Parameter-Efficient Dual-Loss Adaptation with Logit Divergence: A Unified Approach for Adversarial Example Detection and Robust Inference

Zirui Fu , Marco Donato
Tufts University
{zirui.fu, marco.donato}@tufts.edu

Abstract

We present D3Adapter, a threat-aware framework that unifies adversarial example detection (AED) and robust inference. D3Adapter attaches a small library of lightweight adapters to a frozen ResNet backbone; the adapters are trained under two loss functions (cross-entropy and optimized reverse cross-entropy) and complementary objectives (clean training and adversarial training), yielding intentionally distinct logit behaviors. During inference, D3Adapter quantifies inter-adapter logit divergence to produce an agreement-based threat score without any external detector for AED. The same pass also performs robust inference by outputting prediction from the selected adapter when an adversarial example is detected, therefore unifying detection and defense into single-pass forward computation. We evaluate D3Adapter under transfer-based and adaptive white-box attacks, and we study scalability across datasets with varying numbers of classes, showing that unified detection and robust inference can be achieved with predictable overhead proportional to the number of adapters.

1 Introduction

Deep neural networks (DNNs) underpin modern computer vision systems in safety-critical applications such as autonomous driving, medical imaging, and content moderation. Despite their strong performance, DNNs remain vulnerable to adversarial examples, where imperceptible input perturbations can induce incorrect predictions. In practical deployments, the challenge is not only to improve average-case accuracy, but also to make inference trustworthy under worst-case perturbations while respecting real-world constraints on latency, power, and computational complexity.

Existing defenses largely follow two approaches. Adversarial example detection (AED) aims to identify adversarial inputs at inference time and then reject, defer, or route them to a conservative pipeline. AED systems often introduce extra inference steps or additional detector modules, require dataset- or deployment-specific calibration, and can be fragile under adaptive attacks where the detector becomes part of the attack objective [Xu *et al.*, 2018; Metzen *et al.*, 2017; Liang *et al.*, 2018; Carlini and Wagner, 2017]. Robustness enhancement methodologies aims to preserve correct predictions within a perturbation budget by switching training objectives, regularizing or stabilizing inputs and activations, or enhancing model structural complexity. They are often costly to train, may compromise clean accuracy, and are not readily applicable to third-party models without retraining or structural changes [Madry *et al.*, 2018; Shafahi *et al.*, 2019; Tsipras *et al.*, 2019; Zhang *et al.*, 2019]. As a result, detection and robust inference are often implemented as separate workflows, further increasing engineering complexity and inference overhead when both are required [Meng and Chen, 2017].

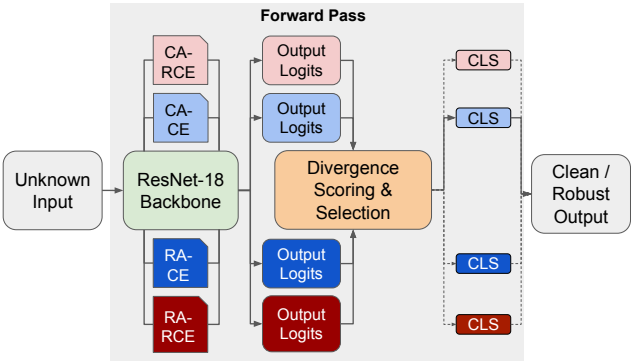

Figure 1: Overview of the D3Adapter framework. Given an input, the model executes inference through a frozen backbone with multiple attached adapters, producing multiple logits. Inter-adapter logit divergence is aggregated as an agreement-based threat score for adversarial example detection (orange). Its output routing mechanism selects the final prediction from the appropriate adapter according to the threat score.

To bridge detection and robustness, Reverse Cross-Entropy (RCE) [Pang *et al.*, 2018] has been explored as a principled mechanism to reshape logit geometry and improve separability for anomaly detection. The RCE-KDE paradigm, which trains a classifier with RCE and then attaches a kernel density estimation (KDE) detector, can yield strong detection performance under controlled settings. Nevertheless, it exhibits two practical obstacles. First, RCE optimization can be unstable and becomes harder to scale as the number of

classes increases. Second, it requires full retraining of the target model and relies on an external KDE module at inference time, which incurs storage cost, calibration effort, and additional deployment complexity. Recent work mitigates the retraining requirement by integrating stabilized RCE variants such as RCE with temperature scaling and weight focusing (RCE-TF) [Fu and Donato, 2025] coorindated with parameter-efficient adapters. However, this line of work still falls short of a detector-free and deployment-friendly unification of detection and robust inference.

This work departs from model architectures using external detectors and instead treats disagreement among adapters trained for different purposes as an endogenous adversarial detection signal. Loss functions and training objectives induce different decision boundaries and logit geometries. As a result, adapters trained under different losses and objectives tend to agree on the predicted label for benign inputs. However, the same set of adapters tends to disagree in systematic ways under adversarial perturbations. Therefore, we propose an inter-adapter logit divergence as an agreement-based threat score to select the most reliable prediction among adapters at inference time. In this way, detection and defense can be unified into a single inference primitive without an extra detector module.

In summary, we propose **D3Adapter**, a Divergence-aware Detection and Defense adapter framework that unifies AED and robust inference. D3Adapter attaches a small library of lightweight adapters to a frozen backbone; adapters are trained under two loss families (CE and RCE-TF, an optimized version of RCE) to target complementary objectives (clean training and adversarial training) with the goal of producing different logit behaviors. During inference, D3Adapter performs a single-pass forward computation through all adapters, quantifies inter-adapter divergence to obtain a threat score, and outputs the prediction from the appropriate adapter based on the outcome of the adversarial detection. The resulting overhead is predictable and scales linearly with the number of adapters, making the framework suitable for resource-constrained deployment. Our main contributions are as follows:

- **Unified inference primitive for detection and defense:** We introduce an agreement-based threat score via inter-adapter logit divergence and a same-pass routing rule that jointly provides adversarial example detection and robust inference without an external detector module.
- **Threat-aware robustness under strong attacks:** We evaluate D3Adapter under transfer-based and adaptive white-box attacks, demonstrating improved detection quality and robust accuracy while preserving clean performance.
- **Scalability and deployment efficiency:** We study scalability across datasets with varying numbers of classes, and show that unified detection and robust inference can be achieved with overhead proportional to the number of adapters. The framework also supports controlled ablations through interchangeable loss families, divergence metrics, and selection rules.

2 Background and Related Work

This section reviews four threads that motivate D3Adapter: (I) loss-based logit geometry shaping, (II) parameter-efficient fine-tuning, (III) threat models and robust evaluation practices, and (IV) divergence-based adversarial detection.

2.1 Loss Shaping for Robustness and Detectability

Standard training typically minimizes cross-entropy, encouraging highly confident predictions. Such overconfident logits can worsen calibration and amplify sensitivity to label noise and distributional outliers, motivating alternative losses that reshape probability geometry. Common strategies include probability tempering and smoothing, e.g., temperature scaling for calibration [Guo *et al.*, 2017] and label smoothing [Szegedy *et al.*, 2016], as well as reweighting hard examples such as focal loss [Lin *et al.*, 2017].

TRADES [Zhang *et al.*, 2019] is a representative surrogate-loss design that trains the model on adversarial examples but optimizes it with a tailored clean-and-adversarial target for clean-robust performance trade-off. Reverse Cross-Entropy (RCE) [Pang *et al.*, 2018] is proposed as a robust training objective to suppress incorrect representations, opposite to normal cross-entropy (CE) loss. RCE enables thresholding- and density-based out-of-order distribution detection. It is also used as a noise-tolerant counterpart of CE loss in Symmetric Cross Entropy [Wang *et al.*, 2019] to stabilize noisy label training. Recent work further modifies RCE-style objectives to improve adversarial robustness and scale on larger label spaces [Fu and Donato, 2025]. It couples optimized RCE-TF with lightweight adapters (introduced in Section 2.2) while the backbone model is frozen, avoiding full-model retraining.

These losses can be incorporated with adversarial training, a common method for robustness enhancement, as an alternative objective in the training process to further increase model's adversarial robustness thanks to their distinguished logits distribution. These collective efforts can deliberately shape model's logit geometry and representation structure, yielding complementary behaviors in calibration, density separation, and adversarial smoothness. We exploit these controllable behaviors to construct multiple, measurable prediction profiles for threat scoring and robust inference, and instantiate them via parameter-efficient fine-tuning, introduced in Section 2.2.

2.2 Parameter-Efficient Fine-Tuning with Adapters

Full-model fine-tuning is often memory- and compute-intensive. Parameter-Efficient Fine-Tuning (PEFT) addresses the problem by keeping a backbone frozen and training small additive modules. In computer vision domain, residual adapters [Rebuffi *et al.*, 2017] were introduced to steer a shared ResNet [He *et al.*, 2016] backbone towards correct inference over multiple domains via lightweight residual branches. Houlsby et al. introduced bottleneck adapters in a transformer architecture, allowing new language tasks to be added without retraining the full model [Houlsby *et al.*, 2019]. This modularity naturally enables maintaining

Dataset	Number of Classes	Adapter Size	CE Acc (%)	RCE-TF Acc (%)	Total Params (M)	Trainable Params (M)	Overhead (%)
CIFAR-10	10	128	90.59	**94.72**	12.11	0.95	7.80
SVHN	10	128	91.12	**92.39**	12.11	0.95	7.80
GTSRB	43	128	99.48	**99.87**	12.13	0.96	7.93
CIFAR-100	100	128	72.76	**77.86**	12.16	0.99	8.15
Tiny-ImageNet	200	128	**70.79**	62.15	12.21	1.04	8.54
Tiny-ImageNet	200	256	**70.87**	67.47	13.14	1.97	14.98

Table 1: Clean accuracy (%) of an ImageNet-pretrained ResNet-18 backbone with adapter-only fine-tuning under CE and RCE-TF losses across various datasets.

an adapter library and dynamically selecting or composing modules: AdapterHub [Pfeiffer *et al.*, 2020] demonstrates a practical ecosystem of pre-trained adapters, while AdapterFusion [Pfeiffer *et al.*, 2021] shows that multiple pretrained adapters can be combined via a lightweight fusion mechanism without modifying the backbone. Other PEFT families such as LoRA inject low-rank updates into linear layers [Hu *et al.*, 2022]. Although LoRA-style low-rank adaptation can be extended beyond Transformer models, we focus on adapter modules because they better support a compact library of independently trained adapter candidates for our victim ResNet-based model.

The paradigm of swapping adapters on a frozen backbone is particularly attractive for deployment as it enables multi-task adaptation without the need of replacing entire models. MEMTI [Donato *et al.*, 2019] and Adapter-ALBERT [Fu *et al.*, 2024] have shown that adapter-equipped networks can improve on-device efficiency by selectively updating only a small parameter subset in faster memory tiers. This modularity and previous AdapterHub result a key enabler for our work: maintaining a small adapter library with deliberately different robustness profiles makes it feasible to both measure disagreement and route to a robustness-oriented adapter under attack.

2.3 Adversarial Attacks and Evaluation

Adversarial attacks are commonly categorized by perturbation norm. L_∞ attacks include one-step FGSM [Goodfellow *et al.*, 2015] and multi-step PGD [Madry *et al.*, 2018], the latter serving as a strong baseline for first-order robustness. In this work, we will primarily use PGD as the main adversarial attacking algorithm. L_2 attacks are typified by C&W [Carlini and Wagner, 2017], and L_1 attacks include the elastic-net formulation EAD [Chen *et al.*, 2018]. Adversarial attacks can also be categorized by attacker's knowledge of the victim model. Transfer-based (black-box) attacks are naive perturbations that are generated by the attacker from the victim model and applied to other similar models. Adaptive (white-box) attacks are conducted when the attacker has knowledge to the victim model and can access its gradients and defense modules, if applicable.

A key lesson from prior work is that detection-style defenses are vulnerable under adaptive attackers: many detectors can be bypassed by optimizing an attack objective tailored to the detector [Carlini and Wagner, 2017], and defenses relying on gradient masking can appear robust while failing under stronger attacks [Athalye *et al.*, 2018]. We therefore report results under both transfer and adaptive settings for a comprehensive performance evaluation of our framework. We also craft a detector-aware adaptive PGD attack to optimize both the model itself and its detection mechanism to stress our divergence threat-scoring signals.

2.4 Divergence-Based Adversarial Detection

Multiple lines of work indicate that disagreement between predictive distributions correlates with epistemic risk under distribution shift. Deep Ensembles exploit member disagreement for uncertainty estimation [Lakshminarayanan *et al.*, 2017], and Bayesian approximations such as MC Dropout use stochastic forward passes to estimate predictive uncertainty [Gal and Ghahramani, 2016]. Consistency regularization methods such as AugMix [Hendrycks *et al.*, 2020] use Jensen-Shannon divergence (JSD) [Lin, 1991] across augmented views to encourage stable predictions.

Taken together, prior uncertainty and consistency literature supports using distributional disagreement as a risk proxy, and divergences such as JSD and symmetric KL (sKL) provide simple, fully differentiable summary statistics on predictive distributions. Motivated by this principle, we use an inter-prediction divergence score as the threat signal in our framework; the exact formulation and ablations are presented in Section 4.

3 Threat-Aware Adapter Library for One-Pass Detection and Defense

To unify AED and robust inference within a single inference procedure, D3Adapter relies on a small library of lightweight adapters attached to a frozen ResNet-18 backbone. The library is intentionally diversified along two axes: loss geometry (CE and RCE-TF) and training objective (clean training and adversarial training). This section first explains the construction of adapter library from an observation of RCE-TF clean performance recovery and the library's behavior under transfer-based and adaptive PGD attacks. The resulting robustness and failure mode divergence later becomes an intrinsic signal for one-pass detection and routing.

3.1 Balancing CE and RCE-TF Clean Accuracy

We start from a ResNet-18 backbone pretrained on ImageNet and extended with bottleneck adapters. For each downstream dataset, we freeze the backbone and train only the adapters

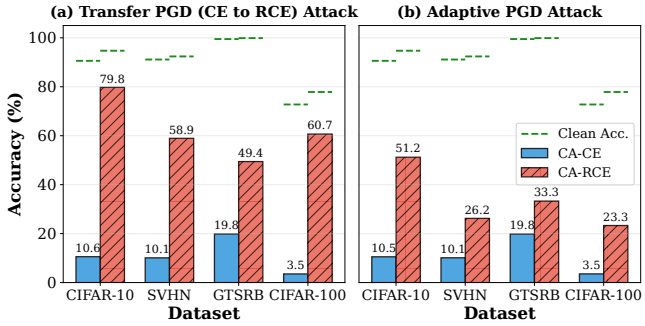

Figure 2: Robust accuracy comparison of CA-CE (blue) and CA-RCE (red) adapters under (a) transfer PGD attack crafted from CE model and (b) adaptive PGD attack crafted by their own loss functions. The dashed green line above each data is the clean accuracy of the corresponding model.

and the dataset-specific classification head. This setting keeps the adapter library lightweight, while ensuring that different candidates share the same feature extractor, which is required by a unified inference path.

A practical requirement is that the clean accuracy remains usable after introducing RCE-TF, since the library must serve both benign and adversarial inputs. Prior RCE-TF work [Fu and Donato, 2025] reports clean CIFAR-10 and CIFAR-100 accuracies with adapter size 64. In our adapter-only training regime, we start with adapter size 128 to improve accuracy and adjust to 256 when RCE-TF accuracy lags behind the CE counterpart. We explicitly measure the clean accuracy versus parameter overhead trade-off across datasets with different numbers of classes, shown in Table 1. It compares CE and RCE-TF adapters on CIFAR-10, SVHN, GTSRB, CIFAR-100, and Tiny-ImageNet, covering class counts from 10 to 200. The RCE-TF adapter outperforms CE adapters with the same size setting on datasets with up to 100 classes. Additionally, the parameter overhead grows proportional to the adapter size and number of classes of the dataset. The overhead becomes non-negligible when adapter size increases to 256, especially on Tiny-ImageNet where the result is atypical in that an RCE-TF adapter with size 256 still underperforms a CE adapter with size 128. On Tiny-ImageNet's CE adapters, increasing adapter size to 256 also does not produce a significant accuracy improvement. This outcome suggests that increasing adapter capacity alone is not sufficient to close the gap in the 200-class regime under a frozen backbone. A plausible explanation is that RCE-TF relies on a reversed logit geometry and a strong separation between the true class and many non-target classes. When the number of classes is larger, the CE/RCE-TF competition among their own targeted classes becomes more significant, and adapter-only training may not have enough strength to reshape representations to the extent required by the RCE-TF objective. In addition, Tiny-ImageNet differs from the ImageNet pretraining distribution and resolution, which can make adapter-only adaptation more sensitive to the choice of loss geometry. In our setting, this combination can limit the benefit of RCE-TF even when the adapter size is increased.

To stay within a conservative overhead and accuracy scenario, we use adapter size 128 as the default and instantiate two clean-trained adapter candidates: a **c**lean-trained **a**dapter by **CE** loss (**CA-CE**) and a **c**lean-trained **a**dapter by **RCE-TF** loss (**CA-RCE**). On CIFAR-100, CA-CE achieves 72.76% clean accuracy, while CA-RCE achieves 77.86%. For later experiments we focus on CIFAR-10, SVHN, GTSRB, and CIFAR-100, and we exclude Tiny-ImageNet due to its different scaling behavior under adapter-only training.

3.2 Transfer and Adaptive PGD Threats: Attack Protocols and Threat Signals

We evaluate the adapters under two threat models that cover both black-box and white-box adversarial attacks.

Transfer-based black-box PGD. In the transfer setting, the attacker crafts adversarial examples on a surrogate model and applies them to the target model without access to the target parameters. Figure 2(a) shows the accuracy of CA-CE and CA-RCE adapters when subjected to a transfer PGD attack with strength $\varepsilon = 8/255$ crafted from the CE model. CA-RCE adapter has shown stronger resilience on the four datasets than the CA-CE counterpart.

Adaptive white-box PGD. In the adaptive setting, the attacker has full access to the target model and adapts the attack to the target loss. For each target, we craft adversarial examples by directly optimizing the corresponding loss used by the target so that the attack is loss-consistent with the target geometry. This setting provides a stricter evaluation than transfer attacks because it removes the surrogate mismatch. Figure 2(b) shows the accuracy of CA-CE and CA-RCE adapters under adaptive PGD attack of the same strength as Figure 2(a). While CA-RCE still shows stronger robustness than CA-CE adapter, both clean-trained candidates are not resilient enough and lose practical reliability under loss-matched adaptive attacks. This necessitates adversarial training to enhance adapter candidates.

Robust candidates and divergence trends lead to Divergence AED. To obtain candidates that remain functional under adaptive attacks, we perform PGD adversarial training and obtain two additional adapters: a CE robust adapter (**RA-CE**) and an RCE-TF robust adapter (**RA-RCE**), under adaptive PGD $\varepsilon = 8/255$. Table 2 shows all four candidates' clean and robust accuracies under adaptive PGD $\varepsilon = 8/255$ in CIFAR-10, SVHN, GTSRB, and CIFAR-100 datasets.

Additionally, Figure 3 shows the robust accuracies of four pairs of RA-CE and RA-RCE adapters, each pair trained under PGD $\varepsilon = \{2, 4, 8, 16\}/255$, under transfer PGD attack with strength from 0 to 32/255. Noticeably, CE and RCE-TF adapters exhibit a systematic performance separation under the same perturbations: on benign inputs the heads tend to agree (left side of the X-axis), while the separation and disagreement become more pronounced as perturbation strength increases.

In summary, the four adapters from different loss-objective correlations exhibit different characteristics and strengths under clean and adversarial settings, and generates unique logits. No single candidate simultaneously optimizes clean utility, transfer robustness, and loss-matched adaptive robustness. This trend suggests that cross-adapter differences in

Dataset	Adapter	Clean / PGD Adv Acc (%)
CIFAR-10	CA-CE	90.59 / 10.53
	CA-RCE	**94.72** / 51.22
	RA-CE	89.08 / 73.25
	RA-RCE	93.47 / **80.40**
SVHN	CA-CE	91.12 / 10.13
	CA-RCE	**92.39** / 26.23
	RA-CE	87.77 / 46.02
	RA-RCE	90.93 / **54.98**
GTSRB	CA-CE	99.48 / 19.82
	CA-RCE	**99.87** / 33.30
	RA-CE	91.18 / 46.91
	RA-RCE	91.02 / **58.12**
CIFAR-100	CA-CE	72.76 / 3.54
	CA-RCE	**77.86** / 23.31
	RA-CE	68.46 / 46.31
	RA-RCE	72.67 / **51.52**

Table 2: Clean and Adversarial Performance of the Four Adapters on CIFAR-10, SVHN, GTSRB, and CIFAR-100. CA adapters show better clean accuracies and RA adapters show better adversarial accuracies in all datasets.

logits and predictions can serve as an intrinsic signal for AED, and Section 4 shows how this signal can guide defense decisions by measuring inter-adapter divergence.

4 Divergence-Aware Adapter Detection Mechanism

A representative RCE-style detection baseline is RCE-KDE [Pang *et al.*, 2018], which pairs an RCE (or RCE-variant) classifier with a KDE detector on pre-classification features. This design introduces an extra inference-time scoring module, which requires constant detector-specific calibration on KDE's bandwidth and threshold throughout the inference period, and expands the differentiable attack surface when the detector is included as objective for adaptive attacks. D3Adapter avoids an external detector and instead uses model-intrinsic disagreement signals produced by the adapter library, enabling one-pass scoring and robust prediction.

4.1 Inter-Adapter Calibration and Divergence Score

All four adapters run on a single frozen ResNet-18 backbone and output logits $z_i(x) \in \mathbb{R}^C$ for any given input x, where the subscript i refers to the i-th adapter and C is the number of classes in the current dataset. Adapter outputs must be normalized before any divergence measurement because CE and RCE-TF induce opposite logit orientations, and because different training objectives can yield different confidence scales.

Logit orientation. To enforce a consistent semantic direction where larger values indicate higher confidence, we apply a loss-family dependent orientation transform:

$$\tilde{z}_i(x) = \begin{cases} z_i(x), & \text{if adapter } i \text{ uses CE,} \\ -z_i(x), & \text{if adapter } i \text{ uses RCE-TF.} \end{cases} \quad (1)$$

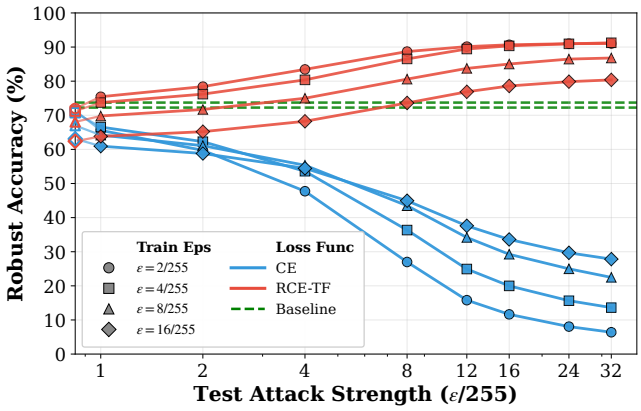

Figure 3: Accuracy inversion effect under transfer PGD attacks on CIFAR-100. CA-CE (blue) adapters are proportionally weaker as the attack strength increases, while CA-RCE (red) adapters are more robust facing stronger attacks. Hollow markers on the Y-axis indicate clean accuracy for each adversarially trained configuration.

Temperature scaling. Even after orientation, adapter confidences are not directly comparable. We therefore fit a per-adapter temperature τ_i on a small clean calibration split by minimizing cross-entropy: Following the concept of temperature scaling [Guo *et al.*, 2017], the calibrated posterior is $p_i(x) = \mathrm{softmax}(\tilde{z}_i(x)/\tau_i)$.

This temperature scaling calibration aligns predictions from different adapters so that divergence values reflect disagreement rather than scale artifacts. The calibration happens before the inference process and the same values are used for all other normalization process. This differs from KDE-related calibration, where frequent sampling for bandwidth selection and density-threshold selection are required to make feature-space density estimates operational as an anomaly score.

Divergence computation. Given calibrated posteriors, we compute pairwise divergences across adapters. For each (i, j) pair of adapters, we compute JSD:

$$\mathrm{JSD}(p_i \| p_j) = \tfrac{1}{2}\,\mathrm{KL}\big(p_i \,\|\, m\big) + \tfrac{1}{2}\,\mathrm{KL}\big(p_j \,\|\, m\big), \quad (2)$$

where $m = \frac{1}{2}\,(p_i + p_j)$ and additionally a sKL term:

$$\mathrm{sKL}(p_i, p_j) = \mathrm{KL}(p_i \| p_j) + \mathrm{KL}(p_j \| p_i). \quad (3)$$

With four adapters, this yields 6 pairwise JSD values and optionally 6 sKL values. To capture the structured separation observed in Section 3, we also compute group-level statistics, including mean divergence between the CE-family and the RCE-family, and within-family divergence between clean-trained and robust-trained variants.

Score fusion. All divergence features are standardized using the same data for temperature scaling calibration. The normalized features are then aggregated into a scalar divergence score $s(x)$. We use equal-weight fusion as a default:

$$s(x) = \mathrm{mean}\big(\hat{\phi}(x)\big), \quad (4)$$

and optionally consider a learned logistic fusion:

$$s(x) = \sigma\big(\mathbf{w}^\top \hat{\phi}(x) + b\big), \quad (5)$$

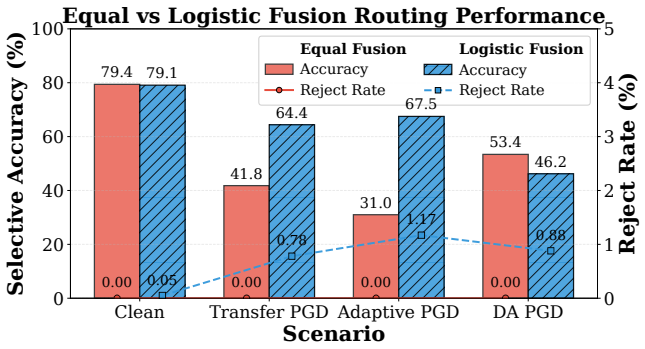

Figure 4: Fusion performance comparison between equal fusion (simple averaging) and logistic fusion (learned weights). Logistic fusion improves accuracy on adaptive attacks through selective rejection. Equal fusion remains zero rejection as all inputs are considered for fusion. Logistic fusion exhibits near-zero rejection on clean inputs (0.05%) and rejection rates below 1% across attack settings.

where $(\mathbf{w}, b)$ are trained on a balanced clean/adversarial subset. Figure 4 shows a performance comparison between fusion modes on clean data, transfer PGD, adaptive PGD, and detector-aware (DA) adaptive PGD attack which will be addressed in Section 4.3 as a stress test to our detection system.

4.2 Robust Prediction Routing Based on Threshold

Unifying detection and robust inference without introducing an extra inference module is beneficial for deployment. D3Adapter uses divergence score $s(x)$ to both detect suspicious inputs and select which adapter logits to trust for the final prediction.

There are two thresholds for the system: θ_{sus} (suspicious) and θ_{rej} (rejection). The thresholds are defined using the clean-score quantiles Q from the same calibration data used for temperature scaling and divergence feature standardization:

$$\theta_{\text{sus/rej}} = Q_q\big(s(x) \mid x \in \mathcal{D}_{\text{clean}}\big), \tag{6}$$

with typical $q = 0.98$ for θ_{sus} and $q = 0.995$ for θ_{rej}. This avoids tuning thresholds on a specific attack configuration.

Given $s(x)$ and thresholds $(\theta_{\text{sus}}, \theta_{\text{rej}})$, we perform one-pass routing:

- If $s(x) \le \theta_{\text{sus}}$: treat x as benign and output the prediction from CA-RCE, which has the highest clean accuracy among the four adapters according to Table 2.
- If $\theta_{\text{sus}} < s(x) \le \theta_{\text{rej}}$: treat x as suspicious and output the robust-trained candidate (RA-RCE) with higher confidence.
- If $s(x) > \theta_{\text{rej}}$: optionally abstain when even the selected robust adapter has low confidence. The rejection rate varies according to selected fusion methods and attacking schemes, depicted in Figure 4.

This design directly reuses the adapter logits computed for divergence scoring, and does not require any auxiliary detector forward pass.

4.3 Detector-Aware Adaptive Attack Evaluation

Detection-style defenses must be evaluated under adaptive attacks that differentiate through the full pipeline [Tramèr *et al.*, 2020]. We therefore adopt a detector-aware adaptive PGD objective that jointly maximizes misclassifications while minimizing the divergence score:

$$\max_{\|\delta\|_\infty \le \epsilon} \left[\sum_{i=1}^{4} w_i \,\text{CE}\big(\tilde{z}_i(x+\delta)/\tau_i,\, y\big) \;-\; \gamma\, s\big(x+\delta\big) \right], \tag{7}$$

where $\tilde{z}_i$ and τ_i follow Eq. (1) and temperature scaling, and γ trades off detector evasion against misclassification. We use multi-restart PGD with 5 restarts, 40 steps each, and retain the perturbation with the highest joint objective per sample.

4.4 Controlled Comparison with KDE Detection

For a controlled baseline, we implement a Gaussian KDE detector on the pre-classification features of the frozen backbone with a single adapter. Given clean training features $\{f_k\}_{k=1}^{n} \subset \mathbb{R}^d$, the KDE density estimate at test feature f is

$$\hat{p}(f) = \frac{1}{n\,h^d} \sum_{k=1}^{n} \exp\Big(-\tfrac{1}{2} \big\| \tfrac{f - f_k}{h} \big\|_2^2 \Big), \qquad s_{\text{KDE}}(f) = -\log \hat{p}(f), \tag{8}$$

with bandwidth h chosen by Silverman's rule [Silverman, 1986]. The detection threshold is set at the 5th percentile of training log-densities.

From Figure 5, under detector-aware PGD, where the attacker jointly minimizes the divergence score and maximizes misclassifications, D3Adapter retains strong separation (AUROC > 0.97) on CIFAR-10, SVHN, and GTSRB, whereas KDE degrades to near-random on SVHN (0.52) and remains below 0.65 on CIFAR-10. This gap reflects the structural difficulty of simultaneously suppressing disagreement across four intentionally diverse adapters while also inducing misclassifications.

4.5 Reduced-Adapter Divergence Design

Since executing multiple adapters incurs both parameter and latency overhead, a natural question is whether a smaller adapter set can preserve detection strength. We construct a three-adapter variant by removing RA-CE, the adapter with the narrowest applicable scenario based on the performance observed in Section 3. All other pipeline components, including divergence definition, score aggregation, and threshold calibration, remain unchanged.

Figure 6 compares the three- and four-adapter configurations for D3Adapter, and KDE. On CIFAR-10, the three-adapter variant achieves AUROC of 0.89 under transfer PGD, 0.95 under adaptive PGD, and 0.92 under detector-aware PGD, all within 0.06 of the four-adapter baseline. On GTSRB, the three-adapter variant reaches 0.99 AUROC under adaptive PGD, slightly exceeding the four-adapter result. However, under detector-aware PGD, the four-adapter configuration consistently outperforms the three-adapter variant: AUROC differences are 0.05 on CIFAR-10, 0.11 on SVHN, and 0.06 on GTSRB. This suggests that the fourth adapter,

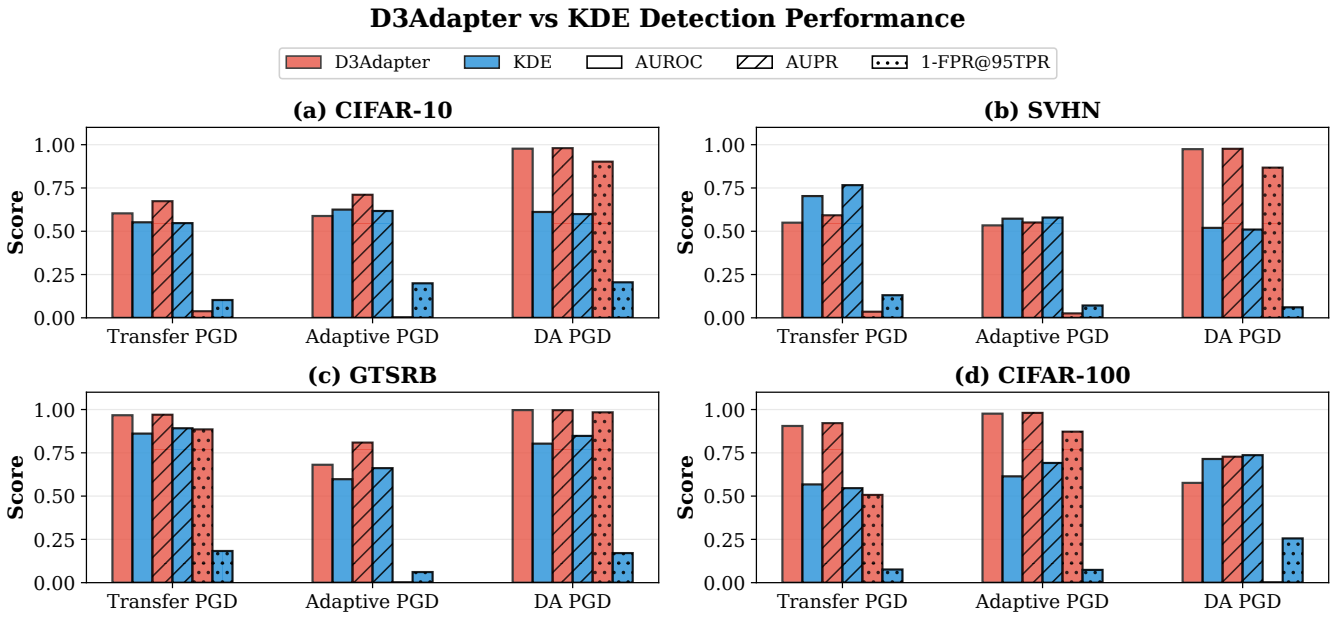

Figure 5: Detection performance comparison between D3Adapter and KDE across four datasets and three threat models. Bars show AUROC, AUPR, and $1 - \mathrm{FPR@95\%TPR}$ (higher is better). Under detector-aware PGD, D3Adapter achieves AUROC gains of 0.37 on CIFAR-10, 0.46 on SVHN, and 0.19 on GTSRB compared to KDE, indicating that multi-adapter divergence is harder for the attacker to suppress than single-feature density. On transfer and adaptive PGD, the two methods perform comparably, with D3Adapter slightly ahead on CIFAR-10 and GTSRB and KDE slightly ahead on SVHN.

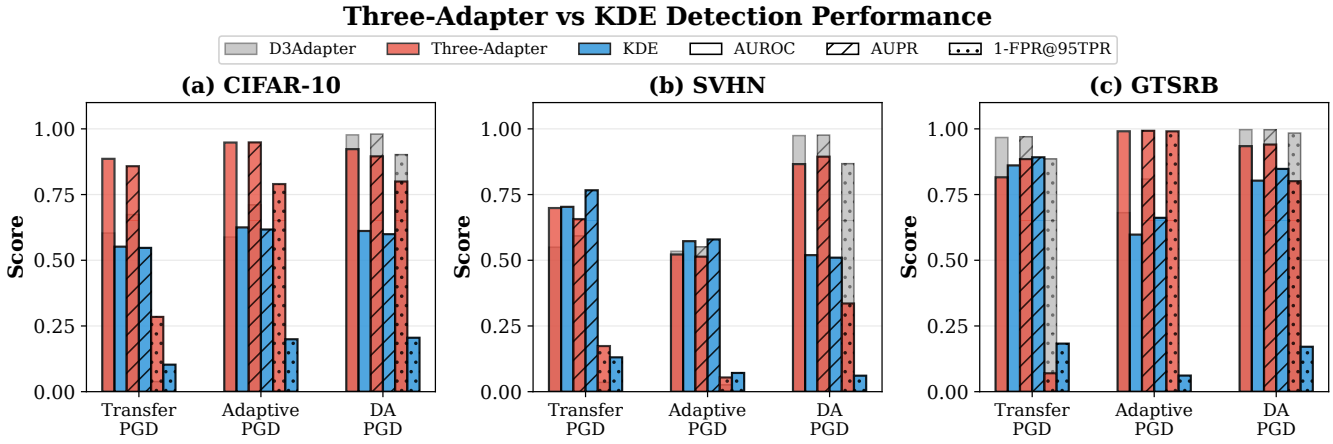

Figure 6: Comparison of detection performance between the full D3Adapter (four-adapter, gray background), the three-adapter variant (red foreground), and KDE (blue) on CIFAR-10, SVHN, and GTSRB. Under transfer and adaptive PGD, the three-adapter variant matches or exceeds the four-adapter baseline on several datasets. Under detector-aware PGD, retaining the fourth adapter yields consistent AUROC gains, confirming that additional adapter diversity raises the cost of evading detection.

despite limited standalone utility, provides an additional constraint that the attacker must satisfy when optimizing against the full detection objective.

5 Conclusion

D3Adapter is a threat-aware framework that unifies AED and robust inference within single forward computation on a frozen backbone, using a small library of lightweight adapters trained under CE and RCE-TF losses with clean and adversarial objectives. Inter-adapter logit divergence provides an agreement-based threat score without any external detector and also guides adapter selection for robust inference. D3Adapter achieves competitive detection quality and robust inference performance under transfer-based and adaptive white-box attacks.

Future work will focus on reducing runtime cost via adapter subset selection and layer-local deployment, improving the calibration of inter-adapter logit divergence for broader class scales, and extending the framework to additional threat models.

References

[Athalye *et al.*, 2018] Anish Athalye, Nicholas Carlini, and David Wagner. Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 274–283. PMLR, 2018.

[Carlini and Wagner, 2017] Nicholas Carlini and David Wagner. Adversarial examples are not easily detected: Bypassing ten detection methods. In *Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security (AISec)*, pages 3–14, 2017.

[Chen *et al.*, 2018] Pin-Yu Chen, Yash Sharma, Huan Zhang, Jinfeng Yi, and Cho-Jui Hsieh. Ead: Elastic-net attacks to deep neural networks via adversarial examples. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 2018.

[Donato *et al.*, 2019] Marco Donato, Lillian Pentecost, David Brooks, and Gu-Yeon Wei. Memti: Optimizing on-chip nonvolatile storage for visual multitask inference at the edge. *IEEE Micro*, 39(5):73–81, 2019.

[Fu and Donato, 2025] Zirui Fu and Marco Donato. Parameter-efficient adversarial example detection and robustness enhancement utilizing optimized reverse-cross entropy. In *2025 IEEE International Conference on AI and Data Analytics (ICAD)*, 2025.

[Fu *et al.*, 2024] Zirui Fu, Aleksandre Avaliani, and Marco Donato. Heterogeneous memory integration and optimization for energy-efficient multi-task nlp edge inference. In *Proceedings of the 29th ACM/IEEE International Symposium on Low Power Electronics and Design (ISLPED)*, pages 1–6, 2024.

[Gal and Ghahramani, 2016] Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *Proceedings of the 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pages 1050–1059. PMLR, 2016.

[Goodfellow *et al.*, 2015] Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In *International Conference on Learning Representations (ICLR)*, 2015.

[Guo *et al.*, 2017] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 1321–1330. PMLR, 2017.

[He *et al.*, 2016] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016.

[Hendrycks *et al.*, 2020] Dan Hendrycks, Norman Mu, Ekin D. Cubuk, Barret Zoph, Justin Gilmer, and Balaji Lakshminarayanan. Augmix: A simple data processing method to improve robustness and uncertainty. In *International Conference on Learning Representations (ICLR)*, 2020.

[Houlsby *et al.*, 2019] Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. Parameter-efficient transfer learning for nlp. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 2790–2799. PMLR, 2019.

[Hu *et al.*, 2022] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. In *International Conference on Learning Representations (ICLR)*, 2022.

[Lakshminarayanan *et al.*, 2017] Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

[Liang *et al.*, 2018] Shiyu Liang, Yixuan Li, and R. Srikant. Enhancing the reliability of out-of-distribution image detection in neural networks. In *International Conference on Learning Representations (ICLR)*, 2018.

[Lin *et al.*, 2017] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 2999–3007, 2017.

[Lin, 1991] Jianhua Lin. Divergence measures based on the shannon entropy. *IEEE Transactions on Information Theory*, 37(1):145–151, 1991.

[Madry *et al.*, 2018] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations (ICLR)*, 2018.

[Meng and Chen, 2017] Dongyu Meng and Hao Chen. Magnet: A two-pronged defense against adversarial examples. In *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security (CCS)*, pages 135–147, 2017.

[Metzen *et al.*, 2017] Jan Hendrik Metzen, Tim Genewein, Volker Fischer, and Bastian Bischoff. On detecting adversarial perturbations. In *International Conference on Learning Representations (ICLR)*, 2017.

[Pang *et al.*, 2018] Tianyu Pang, Chao Du, Yinpeng Dong, and Jun Zhu. Towards robust detection of adversarial examples. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018.

[Pfeiffer *et al.*, 2020] Jonas Pfeiffer, Andreas Rücklé, Clifton Poth, Aishwarya Kamath, Ivan Vulić, Sebastian Ruder, Kyunghyun Cho, and Iryna Gurevych. Adapterhub: A framework for adapting transformers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, 2020.

[Pfeiffer *et al.*, 2021] Jonas Pfeiffer, Aishwarya Kamath, Andreas Rücklé, Kyunghyun Cho, and Iryna Gurevych. Adapterfusion: Non-destructive task composition for transfer learning. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume (EACL)*, 2021.

[Rebuffi *et al.*, 2017] Sylvestre-Alvise Rebuffi, Hakan Bilen, and Andrea Vedaldi. Learning multiple visual domains with residual adapters. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

[Shafahi *et al.*, 2019] Ali Shafahi, Mahyar Najibi, Amin Ghiasi, Zheng Xu, John Dickerson, Christoph Studer, Larry S. Davis, Gavin Taylor, and Tom Goldstein. Adversarial training for free! In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

[Silverman, 1986] Bernard W. Silverman. *Density Estimation for Statistics and Data Analysis*. Chapman and Hall, 1986.

[Szegedy *et al.*, 2016] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jonathon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2818–2826, 2016.

[Tramèr *et al.*, 2020] Florian Tramèr, Nicholas Carlini, Wieland Brendel, and Aleksander Madry. On adaptive attacks to adversarial example defenses. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, 2020.

[Tsipras *et al.*, 2019] Dimitris Tsipras, Shibani Santurkar, Logan Engstrom, Alexander Turner, and Aleksander Madry. Robustness may be at odds with accuracy. In *International Conference on Learning Representations (ICLR)*, 2019.

[Wang *et al.*, 2019] Yisen Wang, Xingjun Ma, Zaiyi Chen, Yuan Luo, Jinfeng Yi, and James Bailey. Symmetric cross entropy for robust learning with noisy labels. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019.

[Xu *et al.*, 2018] Weilin Xu, David Evans, and Yanjun Qi. Feature squeezing: Detecting adversarial examples in deep neural networks. In *Network and Distributed System Security Symposium (NDSS)*, 2018.

[Zhang *et al.*, 2019] Hongyang Zhang, Yaodong Yu, Jiantao Jiao, Eric P. Xing, Laurent El Ghaoui, and Michael I. Jordan. Theoretically principled trade-off between robustness and accuracy. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 7472–7482. PMLR, 2019.