MoETTA: Test-Time Adaptation Under Mixed Distribution Shifts with MoE-LayerNorm

Xiao Fan[1,2], **Jingyan Jiang**[3*], **Zhaoru Chen**[3], **Fanding Huang**[2], **Xiao Chen**[2], **Qinting Jiang**[2], **Bowen Zhang**[3], **Xing Tang**[3], **Zhi Wang**[2]

[1]College of Computer Science and Technology, Tongji University, Shanghai, China
xiaofan140@gmail.com
[2]Shenzhen International Graduate School, Tsinghua University, Shenzhen, China
xiaofan140@gmail.com, {hfd24, chen-x25, jqt23}@mails.tsinghua.edu.cn, wangzhi@sz.tsinghua.edu.cn
[3]School of Artificial Intelligence, Shenzhen Technology University, Shenzhen, China
jiangjingyan@sztu.edu.cn, chenzhaoru2022@email.szu.edu.cn, zhang_bo_wen@foxmail.com, xing.tang@hotmail.com

Abstract

Test-Time Adaptation (TTA) has proven effective in mitigating performance drops under single-domain distribution shifts by updating model parameters during inference. However, real-world deployments often involve mixed distribution shifts, where test samples are affected by diverse and potentially conflicting domain factors, posing significant challenges even for state-of-the-art TTA methods. A key limitation in existing approaches is their reliance on a unified adaptation path, which fails to account for the fact that optimal gradient directions can vary significantly across different domains. Moreover, current benchmarks focus only on synthetic or homogeneous shifts, failing to capture the complexity of real-world heterogeneous mixed distribution shifts. To address this, we propose **MoETTA**, a novel entropy-based TTA framework that integrates the Mixture-of-Experts (MoE) architecture. Rather than enforcing a single parameter update rule for all test samples, MoETTA introduces a set of structurally decoupled experts, enabling adaptation along diverse gradient directions. This design allows the model to better accommodate heterogeneous shifts through flexible and disentangled parameter updates. To simulate realistic deployment conditions, we introduce two new benchmarks: *potpourri* and *potpourri+*. While classical settings focus solely on synthetic corruptions (i.e., ImageNet-C), potpourri encompasses a broader range of domain shifts—including natural, artistic, and adversarial distortions—capturing more realistic deployment challenges. Additionally, potpourri+ further includes source-domain samples to evaluate robustness against catastrophic forgetting. Extensive experiments across three mixed distribution shifts settings show that MoETTA consistently outperforms strong baselines, establishing new state-of-the-art performance and highlighting the benefit of modeling multiple adaptation directions via expert-level diversity.

Code — https://github.com/AnikiFan/MoETTA
Extended version — https://arxiv.org/abs/2511.13760

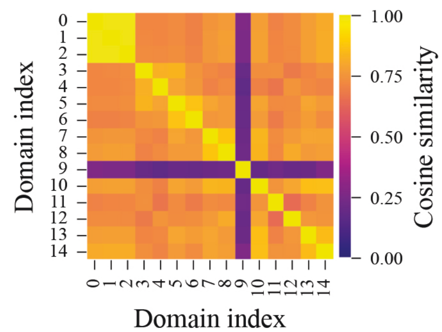

Figure 1. Cosine similarity heatmap of accumulated gradient directions $\boldsymbol{\theta}_i - \boldsymbol{\theta}_{\text{pre}}$, where $\boldsymbol{\theta}_i$ denotes the adapted model parameters on the i-th domain of ImageNet-C (Hendrycks and Dietterich 2019) using Tent (Wang et al. 2021), and $\boldsymbol{\theta}_{\text{pre}}$ is the pre-trained model parameters. Each entry at position (i, j) represents the cosine similarity between adaptation directions from domains i and j. The average similarity in the strictly lower triangle is 0.69, suggesting substantial variation across domains and highlighting the limitation of using a single adaptation direction under mixed distribution shifts.

1 Introduction

Deep learning has achieved remarkable success. However, it typically relies on the assumption that the source domain, on which the model is trained, shares the same distribution as the target domain, where the model is deployed (Ben-David et al. 2010). In practice, this assumption often fails to hold, leading to significant performance degradation under distribution shifts (Wang et al. 2024). Test-Time Adaptation (TTA) (Wang et al. 2021; Niu et al. 2022; Lee et al. 2024; Niu et al. 2023; Deng et al. 2025) has emerged as a promising solution to this problem, enabling models to adapt to unseen test distributions without requiring labeled data from the target domain. TTA methods adapt the model $f(\cdot;\boldsymbol{\theta})$ by updating its parameters through the minimization of an unsupervised loss function $\mathcal{L}(\boldsymbol{\theta})$, e.g., entropy loss (Wang et al. 2021), rotation prediction loss (Gidaris, Singh, and Komodakis 2018) and contrastive-based loss (Liu et al. 2021).

*Corresponding author.

Early works on TTA, such as Tent (Wang et al. 2021) and EATA (Niu et al. 2022), mainly focused on addressing single-domain distribution shifts, where all test samples originate from the same domain. Niu et al. (2023) pointed out that existing TTA methods fail under more realistic settings, with mixed distribution shifts as one of the representative cases. For instance, in real-world deployment scenarios, inference input batches are dynamically constructed from asynchronously arriving, heterogeneous task streams, such as concurrent requests from edge devices, without guarantee of domain consistency (Cang, Chen, and Huang 2024). Although recent studies have begun to explore these challenging "in-the-wild" scenarios, a significant performance gap remains between single-domain and mixed-domain settings. A strong baseline, MGTTA (Deng et al. 2025), reports an accuracy drop from 71.3% to 65.9% in the latter setting, which better reflects the challenges of real-world deployment.

To understand this performance gap, we investigate the underlying causes of this limitation. As shown in Fig. 1, the average cosine similarity between the accumulated gradient directions across the 15 domains of ImageNet-C is only 0.69. This indicates that the gradient directions for different domains are misaligned. Furthermore, assuming that the domain-specific d-dimensional parameters at convergence follow a Gaussian distribution, we theoretically show that the expected cosine similarity between the corresponding accumulated gradients, from the pre-trained parameters to the converged ones, converges to $0.5+\mathcal{O}(1/d)$ (see App. B). This result implies the inevitability of aforementioned misalignment. More critically, in some extreme cases, e.g., domain 9 in Fig. 1, the cosine similarity with other domains can be nearly zero, leading to slow learning and poor generalization (Jacobs et al. 1991). This phenomenon reveals a fundamental limitation of current methods in mixed distribution shifts scenarios: *by enforcing a shared adaptation direction, they fail to accommodate domain-specific gradient signals, which can be inconsistent or even conflicting.*

To address this challenge, we propose **MoETTA**, a novel test-time adaptation framework that incorporates Mixture-of-Experts (MoE) paradigm (Jacobs et al. 1991; Jordan and Jacobs 1994; Shazeer et al. 2017a; Dai et al. 2024) into entropy-based test-time adaptation. Our key **insight** is that, rather than enforcing a single, unified adaptation path, leveraging a diverse set of experts to represent multiple adaptation solutions within one model is particularly advantageous for mixed distribution shifts. Specifically, MoETTA reinterprets the LayerNorm (Ba, Kiros, and Hinton 2016) parameters in Vision Transformer (ViT) (Dosovitskiy et al. 2021) as a set of structurally decoupled expert branches. These experts offer distinct parameterizations, enabling the model to accommodate diverse gradient update directions during inference. Unlike the static domain knowledge vectors used by Chen et al. (2024a), our experts are dynamically updated during inference, allowing the model to flexibly absorb distribution-specific signals. What's more, through the routing mechanism, our methods can handle samples in a sample-wise way, which can further decompose diverse gradient directions brought by mixed distribution shifts.

To reflect more realistic deployment scenarios, we propose two new benchmarks: *potpourri* and *potpourri+*. Compared with the mixed distribution shifts setting proposed by Niu et al. (2023), which we refer to as the *classical mixed distribution shifts*, the potpourri setting includes samples from not only ImageNet-C (Hendrycks and Dietterich 2019) but also ImageNet-R (Hendrycks et al. 2020), ImageNet-A (Hendrycks et al. 2021), and ImageNet-Sketch (Wang et al. 2019), forming a heterogeneous mixture of distribution shifts. Beyond synthetic corruptions such as noise, blur, weather, and digital artifacts from ImageNet-C, potpourri introduces additional challenges from natural, artistic, and adversarial distortions, offering a more comprehensive testbed.

We further propose potpourri+, which augments potpourri with ImageNet validation samples to evaluate methods' resilience to catastrophic forgetting when simultaneously handling in-distribution (ID) and out-of-distribution (OOD) data—a common challenge in practical applications. Our main contributions are summarized as follows:

- We identify a key limitation of existing TTA methods under mixed distribution shifts and address it with MoETTA, an entropy-based method that leverages expert routing to model diverse adaptation directions.
- We propose two novel evaluation settings, i.e., potpourri and potpourri+, which better simulate realistic deployment conditions involving mixed distribution shifts.
- MoETTA achieves state-of-the-art performance on both existing mixed distribution shifts benchmark and the newly proposed potpourri and potpourri+ settings.

2 Related Work

Entropy-Based Test-Time Adaptation. Entropy-Based Test-Time Adaptation (Wang et al. 2021; Niu et al. 2022, 2023; Lee et al. 2024; Deng et al. 2025; Wang et al. 2022; Lee, Yoon, and Hwang 2024) refers to a class of TTA methods that enhance the robustness of pre-trained models under distribution shifts by minimizing prediction entropy (Grandvalet and Bengio 2004) during test-time adaptation.

Tent (Wang et al. 2021), EATA (Niu et al. 2022), SAR (Niu et al. 2023), and DeYO (Lee et al. 2024) adapt only the affine parameters of normalization layers. Among them, EATA, SAR, and DeYO improve the reliability of gradient signals through selective sample filtering, while SAR further incorporates Sharpness-Aware Minimization (SAM) (Foret et al. 2021) to enhance generalization under severe shifts. MGTTA (Deng et al. 2025) introduces a meta-gradient generator (MGG) to guide affine parameter updates.

While our method also belongs to the family of entropy-minimization-based approaches, it differs fundamentally in its architectural flexibility and update strategy. Some recent works have also explored structural modifications to enhance test-time adaptation. Notably, BECoTTA (Lee, Yoon, and Hwang 2024) proposes a MoE framework by incorporating domain-specific low-rank modules (LoRA) (Hu et al. 2022) as experts, aiming to address the challenges of Continual Test-Time Adaptation, where the dominant domain may gradually evolve over time. In contrast, our approach adopts LayerNorm as the expert unit and specifically targets the more challenging setting of mixed distribution shifts.

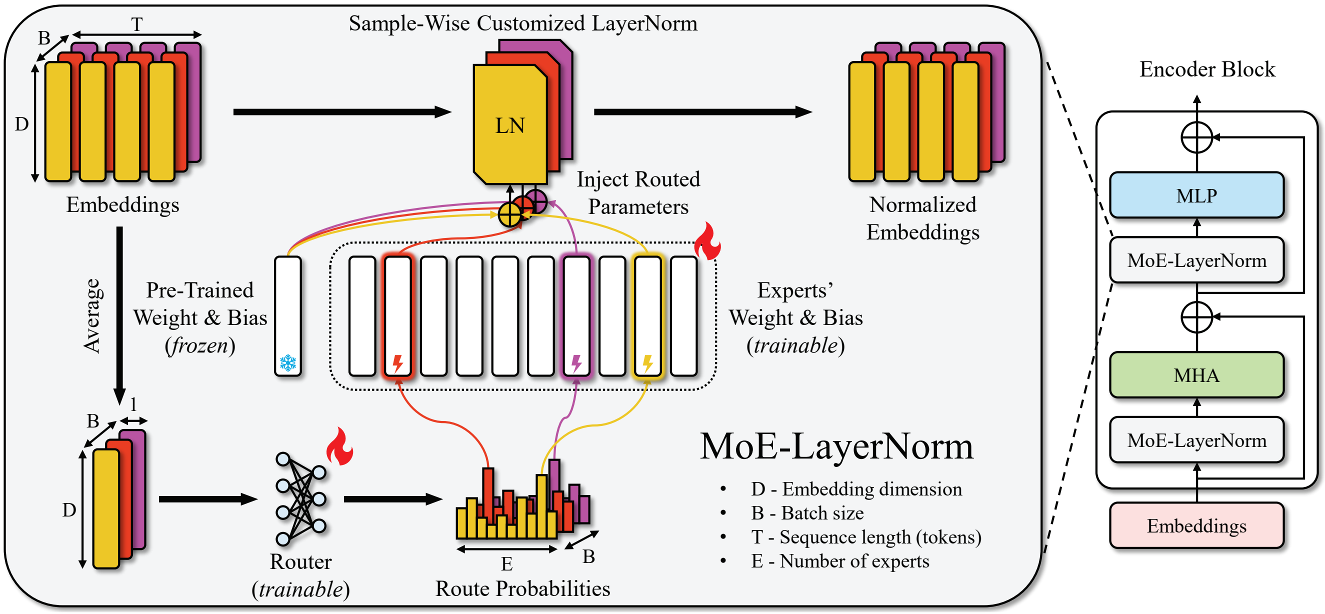

Figure 2. *Method Overview*. We replace the LayerNorm modules in the encoder blocks of a Vision Transformer (ViT) (Dosovitskiy et al. 2021) with our proposed MoE-LayerNorm. Colors of the embeddings and their routed components are used illustratively to suggest that the samples originate from different domains. For each input embedding, we first compute the mean across the token (sequence) dimension. This averaged vector is fed into a router to obtain routing probabilities, and the expert with the highest probability is selected. Its parameters are then added to the frozen pre-trained LayerNorm parameters to form a sample-specific LayerNorm. Finally, each token embedding is normalized using this sample-wise customized LayerNorm. A PyTorch-style (Paszke et al. 2019) pseudo code of the forward pass for MoE-LayerNorm can be found in App. A.

Mixture of Experts. Mixture of Experts was initially proposed for its ability to handle diverse subtasks by routing inputs to multiple expert networks through a gating mechanism (Jacobs et al. 1991). In recent years, MoE has gained significant traction in natural language processing (NLP) (Shazeer et al. 2017a; Dai et al. 2024; DeepSeek-AI 2024; Fedus, Zoph, and Shazeer 2022; Jiang et al. 2024; Shazeer et al. 2017b), as it enables large-scale model capacity while maintaining a limited computational footprint during inference. For example, Fedus, Zoph, and Shazeer (2022) proposed a Mixture-of-Experts layer for natural language processing, where multiple feed-forward networks with distinct parameters are treated as experts. During inference, the top-1 expert is selected based on the output of a router, which consists of a single fully connected layer.

Similar to the Deep Mixture of Experts (DMoE) framework proposed by Eigen, Ranzato, and Sutskever (2014), our method introduces an MoE structure at every layer of the model to support diverse computation pathways. However, a key distinction lies in how the experts are implemented: while DMoE uses full linear transformations followed by nonlinearities as experts, we adopt LayerNorm instances as lightweight experts. This design choice significantly reduces the computational overhead, as only the affine parameters of LayerNorm are adapted per expert. As a result, *our approach preserves expert diversity while maintaining high efficiency, making it well-suited for test-time adaptation*.

3 Methodology

We begin this section by formally defining the mixed distribution shifts setting. We then introduce the core component of our method, the MoE-LayerNorm module, as illustrated in Fig. 2. Finally, we detail the overall adaptation procedure.

3.1 Problem Statement

Consider a pre-trained model $f(\cdot\,;\boldsymbol{\theta})$, where $\boldsymbol{\theta}$ is learned from a labeled dataset $\{(\boldsymbol{x}_n, y_n)\}_{n=1}^{N}$ drawn from a source distribution $\mathcal{P}(\boldsymbol{x}, y)$. Once deployed, without access to ground-truth labels y, the model processes a stream of input sample batches $\mathcal{B}_0, \mathcal{B}_1, \ldots$, each of size B, drawn from a target distribution that typically differs from $\mathcal{P}(\boldsymbol{x}, y)$, resulting in distribution shifts that can significantly degrade performance. In the conventional single-domain setting, all test samples—both within and across batches—are assumed to come from a single target distribution $\mathcal{Q}(\boldsymbol{x}, y)$, with $\mathcal{Q} \neq \mathcal{P}$. In contrast, the more realistic mixed-domain setting allows each batch to contain samples from multiple target distributions $\mathcal{Q}_1(\boldsymbol{x}, y), \mathcal{Q}_2(\boldsymbol{x}, y), \ldots$, capturing diverse, domain-specific shifts encountered during deployment.

3.2 MoE-LayerNorm

Key Design Choices. Our method adapts MoE to the test-time adaptation setting through several key design choices:

1. To minimize computational overhead—a critical requirement for TTA—we designate the LayerNorm parameters

as experts instead of employing costly feed-forward networks. During adaptation, only the router and the expert-specific LayerNorm parameters are updated, ensuring lightweight and efficient optimization.

2. To promote diversity across experts while maintaining efficiency, we follow the top-1 routing strategy adopted by Fedus, Zoph, and Shazeer (2022), activating only one expert per sample. This design also avoids the need to merge multiple experts' output, thereby preventing parameter interference and preserving their individuality.
3. Given the limited number of samples at test-time, we construct the effective LayerNorm parameters used for normalizing each sample as the sum of the corresponding expert's parameters (initialized as zeros) selected by the router and the frozen pre-trained LayerNorm parameters, which we refer to as the *shared expert*. The inclusion of the shared expert provides domain-invariant knowledge, such as semantic representations learned from ID samples. In addition, it serves as a common foundation across all samples, which helps reduce redundancy among experts and encourages them to develop complementary adaptation behaviors (Dai et al. 2024).

Routing Mechanism. Our router is a linear projector that takes the mean sample embedding over the token dimension as input and outputs a vector whose dimension equals the number of experts. It is initialized with Xavier initialization (Glorot and Bengio 2010) at the start of adaptation.

Following Fedus, Zoph, and Shazeer (2022), in addition to task loss, we update routers with a differentiable load balancing loss. Given N experts indexed from 1 to N and the t-th batch $\mathcal{B}_t$, the auxiliary loss is computed as the scaled dot-product between two vectors $\boldsymbol{F}$ and $\boldsymbol{P}$:

$$\mathcal{L}_{\text{load balancing}} = N \times \sum_{i=1}^{N} \boldsymbol{F}_i \times \boldsymbol{P}_i. \tag{1}$$

Here, $\boldsymbol{F}_i$ denotes the fraction of samples in $\mathcal{B}_t$ routed to expert i, and $\boldsymbol{P}_i$ represents the average router-assigned probability for expert i w.r.t. samples in $\mathcal{B}_t$:

$$\boldsymbol{F}_i = \frac{1}{|\mathcal{B}_t|}\sum_{\boldsymbol{x}\in\mathcal{B}_t} \mathbb{I}_{\{\arg\max_k \boldsymbol{p}_k(\boldsymbol{x})=i\}}, \boldsymbol{P}_i = \frac{1}{|\mathcal{B}_t|}\sum_{\boldsymbol{x}\in\mathcal{B}_t} \boldsymbol{p}_i(\boldsymbol{x}), \tag{2}$$

where $\boldsymbol{p}(\boldsymbol{x})$ denotes the full routing probability vector produced by the router for the sample $\boldsymbol{x}$, while $\boldsymbol{p}_i(\boldsymbol{x})$ refers to its i-th component.

The aforementioned loss primarily encourages balanced expert utilization across samples, as it reaches its minimum value 1 when all elements in $\boldsymbol{F}$ are equal. A theoretical proof is provided in App. C.

However, in the test-time adaptation setting, balanced routing alone is insufficient. We also aim for the router to establish a consistent mapping that aligns inputs with appropriate experts based on shared adaptation patterns. In other words, it is desirable for samples with similar characteristics to be routed to the same expert, enabling each expert to adapt along distinct update directions rather than responding to a heterogeneous mixture of inputs.

To support such specialization, we adopt the following strategy: during the forward pass, only the expert corresponding to the highest routing probability $p := \arg\max_i \boldsymbol{p}_i$ is activated. To ensure that the router remains trainable under the entropy-based task loss, we apply a gradient-preserving trick by scaling the selected expert output with $p/\operatorname{stopgrad}(p)$. This operation does not affect the forward computation, but allows gradients to flow from the entropy loss into the router. As a result, routing decisions are optimized jointly with model predictions, encouraging experts to develop diverse and input-sensitive behaviors over time.

For the t-th batch $\mathcal{B}_t$, to maintain a balance between the load balancing loss $\mathcal{L}_{\text{load balancing}}$ and the entropy loss during adaptation, we scale $\mathcal{L}_{\text{load balancing}}$ using a dynamic weight

$$\alpha_t = \begin{cases} \lambda \times E^0_{\text{avg}}, & t=0,\\ \alpha_{t-1} \times \dfrac{E^t_{\text{avg}}}{E^{t-1}_{\text{avg}}}, & t\geq 1, \end{cases} \tag{3}$$

where λ is a hyper-parameter that controls the relative importance of the load balancing loss w.r.t. the entropy loss, and E^t_{avg} denotes the average entropy loss of all test samples up to the t-th batch.

3.3 Overall Procedure

Given an image $\boldsymbol{x} \in \mathbb{R}^{H\times W\times C}$, we divide this image into $N = HW/P^2$ patches $\{\boldsymbol{x}_p^i\}_{i=1}^N$, each with shape $P\times P\times C$. Next, we apply D filters with the same shape to each patch to create a patch embedding $\tilde{\boldsymbol{z}}_0^i \in \mathbb{R}^{D\times 1} (i=1,\dots,N)$. We then concatenate the patch embedding with the class embedding $\boldsymbol{x}_{\text{class}} = \tilde{\boldsymbol{z}}_0^0$. Finally, we add the position encoding $\boldsymbol{E}_{\text{pos}} \in \mathbb{R}^{D\times(N+1)}$ to obtain the final embedding at layer 0. With a L-layers ViT, the forward pass of MoETTA can be summarized as Eqs. (4)–(7), where MSA stands for Multi-head Self-Attention (Vaswani et al. 2017) and MLP stands for Multi-Layer Perceptron.

$$\boldsymbol{z}_0 = \left[\boldsymbol{x}_{\text{class}};\, \boldsymbol{x}_p^1\boldsymbol{E};\, \boldsymbol{x}_p^2\boldsymbol{E};\, \dots;\, \boldsymbol{x}_p^N\boldsymbol{E}\right] + \boldsymbol{E}_{\text{pos}}, \tag{4}$$

$$\boldsymbol{z}'_\ell = \mathrm{MSA}\big(\text{MoE-LayerNorm}(\boldsymbol{z}_{\ell-1})\big) + \boldsymbol{z}_{\ell-1}, \tag{5}$$

$$\boldsymbol{z}_\ell = \mathrm{MLP}\big(\text{MoE-LayerNorm}(\boldsymbol{z}'_\ell)\big) + \boldsymbol{z}'_\ell, \tag{6}$$

$$\boldsymbol{p}(y\,|\,\boldsymbol{x}) = \mathrm{Softmax}\left(\mathrm{MLP}\left(\text{MoE-LayerNorm}(\boldsymbol{z}_L^0)\right)\right), \tag{7}$$

where $\boldsymbol{E}\in\mathbb{R}^{D\times(P^2C)}$, $\boldsymbol{E}_{\text{pos}}\in\mathbb{R}^{D\times(N+1)}$ and $\ell=1,\dots,L$.

Inspired by Chen et al. (2024b) and Niu et al. (2022), we filter out samples with high entropy loss using a dynamic threshold and re-weight the remaining samples by their entropy loss. For the t-th batch $\mathcal{B}_t$, once the forward pass is finished, we calculate the sample-selection threshold

$$E^t_{\max} = \begin{cases} E^0_{\text{avg}}, & t=0,\\ E^{t-1}_{\max} \times \dfrac{E^t_{\text{avg}}}{E^{t-1}_{\text{avg}}}, & t\geq 1. \end{cases} \tag{8}$$

We then back-propagate the following multi-term loss:

$$\frac{1}{|\mathcal{S}_t|}\sum_{\boldsymbol{x}\in\mathcal{S}_t} \underbrace{\exp\left[\mathrm{E}_0 - \mathrm{Ent}(\boldsymbol{x})\right]}_{\text{Entropy re-weighting}} \underbrace{\mathrm{Ent}(\boldsymbol{x})}_{\text{Entropy loss}} + \alpha_t \sum_{i=1}^{M} \mathcal{L}^i_{\text{load balancing}}, \tag{9}$$

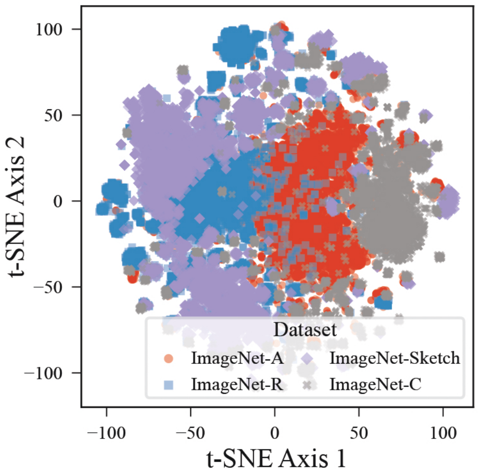

Figure 3. t-SNE (van der Maaten and Hinton 2008) projection of CLS token embeddings $\boldsymbol{z}_L^0$ in Eq. (7), from ViT-B/16 on 7,500 samples per dataset from ImageNet-A, ImageNet-R, ImageNet-Sketch, and ImageNet-C. While the classical mixed distribution shifts (ImageNet-C only) occupy a relatively narrow region (gray), our proposed potpourri benchmark introduces greater semantic and stylistic diversity by incorporating three additional variants. This results in a more heterogeneous and realistic evaluation setting for TTA.

where $\mathrm{Ent}(\boldsymbol{x})$ denotes the entropy of the posterior probability distribution $\boldsymbol{p}(y \mid \boldsymbol{x})$ in Eq. (7), α_t, defined by Eq. (3), is the trade-off coefficient that balances the entropy-minimization term against the load balancing loss, M is the number of MoE-LayerNorm and $\mathcal{L}^i_{\text{load balancing}}$ is the load balancing loss for the router corresponding to the i-th MoE-LayerNorm. The set of samples that participate in the entropy term is

$$\mathcal{S}_t = \left\{\boldsymbol{x} \mid \mathrm{Ent}(\boldsymbol{x}) < E^t_{\max} \wedge \boldsymbol{x} \in \mathcal{B}_t\right\}, \tag{10}$$

i.e., the "reliable" samples whose entropy falls below the current threshold. Given that there are only two degrees of freedom among the learning rate, λ in Eq. (3), and E_0 in Eq. (9), we fix E_0 as a constant and tune the others as hyperparameters throughout this work.

To clarify, among all symbols in Eqs. (1)–(10), only two of them need to be manually specified: λ and M, determined by the chosen MoE-LayerNorm replacement strategy. In Sec. 5.6, we further demonstrate MoETTA's robustness to λ and provide practical guidelines for the replacement strategy. Overall, MoETTA remains flexible yet easy to tune. The pseudocode of the entire procedure is presented in App. A.

4 *Potpourri* and *Potpourri*+ Benchmarks

Since Niu et al. (2023) highlighted the difficulty of addressing mixed distribution shifts and adopted a mixture of ImageNet-C as the evaluation setting, this benchmark has become the *de facto* standard for assessing the robustness of TTA methods under mixed-domain scenarios (Lee et al. 2024; Su et al. 2024; Deng et al. 2025; Tan et al. 2024). While ImageNet-C contains 15 corruption types grouped into four categories—noise, blur, weather, and digital—it still falls short in capturing the full diversity and severity of real-world distribution shifts. For example, natural renditions can cause substantial misalignment in texture and local image statistics (Hendrycks et al. 2020), and spurious background changes may appear even when the object of interest remains visually consistent (Wiles et al. 2022).

To close this gap, we propose the potpourri benchmark, which comprises a heterogeneous mixture of samples from ImageNet-C (Hendrycks and Dietterich 2019), ImageNet-R (Hendrycks et al. 2020), ImageNet-Sketch (Wang et al. 2019), and ImageNet-A (Hendrycks et al. 2021). ImageNet-R evaluates robustness against stylistic renditions, while ImageNet-A and ImageNet-Sketch probe vulnerability to spurious correlations and semantic abstraction. As shown in Fig. 3, these datasets complement ImageNet-C and together span a wider and more diverse set of domain shifts. This composition enables a more comprehensive and realistic evaluation of TTA methods.

In addition, test-time data streams may occasionally include samples from the source distribution. However, TTA models often suffer from catastrophic forgetting on such ID samples after being adapted to OOD samples (Niu et al. 2022). To address this, we introduce potpourri+, an extension of potpourri that includes validation samples from the original ImageNet dataset. It provides a more faithful assessment of a method's ability to balance adaptation and retention during mixed-domain deployment.

5 Experiments

5.1 Experimental Protocol

We evaluate our method under three settings: the classical mixed distribution shifts, and the proposed potpourri and potpourri+ benchmarks. Details of the related datasets are provided in App. D. All evaluation settings used in this section consist of ImageNet-C at corruption severity level 5, with a batch size of 64. In this work, all pre-trained models are obtained from timm (Wightman 2019) library. Unless otherwise stated, all experiments use ViT-B/16.

We compare MoETTA with the following state-of-the-art test-time adaptation methods: Tent (Wang et al. 2021), EATA (Niu et al. 2022), SAR (Niu et al. 2023), DeYO (Lee et al. 2024), CoTTA (Wang et al. 2022), MGTTA (Deng et al. 2025), and BECoTTA (Lee, Yoon, and Hwang 2024). Additional implementation details and baseline configurations are provided in App. E.

5.2 Robustness to Mixed Distribution Shifts

We evaluate MoETTA on three mixed distribution shifts benchmarks using both Transformer-based models (ViT (Dosovitskiy et al. 2021)) and convolution-based models (ConvNeXt (Liu et al. 2022)). As shown in Tab. 1, MoETTA achieves *state-of-the-art performance across all six settings*, demonstrating its effectiveness in improving the robustness of diverse architectures under mixed distribution shifts. To assess scalability, we further apply MoETTA to ViT-L/16 (304M parameters) in addition to ViT-B/16 (86M parameters); details are provided in App. I.

Model	Setting	Noadapt	Tent	EATA	CoTTA	SAR	DeYO	MGTTA	BECoTTA	Ours
ViT -B/16	Classical	55.52	$63.20_{0.08}$	$64.28_{0.09}$	$60.53_{0.57}$	$60.76_{0.04}$	$63.97_{0.04}$	$\underline{66.20}_{0.01}$	$61.57_{0.08}$	$\mathbf{67.20}_{0.03}$
	Pot.	54.18	$60.99_{0.05}$	$61.99_{0.11}$	$59.67_{1.21}$	$58.71_{0.03}$	$61.66_{0.02}$	$\underline{62.98}_{0.26}$	$59.08_{0.86}$	$\mathbf{65.12}_{0.08}$
	Pot.+	55.92	$62.28_{0.03}$	$63.17_{0.06}$	$59.26_{0.68}$	$59.99_{0.07}$	$62.90_{0.03}$	$\underline{64.35}_{0.07}$	$58.87_{3.23}$	$\mathbf{66.15}_{0.06}$
Conv. -B	Classical	54.81	$58.88_{0.06}$	$\underline{64.50}_{0.06}$	$59.65_{0.04}$	$61.67_{2.65}$	$64.32_{0.03}$	-	$50.16_{7.96}$	$\mathbf{67.40}_{0.02}$
	Pot.	53.91	$58.23_{0.05}$	$\underline{62.69}_{0.07}$	$58.57_{0.00}$	$61.16_{0.41}$	$62.46_{0.07}$	-	$28.28_{20.54}$	$\mathbf{65.70}_{0.05}$
	Pot.+	55.69	$59.69_{0.04}$	$\underline{63.94}_{0.07}$	$60.02_{0.02}$	$62.72_{0.10}$	$63.57_{0.06}$	-	$48.92_{9.35}$	$\mathbf{66.68}_{0.07}$

Table 1. Accuracy comparison (%, ↑) under different mixed distribution settings. Conv. stands for ConvNeXt. Noadapt refers to the evaluation of the model without adaptation. Results for MGTTA applied to ConvNeXt are not reported, since pre-training MGG for architectures with multiple normalization dimensions is non-trivial. For compactness, the notation a_b denotes $a \pm b$, where b represents the standard deviation calculated from the results of three different random seeds. The best performance is highlighted in **bold** and the second best is indicated by underlining. This convention is followed in all subsequent tables.

Method	#Act. params per sample	#Fwd	#Bwd	Used time
Noadapt	0	100%	0%	100%
Tent	0.04M	100%	100%	226%
EATA	0.04M	100%	80%	239%
SAR	0.03M	199%	175%	440%
DeYO	0.04M	196%	53%	317%
CoTTA	86.42M	199%	100%	798%
MGTTA	2.80M	100%	100%	227%
BECoTTA	0.13M	100%	86%	334%
Ours	0.23M	100%	76%	247%

Table 2. Computation efficiency comparison, showing per-sample activated parameter count; forward and backward propagation counts (as a percentage of total number of samples); and computation time *relative to the Noadapt baseline*. All statistics are measured under potpourri+ setting.

	Classical	Pot.	Pot.+	Avg.
Full method	**67.25**	**65.14**	**66.21**	**66.20**
Loss Components				
w/o Sample selection	<u>67.04</u>	<u>64.01</u>	57.61	<u>62.89</u>
w/o Entropy re-weight	62.86	60.51	<u>61.79</u>	61.72
w/o $\mathcal{L}_{\text{load balancing}}$	26.27	16.27	21.29	21.28
MoE Architecture				
w/o Grad to router	<u>65.17</u>	<u>62.80</u>	<u>63.92</u>	<u>63.96</u>
w/o Sample-wise router	28.69	28.60	24.96	27.42
w/o MoE-LayerNorm	22.38	17.94	26.93	22.42
w/o Layer-wise router	17.40	27.09	15.18	19.89

Table 3. Ablation analyses. The first group evaluates the effect of different loss components, and the second group evaluates the effect of MoE architectural choices. Accuracy (%, ↑) is reported under the classical mixed distribution shifts.

Furthermore, *MoETTA achieves these results without requiring any additional samples for pre-training or the calculation of statistical information*. In contrast, MGTTA, a strong baseline, relies on both extra OOD and ID samples.

5.3 Computation Efficiency

As shown in Tab. 2, MoETTA incurs modest overhead, its runtime is 247% of the model without adaptation, only slightly above the overheads of Tent, EATA, and MGTTA.

5.4 Ablation Analyses

Ablation Analyses on Loss Components. Removing either the sample selection mechanism, i.e., Eq. (10), or the entropy re-weighting part in Eq. (9) leads to moderate performance drops, showing their effectiveness in filtering reliable samples and emphasizing confident predictions. In contrast, removing the load balancing loss leads to a dramatic degradation across all settings, confirming its essential role in maintaining routing stability and preventing collapse.

Ablation Analyses on MoE Architecture. Disabling the gradient flow from the entropy loss to the router reduces the model's ability to learn informative routing strategies, resulting in a notable performance drop. Disabling sample-wise router causes the same expert to be used across all samples in a batch, which leads to severe performance collapse, highlighting the importance of input-adaptive routing. Removing layer-wise router enforces fixed expert assignment across all MoE-LayerNorm layers, limiting the model's flexibility and hurting performance. Eliminating the entire MoE-LayerNorm module results in unstable and ineffective adaptation, confirming its necessity.

5.5 Analysis of Expert Parameter Diversity

Fig. 4 shows that, during adaptation, experts' parameters within the same MoE-LayerNorm diverge over time. This effect is particularly pronounced in shallow layers, reflected by the decreasing cosine similarity in Fig. 4(a-b). These results suggest that experts learn distinct parameter update trajectories, rather than converging to similar representations.

Importantly, such diversity is not pre-imposed but emerges naturally from our design: experts are structurally decoupled and trained with separate gradient signals via the routing mechanism. The final-state statistics in Fig. 4(c) confirm this decoupling, with many layers exhibiting low simi-

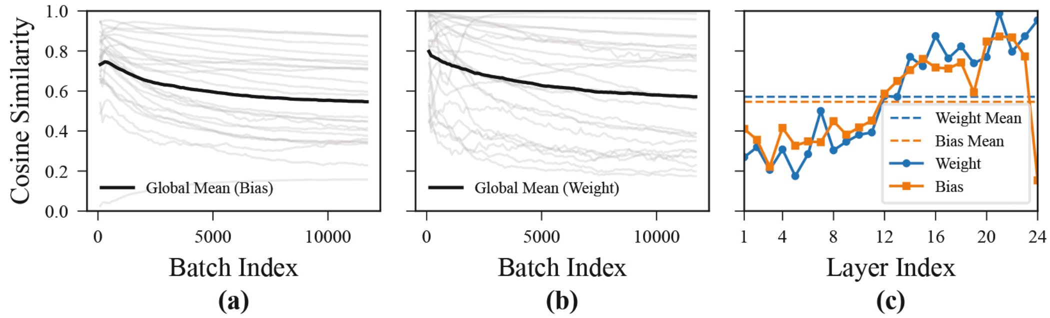

Figure 4. Evolution and final-state statistics of expert parameter similarity across MoE-LayerNorms adapted under the classical mixed distribution shifts. (a–b) track the average pairwise cosine similarity of expert bias and weight parameters during adaptation, with each gray curve representing one MoE-LayerNorm and the bold black line denoting the global mean across layers. (c) reports the final average similarity for each layer at the end of adaptation phase.

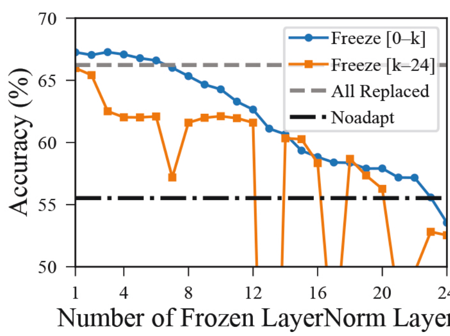

Figure 5. Effect of partially replacing LayerNorm with MoE-LayerNorm on accuracy.

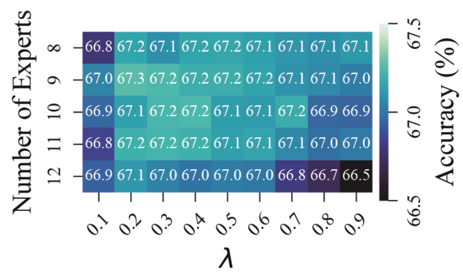

Figure 6. Grid search results over the number of experts and the trade-off coefficient λ in Eq. (3).

larity among their experts.

This parameter-level diversity allows MoETTA to internalize multiple adaptation modes within a single model, enabling it to better accommodate heterogeneous input batches without explicit domain labels. A layer-wise visualization of expert similarity is included in App. G.

5.6 Hyper-Parameter Sensitivity

In this subsection, all experiments are conducted under the classical mixed distribution shifts.

Effect of Replacement Strategy. As shown in Fig. 5, we compare two partial replacement strategies for MoE-LayerNorm: (1) freezing shallow LayerNorm layers indexed 0–k, and (2) freezing deeper layers indexed k—24. Freezing only the shallow layers (indexed 0–k, $k = 0, \ldots, 5$) consistently outperforms full replacement. This is consistent with prior findings that shallow ViT layers primarily encode low-level, domain-invariant features such as color (Dorszewski et al. 2025), which benefit from being preserved. In contrast, allowing deeper layers—responsible for high-level semantics—to adapt leads to more effective test-time adaptation. Freezing middle layers, which are known to be critical for fine-tuning (Valeriani et al. 2023), results in a sharp performance drop. Similarly, freezing only the deeper layers hinders the model's ability to adjust to distribution shifts, often leading to degraded performance or even collapse.

Effect of the Number of Experts and λ in Eq. (3). As shown in Fig. 6, MoETTA performs best when the trade-off coefficient λ is small, suggesting that excessive focus on load balancing can impede effective adaptation. This underscores the value of letting the router flexibly assign experts based on input characteristics rather than enforcing uniform expert usage. The number of experts also affects performance. While MoETTA is robust across a range of choices, using 9 experts achieves the best accuracy under the classical mixed distribution shifts.

6 Conclusion

We study TTA under mixed distribution shifts, where a single update direction across diverse samples often leads to suboptimal adaptation. For more realistic evaluation, we introduce two benchmarks, potpourri and potpourri+, combining multiple shift types within a single test stream.

To overcome the limitations of conventional TTA, we propose MoETTA, a lightweight framework that reparameterizes LayerNorm into a mixture of decoupled expert branches. Allowing experts to follow distinct adaptation trajectories lets MoETTA capture multiple modes of test-time behavior. Empirically, this expert diversity improves robustness and stability under complex mixed distribution shifts, consistently outperforming prior TTA methods.

Acknowledgements

This study was conducted entirely during Xiao Fan's internship at Tsinghua University and was supported by the National Key Research and Development Program of China (Grant No. 2023YFF0905502), the National Natural Science Foundation of China (Grant Nos. 92467204 and 62472249), the Shenzhen Science and Technology Program (Grant Nos. JCYJ20220818101014030, KJZD20240903102300001, and JCYJ20250604145014018), and the Natural Science Foundation for Top Talents of Shenzhen Technology University (Grant No. GDRC202413).

References

Ba, J. L.; Kiros, J. R.; and Hinton, G. E. 2016. Layer Normalization. *arXiv preprint arXiv:1607.06450*.

Ben-David, S.; Blitzer, J.; Crammer, K.; Kulesza, A.; Pereira, F.; and Vaughan, J. 2010. A theory of learning from different domains. *Machine Learning*, 79: 151–175.

Cang, Y.; Chen, M.; and Huang, K. 2024. Joint Batching and Scheduling for High-Throughput Multiuser Edge AI With Asynchronous Task Arrivals. *IEEE Transactions on Wireless Communications*, 23(10): 13782–13795.

Chen, G.; Niu, S.; Chen, D.; Zhang, S.; Li, C.; Li, Y.; and Tan, M. 2024a. Cross-device collaborative test-time adaptation. In *Proceedings of the 38th International Conference on Neural Information Processing Systems*, NIPS '24. Red Hook, NY, USA: Curran Associates Inc. ISBN 9798331314385.

Chen, Y.; Niu, S.; Xu, S.; Song, H.; Wang, Y.; and Tan, M. 2024b. Towards Robust and Efficient Cloud-Edge Elastic Model Adaptation via Selective Entropy Distillation. In *International Conference on Learning Representations*.

Dai, D.; Deng, C.; Zhao, C.; Xu, R.; Gao, H.; Chen, D.; Li, J.; Zeng, W.; Yu, X.; Wu, Y.; Xie, Z.; Li, Y.; Huang, P.; Luo, F.; Ruan, C.; Sui, Z.; and Liang, W. 2024. DeepSeekMoE: Towards Ultimate Expert Specialization in Mixture-of-Experts Language Models. In Ku, L.-W.; Martins, A.; and Srikumar, V., eds., *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 1280–1297. Bangkok, Thailand: Association for Computational Linguistics.

DeepSeek-AI. 2024. DeepSeek-V3 Technical Report. arXiv:2412.19437.

Deng, Q.; Niu, S.; Zhang, R.; Chen, Y.; Zeng, R.; Chen, J.; and Hu, X. 2025. Learning to Generate Gradients for Test-Time Adaptation via Test-Time Training Layers. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 16235–16243.

Dorszewski, T.; Tětková, L.; Jenssen, R.; Hansen, L. K.; and Wickstrøm, K. K. 2025. From colors to classes: Emergence of concepts in vision transformers. In *World Conference on Explainable Artificial Intelligence*, 28–47. Springer.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; Uszkoreit, J.; and Houlsby, N. 2021. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. *ICLR*.

Eigen, D.; Ranzato, M.; and Sutskever, I. 2014. Learning Factored Representations in a Deep Mixture of Experts. arXiv:1312.4314.

Fedus, W.; Zoph, B.; and Shazeer, N. 2022. Switch transformers: scaling to trillion parameter models with simple and efficient sparsity. *J. Mach. Learn. Res.*, 23(1).

Foret, P.; Kleiner, A.; Mobahi, H.; and Neyshabur, B. 2021. Sharpness-aware Minimization for Efficiently Improving Generalization. In *International Conference on Learning Representations*.

Gidaris, S.; Singh, P.; and Komodakis, N. 2018. Unsupervised Representation Learning by Predicting Image Rotations. In *International Conference on Learning Representations*.

Glorot, X.; and Bengio, Y. 2010. Understanding the difficulty of training deep feedforward neural networks. In Teh, Y. W.; and Titterington, M., eds., *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, volume 9 of *Proceedings of Machine Learning Research*, 249–256. Chia Laguna Resort, Sardinia, Italy: PMLR.

Grandvalet, Y.; and Bengio, Y. 2004. Semi-supervised Learning by Entropy Minimization. In *Advances in Neural Information Processing Systems 17 [Neural Information Processing Systems, NIPS 2004, December 13-18, 2004, Vancouver, British Columbia, Canada]*, 529–536.

Hendrycks, D.; Basart, S.; Mu, N.; Kadavath, S.; Wang, F.; Dorundo, E.; Desai, R.; Zhu, T. L.; Parajuli, S.; Guo, M.; Song, D. X.; Steinhardt, J.; and Gilmer, J. 2020. The Many Faces of Robustness: A Critical Analysis of Out-of-Distribution Generalization. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, 8320–8329.

Hendrycks, D.; and Dietterich, T. 2019. Benchmarking Neural Network Robustness to Common Corruptions and Perturbations. In *International Conference on Learning Representations*.

Hendrycks, D.; Zhao, K.; Basart, S.; Steinhardt, J.; and Song, D. 2021. Natural adversarial examples. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 15262–15271.

Hu, E. J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; Chen, W.; et al. 2022. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2): 3.

Jacobs, R. A.; Jordan, M. I.; Nowlan, S. J.; and Hinton, G. E. 1991. Adaptive mixtures of local experts. *Neural computation*, 3(1): 79–87.

Jiang, A. Q.; Sablayrolles, A.; Roux, A.; Mensch, A.; Savary, B.; Bamford, C.; Chaplot, D. S.; de las Casas, D.; Hanna, E. B.; Bressand, F.; Lengyel, G.; Bour, G.; Lample, G.; Lavaud, L. R.; Saulnier, L.; Lachaux, M.-A.; Stock, P.; Subramanian, S.; Yang, S.; Antoniak, S.; Scao, T. L.; Gervet, T.; Lavril, T.; Wang, T.; Lacroix, T.; and Sayed, W. E. 2024. Mixtral of Experts. arXiv:2401.04088.

Jordan, M. I.; and Jacobs, R. A. 1994. Hierarchical mixtures of experts and the EM algorithm. *Neural computation*, 6(2): 181–214.

Lee, D.; Yoon, J.; and Hwang, S. J. 2024. BECoTTA: input-dependent online blending of experts for continual test-time adaptation. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org.

Lee, J.; Jung, D.; Lee, S.; Park, J.; Shin, J.; Hwang, U.; and Yoon, S. 2024. Entropy is not Enough for Test-Time Adaptation: From the Perspective of Disentangled Factors. In *The Twelfth International Conference on Learning Representations*.

Liu, Y.; Kothari, P.; van Delft, B.; Bellot-Gurlet, B.; Mordan, T.; and Alahi, A. 2021. TTT++: When Does Self-Supervised Test-Time Training Fail or Thrive? In Ranzato, M.; Beygelzimer, A.; Dauphin, Y.; Liang, P.; and Vaughan, J. W., eds., *Advances in Neural Information Processing Systems*, volume 34, 21808–21820. Curran Associates, Inc.

Liu, Z.; Mao, H.; Wu, C.-Y.; Feichtenhofer, C.; Darrell, T.; and Xie, S. 2022. A convnet for the 2020s. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 11976–11986.

Niu, S.; Wu, J.; Zhang, Y.; Chen, Y.; Zheng, S.; Zhao, P.; and Tan, M. 2022. Efficient test-time model adaptation without forgetting. In *International conference on machine learning*, 16888–16905. PMLR.

Niu, S.; Wu, J.; Zhang, Y.; Wen, Z.; Chen, Y.; Zhao, P.; and Tan, M. 2023. Towards Stable Test-Time Adaptation in Dynamic Wild World. In *International Conference on Learning Representations*.

Paszke, A.; Gross, S.; Massa, F.; Lerer, A.; Bradbury, J.; Chanan, G.; Killeen, T.; Lin, Z.; Gimelshein, N.; Antiga, L.; Desmaison, A.; Köpf, A.; Yang, E.; DeVito, Z.; Raison, M.; Tejani, A.; Chilamkurthy, S.; Steiner, B.; Fang, L.; Bai, J.; and Chintala, S. 2019. *PyTorch: an imperative style, high-performance deep learning library*, 12. Red Hook, NY, USA: Curran Associates Inc.

Shazeer, N.; Mirhoseini, A.; Maziarz, K.; Davis, A.; Le, Q.; Hinton, G.; and Dean, J. 2017a. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. *arXiv preprint arXiv:1701.06538*.

Shazeer, N.; Mirhoseini, A.; Maziarz, K.; Davis, A.; Le, Q.; Hinton, G.; and Dean, J. 2017b. Outrageously Large Neural Networks: The Sparsely-Gated Mixture-of-Experts Layer. In *International Conference on Learning Representations*.

Su, Z.; Guo, J.; Yao, K.; Yang, X.; Wang, Q.; and Huang, K. 2024. Unraveling Batch Normalization for Realistic Test-Time Adaptation. In Wooldridge, M. J.; Dy, J. G.; and Natarajan, S., eds., *Thirty-Eighth AAAI Conference on Artificial Intelligence, AAAI 2024, Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence, IAAI 2024, Fourteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2014, February 20-27, 2024, Vancouver, Canada*, 15136–15144. AAAI Press.

Tan, M.; Chen, G.; Wu, J.; Zhang, Y.; Chen, Y.; Zhao, P.; and Niu, S. 2024. Uncertainty-Calibrated Test-Time Model Adaptation Without Forgetting. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47: 6274–6289.

Valeriani, L.; Doimo, D.; Cuturello, F.; Laio, A.; Ansuini, A.; and Cazzaniga, A. 2023. The geometry of hidden representations of large transformer models. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23. Red Hook, NY, USA: Curran Associates Inc.

van der Maaten, L.; and Hinton, G. E. 2008. Visualizing Data using t-SNE. *Journal of Machine Learning Research*, 9: 2579–2605.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, L.; and Polosukhin, I. 2017. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, 6000–6010. Red Hook, NY, USA: Curran Associates Inc. ISBN 9781510860964.

Wang, D.; Shelhamer, E.; Liu, S.; Olshausen, B.; and Darrell, T. 2021. Tent: Fully Test-Time Adaptation by Entropy Minimization. In *International Conference on Learning Representations*.

Wang, H.; Ge, S.; Xing, E. P.; and Lipton, Z. C. 2019. Learning robust global representations by penalizing local predictive power. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems*. Red Hook, NY, USA: Curran Associates Inc.

Wang, Q.; Fink, O.; Van Gool, L.; and Dai, D. 2022. Continual Test-Time Domain Adaptation. In *Proceedings of Conference on Computer Vision and Pattern Recognition*.

Wang, Z.; Luo, Y.; Zheng, L.; Chen, Z.; Wang, S.; and Huang, Z. 2024. In Search of Lost Online Test-Time Adaptation: A Survey. *International Journal of Computer Vision*, 133: 1106–1139.

Wightman, R. 2019. PyTorch Image Models. https://github.com/rwightman/pytorch-image-models.

Wiles, O.; Gowal, S.; Stimberg, F.; Rebuffi, S.-A.; Ktena, I.; Dvijotham, K. D.; and Cemgil, A. T. 2022. A Fine-Grained Analysis on Distribution Shift. In *International Conference on Learning Representations*.