Shedding More Light on Robust Classifiers under the lens of Energy-based Models

Mujtaba Hussain Mirza[1], Maria Rosaria Briglia[1], Senad Beadini[2], and Iacopo Masi[1]

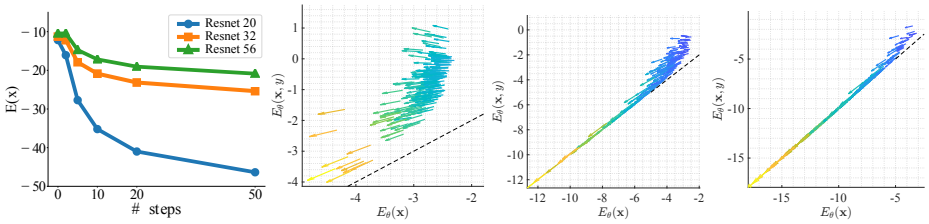
[1] OmnAI Lab, CS Department, Sapienza University of Rome, Italy
[2] Eustema S.p.A. Italy
{mirza,briglia,masi}@di.uniroma1.it
s.beadini@eustema.it

Abstract. By reinterpreting a robust discriminative classifier as Energy-based Model (EBM), we offer a new take on the dynamics of adversarial training (AT). Our analysis of the energy landscape during AT reveals that *untargeted* attacks generate adversarial images much more in-distribution (lower energy) than the original data *from the point of view of the model*. Conversely, we observe the opposite for *targeted* attacks. On the ground of our thorough analysis, we present new theoretical and practical results that show how interpreting AT energy dynamics unlocks a better understanding: (1) AT dynamic is governed by three phases and robust overfitting occurs in the third phase with a drastic divergence between natural and adversarial energies (2) by rewriting TRADES loss in terms of energies, we show that TRADES implicitly alleviates overfitting by means of aligning the natural energy with the adversarial one (3) we empirically show that all recent state-of-the-art robust classifiers are smoothing the energy landscape and we reconcile a variety of studies about understanding AT and weighting the loss function under the umbrella of EBMs. Motivated by rigorous evidence, we propose Weighted Energy Adversarial Training (WEAT), a novel sample weighting scheme that yields robust accuracy matching the state-of-the-art on multiple benchmarks such as CIFAR-10 and SVHN and going beyond in CIFAR-100 and Tiny-ImageNet. We further show that robust classifiers vary in the intensity and quality of their generative capabilities, and offer a simple method to push this capability, reaching a remarkable Inception Score (IS) and FID using a robust classifier without training for generative modeling. The code to reproduce our results is available at github.com/OmnAI-Lab/Robust-Classifiers-under-the-lens-of-EBM.

Keywords: robustness · adversarial training · energy-based models

1 Introduction

Ten years ago the seminal paper by Szegedy *et al*. [50] was released arguing about "intriguing properties of neural networks". Those properties revealed that deep nets exhibit unconventional traits concerning their abrupt transitions w.r.t. to small perturbations of the input, i.e. adversarial attacks. During the last decade,



(a) $E_{\boldsymbol{\theta}}(\mathbf{x})$ vs PGD steps **(b)** Epoch 1 **(c)** Epoch 50 **(d)** Epoch 100

Fig. 1: (a) PGD untargeted attacks create points that heavily bias the energy landscape. Plot shows $E_{\boldsymbol{\theta}}(\mathbf{x})$ in function of PGD steps, across non-robust networks of various depths on CIFAR-10. CIFAR-100 is available in supp. material. **(b, c, d)** $E_{\boldsymbol{\theta}}(\mathbf{x}, y)$ in function of $E_{\boldsymbol{\theta}}(\mathbf{x})$ for a subset of CIFAR-10 training data at various stages during SAT with PGD 5 iterations. Note that the axes across figures are not in the same range for clarity. Each arrow represents the "drift" induced in the energy by the attacks: the base of the arrow is the natural data, while the tip is after the attack. The dashed black line corresponds to zero cross-entropy when $E_{\boldsymbol{\theta}}(\mathbf{x}, y) = E_{\boldsymbol{\theta}}(\mathbf{x})$. The plot can only take values above that line. Color-coded by intensity of the drift less more intense.

a plethora of algorithms have been proposed to enforce robustness in a classifier, mainly relying on adversarial training (AT) [20, 36, 52, 62] or to certify a prediction [32, 33] using randomized smoothing [8]. Improvements of AT have been reported on multiple axes: less training time [45]; more data improves robustness either from a real data distribution [5] or generated via a denoising diffusion process [22, 53]; variations such as TRADES [62] and MART [52] and in some cases solutions that are less robust than the baseline, GAIRAT [64]. The training process has also been studied from the point of view of overfitting [39]. Standard benchmarks have been proposed [9] such as `RobustBench`. Despite all these efforts, except a few rare cases [54], no notable algorithmic improvement has been reported in these years, with AT hitting a plateau in performance [21]: thus, it is not a surprise that top performing methods attain robustness simply pouring more data [5,53] or designing better architectures [38]. Regardless of performance, very little attention has been placed to *understanding* the role of AT and to *demystifying* some unexpected capabilities of robust classifiers, such as generative capability and better calibration abilities. The only work that adventures connecting robust model with generative is [66] setting the foundation to interpret AT as an Energy-based Model (EBM) [23]. Despite adversarial attacks have been recognized as input points that cross the decision boundary—thus impacting $p_{\boldsymbol{\theta}}(y|\mathbf{x})$—following [2], we illustrate a surprising yet strong correlation with $p_{\boldsymbol{\theta}}(\mathbf{x})$ for untargeted PGD attacks [36]. Going beyond [2], we extend the analysis to a vast pool of attacks such as untargeted PGD [36], targeted attacks, CW [4], TRADES (KL divergence) [62] AutoAttack [11] and show that different attacks induce difference shifts in the energy landscape. We go beyond the study of [51, 66] by offering a novel interpretation of TRADES [62] as an EBM. This interpretation sheds light on how TRADES outperforms SAT [36] by mitigating robust overfitting, and provides a more fine-grained analysis on the generative

capabilities of robust classifiers. We finally bring our insights about the energy landscape into the training dynamics discovering a new property that it is not explicitly enforced by AT: the more a classifier is robust, the smoother is its energy landscape; the model attains this implicitly by reconciling the range of energies of natural data with those of adversarial data. To show how untargeted Projected Gradient Descent (PGD) bends the energy landscape, following [2], we attacked *non-robust* residual classifiers with PGD and recorded the average energy of the adversarial points in functions of the PGD steps. Fig. 1a shows also a strong dependency between the number of iterations taken and the marginal energy tending to be negative. Note that although there is a steep decrease in the energy, the attack is still norm-bounded in the input by $\boldsymbol{\epsilon}$. We also note how attacks to deeper models bend way less the energy. The dynamic on how AT compensates for the perturbations shown in Fig. 1a can be grasped in Figs. 1b to 1d. The figure offers a visualization of the dynamics of the changes in joint energy $E_{\boldsymbol{\theta}}(\mathbf{x}, y)$ and marginal $E_{\boldsymbol{\theta}}(\mathbf{x})$ during standard adversarial training[3] (SAT) with PGD with 5 iterations. Fig. 1 shows that not only the model has to correct its prediction about a class yet has also to compensate for abrupt changes in the energy values. This figure offers important insights such as in the beginning of the training—Fig. 1b—for most of the samples holds $E_{\boldsymbol{\theta}}(\mathbf{x}, y) > E_{\boldsymbol{\theta}}(\mathbf{x})$, indicating high loss; note also how the more we approach zero loss, the easier it is to bend the energy. Then in the middle of AT—Fig. 1c—most of the vectors are at zero loss and the intensity of the attack on the energy decreases. Finally, the end of the training—Fig. 1d: attacks are successful when $E_{\boldsymbol{\theta}}(\mathbf{x})$ is high (top-right) yet even though a lot of samples have loss close to zero, i.e. $E_{\boldsymbol{\theta}}(\mathbf{x}, y) \approx E_{\boldsymbol{\theta}}(\mathbf{x})$ the attack manages still to distort the energy significantly. Leveraging on the limits of the prior art, we make the following contributions:

- ◇ We empirically show a curious effect: all top performing models in `RobustBench` share the same property of having a smooth *marginal* energy landscape. An increase in the model's robustness is correlated with a decrease of $E_{\boldsymbol{\theta}}(\mathbf{x}) - E_{\boldsymbol{\theta}}(\mathbf{x}^\star)$, which conveys energy landscape smoothness in the neighborhood of real data samples. We also explain overfitting as a drastic divergence between natural and adversarial energies.
- ◇ We further offer experiments that demystify the role of misclassification [34,52] and reconnect AT with energy and give a better explanation for the transferability of AT w.r.t. to the training samples [35]. We theoretically show how rewriting TRADES as EBM can better explain its capabilities.
- ◇ Guided by our analysis and theoretical results, we propose Weighted Energy Adversarial Training (WEAT) that yields robust accuracy matching the SOTA on CIFAR-10, and SVHN going beyond in CIFAR-100 and Tiny-ImageNet. We further show how we can push the generative capabilities of robust classifiers reaching a remarkable Inception Score (IS) and FID just using a single robust classifier, without training for generative modeling.

[3] We refer to adversarial training (AT) as a generic procedure that regards all methods for robust classifiers (SAT, TRADES, MART) while SAT indicates Standard AT [36].

2 Prior Work

Adversarial Robustness. The robustness of neural networks is a crucial topic in deep learning. Despite intensive efforts, AT [36], which incorporates adversarial examples into training, remains the most effective empirical strategy. This method has attracted considerable interest and several modifications. [62] proposed TRADES, leveraging the Kullback-Leibler (KL) divergence to balance the trade-off between standard and robust accuracy. Additionally, there are studies dedicated to exploring how DNN architecture impacts robustness [38].

Robust Classifiers and EBM. A recent connection between robust and generative models is presented in [23]. The Joint Energy-based Model (JEM) [23] reformulates the traditional softmax classifier into an EBM for hybrid discriminative-generative modeling. In [57], JEM++ was introduced to enhance training stability and speed. Subsequently, [66] established an initial link between adversarial training and energy-based models, illustrating how they manipulate the energy function differently yet share a comparable contrastive approach. Generative capabilities of robust classifiers have been studied in other works [17,51,58,59] and even employed in inverse problems [40] or controlled image synthesis [41].

Mitigation of Robust Overfitting and Additional Data. [44] first investigated robust overfitting, arguing for the need for large datasets for robust generalization. Subsequent studies have shown that larger datasets are crucial for robust models, providing empirical evidence that supports this finding: [22] illustrated that training with synthesized images from generative models leads to an improvement in robustness. [53] demonstrated that using synthesized images from more advanced generative models, such as diffusion models [26], leads to superior adversarial robustness, setting a new state-of-the-art in robust accuracy. Recently, [16] hypothesizes overfitting is due to difficult samples (hard to fit) that are closer to the decision boundary, and the network ends up memorizing instead of learning. [37] explains overfitting using their optimization objective (Self-COnsistent Robust Error (SCORE). Other works like AWP [54] adversarially perturb both inputs and model weights. [27] optimizes the trajectories of adversarial training considering its dynamics, while others [6,7,16,46,49,54] link it to the flatness of the loss function. Orthogonal to all aforementioned works, we show that overfitting is actually linked to the model, drastically increasing the discrepancy between natural and adversarial energies. Our work is connected to [60] which ascribes overfitting to data with low loss values. Nevertheless, with our formulation, we can actually show that *low* loss values correspond to attacks that bend the energy even *more* than higher values, see Figs. 1c and 1d.

Weighting the Samples in Adversarial Training. MART [52] started a line of research that shows improvement by weighting the samples in AT. GAIRAT [64] follows the trend, though was proved to be non-robust [25]. Several fixes to [64] have been proposed, such as continuous probabilistic margin (PM) [34] or weighting with entropy [28]. Unlike previous methods, we offer a new way to weight the samples using the marginal energy, which is a quantity not related to the labels and more connected with the hidden generative model inside classifiers.

3 Method

We will give an overview of the settings for adversarial attacks in a white-box scenario. Moving on, we are going to explore the modeling of data density and standard discriminative classifiers using Energy-based Models (EBMs).

Preliminaries and Objective. Consider a set of labeled images $X = \{(\mathbf{x}, y)|\mathbf{x} \in \mathbb{R}^d \text{ and } y \in \{1, .., K\}\}$, assuming that each $(\mathbf{x}, y)$ is generated from an underlying distribution $\mathcal{D}$; let be $\boldsymbol{\theta} : \mathbb{R}^d \to \mathbb{R}^K$ a classifier implemented with a DNN. The problem of learning a robust classifier can be modeled AT [36] by solving $\min_{\boldsymbol{\theta}} \mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}}\Big[\max_{\boldsymbol{\delta}\in\mathcal{S}} \mathcal{L}\big(\boldsymbol{\theta}(\mathbf{x}+\boldsymbol{\delta}),\ y\big)\Big]$, where $\mathcal{L}$ is cross-entropy loss and $\mathcal{S} = \{\boldsymbol{\delta} \in \mathbb{R}^d : ||\boldsymbol{\delta}||_p \leq \epsilon\}$ is a set of feasible ℓ_p perturbations. In this process, the attacker optimizes an adversarial point, denoted as $\mathbf{x}^\star \doteq \mathbf{x}+\boldsymbol{\delta} \in \mathbb{R}^d$ in the input space by either increasing the loss in the output space (untargeted attack) or prompting a confident incorrect label (targeted attack). For ℓ_∞, the perturbation is usually built via PGD [36]: $\mathbf{x}^\star = \mathbb{P}_\epsilon\left[\mathbf{x}^\star + \alpha\ \text{sign}\left(\ \nabla_{\mathbf{x}^\star}\mathcal{L}(\boldsymbol{\theta}(\mathbf{x}^\star),\ y)\ \right)\right]$, where $\mathbb{P}_\epsilon$ projects into the surface of $\mathbf{x}$'s neighbor ϵ-ball while α is the step size.

Discriminative Models as EBM. Energy-based models (EBM) [31] are based on the assumption that any probability density function $p(\mathbf{x})$ can be defined through a Boltzmann distribution as $p_{\boldsymbol{\theta}}(\mathbf{x}) = \frac{\exp\left(-E_{\boldsymbol{\theta}}(\mathbf{x})\right)}{Z(\boldsymbol{\theta})}$ where $E_{\boldsymbol{\theta}}(\mathbf{x})$ is known as energy, that maps each input $\mathbf{x}$ to a scalar. $Z(\boldsymbol{\theta}) = \int \exp(-E_{\boldsymbol{\theta}}(\mathbf{x}))\, d\mathbf{x}$ is the normalizing constant, such that $p_{\boldsymbol{\theta}}(\mathbf{x})$ is a proper probability density function. In the same manner, we can define the joint probability $p_{\boldsymbol{\theta}}(\mathbf{x}, y)$ in terms of energy and combining all together, we can write a traditional discriminative classifier in terms of energy and normalizing constants like:

$$p_{\boldsymbol{\theta}}(y|\mathbf{x}) = \frac{p_{\boldsymbol{\theta}}(\mathbf{x}, y)}{p_{\boldsymbol{\theta}}(\mathbf{x})} = \frac{\exp\left(-E_{\boldsymbol{\theta}}(\mathbf{x}, y)\right) Z_{\boldsymbol{\theta}}}{\exp\left(-E_{\boldsymbol{\theta}}(\mathbf{x})\right) \hat{Z}_{\boldsymbol{\theta}}} = \frac{\exp\left(\boldsymbol{\theta}(\mathbf{x})[y]\right)}{\sum_{k=1}^{K} \exp \boldsymbol{\theta}(\mathbf{x})[k]}, \tag{1}$$

where $\hat{Z}_{\boldsymbol{\theta}}$ is the normalizing constant of $p_{\boldsymbol{\theta}}(\mathbf{x}, y)$, $Z_{\boldsymbol{\theta}} = \hat{Z}_{\boldsymbol{\theta}}$ [66] and $\boldsymbol{\theta}[i]$ is i^{th} logit. Observing Eq. (1), we can deduce the definition of the energy functions as:

$$E_{\boldsymbol{\theta}}(\mathbf{x}, y) = -\log \exp\left(\boldsymbol{\theta}(\mathbf{x})[y]\right) \ \text{ and } E_{\boldsymbol{\theta}}(\mathbf{x}) = -\log \sum_{k=1}^{K} \exp\left(\boldsymbol{\theta}(\mathbf{x})[k]\right). \tag{2}$$

This framework offers a versatile approach to consider a generative model within any DNN by leveraging their logits [23].

3.1 Reconnecting Attacks with the Energy

Different Attacks Induce diverse Energy Landscapes. Following [66] and using Eq. (2), we get that the cross-entropy (CE) loss $\mathcal{L}_{\text{CE}}(\mathbf{x}, y; \boldsymbol{\theta}) = -\log\big(p_{\boldsymbol{\theta}}(y|\mathbf{x})\big) = -\boldsymbol{\theta}(\mathbf{x})[y] + \log \sum_{k=1}^{K} \exp\left(\boldsymbol{\theta}(\mathbf{x})[k]\right)$ and thus we can express it with energy as:

$$\mathcal{L}_{\text{CE}}(\mathbf{x}, y; \boldsymbol{\theta}) = \underbrace{-\boldsymbol{\theta}(\mathbf{x})[y]}_{E_{\boldsymbol{\theta}}(\mathbf{x}, y)} \underbrace{+ \log \sum_{k=1}^{K} \exp\left(\boldsymbol{\theta}(\mathbf{x})[k]\right)}_{-E_{\boldsymbol{\theta}}(\mathbf{x})} = E_{\boldsymbol{\theta}}(\mathbf{x}, y) - E_{\boldsymbol{\theta}}(\mathbf{x}). \tag{3}$$

Note by definition Eq. (3) ≥ 0 and the loss is zero when $E_{\boldsymbol{\theta}}(\mathbf{x}, y) = E_{\boldsymbol{\theta}}(\mathbf{x})$. To see how the loss used in adversarial attacks induces different changes in the energies, we can consider the maximization of Eq. (3) performed during *untargeted* PGD. At each step, PGD shifts the input by two terms $\nabla_{\mathbf{x}^\star} E_{\boldsymbol{\theta}}(\mathbf{x}^\star, y) - \nabla_{\mathbf{x}^\star} E_{\boldsymbol{\theta}}(\mathbf{x}^\star)$: a *positive* direction of $E_{\boldsymbol{\theta}}(\mathbf{x}, y)$ and a *negative* direction $E_{\boldsymbol{\theta}}(\mathbf{x})$. As found by [2], untargeted PGD finds input points that fool the classifier—high joint energy—yet are even more likely than natural data—very low marginal energy. Note that by "more likely", *we mean from the perspective of the model*, as ℓ_p attacks are known to be out of distribution and orthogonal to data manifold $p_{\text{data}}(\mathbf{x})$ [48]. To make a connection with recent denoising score-matching [47] and diffusion models [14], we can see how PGD is heavily biased by the score function i.e. $\nabla_{\mathbf{x}} \log p_{\boldsymbol{\theta}}(\mathbf{x})$ since $\nabla_{\mathbf{x}} \log p_{\boldsymbol{\theta}}(\mathbf{x}) = \nabla_{\mathbf{x}} - E_{\boldsymbol{\theta}}(\mathbf{x}) - \nabla_{\mathbf{x}} \log Z_{\boldsymbol{\theta}} = -\nabla_{\mathbf{x}} E_{\boldsymbol{\theta}}(\mathbf{x})$ where the last identity follows since $\nabla_{\mathbf{x}} \log Z_{\boldsymbol{\theta}} = 0$. On the contrary, it is interesting to reflect on how the dynamic is flipped for *targeted* attacks: assuming we target y_t, $-\nabla_{\mathbf{x}^\star} E_{\boldsymbol{\theta}}(\mathbf{x}^\star, y_t) + \nabla_{\mathbf{x}^\star} E_{\boldsymbol{\theta}}(\mathbf{x}^\star)$, the optimization lowers the joint energy yet produces new points in the opposite direction of the score—out of distribution. To empirically prove it, in addition to Figs. 1b to 1d, we probe a state-of-the-art (SOTA) non-robust model from `RobustBench` [9], namely WideResNet-28-10 and report in Fig. 2a the distribution of the marginal energies and in Fig. 2b the distribution of conditional. We employ a diverse set of state-of-the-art untargeted and targeted attacks, mainly from AutoAttack (AA) [11]. We can see how PGD drastically shifts $E_{\boldsymbol{\theta}}(\mathbf{x})$ to the left; notice also how the distributions $E_{\boldsymbol{\theta}}(\mathbf{x}, y)$ are pushed to the right, coherent with the attack logic of decreasing $p(y|\mathbf{x})$, indeed the robust accuracy is 0%. TRADES instead performs similar for $E_{\boldsymbol{\theta}}(\mathbf{x})$ yet the robust accuracy is surprisingly 30%. We can notice how $E_{\boldsymbol{\theta}}(\mathbf{x}, y)$ is divided in two modes: one mode on the right when the attack is successful; vice versa, the one on the left is actually capturing ground-truth logits that *increase* after the attack; in other words, for a small part of the data TRADES helps the classification. APGD is more subtle, as a tiny fraction of test points share similar values to natural data. The situation is flipped for targeted attacks: APGD-T moves the $E_{\boldsymbol{\theta}}(\mathbf{x})$ energy to the right so to push $E_{\boldsymbol{\theta}}(\mathbf{x}, y)$ to the target label, thereby creates points that are more out-of-distribution compared to natural samples. This behavior was already noted in [51] but not yet shown empirically for multiple SOTA attacks. FAB (Fast Adaptive Boundary) [10] behaves similarly to a targeted attack. Square and Carlini Wagner (CW) [4] are very subtle since the marginal energy completely overlaps the natural: this is visible for attacks like CW and APG-DLR that uses DLR (Difference of Logits Ratio) thereby causing less deformation in logit's space by attacking the margin. Targeted Carlini Wagner (CW-T) minimizes $\max\big(\max[\boldsymbol{\theta}(\mathbf{x}^\star)[i] : i \neq t] - \boldsymbol{\theta}(\mathbf{x}^\star)[t], -\kappa\big)$ for a target class t, decreasing the competing logit (mostly likely the gt class y) or increasing t logit. Our experiments show the former. Unlike Fig. 2b-CW, $E_{\boldsymbol{\theta}}(\mathbf{x}, y)$ now has two modes: the small one is random target labels, hard to optimize, thus overlapping with clean data. The bound on the perturbation limits the changes in $E_{\boldsymbol{\theta}}(\mathbf{x})$ because, unlike CE, there is no explicit term that pushes it to the left, so $E_{\boldsymbol{\theta}}(\mathbf{x})$ plot is similar to Fig. 2a-CW. Further details are in supp. material.

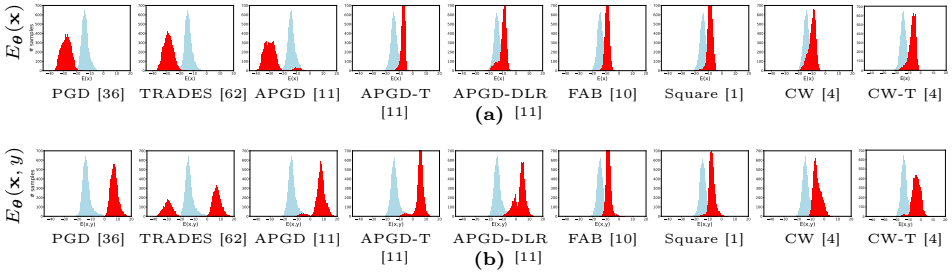


Fig. 2: **(a)** Distributions of the $E_{\boldsymbol{\theta}}(\mathbf{x})$ and **(b)** the $E_{\boldsymbol{\theta}}(\mathbf{x}, y)$ of adversarial and natural inputs for several adversarial perturbations both untargeted and targeted (-T), on CIFAR-10 test set, using a non-robust model. ■ indicates adv. and ■ natural data.

3.2 How Adversarial Training Impacts the Energy of Samples

Connecting Robust Overfitting with Energy Divergence. We find energy plays a key factor in understanding the behavior of AT, especially in the context of robust overfitting. To show this, we conduct an experiment comparing the energies of samples in the *training set* with their corresponding adversarial counterparts at each epoch during AT. Given an input image $\mathbf{x}$ and its corresponding adversarial example $\mathbf{x}^{\star}$, we measure the difference between their marginal energies, $E_{\boldsymbol{\theta}}(\mathbf{x}) - E_{\boldsymbol{\theta}}(\mathbf{x}^{\star})$, denoted by $\Delta E_{\boldsymbol{\theta}}(\mathbf{x})$. When using SAT [36], we find that the training is divided into three phases where in first two phases, the energies of original and adversarial samples exhibit comparable values. However, in the third phase, the energies $E_{\boldsymbol{\theta}}(\mathbf{x})$ and $E_{\boldsymbol{\theta}}(\mathbf{x}^{\star})$ begin to diverge from each other, implied by the steep decrease of $\Delta E_{\boldsymbol{\theta}}(\mathbf{x})$. Concurrently, we observe a simultaneous increase in test error for adv. data at this point as shown in Fig. 3, indicating robust overfitting. Thus, to alleviate robust overfitting, it seems imperative to maintain similarity in energies between original and adversarial samples, thereby smoothing the energy landscape around each sample. Interestingly, reinterpreting TRADES [62] as EBM reveals that TRADES is essentially achieving the desired objective, towards a notable mitigation of overfitting.

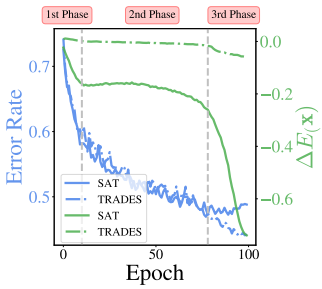


Fig. 3: Three phases in the energy dynamics while training: overfitting happens in the last, with a steep fall in $\Delta E_{\boldsymbol{\theta}}(\mathbf{x})$ for SAT. For TRADES, it stays almost constant.

Interpreting TRADES as Energy-based Model. Going beyond prior work [2, 23, 51, 66], we reinterpret TRADES objective [62] as an EBM. TRADES loss is as follows:

$$\min_{\boldsymbol{\theta}} \left[\mathcal{L}_{\text{CE}}\big(\boldsymbol{\theta}(\mathbf{x}), y\big) + \beta \max_{\boldsymbol{\delta} \in \Delta} \text{KL}\left(p(y|\mathbf{x}) \middle|\middle| p(y|\mathbf{x}^{\star}) \right) \right], \tag{4}$$

where $\mathrm{KL}(\cdot,\cdot)$ is the KL divergence between the conditional probability over classes $p(y|\mathbf{x})$ that acts as reference distribution and probability over classes for generated points $p(y|\mathbf{x}^\star)$, the loss $\mathcal{L}$ is CE loss and $p(y|\mathbf{x})$ is from Eq. (1).

Proposition 1. *The KL divergence between two discrete distributions $p(y|\mathbf{x})$ and $p(y|\mathbf{x}^\star)$ can be interpreted using EBM as* [4]*:*

$$\underbrace{\mathbb{E}_{k\sim p(y|\mathbf{x})}\Big[E_{\boldsymbol{\theta}}(\mathbf{x}^\star,k)-E_{\boldsymbol{\theta}}(\mathbf{x},k)\Big]}_{\text{conditional term weighted by classifier prob.}}+\underbrace{E_{\boldsymbol{\theta}}(\mathbf{x})-E_{\boldsymbol{\theta}}(\mathbf{x}^\star)}_{\text{marginal term}}. \tag{5}$$

Corollary 1. *TRADES object can be written as EBM as:*

$$E_{\boldsymbol{\theta}}(\mathbf{x},y)+(\beta-1)E_{\boldsymbol{\theta}}(\mathbf{x})-\beta\Big\{E_{\boldsymbol{\theta}}(\mathbf{x}^\star)+\mathbb{E}_{p(y|\mathbf{x})}\Big[E_{\boldsymbol{\theta}}(\mathbf{x},k)-E_{\boldsymbol{\theta}}(\mathbf{x}^\star,k)\Big]\Big\}. \tag{6}$$

By writing the KL divergence as Eq. (5), we can better see analogies and differences with SAT. Similarly to SAT, TRADES has to push down $E_{\boldsymbol{\theta}}(\mathbf{x}^\star)$ yet it does so considering a reference fixed energy value which is given by the corresponding natural data $E_{\boldsymbol{\theta}}(\mathbf{x})$. At the same time, they both have to push up $E_{\boldsymbol{\theta}}(\mathbf{x}^\star,k)$ yet TRADES attack only increases the loss when $E_{\boldsymbol{\theta}}(\mathbf{x}^\star,k) > E_{\boldsymbol{\theta}}(\mathbf{x},k)$ for k classes. Furthermore, a big difference resides in the training dynamics: while AT is agnostic to the dynamics, TRADES uses the classifier prediction as weighted average: at the beginning of the training $p(y|\mathbf{x})$ is uniform, being the conditional part averaged across all classes, so the attack is not really affecting any class in particular. Instead, at the end of the training when $p(y|\mathbf{x})$ may distribute more like a one-hot encoding, TRADES will consider the most likely class.

Better Robust Models Have Smoother Energy Landscapes. Smoothness is a well-established concept in robustness, where a smooth loss landscape suggests that for small perturbations $\boldsymbol{\delta}$, the difference in loss $|\mathcal{L}_{\boldsymbol{\theta}}(\mathbf{x})-\mathcal{L}_{\boldsymbol{\theta}}(\mathbf{x}+\boldsymbol{\delta})|$ remains small ($<\epsilon$) wrt the input $\mathbf{x}$. We show a link between Energy and Loss in Eq. (3). PGD-like attacks drastically bend the energy surface–see Fig. 2–thereby the model needs to reconcile the adv. energy with the natural. This reconciliation yields the smoothness. The intuition is that classifiers may tend towards the data distribution to some extent yet the attacks generate new points out of manifold. The model has now to align these two distributions and it is forced to smooth the two energies to keep classifying both correctly. Once $E_{\boldsymbol{\theta}}(\mathbf{x})$ smoothness does not hold, the model is incapable of performing the alignment. $E_{\boldsymbol{\theta}}(\mathbf{x})$ smoothness is also a desirable property of EBMs. Over the past few years, various strategies have emerged to enhance robustness, some techniques weight the training samples like MART [52], GAIR-RST [64] and some focus on smoothing the weight loss landscape, AWP [54]. Furthermore, recent state-of-the-art [12,53] leverage synthesized data to increase robustness even further. Upon analyzing

[4] Proofs of Proposition 1 and Corollary 1 are in the supplementary material.

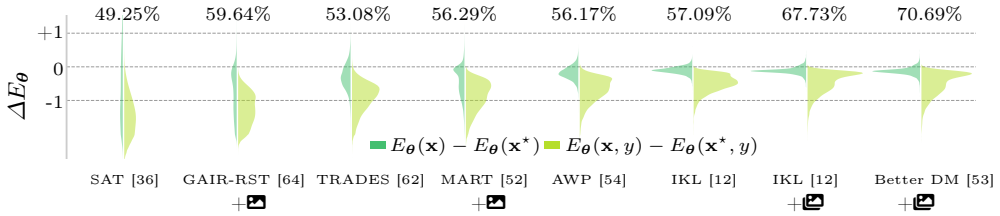


Fig. 4: Difference in the energy between natural data $\mathbf{x}$ and $\mathbf{x}^{\star}$ for state-of-the-art methods in adversarial robustness. For each method we show the signed difference between $\mathbf{x}$ and $\mathbf{x}^{\star}$ for both $E_{\boldsymbol{\theta}}(\mathbf{x})$ and $E_{\boldsymbol{\theta}}(\mathbf{x}, y)$, on top of each method we report the robust accuracy from [11]. The vertical axis is in *symmetric log scale*. The increase in robust accuracy correlates well with $\Delta E_{\boldsymbol{\theta}}(\mathbf{x})$ approaching zero and reducing the spread of the distribution. + indicates training with generated images by [53], while the + indicates training with additional data by [5] for the CIFAR-10 dataset.

the distributions of $\Delta E_{\boldsymbol{\theta}}(\mathbf{x})$ and $\Delta E_{\boldsymbol{\theta}}(\mathbf{x}, y)$ for all test samples, we observed that as the model's robustness increases, the energy distribution tended to approach zero, as depicted in Fig. 4. From the figure is also clear the smoothing effect of TRADES compared to SAT also visible in Fig. 3.

AT in function of High vs Low Energy Samples. Several studies have highlighted the unequal impact of samples in AT: [15,52,63] focus on the importance of samples in relation to their correct or incorrect classification, while [34, 64] suggest that samples near the decision boundaries are regarded as more critical. We can comprehensively explain such findings as well as others [35,60] using our framework. We begin by investigating MART, which employs Misclassification-Aware Regularization (MAR), focusing on the significance of samples categorized by their correct or incorrect classification. We do a proof-of-concept experiment closely resembling MART's [52] where we initially start from a robust model trained with SAT [36]. Unlike [52], we opted to make subsets based on their energy values. We selected two subsets from the natural training dataset: one comprising high-energy examples but excluding misclassifications; another with low-energy samples of correctly classified examples. All the subsets are created considering the initial values from the robust SAT classifier. We trained again the same networks from scratch without these subsets[5]. Subsequently, we assessed the robustness against PGD [36] on the test dataset. Our findings indicate that removing high-energy correct samples has a similar impact to removing incorrectly classified samples, as shown in Fig. 5a. Additionally, we observed that most incorrectly classified samples exhibit higher energies, suggesting that robustness reduction is likely due to their high energy values and not to their incorrect classification. On another axis, we reinterpret weighting schemes like MAIL [34]: it uses Probabilistic Margins (PMs) to weight samples, with optimal results attained when calculated on adversarial points. Interestingly, our analysis reveals a good correlation between the PM and $E_{\boldsymbol{\theta}}(\mathbf{x}, y)$ while there is less correlation with $E_{\boldsymbol{\theta}}(\mathbf{x})$ showing that a weighting scheme based on energy is not the same

[5] Additional details about statistics in the supplementary material.

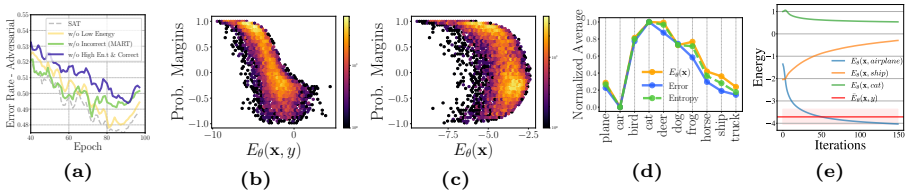


Fig. 5: (a) Not perturbing high-energy samples (correctly classified) increases robust error akin to not perturbing incorrectly classified samples shown in [52]. **(b)** Probabilistic Margins (PMs) in function of $E_{\boldsymbol{\theta}}(\mathbf{x}, y)$ **(c)** and of $E_{\boldsymbol{\theta}}(\mathbf{x})$ **(d)** Relationship between error rate, entropy and energy **(e)** Trend of $E_{\boldsymbol{\theta}}(\mathbf{x}, y)$ during the generative steps.

as PMs—Figs. 5b and 5c. Using our formulation, we can also explain recent research [35] revealing that robustness can transfer to other classes never attacked during AT. Findings from [35] indicate that classes that are harder to classify show better transfer of robustness to other classes. Moreover, they found that classes with high error rates happen to have high entropy. Our analysis shows that the same classes with high error rates[6] also display higher energy as shown in Fig. 5d. Thus, we can infer that classes with higher energy levels better facilitate robustness transferability. Finally, [60], by investigating robust overfitting, identifies that some small-loss data samples lead to overfitting. We can argue that this finding can also be explained in terms of energy, where samples with low loss correspond to high energy samples, as illustrated in Fig. 1. Building upon our findings we propose a simple weighting scheme dubbed *Weighted Energy Adversarial Training (WEAT)*. Our exploration concludes with the realization that low-energy samples tend to overfit, while high-energy samples contribute more significantly to robustness. Thus, we advocate for weighting the loss based on the energy metric $E_{\boldsymbol{\theta}}(\mathbf{x})$, wherein high-energy samples are assigned greater weight and low-energy samples are weighted less. Exploiting $E_{\boldsymbol{\theta}}(\mathbf{x})$ instead of $E_{\boldsymbol{\theta}}(\mathbf{x}, y)$ or PMs for weighting samples eliminates the need for a burn-in period required by [34,64], as it operates independently of class labels. To implement WEAT, we adopt TRADES [62] (WEAT_{NAT}), and a similar approach where we apply CE loss to adversarial data (WEAT_{ADV}). We utilize KL divergence as the inner loss to generate adversarial samples, and unlike [34,52] we weight the entire outer loss (both CE+KL) with a scalar function as $\log(1 + \exp(|E_{\boldsymbol{\theta}}(\mathbf{x})|))^{-1}$ that weights more the samples close to zero energy and decays very fast. More importantly, while weighting the loss, $E_{\boldsymbol{\theta}}(\mathbf{x})$ is detached from the computational graph so that the weighting branch does not backpropagate, to avoid trivial solutions.

Impact of the Energy in the Generative Capabilities. Though generative capabilities have been the subject of previous investigations [23,51,66], we find that the optimization for adversarial perturbations is crucial to develop the generative model. A key factor is on how different losses bend the energy landscape—i.e. CE vs KL divergence. Despite recent methods [18] report that robustness goes "hand in hand" with perceptually aligned gradients (PAG), we find out the gen-

[6] We report probabilistic error rate $1 - p(y|\mathbf{x})$, contrary to hard error rates in [35].

erative capabilities for all recent approaches [52, 62] are less "intense", requiring more iterations to produce meaningful images. We suspect this could be due to usage of KL divergence instead of CE, aiming at better robustness. Surprisingly, we find that even SOTA robust classifiers trained on millions of synthetic images from diffusion models [53] using TRADES have *less* intense generative performance than the "old" model by [43]. We propose a new simple inference technique that pushes their generative capabilities, lifting generation to high standard, despite no actual training towards generative modeling. We do so by means of a proper initialization of the Stochastic Gradient Langevin Dynamics (SGLD) MCMC, by starting the chain close to the class manifold instead of random noise like JEM [23, 66] or from multivariate Gaussian per class [43]. We sample from principal components per class weighted by their singular values to generate the main low-frequency content near the class manifold and let SGLD add the high-frequency part without leaving the manifold. To do so, we take very small steps yet we use the inertia of the chain to greatly speed up the descent:

$$\begin{cases} \boldsymbol{\nu}_{n+1} = \zeta \boldsymbol{\nu}_n - \frac{1}{2}\eta \nabla_{\mathbf{x}} E_{\boldsymbol{\theta}}(\mathbf{x}, y) \quad \text{with} \quad \boldsymbol{\nu}_0 = \mathbf{0} \\ \mathbf{x}_{n+1} = \mathbf{x}_n + \boldsymbol{\nu}_{n+1} + \boldsymbol{\varepsilon} \quad \text{with} \quad \mathbf{x}_0 = \boldsymbol{\mu}_y + \sum_i \lambda_i \boldsymbol{\alpha}_i \mathbf{U}_i^y \end{cases} \tag{7}$$

where the initialization stochasticity comes from $\boldsymbol{\alpha} \sim \mathcal{N}(0; \sigma)$, then $\boldsymbol{\mu}_y$ and $\mathbf{U}^y$ are the mean and the principal components per class y and λ_i is the singular value associated to each component. We add regular noise in the SGLD chain as $\boldsymbol{\varepsilon} \sim \mathcal{N}(0, \gamma \mathbf{I})$, η is the step size and ζ the friction coefficient. We use the same loss as in [66] which is class dependent and allows us to samples from $p(\mathbf{x}|y)$. During SGLD steps, shown in Fig. 5e, the energy $E_{\boldsymbol{\theta}}(\mathbf{x}, y)$ associated to the class we want to generate decreases, while joint energies for other classes increase. Note how the target energy converges to the average joint energy $\bar{E}_{\boldsymbol{\theta}}(\mathbf{x}, y)$, computed all over CIFAR-10 training samples belonging to the desired class.

4 Experimental Evaluation

In this section, we pursue two distinct avenues of investigation. Firstly, we conduct an in-depth comparison of model robustness, demonstrating the effectiveness of our method, WEAT. Concurrently, we evaluate the quality of images generated by the existing state-of-the-art robust classifiers. We quantitatively assess image quality and diversity using established metrics like IS [42], FID [24], KID [3] and LPIPS [65], evaluated on $50,000$ images. Using those, we aim to illustrate the importance of initialization in SGLD, the impact of different sampling approaches and importance of momentum.

Datasets and Network Architecture. We train WEAT on four standard benchmark datasets: CIFAR-10, CIFAR-100 [30], SVHN [61] and Tiny-ImageNet [13] using ResNet-18. When possible, we also trained the competitive methods under the same settings for fairness. Additionally, we use CIFAR-10 to assess the generative capabilities of various SOTA robust classifiers from RobustBench. The implementation details can be found in the supplementary material.

Defence method	CIFAR-10			CIFAR-100			SVHN		
	Natural	PGD	AA	Natural	PGD	AA	Natural	PGD	AA
SAT [36]	82.43±.66	49.03±.46	45.37±.41	54.78±1.03	23.89±.18	20.99±.28	**93.22±.20**	50.54±.35	44.87±.30
TRADES [62]	82.91±.14	52.65±.16	49.46±.20	56.31±.28	28.53±.22	24.29±.16	89.09±.49	55.52±.29	48.13±1.10
MAIL-TR. [34]	81.63±.25	53.09±.22	49.42±.16	56.30±.14	28.79±.19	24.24 ±.07	89.65±.34	54.94±.47	47.48±1.73
WEAT$_{\text{NAT}}$	**83.36±.15**	52.43±.12	49.02±.21	**59.07±.59**	29.71±.22	24.88±.25	88.65±.77	55.31±.51	48.61±.49
WEAT$_{\text{ADV}}$	81.00±.17	**53.35±.07**	**49.75±.04**	56.57±.15	**30.90±.18**	**25.63±.15**	87.66±.62	**56.40±.37**	**49.60±.29**

(a)

Defence method	Clean Acc.	PGD	AA
SAT [36] ∗	48.09	—	16.46
TRADES [62]	49.15	21.92	17.24
MART [52] ∗	45.51	—	17.79
DyART [55] ∗	49.71	—	18.02
MAIL-TRADES [34]	48.72	21.98	17.03
WEAT$_{\text{NAT}}$	**52.73**	23.42	17.35
WEAT$_{\text{ADV}}$	49.54	**24.39**	**18.45**

(b)

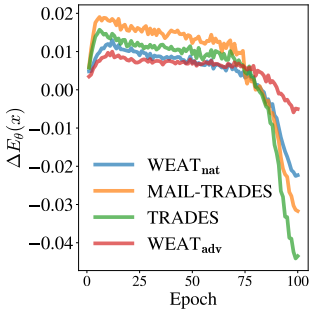
	Inner loss	Outer loss	β	Weight fun. w	Clean Acc.	PGD	AA
	CE	BCE($\mathbf{x}^\star$) + β·w·KL	5	MART [52]	54.09	28.24	23.63
	CE	CE($\mathbf{x}^\star$) + β·w·KL	5	MART [52]	54.03	27.32	23.71
	CE	CE($\mathbf{x}^\star$) + β·KL	6	—	53.55	28.93	23.97
	KL	CE($\mathbf{x}^\star$) + β·KL	6	—	55.45	29.38	24.59
†	KL	CE($\mathbf{x}$) + β·KL	6	—	56.31	28.53	24.29
	KL	w·CE($\mathbf{x}$) + β·w·KL	5	PM$_{adv}$ [34]	56.45	27.74	23.26
†	KL	w·CE($\mathbf{x}$) + β·w·KL	6	**WEAT$_{\text{NAT}}$**	**59.07**	29.71	24.88
†	KL	w·CE($\mathbf{x}^\star$) + β·w·KL	6	**WEAT$_{\text{ADV}}$**	57.31	**30.64**	**25.43**

(c)

Table 1: (a) Results on CIFAR-10, CIFAR-100, and SVHN. **(b)** Results on Tiny-ImageNet, rows marked with ∗ are mean values from [55]. **(c)** Ablation study on CIFAR-100 with loss and weighting scheme. w is the weighting method. Rows marked with † show mean values from 5 trials, similar to Tab. 1a.

4.1 Quantitative Results

Ablation Study. In Tab. 1c we assessed the impact of different inner and outer loss functions, starting with MART where we replaced boosted cross entropy (BCE) with CE. Using BCE improved accuracy with PGD, but not with AA. If we do not weight the samples, the KL divergence as inner loss outperformed CE, showing improvements in both clean accuracy and AA. Similar to our approach, we also explored weighting the entire loss with PM$_{adv}$ in MAIL-TRADES, but observed a degradation in performance. With same β, WEAT$_{\text{ADV}}$ showed superior robustness, while WEAT$_{\text{NAT}}$ excelled in clean accuracy, yet still has better robustness than existing approaches. We then study the individual components contributing to our generative framework's performance and their respective impacts in Tab. 2b. We analyze different initializations for the same model [53], compare results of classifiers trained under different threat models using ℓ_∞ and ℓ_2 norms and finally explore generative capabilities of a set of various robust classifiers. Our method provides a better initialization compared to others in ℓ_2 norm setting, reaching impressive results in the generation considering that samples are produced by a robust classifier, not trained optimizing its generation.
Comparison with the State-of-the-Art. WEAT's results are summarized in Tab. 1a for CIFAR-10/100 and SVHN, where for each method we report mean and standard deviation from five models trained with different seeds. In Tab. 1b for Tiny-ImageNet, due to computational limitations, we present results from a single run. We report the accuracies on natural examples and adversarial ex-



(a)

Method	IS ↑	FID ↓	KID ↓	LPIPS ↓
Initialization with [53], ℓ_∞				
Random [23, 60]	1.82	357.21	11.19	0.39
Gaussian [43, 57]	7.18	**64.98**	**2.02**	**0.18**
PCA - **Ours**	**7.66**	97.38	2.15	0.20
Initialization with [53], ℓ_2				
Gaussian [43, 57]	8.75	**27.71**	0.56	0.18
PCA - **Ours**	**8.97**	30.74	**0.51**	**0.18**
Classifier, Eq. (7), ℓ_∞				
SAT [43]	7.96	72.15	1.03	0.21
TRADES [62]	7.19	72.51	1.31	0.22
MART [52]	**8.11**	**66.98**	**1.03**	**0.20**
Better DM [53]	7.66	97.38	2.15	**0.20**
Classifier, Eq. (7), ℓ_2				
SAT [43]	8.58	45.19	**0.49**	0.19
Better DM [53]	**8.97**	**30.74**	0.51	**0.18**

(b)

Method	FID ↓	IS ↑
Hybrid models		
JEM [23]	38.4	8.76
DRL [19]	9.60	8.58
JEAT [66]	38.24	8.80
JEM ++ [57]	37.1	8.29
SADA-JEM [59]	**9.41**	8.77
M-EBM [58]	21.1	7.20
Robust classifiers		
PreJEAT [66]	56.85	7.91
SAT [43]	—	7.5
Ours, Eq. (7)		
SAT [43]	45.19	8.58
Better DM [53]	30.74	**8.97**

(c)

Table 2: (a) While training our models on CIFAR-100, WEAT has lower $\Delta E_{\boldsymbol{\theta}}(\mathbf{x})$ compared to other approaches suggesting lesser robust overfitting, see also Fig. 3 **(b)** Ablation study for different components of our framework using *only robust classifiers*. Adopting ℓ_2 leads to major improvements in metrics. **(c)** Model [53] overcomes SOTA generative abilities, topping IS and matching FID of even certain hybrid models.

amples obtained using PGD [36] with 20 steps (step size $\alpha = 2/255$), and Auto Attack (AA) [11] for robustness evaluation. WEAT outperforms existing similar methods across all datasets, with WEAT_{NAT} showing superior clean accuracy and comparable robust accuracy, while WEAT_{ADV} achieves the highest robust accuracy overall but with a slight reduction in clean accuracy. With Tiny-ImageNet, our results outperform [55] without any extra computational cost, unlike their approach which incurs costs up to twice that of TRADES [62]. Our approach exhibits lesser robust overfitting compared to other approaches as it weights low-energy samples less, resulting in a lower $\Delta E_{\boldsymbol{\theta}}(\mathbf{x})$ as shown in Tab. 2a. Regarding image generation, we conduct experiments in producing synthetic images, whose results are shown in Tabs. 2b and 2c. Our findings demonstrate that integrating momentum in the SGLD framework, along with the PCA initialization, improves image quality beyond conventional SGLD. Our method reaches the highest IS and is able to exceed FID performance of robust classifiers as well as the majority of the listed SOTA hybrid models, trained *explicitly* for generation.

4.2 Qualitative Results

Ablation Study. Fig. 6 (bottom row) shows that starting the chain from Random Noise [23, 60], leads to unrealistic images, with saturated colors and no object's shape, while beginning from a Gaussian per class, employed in [43,57], images are coherently generated to the label yet with low fidelity due to the highly saturated colors. With our method, images achieve higher quality and realism, being more aligned with the data manifold. The improvement is even more visible when we combine the momentum and small step size with our init, thereby using Eq. (7). Our method allows generating realistic images, close to the natural distribution, just using a robust classifier trained with AT.

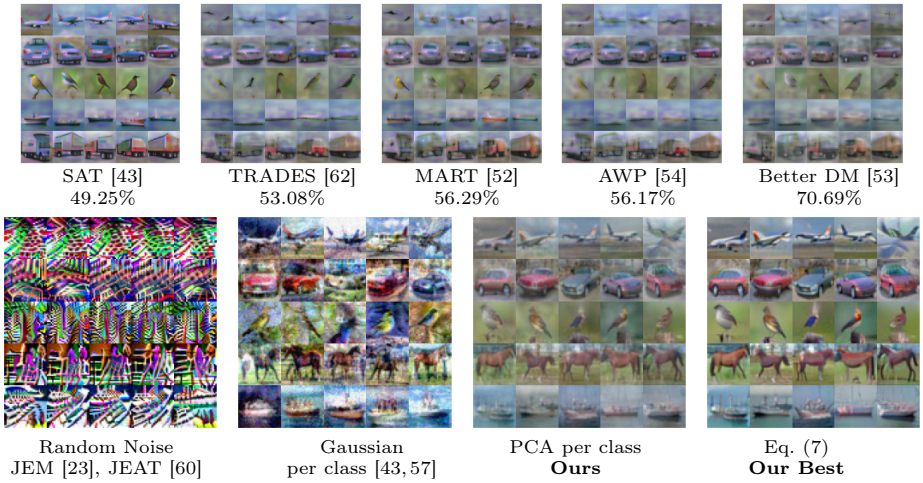


Fig. 6: (Top) Images generated from different robust classifiers with our proposed PCA init, while comparing their robust accuracies with generative capability. (Bottom) Different init in SGLD MCMC using the same model [53]. Random noise offers overly noisy init. Our PCA-based init shines in variability and smooth images, allowing us to match SOTA generative performance *just using a discriminative robust classifier*.

Comparison with the State-of-the-Art. As shown in Fig. 6 (top row), robust classifiers differ in their generation abilities. Surprisingly, using our initialization, the "old" model SAT [43] has more intense capabilities than recent models trained with TRADES, despite its lower robust accuracy. Compared to TRADES, SAT guides the SGLD chain to saturate more quickly, thereby converging faster to oversaturated images where the class signal is over-dominant. Fig. 6 (bottom row) compares different initialization methods, fixing the same classifier as [53], e.g. Random Noise [21, 60] and Gaussian per class [57]. Our PCA initialization, with a proper selection of parameters, robust classifiers can synthesize realistic and smooth images, with no need for generative retraining.

5 Conclusions and Future Work

This work aims at enhancing the understanding of robust classifiers via EBMs. We propose a sample weighting scheme, achieving SOTA results across popular benchmark datasets. Future work aims to modify the energy weighting function to account for the energy distribution of the data and applying the EBM framework to explain score-based Unrestricted Adversarial Examples (UAE) [29, 56].
Potential Negative Societal Impact. Although perceived as resistant to attacks, robust models are often viewed as benign but could have a potential negative effect if they are invariant to perturbation meant to protect privacy. Moreover, the possibility of "inverting" a robust classifier so easily makes it more prone to expose its training data, thereby possibly causing problem of privacy.

Acknowledgment. This work was supported by projects PNRR MUR PE0000013-FAIR under the MUR National Recovery and Resilience Plan funded by the European Union - NextGenerationEU and PRIN 2022 project 20227YET9B "AdVVent" CUP code B53D23012830006. It was also partially supported by Sapienza research projects "Prebunking", "Adagio", and "Risk and Resilience factors in disadvantaged young people: a multi-method study in ecological and virtual environments". Computing was supported by CINECA cluster under project Ge-Di HP10CRPUVC and the Sapienza Computer Science Department cluster.

References

1. Andriushchenko, M., Croce, F., Flammarion, N., Hein, M.: Square attack: a query-efficient black-box adversarial attack via random search. In: ECCV (2020)
2. Beadini, S., Masi, I.: Exploring the connection between robust and generative models. In: Italian Conference on AI - Ital-IA - Workshop on AI for Cybersecurity (2023)
3. Binkowski, M., Sutherland, D.J., Arbel, M., Gretton, A.: Demystifying MMD gans. In: ICLR (2018)
4. Carlini, N., Wagner, D.: Towards evaluating the robustness of neural networks. In: IEEE Symposium on Security and Privacy (SP) (2017)
5. Carmon, Y., Raghunathan, A., Schmidt, L., Liang, P., Duchi, J.: Unlabeled data improves adversarial robustness. In: NeurIPS (2019)
6. Chen, T., Zhang, Z., Liu, S., Chang, S., Wang, Z.: Robust overfitting may be mitigated by properly learned smoothening. In: ICLR (2020)
7. Chen, T., Zhang, Z., Wang, P., Balachandra, S., Ma, H., Wang, Z., Wang, Z.: Sparsity winning twice: Better robust generaliztion from more efficient training. arXiv preprint arXiv:2202.09844 (2022)
8. Cohen, J., Rosenfeld, E., Kolter, Z.: Certified adversarial robustness via randomized smoothing. In: ICML. pp. 1310–1320. PMLR (2019)
9. Croce, F., Andriushchenko, M., Sehwag, V., Debenedetti, E., Flammarion, N., Chiang, M., Mittal, P., Hein, M.: Robustbench: a standardized adversarial robustness benchmark. In: ICLR Workshops 2021 - Workshop on Security and Safety in Machine Learning Systems (2021)
10. Croce, F., Hein, M.: Minimally distorted adversarial examples with a fast adaptive boundary attack. In: ICML (2020)
11. Croce, F., Hein, M.: Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks. In: ICML (2020)
12. Cui, J., Tian, Z., Zhong, Z., Qi, X., Yu, B., Zhang, H.: Decoupled kullback-leibler divergence loss. arXiv preprint arXiv:2305.13948 (2023)
13. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: CVPR (2009)
14. Dhariwal, P., Nichol, A.: Diffusion models beat gans on image synthesis. In: NeurIPS (2021)
15. Ding, G.W., Sharma, Y., Lui, K.Y.C., Huang, R.: Mma training: Direct input space margin maximization through adversarial training. In: ICLR (2020)
16. Dong, Y., Xu, K., Yang, X., Pang, T., Deng, Z., Su, H., Zhu, J.: Exploring memorization in adversarial training. In: ICLR (2021)
17. Foret, P., Kleiner, A., Mobahi, H., Neyshabur, B.: Sharpness-aware minimization for efficiently improving generalization. In: ICLR (2021)

18. Ganz, R., Kawar, B., Elad, M.: Do perceptually aligned gradients imply robustness? In: ICML (2023)
19. Gao, R., Song, Y., Poole, B., Wu, Y.N., Kingma, D.P.: Learning energy-based models by diffusion recovery likelihood. In: ICLR (2021)
20. Goodfellow, I., Shlens, J., Szegedy, C.: Explaining and harnessing adversarial examples. In: ICLR (2015)
21. Gowal, S., Qin, C., Uesato, J., Mann, T., Kohli, P.: Uncovering the limits of adversarial training against norm-bounded adversarial examples. arXiv preprint arXiv:2010.03593 (2020)
22. Gowal, S., Rebuffi, S.A., Wiles, O., Stimberg, F., Calian, D.A., Mann, T.A.: Improving robustness using generated data. In: NeurIPS (2021)
23. Grathwohl, W., Wang, K.C., Jacobsen, J.H., Duvenaud, D., Norouzi, M., Swersky, K.: Your classifier is secretly an energy based model and you should treat it like one. In: ICLR (2020)
24. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: Gans trained by a two time-scale update rule converge to a local nash equilibrium. In: NeurIPS. vol. 30 (2017)
25. Hitaj, D., Pagnotta, G., Masi, I., Mancini, L.V.: Evaluating the robustness of geometry-aware instance-reweighted adversarial training. arXiv preprint arXiv:2103.01914 (2021)
26. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. In: NeurIPS. vol. 33 (2020)
27. Huang, T., Liu, S., Chen, T., Fang, M., Shen, L., Menkovski, V., Yin, L., Pei, Y., Pechenizkiy, M.: Enhancing adversarial training via reweighting optimization trajectory. In: Joint European Conference on Machine Learning and Knowledge Discovery in Databases (2023)
28. Kim, M., Tack, J., Shin, J., Hwang, S.J.: Entropy weighted adversarial training. In: ICML Workshop on Adversarial Machine Learning (2021)
29. Kollovieh, M., Gosch, L., Scholten, Y., Lienen, M., Günnemann, S.: Assessing robustness via score-based adversarial image generation. In: ICLR (2024)
30. Krizhevsky, A., Hinton, G., et al.: Learning multiple layers of features from tiny images. Tech. rep., Citeseer (2009)
31. LeCun, Y., Chopra, S., Hadsell, R., Ranzato, M., Huang, F.: A tutorial on energy-based learning. Predicting structured data **1**(0) (2006)
32. Lecuyer, M., Atlidakis, V., Geambasu, R., Hsu, D., Jana, S.: Certified robustness to adversarial examples with differential privacy. In: IEEE Symposium on Security and Privacy (SP) (2019)
33. Li, B., Chen, C., Wang, W., Carin, L.: Second-order adversarial attack and certifiable robustness. arXiv preprint arXiv:1809.03113 (2018)
34. Liu, F., Han, B., Liu, T., Gong, C., Niu, G., Zhou, M., Sugiyama, M., et al.: Probabilistic margins for instance reweighting in adversarial training. In: NeurIPS (2021)
35. Losch, M., Omran, M., Stutz, D., Fritz, M., Schiele, B.: On adversarial training without perturbing all examples. In: ICLR (2024)
36. Madry, A., Makelov, A., Schmidt, L., Tsipras, D., Vladu, A.: Towards deep learning models resistant to adversarial attacks. In: ICLR (2018)
37. Pang, T., Lin, M., Yang, X., Zhu, J., Yan, S.: Robustness and accuracy could be reconcilable by (proper) definition. In: ICML. pp. 17258–17277. PMLR (2022)
38. Peng, S., Xu, W., Cornelius, C., Hull, M., Li, K., Duggal, R., Phute, M., Martin, J., Chau, D.H.: Robust principles: Architectural design principles for adversarially robust CNNs. In: BMVC (2023)

39. Rice, L., Wong, E., Kolter, Z.: Overfitting in adversarially robust deep learning. In: ICML (2020)
40. Rojas-Gomez, R.A., Yeh, R.A., Do, M.N., Nguyen, A.: Inverting adversarially robust networks for image synthesis. arXiv preprint arXiv:2106.06927 (2021)
41. Rouhsedaghat, M., Monajatipoor, M., Kuo, C.C.J., Masi, I.: MAGIC: Mask-guided image synthesis by inverting a quasi-robust classifier. In: AAAI Conference on Artificial Intelligence (2023)
42. Salimans, T., Goodfellow, I., Zaremba, W., Cheung, V., Radford, A., Chen, X., Chen, X.: Improved techniques for training gans. In: NeurIPS (2016)
43. Santurkar, S., Tsipras, D., Tran, B., Ilyas, A., Engstrom, L., Madry, A.: Image synthesis with a single (robust) classifier. In: NeurIPS (2019)
44. Schmidt, L., Santurkar, S., Tsipras, D., Talwar, K., Madry, A.: Adversarially robust generalization requires more data. In: NeurIPS (2018)
45. Shafahi, A., Najibi, M., Ghiasi, M.A., Xu, Z., Dickerson, J., Studer, C., Davis, L.S., Taylor, G., Goldstein, T.: Adversarial training for free! In: NeurIPS (2019)
46. Singla, V., Singla, S., Feizi, S., Jacobs, D.: Low curvature activations reduce overfitting in adversarial training. In: ICCV (2021)
47. Song, Y., Ermon, S.: Generative modeling by estimating gradients of the data distribution. In: NeurIPS (2019)
48. Stutz, D., Hein, M., Schiele, B.: Disentangling adversarial robustness and generalization. In: CVPR. IEEE Computer Society (2019)
49. Stutz, D., Hein, M., Schiele, B.: Relating adversarially robust generalization to flat minima. In: ICCV (2021)
50. Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., Fergus, R.: Intriguing properties of neural networks. In: ICLR (2014)
51. Wang, Y., Wang, Y., Yang, J., Lin, Z.: A unified contrastive energy-based model for understanding the generative ability of adversarial training. In: ICLR (2022)
52. Wang, Y., Zou, D., Yi, J., Bailey, J., Ma, X., Gu, Q.: Improving adversarial robustness requires revisiting misclassified examples. In: ICLR (2020)
53. Wang, Z., Pang, T., Du, C., Lin, M., Liu, W., Yan, S.: Better diffusion models further improve adversarial training. In: ICML (2023)
54. Wu, D., Xia, S.T., Wang, Y.: Adversarial weight perturbation helps robust generalization. In: NeurIPS (2020)
55. Xu, Y., Sun, Y., Goldblum, M., Goldstein, T., Huang, F.: Exploring and exploiting decision boundary dynamics for adversarial robustness. In: ICLR (2022)
56. Xue, H., Araujo, A., Hu, B., Chen, Y.: Diffusion-based adversarial sample generation for improved stealthiness and controllability. In: NeurIPS (2023)
57. Yang, X., Ji, S.: Jem++: Improved techniques for training jem. In: ICCV. pp. 6494–6503 (2021)
58. Yang, X., Ji, S.: M-ebm: Towards understanding the manifolds of energy-based models. In: Pacific-Asia Conference on Knowledge Discovery and Data Mining. pp. 291–302. Springer (2023)
59. Yang, X., Su, Q., Ji, S.: Towards bridging the performance gaps of joint energy-based models. In: CVPR. pp. 15732–15741 (2023)
60. Yu, C., Han, B., Shen, L., Yu, J., Gong, C., Gong, M., Liu, T.: Understanding robust overfitting of adversarial training and beyond. In: ICML (2022)
61. Yuval, N.: Reading digits in natural images with unsupervised feature learning. In: NIPS Workshop on Deep Learning and Unsupervised Feature Learning (2011)
62. Zhang, H., Yu, Y., Jiao, J., Xing, E.P., Ghaoui, L.E., Jordan, M.I.: Theoretically principled trade-off between robustness and accuracy. In: ICML (2019)

63. Zhang, J., Xu, X., Han, B., Niu, G., Cui, L., Sugiyama, M., Kankanhalli, M.: Attacks which do not kill training make adversarial learning stronger. In: ICML. pp. 11278–11287. PMLR (2020)
64. Zhang, J., Zhu, J., Niu, G., Han, B., Sugiyama, M., Kankanhalli, M.: Geometry-aware instance-reweighted adversarial training. In: ICLR (2020)
65. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 586–595 (2018)
66. Zhu, Y., Ma, J., Sun, J., Chen, Z., Jiang, R., Chen, Y., Li, Z.: Towards understanding the generative capability of adversarially robust classifiers. In: ICCV (2021)