# TEA: Test-time Energy Adaptation

Yige Yuan[1,2], Bingbing Xu[1*], Liang Hou[3], Fei Sun[1], Huawei Shen[1,2], Xueqi Cheng[1,2]

[1]CAS Key Laboratory of AI Safety, Institute of Computing Technology, Chinese Academy of Sciences
[2]University of Chinese Academy of Sciences [3]Kuaishou Technology

{yuanyige20z, xubingbing, sunfei, shenhuawei, cxq}@ict.ac.cn, lianghou96@gmail.com

## Abstract

*Test Time Adaptation (TTA) aims to improve model generalizability when test data diverges from training distribution, with the distinct advantage of not requiring access to training data and processes, especially valuable in the context of pre-trained models. However, current TTA methods fail to address the fundamental issue: covariate shift, i.e., the decreased generalizability can be attributed to the model's reliance on the marginal distribution of the training data, which may impair model calibration and introduce confirmation bias. To address this, we propose a novel energy-based perspective, enhancing the model's perception of target data distributions without requiring access to training data or processes. Building on this perspective, we introduce **Test-time Energy Adaptation (TEA)**, which transforms the trained classifier into an energy-based model and aligns the model's distribution with the test data's, enhancing its ability to perceive test distributions and thus improving overall generalizability. Extensive experiments across multiple tasks, benchmarks and architectures demonstrate TEA's superior generalization performance against state-of-the-art methods. Further in-depth analyses reveal that TEA can equip the model with a comprehensive perception of test distribution, ultimately paving the way toward improved generalization and calibration[1].*

## 1. Introduction

Deep neural networks, despite their remarkable performance under the assumption of independent and identically distributed (i.i.d.) training and test data [26, 33, 59], significantly degrade in real-world scenarios where unseen test data diverges from the training distribution. This limitation, known as distribution shift or domain shift, emphasizes the pressing need for generalizability across shifted test distributions [49, 57, 61, 67]. To tackle this issue, recent studies

---

*Corresponding author

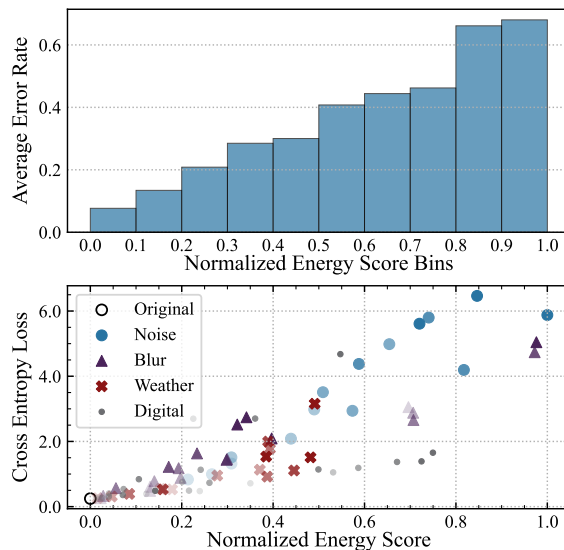[1]Code is available at https://github.com/yuanyige/tea



Figure 1. Performance vs. energy on model trained with original distribution, tested across various shifted distributions. **Upper**: error rate change within energy score groups. **Lower**: loss variation with energy scores, each point denoting a distribution. Marker styles and opacity reflect distribution types and divergence.

of Test Time Adaptation (TTA) [37, 60, 70] began considering the incorporation of unlabeled test data and leveraging it to fine-tune the source model in an unsupervised manner. This paradigm has garnered significant attention due to its ability to operate without access to training data or involvement with the training process. In the era of large open-source models [9, 58, 63], where models are publicly available but the training data and training process remain inaccessible due to privacy and resource restrictions [6, 23], TTA emerges as especially beneficial and practical.

Existing TTA methods can be broadly categorized into three classes [37]. Normalization-based methods [41, 52] adjust the BatchNorm statistics of the model with test data statistics. Entropy-based methods [45, 46, 60] fine-tune the model by minimizing the prediction entropy. Pseudo-labeling-based methods [34, 36] utilize test-time generated labels for updates. While these methods have been empiri-

cally effective, these methods all fail to address a fundamental issue: covariate shift. Specifically, the decrease in generalization ability on test data with distribution shift can be attributed to the model's reliance on the marginal distribution of the training data. However, previous TTA methods do not address this shift due to their lack of connection with marginal distributions, impairing model calibration [14, 24] and introducing confirmation bias [1].

To combat the above challenges, we propose a novel way rooted in an energy-based perspective. Within this way, energy is defined as an unnormalized probability assigned to a sample, where a lower score corresponds to a higher likelihood of that sample within a distribution [32, 54]. Proposing such a way to improve test time adaptation is twofold.

Firstly, test samples that correspond to **lower energy** within the model's distribution tend to exhibit **higher performance**. This is demonstrated by examining the energy scores of various test datasets in relation to a model trained on a specific training distribution. As depicted in Fig. 1, an increase in the divergence between the test and training distributions is accompanied by a drastic escalation in energy scores, leading to a significant degradation in performance. Secondly, the energy-based way can **address covariate shift** under TTA via directly injecting the model with a comprehensive **perception of test distribution**. Addressing covariate shift in TTA is particularly challenging, as it is neither feasible to access the training dataset to align the marginal distribution between training and testing [53], nor possible to modify the training process to mitigate the influence of marginal training distribution [48]. Under such circumstances, the energy-based way can directly manipulate the trained model's likelihood landscape [39] via an implicit distribution modeling process without requiring training process and training data, becoming a promising way. This stands in contrast to other models such as GANs [12, 19], Flows [50], and VAEs [28] which are advantageous only when the training data are accessible.

Building on the above energy-based way, we propose **T**est-time **E**nergy **A**daptation, abbreviated as **TEA**, which constructs an energy-based model from the trained (classifier) model by reinterpreting the negative log-sum-exp of logits as an energy function, and employs Contrastive Divergence [18] as the adaptation objective to decrease the energy of test samples while increase that of samples generated by Stochastic Gradient Langevin Dynamics [62]. This approach prevents a trivial solution that indiscriminately reduces the energy across the entire data space to ensure an increased likelihood for target test samples within the model's distribution. TEA enables a gradual alignment between the distributions of the trained model and the test data, bolstering the trained model's perception of the test distribution and paving the way for superior adaptability and performance when confronted with the corresponding test data.

We investigate the effectiveness of TEA under image corruption and domain generalization on four popular benchmarks CIFAR-10, CIFAR-100, TinyImageNet and PACS, across three architectures WRN-28-10, ResNet-50 and ResNet-18. Experimental results underscore that TEA significantly outperforms current best-performing TTA methods in terms of generalizability, with an average increment of 4.7%. We further reveal that TEA can equip the model with a comprehensive perception of the test distribution. This, in turn, significantly improves the generalization and calibration of the trained model.

Our main contributions include:
- **Promising Way**: We propose a new energy-based way for test time adaptation, which marks a departure from traditional methods and sheds light on potential avenues for mitigating the impact of distribution shifts.
- **Innovative Method**: We propose TEA to decrease the energy of the test data within the model's distribution, thereby equipping the trained model with a perception of the test distribution and enhancing generalizability.
- **Extensive Experiments**: Experiments across extensive settings validate TEA's superiority over current leading methods. Further in-depth analyses extend the understanding of how energy reduction enhances the model's perception of test distribution, ultimately paving the way toward improved generalization and calibration.

## 2. Related Work

**Test Time Adaptation** Test Time Adaptation (TTA) [37] is a paradigm aiming to enhance a model's generalizability on specific test data through unsupervised fine-tuning with these data. Note that the model is originally trained on a distinct training dataset, which is not available during the adaptation phase. Approaches like TTT [55] adapt models through self-supervised proxy task during testing but require the training of the same proxy task during training procedure. DDA [11, 65] explores adapting the test data, yet faces limitations due to model structure and training constraints. Recent research [60] highlights a scenario where the training process and training data is entirely agnostic, leading to three main categories of approaches: For normalization-based, BN [52] adapts the BatchNorm [22] statistics with test data. DUA [41] uses a tiny fraction of test data and its augmentation for BatchNorm statistics adaptation. For entropy-based, TENT [60] fine-tunes BatchNorm layers using entropy minimization during the test phase. EATA [45] employs a Fisher regularizer to limit excessive model parameter changes. SAR [46] removes high-gradient samples and promotes flat minimum weights. For pseudo-labeling-based, PL [34] fine-tunes parameters using confident pseudo labels. SHOT [36] combines entropy minimization methods with pseudo labeling.
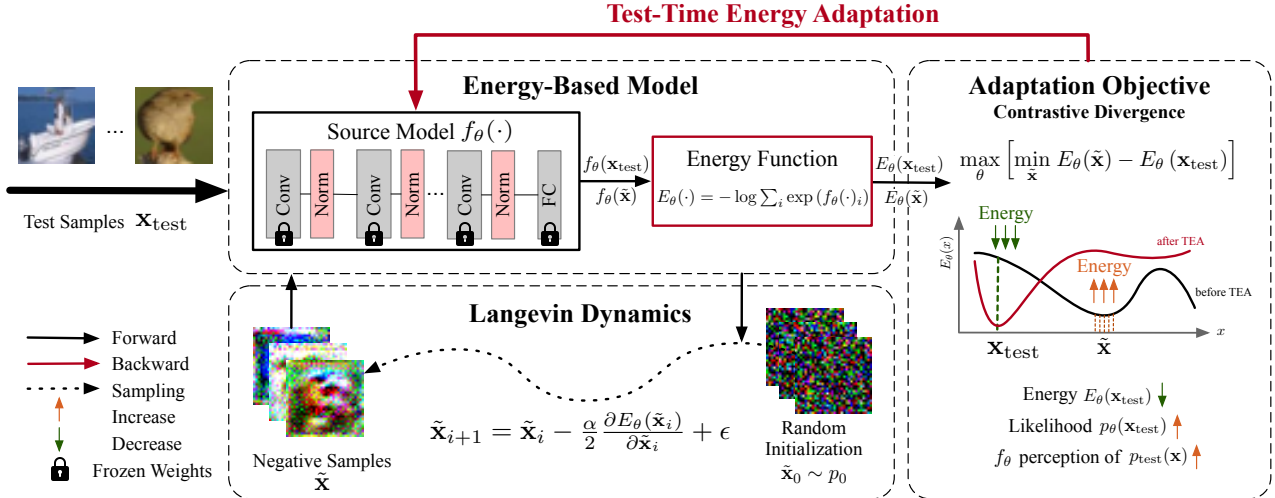
Figure 2. **Overview of Test-time Energy Adaptation (TEA)**. Given a trained model (classifier) and in-coming test data, TEA directly integrates test data distribution into the trained classifier by fine-tuning its normalization layers through energy-based training: TEA constructs an Energy-Based Model from the classifier by reinterpreting the negative log-sum-exp of logits as an energy function, and employs Contrastive Divergence as the adaptation objective which decrease the energy of test samples while increase the energy of negative samples generated by Langevin Dynamics. This adaptation increases the likelihood of test samples under the classifier's distribution, enabling a gradual alignment between the distributions of the trained classifier and the test data, thereby enhancing generalizability.

**Energy Based Model** Energy-Based Models (EBMs) are a type of non-normalized probabilistic models. Unlike most other probabilistic models, EBMs do not necessitate the normalizing constant to be tractable [32, 54] and do not require an explicit neural network for sample generation, implying the generation process is implicit [7]. These lead to increased flexibility in parameterization and allow for modeling a wider range of probability distributions. Due to their flexibility, EBMs can construct hybrid models with both discriminative and generative capabilities, integrating the generative competencies into discriminative models without sacrificing their discriminative capabilities [7, 13, 15]. Among these, JEM [13] is particularly representative, reinterpreting classifiers as an EBM and achieving impressive results in both classification and generation.

## 3. Method

In this section, we detail our method, Test-time Energy Adaptation (TEA). Initially, we present a thorough description of the notation and overall architecture in Sec. 3.1, after which we proceed to explain energy adaption (Sec. 3.2) and modulation parameters (Sec. 3.3), respectively. Furthermore, we engage in a discussion about the difference between our method and entropy-based adaptation in Sec. 3.4.

### 3.1. Notation and Overall Architecture

The labeled training dataset is denoted as $\{(\mathbf{x}_{\text{train}}, y_{\text{train}})\} \subset \mathcal{X} \times \mathcal{Y}$ and the unlabeled test data are represented by $\mathbf{x}_{\text{test}} \in \mathcal{X}$, where $\mathcal{X}$ and $\mathcal{Y}$ are data and label spaces. The respective marginal distributions of the training and test data are given

by $\mathbf{x}_{\text{train}} \sim p_{\text{train}}(\mathbf{x})$ and $\mathbf{x}_{\text{test}} \sim p_{\text{test}}(\mathbf{x})$. A classifier model trained on the training dataset and parameterized by $\theta$, is denoted as $f_\theta : \mathcal{X} \rightarrow \mathcal{Y}$. The data distribution learned by this trained classifier is denoted by $p_\theta(\mathbf{x})$, which will be referred to as the model distribution henceforth.

The overall framework of TEA is depicted in Fig. 2. The motivation behind TEA is rooted in the issue of covariate shifts [25], where the degradation of model generalization on test data $\mathbf{x}_{\text{test}}$ is attributed to the model's reliance on the training distribution $p_{\text{train}}(\mathbf{x})$. To overcome this issue without accessing the training data or training process, TEA directly integrates the test data distribution into the trained model. TEA constructs an energy-based model from the trained classifier by reinterpreting the negative log-sum-exp of logits as an energy function. Through this lens, TEA employs contrastive divergence [18] as the adaptation objective, which serves to decrease the energy (increase the likelihood) of the test samples under the model distribution $p_\theta(\mathbf{x})$. This adaptation enables the gradual alignment of distribution between the trained model and test data, thereby bolstering the model's perception of the test distribution and enhancing generalizability. Following previous TTA methods, TEA freezes the majority of the model's parameters, permitting only minor adjustments for efficient adaptation.

### 3.2. Energy Adaptation for Test Distribution

Enhancing the model's perception of test distribution from an energy-based perspective involves two key steps: constructing the energy-based model and optimizing it.

**Constructing the energy-based model.** Energy-based models (EBMs) [13, 32, 54] represent a class of proba-

bilistic models that are characterized by an energy function. Consider a sample $\mathbf{x} \in \mathbb{R}^D$, the energy function $E : \mathbb{R}^D \to \mathbb{R}$ maps each sample into an energy that can be considered as an unnormalized probability, with lower scores indicating higher likelihoods [7]. Consequently, the probability density $p(\mathbf{x})$, as defined by EBM, can be expressed using the Boltzmann distribution [38], as shown in Eq. (1), where the partition function $Z = \int \exp(-E(\mathbf{x})) \, d\mathbf{x}$ serves to normalize the probability density.

$$p(\mathbf{x}) = \exp(-E(\mathbf{x}))/Z. \qquad (1)$$

Constructing an energy-based model from a trained classifier $f_\theta$ is founded on the fundamental analysis that an energy-based framework inherently underlies any discriminative model [13]. In this framework, the energy of one sample for a corresponding class can be represented as its logit produced by the trained classifier, denoted by $E_\theta(\mathbf{x}, y) = -f_\theta(\mathbf{x})[y]$. Therefore, the joint probability distribution of $\mathbf{x}$ and $y$ can be defined as,

$$p_\theta(\mathbf{x}, y) = \exp(f_\theta(\mathbf{x})[y])/Z_\theta, \qquad (2)$$

then the distribution of $\mathbf{x}$ can be obtained by marginalizing over $y$, as shown below,

$$p_\theta(\mathbf{x}) = \sum_y p_\theta(\mathbf{x}, y) = \sum_y \exp(f_\theta(\mathbf{x})[y])/Z_\theta. \qquad (3)$$

By substituting Eq. (3) into Eq. (1), we can obtain the form of the energy function as follows:

$$E_\theta(\mathbf{x}) = -\log \sum_y \exp(f_\theta(\mathbf{x})[y]). \qquad (4)$$

Following the aforementioned steps, we repurpose and reinterpret the logits produced by the trained classifier to establish an energy-based model and define the energy function as the negative log-sum-exp of the logits.

After the logits reinterpretation, we can construct a energy-based model for the test data $\mathbf{x}_{\text{test}}$ using the trained classifier, where $Z_\theta = \int \sum_y \exp(f_\theta(\mathbf{x})[y]) \, d\mathbf{x}$.

$$p_\theta(\mathbf{x}_{\text{test}}) = \frac{\exp(-E_\theta(\mathbf{x}_{\text{test}}))}{Z_\theta} = \frac{\sum_y \exp(f_\theta(\mathbf{x}_{\text{test}})[y])}{Z_\theta}. \qquad (5)$$

**Optimizing the energy-based model.** Optimizing Eq. (5) is challenging as the partition function $Z_\theta$ necessitates integration across the whole input space of $\mathbf{x}$, typically making it computationally intractable. Thus, direct maximizing the log-likelihood of test data $\log p_\theta(\mathbf{x}_{\text{test}})$ presents significant difficulties when training the parameter $\theta$ of our energy-based model. To overcome this difficulty, we propose to use contrastive divergence [3, 18] via estimating the gradient of the log-likelihood,

$$\frac{\partial \log p_\theta(\mathbf{x}_{\text{test}})}{\partial \theta} = \mathbb{E}_{\tilde{\mathbf{x}} \sim p_\theta} \left[ \frac{\partial E_\theta(\tilde{\mathbf{x}})}{\partial \theta} \right] - \frac{\partial E_\theta(\mathbf{x}_{\text{test}})}{\partial \theta}. \qquad (6)$$

In Eq. (6), the notation $\tilde{\mathbf{x}} \sim p_\theta$ denotes a random sample drawn from the distribution over $\mathbf{x}$, which is defined by the model's distribution. The sampling procedure can be

---

**Algorithm 1:** Test-time Energy Adaptation

**Input:** $f_\theta$, $N$, $T$, $p_0$, $\alpha$, $\beta$, $\mathbf{x}_{\text{test}}$
**Output:** $f_\theta(\mathbf{x}_{\text{test}})$

1   $E_\theta(\cdot) \leftarrow -\log \sum_y \exp(f_\theta(\cdot)[y])$
2   **for** $i \leftarrow 0, 1, \ldots, N-1$ **do**
3     $\tilde{\mathbf{x}}_0 \leftarrow \text{sample}(p_0)$
4     **for** $t \leftarrow 0, 1, \ldots, T-1$ **do**
5       $\epsilon \leftarrow \text{sample}(\mathcal{N}(\mathbf{0}, \mathbf{I}))$
6       $\tilde{\mathbf{x}}_{t+1} \leftarrow \tilde{\mathbf{x}}_t - \frac{\alpha}{2} \frac{\partial E_\theta(\tilde{\mathbf{x}}_t)}{\partial \tilde{\mathbf{x}}_t} + \alpha\epsilon$
7     **end**
8     $\tilde{\mathbf{x}} \leftarrow \tilde{\mathbf{x}}_{T-1}$
9     $\theta \leftarrow \theta - \beta \nabla_\theta [E_\theta(\mathbf{x}_{\text{test}}) - E_\theta(\tilde{\mathbf{x}})]$
10 **end**
11 **return** $f_\theta(\mathbf{x}_{\text{test}})$

---

performed through Stochastic Gradient Langevin Dynamics (SGLD) [62], which iteratively generates samples by using the gradient information [13, 44]. In this context, $p_0$ represents an initial noise distribution, $\alpha$ denotes the step-size, and $t = 0, 1, \ldots, T-1$ is the time step. After $T$ steps of updating, a fictitious sample, generated by the energy-based model governed by the classifier, can be obtained.

$$\tilde{\mathbf{x}}_{t+1} = \mathbf{x}_t - \frac{\alpha}{2} \frac{\partial E_\theta(\tilde{\mathbf{x}}_t)}{\partial \tilde{\mathbf{x}}_t} + \sqrt{\alpha}\epsilon, \quad \epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I}), \qquad (7)$$

this sampling process in Eq. (7) essentially optimizes the sample by moving in the direction of energy reduction. Consequently, the objective of Eqs. (6) and (7) can be fundamentally understood as a min-max game (Eq. (8)), which minimizes the energy derived from the incoming test samples while concurrently amplifying the energy of fictitious samples obtained via SGLD from the classifier's distribution. Significantly, the latter plays a pivotal role in preventing a collapse towards a trivial solution where energy is indiscriminately minimized throughout the entire space, rather than focusing on the target test data.

$$\max_\theta \left[ \min_{\tilde{\mathbf{x}}} E_\theta(\tilde{\mathbf{x}}) - E_\theta(\mathbf{x}_{\text{test}}) \right]. \qquad (8)$$

By adapting through this objective, the classifier's distribution continually converges towards the test distribution. The convergence point of the min-max game is reached when the energy of samples drawn from the classifier's distribution matches the energy of the test samples. At this stage, the classifier reaches a low-energy steady state with respect to the test data. The likelihood landscape of the classifier then exhibits a higher probability for samples that are consistent with the test data distribution, and conversely, a lower probability for samples that deviate from it, leading to a comprehensive perception of the classifier towards this test distribution and ultimately enhancing its generalizability. The pseudocode for TEA can be found in Algorithm 1.

## 3.3. Modulation Parameters

As outlined in Eq. (8), TEA requires updating the parameters of the trained model using the aforementioned energy adaptation to adjust to test data. In line with previous methods, we opt to update the parameters of the normalization layer due to the following two factors: (1) **Practicality and Efficiency**: In the era of large-scale models, the practice of fine-tuning a selected group of parameters has gained prominence [20]. For both practicality and efficiency, it's crucial to avoid updating all parameters as this process would be excessively time-consuming. Note that the parameters of the normalization layer account for a modest 1% of total model parameters [60], making their update become more manageable. (2) **Direct impact on data distribution**: The parameters within normalization layers possess the potential to capture intrinsic features of data, thereby exerting direct influence upon the corresponding data distribution. As evidenced by [21], simple modulation to the mean and variance in a generator's normalization layers can modify image style, underlining the normalization parameters' profound impact on data distribution. This aligns well with our goal of manipulating the energy of test data to make it compatible with the model distribution.

## 3.4. Discussion

As entropy-based adaptation has been the representative adaptation method, we further discuss the difference between TEA and entropy-based adaptation. Intriguingly, TEA may have a connection with entropy-based adaptation, given that the negative entropy $\text{NegEntropy}(\mathbf{x}) = \sum_i x_i \log x_i$ is the convex conjugate of the free energy $\text{LogSumExp}(\mathbf{x}) = \log \sum_i \exp(x_i)$, as established in the literature [2, 43]. Despite this connection, entropy-based methods, which apply softmax normalization in the label space $y$ and strive to minimize entropy, can result in diminished uncertainty within classification probabilities, leading to compromised model calibration. In contrast, TEA, utilizing the log-sum-exp function within the data space $\mathbf{x}$ can not only effectively avoid the pitfalls associated with entropy-based methods, but also improve calibration by introduces uncertainty to each class. This conjecture has been substantiated through experiments (refer to Sec. 4.3.3).

## 4. Experiment

In the following sections, we compare TEA with state-of-the-art methods across various tasks, benchmarks and architectures. Then, we delve into deeper understanding of our method by exploring its desirable properties and identifying significant components that contribute to its improvements. Due to the space limitations, more comprehensive experiments including full results on corruptions and other analyses, are provided in Appendix Sec. 7.

## 4.1. Experimental Setup

**Datasets and Metrics** We focus on two tasks to verify the performance of TEA: generalization on image corruption and domain generalization. Image corruption includes clean and corrupted datasets from CIFAR-10(C), CIFAR-100(C) and TinyImageNet-200(C) [17, 29, 31], incorporating 15 types of corruption at 5 severity levels. Domain generalization considers PACS [35], encompassing 4 domains (Photo, Art Painting, Cartoon, Sketch) across 7 categories. We follow [68] to use Accuracy and mean Corruption Error (mCE) [17] as evaluation metrics. The evaluations are conducted at both the most severe level and the average of all severity levels to ensure thorough analysis.

**Backbones and Baselines** We use two architectures for image corruption: WideResNet-28-10 [69] with Batch-Norm [22], and ResNet-50 [16] with GroupNorm [64], consistent with the implementations of TENT [60] and SAR [46]. We use ResNet-18 [16] for domain generalization. We evaluate our method against eight leading TTA methods across three categories: (1) Normalization-based methods: BN [52] and DUA [41]. (2) Entropy-based methods: TENT [60], ETA, EATA [45], and SAR [46]; (3) Pseudo-labeling-based methods: PL [34] and SHOT [36]. Source denotes the original model without any adaptation.

**Implementation** We implement methods based on Py-Torch [47]. Consistency in model weight is ensured by the RobustBench protocol [5], which provides pre-trained weights for the WideResNet-28-10 (BatchNorm) on CIFAR-10. In the case where RobustBench weights are unavailable, we train models in accordance with the guidelines specified in [69]. All adaptation employ Adam [27], except for SAR, which originally uses SAM [10] with SGD [51]. Baselines are replicated using their original hyper-parameters, except when these were unspecified. More details and setups are deferred to Appendix Sec. 8.

## 4.2. Adaptation Results

This section compares the generalizability of TEA with baselines in image corruption and domain generalization.

**Image Corruption** As reported in Tab. 1, we conducted experiments on three benchmarks against eight baselines for corruption scenarios. TEA markedly surpasses all baselines on the vast majority of datasets and severity levels. Specifically, TEA outshines the best-performing baseline by an average of 4.7% at the most severe level. The only exception is on TinyImageNet across all levels, where TEA ranks second, merely trailing by a minimal margin 0.1%. We further incorporate the ResNet50 with GroupNorm for a broader validation. Tab. 2 indicates that TEA still delivers the best

Table 1. Comparisons of TEA and baselines for image corruption on CIFAR-10(C), CIFAR-100(C), and TinyImageNet(C) using WRN-28-10 with BatchNorm. Accuracy and mCE are evaluated at the most severe level and across all levels with asterisk (*) indicating the results are taken from the original paper [56]. The best adaptation results are highlighted in **boldface**.

| WRN-28-10 BatchNorm | | CIFAR-10(C) | | | | | CIFAR-100(C) | | | | | Tiny-ImageNet(C) | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Clean | Corr Severity 5 | | Corr Severity 1-5 | | Clean | Corr Severity 5 | | Corr Severity 1-5 | | Clean | Corr Severity 5 | | Corr Severity 1-5 | |
| | | Acc (↑) | Acc (↑) | mCE (↓) | Acc (↑) | mCE (↓) | Acc (↑) | Acc (↑) | mCE (↓) | Acc (↑) | mCE (↓) | Acc (↑) | Acc (↑) | mCE (↓) | Acc (↑) | mCE (↓) |
| Source | | 94.77 | 56.47 | 100.00 | 73.45 | 100.00 | 81.79 | 35.39 | 100.00 | 52.12 | 100.00 | 63.19 | 21.21 | 100.00 | 34.13 | 100.00 |
| Norm | BN [52] | 93.97 | 79.56 | 52.65 | 85.63 | 60.00 | 80.83 | 60.06 | 63.54 | 68.11 | 69.42 | 45.04 | 27.74 | 93.42 | 34.27 | 100.96 |
| | DUA* [41] | - | 80.10 | 50.78 | - | - | - | - | - | - | - | - | - | - | - | - |
| Pseudo | PL [34] | 93.75 | 51.42 | 106.98 | 72.62 | 99.37 | 80.52 | 53.40 | 72.12 | 64.53 | 75.29 | 47.84 | 28.26 | 91.22 | 39.83 | 91.67 |
| | SHOT [36] | 93.25 | 74.77 | 63.19 | 82.35 | 72.61 | 80.52 | 56.53 | 68.01 | 66.00 | 73.28 | 47.95 | 29.14 | 90.16 | **40.01** | **91.41** |
| Entropy | TENT [60] | 93.66 | 81.41 | 48.13 | 86.75 | 56.17 | 80.14 | 63.09 | 59.42 | 69.47 | 67.80 | 39.54 | 26.31 | 95.52 | 32.03 | 104.49 |
| | ETA [45] | 93.96 | 79.58 | 52.64 | 85.63 | 59.99 | 80.65 | 59.82 | 64.52 | 67.17 | 72.40 | 43.20 | 27.28 | 94.12 | 33.46 | 102.25 |
| | EATA [45] | 93.96 | 79.59 | 52.62 | 85.64 | 59.98 | 80.68 | 60.24 | 63.75 | 67.48 | 71.66 | 43.42 | 27.28 | 94.09 | 33.47 | 102.24 |
| | SAR [46] | 93.97 | 79.77 | 51.94 | 85.83 | 58.97 | 80.84 | 62.95 | 59.37 | 70.01 | 65.99 | 41.58 | 28.21 | 92.82 | 34.60 | 100.47 |
| Energy | TEA | **94.09** | **83.34** | **43.69** | **87.88** | **52.00** | **80.88** | **65.10** | **56.07** | **71.22** | **63.72** | **51.65** | **31.67** | **87.99** | 39.96 | 92.12 |

Table 2. Comparisons for image corruption on CIFAR-10(C), CIFAR-100(C), and TinyImageNet(C) using ResNet-50 with GroupNorm across all severity levels. Best results are in **boldface**.

| ResNet50 GroupNorm | | CIFAR-10(C) | | CIFAR-100(C) | | Tiny-ImageNet(C) | |
|---|---|---|---|---|---|---|---|
| | | Acc (↑) | mCE (↓) | Acc (↑) | mCE (↓) | Acc (↑) | mCE (↓) |
| Source | | 78.71 | 100.00 | 54.98 | 100.00 | 26.64 | 100.00 |
| Pseudo | PL | 79.43 | 94.76 | 56.68 | 96.02 | 26.60 | 99.92 |
| | SHOT | 81.98 | 86.65 | 58.31 | 93.45 | 29.11 | 96.73 |
| Entropy | TENT | 77.29 | 102.88 | 56.34 | 96.88 | 26.65 | 99.94 |
| | ETA | 78.68 | 100.09 | 56.72 | 96.37 | 29.25 | 96.42 |
| | EATA | 78.70 | 100.02 | 56.76 | 96.28 | 29.25 | 96.42 |
| | SAR | 78.78 | 99.65 | 55.28 | 99.33 | 27.05 | 99.41 |
| Energy | TEA | **83.05** | **79.09** | **59.67** | **89.32** | **30.41** | **94.81** |

Table 3. Single source domain generalization comparisons on PACS datasets using ResNet-18 with BatchNorm in terms of Accuracy. The best adaptation results are highlighted in **boldface**.

| Source Domain | Method | Target Domain | | | | Avg |
|---|---|---|---|---|---|---|
| | | Photo | Art | Cartoon | Sketch | |
| Photo | Source | - | 26.76 | 22.40 | 16.62 | 21.93 |
| | BN | - | 26.66 | 27.94 | 15.96 | 23.52 |
| | TENT | - | 26.95 | 29.86 | 17.54 | 24.78 |
| | EATA | - | 26.66 | 28.11 | 15.98 | 23.59 |
| | SAR | - | 26.71 | 28.41 | 15.98 | 23.70 |
| | SHOT | - | 26.61 | 29.86 | 20.92 | 25.80 |
| | TEA | - | **28.81** | **33.62** | **20.49** | **27.64** |
| Art | Source | 49.04 | - | 36.43 | 24.48 | 36.65 |
| | BN | 46.65 | - | 28.28 | 22.73 | 32.55 |
| | TENT | 50.78 | - | 30.12 | 24.61 | 35.17 |
| | EATA | 46.83 | - | 29.31 | 23.42 | 33.19 |
| | SAR | 47.90 | - | 33.02 | 26.27 | 35.73 |
| | SHOT | 50.24 | - | 34.30 | **29.37** | 37.97 |
| | TEA | **56.29** | - | **38.57** | 28.71 | **41.19** |
| Cartoon | Source | 42.69 | 29.79 | - | 29.47 | 33.98 |
| | BN | 28.68 | 25.15 | - | 20.87 | 24.90 |
| | TENT | 30.96 | 23.34 | - | 22.65 | 25.65 |
| | EATA | 28.80 | 25.10 | - | **25.04** | 26.31 |
| | SAR | 29.70 | 25.78 | - | 21.51 | 25.66 |
| | SHOT | **37.72** | 22.66 | - | 23.14 | 27.84 |
| | TEA | 36.05 | **31.44** | - | 22.88 | **30.12** |
| Sketch | Source | 19.94 | 18.70 | 32.21 | - | 23.62 |
| | BN | 13.47 | 17.14 | 29.86 | - | 20.16 |
| | TENT | 13.53 | 17.38 | 29.52 | - | 20.14 |
| | EATA | 13.17 | 17.33 | 30.08 | - | 20.19 |
| | SAR | 13.29 | 18.80 | 29.95 | - | 20.68 |
| | SHOT | **19.76** | 18.75 | 30.46 | - | 22.99 |
| | TEA | 19.64 | **21.24** | **33.19** | - | **24.69** |

performance in both average accuracy and mCE. These results underscore the efficacy of TEA in handling image corruption scenarios, ensuring its universality for model architecture or normalization techniques.

**Domain Generalization** Tab. 3 provides a comparison between TEA and state-of-the-art TTA methods on the PACS dataset. It is evident that when trained on the photo and art domains, TEA exhibits substantial improvements over the best-performing baseline, achieving increases of 7.1% and 8.5%, respectively. Compared to photo and art, cartoon and sketch domains may exhibit greater domain discrepancies, posing significant challenges to model generalization. Despite these challenging conditions, in the adaptation from cartoon to art and the sketch domain, TEA achieved improvements when all baselines experienced significant declines. TENT shows decline of 14.73% compared to the source in sketch, whereas TEA bucks the trend and enhances performance by 4.5%, highlighting its stability in the face of severe domain shifts.

The results from both tasks highlight TEA's superior generalizability. These effectiveness may originate from the reduced energy, the enhanced distribution perception, and the improved calibration. In next section, we will delve into

these aspects for further analysis and discussion.

## 4.3. Analysis and Discussion

In this section, we delve into the mechanisms driving TEA's effectiveness and explore its desirable properties. Specifically, we studied three key aspects of TEA: (1) The correlation between energy reduction and generalizability enhancement; (2) The distribution perception and generation capabilities; (3) The confidence calibration improvements.
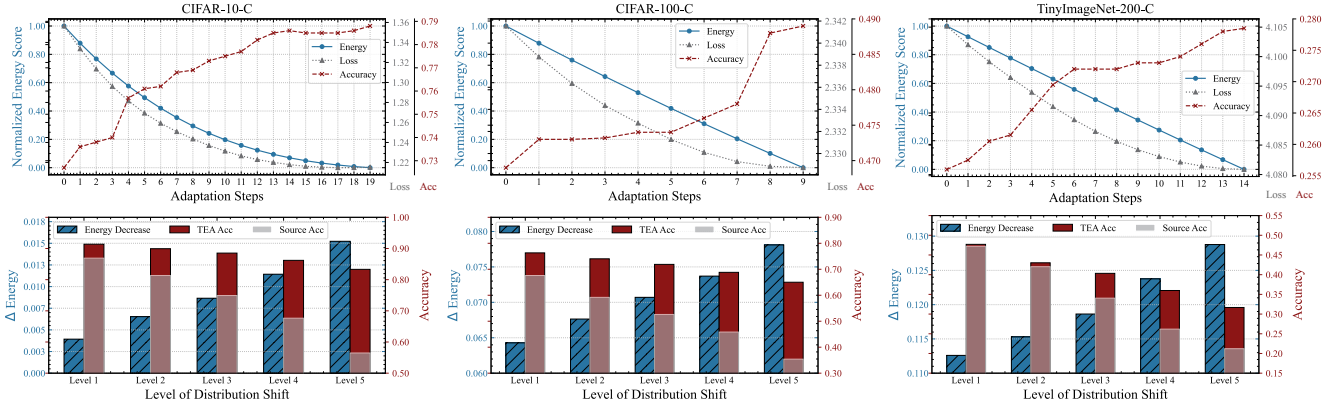
Figure 3. This illustration captures the energy reduction and generalizability enhancement achieved by TEA across CIFAR-10-C, CIFAR-100-C, and TinyImageNet-200-C, displayed from left to right. The **upper** set of graphs trace the evolution of energy score, corresponding loss and accuracy in response to incrementally increasing TEA adaptation steps. The **lower** set uncovers the extent of energy reduction and the consequent performance improvement before and after executing TEA adaptation, under different levels of distribution shift.

### 4.3.1 Relation between TEA's Energy Reduction and Generalizability Enhancement

This experiment validate TEA's energy reduction capability, and its impact on generalizability, spanning two scenarios: (1) In the first scenario, we focus on the trends of energy scores, loss, and accuracy on the same test data over increasing adaptation steps. (2) In the second scenario, we explore how the extent of energy reduction correlates with performance improvements, before and after adaptation, using varied test data that exhibit different distribution shifts.

The results from both scenarios are depicted in Fig. 3. It is observable that $(i)$ As the iteration step of TEA increases, there is a consistent reduction in energy and corresponding decrease in loss, coupled with an ongoing enhancement in accuracy. $(ii)$ As the distribution shift increases, TEA's energy reduction becomes more pronounced. Concurrently, the enhancement in performance over the baseline also increases. $(iii)$ As the distribution shift increases, there is a sharp degradation observed in the baseline performance. However, the model adapted via TEA maintains its stability and robustness, demonstrating resilience against strong distribution shifts. In summary, these aforementioned trends are consistently observed across three datasets, demonstrating TEA's significant effectiveness in reducing energy. Notably, a greater reduction in energy correlates with an increased improvement in the model's generalizability

### 4.3.2 TEA's Distribution Perception and Generation

This experiment aims to validate the capability of TEA in modeling test data distribution. We framed the experiment within two scenarios: identical training and testing distributions, and shifted training and testing distributions. (1) The first scenario include four settings: $(i)$ Source model on MNIST training set, TEA adaptation on MNIST test set;
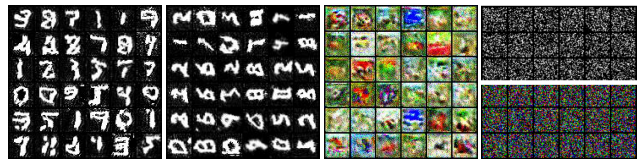


Figure 4. Test distribution perception visualization for identical training and testing distributions on MNIST and CIFAR-10.
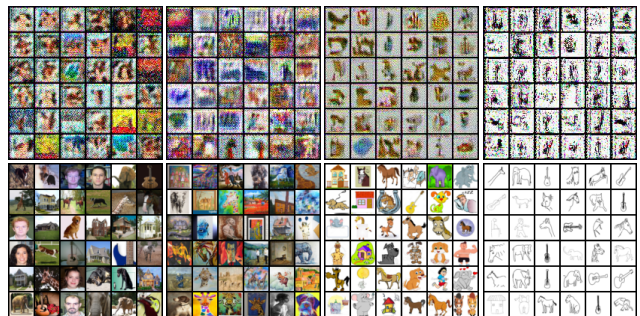


Figure 5. Test distribution perception visualization (**upper**) and real samples (**lower**) on shifted distribution: A model trained on PACS-A dataset then individually tested with TEA adaptation across PACS-P, PACS-A, PACS-C, PACS-S datasets.

$(ii)$ Source model on MNIST training set, TEA adaptation on 90-degree rotated MNIST test set; $(iii)$ Source model on CIFAR-10 training set, TEA adaptation on CIFAR-10 test set; $(iv)$ Source models on MNIST or CIFAR-10 training sets, tested without TEA adaptation. (2) In the second scenario, source model was trained on the PACS-A dataset and individually tested with TEA adaptation on PACS-P, PACS-A, PACS-C and PACS-S datasets.

The outcomes from both scenarios are respectively illustrated in Fig. 4 and Fig. 5. Fig. 4 (1,3) indicates that, in scenarios where the distributions are closely identical, TEA has the potential to reconstruct samples that maintain discernible patterns. Fig. 4 (2) reveals that the sample distri-

bution indeed originates from the test data rather than provoking recollections from the training datasets. Fig. 4 (4) confirms that models without adaptation, or those adapted using other methods, fail to effectively characterize the testing distributions on both the simple MNIST dataset (upper) and the more complex CIFAR-10 dataset (lower). Drawing conclusions from Fig. 5, we can infer that under significant distribution/domain shift, our method can still characterize the key features of the shifted test distribution, such as style, texture and color schemes.

In summary, TEA endows the model with generative ability for test data via energy-based fine-tuning of the normalization layers. This ability may incorporates generative self-supervised information from the test data into the model and improves the model's thorough understanding of the test distribution, which in turn strengthens its generalization performance on that distribution.

### 4.3.3   TEA's Improvements in Confidence Calibration

This experiment compares model calibration across the source model, the entropy-based method TENT, the pseudo-labeling-based method SHOT, and our energy-based TEA, on CIFAR-10. In accordance with the protocol in [4, 14], we illustrate reliability histogram and compute two scalar summary statistics: Expected Calibration Error (ECE) and Maximum Calibration Error (MCE) [14], to evaluate calibration implemented by torchmetrics. The procedures are implemented as follows: For the reliability histogram, we divided the model's predictions into ten bins based on the confidence score of the highest probability class and calculated the average accuracy for each bin.

The results are depicted in Fig. 6. From the perspective of the histogram, an optimally calibrated model should have its bar graph in a diagonal shape to achieve the smallest gap area. However, the bars for TENT and SHOT are observed to fall significantly below this line, manifesting an even inferior performance compared to the source model without any adaptation. These phenomena provide evidence that both the entropy-based methods and the pseudo-label-based method could potentially harm the confidence calibration by inducing overconfidence in their predictions. In stark contrast, our TEA has significantly narrowed the gap area and improved alignment with the diagonal line. For quantitative metrics, ECE and MCE, TEA has improved by 2.43% and 18.31% respectively compared to the source model.

The improvement in calibration of TEA over competitors may come from that neural networks inherently tend to be overconfident [14]. The softmax function enforces exponential normalization among classes, which tends to amplify the probability of dominant classes, thus inherently not advantageous for calibration. TENT and SHOT exacerbate this dominance of certain classes by reducing the un-
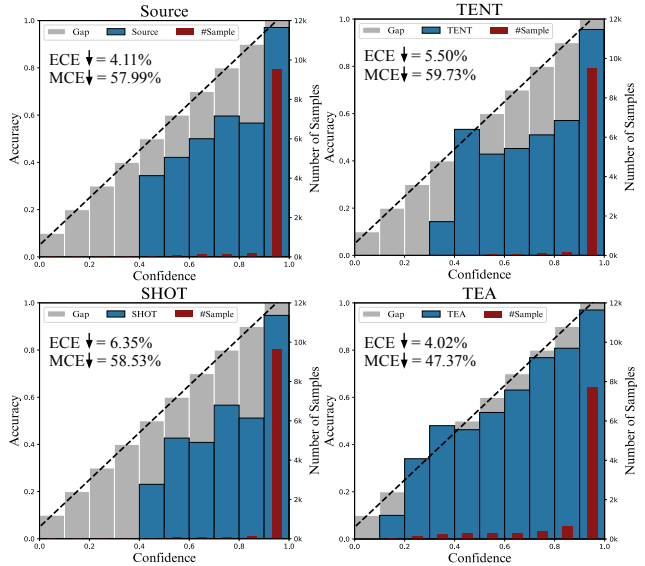


Figure 6. Calibration comparison between TEA and baselines on CIFAR-10 dataset. In an ideal scenario for optimal calibration, blue bars should align with the diagonal line, and a smaller grey gap area is preferred. Quantitative measures are provided via ECE and MCE metrics, where lower values indicate better calibration.

certainty of class probabilities, further amplifying the overconfidence of the classifier. On the contrary, TEA dose not perform normalization in the label space, but maximizes the log-sum-exp of classifier logits, which essentially introduces a certain level of uncertainty to each class and empowers TEA with the ability to enhance calibration.

## 5. Conclusion

To achieve test time adaptation, we introduce an innovative energy-based perspective to mitigate the impact derived from distribution shifts. The proposed TEA aims to decrease the energy of the test data within the pretrained model's distribution. TEA guides the model towards achieving a harmonious low-energy equilibrium state for the test data, which mitigates the model's distribution discrepancy and boosts its generalizability towards test distributions. Comprehensive experiments across multiple tasks, benchmarks, and architectures confirm TEA's superiority over current leading methods. Further in-depth analyses of TEA's underlying mechanisms deepen our understanding of how energy reduction can enhance the model's perception of the test distribution, ultimately paving the way for improved generalization and calibration.

## Acknowledgments

# References

[1] Eric Arazo, Diego Ortego, Paul Albert, Noel E. O'Connor, and Kevin McGuinness. Pseudo-labeling and confirmation bias in deep semi-supervised learning, 2020. 2

[2] Stephen P Boyd and Lieven Vandenberghe. Convex optimization. Cambridge university press, 2004. 5

[3] Miguel A Carreira-Perpinan and Geoffrey Hinton. On contrastive divergence learning. In International workshop on artificial intelligence and statistics, pages 33–40. PMLR, 2005. 4

[4] Dian Chen, Dequan Wang, Trevor Darrell, and Sayna Ebrahimi. Contrastive test-time adaptation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 295–305, 2022. 8

[5] Francesco Croce, Maksym Andriushchenko, Vikash Sehwag, Edoardo Debenedetti, Nicolas Flammarion, Mung Chiang, Prateek Mittal, and Matthias Hein. Robustbench: a standardized adversarial robustness benchmark. In Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track, 2021. 5

[6] Sabrina De Capitani Di Vimercati, Sara Foresti, Giovanni Livraga, and Pierangela Samarati. Data privacy: Definitions and techniques. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, 20(06):793–817, 2012. 1

[7] Yilun Du and Igor Mordatch. Implicit generation and generalization in energy-based models. arXiv preprint arXiv:1903.08689, 2019. 3, 4

[8] Yilun Du, Shuang Li, Joshua Tenenbaum, and Igor Mordatch. Improved contrastive divergence training of energy-based models. In Proceedings of the 38th International Conference on Machine Learning, pages 2837–2848. PMLR, 2021. 2

[9] Zhengxiao Du, Yujie Qian, Xiao Liu, Ming Ding, Jiezhong Qiu, Zhilin Yang, and Jie Tang. Glm: General language model pretraining with autoregressive blank infilling. arXiv preprint arXiv:2103.10360, 2021. 1

[10] Pierre Foret, Ariel Kleiner, Hossein Mobahi, and Behnam Neyshabur. Sharpness-aware minimization for efficiently improving generalization. In International Conference on Learning Representations, 2021. 5

[11] Jin Gao, Jialing Zhang, Xihui Liu, Trevor Darrell, Evan Shelhamer, and Dequan Wang. Back to the source: Diffusion-driven adaptation to test-time corruption. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pages 11786–11796, 2023. 2

[12] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. Communications of the ACM, 63(11):139–144, 2020. 2

[13] Will Grathwohl, Kuan-Chieh Wang, Joern-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 International Conference on Learning Representations, 2020. 3, 4, 2

[14] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In International conference on machine learning, pages 1321–1330. PMLR, 2017. 2, 8

[15] Tian Han, Erik Nijkamp, Xiaolin Fang, Mitch Hill, Song-Chun Zhu, and Ying Nian Wu. Divergence triangle for joint training of generator model, energy-based model, and inferential model. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 8670–8679, 2019. 3

[16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 770–778, 2016. 5

[17] Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. In International Conference on Learning Representations, 2019. 5, 1, 2

[18] Geoffrey E Hinton. Training products of experts by minimizing contrastive divergence. Neural computation, 14(8):1771–1800, 2002. 2, 3, 4

[19] Liang Hou, Qi Cao, Yige Yuan, Songtao Zhao, Chongyang Ma, Siyuan Pan, Pengfei Wan, Zhongyuan Wang, Huawei Shen, and Xueqi Cheng. Augmentation-aware self-supervision for data-efficient gan training. arXiv preprint arXiv:2205.15677, 2022. 2

[20] Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In International Conference on Learning Representations, 2022. 5

[21] Xun Huang and Serge Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In Proceedings of the IEEE international conference on computer vision, pages 1501–1510, 2017. 5

[22] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In International conference on machine learning, pages 448–456. pmlr, 2015. 2, 5

[23] Priyank Jain, Manasi Gyanchandani, and Nilay Khare. Big data privacy: a technological perspective and review. Journal of Big Data, 3:1–25, 2016. 1

[24] Minguk Jang, Sae-Young Chung, and Hye Won Chung. Test-time adaptation via self-training with nearest neighbor information. In The Eleventh International Conference on Learning Representations, 2023. 2

[25] Jing Jiang. A literature survey on domain adaptation of statistical classifiers. URL: http://sifaka. cs. uiuc. edu/jiang4/domainadaptation/survey, 3(1-12):3, 2008. 3

[26] Michael I Jordan and Tom M Mitchell. Machine learning: Trends, perspectives, and prospects. Science, 349(6245):255–260, 2015. 1

[27] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings, 2015. 5

[28] Diederik P. Kingma and Max Welling. Auto-encoding variational bayes. In 2nd International Conference on Learning

Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings, 2014. 2

[29] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009. 5, 1

[30] Jogendra Nath Kundu, Akshay R Kulkarni, Suvaansh Bhambri, Deepesh Mehta, Shreyas Anand Kulkarni, Varun Jampani, and Venkatesh Babu Radhakrishnan. Balancing discriminability and transferability for source-free domain adaptation. In International Conference on Machine Learning, pages 11710–11728. PMLR, 2022. 2

[31] Ya Le and Xuan Yang. Tiny imagenet visual recognition challenge. CS 231N, 7(7):3, 2015. 5, 1

[32] Yann LeCun, Sumit Chopra, Raia Hadsell, M Ranzato, and Fujie Huang. A tutorial on energy-based learning. Predicting structured data, 1(0), 2006. 2, 3

[33] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. nature, 521(7553):436–444, 2015. 1

[34] Dong-Hyun Lee et al. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In Workshop on challenges in representation learning, ICML, page 896. Atlanta, 2013. 1, 2, 5, 6

[35] Da Li, Yongxin Yang, Yi-Zhe Song, and Timothy M Hospedales. Deeper, broader and artier domain generalization. In Proceedings of the IEEE international conference on computer vision, pages 5542–5550, 2017. 5, 2

[36] Jian Liang, Dapeng Hu, and Jiashi Feng. Do we really need to access the source data? source hypothesis transfer for unsupervised domain adaptation. In International conference on machine learning, pages 6028–6039. PMLR, 2020. 1, 2, 5, 6

[37] Jian Liang, Ran He, and Tieniu Tan. A comprehensive survey on test-time adaptation under distribution shifts. arXiv preprint arXiv:2303.15361, 2023. 1, 2

[38] EM Lifshitz and Lev Davidovich Landau. Statistical physics, course of theoretical physics. In Part 2: Theory of the Condensed State. Butterworth-Heinemann Pergamon, London, 1980. 4

[39] Fu Lin, Rohit Mittapalli, Prithvijit Chattopadhyay, Daniel Bolya, and Judy Hoffman. Likelihood landscapes: A unifying principle behind many adversarial defenses. In Computer Vision–ECCV 2020 Workshops: Glasgow, UK, August 23–28, 2020, Proceedings, Part I 16, pages 39–54. Springer, 2020. 2

[40] Weijian Luo, Hao Jiang, Tianyang Hu, Jiacheng Sun, Zhenguo Li, and Zhihua Zhang. Training energy-based models with diffusion contrastive divergences. arXiv preprint arXiv:2307.01668, 2023. 2

[41] M Jehanzeb Mirza, Jakub Micorek, Horst Possegger, and Horst Bischof. The norm must go on: Dynamic unsupervised domain adaptation by normalization. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 14765–14775, 2022. 1, 2, 5, 6

[42] Weili Nie, Arash Vahdat, and Anima Anandkumar. Controllable and compositional generation with latent-space energy-based models. Advances in Neural Information Processing Systems, 34:13497–13510, 2021. 2

[43] Frank Nielsen and Ke Sun. Guaranteed bounds on information-theoretic measures of univariate mixtures using piecewise log-sum-exp inequalities. Entropy, 18(12):442, 2016. 5

[44] Erik Nijkamp, Mitch Hill, Song-Chun Zhu, and Ying Nian Wu. Learning non-convergent non-persistent short-run mcmc toward energy-based model. Advances in Neural Information Processing Systems, 32, 2019. 4

[45] Shuaicheng Niu, Jiaxiang Wu, Yifan Zhang, Yaofo Chen, Shijian Zheng, Peilin Zhao, and Mingkui Tan. Efficient test-time model adaptation without forgetting. In International conference on machine learning, pages 16888–16905. PMLR, 2022. 1, 2, 5, 6

[46] Shuaicheng Niu, Jiaxiang Wu, Yifan Zhang, Zhiquan Wen, Yaofo Chen, Peilin Zhao, and Mingkui Tan. Towards stable test-time adaptation in dynamic wild world. In The Eleventh International Conference on Learning Representations, 2023. 1, 2, 5, 6

[47] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In Advances in Neural Information Processing Systems 32, pages 8024–8035. Curran Associates, Inc., 2019. 5

[48] Francesco Pinto, Harry Yang, Ser Nam Lim, Philip Torr, and Puneet Dokania. Using mixup as a regularizer can surprisingly improve accuracy &amp; out-of-distribution robustness. In Advances in Neural Information Processing Systems, pages 14608–14622. Curran Associates, Inc., 2022. 2

[49] Joaquin Quinonero-Candela, Masashi Sugiyama, Anton Schwaighofer, and Neil D Lawrence. Dataset shift in machine learning. Mit Press, 2008. 1

[50] Danilo Rezende and Shakir Mohamed. Variational inference with normalizing flows. In International conference on machine learning, pages 1530–1538. PMLR, 2015. 2

[51] Herbert Robbins and Sutton Monro. A stochastic approximation method. The annals of mathematical statistics, pages 400–407, 1951. 5

[52] Steffen Schneider, Evgenia Rusak, Luisa Eck, Oliver Bringmann, Wieland Brendel, and Matthias Bethge. Improving robustness against common corruptions by covariate shift adaptation. Advances in neural information processing systems, 33:11539–11551, 2020. 1, 2, 5, 6

[53] Hidetoshi Shimodaira. Improving predictive inference under covariate shift by weighting the log-likelihood function. Journal of statistical planning and inference, 90(2):227–244, 2000. 2

[54] Yang Song and Diederik P Kingma. How to train your energy-based models. arXiv preprint arXiv:2101.03288, 2021. 2, 3

[55] 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 Proceedings of the 37th International Conference on Machine Learning, pages 9229–9248. PMLR, 2020. 2

[56] Yushun Tang, Ce Zhang, Heng Xu, Shuoshuo Chen, Jie Cheng, Luziwei Leng, Qinghai Guo, and Zhihai He. Neuro-modulated hebbian learning for fully test-time adaptation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 3728–3738, 2023. 6

[57] Rohan Taori, Achal Dave, Vaishaal Shankar, Nicholas Carlini, Benjamin Recht, and Ludwig Schmidt. Measuring robustness to natural distribution shifts in image classification. Advances in Neural Information Processing Systems, 33: 18583–18599, 2020. 1

[58] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. arXiv preprint arXiv:2302.13971, 2023. 1

[59] Vladimir Vapnik. The nature of statistical learning theory. Springer science & business media, 1999. 1

[60] Dequan Wang, Evan Shelhamer, Shaoteng Liu, Bruno Olshausen, and Trevor Darrell. Tent: Fully test-time adaptation by entropy minimization. In International Conference on Learning Representations, 2021. 1, 2, 5, 6

[61] Jindong Wang, Cuiling Lan, Chang Liu, Yidong Ouyang, Tao Qin, Wang Lu, Yiqiang Chen, Wenjun Zeng, and Philip Yu. Generalizing to unseen domains: A survey on domain generalization. IEEE Transactions on Knowledge and Data Engineering, 2022. 1

[62] Max Welling and Yee W Teh. Bayesian learning via stochastic gradient langevin dynamics. In Proceedings of the 28th international conference on machine learning (ICML-11), pages 681–688, 2011. 2, 4

[63] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. Huggingface's transformers: State-of-the-art natural language processing. arXiv preprint arXiv:1910.03771, 2019. 1

[64] Yuxin Wu and Kaiming He. Group normalization. In Proceedings of the European conference on computer vision (ECCV), pages 3–19, 2018. 5

[65] Zehao Xiao, Xiantong Zhen, Shengcai Liao, and Cees G. M. Snoek. Energy-based test sample adaptation for domain generalization. In The Eleventh International Conference on Learning Representations, 2023. 2

[66] Xiulong Yang and Shihao Ji. Jem++: Improved techniques for training jem. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 6494–6503, 2021. 2

[67] Yige Yuan, Bingbing Xu, Huawei Shen, Qi Cao, Keting Cen, Wen Zheng, and Xueqi Cheng. Towards generalizable graph contrastive learning: An information theory perspective. arXiv preprint arXiv:2211.10929, 2022. 1

[68] Yige Yuan, Bingbing Xu, Bo Lin, Liang Hou, Fei Sun, Huawei Shen, and Xueqi Cheng. Pde+: Enhancing generalization via pde with adaptive distributional diffusion. arXiv preprint arXiv:2305.15835, 2023. 5

[69] Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. In British Machine Vision Conference 2016. British Machine Vision Association, 2016. 5

[70] Hao Zhao, Yuejiang Liu, Alexandre Alahi, and Tao Lin. On pitfalls of test-time adaptation. In ICLR 2023 Workshop on Pitfalls of limited data and computation for Trustworthy ML, 2023. 1