90.83	0.91	90.72	77.66	0.36	78.72	68.66	5.27	67.27	85.55	0.63	85.02	73.83	0.72	74.00
	0.05	91.00	1.27	91.09	79.18	0.18	80.73	37.66	0.00	40.90	88.25	1.25	86.68	67.00	0.72	64.72
	0.1	87.66	26.90	65.81	80.66	0.18	83.45	28.50	3.27	30.36	89.48	1.87	88.46	42.67	0.54	40.00
	0.2	88.50	100.0	0.00	80.83	0.55	83.09	48.67	3.09	51.27	84.33	99.84	0.15	22.66	0.72	15.81
Average		**90.63**	68.33	29.10	78.92	0.60	**70.08**	54.94	20.77	38.25	79.13	52.87	40.65	67.35	**0.43**	30.42

Table 4: Reference set size ablation on CIFAR-10 and GTSRB.

Dataset	Ref. Size	P. Ratio	BadNet			Blended			SSBA			Refool			LowFreq			Ftrojan			TITIM		
			C-Acc ↑	ASR ↓	RA ↑	C-Acc ↑	ASR ↓	RA ↑	C-Acc ↑	ASR ↓	RA ↑	C-Acc ↑	ASR ↓	RA ↑	C-Acc ↑	ASR ↓	RA ↑	C-Acc ↑	ASR ↓	RA ↑	C-Acc ↑	ASR ↓	RA ↑
GTSRB	5%	0.005	96.95	0.00	97.04	97.44	0.70	56.21	97.33	0.00	95.79	97.44	0.01	67.98	97.63	0.00	63.34	97.27	0.00	97.35	97.52	0.00	97.44
		0.01	96.72	0.00	96.78	96.94	0.07	63.89	97.36	0.02	95.18	97.36	0.01	69.35	97.41	0.00	64.08	97.26	0.00	97.36	97.45	0.00	97.56
		0.05	97.66	0.00	97.75	97.26	2.20	60.07	97.21	0.00	95.81	97.52	0.01	66.34	97.64	0.00	52.72	97.33	0.00	97.26	97.17	0.01	97.22
		0.10	97.67	0.00	97.71	97.14	0.91	56.49	97.71	0.00	96.85	97.76	0.07	67.93	97.15	0.00	57.15	97.86	0.00	97.93	96.98	0.00	97.19
		0.20	97.71	0.00	97.84	97.51	0.16	60.41	97.67	0.00	95.78	97.79	0.05	68.42	97.28	0.00	63.02	96.84	0.00	96.87	97.69	0.00	97.88
	10%	0.005	98.07	0.00	98.08	97.76	0.05	57.22	97.52	0.02	95.96	97.61	0.02	69.96	97.85	0.07	68.08	97.29	0.00	97.41	98.16	0.00	97.09
		0.01	97.68	0.00	97.69	97.02	2.25	59.92	97.41	0.00	95.99	97.77	0.04	69.77	97.48	0.18	66.17	97.43	0.00	97.40	97.56	0.00	97.66
		0.05	97.55	0.00	97.51	97.71	4.15	56.56	97.25	0.00	95.33	97.98	0.07	67.93	97.45	0.50	62.90	97.33	0.00	97.35	97.44	0.00	97.46
		0.10	98.28	0.00	98.29	97.94	1.23	60.70	96.84	0.02	95.24	98.08	0.06	68.53	97.66	0.00	67.55	97.39	0.00	97.37	98.02	0.00	97.93
		0.20	98.00	0.00	98.01	98.06	3.73	59.49	98.24	0.01	96.22	98.21	0.11	68.67	97.86	0.03	63.10	97.98	0.00	97.92	97.81	0.00	97.85
CIFAR-10	5%	0.005	88.05	0.18	91.45	88.24	1.01	73.94	88.20	0.24	88.86	88.36	0.18	83.67	88.16	0.91	86.83	88.14	0.35	91.26	88.16	0.26	84.91
		0.01	89.45	0.32	91.05	90.82	9.22	71.22	89.62	0.34	89.58	90.30	0.22	84.00	90.59	2.96	86.96	90.10	0.52	90.54	91.69	1.90	82.04
		0.05	89.75	0.15	92.66	88.88	0.40	74.06	89.42	0.21	90.66	89.79	0.16	86.28	89.44	0.31	89.65	88.80	0.13	92.26	89.82	0.81	85.08
		0.10	90.10	0.09	93.24	90.31	0.67	77.10	90.46	0.18	91.55	90.61	0.20	87.05	90.28	0.44	91.05	89.97	0.20	93.35	90.31	0.71	85.18
		0.20	90.17	0.15	93.17	89.85	0.20	78.45	89.11	0.11	91.31	89.72	0.18	86.90	89.89	0.28	90.54	89.22	0.17	92.83	90.08	0.38	85.90
	10%	0.005	90.18	0.51	91.22	90.66	1.96	75.75	90.46	0.62	90.03	91.08	0.32	84.98	90.42	1.00	89.48	91.54	0.74	91.43	90.51	2.61	82.36
		0.01	88.34	0.57	89.33	88.43	1.98	74.44	88.70	0.71	88.01	90.32	0.33	85.75	88.10	1.37	86.65	89.10	0.81	89.82	89.51	3.90	83.52
		0.05	90.64	0.23	92.95	90.85	0.80	78.25	91.19	0.24	91.50	90.83	0.25	85.83	90.92	0.57	90.73	91.08	0.27	93.23	90.94	0.43	85.50
		0.10	91.66	0.24	93.57	92.05	0.78	78.02	91.82	0.22	91.77	91.37	0.12	87.52	91.94	0.85	90.37	91.94	0.24	93.61	91.80	1.28	87.21
		0.20	91.21	0.25	93.26	90.72	0.87	77.38	91.26	0.33	91.44	91.22	0.45	86.82	91.07	0.72	90.67	91.36	0.30	93.56	90.86	1.17	85.90

reported in seconds. Detailed implementation and training settings for each defense are provided in appendix section B.

ABL achieves the lowest total runtime on GTSRB and ImageNet-12, while DAR is the fastest method on CIFAR-10 and the second fastest on the other two datasets. Despite its higher runtime than ABL on GTSRB and ImageNet-12, DAR provides a substantially better defense-performance tradeoff. On GTSRB, DAR achieves 97.71 C-Acc, 0.36 ASR, and 83.15 RA, compared with 59.96 C-Acc, 50.04 ASR, and 35.44 RA for ABL. Similarly, on ImageNet-12, DAR achieves 78.92 C-Acc, 0.60 ASR, and 70.08 RA, compared with 54.94 C-Acc, 20.77 ASR, and 38.25 RA for ABL. DAR is also substantially faster than DBD across all three datasets and requires less runtime than ESTI. DBD incurs the largest computational cost, primarily due to its semi-supervised training stage, while ESTI requires multiple auxiliary training stages. Overall, DAR provides a favorable balance between defense effectiveness and practical runtime.

4.5 Adaptive Attack Evaluation

An important property of a backdoor defense is its robustness against adaptive attackers who are aware of the defense mechanism. We therefore construct a defense-aware adaptive attack specifically targeting DAR. We consider a stronger attacker than assumed in our threat model by granting access to both the feature extractor and the trusted clean reference set. Similar to BadNet (Gu et al., 2019), the attack uses a patch-based trigger, but optimizes the patch so that the feature representations of poisoned samples move closer to those of the target-class reference samples. The goal is to reduce the distributional discrepancy exploited by DAR, potentially allowing poisoned samples to receive larger weights during distribution alignment. After generating the adaptive trigger, poisoning is

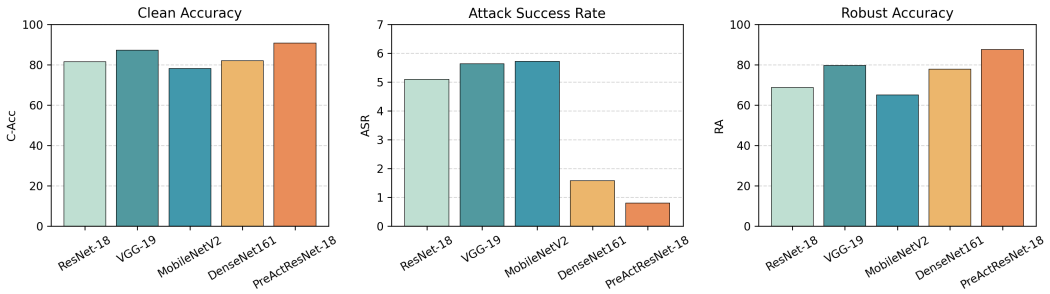

Figure 3: Average performance of DAR across different model architectures on CIFAR-10. Results are averaged over all attack and poison-ratio settings.

Table 5: Average wall-clock runtime over five runs (seconds).

Method	Stage	CIFAR-10 Total	CIFAR-10 Sec./Epoch	GTSRB Total	GTSRB Sec./Epoch	ImageNet-12 Total	ImageNet-12 Sec./Epoch
DAR	Feature Extractor Training	622.49	6.22	479.70	9.59	12,575.71	41.91
	Generating Weights	195.05		582.89		584.60	
	Training Robust Model	626.66	6.27	516.83	10.33	11,979.74	39.93
	Total	**1,444.20**		1,579.42		25,140.04	
ABL	Pre-Train	150.22	7.51	118.83	11.88	1,774.54	44.36
	Isolate Data	108.91		116.91		249.47	
	Finetuning/Unlearning	1,222.02	15.28	868.00	34.72	2,505.42	71.58
	Total	1,481.15		**1,103.74**		**4,529.43**	
DBD	Feature Extractor	2,771.89	14.58	3,522.93	35.23	13,736.09	137.36
	Semi-Supervised Training	12,051.60	120.52	10,742.13	107.42	104,942.22	1,049.42
	Total	14,823.49		14,265.05		118,678.31	
ESTI	Pseudo-Clean Model Training	1,634.08	8.17	1,092.52	5.46	5,035.85	25.18
	Poison Model Training	571.29	3.81	528.57	3.52	3,517.14	23.44
	Dataset Splitting	55.79	0.56	40.72	0.40	980.76	9.88
	Trap Model Training	2,814.63	7.51	2,617.58	6.98	16,672.78	47.64
	Total	5,075.78		4,279.39		26,206.53	

performed following the standard BadNet procedure. Details of the trigger optimization algorithm and example adaptive triggers are provided in appendix section D.

As shown in Table 6, DAR remains effective under the adaptive attack across all three datasets and poison ratios. The average ASR remains below 0.4% on all three datasets, while DAR maintains high C-Acc and RA. In particular, the average ASR is 0.37, 0.00, and 0.40 on CIFAR-10, GTSRB, and ImageNet-12, respectively. These results indicate that explicitly reducing the feature discrepancy between poisoned and clean reference samples is insufficient to circumvent DAR under the evaluated settings.

Table 6: DAR performance under the adaptive attack.

P. Ratio	CIFAR-10 C-Acc ↑	CIFAR-10 ASR ↓	CIFAR-10 RA ↑	GTSRB C-Acc ↑	GTSRB ASR ↓	GTSRB RA ↑	ImageNet-12 C-Acc ↑	ImageNet-12 ASR ↓	ImageNet-12 RA ↑
0.005	90.76	0.84	91.32	98.18	0.00	98.16	74.46	0.55	78.55
0.01	89.20	0.25	90.92	97.74	0.00	97.93	78.16	0.36	78.73
0.05	91.56	0.27	93.47	97.45	0.00	97.51	76.00	0.72	77.63
0.10	91.37	0.21	93.32	98.29	0.00	98.28	80.33	0.18	81.63
0.20	91.21	0.27	93.13	98.26	0.00	98.26	82.83	0.18	84.63

5 Conclusion

We presented Distribution-Aligned Reweighting (DAR), an in-training backdoor defense that uses a trusted clean reference set to guide class-wise distribution alignment. DAR estimates sample weights through kernel mean matching and incorporates them into standard model training through weighted sampling, without relying on assumptions about differences in the learning dynamics of clean and poisoned samples. Experiments on CIFAR-10, GTSRB, and ImageNet-12 demonstrate low average attack success rates and the highest average robust accuracy among the evaluated defenses. Additional analyses examine the effects of reference set size and model architecture, characterize the runtime–performance tradeoff, and demonstrate effectiveness against the evaluated adaptive attack. These results support distribution-guided reweighting as a practical approach to backdoor defense.

AI USE STATEMENT

In this work, we used generative AI tools to assist with drafting and polishing portions of the manuscript based on author-provided descriptions and human-written drafts, and to assist with debugging research code. We did not use generative AI tools to develop the research ideas, methodology, experimental design, analysis, or conclusions of this work. All AI-assisted text was reviewed and revised by the authors, and AI-assisted code was verified and tested for correctness. We take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

To facilitate reproducibility, we provide the source code, experimental settings, and scripts used to conduct the experiments. In this work, we used the BackdoorBench (Wu et al., 2022) repository for carrying out the attacks and defenses used in the experiments. The provided code is compatible with the BackdoorBench and can be easily integrated into the repository to ensure reproducibility.

REFERENCES

Mauro Barni, Kassem Kallas, and Benedetta Tondi. A new backdoor attack in cnns by training set corruption without label poisoning. In *IEEE International Conference on Image Processing*, pp. 101–105. IEEE, 2019.

Alfredo Canziani, Adam Paszke, and Eugenio Culurciello. An analysis of deep neural network models for practical applications. *arXiv preprint arXiv:1605.07678*, 2016.

Nicholas Carlini, Matthew Jagielski, Christopher A Choquette-Choo, Daniel Paleka, Will Pearce, Hyrum Anderson, Andreas Terzis, Kurt Thomas, and Florian Tramèr. Poisoning web-scale training datasets is practical. In *IEEE Symposium on Security and Privacy*, pp. 407–425, 2024.

Xinyun Chen, Chang Liu, Bo Li, Kimberly Lu, and Dawn Song. Targeted backdoor attacks on deep learning systems using data poisoning. *arXiv preprint arXiv:1712.05526*, 2017.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255. IEEE, 2009.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pp. 4171–4186, 2019.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021.

Micah Goldblum, Dimitris Tsipras, Chulin Xie, Xinyun Chen, Avi Schwarzschild, Dawn Song, Aleksander Madry, Bo Li, and Tom Goldstein. Dataset security for machine learning: Data poisoning, backdoor attacks, and defenses. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(2):1563–1580, 2022.

Arthur Gretton, Alex Smola, Jiayuan Huang, Marcel Schmittfull, Karsten Borgwardt, and Bernhard Schölkopf. Covariate shift by kernel mean matching. In Joaquin Quiñonero-Candela, Masashi Sugiyama, Andreas Schwaighofer, and Neil D. Lawrence (eds.), *Dataset Shift in Machine Learning*, pp. 131–160. MIT Press, 2008.

Tianyu Gu, Brendan Dolan-Gavitt, and Siddharth Garg. Badnets: Identifying vulnerabilities in the machine learning model supply chain. *IEEE Access*, 7:47230–47244, 2019.

Jonathan Hayase, Weihao Kong, Raghav Somani, and Sewoong Oh. Spectre: defending against backdoor attacks using robust statistics. In *International Conference on Machine Learning*, volume 139, pp. 4129–4139, 18–24 Jul 2021.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2016a.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Identity mappings in deep residual networks. In *European Conference on Computer Vision*, pp. 630–645. Springer, 2016b.

Gao Huang, Zhuang Liu, Laurens van der Maaten, and Kilian Q. Weinberger. Densely connected convolutional networks. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

Kunzhe Huang, Yiming Li, Baoyuan Wu, Zhan Qin, and Kui Ren. Backdoor defense via decoupling the training process. In *International Conference on Learning Representations*, 2021.

Takafumi Kanamori, Shohei Hido, and Masashi Sugiyama. A least-squares approach to direct importance estimation. *Journal of Machine Learning Research*, 10:1391–1445, 2009.

Alex Krizhevsky and Geoffrey Hinton. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009.

Yao Li, Sarah C. Van Alsten, Dong Neuck Lee, Taebin Kim, Benjamin C. Calhoun, Charles M. Perou, Sara E. Wobker, J. S. Marron, Katherine A. Hoadley, and Melissa A. Troester. Visual intratumor heterogeneity and breast tumor progression. *Cancers*, 16(13), 2024. ISSN 2072-6694.

Yige Li, Xixiang Lyu, Nodens Koren, Lingjuan Lyu, Bo Li, and Xingjun Ma. Anti-backdoor learning: Training clean models on poisoned data. *Advances in Neural Information Processing Systems*, 34:14900–14912, 2021a.

Yiming Li, Tong Li, et al. Reflection backdoor: A natural backdoor attack on deep neural networks. In *European Conference on Computer Vision*, 2020.

Yuezan Li, Yiming Li, Baoyun Wu, Longkang Li, Ran He, and Siwei Lyu. Invisible backdoor attack with sample specific triggers. In *International Conference on Computer Vision*, 2021b.

Chenhao Lin, Chenyang Zhao, Shiwei Wang, Longtian Wang, Chao Shen, and Zhengyu Zhao. Revisiting Training-Inference trigger intensity in backdoor attacks. In *USENIX Security Symposium*, pp. 6359–6378. USENIX Association, 2025.

Li Liu, Gang Feng, Denis Beautemps, and Xiao-Ping Zhang. Re-synchronization using the hand preceding model for multi-modal fusion in automatic continuous cued speech recognition. *IEEE Transactions on Multimedia*, 23:292–305, 2020.

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *International Conference on Computer Vision*, 2021.

Rui Min, Zeyu Qin, Li Shen, and Minhao Cheng. Towards stable backdoor purification through feature shift tuning. In *Advances in Neural Information Processing Systems*, volume 36, pp. 75286–75306, 2023.

XuanLong Nguyen, Martin J. Wainwright, and Michael I. Jordan. Estimating divergence functionals and the likelihood ratio by convex risk minimization. *IEEE Transactions on Information Theory*, 56(11):5847–5861, November 2010. ISSN 1557-9654. doi: 10.1109/tit.2010.2068870. URL `http://dx.doi.org/10.1109/TIT.2010.2068870`.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pp. 8748–8763, 2021.

Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

Virat Shejwalkar, Amir Houmansadr, Peter Kairouz, and Daniel Ramage. Back to the drawing board: A critical evaluation of poisoning attacks on production federated learning. In *IEEE Symposium on Security and Privacy*, pp. 1354–1371, 2022.

Hidetoshi Shimodaira. Improving predictive inference under covariate shift by weighting the log-likelihood function. *Journal of Statistical Planning and Inference*, 90(2):227–244, 2000.

Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations*, 2015.

Johannes Stallkamp, Marc Schlipsing, Jan Salmen, and Christian Igel. The german traffic sign recognition benchmark: A multi-class classification competition. In *Proceedings of the 2011 International Joint Conference on Neural Networks (IJCNN)*, pp. 1453–1460. IEEE, 2011.

Brandon Tran, Jerry Li, and Aleksander Madry. Spectral signatures in backdoor attacks. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

Alexander Turner, Dimitris Tsipras, and Aleksander Madry. Clean-label backdoor attacks. 2018.

Bolun Wang, Yuanshun Yao, Shawn Shan, Huiying Li, Bimal Viswanath, Haitao Zheng, and Ben Y Zhao. Neural cleanse: Identifying and mitigating backdoor attacks in neural networks. In *IEEE Symposium on Security and Privacy*, pp. 707–723, 2019.

Tong Wang, Yuan Yao, Feng Xu, Shengwei An, Hanghang Tong, and Ting Wang. An invisible black-box backdoor attack through frequency domain. In *European Conference on Computer Vision*, pp. 396–413. Springer, 2022.

Baoyuan Wu, Hongrui Chen, Mingda Zhang, Zihao Zhu, Shaokui Wei, Danni Yuan, and Chao Shen. Backdoorbench: A comprehensive benchmark of backdoor learning. In *Advances in Neural Information Processing Systems*, 2022.

Dongxian Wu and Yisen Wang. Adversarial neuron pruning purifies backdoored deep models. In *Advances in Neural Information Processing Systems*, volume 34, pp. 16913–16925, 2021.

Hongrui Yu, Lu Qi, Wanyu Lin, Jian Chen, Hailong Sun, and Chengbin Sun. Backdoor defense via enhanced splitting and trap isolation. In *International Conference on Computer Vision*, pp. 1708–1717, 2025.

Yi Zeng, Won Park, Z Morley Mao, and Ruoxi Jia. Rethinking the backdoor attacks' triggers: A frequency perspective. In *International Conference on Computer Vision*, 2021.