Multi-Hypothesis Test-Time Adaptation to Mitigate Underspecification

Afshar Shamsi[1], Xiao-Yu Guo[2], Hamid Alinejad-Rokny[3]
Arash Mohammadi[1], Damien Teney[4], and Ehsan Abbasnejad[5]

[1] Concordia University, Canada
{afshar.shamsi,arash.mohammadi}@concordia.ca
[2] Australian Institute for Machine Learning, Australia
xiaoyu.guo@gmail.com
[3] University of New South Wales, Australia
h.alinejad@unsw.edu.au
[4] Idiap Research Institute, Switzerland
damien.teney@idiap.ch
[5] Monash University, Australia
Ehsan.Abbasnejad@monash.edu

Abstract. Test-Time Adaptation (TTA) seeks to improve model robustness under distribution shifts by adapting parameters using unlabeled target data. However, in the absence of supervision, entropy-based adaptation is fundamentally underconstrained: multiple distinct parameter updates can achieve similarly low entropy while inducing drastically different decision boundaries. This phenomenon, known as underspecification, renders standard TTA brittle and prone to collapse into spurious modes. In this work, we reinterpret TTA through a posterior-inspired lens induced by entropy minimization, where low-entropy solutions define a pseudo-likelihood over parameters. Instead of committing to a single point estimate, we introduce a particle-based diversification framework that explores multiple plausible adaptation trajectories simultaneously. Our method can be viewed as a structured exploration of multiple plausible adaptation solutions, implemented through multi-level diversification at the output, parameter, optimizer, and input levels. Crucially, the framework acts as a plug-and-play wrapper compatible with existing TTA methods. Extensive experiments on challenging benchmarks demonstrate consistent gains in stability and robustness, achieving improvements of 3–4% under mixed shifts, 2–3% with batch size one, and 1–2.5% under label shifts, outperforming state-of-the-art baselines. Our results suggest that treating TTA as a multi-hypothesis inference problem, rather than a single-point optimization task, is key to mitigating underspecification and enabling reliable real-world deployment.

Keywords: Test-Time Adaptation · Underspecification · Model Diversification

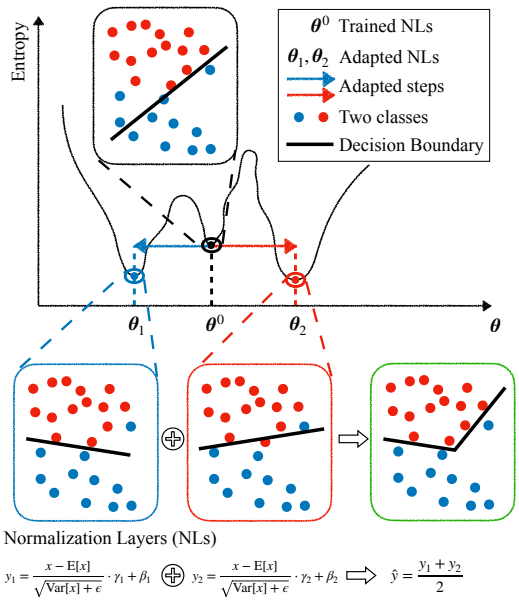


Fig. 1: Conceptual illustration of multi-hypothesis test-time adaptation. Entropy minimization can yield multiple low-entropy solutions in the adaptation landscape. Standard TTA follows a single trajectory from θ^0, which may converge to a suboptimal decision boundary. Our framework instead maintains multiple adaptation particles (θ_1, θ_2) by adapting separate normalization parameters. Aggregating their predictions produces a more stable adaptation and mitigates underspecification.

1 Introduction

Deep learning models have achieved remarkable performance across a range of tasks when training and test data are drawn from the same distribution. However, their performance often degrades sharply when deployed in the wild, where distribution shifts are inevitable [14, 16]. Test-time adaptation (TTA) has emerged as a promising strategy to bridge this train–test gap by adapting the model to the target distribution using only unlabeled test data [2, 11, 18–22].

Despite its appeal, standard TTA optimizes a single entropy-based objective in the absence of ground-truth supervision. This renders the adaptation process fundamentally underconstrained: multiple parameter configurations can achieve similarly low entropy on the target data while corresponding to qualitatively different decision boundaries. This phenomenon, underspecification, has been shown to induce instability and spurious feature reliance in deep models. From an optimization perspective, entropy minimization may admit multiple local minima of comparable value, yet these solutions need not generalize equally well under shift. We argue that existing TTA methods implicitly compute a single maximum a posteriori (MAP) solution under an entropy-induced pseudo-

likelihood. In underspecified regimes, where multiple low-entropy configurations exist, such point-estimate adaptation is inherently unstable: small perturbations in optimization trajectory can lead to qualitatively different decision boundaries. Consequently, entropy minimization under shift should be viewed not as a well-posed optimization problem, but as a multi-hypothesis inference problem.

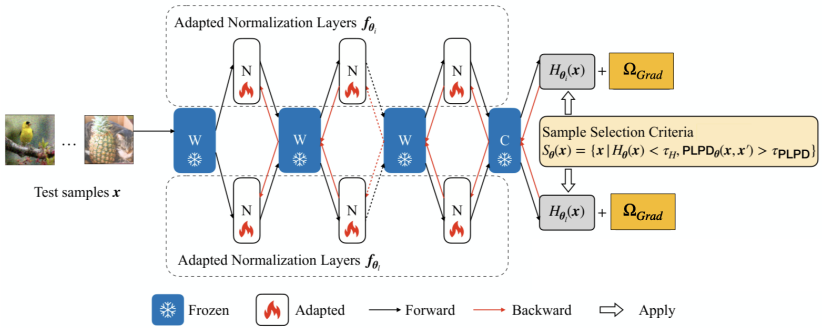


Fig. 2: An illustration of the proposed method for the case of gradient diversity diversification. During the adaptation, we only update the normalization layers (N) and keep the rest of layers (W and C) frozen. When a batch of test samples come, we first identify non-harmful ones (see Appendix). We only perform backward propagation on selected samples with gradient diversity measure to push normalization layers away from one another.

In this work, we reformulate TTA as a population-based inference problem and introduce a particle-driven diversification framework to approximate the posterior over adaptation parameters induced by entropy minimization. Rather than committing to a single entropy-minimizing trajectory, we maintain a population of interacting adaptation particles that explore distinct low-entropy regions of the parameter space while being regularized to prevent premature mode collapse. Each particle follows a unique optimization trajectory under entropy minimization, guided by multi-level diversification mechanisms that operate at the output, parameter, optimizer, and input levels. This perspective transforms TTA from a brittle point-estimation procedure into a controlled multi-trajectory exploration process. By preserving multiple plausible adaptation hypotheses and aggregating their predictions, our framework mitigates instability arising from underspecification and reduces reliance on spurious shortcuts. Importantly, the proposed approach is model-agnostic and acts as a plug-and-play wrapper that can enhance a broad class of existing TTA methods. We evaluate our framework on challenging benchmarks such as ImageNet-C [6] under mixed corruptions, label shifts, and batch size one settings. Across all scenarios, our method consistently improves robustness and stability, outperforming state-of-the-art TTA baselines by 1–4%. These results suggest that treating TTA as a multi-hypothesis inference problem is crucial for reliable deployment under distribution shift.

In Figure 1, we illustrate the entropy landscape with respect to the model's parameters $\boldsymbol{\theta}$ for a binary classification task. Suppose that $\boldsymbol{\theta}^0$ represents all parameters of the normalization layers in the source model. Without adaptation, the decision boundary is likely to be inadequate under Out-of-Distribtion (OOD) conditions. At test time, TTA approaches typically update the model by minimizing entropy to align target data with source data. Prior TTA approaches update $\boldsymbol{\theta}^0$ to either $\boldsymbol{\theta}_1$ or $\boldsymbol{\theta}_2$, optimizing the model toward low-entropy directions. Although these methods can improve overall performance, their decision boundaries remain suboptimal and can introduce additional errors. These issues are closely associated with the broader challenge of underspecification, which has been mitigated in other contexts using strategies such as data augmentation and ensembling. However, their integration into TTA remains largely unexplored. Although [8] indirectly used augmentation-based filtering to highlight robust features, this was not aimed at resolving underspecification itself. In contrast, our work directly tackles the root of underspecification by introducing a diversification-based adaptation framework that encourages representational diversity and enhances model robustness during TTA.

In this paper, we build on these observations to explain why existing TTA methods can be unreliable and to develop a more stable alternative. Specifically, starting from a source model with normalization parameters $\boldsymbol{\theta}^0$, we 1) initialize a collection of normalization parameter sets $\Theta = \{\boldsymbol{\theta}_i\}_{i=1}^N$ from $\boldsymbol{\theta}^0$ while keeping all non-normalization layers frozen (see Figure 2), 2) use N optimizers with distinct hyperparameters to adapt different particles $\boldsymbol{\theta}_i$, and 3) apply diversification regularization to prevent particle collapse. At inference time, we aggregate predictions from the N adapted parameter sets $\{\boldsymbol{\theta}_i\}_{i=1}^N$, thereby combining multiple plausible adaptation hypotheses instead of committing to a single adapted solution. Each $\boldsymbol{\theta}_i$ follows a distinct entropy-minimization trajectory, and their interaction encourages exploration of different low-entropy regions. Aggregating these interacting particles reduces sensitivity to spurious modes induced by underspecification. Empirically, this population-based inference consistently improves stability and robustness over single-model entropy minimization, outperforming representative TTA baselines including Tent [19], SAR [12], and DeYO [8].

Summary of Contributions:

- We reinterpret entropy-based test-time adaptation as an implicit posterior inference problem under underspecification, highlighting the limitations of single-point MAP adaptation.
- We propose a particle-based diversification framework that approximates posterior exploration over adaptation parameters through multi-level diversification (output, parameter, optimizer, and input levels).
- We demonstrate that controlled particle interactions mitigate collapse into spurious modes and consistently improve robustness across mixed shifts, label shifts, and batch size one scenarios.
- Extensive ablation studies reveal that gradient-based diversification provides the strongest stabilization effect, offering practical guidance for reliable test-time adaptation.

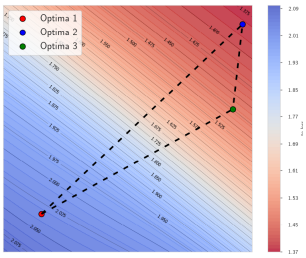


Fig. 3: The interpolation of three optima in the test loss landscape is visualized, where blue regions indicate high error/loss and red regions correspond to low error/loss.

2 Preliminaries

We first introduce the test-time adaptation setting, the entropy minimization objective commonly used for unlabeled target data, and the evaluation protocols adopted in our experiments. This provides the notation and experimental context needed to motivate our proposed diversified adaptation framework.

2.1 Test Time Adaptation

Suppose that we have a training (source) dataset $\mathcal{D}^s = \{(\mathbf{x}_j^s, y_j^s)\}_{j=1}^{N^s}$ where $\mathbf{x}_j^s \in \mathcal{X}^s$ and $y_j^s \in \mathcal{Y}^s$, N^s is the number of training instances, and a testing (target) set $\mathcal{D}^t = \{(\mathbf{x}_j^t, y_j^t)\}_{j=1}^{N^t}$ where $\mathbf{x}_j^t \in \mathcal{X}^t$ and $y_j^t \in \mathcal{Y}^t$, N^t is the number of test instances. A source model $\boldsymbol{f}_{\boldsymbol{\theta}}(\cdot)$ with parameters $\boldsymbol{\theta}$ is first trained on $\mathcal{D}^s$ and then evaluated on $\mathcal{D}^t$. When $\mathcal{D}^t$ involves OOD, $\boldsymbol{f}_{\boldsymbol{\theta}}(\cdot)$ may not generalize well. Test Time Adaptation approaches adapt the model $\boldsymbol{f}_{\boldsymbol{\theta}}(\cdot)$ on the target dataset $\mathcal{D}^t = \{\mathbf{x}_j^t\}_{j=1}^{N^t}$ without any labels, since labels y_j^t are not accessible at test time. In this paper, we focus on TTA approaches that only reuse the normalization layers of the source model; therefore, without loss of generality, we use $\boldsymbol{\theta}$ to represent all parameters of the normalization layers.

2.2 Entropy Minimization

Without the need for additional labeled data, previous TTA methods [12, 19] utilize the entropy of model predictions as the objective function to adapt the model $\boldsymbol{f}_{\boldsymbol{\theta}}(\cdot)$:

$$\min H_{\boldsymbol{\theta}}(\mathcal{X}^t) = \min \sum_{j}^{N^t} H(\boldsymbol{f}_{\boldsymbol{\theta}}(\mathbf{x}_j^t)), \tag{1}$$

where $H_{\boldsymbol{\theta}}(\cdot)$ is entropy and measured by the Shannon entropy [17]:

$$H_{\boldsymbol{\theta}}(\mathbf{x}) = -\mathbf{p}_{\boldsymbol{\theta}}(\mathbf{x}) \cdot \log \mathbf{p}_{\boldsymbol{\theta}}(\mathbf{x}) = -\sum_{c}^{C} p_{\boldsymbol{\theta}}(x)_c \log p_{\boldsymbol{\theta}}(x)_c, \tag{2}$$

Table 1: Multi-seed DeYO solutions under batch size one (Zoom Blur, severity 5). Despite nearly identical final entropy values, different random seeds lead to substantial variance in final accuracy and distinct parameter endpoints. This empirical evidence supports the presence of underspecification in entropy-based test-time adaptation.

Seed	Final Entropy ↓	Acc. (%) ↑	$\|\theta^{\text{final}} - \theta^0\|_2$	Mean Pairwise $\|\Delta\theta\|_2$
0	0.247	45.1	0.92	0.51
1	0.241	34.6	0.88	0.47
2	0.245	41.8	0.95	0.54
3	0.249	32.9	0.90	0.49
4	0.244	37.5	0.93	0.52
Mean ± Std	0.245 ± 0.003	38.37 ± 4.80	0.92 ± 0.03	0.51 ± 0.03

where $\mathbf{p}_{\boldsymbol{\theta}}(\mathbf{x}) = \text{softmax}(\boldsymbol{f}_{\boldsymbol{\theta}}(\mathbf{x})) = (p_{\boldsymbol{\theta}}(x)_1, ..., p_{\boldsymbol{\theta}}(x)_C)$, C is the number of class labels for the underlying classification task. Entropy is an unsupervised objective since it only depends on the model predictions instead of annotations.

2.3 Benchmark Methods and Datasets

For comparison, we conducted experiments on several benchmark datasets: **1) The ImageNet-C** [7] benchmark consists of 15 diverse corruption types encompassing 5 severity levels for each type applied to validation images of ImageNet. **2) CIFAR-10-C and CIFAR-100-C** [7] benchmarks provide the same 15 corruption types as ImageNet-C, but all images are derived from CIFAR-10 and CIFAR-100, respectively. We applied our method to mild scenarios proposed by [19] and wild scenarios proposed by [12]. The mild scenario is to adapt models with a batch of test samples that have the same distribution shift type and randomly shuffled label distribution. In contrast, the wild scenario includes three more realistic settings: a) dynamic shifts in the ground-truth test label distribution, which lead to class imbalance at each corruption, b) batch size equal to one (single-sample adaptation), and c) mixtures of multiple distribution shifts. We evaluate three representative test time adaptation baseline methods including Tent [19], SAR [12] and DeYO [8], and employ ResNet-50-GN (group normalization) and ViTBase-LN (layer normalization) [4] for ImageNet-C, CLIP-VitBase-Patch32-LN [15] for CIFAR-10-C and CIFAR-100-C, taking into consideration test time adaptation's utilization across various normalization layers. To further assess robustness under broader distribution shifts, we additionally evaluate on ColoredMNIST, Waterbirds, ImageNet-R [5], and VisDA-2021 [1], which capture spurious correlations and domain shifts beyond corruption-based benchmarks. Detailed results are provided in Appendix.

3 Methodology and Results

Prior TTA approaches adapt a single model at test time by minimizing an unsupervised objective, most commonly, prediction entropy [19], or by combining

entropy minimization with augmentation-based filtering [3, 8, 9]. While these strategies have demonstrated promising improvements under distribution shift, they rely on updating a single parameter instance in the absence of supervision. As discussed in Sec. 1, entropy minimization is inherently underconstrained: multiple parameter configurations can achieve similar low entropy while corresponding to qualitatively different decision boundaries. Updating a single model therefore risks drifting toward suboptimal low-entropy solutions, amplifying simplicity bias and accumulating adaptation errors over time.

To empirically evaluate the stability of single-model entropy-based adaptation, we run DeYO [8] under the batch size 1 protocol multiple times with different random seeds while keeping the hyperparameters fixed. Table 1 shows that the final entropy values are nearly identical across runs (0.245 ± 0.003), yet the resulting accuracies vary substantially ($38.37 \pm 4.80\%$). Furthermore, the adapted normalization parameters converge to different endpoints in parameter space. In particular, $\|\boldsymbol{\theta}^{\text{final}} - \boldsymbol{\theta}^0\|_2$ measures the magnitude of the parameter change from the source model, while the mean pairwise distance $\|\Delta\boldsymbol{\theta}\|_2$ denotes the average ℓ_2 distance between final parameter vectors obtained from different seeds (computed across all seed pairs). These results indicate that entropy minimization does not uniquely specify a robust adaptation outcome: multiple low-entropy solutions correspond to qualitatively different decision boundaries and generalization performance. This instability highlights the need for structured mechanisms that explicitly maintain diverse adaptation hypotheses. To address this limitation, we adopt a population-based formulation of test-time adaptation. Instead of committing to a single adaptation trajectory, we maintain a collection of K interacting adaptation particles

$$\Theta = \{\theta_i\}_{i=1}^{K},$$

where each θ_i represents the normalization parameters adapted from the shared source model θ^0. All non-normalization layers remain frozen and shared across particles, following standard TTA practice [8, 12, 19]. Each particle is initialized from θ^0 and adapted online on the target stream. We define the general population adaptation objective as

$$\mathcal{L}(X) = \frac{1}{K}\sum_{i=1}^{K} \ell(X;\theta_i) + \lambda\,\Omega(\Theta), \tag{3}$$

where $\ell(X;\theta_i)$ denotes the entropy-based adaptation loss for particle i (Eq. 1), and $\Omega(\Theta)$ is a diversification regularizer that prevents premature particle collapse. The key idea is to encourage controlled exploration of multiple low-entropy regions in parameter space while maintaining interaction between particles. Unlike naive ensembling, where models are trained independently and averaged post hoc, our formulation explicitly couples particle dynamics during adaptation. Diversification therefore acts not as a static aggregation trick, but as a mechanism for mitigating underspecification during optimization.

We instantiate $\Omega(\Theta)$ through complementary diversification mechanisms operating at multiple levels: output-level, parameter-level, optimizer-level, and

input-level. Each mechanism targets a distinct failure mode of single-model entropy minimization under distribution shift, as detailed below.

3.1 Output-level Diversification

Output-level diversification promotes disagreement among particles at the prediction level, preventing premature collapse to identical predictive hypotheses under entropy minimization. Let $p_i(X)$ denote the predictive distribution of particle θ_i on input batch X. To encourage predictive diversity, we maximize pairwise divergence between particle outputs. We define the output-level regularizer as:

$$\Omega_{\mathrm{KL}}(\Theta) = -\frac{1}{K(K-1)} \sum_{i \neq j} D_{\mathrm{KL}}(p_i(X) \,\|\, p_j(X)) \,, \tag{4}$$

where $D_{\mathrm{KL}}(\cdot\|\cdot)$ denotes the Kullback–Leibler divergence. The negative sign ensures that minimizing the overall objective in Equation (3) increases predictive divergence between particles. This repulsive interaction encourages particles to occupy distinct low-entropy regions of the solution space, thereby reducing the risk of collective collapse into the same spurious decision boundary. Unlike naive ensembling, where models are trained independently and averaged post hoc, our formulation explicitly couples particle dynamics during adaptation, allowing disagreement to emerge as a controlled exploration mechanism. While naive ensembling already improves robustness over single-model adaptation, we empirically observe that independently trained particles tend to converge toward similar adaptation trajectories under entropy minimization. This trajectory alignment limits the diversity benefits of simple averaging. The proposed diversification mechanisms explicitly prevent such collapse and yield consistent gains beyond naive ensembling.

3.2 Parameter-level Diversification

Parameter-level diversification encourages adaptation particles to follow distinct trajectories rather than collapsing to the same low-entropy solution. We realize this through explicit particle repulsion, gradient-based functional diversity, and optimizer-induced trajectory variation.

3.2.1 Stein Variational Gradient Descent [10]: To further promote structured exploration of multiple adaptation hypotheses, we employ Stein Variational Gradient Descent (SVGD) as a particle interaction mechanism. Unlike standard gradient descent that converges to a single solution, SVGD evolves a set of particles jointly to approximate a target distribution through deterministic updates with repulsive interactions. We define a pseudo-posterior over normalization parameters as

$$\tilde{p}(\theta) \propto \exp(-\ell(X;\theta)) \exp\Big(-\frac{\gamma}{2}\|\theta - \theta^0\|_2^2\Big) \,, \tag{5}$$

where $\ell(X;\theta)$ is the entropy adaptation loss (Eq. 1). The second exponential term acts as an implicit Gaussian prior centered at the source normalization parameters

θ^0, discouraging excessive drift during adaptation. The coefficient $\gamma > 0$ controls the strength of this quadratic anchoring term, balancing adaptation flexibility and retention of source-domain inductive bias. SVGD updates each particle θ_i according to

$$\theta_i \leftarrow \theta_i + \epsilon\, \phi(\theta_i), \tag{6}$$

where $\epsilon > 0$ is the particle update step size, and

$$\phi(\theta_i) = \frac{1}{K} \sum_{j=1}^{K} \left[\kappa(\theta_j, \theta_i)\, \nabla_{\theta_j} \log \tilde{p}(\theta_j) + \nabla_{\theta_j} \kappa(\theta_j, \theta_i) \right]. \tag{7}$$

Here, K denotes the number of particles and $\kappa(\cdot, \cdot)$ is a positive-definite kernel (e.g., an RBF kernel) that induces repulsive interactions between particles. The first term drives particles toward low-entropy regions of the adaptation objective, while the second term prevents collapse by encouraging dispersion in parameter space. In the context of test-time adaptation, this joint update enables particles to explore multiple plausible low-entropy solutions, thereby mitigating underspecification and reducing sensitivity to spurious adaptation modes.

3.2.2 Gradient-based Diversification: To mitigate simplicity bias during test-time adaptation, we promote *functional diversity* by encouraging particles to respond differently to input perturbations. Instead of enforcing parameter-space separation alone, we directly regularize the alignment of input gradients across particles. For particle θ_i, let

$$g_i = \nabla_X \ell(X; \theta_i)$$

denote the gradient of the entropy loss with respect to the input batch X. We define the gradient-based diversification term as

$$\Omega_{\text{Grad}}(\Theta) = -\frac{1}{K(K-1)} \sum_{i \neq j} \langle g_i, g_j \rangle, \tag{8}$$

where $\langle \cdot, \cdot \rangle$ denotes the Euclidean inner product. Minimizing the overall objective in Eq. Equation (3) therefore penalizes large positive gradient alignment between particles. This encourages particles to rely on different input directions during adaptation, promoting diverse decision boundaries under distribution shift. Unlike output-level disagreement, which acts on predictions, or parameter-space repulsion, which operates directly in weight space, gradient-based diversification regularizes the *functional behavior* of particles. By reducing gradient alignment, particles are discouraged from collapsing toward identical adaptation trajectories, thereby improving robustness in underspecified settings. We thoroughly investigate output-level and parameter-level diversification under wild settings; results for mild settings are provided in Appendix.

Comparison on Wild Scenario. We evaluate our population-based adaptation framework on ImageNet-C under three challenging wild settings: (1) batch size 1, (2) label distribution shift, and (3) mixed distribution shifts. Results are reported in Tabs. 2 and 3.

Table 2: Comparisons with baseline methods on ImageNet-C wild senarios at severity level 5 under batch size 1 and label shifts averaged over five random seeds regarding accuracy (%).

	Noise			Blur				Weather				Digital				
Batch Size 1	**Gauss.**	**Shot**	**Impl.**	**Defoc.**	**Glass**	**Motion**	**Zoom**	**Snow**	**Frost**	**Fog**	**Brit.**	**Contr.**	**Elastic**	**Pixel**	**JPEG**	**Avg.**
ResNet-50-GN	18.01	19.78	17.91	19.80	11.41	21.39	24.90	40.41	47.29	33.60	69.31	36.31	18.59	28.41	52.29	30.60
Tent	$28.85_{\pm 0.4}$	$35.95_{\pm 0.3}$	$33.49_{\pm 0.3}$	$16.67_{\pm 0.2}$	$7.19_{\pm 0.3}$	$27.35_{\pm 0.3}$	$30.14_{\pm 0.4}$	$17.97_{\pm 0.2}$	$25.20_{\pm 0.2}$	$2.27_{\pm 0.2}$	$71.88_{\pm 0.1}$	$46.24_{\pm 0.1}$	$7.36_{\pm 0.3}$	$52.72_{\pm 0.3}$	$56.40_{\pm 0.1}$	30.65
SAR	$28.51_{\pm 0.3}$	$30.91_{\pm 0.4}$	$29.27_{\pm 0.1}$	$18.50_{\pm 0.1}$	$15.34_{\pm 0.3}$	$28.80_{\pm 0.3}$	$30.53_{\pm 0.2}$	$44.47_{\pm 0.3}$	$44.27_{\pm 0.2}$	$32.94_{\pm 0.6}$	$72.03_{\pm 0.2}$	$44.67_{\pm 0.1}$	$13.16_{\pm 2.8}$	$47.70_{\pm 0.1}$	$56.23_{\pm 0.1}$	35.82
DeYO	$41.97_{\pm 0.7}$	$44.21_{\pm 0.4}$	$43.03_{\pm 0.7}$	$21.81_{\pm 0.1}$	$23.28_{\pm 0.3}$	$40.87_{\pm 0.1}$	$26.10_{\pm 4.3}$	$53.24_{\pm 0.2}$	$51.22_{\pm 0.2}$	$35.13_{\pm 6.1}$	$72.39_{\pm 0.1}$	$52.50_{\pm 0.2}$	$42.44_{\pm 0.3}$	$59.10_{\pm 0.0}$	$58.73_{\pm 0.1}$	44.40
Naive	$42.54_{\pm 0.4}$	$44.95_{\pm 0.3}$	$43.61_{\pm 0.4}$	$22.47_{\pm 0.1}$	$23.97_{\pm 0.2}$	$41.00_{\pm 0.1}$	$30.90_{\pm 0.3}$	$53.70_{\pm 0.2}$	$51.66_{\pm 0.1}$	$27.93_{\pm 0.2}$	$73.37_{\pm 0.3}$	$52.89_{\pm 0.1}$	$47.01_{\pm 0.2}$	$59.66_{\pm 0.0}$	$59.53_{\pm 0.1}$	45.01
SVGD	$42.57_{\pm 0.3}$	$44.90_{\pm 0.3}$	$43.79_{\pm 0.3}$	$22.72_{\pm 0.1}$	$24.20_{\pm 0.2}$	$40.89_{\pm 0.1}$	$32.04_{\pm 0.3}$	$53.40_{\pm 0.1}$	$51.69_{\pm 0.1}$	$28.41_{\pm 0.1}$	$73.34_{\pm 0.0}$	$53.00_{\pm 0.1}$	$46.39_{\pm 0.1}$	$59.64_{\pm 0.1}$	$59.62_{\pm 0.1}$	45.11
KL	$42.50_{\pm 0.2}$	$44.74_{\pm 0.2}$	$43.70_{\pm 0.3}$	$22.50_{\pm 0.1}$	$24.06_{\pm 0.2}$	$40.77_{\pm 0.1}$	$35.05_{\pm 0.2}$	$53.67_{\pm 0.1}$	$51.80_{\pm 0.2}$	$32.44_{\pm 0.1}$	$73.33_{\pm 0.0}$	$52.97_{\pm 0.1}$	$47.02_{\pm 0.2}$	$59.55_{\pm 0.1}$	$59.58_{\pm 0.1}$	45.58
Grad	$42.73_{\pm 0.2}$	$44.99_{\pm 0.2}$	$43.88_{\pm 0.2}$	$22.49_{\pm 0.1}$	$24.32_{\pm 0.1}$	$41.07_{\pm 0.1}$	$25.63_{\pm 0.2}$	$54.10_{\pm 0.1}$	$52.03_{\pm 0.0}$	$58.88_{\pm 0.1}$	$73.33_{\pm 0.1}$	$53.28_{\pm 0.1}$	$47.69_{\pm 0.2}$	$59.41_{\pm 0.1}$	$59.70_{\pm 0.0}$	**46.90**
VitBase-LN	9.51	6.70	8.21	28.88	23.40	33.91	27.11	15.90	26.48	47.20	54.70	44.11	30.51	44.50	47.81	29.91
Tent	$51.15_{\pm 0.3}$	$51.78_{\pm 0.2}$	$52.50_{\pm 0.2}$	$52.18_{\pm 0.1}$	$47.68_{\pm 0.2}$	$56.28_{\pm 0.3}$	$49.06_{\pm 0.3}$	$8.78_{\pm 0.4}$	$15.92_{\pm 0.1}$	$67.25_{\pm 0.1}$	$73.45_{\pm 0.2}$	$66.64_{\pm 0.3}$	$52.09_{\pm 0.4}$	$64.93_{\pm 0.2}$	$63.98_{\pm 0.1}$	51.58
SAR	$50.25_{\pm 0.4}$	$50.65_{\pm 0.7}$	$51.85_{\pm 0.3}$	$51.61_{\pm 0.2}$	$48.92_{\pm 0.1}$	$56.71_{\pm 0.1}$	$50.55_{\pm 0.3}$	$20.45_{\pm 0.4}$	$54.45_{\pm 0.6}$	$67.42_{\pm 0.9}$	$74.92_{\pm 0.7}$	$65.84_{\pm 0.0}$	$54.68_{\pm 0.1}$	$66.57_{\pm 0.1}$	$64.91_{\pm 0.1}$	55.31
DeYO	$52.61_{\pm 0.7}$	$53.50_{\pm 1.7}$	$53.46_{\pm 0.8}$	$54.81_{\pm 0.1}$	$55.42_{\pm 0.1}$	$61.92_{\pm 0.1}$	$38.37_{\pm 4.8}$	$64.55_{\pm 0.1}$	$62.93_{\pm 0.0}$	$70.91_{\pm 0.1}$	$76.19_{\pm 0.0}$	$60.14_{\pm 0.1}$	$65.17_{\pm 0.1}$	$71.00_{\pm 0.1}$	$67.56_{\pm 0.3}$	60.57
Naive	$53.47_{\pm 0.3}$	$54.67_{\pm 0.2}$	$54.80_{\pm 0.4}$	$56.17_{\pm 0.2}$	$56.51_{\pm 0.1}$	$62.18_{\pm 0.1}$	$47.20_{\pm 0.3}$	$64.91_{\pm 0.3}$	$63.04_{\pm 0.1}$	$70.97_{\pm 0.1}$	$76.64_{\pm 0.1}$	$66.20_{\pm 0.1}$	$66.10_{\pm 0.1}$	$71.42_{\pm 0.1}$	$68.29_{\pm 0.3}$	62.17
SVGD	$54.11_{\pm 0.3}$	$54.78_{\pm 0.2}$	$54.55_{\pm 0.3}$	$57.01_{\pm 0.2}$	$57.08_{\pm 0.1}$	$62.85_{\pm 0.1}$	$51.33_{\pm 0.2}$	$65.01_{\pm 0.2}$	$63.69_{\pm 0.1}$	$72.04_{\pm 0.1}$	$77.18_{\pm 0.1}$	$68.04_{\pm 0.1}$	$66.22_{\pm 0.2}$	$71.73_{\pm 0.1}$	$68.53_{\pm 0.2}$	62.94
KL	$54.83_{\pm 0.3}$	$55.78_{\pm 0.3}$	$55.97_{\pm 0.2}$	$56.46_{\pm 0.3}$	$56.84_{\pm 0.2}$	$62.66_{\pm 0.2}$	$54.78_{\pm 0.2}$	$65.05_{\pm 0.1}$	$63.69_{\pm 0.2}$	$72.30_{\pm 0.2}$	$77.24_{\pm 0.1}$	$68.04_{\pm 0.1}$	$66.00_{\pm 0.0}$	$71.82_{\pm 0.1}$	$68.89_{\pm 0.2}$	**63.36**
Grad	$54.83_{\pm 0.3}$	$55.65_{\pm 0.1}$	$56.13_{\pm 0.2}$	$56.68_{\pm 0.3}$	$57.13_{\pm 0.1}$	$62.86_{\pm 0.1}$	$50.72_{\pm 0.1}$	$64.98_{\pm 0.1}$	$63.48_{\pm 0.0}$	$72.51_{\pm 0.1}$	$77.34_{\pm 0.1}$	$67.93_{\pm 0.1}$	$66.01_{\pm 0.0}$	$72.20_{\pm 0.1}$	$68.76_{\pm 0.0}$	63.14

	Noise			Blur				Weather				Digital				
Label Shifts	**Gauss.**	**Shot**	**Impl.**	**Defoc.**	**Glass**	**Motion**	**Zoom**	**Snow**	**Frost**	**Fog**	**Brit.**	**Contr.**	**Elastic**	**Pixel**	**JPEG**	**Avg.**
ResNet-50-GN	17.9	19.9	17.9	19.7	11.3	21.3	24.9	40.4	47.4	33.6	69.2	36.3	18.7	28.4	52.2	30.6
Tent	$1.67_{\pm 0.2}$	$2.02_{\pm 0.2}$	$2.01_{\pm 0.1}$	$12.17_{\pm 0.3}$	$2.14_{\pm 0.3}$	$23.00_{\pm 0.2}$	$13.74_{\pm 0.4}$	$8.86_{\pm 0.2}$	$13.69_{\pm 0.2}$	$1.00_{\pm 0.1}$	$73.00_{\pm 0.3}$	$48.44_{\pm 0.3}$	$3.51_{\pm 0.2}$	$54.90_{\pm 0.2}$	$57.18_{\pm 0.3}$	21.15
SAR	$28.71_{\pm 0.1}$	$31.19_{\pm 0.4}$	$29.39_{\pm 0.9}$	$18.52_{\pm 0.4}$	$15.10_{\pm 1.2}$	$28.71_{\pm 0.5}$	$29.84_{\pm 3.5}$	$42.86_{\pm 2.6}$	$44.33_{\pm 0.4}$	$35.32_{\pm 0.9}$	$72.07_{\pm 0.1}$	$44.65_{\pm 0.2}$	$14.40_{\pm 1.9}$	$47.04_{\pm 0.5}$	$56.22_{\pm 0.1}$	35.89
DeYO	$40.27_{\pm 0.5}$	$43.58_{\pm 0.2}$	$41.75_{\pm 0.3}$	$22.42_{\pm 0.0}$	$22.69_{\pm 0.2}$	$40.93_{\pm 0.2}$	$14.63_{\pm 2.3}$	$53.19_{\pm 0.4}$	$51.78_{\pm 0.5}$	$1.03_{\pm 0.2}$	$72.89_{\pm 0.1}$	$52.93_{\pm 0.2}$	$47.81_{\pm 1.2}$	$59.48_{\pm 0.0}$	$59.30_{\pm 0.1}$	41.65
Naive	$41.37_{\pm 0.3}$	$44.24_{\pm 0.2}$	$42.43_{\pm 0.3}$	$22.52_{\pm 0.1}$	$23.21_{\pm 0.2}$	$40.95_{\pm 0.2}$	$12.46_{\pm 0.4}$	$52.49_{\pm 0.3}$	$51.83_{\pm 0.4}$	$1.39_{\pm 0.4}$	$73.30_{\pm 0.3}$	$52.91_{\pm 0.2}$	$47.81_{\pm 0.3}$	$59.53_{\pm 0.1}$	$59.59_{\pm 0.2}$	41.74
SVGD	$42.43_{\pm 0.3}$	$44.82_{\pm 0.3}$	$43.66_{\pm 0.3}$	$22.98_{\pm 0.1}$	$23.54_{\pm 0.2}$	$40.74_{\pm 0.2}$	$17.55_{\pm 0.3}$	$53.11_{\pm 0.3}$	$51.97_{\pm 0.2}$	$4.31_{\pm 0.3}$	$73.51_{\pm 0.3}$	$52.97_{\pm 0.2}$	$46.82_{\pm 0.2}$	$59.38_{\pm 0.1}$	$59.58_{\pm 0.2}$	42.49
KL	$42.49_{\pm 0.3}$	$44.87_{\pm 0.3}$	$43.56_{\pm 0.2}$	$22.92_{\pm 0.1}$	$23.70_{\pm 0.2}$	$40.76_{\pm 0.2}$	$17.59_{\pm 0.2}$	$53.07_{\pm 0.3}$	$52.03_{\pm 0.2}$	$1.87_{\pm 0.3}$	$73.38_{\pm 0.3}$	$52.93_{\pm 0.2}$	$46.28_{\pm 0.1}$	$59.18_{\pm 0.2}$	$59.28_{\pm 0.1}$	42.26
Grad	$42.42_{\pm 0.2}$	$44.80_{\pm 0.2}$	$43.64_{\pm 0.2}$	$22.95_{\pm 0.1}$	$23.59_{\pm 0.2}$	$40.79_{\pm 0.1}$	$17.53_{\pm 0.1}$	$53.09_{\pm 0.1}$	$51.97_{\pm 0.0}$	$15.79_{\pm 0.1}$	$73.50_{\pm 0.1}$	$53.00_{\pm 0.2}$	$46.92_{\pm 0.0}$	$59.41_{\pm 0.1}$	$59.61_{\pm 0.0}$	**43.27**
ViTBase-LN	9.42	6.73	8.32	29.11	23.41	34.03	27.01	15.80	26.30	47.41	54.72	43.90	30.50	44.5	47.6	29.9
Tent	$53.41_{\pm 0.2}$	$53.16_{\pm 0.2}$	$54.36_{\pm 0.3}$	$54.34_{\pm 0.4}$	$52.46_{\pm 0.4}$	$58.74_{\pm 0.6}$	$52.83_{\pm 0.2}$	$4.29_{\pm 0.2}$	$7.04_{\pm 0.1}$	$69.46_{\pm 0.3}$	$74.92_{\pm 0.3}$	$67.42_{\pm 0.4}$	$59.97_{\pm 0.2}$	$67.89_{\pm 0.2}$	$66.13_{\pm 0.1}$	53.10
SAR	$51.90_{\pm 0.4}$	$51.47_{\pm 0.7}$	$53.02_{\pm 0.3}$	$51.69_{\pm 0.2}$	$48.57_{\pm 0.1}$	$56.55_{\pm 0.1}$	$50.62_{\pm 0.3}$	$26.45_{\pm 0.4}$	$54.29_{\pm 0.6}$	$68.00_{\pm 0.9}$	$74.28_{\pm 0.7}$	$65.65_{\pm 0.0}$	$55.44_{\pm 0.1}$	$66.68_{\pm 0.1}$	$65.07_{\pm 0.0}$	55.98
DeYO	$53.92_{\pm 0.3}$	$54.86_{\pm 0.7}$	$55.08_{\pm 0.4}$	$54.78_{\pm 0.1}$	$55.70_{\pm 0.2}$	$61.63_{\pm 0.1}$	$53.34_{\pm 0.2}$	$64.31_{\pm 0.4}$	$63.10_{\pm 1.1}$	$70.75_{\pm 0.3}$	$76.81_{\pm 0.1}$	$65.90_{\pm 0.3}$	$65.60_{\pm 0.3}$	$71.65_{\pm 0.2}$	$68.10_{\pm 0.2}$	62.37
Naive	$54.11_{\pm 0.3}$	$54.91_{\pm 0.4}$	$55.41_{\pm 0.4}$	$55.75_{\pm 0.1}$	$56.25_{\pm 0.2}$	$61.95_{\pm 0.1}$	$56.15_{\pm 0.1}$	$64.14_{\pm 0.1}$	$63.03_{\pm 0.3}$	$71.76_{\pm 0.4}$	$76.74_{\pm 0.2}$	$67.47_{\pm 0.2}$	$65.54_{\pm 0.2}$	$71.51_{\pm 0.1}$	$67.95_{\pm 0.2}$	62.84
SVGD	$54.59_{\pm 0.3}$	$55.41_{\pm 0.3}$	$55.94_{\pm 0.3}$	$56.17_{\pm 0.1}$	$56.77_{\pm 0.2}$	$62.50_{\pm 0.1}$	$56.22_{\pm 0.1}$	$64.63_{\pm 0.2}$	$63.46_{\pm 0.3}$	$72.32_{\pm 0.3}$	$77.25_{\pm 0.2}$	$67.54_{\pm 0.2}$	$65.98_{\pm 0.2}$	$71.98_{\pm 0.1}$	$68.55_{\pm 0.1}$	63.29
KL	$54.56_{\pm 0.2}$	$55.33_{\pm 0.2}$	$55.89_{\pm 0.2}$	$56.08_{\pm 0.1}$	$56.52_{\pm 0.1}$	$62.47_{\pm 0.1}$	$56.76_{\pm 0.1}$	$64.65_{\pm 0.2}$	$63.57_{\pm 0.2}$	$72.21_{\pm 0.2}$	$77.24_{\pm 0.1}$	$67.97_{\pm 0.1}$	$65.81_{\pm 0.1}$	$71.98_{\pm 0.1}$	$68.35_{\pm 0.1}$	62.24
Grad	$54.52_{\pm 0.2}$	$55.37_{\pm 0.1}$	$55.89_{\pm 0.1}$	$56.09_{\pm 0.1}$	$56.74_{\pm 0.0}$	$62.47_{\pm 0.1}$	$57.31_{\pm 0.1}$	$64.63_{\pm 0.0}$	$63.55_{\pm 0.2}$	$72.36_{\pm 0.1}$	$77.23_{\pm 0.1}$	$67.80_{\pm 0.0}$	$65.92_{\pm 0.1}$	$72.01_{\pm 0.0}$	$68.52_{\pm 0.1}$	**63.36**

1) Batch size 1. When adaptation is performed on a single test sample, entropy minimization becomes highly unstable due to the lack of batch statistics. Table 2 shows that simply maintaining multiple normalization particles (`Ens`) already improves robustness over single-model adaptation. For ViTBase-LN, the ensemble setting improves average accuracy by 2.23% over DeYO, outperforming all baselines across most corruption types. ResNet-50-GN exhibits similar gains.

Introducing explicit diversification further strengthens performance. Among all strategies, gradient-based diversification (`Grad`) achieves the best results, improving average accuracy by 3.21% for ViTBase-LN and 2.50% for ResNet-50-GN over DeYO. Notably, under severe corruptions such as Fog, `Grad` attains the highest overall accuracy (58.88%), indicating improved stability under extreme underspecification.

2) Label Distribution Shift. Under infinite class imbalance ratio [12], entropy minimization is prone to reinforcing dominant classes. The ensemble baseline (`Ens`) improves over DeYO by 0.47% on ViTBase-LN, suggesting that population-based adaptation mitigates collapse toward biased modes. Applying diversification further enhances robustness: `Grad` achieves improvements of 0.99% (ViTBase-LN) and 1.62% (ResNet-50-GN) over DeYO. These gains indicate that gradient-level functional repulsion reduces overconfidence under skewed label distributions.

3) Mixed Distribution Shifts. We further evaluate performance on mixtures of 15 corruption types at severity level 5. As shown in Tab. 3, gradient-based diversification consistently delivers the strongest robustness. Compared with the best-performing baseline, `Grad` improves accuracy by 3% on ResNet-50-GN and

Table 3: Comparisons with baselines on ImageNet-C at severity 5 under a mixture of 15 corruptions. Results report top-1 and top-5 accuracy (%).

ResNet-50-GN		
Mix Shifts	top@1	top@5
No Adapt	30.61	53.89
Tent	$29.80_{\pm 0.2}$	$30.65_{\pm 0.2}$
SAR	$38.12_{\pm 0.1}$	$59.70_{\pm 0.1}$
DeYO	$31.36_{\pm 1.3}$	$50.78_{\pm 0.1}$
Naive	$\underline{33.06}_{\pm 0.3}$	$\underline{55.91}_{\pm 0.1}$
SVGD	$32.71_{\pm 0.4}$	$55.79_{\pm 0.2}$
KL	$32.66_{\pm 0.3}$	$52.23_{\pm 0.2}$
Grad	$\mathbf{34.36}_{\pm 0.2}$	$\mathbf{56.21}_{\pm 0.1}$

ViTBase-LN		
Mix Shifts	top@1	top@5
No Adapt	29.94	54.21
Tent	$32.36_{\pm 0.3}$	$47.40_{\pm 0.4}$
SAR	$57.78_{\pm 0.1}$	$78.14_{\pm 0.1}$
DeYO	$57.05_{\pm 1.1}$	$78.35_{\pm 1.3}$
Naive	$60.56_{\pm 0.4}$	$79.23_{\pm 0.4}$
SVGD	$\underline{60.84}_{\pm 0.3}$	$80.30_{\pm 0.2}$
KL	$60.37_{\pm 0.3}$	$\underline{80.42}_{\pm 0.3}$
Grad	$\mathbf{61.36}_{\pm 0.2}$	$\mathbf{81.45}_{\pm 0.2}$

Table 4: Accuracy (%) comparison with baseline methods, averaged over five random seeds, on ImageNet-C wild scenarios (severity level 5) using batch size 1 and three particles, each employing a different optimizer.

	Noise			Blur				Weather				Digital				
Batch Size 1	**Gauss.**	**Shot**	**Impl.**	**Defoc.**	**Glass**	**Motion**	**Zoom**	**Snow**	**Frost**	**Fog**	**Brit.**	**Contr.**	**Elastic**	**Pixel**	**JPEG**	**Avg.**
VitBase-LN	9.51	6.70	8.21	28.88	23.40	33.91	27.11	15.90	26.48	47.20	54.70	44.11	30.51	44.50	47.81	29.91
Tent	$51.15_{\pm 0.3}$	$51.78_{\pm 0.2}$	$52.50_{\pm 0.2}$	$52.18_{\pm 0.1}$	$47.68_{\pm 0.2}$	$56.28_{\pm 0.3}$	$49.06_{\pm 0.3}$	$8.78_{\pm 0.4}$	$15.92_{\pm 0.1}$	$67.25_{\pm 0.1}$	$73.45_{\pm 0.2}$	$66.64_{\pm 0.3}$	$52.09_{\pm 0.4}$	$64.93_{\pm 0.2}$	$63.98_{\pm 0.1}$	51.58
SAR	$50.25_{\pm 0.4}$	$50.65_{\pm 0.7}$	$51.85_{\pm 0.3}$	$51.61_{\pm 0.2}$	$48.92_{\pm 0.1}$	$56.71_{\pm 0.1}$	$50.55_{\pm 0.3}$	$20.45_{\pm 0.4}$	$54.45_{\pm 0.6}$	$67.42_{\pm 0.9}$	$74.92_{\pm 0.7}$	$65.84_{\pm 0.0}$	$54.68_{\pm 0.1}$	$66.57_{\pm 0.1}$	$64.91_{\pm 0.1}$	55.31
DeYO	$52.61_{\pm 0.7}$	$53.50_{\pm 1.7}$	$53.46_{\pm 0.8}$	$54.81_{\pm 0.1}$	$55.42_{\pm 0.1}$	$61.92_{\pm 0.1}$	$38.37_{\pm 4.8}$	$64.55_{\pm 0.1}$	$62.93_{\pm 0.0}$	$70.91_{\pm 0.1}$	$76.19_{\pm 0.0}$	$60.14_{\pm 0.1}$	$65.17_{\pm 0.1}$	$71.00_{\pm 0.1}$	$67.56_{\pm 0.3}$	60.57
Naive	$52.95_{\pm 0.5}$	$54.01_{\pm 0.7}$	$54.23_{\pm 0.7}$	$55.21_{\pm 0.1}$	$54.17_{\pm 0.1}$	$60.43_{\pm 0.1}$	$55.86_{\pm 0.9}$	$62.15_{\pm 0.1}$	$61.23_{\pm 0.1}$	$71.28_{\pm 0.1}$	$76.47_{\pm 0.1}$	$67.21_{\pm 0.1}$	$64.36_{\pm 0.0}$	$71.04_{\pm 0.1}$	$67.63_{\pm 0.2}$	61.88
SVGD	$53.87_{\pm 0.5}$	$54.65_{\pm 0.6}$	$54.93_{\pm 0.5}$	$56.57_{\pm 0.1}$	$56.24_{\pm 0.1}$	$62.43_{\pm 0.1}$	$54.21_{\pm 0.7}$	$63.72_{\pm 0.1}$	$62.94_{\pm 0.1}$	$72.33_{\pm 0.1}$	$76.82_{\pm 0.1}$	$68.64_{\pm 0.1}$	$65.33_{\pm 0.0}$	$71.26_{\pm 0.0}$	$68.22_{\pm 0.2}$	62.81
KL	$53.90_{\pm 0.5}$	$54.61_{\pm 0.4}$	$55.15_{\pm 0.4}$	$56.62_{\pm 0.1}$	$55.69_{\pm 0.1}$	$62.13_{\pm 0.1}$	$57.27_{\pm 0.4}$	$63.58_{\pm 0.1}$	$62.45_{\pm 0.1}$	$72.31_{\pm 0.1}$	$76.86_{\pm 0.1}$	$68.49_{\pm 0.1}$	$64.92_{\pm 0.1}$	$71.28_{\pm 0.1}$	$68.43_{\pm 0.1}$	$\underline{62.91}$
Grad	$55.13_{\pm 0.4}$	$55.28_{\pm 0.4}$	$56.22_{\pm 0.2}$	$56.23_{\pm 0.0}$	$55.84_{\pm 0.1}$	$62.02_{\pm 0.0}$	$56.94_{\pm 0.4}$	$63.73_{\pm 0.1}$	$62.24_{\pm 0.0}$	$72.31_{\pm 0.0}$	$76.90_{\pm 0.1}$	$68.78_{\pm 0.0}$	$64.78_{\pm 0.1}$	$71.37_{\pm 0.1}$	$68.47_{\pm 0.0}$	**63.08**

4.32% on ViTBase-LN. These results confirm that particle-level exploration combined with functional repulsion effectively stabilizes adaptation under compound shifts.

Overall, results across all wild scenarios demonstrate that (i) maintaining multiple adaptation particles already improves stability over single-model entropy minimization, and (ii) structured diversification, particularly gradient-based repulsion, provides additional robustness gains by preventing trajectory collapse under underspecified adaptation.

3.2.3 Optimizer Type Diversification: Beyond explicit regularization, we introduce trajectory-level diversification by assigning different optimization algorithms to different particles. While all particles minimize the same entropy objective, the choice of optimizer affects the geometry of parameter updates due to differences in momentum accumulation, adaptive scaling, and weight decay mechanisms. Specifically, for a particle θ_i, the update at iteration t is given by

$$\theta_i^{(t)} = \mathrm{Update}_{\mathrm{opt}_i}\left(\theta_i^{(t-1)}, \nabla_{\theta_i}\ell(X; \theta_i^{(t-1)})\right), \tag{9}$$

where $\mathrm{Update}_{\mathrm{opt}_i}$ denotes the update rule defined by the optimizer assigned to particle i (SGD, Adam, or AdamW), including its internal state variables. Using heterogeneous optimizers induces different curvature sensitivities and step dynamics across particles. For example, SGD follows uniform gradient directions, whereas Adam rescales updates according to estimated second-order statistics, and AdamW decouples weight decay from gradient updates. Consequently, even under identical entropy gradients, particles follow distinct trajectories in parameter space. This optimizer heterogeneity acts as an implicit diversification mechanism, reducing the likelihood that all particles converge toward the same adaptation

Table 5: Comparison with baseline methods on ImageNet-C wild scenarios (severity level 5) under batch size 1, evaluating the impact of augmentation, averaged over five random seeds, on accuracy (%).

	Noise			Blur				Weather				Digital				
Batch Size 1	**Gauss.**	**Shot**	**Impl.**	**Defoc.**	**Glass**	**Motion**	**Zoom**	**Snow**	**Frost**	**Fog**	**Brit.**	**Contr.**	**Elastic**	**Pixel**	**JPEG**	**Avg.**
VitBase-LN	9.51	6.70	8.21	28.88	23.40	33.91	27.11	15.90	26.48	47.20	54.70	44.11	30.51	44.50	47.81	29.91
Tent	$51.15_{\pm0.3}$	$51.78_{\pm0.2}$	$52.50_{\pm0.2}$	$52.18_{\pm0.1}$	$47.68_{\pm0.2}$	$56.28_{\pm0.3}$	$49.06_{\pm0.3}$	$8.78_{\pm0.4}$	$15.92_{\pm0.1}$	$67.25_{\pm0.1}$	$73.45_{\pm0.2}$	$66.64_{\pm0.3}$	$52.09_{\pm0.4}$	$64.93_{\pm0.2}$	$63.98_{\pm0.1}$	51.58
SAR	$50.25_{\pm0.4}$	$50.65_{\pm0.7}$	$51.85_{\pm0.3}$	$51.61_{\pm0.2}$	$48.92_{\pm0.1}$	$56.71_{\pm0.1}$	$50.55_{\pm0.3}$	$20.45_{\pm0.4}$	$54.45_{\pm0.6}$	$67.42_{\pm0.9}$	$74.92_{\pm0.7}$	$65.84_{\pm0.0}$	$54.68_{\pm0.1}$	$66.57_{\pm0.1}$	$64.91_{\pm0.1}$	55.31
DeYO	$52.61_{\pm0.7}$	$53.50_{\pm1.7}$	$53.46_{\pm0.8}$	$54.81_{\pm0.1}$	$55.42_{\pm0.1}$	$61.92_{\pm0.1}$	$38.37_{\pm4.8}$	$64.55_{\pm0.1}$	$62.93_{\pm0.0}$	$70.91_{\pm0.1}$	$76.19_{\pm0.0}$	$60.14_{\pm0.1}$	$65.17_{\pm0.1}$	$71.00_{\pm0.1}$	$67.56_{\pm0.3}$	60.57
Naive+Aug	$54.21_{\pm0.3}$	$55.23_{\pm0.4}$	$55.13_{\pm0.4}$	$55.38_{\pm0.1}$	$55.61_{\pm0.1}$	$62.58_{\pm0.1}$	$43.11_{\pm0.8}$	$65.49_{\pm0.1}$	$64.61_{\pm0.1}$	$71.67_{\pm0.1}$	$76.38_{\pm0.1}$	$68.41_{\pm0.1}$	$66.39_{\pm0.1}$	$71.29_{\pm0.1}$	$68.37_{\pm0.2}$	62.26
SVGD +Aug	$54.72_{\pm0.3}$	$56.14_{\pm0.3}$	$56.17_{\pm0.2}$	$57.10_{\pm0.1}$	$57.52_{\pm0.1}$	$63.50_{\pm0.1}$	$43.13_{\pm0.5}$	$66.33_{\pm0.1}$	$65.14_{\pm0.0}$	$73.36_{\pm0.1}$	$77.61_{\pm0.1}$	$69.26_{\pm0.1}$	$67.61_{\pm0.1}$	$72.64_{\pm0.0}$	$69.24_{\pm0.1}$	<u>63.30</u>
KL +Aug	$54.54_{\pm0.3}$	$55.63_{\pm0.4}$	$55.98_{\pm0.3}$	$56.96_{\pm0.1}$	$57.25_{\pm0.1}$	$63.50_{\pm0.2}$	$41.01_{\pm0.5}$	$66.20_{\pm0.1}$	$64.96_{\pm0.1}$	$73.46_{\pm0.1}$	$77.52_{\pm0.1}$	$69.25_{\pm0.1}$	$68.22_{\pm0.1}$	$72.53_{\pm0.1}$	$69.53_{\pm0.1}$	63.10
Grad +Aug	$56.07_{\pm0.2}$	$56.99_{\pm0.3}$	$57.32_{\pm0.2}$	$57.35_{\pm0.0}$	$57.89_{\pm0.1}$	$63.85_{\pm0.1}$	$60.24_{\pm0.2}$	$66.28_{\pm0.0}$	$64.97_{\pm0.1}$	$73.30_{\pm0.1}$	$77.56_{\pm0.1}$	$69.29_{\pm0.1}$	$67.63_{\pm0.0}$	$72.68_{\pm0.0}$	$69.39_{\pm0.1}$	**64.72**

Table 6: Comparison of baselines and WaTT with our diversified variant of WaTT on CIFAR-100-C in terms of accuracy (%).

CIFAR-100-C	CLIP	TENT	TPT	TDA	DiffTPT	SAR	CLIPArTT	WATT-P	WATT-S	WATT-S+Aug	WATT-S+Aug+Reg.
Gaussian Noise	14.80	$14.38_{\pm0.14}$	$14.03_{\pm0.10}$	$8.20_{\pm0.35}$	$21.40_{\pm0.08}$	$15.85_{\pm0.06}$	$25.32_{\pm0.14}$	$31.28_{\pm0.03}$	$32.07_{\pm0.23}$	$34.07_{\pm0.03}$	$34.91_{\pm0.33}$
Shot noise	16.03	$17.34_{\pm0.27}$	$15.25_{\pm0.17}$	$9.58_{\pm0.43}$	$24.17_{\pm0.49}$	$17.41_{\pm0.05}$	$27.90_{\pm0.05}$	$33.44_{\pm0.11}$	$34.36_{\pm0.11}$	$35.87_{\pm0.13}$	$36.79_{\pm0.26}$
Impulse Noise	13.85	$10.03_{\pm0.13}$	$13.01_{\pm0.13}$	$7.63_{\pm0.19}$	$16.87_{\pm0.24}$	$14.90_{\pm0.09}$	$25.62_{\pm0.09}$	$29.40_{\pm0.11}$	$30.33_{\pm0.03}$	$32.25_{\pm0.05}$	$33.81_{\pm0.19}$
Defocus blur	36.74	$49.05_{\pm0.07}$	$37.07_{\pm0.17}$	$25.59_{\pm0.41}$	$20.30_{\pm0.30}$	$42.00_{\pm0.04}$	$49.88_{\pm0.23}$	$52.32_{\pm0.28}$	$52.99_{\pm0.16}$	$53.71_{\pm0.23}$	$54.16_{\pm0.28}$
Glass blur	14.19	$3.71_{\pm0.07}$	$16.41_{\pm0.02}$	$9.83_{\pm0.05}$	$15.57_{\pm0.46}$	$15.07_{\pm0.02}$	$27.89_{\pm0.03}$	$31.20_{\pm0.12}$	$32.15_{\pm0.30}$	$35.12_{\pm0.30}$	$35.61_{\pm0.33}$
Motion blur	36.14	$46.62_{\pm0.27}$	$37.52_{\pm0.23}$	$28.92_{\pm0.18}$	$21.10_{\pm0.64}$	$39.52_{\pm0.06}$	$47.93_{\pm0.14}$	$49.72_{\pm0.15}$	$50.53_{\pm0.12}$	$51.87_{\pm0.20}$	$51.98_{\pm0.23}$
Zoom blur	40.24	$51.84_{\pm0.15}$	$42.09_{\pm0.11}$	$31.08_{\pm0.36}$	$25.53_{\pm0.36}$	$45.40_{\pm0.06}$	$52.70_{\pm0.06}$	$54.72_{\pm0.04}$	$55.30_{\pm0.22}$	$56.29_{\pm0.05}$	$56.27_{\pm0.11}$
Snow	38.95	$46.71_{\pm0.21}$	$43.25_{\pm0.32}$	$32.94_{\pm0.12}$	$28.83_{\pm0.37}$	$43.56_{\pm0.04}$	$49.72_{\pm0.17}$	$51.79_{\pm0.04}$	$52.77_{\pm0.15}$	$54.09_{\pm0.12}$	$55.21_{\pm0.17}$
Frost	40.56	$44.90_{\pm0.27}$	$43.31_{\pm0.31}$	$34.84_{\pm0.25}$	$31.60_{\pm0.32}$	$41.70_{\pm0.06}$	$49.63_{\pm0.10}$	$53.04_{\pm0.24}$	$53.79_{\pm0.21}$	$54.63_{\pm0.13}$	$54.78_{\pm0.12}$
Fog	38.00	$47.31_{\pm0.18}$	$30.81_{\pm0.31}$	$31.13_{\pm0.16}$	$16.60_{\pm0.43}$	$40.36_{\pm0.14}$	$48.77_{\pm0.04}$	$50.78_{\pm0.21}$	$51.49_{\pm0.21}$	$53.32_{\pm0.12}$	$53.91_{\pm0.16}$
Brightness	48.18	$60.58_{\pm0.18}$	$50.23_{\pm0.11}$	$42.36_{\pm0.36}$	$31.20_{\pm0.16}$	$52.77_{\pm0.14}$	$61.27_{\pm0.08}$	$62.65_{\pm0.25}$	$63.57_{\pm0.21}$	$64.34_{\pm0.34}$	$65.13_{\pm0.37}$
Contrast	29.53	$45.90_{\pm0.18}$	$28.09_{\pm0.13}$	$19.03_{\pm0.32}$	$7.70_{\pm0.22}$	$28.41_{\pm0.11}$	$48.55_{\pm0.24}$	$51.34_{\pm0.14}$	$52.76_{\pm0.27}$	$54.99_{\pm0.20}$	$55.07_{\pm0.26}$
Elastic	26.33	$33.09_{\pm0.08}$	$28.12_{\pm0.17}$	$18.88_{\pm0.24}$	$21.60_{\pm0.54}$	$26.77_{\pm0.14}$	$37.45_{\pm0.30}$	$39.97_{\pm0.30}$	$40.90_{\pm0.43}$	$43.02_{\pm0.22}$	$43.01_{\pm0.27}$
Pixelate	21.98	$26.47_{\pm0.06}$	$20.43_{\pm0.14}$	$14.59_{\pm0.22}$	$22.83_{\pm0.31}$	$23.88_{\pm0.09}$	$33.88_{\pm0.16}$	$39.59_{\pm0.09}$	$40.97_{\pm0.16}$	$44.29_{\pm0.14}$	$44.81_{\pm0.17}$
JPEG comp.	25.91	$29.89_{\pm0.07}$	$28.82_{\pm0.09}$	$17.56_{\pm0.11}$	$31.77_{\pm0.45}$	$27.20_{\pm0.04}$	$36.07_{\pm0.32}$	$38.99_{\pm0.16}$	$39.59_{\pm0.08}$	$41.67_{\pm0.19}$	$42.48_{\pm0.24}$
Mean	29.43	35.19	30.46	22.08	22.89	31.92	41.51	44.68	45.57	47.30	47.86

basin. As shown in Tab. 4, combining optimizer diversification with explicit regularization (SVGD, Grad, or KL) further improves robustness under batch size 1 adaptation.

3.3 Input-Level Diversification

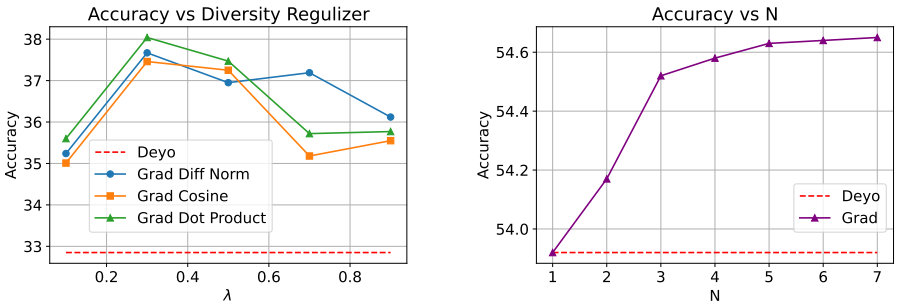
We further introduce diversification at the input level by incorporating structured data augmentations during adaptation. Unlike prior work that employs augmentations primarily for filtering or consistency regularization [8], we use them as optimization-time perturbations that alter the local geometry of the entropy landscape experienced by each particle. Given an input $\mathbf{x}$, we construct a set of transformed views

$$\mathcal{A}(\mathbf{x}) = \{\mathbf{x}, \mathbf{x}^{(h)}, \mathbf{x}^{(v)}\},$$

where $\mathbf{x}^{(h)}$ and $\mathbf{x}^{(v)}$ denote horizontal and vertical flips. For each particle θ_i, adaptation is performed over all views:

$$\ell_{\text{aug}}(X; \theta_i) = \sum_{\mathbf{x}' \in \mathcal{A}(\mathbf{x})} \ell(\mathbf{x}'; \theta_i). \tag{10}$$

The augmented entropy loss replaces $\ell(X; \theta_i)$ in the population objective (Eq. 3), while diversification regularizers remain unchanged.



(a) Comparison of gradient-based diversity variants on ImageNet-C with zoom blur corruption (severity 5).

(b) Impact of the particle collection size N on ImageNet-C with Gaussian corruption (severity 5).

Fig. 4: Ablation studies of diversification design choices on ImageNet-C under the wild test-time adaptation scenario (batch size 1) using a ViTBase-LN backbone.

Input diversification influences adaptation in two complementary ways. First, it stabilizes entropy minimization by preventing overfitting to a single input configuration. Second, because augmentations perturb gradients differently for each particle, the repulsive regularizers (like SVGD, Grad, etc.) amplify trajectory separation in parameter space. Consequently, particles explore distinct local minima corresponding to different augmented perspectives of the target stream. As reported in Tab. 5, combining input diversification with gradient-based repulsion yields the strongest robustness under batch-size-one adaptation. Across corruption types, the proposed configuration consistently outperforms prior TTA methods and improves over DeYO by up to 4%. Furthermore, compared to WaTT [13], which adapts only the text encoder of CLIP, our framework operates directly on the image encoder and achieves an additional 2% average accuracy gain, demonstrating the complementary benefit of particle-level diversification. It should be noted that our framework maintains K adaptation particles whose predictions are aggregated at inference time. As a result, the computational cost scales approximately linearly with K. In our experiments, we use $K = 3$, resulting in roughly a $3\times$ inference overhead compared to single-model TTA methods. However, since only the normalization layers are updated while the remaining network parameters remain frozen, the additional optimization cost remains lightweight. A more detailed discussion is provided in the Appendix.

4 Ablation Study

We conduct ablation studies to examine the main design choices of our wrapper. Specifically, we analyze the formulation of the diversity regularizer, the effect of collection size, and the generality of the proposed objective when used as a plug-in component for an existing entropy-based baseline.

Table 7: Comparisons with baseline methods on ImageNet-C wild scenario at severity level 5 under batch size 1 across part of corruption types regarding mean of accuracy(%).

	Noise			Blur				
Batch Size 1	**Gauss.**	**Shot**	**Impl.**	**Defoc.**	**Glass**	**Motion**	**Zoom**	**Avg.**
VitBase-LN	9.51	6.72	8.21	29.01	23.41	33.87	27.11	19.69
Tent	51.15	51.78	52.50	52.18	47.68	56.28	49.06	51.52
SAR	50.25	50.65	51.85	51.61	48.92	56.71	50.55	51.51
DeYO	52.61	53.50	53.46	54.81	55.42	61.92	38.37	<u>52.87</u>
Tent + Grad	52.23	52.97	53.45	54.38	51.40	58.49	52.54	**53.63**

4.1 Different Gradient Diversity Formulation

We evaluate three formulations of gradient-based diversification: (i) pairwise dot-product alignment (Eq. 8), (ii) ℓ_2 difference between gradients, and (iii) cosine similarity. Figure 4a reports results on ImageNet-C (Zoom Blur, severity level 5) across grid-selected values of the regularization coefficient λ. Among the three variants, the dot-product formulation achieves the highest accuracy across most values of λ. Unlike cosine similarity, which normalizes magnitude information, and difference norms, which penalize scale rather than directional alignment, the dot-product directly discourages shared descent directions. This suggests that reducing gradient alignment, rather than merely increasing gradient magnitude disparity, is more effective for preventing trajectory collapse under entropy minimization.

Performance peaks around $\lambda = 0.3$, indicating a balanced trade-off between entropy minimization and functional repulsion. Very large λ values degrade accuracy, confirming that excessive repulsion can hinder adaptation.

4.2 Sensitivity to Collection Size N

We analyze the effect of the number of particles on Gaussian noise corruption (severity level 5). As shown in Figure 4b, accuracy improves substantially when increasing from $N = 1$ (single-model adaptation) to $N = 3$, demonstrating the benefit of multi-trajectory exploration. Beyond $N = 3$, gains become marginal relative to computational overhead. This saturation behavior indicates that a small number of interacting particles is sufficient to capture multiple plausible low-entropy modes. Accordingly, we adopt $N = 3$ in all experiments to balance robustness and efficiency.

4.3 Plug-in Evaluation on Tent

To verify that gradient diversification is not specific to our ensemble initialization, we integrate the proposed gradient repulsion term into Tent [19]. Table 7 shows that Tent-`Grad` consistently improves over vanilla Tent and achieves competitive performance with DeYO. This demonstrates that gradient-based functional diversification acts as a general stabilization mechanism for entropy-based TTA, independent of the underlying baseline.

5 Conclusion

We revisited entropy-based test-time adaptation through the lens of underspecification and argued that conventional single-model adaptation implicitly performs point-estimate (MAP-like) inference under highly underconstrained objectives. Such point estimates are unstable under distribution shift, often leading to trajectory collapse and reliance on spurious modes. To address this limitation, we proposed a population-based diversification framework that maintains multiple interacting adaptation particles. By introducing structured diversification at the parameter, functional (gradient), optimizer, and input levels, our method transforms TTA from brittle single-trajectory optimization into controlled multi-hypothesis exploration. Extensive experiments across ImageNet-C and related benchmarks demonstrate consistent improvements under challenging settings, including batch-size-one adaptation, label distribution shifts, and mixed corruptions. Ablation studies further confirm that gradient-based functional repulsion is particularly effective in preventing trajectory alignment under entropy minimization. Overall, our results suggest that treating test-time adaptation as a multi-hypothesis inference problem is essential for reliable deployment under distribution shift. We hope this perspective motivates future work toward principled population-based adaptation strategies that improve robustness under underspecification.

References

1. Bashkirova, D., Hendrycks, D., Kim, D., Liao, H., Mishra, S., Rajagopalan, C., Saenko, K., Saito, K., Tayyab, B.U., Teterwak, P., et al.: Visda-2021 competition: Universal domain adaptation to improve performance on out-of-distribution data. In: NeurIPS 2021 Competitions and Demonstrations Track. pp. 66–79. PMLR (2022)
2. Boudiaf, M., Mueller, R., Ben Ayed, I., Bertinetto, L.: Parameter-free online test-time adaptation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8344–8353 (2022)
3. Chen, D., Wang, D., Darrell, T., Ebrahimi, S.: Contrastive test-time adaptation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 295–305 (2022)
4. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An image is worth 16x16 words: Transformers for image recognition at scale. In: 9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021. OpenReview.net (2021), https://openreview.net/forum?id=YicbFdNTTy
5. Hendrycks, D., Basart, S., Mu, N., Kadavath, S., Wang, F., Dorundo, E., Desai, R., Zhu, T., Parajuli, S., Guo, M., et al.: The many faces of robustness: A critical analysis of out-of-distribution generalization. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 8340–8349 (2021)
6. Hendrycks, D., Dietterich, T.: Benchmarking neural network robustness to common corruptions and perturbations. arXiv preprint arXiv:1903.12261 (2019)

7. Hendrycks, D., Dietterich, T.G.: Benchmarking neural network robustness to common corruptions and perturbations. In: 7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019. OpenReview.net (2019), https://openreview.net/forum?id=HJz6tiCqYm
8. Lee, J., Jung, D., Lee, S., Park, J., Shin, J., Hwang, U., Yoon, S.: Entropy is not enough for test-time adaptation: From the perspective of disentangled factors. In: The Twelfth International Conference on Learning Representations (2024), https://openreview.net/forum?id=9w3iw8wDuE
9. Lee, J., Jung, D., Lee, S., Park, J., Shin, J., Hwang, U., Yoon, S.: Entropy is not enough for test-time adaptation: From the perspective of disentangled factors. arXiv preprint arXiv:2403.07366 (2024)
10. Liu, Q., Wang, D.: Stein variational gradient descent: A general purpose bayesian inference algorithm. In: Advances in Neural Information Processing Systems. vol. 29 (2016)
11. Niu, S., Chen, G., Zhao, P., Wang, T., Wu, P., Shen, Z.: Self-bootstrapping for versatile test-time adaptation. arXiv preprint arXiv:2504.08010 (2025)
12. Niu, S., Wu, J., Zhang, Y., Wen, Z., Chen, Y., Zhao, P., Tan, M.: Towards stable test-time adaptation in dynamic wild world. In: Internetional Conference on Learning Representations (2023)
13. Osowiechi, D., Noori, M., Hakim, G.A.V., Yazdanpanah, M., Bahri, A., Cheraghalikhani, M., Dastani, S., Beizaee, F., Ayed, I.B., Desrosiers, C.: Watt: Weight average test-time adaptation of clip. arXiv preprint arXiv:2406.13875 (2024)
14. Quiñonero, J., Sugiyama, M., Schwaighofer, A., Lawrence, N.D.: Dataset Shift in Machine Learning. MIT Press, Cambridge, MA (2009)
15. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I.: Learning transferable visual models from natural language supervision. In: Meila, M., Zhang, T. (eds.) Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event. Proceedings of Machine Learning Research, vol. 139, pp. 8748–8763. PMLR (2021), http://proceedings.mlr.press/v139/radford21a.html
16. Recht, B., Roelofs, R., Schmidt, L., Shankar, V.: Do imagenet classifiers generalize to imagenet? In: International conference on machine learning. pp. 5389–5400. PMLR (2019)
17. Shannon, C.E.: A mathematical theory of communication. The Bell System Technical Journal **27**, 379–423 (1948), http://plan9.bell-labs.com/cm/ms/what/shannonday/shannon1948.pdf
18. Sun, Y., Wang, X., Liu, Z., Miller, J., Efros, A., Hardt, M.: Test-time training with self-supervision for generalization under distribution shifts. In: International conference on machine learning. pp. 9229–9248. PMLR (2020)
19. Wang, D., Shelhamer, E., Liu, S., Olshausen, B., Darrell, T.: Tent: Fully test-time adaptation by entropy minimization. In: International Conference on Learning Representations (2021), https://openreview.net/forum?id=uXl3bZLkr3c
20. Zhang, M., Levine, S., Finn, C.: Memo: Test time robustness via adaptation and augmentation. Advances in neural information processing systems **35**, 38629–38642 (2022)
21. Zhang, Q., Bian, Y., Kong, X., Zhao, P., Zhang, C.: Come: Test-time adaption by conservatively minimizing entropy. arXiv preprint arXiv:2410.10894 (2024)
22. Zhao, H., Liu, Y., Alahi, A., Lin, T.: On pitfalls of test-time adaptation. arXiv preprint arXiv:2306.03536 (2023)