

A Provable Energy-Guided Test-Time Defense Boosting Adversarial Robustness of Large Vision-Language Models

Mujtaba Hussain Mirza¹ Antonio D’Orazio¹ Odelia Melamed² Iacopo Masi¹

¹OmnAI Lab, Computer Science Department, Sapienza University of Rome, Italy

²Weizmann Institute of Science, Israel

{mirza, dorazio, masi}@di.uniroma1.it, odelia.melamed@weizmann.ac.il

Abstract

Despite the rapid progress in multimodal models and Large Visual-Language Models (LVLM), they remain highly susceptible to adversarial perturbations, raising serious concerns about their reliability in real-world use. While adversarial training has become the leading paradigm for building models that are robust to adversarial attacks, Test-Time Transformations (TTT) have emerged as a promising strategy to boost robustness at inference. In light of this, we propose **Energy-Guided Test-Time Transformation (ET3)**, a lightweight, training-free defense that enhances the robustness by minimizing the energy of the input samples. Our method is grounded in a theory that proves our transformation succeeds in classification under reasonable assumptions. We present extensive experiments demonstrating that ET3 provides a strong defense for classifiers, zero-shot classification with CLIP, and also for boosting the robustness of LVLMs in tasks such as Image Captioning and Visual Question Answering. Code is available on [GitHub](#).

1. Introduction

Large vision-language models (LVLMs) have achieved remarkable progress, demonstrating strong multimodal reasoning and zero-shot generalization across tasks such as image captioning, visual question answering, and open-world recognition. Recent LVLMs such as LLaVA [46], OpenFlamingo [5], and Qwen-VL [6] present excellent performance; however, despite their broad generalization, they are still vulnerable to attacks to the visual modality [11, 64], potentially compromising downstream LVLM tasks.

Crucially, many LVLMs rely on the visual encoder from CLIP [65], which provides exceptional flexibility and generalization across diverse visual inputs, but also forms the primary source of vulnerability to image-based adversarial attacks: unnoticeable, carefully crafted perturbations to input images can induce incorrect predictions [28, 84].

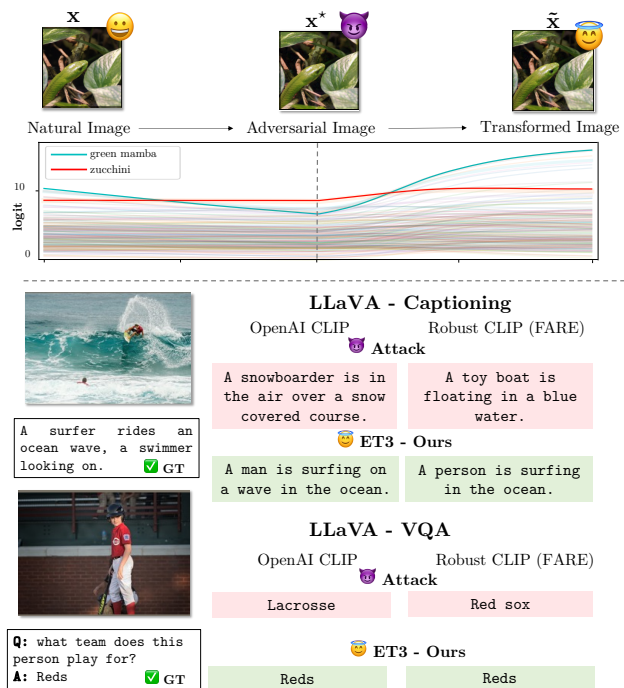


Figure 1. (top) Presenting a natural image green mamba x , and its adversarial image x^* mistakenly classified as a zucchini by a robust classifier f_θ . Given only x^* and f_θ , our ET3 test-time defense produces the correctly classified $\tilde{x}$, thereby boosting adversarial robustness. The plot illustrates the logit’s change: even though the ground-truth class is *not* the second best, ET3 still recovers it. (bottom) ET3 boosts robust accuracy of Large VLMs like LLaVa [46] using standard or even Robust CLIP [72] on both image captioning and Visual QA. *Note in VQA example, the Reds team is a different team from the Red Sox, and refers to the Cincinnati Reds team.*

During the past decade, a plethora of defenses have been proposed to improve adversarial robustness, with adversarial training (AT) [49] emerging as a leading efficient yet time-consuming paradigm. Despite this progress, AT-trained models still exhibit a substantial gap between their

performance on clean and perturbed inputs, particularly under strong or previously unseen attacks. In contrast, another line of work focuses on test-time defenses that achieve robustness at inference. Within this class, adversarial purification methods [79] attempt to denoise and remove adversarial perturbations using auxiliary generative models, while Randomized Smoothing (RS) [16] certifies robustness by averaging predictions over noise-perturbed inputs.

More recently, test-time transformations (TTT) [75, 96] have been proposed to improve classification without explicitly removing perturbations, instead adaptively transforming inputs dynamically at test time. Compared to adversarial training, these approaches defend against previously unseen threats in a plug-and-play manner without requiring model retraining. However, they often incur significant inference overhead, may rely on additional models, and sometimes struggle to scale to stronger attacks. They can also be computationally expensive, as they typically require optimizing additional tokens or parameters or generating augmented views for each sample. In this work, we present our novel and efficient test-time defense **ET3**, which demonstrates superior robustness achieved without such costly adaptation. The ET3 defense transforms the image at inference to boost robustness of CLIP while remaining lightweight and computationally efficient. Furthermore, our approach can be easily incorporated with existing methods to further improve their robustness significantly, incurring only a slight higher inference cost.

ET3 is inspired by energy-based models (EBMs) that were previously linked to robustness [30] and aims to lower the energy of a given input. Requiring extremely short optimization through the vision encoder, the defense is fast and lightweight, having a very minor overhead over the Visual LVM inference time. Fig. 1 presents a few qualitative examples of the ET3 transformation improving the robustness of classification, image captioning, and VQA. The ET3 boosts robustness for a broad range of models and tasks. In our experiments, we demonstrate increased robustness for zero-shot classification using CLIP-based vision language models and for robust image classifiers on the ImageNet dataset. We also show a significant boost in robustness in downstream tasks for LVLMs, even with a single optimization step through the CLIP encoder alone, improving robustness without noticeably impacting inference speed. Finally, we show our approach is robust against adaptive attacks.

To further support the ET3 defense, we theoretically prove its effectiveness on binary classifiers. For a binary classifiers satisfying few assumptions, we theoretically prove that an input transformed by ET3 will be correctly classified. We later provide experimental evidence of a robust model, along with examples of classifiers that have been theoretically proven to be robust in several different settings [26, 53, 54, 87], satisfying these assumptions.

2. Related Work

Adversarial Attacks. Adversarial examples are slightly perturbed inputs to induce erroneous predictions or behaviors in deep neural networks. As various methods exist to create such perturbations [3, 28, 49, 58], within the domain of LVLMs, a growing body of work has started examining their susceptibility to adversarial attacks [11, 64, 71].

Adversarial Defense. Among many defense strategies, adversarial training (AT)—which incorporates adversarial examples during training—remains the most effective empirical approach. AT has also been extended to vision language models (VLM) [51, 63, 67, 72, 92] and diffusion models [68]. In VLM, AT addresses vulnerabilities arising from multi-modal inputs by improving robustness to perturbations in either the visual or textual components. Beyond training-based defenses, offline prompt tuning methods learn a set of input tokens using a training dataset, optimizing them offline to improve accuracy or robustness without updating the backbone [43, 51, 105]. Test-time defenses constitute another direction in which the model itself is not trained to be robust. Instead, robustness is achieved at inference time through additional models or mechanisms. Within this broader class, Randomized Smoothing (RS) comprises a family of methods that certify robustness by averaging predictions over noise-perturbed inputs [16, 42], with various extensions proposed in recent works [1, 48]. Another approach is adversarial purification, which uses an auxiliary generative model to denoise the test image, removing adversarial perturbations while remaining close to the original image [34, 59, 79, 102]. In another line of work, the input is adaptively transformed at the test time [2, 8, 62, 94, 96], aiming for correct classification without explicitly removing adversarial perturbations or constraining a similarity to the original image, as in [8, 96].

Test-Time Transformations for VL tasks. Test-time adaptation methods [23, 27, 82, 83, 89] aims to adapt models to the test data at inference time to further improve their performance. For Vision-Language tasks, these methods usually aim to improve the robustness or generalization by altering the prompt, adjusting the embedding, or leveraging test-time augmentations. [76] introduces Test-Time Prompt Tuning (TPT), which optimizes a text prompt by minimizing prediction entropy across augmented views. While training-free and effective under distribution shift, its multiple augmentations cause significant inference overhead and reduced clean accuracy. C-TPT [101] mitigates this issue by introducing a calibration term that maximizes text-feature dispersion. R-TPT [75] targets adversarial robustness, optimizing the prompt against adversarial perturbations. [104] proposes MTA, which aggregates the embeddings of multiple augmentations using a robust mean-shift procedure. However, all of these methods introduce substantial inference overhead. [96] introduces TTC, which improves zero-

shot adversarial robustness by applying a gradient-based perturbation that maximizes the distance between the adversarial input and its clean embedding in CLIP’s feature space, yet offers limited robustness gains.

Adversarially Robust Models and EBM. Energy-based models (EBMs) [41] are generative models that learn an energy function $E(\mathbf{x})$ which assigns low energy values to inputs $\mathbf{x}$ in the data distribution and high energy values to other inputs. Intuitively, lower energy corresponds to samples that align well with the learned data distribution (on-manifold), while higher energy indicates atypical or out-of-distribution inputs (off-manifold). This conceptual perspective is particularly relevant for adversarial examples, which are known to deviate from the natural data manifold and typically go off the manifold of natural data [56, 74, 81].

Bridging this generative perspective with discriminative tasks, several studies [22, 86, 95] explore the link between generative models such as EBMs and discriminative models. This connection is formalized in the Joint Energy-based Model (JEM) [30], which reformulates the softmax classifier within an energy-based framework. This formulation allows the classifier’s output as energy can enable OOD detection [47]. The connection between learning an energy objective and robustness has been demonstrated in both ways, showing that the addition of EBM objectives to training increases robustness [30], and performing adversarial training has been shown to implicitly learn an EBM [99, 106]. Incorporating explicit energy-based objectives has also been shown to further enhance both robustness and generative performance [35, 55, 85, 100]. The generative aspects of robust classifiers have also been explored in other studies [69, 93, 97, 98].

Theoretically Proven Robustness. There has been a great effort in the theoretical research to characterize robustness guarantees and measure the robustness of a trained classifier. Starting in [73], many have tried to theoretically show adversarial vulnerability. It had been shown for random networks with different natural architectures in [7, 9, 10, 20, 57]. Later, in [26, 87] it had been shown that gradient-based training converge to a non-robust networks while robust network do exist, while even offering a concrete robust classifier as an example. In [54] and [53] the author discuss a different activation function, polynomial ReLU, and prove that training such model converges to a robust classifier.

3. Method

We introduce an *energy-based test-time transformation*, ET3, grounded in the energy-based modeling (EBM) perspective of discriminative classifiers. Unlike other test-time defense approaches that train an explicit generative EBM [34] or auxiliary diffusion/score models [79], ET3 requires no additional model training. It operates directly on

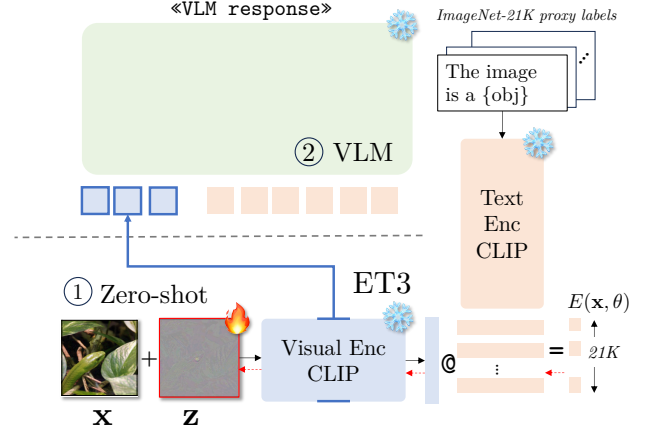


Figure 2. ① ET3 transforms the natural image $\mathbf{x}$ adding a small perturbation $\mathbf{z}$ optimized to lower the energy wrt to ImageNet-21k proxy classes and concepts. This allows robust zero-shot classification; ② the transformed image transfers and protects Large VLM, thereby increasing their robustness. The VLM is *not* used in the optimization, and the optimized image simply transfers to VLM by using the internal representation of the visual encoder.

a pre-trained classifier or visual encoder f_θ that we aim to defend. This is possible because a standard softmax classifier can itself be viewed through the EBM lens, by interpreting its logits as energies [30]. At its core, ET3 applies a lightweight transformation that simply decreases this logit-derived energy for a given input. Our focus on robust models stems from the tight connection between energy-based and robustness objectives. For instance, adding an explicit EBM term during training can enhance adversarial robustness [30]. Conversely, adversarial training has been shown to implicitly induce an EBM with smoother local energy landscapes [55, 99, 106]. For clean inputs, ET3 either preserves the original prediction or can even increase the model’s confidence. For adversarial inputs, it guides the sample back toward the correct classification, effectively enhancing the model’s robustness.

3.1. Energy Definition

Consider a set of d -dimensional labeled images $X = \{(\mathbf{x}, y) \sim \mathcal{D} \mid \mathbf{x} \in \mathbb{R}^d \text{ and } y \in \{1, \dots, K\}\}$, let $f_\theta(\mathbf{x}) : \mathbb{R}^d \rightarrow \mathbb{R}^K$ be a K -class classifier or an encoder parameterized by θ trained on a dataset X . For a sample $\mathbf{x} \in \mathbb{R}^d$, we denote by $f_\theta(\mathbf{x})_k$ the k -th coordinate of the output logits vector $f_\theta(\mathbf{x}) \in \mathbb{R}^K$. We define the energy at $\mathbf{x}$ as:

$$E(\mathbf{x}) = -\log\left(\sum_{k=1}^K \exp(f_\theta(\mathbf{x})_k)\right), \quad (1)$$

which is the negative LogSumExp of the output logits.

3.2. The ET3 Defense

Inspired by the connection between the energy and the perception of the data distribution by f_θ , we present a test-time defense based on energy minimization. Given a test image $\mathbf{x}$ and a defense radius ϵ , our goal is to obtain a modified image $\tilde{\mathbf{x}} = \mathbf{x} + \mathbf{z}$ by iteratively minimizing $E(\cdot, \theta)$. Concretely, we solve the constraint minimization $\tilde{\mathbf{x}} = \operatorname{argmin}_{\mathbf{x}^* \in \mathcal{B}_\epsilon(\mathbf{x})} E(\mathbf{x}^*)$, where $\mathcal{B}_\epsilon(\mathbf{x}) = \{\mathbf{x}^* : \|\mathbf{x}^* - \mathbf{x}\| \leq \epsilon\}$, using a gradient-based multi-step optimization for T steps. Formally, in each step $t \in [1, \dots, T]$ we calculate:

$$\mathbf{x}^{(t)} = \Pi_{\mathcal{B}_\epsilon(\mathbf{x})}(\mathbf{x}^{(t-1)} - \alpha \nabla_{\mathbf{x}} E(\mathbf{x}^{(t-1)})), \quad (2)$$

starting from $\mathbf{x}^{(0)} = \mathbf{x}$. The step size α and number of iterations T are hyperparameters, yet we found our method is fast and works for fewer steps. The projection $\Pi_{\mathcal{B}_\epsilon(\mathbf{x})}(\cdot)$ enforces the defense perturbation in a ℓ_2 ball.

3.3. Extending ET3 to Vision-Language Tasks

We extend ET3 to zero-shot classification and vision-language tasks. The core principle remains the same: minimize the energy function defined in Eq. (1) by refining the input image at test time. Fig. 2 depicts how we apply ET3 to both zero-shot classification and VLMs.

Zero-Shot Classification with CLIP. For zero-shot classification, we apply ET3 to CLIP models, where classification is performed by computing similarity scores between image and text embeddings, which serve as logits for ET3. To compute energy, we consider two distinct label configurations: a refined subset of labels or a vast set of labels, such as the full ImageNet-21K label set [66], yielding slightly improved robustness. We primarily adopt the latter, with further details on label-set selection provided in App. D.4.

Large Vision-Language Models (LVLMs). For multimodal tasks such as image captioning and visual question answering, we apply ET3 to the LLaVA model. Similar to the zero-shot setting, ET3 optimizes the visual input at test time by minimizing its energy through the CLIP vision encoder. After refinement, the visual embeddings from the vision encoder are passed to LLaVA’s projection layer, and the subsequent language generation process remains unchanged. ET3 enhances robustness specifically in the vision modality, even without modifying or fine-tuning the vision encoder to improve its robustness. To further support ET3, next we prove that the ET3 defense method boosts robustness for a binary classifier, where it provably transforms a clean or adversarial sample into a correctly classified input.

4. ET3 defense provably boosts Robust Nets

In this section, we theoretically prove that for a binary classifier $f_\theta(\mathbf{x}) : \mathbb{R}^d \rightarrow \mathbb{R}^2$, under some assumptions, minimizing the energy with an ϵ budget will transform a given

sample to be classified as its original ground truth class for both natural and adversarial samples. First, for simplicity, we assume local linearity in an ϵ -ball around the sample. We note that approximate local linearity has been proved and demonstrated before in the context of adversarial attacks and robust classifier [52]. Second, we assume that the gradient of the energy derived from the ground truth label is larger than the one derived from the opposite label. This assumption is particularly intriguing and arises from the vicinity of the adversarial sample and its corresponding source sample. Details about these assumptions in Remark 4.1.

Data and Model. We consider d dimensional data inputs $\mathbf{x} \in \mathbb{R}^d$, with unknown binary ground truth labels $y_t \in \{-1, 1\}$. The incorrect label is $\hat{y}_t = -y_t$. Our model is a binary classifier with two output logits denoted by $f_\theta(\mathbf{x}) : \mathbb{R}^d \rightarrow \mathbb{R}^2$. We use a two-logit output to adapt to the multiclass classifier regime, for which each possible class relates to a logit. Similarly, for a given input $\mathbf{x} \in \mathbb{R}^d$ and its ground truth label $y_t \in \{-1, 1\}$, if $f_\theta(\mathbf{x})_{y_t} > f_\theta(\mathbf{x})_{\hat{y}_t}$, then $\mathbf{x}$ is correctly classified.

Theorem 4.1. *Let $\mathbf{x} \in \mathbb{R}^d$ be a data sample, and y_t be its ground truth label. Let $f_\theta : \mathbb{R}^d \rightarrow \mathbb{R}^2$ be a binary classifier such that it’s locally linear in $\mathcal{B}_\epsilon(\mathbf{x})$. Denote $r_x = f_\theta(\mathbf{x})_{y_t} - f_\theta(\mathbf{x})_{\hat{y}_t}$, $\mathbf{g}_i = \nabla_{\mathbf{x}} f_\theta(\mathbf{x})_i$ and $e_i = \operatorname{SoftMax}(f(\mathbf{x}))_i$. Let $\epsilon > 0$ a defense budget such that $\epsilon > \frac{-2r_x}{\|\mathbf{g}_{y_t}\|}$. Then, if*

$$C \|e_{\hat{y}_t} \mathbf{g}_{\hat{y}_t}\| < \|e_{y_t} \mathbf{g}_{y_t}\|$$

for

$$C > \max \left\{ \frac{\exp(|r_x|)\epsilon \|\mathbf{g}_{y_t}\|}{\epsilon \|\mathbf{g}_{y_t}\| + 2r_x}, 1 \right\},$$

the ET3 defense transformation $\mathbf{z}$, parametrized by $T = 1$ and $\alpha = \frac{\epsilon}{e_{y_t}(1 + \frac{1}{C}) \|\mathbf{g}_{y_t}\|}$, will satisfy $\|\mathbf{z}\| \leq \epsilon$ and

$$f_\theta(\mathbf{x} + \mathbf{z})_{y_t} > f_\theta(\mathbf{x} + \mathbf{z})_{\hat{y}_t}.$$

Proof Idea. Formally, given a data point $\mathbf{x} \in \mathbb{R}^d$, we denote $\mathbf{g}_i$, the gradient of the network’s i -th logit w.r.t. the input and e_i the gradient of the energy w.r.t. the i -th logit. Then, the gradient of $\operatorname{LogSumExp}$ w.r.t. $\mathbf{x}$ is

$$\begin{aligned} \nabla_{\mathbf{x}} E(\mathbf{x}) &= -\operatorname{SoftMax}(f_\theta(\mathbf{x}))^\top \nabla_{\mathbf{x}} f_\theta(\mathbf{x}) \\ &= -e_{-1} \mathbf{g}_{-1} - e_1 \mathbf{g}_1. \end{aligned}$$

Thus, the purification step is of the form $\mathbf{z} = \alpha(e_{-1} \mathbf{g}_{-1} + e_1 \mathbf{g}_1)$. To see the effect of the transformation on each logit, we use the local linearity and look at the classifier as two locally linear functions $f_{-1}, f_1 : \mathbb{R}^d \rightarrow \mathbb{R}$, one for each logit. We observe that for $i \in \{-1, 1\}$

$$f_i(\mathbf{x} + \mathbf{z}) = f_i(\mathbf{x}) + \mathbf{z}^\top \mathbf{g}_i.$$

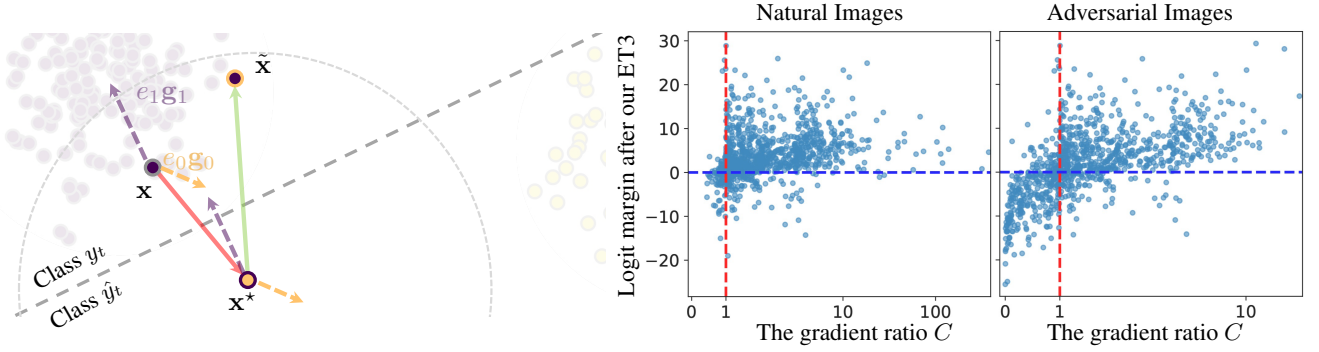


Figure 3. (left) The ET3 defense transformation for adversarial examples. Assuming local linearity of the model in the defense neighborhood $\mathcal{B}_\epsilon(\mathbf{x})$, and a large enough ratio C between the norms of the gradients of the energy through each class logit, $e_0 \mathbf{g}_0$ and $e_1 \mathbf{g}_1$. The adversarial attack, determined to reduce the ground truth logit, follows the negative direction of the larger gradient ($-\mathbf{g}_1$), while our transformation follows its positive direction ($\mathbf{g}_1$), increasing the ground truth logit and pulling the adversarial point back to its ground truth region. Both might also increase the other logit, corresponding to the smaller gradient $\mathbf{g}_0$, that may introduce some smaller deviation. (right) Scatter plot of the ratio between the gradients norms C and the logit margin at the transform image $\tilde{\mathbf{x}}$ on ImageNet robust classifier for 1000 randomly sampled images from ImageNet. For most samples for which $C > 1$, we can see that the purified image is correctly classified (logit margin > 0). One can see a correlation between the norms ratio and the logits difference of the transformed image.

We complete the proof by using the ratio of the gradients' norms, C , to show that multiplying by the larger norm gradient, $\mathbf{z}^\top \mathbf{g}_{y_t}$, will create a larger effect, implying correct classification. Formally, we note that the energy function allows a relaxed assumption on the ratio of the gradients as:

$$\exp(f_\theta(\mathbf{x})_{\hat{y}_t} - f_\theta(\mathbf{x})_{y_t}) C \|\mathbf{g}_{\hat{y}_t}\| < \|\mathbf{g}_{y_t}\|,$$

that is adaptive to the logits of $f_\theta(\mathbf{x})$. When $\mathbf{x}$ is classified correctly, we will have $f_\theta(\mathbf{x})_{\hat{y}_t} - f_\theta(\mathbf{x})_{y_t} \leq 0$ and thus $\exp(f_\theta(\mathbf{x})_{\hat{y}_t} - f_\theta(\mathbf{x})_{y_t}) \leq 1$ allowing a smaller ratio between the gradients to preserve correct classification. Otherwise, the assumed ratio will be weighted by the difference between the logits. The full proof provided in App. E.

Remark 4.1 (The Local Linearity and Gradient Norm Ratio Assumptions). *In Practice.* In the right part of Fig. 3 we present on the x -axis, for each natural image of ImageNet and its respective adversarial example, the norm ratio of the gradients $e_{y_t} \mathbf{g}_{y_t}$ and $e_{y_{adv}} \mathbf{g}_{y_{adv}}$, for which we used the APGD attack [17] and a robust classifier [24]. All inputs are then transformed using our method ET3, and the difference between the resulting output logits is presented in the y -axis. One can see that many natural and adversarial example has high ratio values, i.e., $C \gg 1$. Moreover, one can see that transforming samples with a higher ratio C results in a larger logit difference for the transformed image, indicating successful transformation. For the locally linearity assumption, previous works showed that robust networks are experimentally approximately locally linear [52].

In Theory. When proving robustness of a model, a helpful property is have an approximate locally linear model, allowing a more immediate upper bound on the effectiveness of a worst-case perturbation. To this end, and to also allow correct classification, one idea for a robust network construction will be to assign higher inner products between the data

points and the local gradients corresponding to the ground truth label, and smaller inner product with the others, or assigning a sufficiently large bias. In previous works, [26, 87] construct a robust network enforcing both local linearity with a large radius and a larger gradient of the ground truth label in the settings of almost orthogonal data. In [54] and [53], the network trained using polynomial ReLU is approximately locally linear and with a similar larger ground truth gradient. Sec. 4.1 gives more details on the robust construction from [26] for which both assumptions hold, adapting the classic binary classifier to our two-logit regime.

We note that as there are no assumptions of $\mathbf{x}$, the ET3 defense will be successful for both clean sampled and adversarial data input, yet the assumption are milder on a clean data sample. First, since $r_x > 0$ for a clean sample, the assumption on ϵ will hold for any $\epsilon > 0$. Second, $\exp(f_\theta(\mathbf{x})_{\hat{y}_t} - f_\theta(\mathbf{x})_{y_t})$ will be smaller than 1, inducing a relaxed lower bound for the ratio C .

This theorem extends to the case of multi-step transformation within the same budget. Since the classifier is locally linear, each step to minimize the energy, will still increase the ground truth logit and the gap between it and the other logit. In other words, the theorem shows that

$$f_\theta(\mathbf{x} + \mathbf{z})_{y_t} - f_\theta(\mathbf{x} + \mathbf{z})_{\hat{y}_t} > f_\theta(\mathbf{x})_{y_t} - f_\theta(\mathbf{x})_{\hat{y}_t}$$

for a transformation $\mathbf{z}$ of any size. Therefore, the new input $\mathbf{x} + \mathbf{z}$ will preserve the assumptions on C and ϵ , allowing a multi-step attack with the same budget ϵ to take multiple small gradient steps till accumulated to a sufficient size.

4.1. Theoretical Application

We look at the construction for a robust two-layer ReLU network, stated in [26].

Data. We use the following data distribution $\mathcal{D}_{\text{clusters}}$ on $\mathbb{R}^d \times \{-1, 1\}$ in which we have k clusters with means $\mu^{(1)}, \dots, \mu^{(k)} \in \mathbb{R}^d$ and covariance $\sigma^2 I_d$, where examples in the j -th cluster are labeled $y^{(j)} \in \{-1, 1\}$.

Model. We adapt the two-layer binary classification network into our multi-class network. For a given binary classifier $f_{\theta}(\mathbf{x}) = \sum_{j \in J} v_j \sigma(\mathbf{w}_j^T \mathbf{x} + b_j)$, we define a two-logits output classifier as following. We define $J_+ = \{j : v_j \geq 0\}$ and $J_- = \{j : v_j < 0\}$, and denote for $l \in \{-1, 1\}$

$$\bar{f}_{\theta}(\mathbf{x})_l = l \sum_{j \in J_l} v_j \sigma(\mathbf{w}_j^T \mathbf{x} + b_j),$$

We note that for any $\mathbf{x}$, $f_{\theta}(\mathbf{x}) \equiv \bar{f}_{\theta}(\mathbf{x})_1 - \bar{f}_{\theta}(\mathbf{x})_{-1}$.

In [26], the authors present a robust network, proving its robustness to a large $O(\sqrt{d})$ adversarial perturbation.

For the robust network of width k , they take for any $j \in [k]$, $v_j = y^{(j)}$ and $\mathbf{w}_j = \frac{4\mu^{(j)}}{d}$ and $b_j = -2$ (for wider networks, i.e. $J > k$, we set the extra weights to be zero).

Let $(\mathbf{x}, y) \in \mathcal{D}_{\text{clusters}}$, meaning that $\mathbf{x} = \mu^{(q)} + \xi$ for some $\xi \sim \mathcal{N}(0, \sigma^2 I_d)$ and $y = y^{(q)}$. For the gradient ratio and local linearity assumptions, we look at the analysis for $\mathbf{w}_j^T \mathbf{x} + b_j$ for different $j \in [k]$, and get that for $j = q$ $\mathbf{w}_j^T \mathbf{x} + b_j > 1$, while for $j \neq q$, $\mathbf{w}_j^T \mathbf{x} + b_j < -1$. Thus, denoting $\mathbb{1}_{\{A\}}$ as the indicator function for an event A , we have with high probability (w.h.p.)

$$\mathbf{g}_{y_t} = \frac{\partial \bar{f}_{\theta}(\mathbf{x})_{y_t}}{\partial \mathbf{x}} = \sum_{j \in J_+} \mathbf{w}_j \mathbb{1}_{\{\mathbf{w}_j^T \mathbf{x} + b_j \geq 0\}} = y \mathbf{w}_q,$$

and $\mathbf{g}_{y_t} = 0$, satisfying the ratio assumption. In Theorem 4.1 the authors show that w.h.p. over $\mathbf{x}$, the network is locally linear in $\mathcal{B}_{\epsilon}(\mathbf{x})$ for the large $\epsilon \leq \frac{\sqrt{d}}{8}$. Applying the ET3 defense, we can see that the transformation will be $\mathbf{z} = \alpha \mathbf{w}_q$ for some $\alpha > 0$, which will ensure correct classification for the transformed input.

5. Experiments

We evaluate ET3 in two threat models and in three settings: zero-shot classification with CLIP, vision-language tasks with LVLMs, and classification with image classifiers. The threat models differ in the attacker’s knowledge of the test-time defense: a defense-unaware setting, where we evaluate across all three settings, and a stronger defense-aware setting, where the attacker knows the defense exists and adapts the attack accordingly, for which we restrict evaluation to LVLMs and classifiers. For CLIP, improvements are observed both when ET3 is applied alone and when combined with complementary techniques such as test-time augmentation and test-time prompt tuning. Furthermore, the features obtained from the CLIP vision encoder obtained after applying ET3 consistently enhance the adversarial robustness of downstream LVLMs, demonstrating that ET3 yields consistent robustness gains across all settings.

5.1. Evaluating Zero-Shot Classification

We perform zero-shot classification using CLIP models with the robust vision encoders TeCoA [51] and FARE [72], and evaluate across 15 benchmark datasets, full implementation details are in App. A.1 provided in supp. material.

Attack setup. Unless stated otherwise, following standard practice [72], we use the two attacks from AutoAttack [17]: APGD-CE and APGD-DLR with 100 iterations each.

Results. In Tab. 1 we show our test-time defense ET3 consistently enhances robust accuracy on adversarially robust models (TeCoA and FARE) across 14 datasets under attacks with a perturbation magnitude of $\epsilon_a = 4/255$. These models trained against attacks with either $\epsilon_t = 2/255$ or $4/255$, where ET3 show substantial improvements even with the “weaker” models, which were trained against the weaker attacks. This trend is further illustrated in Fig. 4, where ET3 improves robustness even for larger magnitude attacks—see App. D.3 for per-dataset results and additional details. Next, in Tab. 2 we show that ET3 surpasses similar defense methods in diverse defense scopes. Following the comparison standardization of the previous methods, we report results on a subset of 8 datasets. In the **lightweight defense scope**, ET3 consistently improves robust accuracy over baseline CLIP model and TTC [96], and surpasses even the slower TPT and C-TPT methods [76, 101], despite requiring no training. Within the **multiple augmentation scope**, we show that an easy incorporation of the lightweight ET3 yields further gains to the ensemble that averages predictions across all augmented views. It surpassing MTA [104] and matching the state-of-the-art robustness of R-TPT [75]. In the last **augmentation-based methods with prompt-tuning scope**, we outperform existing methods by similarly combining ET3 with the former state-of-the-art R-TPT. Unless otherwise stated, the ϵ for ET3 is set to 5 for TeCoA and 4 for FARE, with the number of steps fixed at $T = 2$.

5.2. Evaluating Large Vision-Language Models

We evaluate ET3 on LLaVA 1.5-7B using as vision encoder (i) standard CLIP, (ii) TeCoA-robust CLIP [51], and (iii) FARE-robust CLIP [72]. For both robust methods, we employ the two variants trained with $\epsilon = 2/255$ and $\epsilon = 4/255$, marked as ² or ⁴ respectively. We consider image captioning (COCO, Flickr30k) and visual question answering tasks (TextVQA, VQAv2); evaluations on clean data use the full datasets, while adversarial results are computed over 500 perturbed inputs generated with an attack budget of $\epsilon_a = 4/255$, following the evaluation of [72]. Across all tasks, we employ ImageNet-21k text labels [66], to compute the energy. Further details in App. A.2.

Results. Under the defense-unaware setting shown in, Tab. 3, ET3 consistently improves the robustness across all base models while preserving performance on clean data. We further evaluate ET3 under defense-aware setting (avg.

Table 1. **Zero-shot robustness of ET3 across 14 benchmark datasets in the defense-unaware setting.** Comparison of clean and robust accuracy for baseline models versus same models augmented with ET3. Robustness is evaluated against Auto-Attack (AA) at $\epsilon_a = 4/255$.

Model	Method	ImageNet	CalTech	Cars	CIFAR10	CIFAR100	DTD	EuroSAT	FGVC	Flowers	ImageNet-R	ImageNet-S	PCAM	OxfordPets	STL-10	Avg.	Improv.
ViT-L/14 (TeCoA)	Base (Clean)	74.91	78.36	37.83	79.61	50.26	38.03	22.48	11.76	38.41	74.35	54.22	49.95	76.07	93.44	55.69	
	+ ET3 (Clean)	75.20	78.16	37.21	81.27	49.97	38.09	23.31	11.61	39.31	76.66	55.22	50.03	76.18	93.08	56.09	(+0.4)
	Base (Robust)	44.50	60.90	8.50	37.10	21.50	16.50	6.40	2.20	12.60	41.90	32.80	45.70	55.00	74.30	32.85	
	+ ET3 (Robust)	53.00	65.10	13.50	56.40	34.50	22.70	15.80	4.40	22.50	51.90	40.00	51.50	60.90	80.80	40.93	(+8.08)
ViT-L/14 (FARE)	Base (Clean)	70.78	84.70	63.84	77.67	56.53	43.83	18.28	21.96	58.07	80.24	56.74	50.02	87.14	96.04	61.85	
	+ ET3 (Clean)	70.97	84.73	62.96	80.58	55.91	43.94	18.93	21.96	58.86	82.49	57.25	50.02	86.59	96.17	62.24	(+0.39)
	Base (Robust)	34.80	64.20	12.70	34.80	20.20	17.50	11.10	3.00	12.20	40.50	30.60	52.30	50.60	74.30	32.77	
	+ ET3 (Robust)	41.20	69.40	18.40	53.50	33.10	26.80	14.70	8.30	24.20	50.80	37.40	52.30	58.50	79.80	40.60	(+7.83)
ViT-L/14 (TeCoA)	Base (Clean)	80.11	80.67	50.08	87.53	60.69	44.36	26.06	14.04	51.80	80.12	58.43	49.89	80.02	96.08	61.42	
	+ ET3 (Clean)	79.82	79.62	46.33	86.34	59.90	44.15	31.52	13.35	50.02	82.05	58.99	49.99	80.43	94.97	61.25	(-0.17)
	Base (Robust)	37.00	57.40	6.40	31.00	17.90	14.70	7.80	1.00	9.60	36.60	30.90	17.40	50.40	69.10	27.66	
	+ ET3 (Robust)	44.40	63.10	14.10	49.20	32.20	23.40	24.00	4.70	20.50	46.00	38.40	47.10	58.40	75.40	38.64	(+10.98)
ViT-L/14 (FARE)	Base (Clean)	74.48	84.77	70.53	89.52	69.13	50.05	25.39	26.70	70.60	85.52	59.72	50.01	91.06	98.47	67.57	
	+ ET3 (Clean)	74.07	84.60	68.31	89.64	66.64	48.03	32.98	24.90	69.47	86.94	59.86	50.03	90.38	98.00	67.42	(-0.15)
	Base (Robust)	17.80	46.40	5.00	25.70	14.20	11.60	0.40	0.90	7.10	25.60	22.10	19.10	28.10	61.50	20.39	
	+ ET3 (Robust)	25.20	56.40	12.20	43.90	28.70	21.90	25.30	6.50	16.00	35.10	29.40	35.90	38.90	69.10	31.75	(+11.36)

Table 2. **Robustness of ET3 on fine-grained classification in the defense-unaware setting.** Comparison against test-time adaptation techniques across eight fine-grained datasets. All defenses are applied to a TeCoA pre-trained CLIP-ViT-B/32 model and evaluated against PGD-100 ($\epsilon = 4/255$). ET3 consistently improves robustness, both as a standalone method and when combined with other defenses.

Method	Caltech101		Pets		Cars		Flower102		Aircraft		DTD		EuroSAT		UCF101		Avg.		Improv.
	Acc.	Rob.	Acc.	Rob.	Acc.	Rob.	Acc.	Rob.	Acc.	Rob.	Acc.	Rob.	Acc.	Rob.	Acc.	Rob.	Acc.	Rob.	
CLIP (Robust)	78.82	43.45	66.88	15.94	10.20	0.99	30.82	9.09	6.60	0.45	24.53	10.70	14.53	10.78	34.58	6.71	33.37	12.26	
<i>Lightweight Defense using Image Transformation (Vision Input Only)</i>																			
TTC [96]	71.26	44.90	65.00	21.10	10.62	1.39	26.02	8.93	6.93	0.69	23.99	11.38	20.84	12.34	32.91	10.45	32.20	13.90	(+1.64)
ET3 (ours)	79.07	<u>50.59</u>	66.86	<u>27.15</u>	10.32	<u>2.04</u>	28.91	<u>12.10</u>	5.28	<u>0.87</u>	24.00	<u>13.18</u>	13.37	11.17	37.35	<u>16.02</u>	33.15	<u>16.51</u>	(+4.25)
<i>Defenses using Multiple Augmentations (Vision Input Only)</i>																			
Ensemble	73.02	55.66	59.96	38.35	5.55	2.80	26.35	16.12	4.29	1.89	23.82	15.96	12.51	11.01	26.35	14.12	28.98	19.4	(+7.14)
MTA [104] *	79.70	55.70	66.20	31.20	9.00	2.50	29.10	14.00	6.50	1.60	24.40	13.50	13.30	<u>11.20</u>	34.60	12.50	32.90	17.80	(+5.54)
Ensemble + ET3	<u>76.23</u>	<u>61.38</u>	62.03	<u>43.17</u>	5.77	<u>3.03</u>	26.07	<u>18.11</u>	3.93	<u>2.22</u>	22.64	<u>17.55</u>	11.88	11.14	32.62	<u>21.44</u>	30.14	<u>22.26</u>	(+10)
<i>Defenses using Multiple Augmentations + Prompt Tuning (Vision + Text Input)</i>																			
TPT [76] *	79.30	52.70	65.20	27.40	9.60	2.00	27.90	12.30	6.70	1.70	25.50	14.60	12.20	11.20	34.90	10.20	32.70	16.50	(+4.24)
C-TPT [101] *	79.80	47.30	66.10	19.50	10.60	1.30	29.40	10.70	6.40	0.70	26.20	12.40	13.00	11.10	36.40	8.10	33.50	13.90	(+1.64)
R-TPT [75]	75.58	61.01	63.75	41.43	5.57	2.79	26.29	16.57	5.73	2.46	25.59	18.09	11.46	11.30	31.48	17.63	30.68	21.41	(+9.15)
R-TPT+ ET3	79.27	65.07	63.86	45.71	6.21	3.27	26.51	18.72	5.85	3.51	26.42	19.92	11.37	11.16	36.93	24.72	32.05	23.88	(+11.62)

results in Tab. 5; detailed results in App. C.3). ET3 introduces minimal overhead with increase in inference time by as little as 2.3% in single-step defense (details in App. A.3).

5.3. Evaluation with Robust Classifiers

We evaluate ET3 under defense-aware setting on robust classifiers from RobustBench [18] using the ImageNet dataset. We compare our method against other test-time transformation defenses as well as popular strong defenses, including adversarial training and adversarial purification. We assess ET3 under *adaptive attacks*, where the attacker

has full knowledge of the defense and actively attempts to circumvent it as suggested by [19]. Following [19], we adopt a strong attack configuration, targeted APGD-DLR with gradients approximated through the defense via Backward Pass Differentiable Approximation (BPDA) [4], effectively accounting for the entire test-time optimization process. We also perform a transfer attack from the static base model and report the worst-case accuracy (more details in App. A.5). Due to the computational cost of adaptive attacks, we restrict evaluation to ResNet-50 classifiers. For other competing defenses, we show robustness using the

Table 3. **Evaluation of ET3 on LLaVA 1.5-7B with different vision encoders in defense-unaware setting.** Clean and ℓ_∞ -robust performance ($\epsilon_a = 4/255$) using standard CLIP and the TeCoA/FARE backbones adversarially trained with $\epsilon_t = 2/255$ and $\epsilon_t = 4/255$. Clean results use full test sets, while adversarial scores are computed on 500 APGD perturbations following the ensemble protocol of [72]. Across all tasks, ImageNet-21k labels serve as the reference text embeddings for computing the energy $E(\cdot, \theta)$. ET3 performs two gradient descent iterations. COCO and Flickr30k are evaluated with CIDEr for captioning, while TextVQA and VQAv2 report VQA accuracy.

	COCO [44]				Flickr30k [103]				TextVQA [77]				VQAv2 [29]				Average			
	Clean	+ET3	4/255	+ET3	Clean	+ET3	4/255	+ET3	Clean	+ET3	4/255	+ET3	Clean	+ET3	4/255	+ET3	Clean	+ET3	4/255	+ET3
CLIP	115.5	111.3	2.7	61.1	77.5	76.1	1.1	37.1	37.1	34.9	0.2	17.1	74.5	73.1	0.0	40.0	76.2	73.9 (-2.3)	1.0	38.8 (+37.8)
TeCoA ²	98.4	99.0	30.0	56.0	57.1	58.1	14.8	32.1	24.1	24.2	7.8	13.2	66.9	67.7	25.1	40.7	61.6	62.3 (+0.7)	19.4	35.5 (+16.1)
FARE ²	109.9	110.2	32.5	52.7	71.1	71.5	17.5	31.9	31.9	31.9	7.2	15.3	71.7	71.7	24.5	32.6	71.2	71.3 (+0.1)	20.4	33.1 (+12.7)
TeCoA ⁴	88.3	88.2	34.4	53.7	48.6	48.2	19.5	30.2	20.7	21.0	9.5	12.6	63.2	63.2	31.1	42.6	55.2	55.2 (+0.0)	23.6	34.8 (+11.2)
FARE ⁴	102.4	103.0	42.2	56.4	61.6	61.8	23.1	33.5	27.6	27.4	10.2	17.2	68.3	68.3	29.5	42.2	65.0	65.1 (+0.1)	26.3	37.3 (+11.0)

Table 4. **Defense-aware worst-case robustness on ImageNet.** Clean and ℓ_∞ -robust accuracy ($\epsilon_a = 4/255$) of base model. All TTT methods build upon the same base robust model from [70].

Method (Architecture)	Adaptive Attack	Clean	Robust
Adversarial Purification			
DiffPure [59] (WRN-50-2)	DiffBreak [36]	74.22	12.11
DiffPure [59] (DeiT-S)	DiffBreak [36]	73.63	25.00
GDMP [91] (DeiT-S)	DiffBreak [36]	69.14	20.70
TTT — Base: Salman et al. (RN-50) [70]			
Base Model Only	AutoAttack [17]	64.02	34.96
+ Singh et al. [78]	APGD-DLR [17]	63.91	34.68 (-0.28)
+ Kulkarni et al. [40]	IW-WC [40]	64.10	35.48 (+0.52)
+ ET3 (Ours)	Tr. APGD-T + BPDA [19]	63.12	37.70 (+2.74)

strongest attacks appropriate for each method—for example, DiffBreaker [36] for diffusion-based purification defenses, and AutoAttack for robust classifiers.

Results. Tab. 4 reports the clean and worst-case robust accuracies i.e. the lowest accuracies achieved under the strongest attacks reported in the original papers, including both defense and attack evaluations where applicable. Methods are organized by defense family—adversarial purification, and test-time transformation (TTT)—and worst-case robustness is reported using the strongest adaptive attack available for each method. Consistent with prior work, DiffBreak significantly reduces the robustness of diffusion-based purification methods, whereas adversarially trained models maintain high robustness. In the TTT setting, where all methods use the same robust model as their base classifier [70], ET3 performs the best and improves worst-case robust accuracy. To ensure meaningful comparisons, we focus on worst-case robustness under the relevant adaptive attack for each method and omit defenses previously shown to fail under such evaluations [19]. We also present more results under the defense-unaware setting with larger transformer-based architectures in App. C.2.

For this section, we further provide discussion on budget fairness of defense w.r.t to attack, label set ablation, and a single-step variant of ET3 for VLMs in App. D.

Figure 4. **Robust accuracy across increasing attack strengths in the defense-unaware setting.** Average zero-shot accuracy of CLIP over 14 benchmark datasets, showing that ET3 consistently improves the robustness of the TeCoA models trained with different defense strengths (ϵ_t) as the attack strength (ϵ_a) increases.

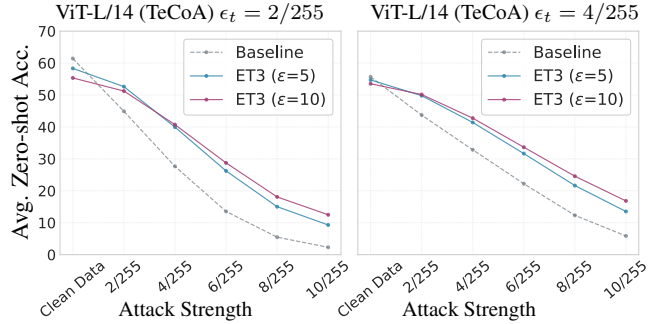


Table 5. **Defense-aware adaptive attack evaluation on LLaVA.** Average robust accuracy over four datasets (200 samples). Legend: B = baseline (no defense); N-Ad. = non-adaptive; Ad. = adaptive; blue = ET3 with ℓ_2 proj.; orange = ET3 with ℓ_∞ proj.

CLIP			FARE ²			TeCoA ²			FARE ⁴			TeCoA ⁴		
B	N-Ad.	Ad.	B	N-Ad.	Ad.	B	N-Ad.	Ad.	B	N-Ad.	Ad.	B	N-Ad.	Ad.
0.9	16.7	4.5	17.6	28.9	25.0	18.3	30.9	25.5	24.1	32.8	31.0	19.7	30.9	28.6
0.9	12.4	3.3	17.6	27.0	23.7	18.3	31.3	26.9	24.1	32.5	31.1	19.7	31.1	28.1

6. Conclusion

In this paper, we present the ET3 defense, a novel test-time transformation method based on energy minimization. We showed extensive experiments of boosting robustness for image classification, zero-shot classification using CLIP, and with Large Vision Language Models, demonstrating superiority over many datasets and downstream tasks. We also present theoretical justification for which our ET3 will provably transform a clean or adversarial image into correctly classified images. A promising future direction is to optimize the network to support conditions that enable the ET3 defense, such as increasing its local linearity radius or the energy gradient ratio.

Acknowledgment. Supported by projects PNRR MUR PE0000013-FAIR under the MUR National Recovery and Resilience Plan funded by the European Union - NextGenerationEU, PRIN 2022 project 20227YET9B “AdVVent” CUP code B53D23012830006. Also partially supported by Sapienza research project BEAT (Better dEep leArning securiTy), bando per la ricerca di Ateneo 2024. We acknowledge CINECA HPC resources via ISCRA (HP10CIWXXH) and EuroHPC (EHPC-DEV-2025D07-037) for computing.

References

- [1] Motasem Alfarra, Adel Bibi, Philip HS Torr, and Bernard Ghanem. Data dependent randomized smoothing. In *Uncertainty in Artificial Intelligence*, pages 64–74. PMLR, 2022. 2
- [2] Motasem Alfarra, Juan C Pérez, Ali Thabet, Adel Bibi, Philip HS Torr, and Bernard Ghanem. Combating adversaries with anti-adversaries. In *AAAI Conference on Artificial Intelligence*, 2022. 2
- [3] Maksym Andriushchenko, Francesco Croce, Nicolas Flammarion, and Matthias Hein. Square attack: a query-efficient black-box adversarial attack via random search. In *ECCV*, 2020. 2
- [4] Anish Athalye, Nicholas Carlini, and David Wagner. Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples. In *ICML*, 2018. 7, 15
- [5] Anas Awadalla, Irena Gao, Josh Gardner, Jack Hessel, Yusuf Hanafy, Wanrong Zhu, Kalyani Marathe, Yonatan Bitton, Samir Gadre, Shiori Sagawa, et al. Openflamingo: An open-source framework for training large autoregressive vision-language models. *arXiv preprint arXiv:2308.01390*, 2023. 1
- [6] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. *arXiv preprint arXiv:2308.12966*, 2023. 1
- [7] Peter Bartlett, Sébastien Bubeck, and Yeshwanth Cherapanamjeri. Adversarial examples in multi-layer random relu networks. In *NeurIPS*, 2021. 3
- [8] Tsachi Blau, Roy Ganz, Chaim Baskin, Michael Elad, and Alex M. Bronstein. Class-conditioned transformation for enhanced robust image classification. In *WACV*, 2025. 2
- [9] Sébastien Bubeck, Yin Tat Lee, Eric Price, and Ilya Razenshteyn. Adversarial examples from computational constraints. In *ICML*, 2019. 3
- [10] Sébastien Bubeck, Yeshwanth Cherapanamjeri, Gauthier Gidel, and Rémi Tachet des Combes. A single gradient step finds adversarial examples on random two-layers neural networks. In *NeurIPS*, 2021. 3
- [11] Nicholas Carlini, Milad Nasr, Christopher A Choquette-Choo, Matthew Jagielski, Irena Gao, Pang Wei W Koh, Daphne Ippolito, Florian Tramèr, and Ludwig Schmidt. Are aligned neural networks adversarially aligned? *NeurIPS*, 2023. 1, 2
- [12] Mehdi Cherti and Romain Beaumont. Clip benchmark. <https://zenodo.org/records/15403103>, 2025. 13
- [13] Mehdi Cherti, Romain Beaumont, Ross Wightman, Mitchell Wortsman, Gabriel Ilharco, Cade Gordon, Christoph Schuhmann, Ludwig Schmidt, and Jenia Jitsev. Reproducible scaling laws for contrastive language-image learning. In *CVPR*, 2023. 13
- [14] Mircea Cimpoi, Subhansu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing textures in the wild. In *CVPR*, 2014. 13
- [15] Adam Coates, Andrew Ng, and Honglak Lee. An analysis of single-layer networks in unsupervised feature learning. In *Proceedings of the fourteenth international conference on artificial intelligence and statistics. JMLR Workshop and Conference Proceedings*, 2011. 13
- [16] Jeremy Cohen, Elan Rosenfeld, and Zico Kolter. Certified adversarial robustness via randomized smoothing. In *ICML*, pages 1310–1320. PMLR, 2019. 2
- [17] Francesco Croce and Matthias Hein. Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks. In *ICML*, 2020. 5, 6, 8, 13
- [18] Francesco Croce, Maksym Andriushchenko, Vikash Sehwag, Edoardo Debenedetti, Nicolas Flammarion, Mung Chiang, Prateek Mittal, and Matthias Hein. Robustbench: a standardized adversarial robustness benchmark. In *ICLR Workshops 2021 - Workshop on Security and Safety in Machine Learning Systems*, 2021. 7, 15
- [19] Francesco Croce, Sven Gowal, Thomas Brunner, Evan Shelhamer, Matthias Hein, and Taylan Cemgil. Evaluating the adversarial robustness of adaptive test-time defenses. In *ICML*. PMLR, 2022. 7, 8, 15, 17
- [20] Amit Daniely and Hadas Shacham. Most relu networks suffer from ℓ_2 adversarial perturbations. In *NeurIPS*, 2020. 3
- [21] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009. 13
- [22] Yilun Du and Igor Mordatch. Implicit generation and modeling with energy based models. *NeurIPS*, 2019. 3
- [23] Nikita Durasov, Assaf Shocher, Doruk Oner, Gal Chechik, Alexei A Efros, and Pascal Fua. IT³: Idempotent test-time training. In *ICML*, 2025. 2
- [24] Logan Engstrom, Andrew Ilyas, Shibani Santurkar, Dimitris Tsipras, Brandon Tran, and Aleksander Madry. Adversarial robustness as a prior for learned representations. *arXiv preprint arXiv:1906.00945*, 2019. 5
- [25] Li Fei-Fei, Rob Fergus, and Pietro Perona. Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. In *CVPR Workshops*, 2004. 13
- [26] Spencer Frei, Gal Vardi, Peter Bartlett, and Nati Srebro. The double-edged sword of implicit bias: Generalization vs. robustness in relu networks. *NeurIPS*, 2023. 2, 3, 5, 6
- [27] Yossi Gandelsman, Yu Sun, Xinlei Chen, and Alexei Efros. Test-time training with masked autoencoders. *NeurIPS*, 2022. 2
- [28] Ian Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In *ICLR*, 2015. 1, 2

- [29] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *CVPR*, 2017. 8, 16, 20
- [30] Will Grathwohl, Kuan-Chieh Wang, Jörn-Henrik Jacobsen, David Duvenaud, Mohammad Norouzi, and Kevin Swersky. Your classifier is secretly an energy based model and you should treat it like one. In *ICLR*, 2020. 2, 3
- [31] Patrick Helber, Benjamin Bischke, Andreas Dengel, and Damian Borth. Eurosat: A novel dataset and deep learning benchmark for land use and land cover classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 2019. 13
- [32] Dan Hendrycks, Norman Mu, Ekin D Cubuk, Barret Zoph, Justin Gilmer, and Balaji Lakshminarayanan. Augmix: A simple data processing method to improve robustness and uncertainty. *arXiv preprint arXiv:1912.02781*, 2019. 13
- [33] Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadam, Frank Wang, Evan Doro, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, et al. The many faces of robustness: A critical analysis of out-of-distribution generalization. In *ICCV*, 2021. 13
- [34] Mitch Hill, Jonathan Craig Mitchell, and Song-Chun Zhu. Stochastic security: Adversarial defense using long-run dynamics of energy-based models. In *ICLR*, 2021. 2, 3
- [35] Kaichao Jiang, He Wang, Xiaoshuai Hao, Xiulong Yang, Aijian Liu, Qi Chu, and Yunfeng Diao. Your classifier can do more: Towards bridging the gaps in classification, robustness, and generation. *arXiv preprint arXiv:2505.19459*, 2025. 3
- [36] A. Kassis, U. Hengartner, and Y. Yu. DiffBreak: Breaking diffusion-based purification with adaptive attacks. In *NeurIPS*, 2025. 8
- [37] Diederik Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR*, 2014. 13
- [38] Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. In *CVPR Workshops*, 2013. 13
- [39] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. Technical report, Citeseer, 2009. 13
- [40] Akshay Kulkarni and Tsui-Wei Weng. Interpretability-guided test-time adversarial defense. In *ECCV*, 2024. 8
- [41] Yann LeCun, Sumit Chopra, Raia Hadsell, M Ranzato, and F Huang. A tutorial on energy-based learning. *Predicting structured data*, 1(0), 2006. 3
- [42] Mathias Lecuyer, Vaggelis Atlidakis, Roxana Geambasu, Daniel Hsu, and Suman Jana. Certified robustness to adversarial examples with differential privacy. In *IEEE Symposium on Security and Privacy (SP)*, 2019. 2
- [43] Lin Li, Haoyan Guan, Jianing Qiu, and Michael Spratling. One prompt word is enough to boost adversarial robustness for pre-trained vision-language models. In *CVPR*, 2024. 2
- [44] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, Lubomir D. Bourdev, Ross B. Girshick, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: common objects in context. *CoRR*, abs/1405.0312, 2014. 8, 16, 20
- [45] Chang Liu, Yinpeng Dong, Wenzhao Xiang, Xiao Yang, Hang Su, Jun Zhu, Yuefeng Chen, Yuan He, Hui Xue, and Shibao Zheng. A comprehensive study on robustness of image classification models: Benchmarking and rethinking. *IJCV*, 2025. 15
- [46] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *NeurIPS*, 2023. 1
- [47] Weitang Liu, Xiaoyun Wang, John Owens, and Yixuan Li. Energy-based out-of-distribution detection. *NeurIPS*, 2020. 3
- [48] Saiyue Lyu, Shadab Shaikh, Frederick Shpilevskiy, Evan Shelhamer, and Mathias Lécuyer. Adaptive randomized smoothing: Certified adversarial robustness for multi-step defences. In *NeurIPS*, 2024. 2
- [49] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *ICLR*, 2018. 1, 2, 13
- [50] Subhansu Maji, Esa Rahtu, Juho Kannala, Matthew Blaschko, and Andrea Vedaldi. Fine-grained visual classification of aircraft. *arXiv preprint arXiv:1306.5151*, 2013. 13
- [51] Chengzhi Mao, Scott Geng, Junfeng Yang, Xin Wang, and Carl Vondrick. Understanding zero-shot adversarial robustness for large-scale models. In *ICLR*, 2023. 2, 6, 13, 21
- [52] Odelia Melamed, Gilad Yehudai, and Adi Shamir. Malt powers up adversarial attacks. In *NeurIPS*, 2024. 4, 5
- [53] Hancheng Min and René Vidal. Can implicit bias imply adversarial robustness? *arXiv preprint arXiv:2405.15942*, 2024. 2, 3, 5
- [54] Hancheng Min and Rene Vidal. Gradient flow provably learns robust classifiers for orthonormal GMMs. In *ICML*, 2025. 2, 3, 5
- [55] Mujtaba Hussain Mirza, Maria Rosaria Briglia, Senad Beadini, and Iacopo Masi. Shedding more light on robust classifiers under the lens of energy-based models. In *ECCV*, 2024. 3
- [56] Mujtaba Hussain Mirza, Maria Rosaria Briglia, Filippo Bartolucci, Senad Beadini, Giuseppe Lisanti, and Iacopo Masi. Understanding adversarial training with energy-based models, 2025. 3
- [57] Andrea Montanari and Yuchen Wu. Adversarial examples in random neural networks with general activations. *arXiv preprint arXiv:2203.17209*, 2022. 3
- [58] Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, and Pascal Frossard. Deepfool: a simple and accurate method to fool deep neural networks. In *CVPR*, 2016. 2
- [59] Weili Nie, Brandon Guo, Yujia Huang, Chaowei Xiao, Arash Vahdat, and Animashree Anandkumar. Diffusion models for adversarial purification. In *ICML*, 2022. 2, 8
- [60] Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In *2008 Sixth Indian conference on computer vision, graphics & image processing*, 2008. 13
- [61] Omkar M Parkhi, Andrea Vedaldi, Andrew Zisserman, and CV Jawahar. Cats and dogs. In *CVPR*, 2012. 13

- [62] Juan C. Pérez, Motasem Alfarra, Guillaume Jeanneret, Laura Rueda, Ali Thabet, Bernard Ghanem, and Pablo Arbeláez. Enhancing adversarial robustness via test-time transformation ensembling. In *ICCV*, 2021. 2
- [63] Brian Pulfer, Yury Belousov, and Slava Voloshynovskiy. Robustness tokens: Towards adversarial robustness of transformers. In *ECCV*, 2024. 2
- [64] Xiangyu Qi, Kaixuan Huang, Ashwinee Panda, Peter Henderson, Mengdi Wang, and Prateek Mittal. Visual adversarial examples jailbreak aligned large language models. In *AAAI Conference on Artificial Intelligence*, 2024. 1, 2
- [65] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, 2021. 1
- [66] Tal Ridnik, Emanuel Ben-Baruch, Asaf Noy, and Lihi Zelnik-Manor. Imagenet-21k pretraining for the masses, 2021. 4, 6
- [67] Elias Abad Rocamora, Christian Schlarmann, Naman Deep Singh, Yongtao Wu, Matthias Hein, and Volkan Cevher. Robustness in both domains: Clip needs a robust text encoder, 2025. 2
- [68] Briglia Maria Rosaria, Mujtaba Hussain Mirza, Giuseppe Lisanti, and Iacopo Masi. Why adversarially train diffusion models? In *ICLR*, 2026. 2
- [69] Mozhdeh Rouhsedaghat, Masoud Monajatipoor, C-C Jay Kuo, and Iacopo Masi. MAGIC: Mask-guided image synthesis by inverting a quasi-robust classifier. In *AAAI Conference on Artificial Intelligence*, 2023. 3
- [70] Hadi Salman, Andrew Ilyas, Logan Engstrom, Ashish Kapoor, and Aleksander Madry. Do adversarially robust imagenet models transfer better? In *NeurIPS*, 2020. 8, 15
- [71] Christian Schlarmann and Matthias Hein. On the adversarial robustness of multi-modal foundation models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3677–3685, 2023. 2
- [72] Christian Schlarmann, Naman Deep Singh, Francesco Croce, and Matthias Hein. Robust CLIP: Unsupervised adversarial fine-tuning of vision embeddings for robust large vision-language models. In *ICML*, 2024. 1, 2, 6, 8, 13, 14, 15, 16
- [73] Adi Shamir, Itay Safran, Eyal Ronen, and Orr Dunkelman. A simple explanation for the existence of adversarial examples with small hamming distance. *arXiv preprint arXiv:1901.10861*, 2019. 3
- [74] Adi Shamir, Odelia Melamed, and Oriel BenShmuel. The dimpled manifold model of adversarial examples in machine learning. *arXiv preprint arXiv:2106.10151*, 2021. 3
- [75] Lijun Sheng, Jian Liang, Zilei Wang, and Ran He. R-tpt: Improving adversarial robustness of vision-language models through test-time prompt tuning. In *CVPR*, 2025. 2, 6, 7, 13, 14
- [76] Manli Shu, Weili Nie, De-An Huang, Zhiding Yu, Tom Goldstein, Anima Anandkumar, and Chaowei Xiao. Test-time prompt tuning for zero-shot generalization in vision-language models. *NeurIPS*, 2022. 2, 6, 7, 13, 14
- [77] Amanpreet Singh, Vivek Natarjan, Meet Shah, Yu Jiang, Xinlei Chen, Devi Parikh, and Marcus Rohrbach. Towards vqa models that can read. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8317–8326, 2019. 8, 16, 20
- [78] Anurag Singh, Mahalakshmi Sabanayagam, Krikamol Muandet, and Debarghya Ghoshdastidar. Robust feature inference: A test-time defense strategy using spectral projections. *Transactions on Machine Learning Research*, 2024. 8
- [79] Yang Song, Taesup Kim, Sebastian Nowozin, Stefano Ermon, and Nate Kushman. Pixeldefend: Leveraging generative models to understand and defend against adversarial examples. In *ICLR*, 2018. 2, 3
- [80] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012. 13
- [81] David Stutz, Matthias Hein, and Bernt Schiele. Disentangling adversarial robustness and generalization. In *CVPR*, 2019. 3
- [82] Yu Sun, Xiaolong Wang, Zhuang Liu, John Miller, Alexei Efros, and Moritz Hardt. Test-time training with self-supervision for generalization under distribution shifts. In *ICML*, 2020. 2
- [83] Yu Sun, Xinhao Li, Karan Dalal, Jiarui Xu, Arjun Vikram, Genghan Zhang, Yann Dubois, Xinlei Chen, Xiaolong Wang, Sanmi Koyejo, Tatsunori Hashimoto, and Carlos Guestrin. Learning to (learn at test time): RNNs with expressive hidden states. In *ICML*, 2025. 2
- [84] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. In *ICLR*, 2014. 1
- [85] Keke Tang, Tianrui Lou, Weilong Peng, Nenglu Chen, Yawen Shi, and Wenping Wang. Effective single-step adversarial training with energy-based models. *IEEE Transactions on Emerging Topics in Computational Intelligence*, 2024. 3
- [86] Zhuowen Tu. Learning generative models via discriminative approaches. In *CVPR*, 2007. 3
- [87] Gal Vardi, Gilad Yehudai, and Ohad Shamir. Gradient methods provably converge to non-robust networks. *Advances in Neural Information Processing Systems*, 35: 20921–20932, 2022. 2, 3, 5
- [88] Bastiaan S Veeling, Jasper Linmans, Jim Winkens, Taco Cohen, and Max Welling. Rotation equivariant cnns for digital pathology. In *International Conference on Medical image computing and computer-assisted intervention*. Springer, 2018. 13
- [89] Dequan Wang, Evan Shelhamer, Shaoteng Liu, Bruno Olshausen, and Trevor Darrell. Tent: Fully test-time adaptation by entropy minimization. In *ICLR*, 2021. 2
- [90] Haohan Wang, Songwei Ge, Zachary Lipton, and Eric P Xing. Learning robust global representations by penalizing local predictive power. *NeurIPS*, 2019. 13
- [91] Jinyi Wang, Zhaoyang Lyu, Dahua Lin, Bo Dai, and Hongfei Fu. Guided diffusion model for adversarial purification. *arXiv preprint arXiv:2205.14969*, 2022. 8

- [92] Sibor Wang, Jie Zhang, Zheng Yuan, and Shiguang Shan. Pre-trained model guided fine-tuning for zero-shot adversarial robustness. In *CVPR*, 2024. [2](#)
- [93] Yifei Wang, Yisen Wang, Jiansheng Yang, and Zhouchen Lin. A unified contrastive energy-based model for understanding the generative ability of adversarial training. In *ICLR*, 2022. [3](#)
- [94] Boxi Wu, Heng Pan, Li Shen, Jindong Gu, Shuai Zhao, Zhifeng Li, Deng Cai, Xiaofei He, and Wei Liu. Attacking adversarial attacks as a defense. *arXiv preprint arXiv:2106.04938*, 2021. [2](#)
- [95] Jianwen Xie, Yang Lu, Song-Chun Zhu, and Yingnian Wu. A theory of generative convnet. In *ICML*. PMLR, 2016. [3](#)
- [96] Songlong Xing, Zhengyu Zhao, and Nicu Sebe. Clip is strong enough to fight back: Test-time counterattacks towards zero-shot adversarial robustness of clip. In *CVPR*, 2025. [2](#), [6](#), [7](#), [13](#), [14](#)
- [97] Xiulong Yang and Shihao Ji. M-ebm: Towards understanding the manifolds of energy-based models. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pages 291–302. Springer, 2023. [3](#)
- [98] Xiulong Yang, Qing Su, and Shihao Ji. Towards bridging the performance gaps of joint energy-based models. In *CVPR*, pages 15732–15741, 2023. [3](#)
- [99] Xuwang Yin, Shiyang Li, and Gustavo K Rohde. Learning energy-based models with adversarial training. In *ECCV*. Springer, 2022. [3](#)
- [100] Xuwang Yin, Claire Zhang, Julie Steele, Nir Shavit, and Tony T Wang. Joint discriminative-generative modeling via dual adversarial training. *arXiv preprint arXiv:2510.13872*, 2025. [3](#)
- [101] Hee Suk Yoon, Eunseop Yoon, Joshua Tian Jin Tee, Mark A. Hasegawa-Johnson, Yingzhen Li, and Chang D. Yoo. C-TPT: Calibrated test-time prompt tuning for vision-language models via text feature dispersion. In *ICLR*, 2024. [2](#), [6](#), [7](#), [13](#), [14](#)
- [102] Jongmin Yoon, Sung Ju Hwang, and Juho Lee. Adversarial purification with score-based generative models. In *International Conference on Machine Learning*, pages 12062–12072. PMLR, 2021. [2](#)
- [103] Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *TACL*, 2:67–78, 2014. [8](#), [16](#), [20](#)
- [104] Maxime Zanella and Ismail Ben Ayed. On the test-time zero-shot generalization of vision-language models: Do we really need prompt learning? In *CVPR*, 2024. [2](#), [6](#), [7](#), [13](#), [14](#)
- [105] Jiaming Zhang, Xingjun Ma, Xin Wang, Lingyu Qiu, Jiaqi Wang, Yu-Gang Jiang, and Jitao Sang. Adversarial prompt tuning for vision-language models. In *ECCV*, 2024. [2](#)
- [106] Yao Zhu, Jiacheng Ma, Jiacheng Sun, Zewei Chen, Rongxin Jiang, Yaowu Chen, and Zhenguo Li. Towards understanding the generative capability of adversarially robust classifiers. In *ICCV*, 2021. [3](#)