DATA-FREE ADVERSARIAL TRAINING: ACHIEVING ROBUSTNESS WITHOUT DATA AND ROBUST TEACHER

Anonymous authors
Paper under double-blind review

ABSTRACT

Deep neural networks are vulnerable to adversarial perturbations, while the most empirically reliable defenses (i.e., adversarial training and robustness distillation) typically require access to real training data and, in the latter case, an additional robust teacher. These extravagant assumptions hinder the deployment of existing defenses when sensitive data and robust teacher are unavailable, while only a naturally trained pretrained model can be accessed. To achieve adversarial robustness when only a naturally pre-trained model is available, we propose *Data-Free Adversarial Training* (DFAT), a robust post-training framework that requires neither real data nor a robust teacher. We theoretically show that a locally accurate student surrogate can reveal attack directions aligned with the natural teacher's local sensitivity, and that robust risk on real data can be controlled by optimizing teacher quality, robust anchoring error, and the discrepancy between synthetic and real distributions. Based on this, DFAT first performs data-free adversarial training to construct a natural student surrogate and learn a surrogate data distribution. It then uses the frozen natural teacher as a clean semantic oracle, and performs self-adversarial training on the student by maximizing and minimizing teacher-student output discrepancies over synthetic local neighborhoods. We further identify two additional challenges amplified in the data-free scenario, namely the fairness challenge and the generalization challenge, and propose to mitigate them from the perspectives of entropy reweighting and probability sparsity, respectively. To further mitigate the performance loss caused by surrogate distribution shift, we incorporate a test-time adaptation (TTA) module to further explore the self-adaptation potential of DFAT. Experiments and analyses demonstrate that DFAT achieves effective and reliable adversarial robustness in the practical setting with neither real data nor robust teacher.

1 INTRODUCTION

The flourishing development of deep learning provides high-performance solutions for modern artificial intelligence. However, existing works demonstrate the vulnerability of deep models to adversarial attacks: adding imperceptible perturbations to input data can cause a well-trained model to make completely wrong predictions Szegedy et al. (2013); Goodfellow et al. (2014), which hinders the reliable deployment of deep models.

As countermeasures, a large number of defense methods have been proposed. Among them, adversarial training Goodfellow et al. (2014); Madry et al. (2017); Jia et al. (2022a) is considered the most reliable empirical defense, where adversarial examples are generated online during the training process and added to the training set. The target model learns robust representations within a min-max optimization framework. Robustness distillation Goldblum et al. (2020); Zhu et al. (2021); Zi et al. (2021) additionally considers the efficiency issues of edge models and the limitations of adversarial training for small models, and proposes distilling the knowledge of a robust large model into a smaller edge model, with its main training framework also following the min-max paradigm.

Although effective, adversarial training and robustness distillation face a common practical challenge: both rely on the complete availability of the target distribution; robustness distillation additionally requires a robust teacher model. However, in practical applications, on the one hand, publicly available pre-trained models are often released without robustness guarantees, since per-

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.

forming robust pre-training on large-scale pre-training datasets is costly, making the acquisition of robust teacher models a luxury. On the other hand, sensitive original pre-training data (e.g., facial images Ren et al. (2022), pedestrian data Ye et al. (2021), medical images Rajpurkar et al. (2022), and financial data Cao (2022)) are often restricted due to privacy concerns and regulatory constraints, and are typically not released together with the pre-trained model, which makes both adversarial training and robustness distillation based on the real distribution a luxury. Therefore, although data-driven robustness training methods under the traditional fully white-box setting have proven effective in laboratory environments, they lack practical deployability. Data-free robustness distillation (DFRD) Zhou et al. (2024); Li et al. (2026) initially addresses the reliance on real data by mining the prior statistics stored in a robust teacher to generate a surrogate distribution, based on which robustness distillation is performed. However, DFRD still relies on the robust teacher as both a supervision signal and a data source. As we analyze, this remains a costly assumption. We provide a more detailed discussion of related work in Appendix A.

Building on existing research, we extend this to a new practical challenge: *Can edge users achieve robust training without relying on either a robust teacher or the real distribution?*

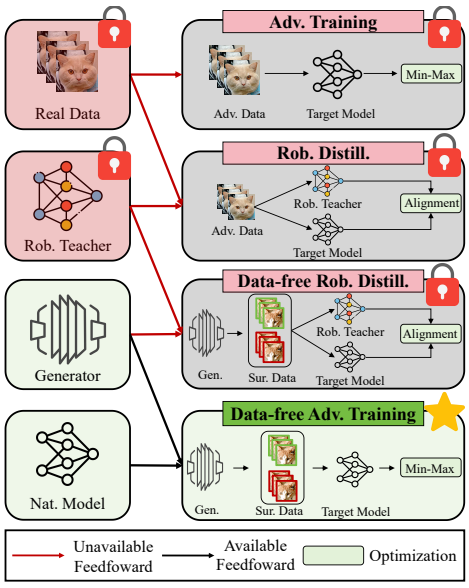


Figure 1: An intuitive comparison of DFAT with current adversarial training, robustness distillation, and data-free robust distillation.

To address the above challenge, we propose data-free adversarial training (DFAT), which aims to achieve robust post-training using only an available naturally pre-trained model. The intuitive difference between DFAT and existing defense methods is shown in Figure 1. Our core assumption is that, under the condition that the supervision signals of the natural model are preserved, a robust teacher may not be necessary, and the natural model can achieve robustness through self-adversarial training based on surrogate data. We prove two key arguments through theoretical derivation: (1) The adversarial direction of the student that is locally close enough to the teacher is similar to that of the teacher. (2) Adversarial training based on sufficient synthetic data that is similar to the real distribution can optimize the robustness upper bound of the target distribution.

Based on the above theory, we propose a two-stage data-free adversarial training model. In stage 1, we treat the naturally pre-trained model as a teacher model and obtain a student copy of the pre-trained model through a data-free knowledge distillation framework. This brings three benefits: (1) The generative model of the surrogate distribution learns the real distribution as closely as possible during stage 1. (2) Subsequent robust training is deployed on the student (target) model, avoiding the loss of the only available reliable supervision signal. (3) The student model learns natural knowledge in stage 1, which not only resembles the natural model as closely as possible but can also be regarded as natural pre-training for the student model, which is beneficial for robust training. In stage 2, the teacher model acts as a frozen oracle, providing reliable soft supervision signals to guide the self-adversarial training performed by the student model. Such a two-stage model is closely linked to the theoretical analysis: it preserves the reliable natural model, obtains a student model that is sufficiently similar to the natural model, and performs adversarial training based on surrogate data. Besides, we further identify two challenges that are amplified in data-free adversarial training as follows:

(a) *Class-wise fairness challenge.* Unlike existing methods that optimize fairness by improving the adversarial training process Sun et al. (2023); Li et al. (2026), we find that the fairness issue in data-free adversarial training actually stems from the fairness of the surrogate distribution, where the simple mean-entropy regularization overlooks samples from more difficult classes. We provide a detailed quantitative analysis of the fairness challenge in Appendix B.1. To mitigate this issue, we incorporate reweighting-based adaptive mean-entropy regularization in stage 1: instead of applying a fixed mean-entropy regularization loss throughout, we optimize the entropy constraint for each class based on the student model's confidence in that class. Specifically, classes with higher

confidence are considered to have relatively saturated knowledge, and thus we reduce the weight of the entropy constraint. Conversely, classes with lower confidence are considered under-learned, and we appropriately increase the entropy constraint weight.

(b) *Synthetic-to-real generalization challenge*. Generalizing from surrogate data to real data is an inherent challenge in data-free learning. We discuss generalization from a novel perspective, namely probability sparsity Zhou et al. (2026). Models with sparse output probabilities have been shown to exhibit better robustness, and are induced to produce interpretable probability distributions: categories semantically similar to the target class are assigned higher probabilities, while probabilities for other categories are approximately zero, which aligns with the intuitive interpretation of semantic generalization. However, we observe that the probability sparsity in data-free adversarial training is inferior compared to data-driven defenses, a phenomenon we define as "probability redundancy". We provide a more detailed explanation and analysis in Appendix B.2. Therefore, in stage 2, we introduce nasty regularization to induce sparse probability in the student model. Further, since DFAT performs training on a surrogate distribution, it inevitably suffers from a distribution shift relative to the real target distribution at the test stage. To further exploit its adaptive capability to unknown distributions, we incorporate an optional test-time adaptation (TTA) stage, which mitigates the performance loss caused by distribution shift while further unlocking DFAT's adaptive potential.

Our main contributions can be summarized as follows:

- We propose theoretical analyses of the adversarial direction and robustness upper bound of the surrogate strategy, based on which we propose data-free adversarial training, which achieves robust post-training without relying on either real data or a robust teacher.
- We identify two significant challenges in data-free adversarial training: the fairness challenge and the generalization challenge, and we mitigate these two issues from novel perspectives. The integration with TTA further unlocks the adaptive potential of DFAT.
- Extensive quantitative experiments and qualitative analyses demonstrate the effectiveness and reliability of our proposed method.

2 METHODOLOGY

2.1 PRELIMINARY

In this subsection, we briefly formalize adversarial attacks and defenses, and provide an initial definition of the main notations used in this paper.

Adversarial attack. Given a well-trained model F_θ parameterized by θ, the objective of the attacker can be formulated as a conditional optimization problem to fool the target model:

$$\arg\max_{\delta_x} \mathcal{L}(F_\theta(x+\delta_x,\theta),y),\ \text{s.t.}\ \|\delta_x\|_p \le \varepsilon, \tag{1}$$

where (x, y) is a data-label pair sample from the joint distribution $\mathcal{X}$, and δ_x denotes the adversarial perturbation, $\mathcal{L}$ represents the target loss (e.g., cross entropy loss), and ε is the upper bound of the perturbation under the l_p-norm (i.e., attack budget).

Adversarial defense aims to maintain the discriminability capability of F_θ under adversarial attacks, i.e., adversarial robustness. Adversarial training is considered to be the most reliable empirical defense, which can be formalized as:

$$\min_\theta \mathbb{E}_{(x,y)\sim\mathcal{X}}\left[\max_{||\delta_x||_p\le\varepsilon}\mathcal{L}(F(x+\delta_x,y),\theta)\right]. \tag{2}$$

Data-free adversarial defense. Current mainstream data-free defense models follow the paradigm of "pseudo-data generation + robust teacher-guided distillation". Although they get rid of the dependence on real datasets, they still rely on the supervision of a robust teacher. We aim to achieve robust training based solely on a natural model without either real datasets or robust teachers.

2.2 DATA-FREE ADVERSARIAL TRAINING

In this subsection, we introduce the proposed DFAT, which mainly includes natural pre-training, robust training, and an optional test-time adaptation stage.

2.2.1 STAGE 1: VANILLA SELF-DISTILLATION AS PRE-TRAINING

We first use the pre-trained natural model F parameterized by θ_F as the teacher to obtain a student copy S parameterized by θ_S under the data-free natural knowledge distillation framework. This can bring several practical benefits: (1) Subsequent training will be directly deployed on the student copy S, avoiding the loss of the only available supervision signal F. (2) Performing natural pre-training on the student model before robust training allows the model to acquire better representation capabilities Suzuki et al. (2023). We follow a similar natural knowledge distillation stage as in DERD Zhou et al. (2024) for pre-training, and the overall loss function can be formulated as:

$$\mathcal{L}_{\text{stage}_1} = \mathcal{L}^G_{\text{stage}_1} + \mathcal{L}^S_{\text{stage}_1}, \tag{3}$$

where $\mathcal{L}^G_{\text{stage}_1}$ and $\mathcal{L}^S_{\text{stage}_1}$ represent the loss of the generator G and student copy S in stage 1. The specific form of $\mathcal{L}^G_{\text{stage}_1}$ is:

$$\mathcal{L}^G_{\text{stage1}} = \lambda_{\text{oh}}\mathcal{L}_{\text{oh}} + \lambda_{\text{a}}\mathcal{L}_{\text{a}} + \lambda_{\text{BN}}\mathcal{L}_{\text{BN}} + \lambda_{\text{ie}}\mathcal{L}_{\text{ie}}, \tag{4}$$

where λ_{oh}, λ_{a}, λ_{BN}, and λ_{ie} are the hyperparameters, and the specific forms of each component are:

$$\begin{cases} \mathcal{L}_{\text{oh}} = \dfrac{1}{n}\displaystyle\sum_i^n \mathcal{L}_{\text{cross}}\left(F(G(z^i)), \hat{y}^i\right), \mathcal{L}_{\text{a}} = -\dfrac{1}{n}\displaystyle\sum_i^n \left\|\text{Feat}(G(z^i))\right\|_1, \\ \mathcal{L}_{\text{BN}}(G(z)) = \displaystyle\sum_l \left\|\mu_l(G(z)) - \mathbb{E}\left(\mu_l(x) \mid \mathcal{X}\right)\right\|_2 + \sum_l \left\|\sigma_l^2(G(z)) - \mathbb{E}\left(\sigma_l^2(x) \mid \mathcal{X}\right)\right\|_2, \\ \mathcal{L}_{\text{ie}} = -\mathcal{H}_{\text{info}}\left(\dfrac{1}{n}\sum \text{y}_F^c\right) = -\left(-\dfrac{1}{K}\displaystyle\sum_{c=1}^K p_c \log p_c\right) = \dfrac{1}{K}\displaystyle\sum_{c=1}^K p_c \log p_c, \end{cases} \tag{5}$$

where z, $\text{Feat}(\cdot)$, μ_l and σ_l represents the stochastic input, hidden layer features, mean and variance value of the input data. The function of each component can be summarized as:

- $\mathcal{L}_{\text{oh}}$ constrains the generator to produce surrogate data according to the pseudo labels.
- $\mathcal{L}_{\text{BN}}$ optimizes the statistical characteristics of surrogate data by leveraging the statistics stored in the batch normalization (BN) layers of the naturally pre-trained model, while $\mathcal{L}_{\text{a}}$ aims to avoid trivial solutions by ensuring the activation of feature maps.
- $\mathcal{L}_{\text{ie}}$ induces the model to generate outputs with a uniform inter-class distribution, ensuring that the number of generated samples for each class remains roughly consistent.

Based on the pseudo data $G(z)$, the optimization goal of the student in stage 1 is to learning natural knowledge from the teacher model:

$$\mathcal{L}^S_{\text{stage}_1} = \frac{1}{n}\sum_i^n \mathcal{KL}\left(S(G(z^i)), F(G(z^i))\right), \tag{6}$$

where $\mathcal{KL}(\cdot,\cdot)$ is the KL-divergence. After stage 1, the student $S(\cdot)$ initially obtains the natural knowledge, and $G(\cdot)$ acquires the ability to generate pseudo natural data as $G(z)$.

2.2.2 STAGE 2: FAIR & GENERALIZABLE ROBUSTNESS LEARNING

Based stage 1, we achieve fair and generalizable robust training in this section.

Surrogate data balancing. We consider optimizing the fairness challenge from the perspectives of quantity (weighted entropy constraint) and quality (weighted one-hot loss), respectively. We plan to mine hard samples for distillation from the perspective of image quality, and mine hard samples for adversarial training from the perspective of quantity. This is because data-free knowledge distillation has been considered to be highly correlated with image quality Fang et al. (2019; 2021); Yin et al. (2020), while adversarial training is more sensitive to the data amount Wang et al. (2023).

First, in terms of the quality of generated samples for distillation hard data, we define the learning difficulty of a sample based on the output discrepancy between the teacher and student models during the learning process as $s_i = \|F(G(z_i) - S(G(z_i)\|_1$. After obtaining the difficulty score s_i for each sample, it is normalized into a sample weight as $\alpha_i = \frac{\exp(s_i/\tau_s)}{\frac{1}{n}\sum_{j=1}^n \exp(s_j/\tau_s)}$, where τ_s is a temperature

coefficient. In this way, each sample obtains a corresponding difficulty weight, and the mean of the weights is reasonably controlled to be 1, avoiding excessive drift in the weight values. Then, we have the weighted one-hot loss of $G(\cdot)$:

$$\mathcal{L}_{\text{oh}}^{w} = \frac{1}{n}\sum_{i=1}^{n} \alpha_i \cdot \mathcal{L}_{\text{cross}}\left(F\left(G\left(z^i\right)\right), \hat{y}^i\right), \tag{7}$$

where $\hat{y}^i$ represents a pseudo target label, and n is the number of pseudo samples in a mini-batch. This loss function aims to train $G(\cdot)$ to generate pseudo data from random inputs z_i, ensuring that the pseudo data $G(z^i)$ is correctly classified by the teacher model $F(\cdot)$ (i.e., the discriminator) with the corresponding pseudo target labels $\hat{y}^i$. In this way, the optimization focus of generated pseudo image quality is adaptively adjusted.

Secondly, in terms of the quantity of generated samples for adversarial hard data, we consider adaptively adjusting the intra-class quantity of generated surrogate data. Specifically, vanilla $\mathcal{L}_{\text{ie}}$ naively keeps the quantities of all categories consistent, ignoring the varying adversarial difficulty across different categories. To measure the current adversarial difficulty, similarly, we use the normalized output difference of the teacher model between clean surrogate samples and adversarial surrogate samples as the adversarial difficulty index. For class c, we define the adversarial difficulty as $d_c = \mathbb{E}_{x\sim\mathcal{D}_c}\left\|F(G(z_i)) - F(G(z_i)+\delta)\right\|_1$. Then, to encourage the generator to produce more samples from difficult classes, we assign smaller weights to classes with larger teacher disagreement, thereby relaxing the uniformity constraint for those classes as $w_c = \frac{\exp(-\gamma\cdot d_c)}{\sum_{j=1}^{K}\exp(-\gamma\cdot d_j)} \times K$, where γ controls the strength of the relaxation. K serves to make the mean of the weights equal to 1, thereby keeping the overall scale of the loss roughly comparable to that of the original L_{ie}. Then we have the weighted information entropy loss as:

$$\mathcal{L}_{\text{ie}}^{w} = \frac{1}{K}\sum_{c=1}^{K} w_c \cdot p_c \log p_c, \tag{8}$$

where the generator is encouraged to adaptively generate more samples for categories with greater adversarial difficulty. To sum up, the the loss function for $G(\cdot)$ in the stage 2 can be summarized as:

$$\mathcal{L}_{\text{stage}_2}^{G} = \lambda_{\text{h}}^{w}\mathcal{L}_{\text{oh}}^{w} + \lambda_{\text{a}}\mathcal{L}_{\text{a}} + \lambda_{\text{BN}}\mathcal{L}_{\text{BN}} + \lambda_{\text{ie}}^{w}\mathcal{L}_{\text{ie}}^{w}, \tag{9}$$

where the feature activation regularization $\mathcal{L}_a$ and statistical prior regularization $\mathcal{L}_{\text{BN}}$ is is retained.

Self adversarial training (SAT). Since there is no supervision from a robust teacher, existing internal maximization approach based on the output discrepancy between a robust teacher and the student model is unavailable. Furthermore, directly generating adversarial examples based on synthetic data and hard labels has been proven to be suboptimal Lee et al. (2024). We propose a self-adversarial paradigm, which generates adversarial perturbations based on the output discrepancy between the student copy natural teacher. Such a manner ensures the availability of supervision signals while introducing soft labels for better generalization. The inner maximization can be formulated as:

$$\arg\max_{\delta} Dis(S(\hat{x}+\delta), T(\hat{x})), \tag{10}$$

where $\hat{x} = G(z)$ represent the surrogate data. Then, the outer minimization based on the inner maximization comprises three key components: natural distillation (ND), adversarial distillation (AD), and prediction smoothing (PS) terms:

$$\arg\min_{\theta_S} \lambda_{\text{ND}} \cdot Dis(S(\hat{x}), F(\hat{x})) + \lambda_{\text{AD}} \cdot Dis(S(\hat{x}+\delta), F(\hat{x})) + \lambda_{\text{PS}} \cdot Dis(S(\hat{x}), F(\hat{x}+\delta)), \tag{11}$$

where λ_{ND}, λ_{AD}, and λ_{PS} are hyper-parameters for the ND, AD and PS term, respectively. Integrating equation 10 and equation 11 yields the self-adversarial training loss $\mathcal{L}_{\text{SAT}}$.

Nasty regularization. Both NT Ma et al. (2021) and NAT Zhou et al. (2026) have demonstrated that, selecting a vanilla training counterpart of the target model as an adversary model and performing nasty regularization based on it can bring probability sparsity to the target model. However, due to the lack of real data, the method for obtaining the adversary and the form of nasty regularization cannot be directly transferred. We select a frozen copy of the student model obtained in stage 1 as the adversary model $S^{\text{a}}(\cdot)$, and formalize the nasty regularization under data-free conditions as:

$$\mathcal{L}_{\text{nasty}} = -\lambda_{\text{nasty}} \cdot (\mathcal{KL}(S\left(\hat{x}+\delta\right), S^{\text{a}}(\hat{x})) + \mathcal{KL}(S(\hat{x}), S^{\text{a}}(\hat{x}))). \tag{12}$$

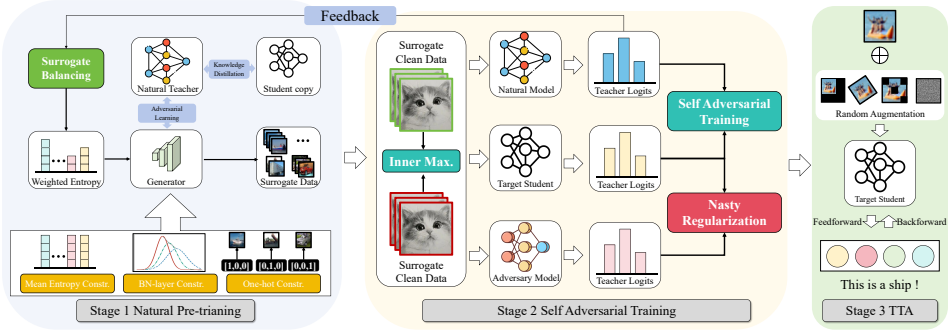


Figure 2: The proposed DFAT method framework, which mainly consists of three stages.

In such a paradigm, while learning robust objectives, the target model is pushed away from the redundant probability preferences output by $S^{\mathrm{a}}(\cdot)$, thereby achieving probabilistically sparse adversarial robustness. Unlike NAT, we only feed natural surrogate samples to the adversary model. This is because, on the one hand, the adversary model is discriminative only toward natural surrogate samples, but performs poorly for adversarial surrogate samples. Findings in NAT Zhou et al. (2026) indicate that regularization is more effective when the adversary model is coupled with the input data. On the other hand, the probability redundancy of a naturally trained adversary model for natural surrogate samples aligns better with the definition, whereas its output probabilities for adversarial surrogate samples tend to be erroneous or completely stochastic.

Then, the overall loss function of the student model in stage 2 can be integrated as:

$$\mathcal{L}^{S}_{\mathrm{stage}_2} = \mathcal{L}_{\mathrm{SAT}} + \mathcal{L}_{\mathrm{nasty}}. \tag{13}$$

2.2.3 Optional Stage 3: Test-time Adaptation

Since the student copy is trained on a surrogate distribution that approximates the real one, distribution shifts during testing are inevitable. A simple yet effective way to exploit information from the real distribution without additional training is test-time adaptation (TTA). However, existing TTA methods mainly rely on auxiliary tasks during training, which introduces optimization challenges when applied to surrogate data. DENT Wang et al. (2021) offers a plug-and-play TTA approach that fine-tunes the model by maximizing its output confidence for each test sample, without requiring any auxiliary tasks during the training phase. In light of this, we incorporate DENT as an optional extension to further enhance DFAT performance at test time. To stabilize the adaptation process, we augment DENT by introducing clean-adversarial data pair augmentations during the adaptation phase while restricting evaluation to the original input. This integration not only improves robustness but also serves as a validation of DFAT adaptability and stability under distribution shifts. To the best of our knowledge, we are the first work to consider the usability of combining data-free learning with TTA. Integrating the three proposed stages yields the final model, as shown in Figure 2.

2.3 Theoretical Analysis

Intuitively, our analysis mainly includes two theoretical results:

(a) *Feasibility of student copy as a surrogate for attack direction* (Theorem 1). When the student model is sufficiently locally close to the teacher model, the adversarial direction of the student model is similar to that of the teacher model. This supports the feasibility of performing adversarial training on a student copy that is sufficiently similar to the teacher model.

(b) *The upper bound of data-free robustness generalization risk* (Theorem 2). The risk upper bound of the student model performing robust training on surrogate data is influenced by the teacher quality, teacher confidence, empirical robust anchoring risk, and distribution discrepancy. We argue that the quality and confidence of a naturally trained teacher induce a controllable risk upper bound, while the empirical robust anchoring risk term and the distribution discrepancy term support the feasibility of performing adversarial training on the surrogate distribution.

We provide the detailed analysis in Appendix C and the proofs in Appendix D.

Table 1: Restult of CIFAR-10 on ResNet-18 (RN-18) and MobileNet-v2 (MN-V2).

Backbones	Methods	Clean			FGSM			PGD			CW			AA		
		Avg.	Worst	NSD	Avg.	Worst	NSD	Avg.	Worst	NSD	Avg.	Worst	NSD	Avg.	Worst	NSD
RN-18	ZSKT	65.48	32.20	0.266	29.80	5.00	0.563	22.88	3.40	0.624	18.48	2.60	0.696	17.16	2.20	0.740
	CMI	68.29	51.20	0.161	33.34	19.50	0.337	28.94	15.80	0.393	27.63	14.90	0.409	26.85	14.60	0.419
	DeepInv	71.22	58.90	0.124	38.60	22.50	0.303	27.95	15.00	0.375	27.23	14.50	0.379	25.19	11.60	0.422
	Fast	72.20	57.80	0.132	43.96	26.90	0.295	39.64	20.10	0.339	33.96	18.40	0.388	33.24	17.70	0.401
	DERD	73.91	53.70	0.158	39.86	22.90	0.341	30.55	13.90	0.417	29.98	13.60	0.421	28.61	11.80	0.443
	DFARD	67.93	45.90	0.247	41.27	18.40	0.458	37.34	15.20	0.491	35.47	13.30	0.513	34.48	12.00	0.552
	DFHL	74.42	58.60	0.129	41.67	24.30	0.298	36.99	19.30	0.344	37.33	17.80	0.348	36.39	18.50	0.351
	FERD	79.26	65.90	0.105	51.09	31.30	0.237	42.24	20.40	0.331	41.73	19.30	0.339	40.12	20.80	0.325
	DFAT (Ours)	82.41	67.01	**0.101**	51.34	32.41	0.196	41.73	20.88	0.361	40.60	18.76	0.335	37.79	18.93	0.364
	DFAT + DENT	**85.44**	**69.85**	0.112	**56.41**	**34.84**	**0.189**	**50.32**	**27.41**	**0.327**	**50.09**	**29.61**	**0.296**	**45.41**	**27.41**	**0.291**
MN-V2	ZSKT	54.69	16.10	0.394	23.62	3.90	0.663	19.83	3.00	0.702	15.86	1.90	0.725	14.85	1.60	0.749
	CMI	60.70	38.60	0.205	27.90	12.50	0.393	19.70	8.50	0.482	19.57	8.10	0.461	18.09	7.60	0.506
	Deepinv	62.77	46.00	0.170	31.66	14.50	0.306	19.13	7.00	0.403	19.01	6.90	0.407	15.39	5.40	0.471
	Fast	58.60	43.40	0.198	29.91	15.70	0.374	24.28	11.60	0.444	19.46	7.50	0.535	18.35	6.80	0.563
	DERD	69.80	48.70	0.175	39.99	22.10	0.329	28.82	15.10	0.372	27.97	14.00	0.386	25.80	12.00	0.403
	DFARD	66.32	40.10	0.273	38.71	12.60	0.519	33.84	9.20	0.562	32.76	8.70	0.575	31.81	8.10	0.584
	DFHL	70.83	50.60	0.153	37.22	16.70	0.320	33.27	14.20	0.358	33.60	14.30	0.353	32.58	13.70	0.368
	FERD	78.34	64.10	0.108	53.04	33.40	0.229	41.63	20.80	0.304	41.43	20.50	0.314	38.06	20.30	0.349
	DFAT (Ours)	75.71	62.71	0.095	51.43	30.97	0.235	34.99	16.41	0.351	34.26	16.93	0.337	33.34	17.50	0.362
	DFAT + DENT	**79.81**	**64.36**	**0.108**	**54.61**	**35.95**	**0.220**	**43.41**	**25.62**	**0.259**	**42.14**	**24.57**	**0.274**	**41.72**	**23.64**	**0.261**

Table 2: Restult of CIFAR-100 on ResNet-18 (RN-18) and MobileNet-v2 (MN-V2).

Backbones	Methods	Clean			FGSM			PGD			CW			AA		
		Avg.	Worst	NSD	Avg.	Worst	NSD	Avg.	Worst	NSD	Avg.	Worst	NSD	Avg.	Worst	NSD
RN-18	ZSKT	36.44	6.40	0.522	12.69	0.10	0.985	10.17	0.00	1.083	7.12	0.00	1.273	6.01	0.00	1.361
	CMI	51.60	27.20	0.295	25.40	5.60	0.624	22.63	3.50	0.686	18.94	1.90	0.785	17.72	1.30	0.824
	DeepInv	53.44	27.20	0.298	26.14	5.60	0.630	23.11	3.60	0.699	19.05	2.20	0.792	17.49	1.60	0.857
	Fast	50.69	23.50	0.323	26.45	5.50	0.634	24.13	4.50	0.687	19.44	2.20	0.806	18.24	1.70	0.859
	DERD	54.01	28.10	0.295	27.83	5.20	0.631	24.78	4.70	0.689	22.56	3.20	0.776	21.01	1.40	0.856
	DFARD	50.43	21.80	0.340	26.44	4.80	0.644	24.22	3.90	0.693	21.56	2.50	0.781	20.21	1.50	0.828
	DFHL	46.27	18.70	0.370	18.73	2.60	0.799	15.72	1.30	0.933	16.33	1.50	0.896	15.23	1.10	0.954
	FERD	**55.91**	29.00	0.289	29.23	5.70	0.622	26.10	4.90	0.681	23.39	3.30	0.754	21.79	1.70	0.805
	DFAT (Ours)	49.54	25.77	0.271	25.61	6.79	**0.617**	25.31	5.21	0.671	24.11	4.19	0.745	19.46	1.61	0.818
	DFAT + DENT	53.67	**30.14**	**0.269**	**29.75**	**7.01**	0.625	**26.98**	**5.94**	**0.665**	**25.27**	**5.01**	**0.737**	**21.94**	**3.99**	**0.653**
MN-V2	ZSKT	30.79	2.70	0.650	13.66	0.00	0.975	12.44	0.00	1.007	9.54	0.00	1.124	8.62	0.00	1.208
	CMI	35.80	6.00	0.425	13.82	0.60	0.828	9.91	0.40	0.966	9.06	0.00	1.000	8.04	0.00	1.042
	Deepinv	38.10	8.00	0.447	13.97	1.00	0.920	9.98	1.00	1.123	8.12	0.00	1.260	6.67	0.00	1.416
	Fast	39.00	6.00	0.441	16.42	1.40	0.821	14.80	0.80	0.913	10.44	0.00	1.085	9.53	0.00	1.139
	DERD	54.12	29.10	0.291	26.31	5.40	0.646	22.39	3.00	0.742	20.59	2.70	0.774	18.92	1.70	0.832
	DFARD	49.02	20.60	0.340	24.62	4.00	0.668	22.42	3.10	0.722	20.65	2.10	0.778	19.30	1.60	0.826
	DFHL	42.08	11.30	0.436	16.25	0.90	0.914	13.55	0.40	1.015	14.06	0.70	0.980	12.99	0.40	1.051
	FERD	56.14	29.20	0.285	**29.63**	6.10	0.611	24.48	3.10	0.713	23.20	3.30	0.743	21.20	2.30	0.793
	DFAT (Ours)	49.41	25.41	0.286	24.42	5.84	0.596	22.62	3.02	0.724	22.16	2.94	**0.732**	18.86	1.99	**0.778**
	DFAT + DENT	**56.54**	**29.55**	**0.257**	28.47	**6.42**	**0.591**	**25.88**	**3.49**	**0.701**	**24.75**	**3.47**	0.739	**23.39**	**2.90**	0.785

3 EXPERIMENTS

3.1 EXPERIMENTAL SETTINGS

Datasets & models. To facilitate performance comparison, we adopt three widely used benchmark datasets: CIFAR-10, CIFAR-100, and Tiny-ImageNet. Regarding model selection, to control single variables, we follow the student model configurations used in current data-free robust distillation works Zhou et al. (2024); Li et al. (2026), where ResNet-18 He et al. (2016) and MobileNet-v2 Sandler et al. (2018) are the backbones for CIFAR-10 and CIFAR-100, and adopt PreActResNet-18 as the backbone for Tiny-ImageNet. Since adversarial training inherently poses greater optimization challenges for smaller models, and to further validate the effectiveness of DFAT, we conduct additional experiments on CIFAR-10 using Wide-ResNet-34-10 to verify its efficacy.

Comparison protocol. We mainly compare our method with existing data-free robustness distillation methods, including ZSKT Micaelli & Storkey (2019), CMI Fang et al. (2021), DeepInv Yin et al. (2020), Fast Fang et al. (2022), DERD Zhou et al. (2024), DFARD Wang et al. (2024), DFHL Yuan et al. (2024) and FERD Li et al. (2026). Note that the first four methods were originally designed as data-free knowledge distillation without robustness. Following FERD, adversarial loss based on RSLAD are integrated into these methods to realize their corresponding robustness distillation versions. The latter four methods belong to data-free robustness distillation, whose frameworks include an available robust teacher. Considering that DFAT is essentially data-free adversarial training, the absence of teacher guidance makes the task more challenging, and adversarial training also poses greater optimization difficulties for smaller models. To further validate the performance potential of

Table 3: Results of CIFAR-10 on WRN-3410.

Methods	Clean		AA	
	Avg.	Worst	Avg.	Worst
DERD	69.00	45.70	33.50	15.10
DFARD	70.81	49.10	34.38	14.60
DFHL	70.89	52.60	36.59	15.60
FERD	78.23	61.70	38.67	16.00
DFAT (Ours)	78.81	61.42	39.48	16.79
DFAT + DENT	**80.04**	**64.93**	**42.07**	**17.96**

Table 4: Black-box evaluation on CIFAR-10.

Methods	ResNet-18			MobileNet-V2		
	PGD	CW	Square	PGD	CW	Square
ZSKT	32.41	32.03	27.19	27.01	26.77	23.96
CMI	37.52	36.07	31.36	29.47	28.32	24.39
DFARD	48.42	47.98	41.69	43.87	42.22	37.91
FERD	50.78	48.62	43.72	47.37	46.96	39.52
DFAT (Ours)	47.01	46.23	41.79	41.26	40.39	35.93
DFAT + DENT	**55.41**	**55.07**	**48.94**	**51.36**	**49.75**	**44.97**

Table 5: Ablation analysis on CIFAR-10.

Modules		Attacks					
		Clean	FGSM	PGD	CW	AA	Avg.
Stage 1		89.76	7.51	1.85	2.03	0.31	20.29
Stage 2	SAT	78.92	46.62	35.79	33.26	29.89	44.89
	+ L_w^{ch}	83.63	46.25	35.82	32.93	28.29	45.38
	+ L_w^{ie}	80.14	48.92	38.64	36.77	32.06	47.30
	+ L_{nasty}	83.09	51.09	40.79	39.72	37.01	50.34
DFAT		**82.41**	**51.34**	**41.73**	**40.60**	**37.79**	**50.77**

Table 6: Evaluation of the natural models.

Datasets	Backbones	Attack				
		Clean	FGSM	PGD	CW	AA
CIFAR-10	ResNet-18	91.74	7.46	2.08	2.71	0.00
	MobileNet-v2	88.64	7.28	2.12	2.64	0.00
CIFAR-100	ResNet-18	69.42	5.98	1.76	1.73	0.00
	MobileNet-v2	66.79	5.54	1.69	1.86	0.00
Tiny-ImageNet	PreActResNet-18	58.39	5.29	1.51	1.66	0.00

DFAT, we also provide experimental results deployed on Wide-ResNet-34-10. Please further refer to Appendix E.1 for more implementation details.

3.2 Main Results

We report the main white-box results of DFAT in Table 1, Table 2 & 3 and Table 7 (in Appendix E.2) for CIFAR-10, CIFAR-100 and Tiny-ImageNet, respectively. The experimental results demonstrate that, compared to the naive transfer of data-free knowledge distillation, DFAT achieves significantly superior robustness. Coupled with test-time adaptation, it can even surpass the performance of specifically designed robustness distillation methods. To verify the reliability of DFAT incorporating the TTA, we follow existing settings and supplement the evaluation of adaptive attacks in Appendix E.3. Furthermore, we also report the evaluation of black-box attacks in Table 4, which verifies its effectiveness against black-box attacks.

3.3 Ablation Analysis

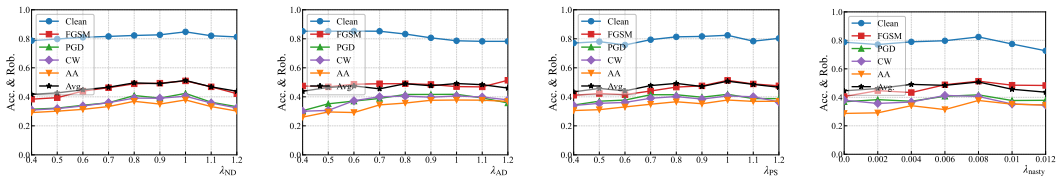
The effectiveness of each module. The ablation analysis results for each module based on CIFAR-10 are shown in Table 5. Intuitively, the natural knowledge distillation in stage 1 equips the target model with natural knowledge, while the self-adversarial training in stage 2 brings considerable adversarial robustness to the target model. The reweighting and nasty regularization further contribute robustness gains from the perspectives of fairness and generalization, respectively, building upon the previous stages. In summary, the function of each module aligns with our expectations.

Ablation analysis of hyperparameters. For stage 1, we adopt hyperparameters consistent with DERD and do not perform redundant ablation analyses. We mainly analyze λ_{ND}, λ_{AD}, and λ_{PS} in Eq. equation 11 and λ_{a} in Eq. equation 12. Our ablation analyses are primarily conducted based on the magnitude of hyperparameters in the baseline model, and the detailed experimental results are shown in Figure 3. In short, despite some fluctuations, the hyperparameter choices under our default settings generally achieve empirically optimal performance.

3.4 Qualitative Analysis

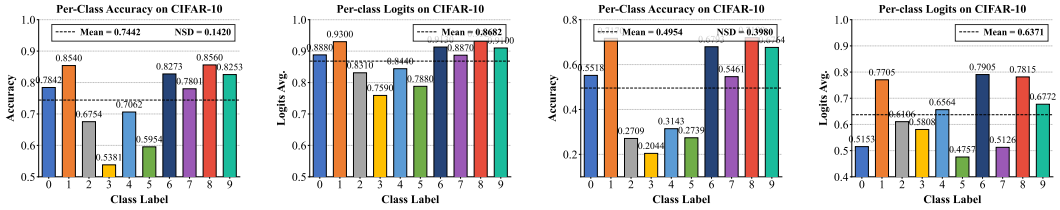
Verification of model fairness. We here verify that the target model trained with reweighting exhibits better fairness. For comparison, we report the class-wise accuracies on both natural and adversarial samples for models with and without the reweighting module, respectively. The experimental results are shown in Figure 4, where the mean, worst-case result, and standard deviation are all reported. Intuitively, the reweighting module significantly improves the robustness performance of the target model by addressing the fairness challenge, with notable improvements in both natural accuracy and robust accuracy. This demonstrates the effectiveness of our reweighting module.

Verification of probability sparsity. We further verify that the model trained with nasty regularization exhibits significant probability sparsity, and the experimental results are shown in Figure 5. In



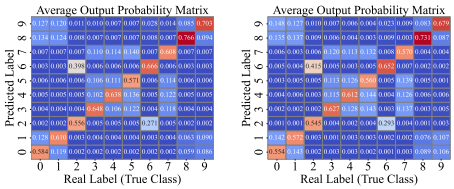
(a) Abla. on λ_{ND} (b) abla. on λ_{AD} (c) Abla. on λ_{PS} (d) abla. on λ_{nasty}

Figure 3: Ablation analysis on the hyperparameters.



(a) Base. on clean (b) Base. on Adv. (c) DFAT on clean (d) DFAT on Adv.

Figure 4: Ablation analysis on the fairness challenge.



(a) w/o $\mathcal{L}_{\text{nasty}}$ (b) with $\mathcal{L}_{\text{nasty}}$

Figure 5: Ablation analysis on $\mathcal{L}_{\text{nasty}}$.



(a) Ori. CIFAR-10 (b) Surrogate in DFAT

Figure 6: Surrogate data of DFAT.

sharp contrast to Figure 8 and Figure 9 in Appendix B.1, the model trained with $\mathcal{L}_{\text{nasty}}$ demonstrates pronounced probability sparsity: the non-target probabilities are concentrated on classes similar to the target class, while the probabilities of dissimilar classes approach zero. This provides an explanation for the generalization ability of DFAT from the perspective of probability sparsity.

Surrogate data visualization. We provide visualizations of the surrogate dataset to offer a more intuitive understanding of DFAT, as shown in Figure 6. The data in the green boxes are natural real/surrogate data, while those in the red boxes are adversarial real/surrogate data. We also provide a comparison with current methods in Figure 10 of Appendix E.4. In short, the generated data from stage 1 in DFAT do not suffer from mode collapse due to the dynamic weights, and they remain consistent with the class semantics. This provides a certain pixel-level interpretability for DFAT.

4 Conclusion

In this paper, we propose DFAT, which achieves robust post-training without relying on either real data or a robust teacher. Specifically, we theoretically demonstrate the feasibility of adversarial training without a robust teacher and of executing adversarial training based on surrogate data, and we propose a two-stage model. The target model performs natural pre-training in stage 1 to acquire natural knowledge, and performs self-adversarial training in stage 2 to obtain robust representations. Furthermore, we identify two unique challenges in DFAT: fairness and generalization, and we propose to mitigate these two challenges from the perspectives of generated sample reweighting and nasty regularization, respectively. The integration with test-time adaptation methods further boosts the stability and adaptive potential of DFAT. Qualitative and quantitative experiments demonstrate the effectiveness and interpretability of the proposed DFAT. We hope this work further advances robustness research under data privacy conditions.

5 THE USE OF LARGE LANGUAGE MODELS (LLMS)

LLMs are used sparingly and only to assist with proofreading and improving the linguistic fluency of a few sections of this paper (such as *Related Work*, *Experiments*, and *Appendix*), i.e., to aid or polish writing. All scientific contributions and core idea are the work of the authors. This assistance poses no issues of scientific ethics or misconduct.

REFERENCES

Wieland Brendel, Jonas Rauber, and Matthias Bethge. Decision-based adversarial attacks: Reliable attacks against black-box machine learning models. *arXiv preprint arXiv:1712.04248*, 2017.

Longbing Cao. Ai in finance: challenges, techniques, and opportunities. *ACM Computing Surveys (CSUR)*, 55(3):1–38, 2022.

Nicholas Carlini and David Wagner. Towards evaluating the robustness of neural networks. In *2017 IEEE Symposium on Security and Privacy (SP)*, pp. 39–57, 2017.

Pin-Yu Chen, Huan Zhang, Yash Sharma, Jinfeng Yi, and Cho-Jui Hsieh. Zoo: Zeroth order optimization based black-box attacks to deep neural networks without training substitute models. In *Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security*, pp. 15–26, 2017.

Francesco Croce and Matthias Hein. Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks. In *International Conference on Machine Learning*, pp. 2206–2216, 2020.

Yinpeng Dong, Fangzhou Liao, Tianyu Pang, Hang Su, Jun Zhu, Xiaolin Hu, and Jianguo Li. Boosting adversarial attacks with momentum. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 9185–9193, 2018.

Gongfan Fang, Jie Song, Chengchao Shen, Xinchao Wang, Da Chen, and Mingli Song. Data-free adversarial distillation. *arXiv preprint arXiv:1912.11006*, 2019.

Gongfan Fang, Jie Song, Xinchao Wang, Chengchao Shen, Xingen Wang, and Mingli Song. Contrastive model inversion for data-free knowledge distillation. *arXiv preprint arXiv:2105.08584*, 2021.

Gongfan Fang, Kanya Mo, Xinchao Wang, Jie Song, Shitao Bei, Haofei Zhang, and Mingli Song. Up to 100x faster data-free knowledge distillation. In *Proceedings of the AAAI conference on artificial intelligence*, volume 36, pp. 6597–6604, 2022.

Micah Goldblum, Liam Fowl, Soheil Feizi, and Tom Goldstein. Adversarially robust distillation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pp. 3996–4003, 2020.

Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778, 2016.

Zhichao Huang, Yanbo Fan, Chen Liu, Weizhong Zhang, Yong Zhang, Mathieu Salzmann, Sabine Süsstrunk, and Jue Wang. Fast adversarial training with adaptive step size. *IEEE Transactions on Image Processing*, 2023.

Xiaojun Jia, Yong Zhang, Baoyuan Wu, Ke Ma, Jue Wang, and Xiaochun Cao. Las-at: adversarial training with learnable attack strategy. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13398–13408, 2022a.

Xiaojun Jia, Yong Zhang, Baoyuan Wu, Jue Wang, and Xiaochun Cao. Boosting fast adversarial training with learnable adversarial initialization. *IEEE Transactions on Image Processing*, 31: 4417–4430, 2022b.

Hyeyoon Lee, Kanghyun Choi, Dain Kwon, Sunjong Park, Mayoore Selvarasa Jaiswal, Noseong Park, Jonghyun Choi, and Jinho Lee. Datafreeshield: Defending adversarial attacks without training data. *arXiv preprint arXiv:2406.15635*, 2024.

Zhengxiao Li, Liming Lu, Xu Zheng, Si-Yuan Liang, Taric Chen, Yongbin Zhou, and Shuchao Pang. Ferd: Fairness-enhanced data-free adversarial robustness distillation. In *International Conference on Learning Representations*, volume 2026, pp. 152021–152043, 2026.

Haoyu Ma, Tianlong Chen, Ting-Kuei Hu, Chenyu You, Xiaohui Xie, and Zhangyang Wang. Undistillable: Making a nasty teacher that cannot teach students. *arXiv preprint arXiv:2105.07381*, 2021.

Haoyu Ma, Yifan Huang, Hao Tang, Chenyu You, Deying Kong, and Xiaohui Xie. Sparse logits suffice to fail knowledge distillation. In *ICLR 2022 Workshop on PAIR2Struct: Privacy, Accountability, Interpretability, Robustness, Reasoning on Structured Data*, 2022. URL https://openreview.net/forum?id=BxZgduuND15.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017.

Paul Micaelli and Amos Storkey. Zero-shot knowledge transfer via adversarial belief matching. *Advances in neural information processing systems*, 32, 2019.

Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, and Pascal Frossard. Deepfool: a simple and accurate method to fool deep neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2574–2582, 2016.

Pranav Rajpurkar, Emma Chen, Oishi Banerjee, and Eric J Topol. Ai in health and medicine. *Nature medicine*, 28(1):31–38, 2022.

Min Ren, Yuhao Zhu, Yunlong Wang, and Zhenan Sun. Perturbation inactivation based adversarial defense for face recognition. *IEEE Transactions on Information Forensics and Security*, 17:2947–2962, 2022.

Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4510–4520, 2018.

Changhao Shi, Chester Holtz, and Gal Mishne. Online adversarial purification based on self-supervision. *arXiv preprint arXiv:2101.09387*, 2021.

Chunyu Sun, Chenye Xu, Chengyuan Yao, Siyuan Liang, Yichao Wu, Ding Liang, Xianglong Liu, and Aishan Liu. Improving robust fariness via balance adversarial training. In *Proceedings of the AAAI conference on artificial intelligence*, volume 37, pp. 15161–15169, 2023.

Satoshi Suzuki, Shin'ya Yamaguchi, Shoichiro Takeda, Sekitoshi Kanai, Naoki Makishima, Atsushi Ando, and Ryo Masumura. Adversarial finetuning with latent representation constraint to mitigate accuracy-robustness tradeoff. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4367–4378. IEEE, 2023.

Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013.

Dequan Wang, An Ju, Evan Shelhamer, David Wagner, and Trevor Darrell. Fighting gradients with gradients: Dynamic defenses against adversarial attacks. *arXiv preprint arXiv:2105.08714*, 2021.

Yisen Wang, Difan Zou, Jinfeng Yi, James Bailey, Xingjun Ma, and Quanquan Gu. Improving adversarial robustness requires revisiting misclassified examples. In *International Conference on Learning Representations*, 2019.

Yuzheng Wang, Zhaoyu Chen, Dingkang Yang, Pinxue Guo, Kaixun Jiang, Wenqiang Zhang, and Lizhe Qi. Out of thin air: Exploring data-free adversarial robustness distillation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 5776–5784, 2024.

Zekai Wang, Tianyu Pang, Chao Du, Min Lin, Weiwei Liu, and Shuicheng Yan. Better diffusion models further improve adversarial training. In *International Conference on Machine Learning*, pp. 36246–36263, 2023.

Cihang Xie, Zhishuai Zhang, Yuyin Zhou, Song Bai, Jianyu Wang, Zhou Ren, and Alan L Yuille. Improving transferability of adversarial examples with input diversity. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2730–2739, 2019.

Han Xu, Xiaorui Liu, Yaxin Li, Anil Jain, and Jiliang Tang. To be robust or to be fair: Towards fairness in adversarial training. In *International conference on machine learning*, pp. 11492–11501. PMLR, 2021.

Mang Ye, Jianbing Shen, Gaojie Lin, Tao Xiang, Ling Shao, and Steven CH Hoi. Deep learning for person re-identification: A survey and outlook. *IEEE transactions on pattern analysis and machine intelligence*, 44(6):2872–2893, 2021.

Hongxu Yin, Pavlo Molchanov, Jose M Alvarez, Zhizhong Li, Arun Mallya, Derek Hoiem, Niraj K Jha, and Jan Kautz. Dreaming to distill: Data-free knowledge transfer via deepinversion. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8712–8721. IEEE, 2020.

Shenglin Yin, Kelu Yao, Sheng Shi, Yangzhou Du, and Zhen Xiao. Again: Adversarial training with attribution span enlargement and hybrid feature fusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 20544–20553, 2023.

Xiaojian Yuan, Kejiang Chen, Wen Huang, Jie Zhang, Weiming Zhang, and Nenghai Yu. Data-free hard-label robustness stealing attack. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 6853–6861, 2024.

Xinli Yue, Mou Ningping, Qian Wang, and Lingchen Zhao. Revisiting adversarial robustness distillation from the perspective of robust fairness. *Advances in Neural Information Processing Systems*, 36:30390–30401, 2023.

Hongyang Zhang, Yaodong Yu, Jiantao Jiao, Eric Xing, Laurent El Ghaoui, and Michael Jordan. Theoretically principled trade-off between robustness and accuracy. In *International Conference on Machine Learning*, pp. 7472–7482, 2019.

Yanghao Zhang, Tianle Zhang, Ronghui Mu, Xiaowei Huang, and Wenjie Ruan. Towards fairness-aware adversarial learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 24746–24755, 2024.

Shiji Zhao, Ranjie Duan, Xizhe Wang, and Xingxing Wei. Improving adversarial robust fairness via anti-bias soft label distillation. *Advances in Neural Information Processing Systems*, 37:89125–89149, 2024.

Yuhang Zhou, Yushu Zhang, Leo Yu Zhang, and Zhongyun Hua. Derd: data-free adversarial robustness distillation through self-adversarial teacher group. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pp. 10055–10064, 2024.

Yuhang Zhou, Zhongyun Hua, Zhaoquan Gu, Keke Tang, Rushi Lan, Yushu Zhang, Qing Liao, and Leo Zhang. Nasty adversarial training: A probability sparsity perspective for robustness enhancement. In *International Conference on Learning Representations*, volume 2026, pp. 41955–41976, 2026.

Jianing Zhu, Jiangchao Yao, Bo Han, Jingfeng Zhang, Tongliang Liu, Gang Niu, Jingren Zhou, Jianliang Xu, and Hongxia Yang. Reliable adversarial distillation with unreliable teachers. *arXiv preprint arXiv:2106.04928*, 2021.

Bojia Zi, Shihao Zhao, Xingjun Ma, and Yu-Gang Jiang. Revisiting adversarial robustness distillation: Robust soft labels make student better. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 16443–16452, 2021.

Junhua Zou, Zhisong Pan, Junyang Qiu, Xin Liu, Ting Rui, and Wei Li. Improving the transferability of adversarial examples with resized-diverse-inputs, diversity-ensemble and region fitting. In *European Conference on Computer Vision*, pp. 563–579, 2020.

A APPENDIX

A RELATED WORKS

A.1 ADVERSARIAL ATTACK

Adversarial attacks aim to introduce small, human-imperceptible perturbations to the input, which can mislead a well-trained model into producing incorrect predictions. Depending on the level of information accessible to the attacker, adversarial attacks can be generally categorized into white-box and black-box settings. In white-box attacks, the attacker has complete access to the target model. Common approaches include gradient-based methods (Goodfellow et al., 2014; Dong et al., 2018; Madry et al., 2017), classifier-based methods (Moosavi-Dezfooli et al., 2016), and optimization-based techniques (Carlini & Wagner, 2017). In contrast, black-box attacks assume limited prior knowledge of the target model and are typically divided into score-based (Chen et al., 2017), decision-based (Brendel et al., 2017), and transfer-based attacks (Xie et al., 2019; Zou et al., 2020). Among these, transfer-based attacks involve training a surrogate model to approximate the behavior of the target model and are widely used to evaluate the black-box adversarial robustness of defense methods.

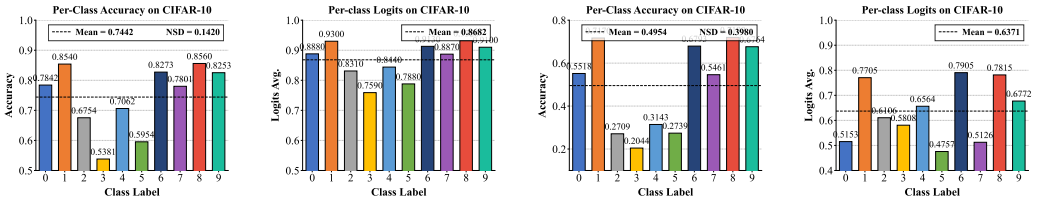
A.2 ADVERSARIAL DEFENSES

Adversarial defenses aim to improve a model's accuracy on adversarial examples. AT (Goodfellow et al., 2014; Zhang et al., 2019; Dong et al., 2018; Madry et al., 2017; Jia et al., 2022a; Wang et al., 2019) is widely regarded as one of the most effective defenses. RD (Goldblum et al., 2020; Zhu et al., 2021; Zi et al., 2021) is later proposed to transfer robustness from a large, robust model to a smaller student model. Recently, AT has been further explored, such as fairness issues (Zhang et al., 2024), attention maps (Yin et al., 2023), and learnable strategies (Jia et al., 2022a). Data-free adversarial defense (Wang et al., 2024; Zhou et al., 2024; Lee et al., 2024) is explored to achieve adversarial robustness in scenarios with limited data. Additionally, advanced internal maximization methods like fast adversarial training (FAT) (Jia et al., 2022b; Huang et al., 2023) have been developed to enhance the efficiency of AT without significantly sacrificing robustness.

Data-free adversarial defense aims to achieve adversarial robustness without access to real data. DFARD Wang et al. (2024) is the first to propose an interactive temperature adjustment strategy to address the challenge of poor transferability of synthetic data, while DERD Zhou et al. (2024) introduces a prior-guided expert strategy. DFHL Yuan et al. (2024) proposes the concept of high-entropy samples, exploiting the potential of data-free defense from the perspective of hard samples. FERD Li et al. (2026) is the first to consider the fairness issue in data-free robust distillation. Furthermore, although specially designed data-free robust distillation methods are limited, data-free knowledge distillation methods Micaelli & Storkey (2019); Fang et al. (2021); Yin et al. (2020) can be easily adapted into data-free robustness distillation by introducing adversarial examples during the sample generation stage, and replacing the teacher model with a robust teacher. However, all the above methods face a common challenge: every paradigm is confined to robust distillation, which, while eliminating reliance on real data, still depends on a robust teacher. As analyzed in Section 1, a robust large-model teacher remains a luxury assumption. We propose DFAT, which aims to achieve robust post-training based solely on a naturally pre-trained model without real data or robust teacher.

A.3 FAIRNESS CHALLENGE IN ADVERSARIAL ROBUSTNESS

Fairness challenge is a common challenge in data-driven defense methods, reflected in the fact that the target model exhibits high robustness for some classes but poor robustness for others. FRL Xu et al. (2021) first introduces "robust fairness" to evaluate the robustness gap across classes, and proposes a fairness constraint mechanism. BAT Sun et al. (2023) identifies significant unfairness between "source classes" and "target classes" during adversarial training and mitigates it by adjusting attack intensities for each class and applying uniform distribution constraints. FairARD Yue et al. (2023) improves the robust fairness of student models by increasing the weight of hard classes. ABSLD Zhao et al. (2024) proposes adjusting the inter-class smoothness of sample soft labels during training to adaptively reduce class gaps. However, all these methods only consider fairness issues in data-driven adversarial defense. FERD Li et al. (2026) further addresses fairness challenges in a



(a) Acc. on Clean (b) Conf. on Clean (c) Acc. on Adv. (d) Conf. on Adv.

Figure 7: Validation of the fairness challenges based on DERD.

data-free distillation framework, but it still relies on a robust teacher. Our work is the first to consider fairness challenges in a data-free adversarial training framework that has neither real data nor a robust teacher.

A.4 Probability Sparsity

Probabilistic sparsity initially emerged as a byproduct of intellectual property protection. Ma et al. Ma et al. (2021) propose nasty training to prevent unauthorized knowledge distillation. Later work Ma et al. (2022) further explained that the non-distillable property originates from probabilistic sparsity. Zhou et al. Zhou et al. (2026) then introduce nasty regularization into adversarial training and propose Nasty Adversarial Training (NAT), attributing probabilistic sparsity to higher-order constraint terms in nasty regularization. NAT presents an interesting perspective: on one hand, probabilistic sparsity significantly improves the robustness performance of adversarial training. On the other hand, models with probabilistic sparsity exhibit better interpretability: the model tends to distribute redundant probability to classes that are similar to the target class. For example, when the target class is "cat", the model tends to allocate redundant probability to "dog", "deer", "horse", and other animal-related categories. Contrary to probabilistic sparsity, we define the preference for non-sparse probabilities across classes other than the target class as "probabilistic redundancy", and we find that probabilistic redundancy is more pronounced in data-free adversarial training. We propose to improve probabilistic sparsity in the data-free scenario for the first time.

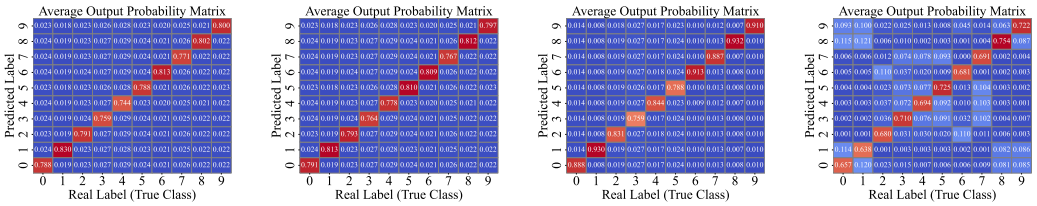
B Observation of Fairness and Generalizability

In this section, we preliminarily verify the two prominent challenges we propose that exist in data-free adversarial training, namely the fairness challenge and the probability redundancy challenge.

B.1 Fairness Challenge

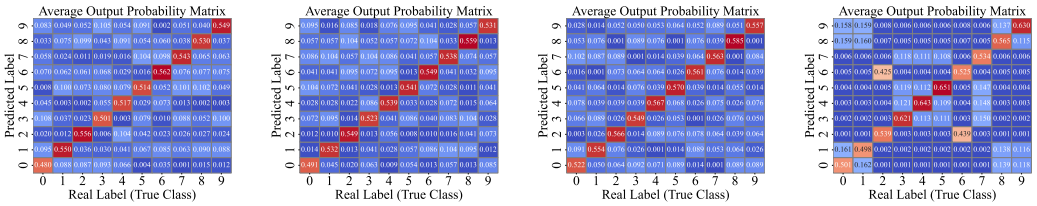
Existing works Li et al. (2026) typically validate fairness directly on test data, and we offer a novel perspective that, the fairness challenge of data-free defense models may originate from the generated surrogate data. Current methods incorporate mean entropy regularization to ensure inter-class quantity balance in surrogate data. However, it may be exactly this mean entropy regularization that induces unfair preference, since the difficulty for the target model to learn different classes is clearly distinct. Therefore, instead of repeatedly validating the student's unfair inter-class accuracy on the test data, we validate the inter-class difficulty disparity of the surrogate distribution.

Using CIFAR-10 as the target task dataset, we validate this hypothesis by statistically analyzing the student model's accuracy and confidence (probability) for each class. To avoid interference from reweighting algorithms, we select DERD Zhou et al. (2024), a SOTA method that does not consider fairness. We first count the number of samples for each class among the 10,000 surrogate data points. Due to the presence of mean entropy regularization, the sample count for each class is essentially balanced. However, as shown in Figure 7, the student model exhibits significant disparities in both average accuracy (Figure 7a) and average confidence (Figure 7b) across different classes. As a preliminary validation, based on the generated surrogate data and vanilla PGD attack, we also conduct an initial verification of the fairness challenge of the student model on the adversarial surrogate distribution. Similar to the vanilla surrogate distribution, whether from the perspective of per-class



(a) DERD (b) FERD (c) AT (d) NAT

Figure 8: Validating probability redundancy on clean data. We take CIFAR-10 as an example, where labels '0' to '9' correspond to the ten class labels in CIFAR-10, i.e., "airplane", "automobile", "bird", "cat", "deer", "dog", "frog", "horse", "ship", and "truck", respectively.



(a) DERD (b) FERD (c) AT (d) NAT

Figure 9: Validating probability redundancy on adversarial data.

average accuracy (Figure 7c) or confidence (Figure 7d), the student model demonstrates a notable imbalance in inter-class fairness on the adversarial surrogate distribution. The above experimental results preliminarily demonstrate that, in addition to the fairness preference of algorithms currently concerned within the community, the surrogate distribution is also an important attribution factor for the fairness of the target model in the data-free robust training framework.

B.2 Probability Redundancy Challenge

As discussed in Section 1, probability sparsity/redundancy is highly correlated with model generalization. Therefore, we statistically analyze the average probability assignment of the target/student model on clean/adversarial data across different class within the current data-free robust training framework, as shown in Figure 8. We also provide the confusion matrices of standard AT Madry et al. (2017) and NAT Zhou et al. (2026) for comparison. Here, the horizontal axis represents the target label, and each probability denotes the average output probability over all samples belonging to that class of the vertical axis. Note that the samples used for computing the average probability include both ordinary clean samples and adversarial examples generated based on PGD.

Intuitively, compared to data-driven adversarial defense methods, the probability redundancy challenge in data-free settings is more significant. This is reflected in two aspects: on one hand, more probability mass is assigned to non-target classes, leading to insufficient confidence in the target class. On the other hand, classes semantically unrelated to the target class also receive more probability allocation. For example, when the target class is '0', i.e., *airplane*, redundant probability mass is allocated to irrelevant classes such as *cat* and *dog*. This indicates that the model has not adaptively learned sparse probabilities and generalizable semantics.

C Theoretical Analysis

We provide two complementary results. The first result explains why a data-free distilled student copy can act as a surrogate for discovering vulnerable directions, without requiring a robust teacher. The second result shows how the real-distribution robust risk is controlled by the teacher quality, robust anchoring error on synthetic data, and the coverage discrepancy between synthetic and real distributions. All proofs are deferred to Appendix D.

C.1 Student Copy as a Surrogate for Attack Directions

Let F be the naturally trained teacher and S_0 be its frozen data-free distilled copy. We simplify the representation of the adversarial perturbation as δ, and define the teacher's local output sensitivity and the teacher–copy surrogate disagreement as:

$$V_F(x) = \max_{\|\delta\|_p \le \epsilon} \mathrm{TV}\left(p_F(x), p_F(x+\delta)\right), \tag{A1}$$

and

$$V_{F,S_0}(x) = \max_{\|\delta\|_p \le \epsilon} \mathrm{TV}\left(p_F(x), p_{S_0}(x+\delta)\right), \tag{A2}$$

respectively. Suppose that the distilled copy locally approximates the teacher as:

$$\sup_{\|\delta\|_p \le \epsilon} \mathrm{KL}\left(p_F(x+\delta) \| p_{S_0}(x+\delta)\right) \le \kappa(x). \tag{A3}$$

Let $\alpha(x) = \sqrt{\kappa(x)/2}$. In addition, assume that the surrogate probabilities are lower bounded by a constant $\rho > 0$.

Theorem 1 (Student-copy attack-direction surrogate)**.** *For every input x, the surrogate disagreement approximates the teacher's local sensitivity as:*

$$|V_{F,S_0}(x) - V_F(x)| \le \alpha(x). \tag{A4}$$

Moreover, we define:

$$K_0(x,\delta) = \mathrm{KL}\left(p_F(x) \| p_{S_0}(x+\delta)\right), \tag{A5}$$

and let $\hat{\delta}$ be a $\xi(x)$-approximate maximizer of K_0:

$$K_0(x,\hat{\delta}) \ge \max_{\|\delta\|_p \le \epsilon} K_0(x,\delta) - \xi(x). \tag{A6}$$

Then we have:

$$\mathrm{TV}\left(p_F(x), p_F(x+\hat{\delta})\right) \ge \frac{1}{2}\sqrt{\rho\left[2\big(V_F(x) - \alpha(x)\big)_+^2 - \xi(x)\right]_+} - \alpha(x). \tag{A7}$$

Interpretation. The theorem does not claim that S_0 possesses robustness. Instead, S_0 serves as a local probe of the natural teacher. When the data-free distillation error is small, the directions that maximize the teacher–copy disagreement closely match the directions along which the teacher itself changes most rapidly. Therefore, maximizing the teacher–copy KL divergence provides informative adversarial directions without access to a robust teacher. The approximation becomes tighter when the local copy error $\alpha(x)$ and the inner optimization error $\xi(x)$ decrease.

C.2 Data-Free Robust Generalization

Let $\hat{y}_F(x) = \arg\max_c p_F^c(x)$ be the teacher pseudo-label and define the teacher margin as:

$$\gamma_F(x) = p_F^{\hat{y}_F(x)}(x) - \max_{c \ne \hat{y}_F(x)} p_F^c(x). \tag{A8}$$

For the robustly trained student S_{θ_S}, define the worst-case robust anchoring error as:

$$B_{\theta_S}(x) = \max_{\|\delta\|_p \le \epsilon} \mathrm{KL}\left(p_F(x) \| p_{S_{\theta_S}}(x+\delta)\right). \tag{A9}$$

For $M \ge \gamma_0^2/2$, let:

$$\bar{B}_{\theta_S,M}(x) = \min\{B_{\theta_S}(x), M\}, \tag{A10}$$

and let:

$$\widehat{\mathcal{R}}_{\mathcal{D}_G,M}(S_{\theta_S}, F) = \frac{1}{n}\sum_{i=1}^{n} \bar{B}_{\theta_S,M}(x_i), \qquad x_i \overset{\text{i.i.d.}}{\sim} Q_G. \tag{A11}$$

We further define the task-dependent discrepancy as:

$$\mathrm{Disc}_M(P_X, Q_G) = \sup_{f \in \mathcal{F}_M} \left|\mathbb{E}_{P_X}[f(x)] - \mathbb{E}_{Q_G}[f(x)]\right|, \tag{A12}$$

where

$$\mathcal{F}_M = \left\{x \mapsto \bar{B}_{\theta_S,M}(x) : S_{\theta_S} \in \mathcal{H}\right\}. \tag{A13}$$

Theorem 2 (Data-free robust generalization bound)**.** *Assume that the generator is fixed before sampling and that $\mathcal{D}_G = \{x_i\}_{i=1}^n$ consists of independent samples from Q_G. For any $\gamma_0 > 0$, with probability at least $1 - \delta_0$, the following holds simultaneously for all $S_\theta \in \mathcal{H}$:*

$$R_P^{\mathrm{rob}}(S_{\theta_S}) \leq$$

$$R_P(F) + \mathbb{P}_{P_X}\left[\gamma_T(x) \leq \gamma_0\right] + \frac{2}{\gamma_0^2}\left[\widehat{\mathcal{R}}_{\mathcal{D}_G,M}(S_{\theta_S}, T) + 2\mathfrak{R}_n(\mathcal{F}_M) + M\sqrt{\frac{\log(1/\delta_0)}{2n}} + \mathrm{Disc}_M(P_X, Q_G)\right]. \tag{A14}$$

Interpretation. The bound decomposes the real-data robust risk into four intuitive sources. First, $R_P(F)$ is the natural error inherited from the teacher: a student cannot be guaranteed correct when its semantic anchor is wrong. Second, low-margin teacher samples are intrinsically ambiguous, and hence are difficult to certify as robust. Third, the empirical robust anchoring term measures whether the student preserves the teacher's clean prediction throughout synthetic local neighborhoods. Finally, $\mathrm{Disc}_M(P_X, Q_G)$ quantifies the coverage gap between synthetic and real data. Consequently, robust generalization is improved by a more accurate and confident teacher, lower robust anchoring loss, a less complex hypothesis class, more synthetic samples, and better generator coverage.

D Proofs of Theoretical Results

In this appendix, we provide the complete proofs of Theorems 1 and 2.

D.1 Proof of Theorem 1

We first formalize the conditions required by Theorem 1 and establish two auxiliary lemmas.

Assumption 1 (Local surrogate approximation)**.** *For every $x \in \mathcal{X}$, there exists $\kappa(x) \geq 0$ such that:*

$$\sup_{\|\delta\|_p \leq \epsilon} \mathrm{KL}\left(p_F(x+\delta) \| p_{S_0}(x+\delta)\right) \leq \kappa(x). \tag{A15}$$

We define:

$$\alpha(x) = \sqrt{\kappa(x)/2}. \tag{A16}$$

Assumption 2 (Probability floor of the surrogate)**.** *There exists a constant $\rho > 0$ such that, for every class $c \in \{1, \ldots, C\}$ and every relevant input $u \in \mathcal{B}_\epsilon(x)$ we have:*

$$p_{S_0}^c(u) \geq \rho. \tag{A17}$$

Assumption 1 requires the distilled copy S_0 to approximate F locally, but does not require either model to be robust. Assumption 2 prevents the surrogate probabilities from becoming exactly zero. It can be satisfied in practice through temperature smoothing, logit clipping, or numerical probability clipping.

Lemma 1 (Pinsker's inequality)**.** *For any two probability distributions $p, q \in \Delta^{C-1}$:*

$$\mathrm{TV}(p, q) \leq \sqrt{\frac{1}{2}\mathrm{KL}(p\|q)}. \tag{A18}$$

Equivalently, we have:

$$\mathrm{KL}(p\|q) \geq 2\,\mathrm{TV}^2(p, q). \tag{A19}$$

Proof. Let

$$A = \{c : p_c \geq q_c\}. \tag{A20}$$

By the definition of total variation, we have:

$$\mathrm{TV}(p, q) = p(A) - q(A). \tag{A21}$$

Let $a = p(A)$ and $b = q(A)$. By the data-processing inequality for KL divergence under the binary mapping, $c \mapsto \mathbb{1}\{c \in A\}$, we have:

$$\mathrm{KL}(p\|q) \geq a \log \frac{a}{b} + (1-a) \log \frac{1-a}{1-b}. \tag{A22}$$

For fixed b, we define:

$$h(a) = a \log \frac{a}{b} + (1-a) \log \frac{1-a}{1-b} - 2(a-b)^2. \tag{A23}$$

Then, considering that:

$$h''(a) = \frac{1}{a(1-a)} - 4 \geq 0, \tag{A24}$$

and $h(b) = h'(b) = 0$, it follows that $h(a) \geq 0$. Therefore, we have:

$$\mathrm{KL}(p\|q) \geq 2(a-b)^2 = 2\,\mathrm{TV}^2(p,q), \tag{A25}$$

which proves the result. □

Lemma 2 (Two-sided relation between KL and TV)**.** *Under Assumption 2, for any probability distributions $p, q \in \Delta^{C-1}$ satisfying $q_c \geq \rho$ for every c, we have:*

$$2\,\mathrm{TV}^2(p,q) \leq \mathrm{KL}(p\|q) \leq \frac{4}{\rho}\,\mathrm{TV}^2(p,q). \tag{A26}$$

Proof. The lower bound follows immediately from Lemma 1. For the upper bound, using $\log u \leq u - 1$, we obtain:

$$\begin{aligned}
\mathrm{KL}(p\|q) &= \sum_{c=1}^{C} p_c \log \frac{p_c}{q_c} \\
&\leq \sum_{c=1}^{C} p_c \left(\frac{p_c}{q_c} - 1 \right) \\
&= \sum_{c=1}^{C} \frac{p_c^2}{q_c} - 1 \\
&= \sum_{c=1}^{C} \frac{(p_c - q_c)^2}{q_c} \\
&\leq \frac{1}{\rho} \sum_{c=1}^{C} (p_c - q_c)^2 \\
&= \frac{1}{\rho} \|p - q\|_2^2 \\
&\leq \frac{1}{\rho} \|p - q\|_1^2 \\
&= \frac{4}{\rho}\,\mathrm{TV}^2(p,q).
\end{aligned} \tag{A27}$$

This proves the upper bound. □

Proof of Theorem 1. We first prove the approximation result in Eq. equation A4.

By Assumption 1 and Lemma 1, for every admissible perturbation δ, we have:

$$\begin{aligned}
&\mathrm{TV}\left(p_F(x+\delta), p_{S_0}(x+\delta)\right) \\
&\leq \sqrt{\frac{1}{2}\,\mathrm{KL}\left(p_F(x+\delta)\|p_{S_0}(x+\delta)\right)} \\
&\leq \sqrt{\kappa(x)/2} = \alpha(x).
\end{aligned} \tag{A28}$$

Using the triangle inequality of total variation distance, we have:

$$\begin{aligned}
&\mathrm{TV}\left(p_F(x), p_{S_0}(x+\delta)\right) \\
&\quad \leq \mathrm{TV}\left(p_F(x), p_F(x+\delta)\right) + \mathrm{TV}\left(p_F(x+\delta), p_{S_0}(x+\delta)\right) \\
&\quad \leq \mathrm{TV}\left(p_F(x), p_F(x+\delta)\right) + \alpha(x).
\end{aligned} \tag{A29}$$

Taking the maximum over all $\|\delta\|_p \le \epsilon$ yields:

$$V_{F,S_0}(x) \le V_F(x) + \alpha(x). \tag{A30}$$

Conversely, the triangle inequality also implies:

$$\begin{aligned} &\mathrm{TV}\left(p_F(x), p_F(x+\delta)\right) \\ &\quad \le \mathrm{TV}\left(p_F(x), p_{S_0}(x+\delta)\right) + \mathrm{TV}\left(p_{S_0}(x+\delta), p_F(x+\delta)\right) \\ &\quad \le \mathrm{TV}\left(p_F(x), p_{S_0}(x+\delta)\right) + \alpha(x). \end{aligned} \tag{A31}$$

Maximizing over δ gives:

$$V_F(x) \le V_{F,S_0}(x) + \alpha(x). \tag{A32}$$

Combining Eqs. equation A30 and equation A32, we obtain:

$$|V_{T,S_0}(x) - V_T(x)| \le \alpha(x), \tag{A33}$$

which proves Eq. equation A4.

We then prove the attack-quality result in Eq. equation A7. Recall that:

$$K_0(x,\delta) = \mathrm{KL}\left(p_F(x) \| p_{S_0}(x+\delta)\right). \tag{A34}$$

Let δ_{TV} be a maximizer of the surrogate disagreement:

$$\delta_{\mathrm{TV}} \in \arg\max_{\|\delta\|_p \le \epsilon} \mathrm{TV}\left(p_F(x), p_{S_0}(x+\delta)\right). \tag{A35}$$

Therefore, we have:

$$\mathrm{TV}\left(p_F(x), p_{S_0}(x+\delta_{\mathrm{TV}})\right) = V_{F,S_0}(x). \tag{A36}$$

By the lower bound in Lemma 2:

$$K_0(x,\delta_{\mathrm{TV}}) \ge 2V_{F,S_0}^2(x). \tag{A37}$$

Since $\hat{\delta}$ is a $\xi(x)$-approximate maximizer of K_0, we have:

$$\begin{aligned} K_0(x,\hat{\delta}) &\ge \max_{\|\delta\|_p \le \epsilon} K_0(x,\delta) - \xi(x) \\ &\ge K_0(x,\delta_{\mathrm{TV}}) - \xi(x) \\ &\ge 2V_{F,S_0}^2(x) - \xi(x). \end{aligned} \tag{A38}$$

Applying the upper bound in Lemma 2 to $p_F(x)$ and $p_{S_0}(x+\hat{\delta})$ gives:

$$K_0(x,\hat{\delta}) \le \frac{4}{\rho}\,\mathrm{TV}^2\left(p_T(x), p_{S_0}(x+\hat{\delta})\right). \tag{A39}$$

Combining Eqs. equation A38 and equation A39, we obtain:

$$\mathrm{TV}\left(p_F(x), p_{S_0}(x+\hat{\delta})\right) \ge \frac{1}{2}\sqrt{\rho\left[2V_{F,S_0}^2(x) - \xi(x)\right]_+}. \tag{A40}$$

From Eq. equation A4,

$$V_{F,S_0}(x) \ge \left(V_F(x) - \alpha(x)\right)_+. \tag{A41}$$

Moreover, the reverse triangle inequality and Eq. equation A28 imply:

$$\begin{aligned} &\mathrm{TV}\left(p_F(x), p_F(x+\hat{\delta})\right) \\ &\quad \ge \mathrm{TV}\left(p_F(x), p_{S_0}(x+\hat{\delta})\right) - \mathrm{TV}\left(p_{S_0}(x+\hat{\delta}), p_F(x+\hat{\delta})\right) \\ &\quad \ge \mathrm{TV}\left(p_F(x), p_{S_0}(x+\hat{\delta})\right) - \alpha(x). \end{aligned} \tag{A42}$$

Substituting Eqs. equation A40 and equation A41 into Eq. equation A42 yields:

$$\mathrm{TV}\left(p_F(x), p_F(x+\hat{\delta})\right) \ge \frac{1}{2}\sqrt{\rho\left[2\left(V_F(x) - \alpha(x)\right)_+^2 - \xi(x)\right]_+} - \alpha(x), \tag{A43}$$

which proves Eq. equation A7. □

D.2 Proof of Theorem 2

We now prove the data-free robust generalization bound. Define the natural classification risk of the teacher as:

$$R_P(F) = \mathbb{P}_{(x,y)\sim P}\left[\hat{y}_F(x) \neq y\right]. \tag{A44}$$

Assumption 3 (Fixed generator and independent synthetic samples)**.** *The generator G is fixed before the synthetic training set is sampled. The synthetic set $\mathcal{D}_G = \{x_i\}_{i=1}^n$ consists of independent samples:*

$$x_i \overset{\text{i.i.d.}}{\sim} Q_G. \tag{A45}$$

Furthermore, the clipping threshold satisfies:

$$M \geq \frac{\gamma_0^2}{2}. \tag{A46}$$

Lemma 3 (Teacher-margin preservation)**.** *Suppose that an example (x, y) satisfies:*

$$\hat{y}_F(x) = y, \qquad \gamma_F(x) > \gamma_0, \qquad B_{\theta_S}(x) < \frac{\gamma_0^2}{2}. \tag{A47}$$

Then, for every perturbation δ satisfying $\|\delta\|_p \leq \epsilon$, we have:

$$\arg\max_c p^c_{S_{\theta_S}}(x+\delta) = y. \tag{A48}$$

Proof. By the definition of $B_\theta(x)$, for every admissible δ, we have:

$$\mathrm{KL}\left(p_F(x) \| p_{S_{\theta_S}}(x+\delta)\right) \leq B_{\theta_S}(x) < \frac{\gamma_0^2}{2}. \tag{A49}$$

Using Lemma 1, we obtain:

$$\begin{aligned} &\mathrm{TV}\left(p_F(x), p_{S_{\theta_S}}(x+\delta)\right) \\ &\leq \sqrt{\frac{1}{2}\,\mathrm{KL}\left(p_F(x) \| p_{S_{\theta_S}}(x+\delta)\right)} \\ &< \frac{\gamma_0}{2}. \end{aligned} \tag{A50}$$

For any two probability distributions p and q, every coordinate satisfies:

$$|p_c - q_c| \leq \mathrm{TV}(p, q). \tag{A51}$$

Hence, for any competing class $c \neq \hat{y}_T(x)$, we have:

$$\begin{aligned} &p^{\hat{y}_F(x)}_{S_{\theta_S}}(x+\delta) - p^c_{S_{\theta_S}}(x+\delta) \\ &\geq p_F^{\hat{y}_F(x)}(x) - p_F^c(x) - 2\,\mathrm{TV}\left(p_F(x), p_{S_{\theta_S}}(x+\delta)\right) \\ &> \gamma_F(x) - \gamma_0 > 0. \end{aligned} \tag{A52}$$

Thus, $\hat{y}_F(x)$ remains the student's top class throughout the entire perturbation set. Since $\hat{y}_F(x) = y$, the student is robustly correct at (x, y). □

Lemma 4 (Uniform generalization on synthetic data)**.** *Let*

$$\mathcal{F}_M = \left\{x \mapsto \bar{B}_{\theta_S, M}(x) : S_{\theta_S} \in \mathcal{H}\right\}, \tag{A53}$$

where every $f \in \mathcal{F}_M$ takes values in $[0, M]$. Define the expected Rademacher complexity as:

$$\mathfrak{R}_n(\mathcal{F}_M) = \mathbb{E}_{\mathcal{D}_G, \boldsymbol{\sigma}}\left[\sup_{f\in\mathcal{F}_M} \frac{1}{n}\sum_{i=1}^n \sigma_i f(x_i)\right], \tag{A54}$$

where $\sigma_1, \ldots, \sigma_n$ are independent Rademacher random variables. Under Assumption 3, with probability at least $1 - \delta_0$, the following holds simultaneously for all $S_{\theta_S} \in \mathcal{H}$:

$$\mathbb{E}_{Q_G}[\bar{B}_{\theta_S, M}(x)] \leq \widehat{\mathcal{R}}_{\mathcal{D}_G, M}(S_{\theta_S}, T) + 2\mathfrak{R}_n(\mathcal{F}_M) + M\sqrt{\frac{\log(1/\delta_0)}{2n}}. \tag{A55}$$

Proof. We define

$$\Phi(\mathcal{D}_G) = \sup_{f \in \mathcal{F}_M} \left(\mathbb{E}_{Q_G}[f(x)] - \frac{1}{n} \sum_{i=1}^{n} f(x_i) \right). \tag{A56}$$

By the standard symmetrization argument, we have:

$$\mathbb{E}_{\mathcal{D}_G}[\Phi(\mathcal{D}_G)] \leq \mathbb{E}_{\mathcal{D}_G, \mathcal{D}'_G} \left[\sup_{f \in \mathcal{F}_M} \frac{1}{n} \sum_{i=1}^{n} \big(f(x'_i) - f(x_i) \big) \right] \leq 2\mathfrak{R}_n(\mathcal{F}_M), \tag{A57}$$

where $\mathcal{D}'_G = \{x'_i\}_{i=1}^n$ is an independent ghost sample from Q_G.

Replacing one element of $\mathcal{D}_G$ can change $\Phi(\mathcal{D}_G)$ by at most M/n, because every $f \in \mathcal{F}_M$ takes values in $[0, M]$. Therefore, with probability at least $1 - \delta_0$, McDiarmid's inequality gives:

$$\Phi(\mathcal{D}_G) \leq \mathbb{E}_{\mathcal{D}_G}[\Phi(\mathcal{D}_G)] + M\sqrt{\frac{\log(1/\delta_0)}{2n}}. \tag{A58}$$

Combining Eqs. equation A57 and equation A58, and choosing $f(x) = \bar{B}_{\theta_S, M}(x)$, proves Eq. equation A55. □

Proof of Theorem 2. Define the robust error event as:

$$\mathcal{E}_{\text{rob}} = \left\{ (x, y) : \exists \delta,\ \|\delta\|_p \leq \epsilon,\ \arg\max_c p^c_{S_{\theta_S}}(x + \delta) \neq y \right\}. \tag{A59}$$

By Lemma 3, if a robust error occurs, at least one of the following events must occur:

$$\begin{aligned} \mathcal{E}_T &= \{\hat{y}_T(x) \neq y\}, \\ \mathcal{E}_{\text{margin}} &= \{\gamma_T(x) \leq \gamma_0\}, \\ \mathcal{E}_B &= \{B_{\theta_S}(x) \geq \gamma_0^2/2\}. \end{aligned} \tag{A60}$$

Consequently, we have:

$$\mathcal{E}_{\text{rob}} \subseteq \mathcal{E}_T \cup \mathcal{E}_{\text{margin}} \cup \mathcal{E}_B. \tag{A61}$$

Applying the union bound gives:

$$R_P^{\text{rob}}(S_{\theta_S}) \leq R_P(F) + \mathbb{P}_{P_X}\left[\gamma_F(x) \leq \gamma_0\right] + \mathbb{P}_{P_X}\left[B_{\theta_S}(x) \geq \frac{\gamma_0^2}{2}\right]. \tag{A62}$$

Since $M \geq \gamma_0^2/2$, the event $B_{\theta_S}(x) \geq \gamma_0^2/2$ implies:

$$\bar{B}_{\theta_S, M}(x) \geq \frac{\gamma_0^2}{2}. \tag{A63}$$

Markov's inequality therefore yields:

$$\mathbb{P}_{P_X}\left[B_{\theta_S}(x) \geq \frac{\gamma_0^2}{2}\right] \leq \mathbb{P}_{P_X}\left[\bar{B}_{\theta_S, M}(x) \geq \frac{\gamma_0^2}{2}\right] \leq \frac{2}{\gamma_0^2}\mathbb{E}_{P_X}[\bar{B}_{\theta_S, M}(x)]. \tag{A64}$$

Substituting Eq. equation A64 into Eq. equation A62, we have:

$$R_P^{\text{rob}}(S_\theta) \leq R_P(F) + \mathbb{P}_{P_X}\left[\gamma_F(x) \leq \gamma_0\right] + \frac{2}{\gamma_0^2}\mathbb{E}_{P_X}[\bar{B}_{\theta_S, M}(x)]. \tag{A65}$$

By the definition of the task-dependent discrepancy, we have:

$$\mathbb{E}_{P_X}[\bar{B}_{\theta_S, M}(x)] \leq \mathbb{E}_{Q_G}[\bar{B}_{\theta_S, M}(x)] + \text{Disc}_M(P_X, Q_G). \tag{A66}$$

Finally, applying Lemma 4 to the first term on the right-hand side of Eq. equation A66 gives:

$$\mathbb{E}_{P_X}[\bar{B}_{\theta_S, M}(x)] \leq \widehat{\mathcal{R}}_{\mathcal{D}_G, M}(S_{\theta_S}, T) + 2\mathfrak{R}_n(\mathcal{F}_M) + M\sqrt{\frac{\log(1/\delta_0)}{2n}} + \text{Disc}_M(P_X, Q_G). \tag{A67}$$

Substituting Eq. equation A67 into Eq. equation A65 proves Eq. equation A14. □

Table 7: Restult of Tiny-ImageNet on PreActResNet-18.

Methods	Clean			FGSM			PGD			CW		AA	
	Avg.	Worst	NSD	Avg.	Worst	NSD	Avg.	Worst	NSD	Avg.	NSD	Avg.	NSD
CMI	38.56	11.90	0.445	16.46	1.10	0.785	13.43	0.00	0.882	9.93	1.109	9.23	1.152
DeepInv	36.25	9.70	0.465	13.46	0.50	0.861	10.95	0.00	0.958	7.32	1.202	6.58	1.288
Fast	33.76	7.80	0.473	12.94	0.40	0.804	11.19	0.00	0.846	7.66	1.087	7.11	1.138
DERD	36.32	9.40	0.470	16.10	1.60	0.737	14.32	0.20	0.779	10.68	0.977	9.67	1.044
DFARD	39.37	11.60	0.448	17.66	1.30	0.750	14.44	0.30	0.804	11.80	0.967	10.84	1.046
DFHL	23.13	1.10	0.719	6.87	0.00	1.368	5.51	0.00	1.555	5.43	1.569	4.95	1.654
FERD	39.64	12.30	0.439	18.00	2.10	**0.722**	**15.94**	0.40	0.783	12.19	**0.937**	11.24	1.007
DFAT (Ours)	32.41	8.16	0.457	15.39	1.60	0.758	13.77	**0.40**	0.776	10.96	0.950	10.39	1.112
DFAT + DENT	**39.76**	**12.63**	**0.419**	**18.14**	**2.40**	0.726	15.79	0.60	**0.750**	**12.63**	0.964	**11.30**	**1.007**

Table 8: Evaluation of adaptive attacks on ResNet-18 for CIFAR-10. A-FGSM means the adaptive FGSM attack that targets the DENT objective function, and a similar notation applies to other attacks. The prefix "A-" means that the attack is adaptive.

Defense	Clean	A-FGSM	A-$\mathrm{PGD_S}$	A-CW	Avg.
Natural	91.74	6.41	2.52	2.26	24.98
DFAT	82.41	49.06	40.68	39.04	52.79
DFAT + DENT	**85.44**	**54.13**	**49.27**	**48.53**	**59.34**

E Extended Experiments

E.1 Detailed Settings

Implementation Details. We employ the SGD optimizer with a momentum of 0.9 and weight decay of 5e-4 to train both the naturally trained teacher and student copy, and the batch size is $m = 512$. For the student copy, stage 1 includes 2000 epochs of natural training, followed by stage 2, which consists of 200 epochs of robustness learning. The initial learning rates are set to 0.001. Table 6 provide the evaluation results for the naturally pre-trained models. The hyper-parameters in stage 1 follow the settings in DERD Zhou et al. (2024), while those in stage 2 are set as $\lambda_{\mathrm{ND}} = 1$, $\lambda_{\mathrm{AD}} = 1$, $\lambda_{\mathrm{PS}} = 0.01$, $\lambda_{\mathrm{a}} = 0.006$. For the attack settings, following the common settings Li et al. (2026); Zhou et al. (2024), four commonly used the white-box attacks are evaluated including: FGSM Goodfellow et al. (2014), $\mathrm{PGD_S}$ Madry et al. (2017), $\mathrm{PGD_T}$ Zhang et al. (2019), CW Carlini & Wagner (2017), and AutoAttack Croce & Hein (2020). The L_∞ norm attack budget is $\epsilon = 8/255$, the perturbation step size is $\eta_1 = 2/255$, and the number of iterations is $K = 10$. Additionally, besides white-box attacks, we also briefly evaluate the student's robustness against black-box attacks on CIFAR-10, including transfer-based and query-based black-box attacks.

E.2 Experiments on Tiny-ImageNet

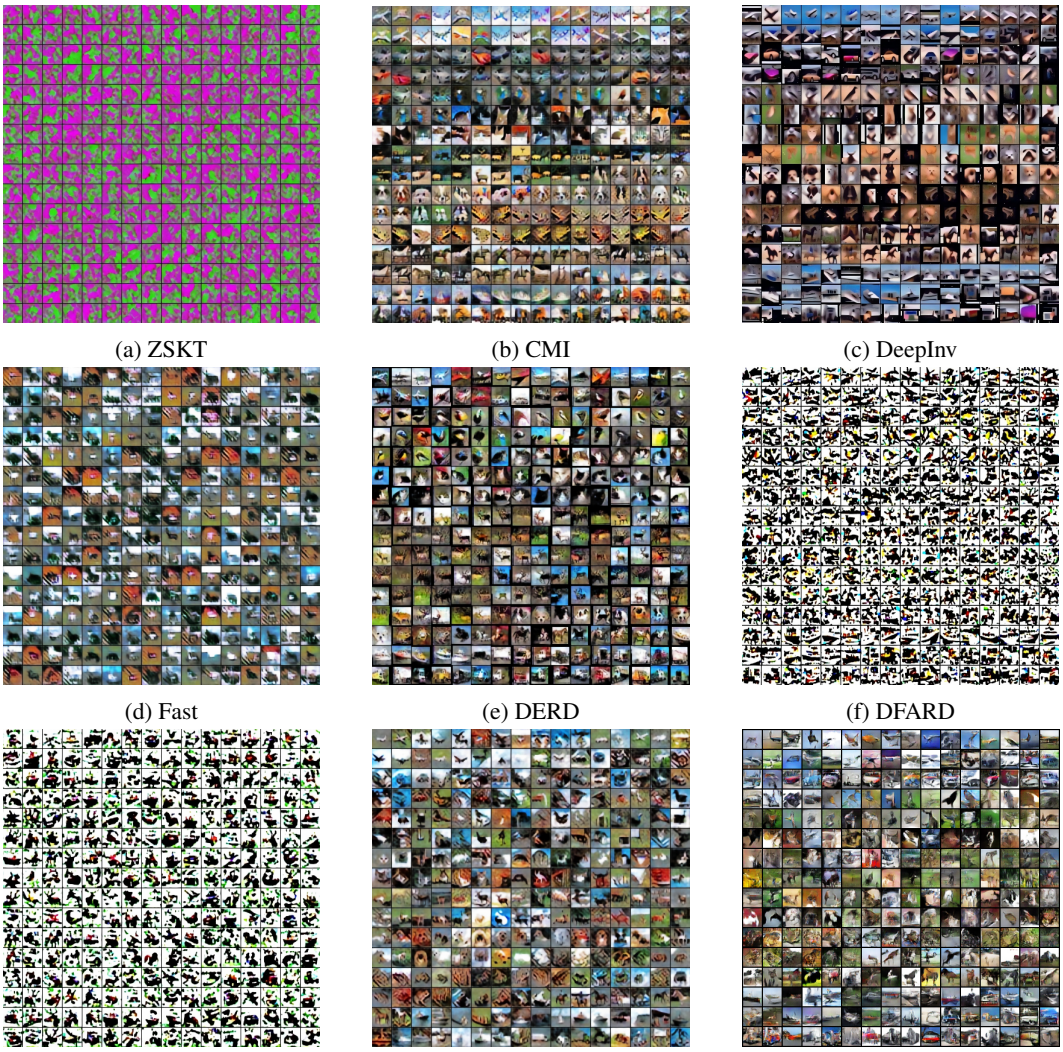
In this subsection, we provide experimental results on Tiny-ImageNet and compare them with several current methods. The experimental results are shown in Figure 7, where DFAT still demonstrates a certain degree of superiority.

E.3 Evaluation on the adaptive attacks

A common evaluation paradigm for assessing the reliability of the TTA-based method family is adaptive attacks Shi et al. (2021); Wang et al. (2021). To further demonstrate the stability of combining DFAT with TTA, we provide a brief discussion of adaptive attacks here. Specifically, an adaptive attacker not only attacks the model itself, but also incorporates the optimization objective of TTA into the attack target, which can be formulated as:

$$\arg\max_{\delta_x} \mathcal{L}_{\mathrm{CE}} + \mathcal{L}_{\mathrm{DENT}}, \tag{A68}$$

where $\mathcal{L}_{\mathrm{CE}}$ is the traditional cross-entropy loss, and $\mathcal{L}_{\mathrm{DENT}}$ is the optimization loss of the TTA stage Wang et al. (2021). We construct adaptive attacks based on the commonly used FGSM, PGD,



(a) ZSKT (b) CMI (c) DeepInv

(d) Fast (e) DERD (f) DFARD

(g) DFHL (h) FERD (i) **DFAT (ours)**

Figure 10: Visualization of the generated surrogate data.

and CW methods, and the experimental results are shown in Table 8. Intuitively, the combination of DFAT and DENT exhibits considerable robustness against adaptive attacks as well, demonstrating its reliable robustness without suffering from the "obfuscated gradient" vulnerability.

E.4 Surrogate data visualization

We here provide visualization results of the surrogate data from existing data-free knowledge distillation/robustness distillation methods. It is worth noting that methods incorporating $\mathcal{L}_{\mathrm{BN}}$ exhibit significantly better semantic representations (i.e., CMI, DERD, FERD, and DFAT), whereas methods without statistical prior support show slightly inferior visual interpretability. Furthermore, compared to FERD, DERD and DFAT incorporate the feature activation of L_{a}, which appears visually smoother, while FERD seems to exhibit a certain texture preference. In summary, the surrogate samples of DFAT conform to class semantics, providing a certain degree of interpretability from the pixel perspective.

F LIMITATION AND DISCUSSION

The proposed DFAT in this paper aims to eliminate the dependence of current data-free defense methods on robust teachers, i.e., transitioning from data-free robust distillation to data-free adversarial training. Although effective, it still has certain limitations. (1) Under the same conditions (i.e., without the TTA module), the performance of DFAT remains significantly lower than that of data-driven methods, which may still be a common challenge in the data-free community. (2) The acquisition of surrogate data still relies on considerable computational resources and time consumption, reducing the time efficiency of the data-free method family. Although time efficiency is not the focus of this paper, higher defense efficiency is also a hot topic in the adversarial community. (3) Current data-free methods are still mainly focused on classification tasks. Data-free defense methods for multimodal large models and the "pre-training + fine-tuning" paradigm still require further research. We intend to leave the above issues for future work.