

Shared & Domain Self-Adaptive Experts with Frequency-Aware Discrimination for Continual Test-Time Adaptation

Jianchao Zhao¹, Chenhao Ding³, SongLin Dong^{2*}, Jiangyang Li¹,
Qiang Wang¹, Yuhang He¹, Yihong Gong^{1,2}

¹ National Key Laboratory of Human-Machine Hybrid Augmented Intelligence,
National Engineering Research Center for Visual Information and Applications,
Institute of Artificial Intelligence and Robotics, Xi'an Jiaotong University

² Faculty of Computility Microelectronics, Shenzhen University of Advanced Technology

³ School of Software Engineering, Xi'an Jiaotong University

{zhaojianchao, dch225739, ljy1021,qwang}@stu.xjtu.edu.cn, dongsl@suat-sz.edu.cn,
heyuhang@xjtu.edu.cn, ygong@mail.xjtu.edu.cn

Abstract

This paper focuses on the Continual Test-Time Adaptation (CTTA) task, aiming to enable an agent to continuously adapt to evolving target domains while retaining previously acquired domain knowledge for effective **reuse** when those domains **reappear**. Existing shared-parameter paradigms struggle to balance adaptation and forgetting, leading to decreased efficiency and stability. To address this, we propose a frequency-aware shared and self-adaptive expert framework, consisting of two key components: (i) a dual-branch expert architecture that extracts general features and dynamically models domain-specific representations, effectively reducing cross-domain interference and repetitive learning cost; and (ii) an online Frequency-aware Domain Discriminator (FDD), which leverages the robustness of low-frequency image signals for online domain shift detection, guiding dynamic allocation of expert resources for more stable and realistic adaptation. Additionally, we introduce a Continual Repeated Shifts (CRS) benchmark to simulate periodic domain changes for more realistic evaluation. Experimental results show that our method consistently outperforms existing approaches on both classification and segmentation CTTA tasks under standard and CRS settings, with ablations and visualizations confirming its effectiveness and robustness.

1 Introduction

Test-Time Adaptation (TTA) (Liang, He, and Tan 2025; Han et al. 2025) enables pretrained models to adapt to unknown data distributions during inference, demonstrating significant potential for practical applications (Dong et al. 2024b; Gong and Wang 2024; Peng et al. 2024, 2025; Wang et al. 2024). Traditional TTA methods primarily focus on static target domains that assume a fixed distribution. However, in real-world scenarios such as autonomous driving, models encounter environmental variations caused by changes in weather, sensors, or location over time, requiring agents to adapt to these evolving conditions continuously. Motivated by this, Continual Test-Time Adaptation (CTTA) (Wang

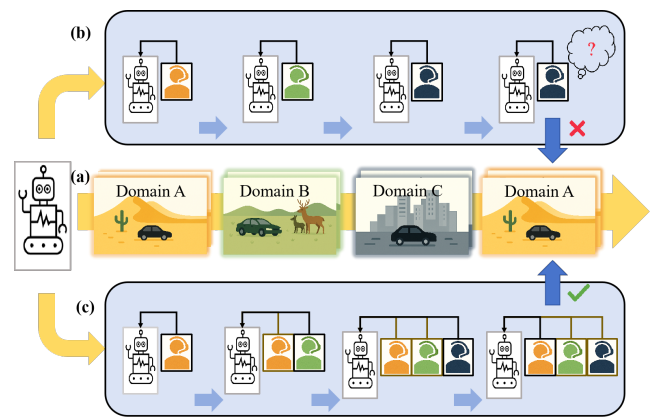


Figure 1: A New Benchmark and Comparison of Methods for CTTA. (a) Our proposed CRS (Continual Repeated Shifts) benchmark simulates realistic, cyclic domain shifts at test time. (b) Existing shared-parameter methods struggle to balance adaptation and memory, leading to forgetting and redundant relearning when domains recur. (c) Our method adds domain self-adaptive experts, enabling knowledge retention and fast adaptation to recurring domains.

et al. 2022a), which aligns closely with practical demands, has attracted increasing attention by enabling models to effectively handle a sequence of continually changing unknown distributions.

Existing mainstream CTTA methods are divided into two categories: (i) model-based paradigm: this paradigm (Döbler, Marsden, and Yang 2023; Liu et al. 2024) updates the entire or partial modules of the model during testing by generating pseudo-labels and leveraging consistency or self-supervised objectives to achieve online adaptation to new domains; (ii) parameter-tuning based: this paradigm (Liu et al. 2023; Gan et al. 2023; Song et al. 2023) is based on parameter-efficient fine-tuning (PEFT) strategies (Hu et al. 2021; Ding et al. 2025), which introduce learnable prompts or LoRA modules into the model to adapt to continuously changing target domain knowledge. Meth-

*Corresponding author.

ods of paradigm (i) usually incur high computational costs and lack fine-grained modeling of domain discrepancies, leading to low adaptation efficiency or unstable performance prone to forgetting during continual adaptation. Therefore, we focus on an in-depth study of paradigm (ii).

Paradigm (ii) methods typically achieve continual adaptation to unseen domains by relying on shared tuning parameters, which have inherent limitations. Consider an autonomous-driving vision model (Fig. 1) which must continually recognize and interpret road scenes: over time it traverses recurring environments—deserts, grasslands, urban areas. Ideally, the model should adapt rapidly to new environments while retaining domain-specific knowledge to reuse when the same environment reappears. However, these methods often suffer from forgetting prior domain knowledge while adapting to new domains, leading to an inherent trade-off between adaptation and memory retention (Dong et al. 2024a, 2025; Wang et al. 2025). When prior domains reappear, the model is forced to relearn what it has already forgotten, hurting efficiency and stability. Although some approaches (Liu et al. 2023; Gan et al. 2023) attempt to disentangle domain-shared and domain-specific features within the shared parameter space, the training process still relies on a unified set of parameters. As a result, the underlying conflict between adaptation and memory retention remains unresolved.

To address this, we propose a frequency-discriminative shared & self-adaptive domain expert model. It has two key components: (i) a dual-branch expert architecture; and (ii) an online frequency-aware domain discriminator (FDD). Specifically, the first module comprises a shared expert and domain self-adaptive experts. The shared expert continuously adapts to extract task-relevant general features, ensuring stable transfer of discriminative knowledge across domains, while domain self-adaptive experts are dynamically introduced upon detecting new domains to model domain-specific features, mitigating cross-domain interference and reducing redundant learning. To achieve accurate domain shift detection and robust expansion, the FDD module leverages low-frequency image signals for real-time domain drift perception. It adaptively allocates the most suitable expert resources for new domains while preserving existing domain knowledge, enabling more realistic and stable continuous test-time adaptation.

Furthermore, current CTTA experiments learn each domain only once, ignoring realistic repeated shifts (e.g., rain→sun→rain). We therefore propose a more realistic benchmark, Continual Repeated Shifts (CRS), constructed by cyclically combining four ImageNet-derived datasets to simulate periodic domain repetitions (see Fig.1 (b)), with extended versions ImageNet+ and ImageNet++. Compared to conventional CTTA tasks, CRS shows larger domain discrepancies and periodic distribution shifts, more faithfully replicating dynamic real-world environments.

The key contributions are as follows: we identify the limitations of existing paradigms in managing the trade-off between forgetting and adaptation, and propose a domain shared & self-adaptive expert with frequency-aware discrimination to address this challenge. We design a realistic CTTA

benchmark, Continual Repeated Shifts (CRS), which captures the recurrent nature of domain shifts in real-world scenarios. Extensive experiments on classification (general and CRS settings) and semantic segmentation demonstrate that our method consistently outperforms others.

2 Related Work

Parameter-Efficient Fine-Tuning. Parameter-efficient fine-tuning (PEFT) methods adapt large pre-trained models to downstream tasks with few trainable parameters and low compute. Representative approaches include Adapters (Houlsby et al. 2019), Low-Rank Adaptation (LoRA) (Hu et al. 2021), and Prompt Tuning (Jia et al. 2022). Owing to its plug-and-play insertion and tiny memory footprint, LoRA is widely adopted in practice. Building on LoRA, Mixture-of-Experts (MoE) (Shazeer et al. 2017; Wang et al. 2022b) splits the low-rank update into multiple experts and employs a gating mechanism to activate only a few per input, preserving scalability while reducing inference FLOPs. Recent work includes DA-MoE (Aghdam et al. 2024), which allocates a variable number of experts to each token based on salience for fine-grained compute–confidence trade-offs, and DeepSeekMoE (Zhang et al. 2024), which introduces fine-grained expert partitioning with shared sub-experts to cut redundancy and boost specialization. Inspired by DeepSeekMoE, we propose a hybrid architecture combining shared and self-adaptive experts, well-suited to tasks with complex, continuous domain drift.

Continual Test-Time Adaptation. Continual Test-Time Adaptation (CTTA) extends the TTA paradigm by addressing distribution shifts across a sequence of unseen target domains. CoTTA (Wang et al. 2022a) first addresses this task by introducing a teacher-averaged pseudo-labeling strategy and stochastic parameter restoration. Subsequent methods build on this foundation. RMT (Döbler, Marsden, and Yang 2023) replaces standard cross-entropy with symmetric cross-entropy to improve gradient stability and combines it with contrastive learning to preserve feature alignment with the source. EcoTTA (Song et al. 2023) addresses memory efficiency by introducing meta-networks that adapt lightweight layers and employ self-distillation to preserve source knowledge. BECoTTA (Lee, Yoon, and Hwang 2024) proposes a mixture-of-experts framework that leverages domain-adaptive routing to minimize parameter updates while maintaining adaptation performance. Meanwhile, VDP (Gan et al. 2023) and VIDA (Liu et al. 2023) attempt to decouple task-relevant and domain-specific knowledge through prompt-based or adapter-based mechanisms. Despite these advances, existing methods still struggle to strike an optimal balance between rapid adaptation and forgetting mitigation due to shared parameters, and also struggle to cope with more realistic and complex scenarios.

3 Method

3.1 Preliminary

Problem Formulation. In Continual Test-Time Adaptation (CTTA), the model $q_\theta(y|x)$ is first pre-trained on a

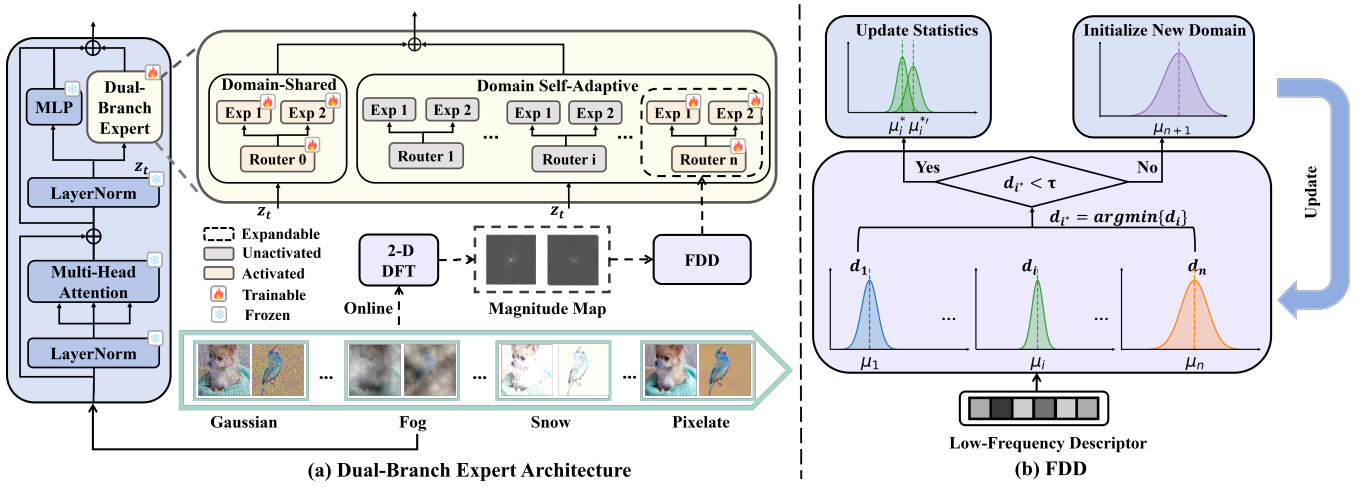


Figure 2: Overall framework. (a) The Dual-Branch Expert Architecture combines a shared expert with a dynamically expandable set of domain self-adaptive experts. The appropriate domain self-adaptive expert is selected based on input, guided by the online Frequency-aware Domain Discriminator (FDD). (b) FDD detects domain shifts using low-frequency features from 2D Fourier-transformed inputs, updating domain statistics or initializing new experts when unseen domains are encountered.

labeled source domain $\mathcal{D}_S = \{(x_s, y_s)\}$. After deployment, the model is adapted to a sequence of target domains $\{\mathcal{D}_{T_i}\}_{i=1}^n$, where n denotes the number of continual shifts, and the distributions of the target domains $\mathcal{D}_{T_1}, \mathcal{D}_{T_2}, \dots, \mathcal{D}_{T_n}$ evolve over time. The CTTA protocol follows three key assumptions: (1) access to the source domain \mathcal{D}_S is strictly prohibited after deployment, (2) each target domain sample $x \in \mathcal{D}_T$ can be observed only once during adaptation, and (3) no ground-truth labels are available for the target domains. Under these constraints, the objective of CTTA is to adapt the pre-trained model q_θ to evolving target domains while maintaining performance and preserving recognition ability on previously seen distributions.

Mixture-of-Experts. A MoE module consists of a router and M experts E_1, \dots, E_M . Given an input $\mathbf{z} \in \mathbb{R}^D$ (where D is the feature dimension), the router computes gating scores $\mathbf{g} = W_r \mathbf{z} + \mathbf{b}_r$ using learnable parameters $W_r \in \mathbb{R}^{M \times D}$ and $\mathbf{b}_r \in \mathbb{R}^M$, and normalizes them via softmax to obtain mixture weights $\alpha \in \mathbb{R}^M$.

Each expert E_i is a two-layer low-rank network:

$$E_i(\mathbf{z}) = \sigma(\mathbf{z}W_i^{\text{down}})W_i^{\text{up}}, \quad (1)$$

where $W_i^{\text{down}} \in \mathbb{R}^{D \times r}$ and $W_i^{\text{up}} \in \mathbb{R}^{r \times D}$ are learnable projection matrices, and $r \ll D$ is the bottleneck dimension. The activation function $\sigma(\cdot)$ is typically ReLU or GELU.

The final MoE output is a weighted sum of expert outputs:

$$\text{MoE}(\mathbf{z}) = \sum_{i=1}^M \alpha_i \cdot E_i(\mathbf{z}). \quad (2)$$

3.2 Dual-Branch Expert Architecture

As illustrated in Fig. 2, our dual-branch architecture comprises (i) a *domain-shared expert branch* that encodes

domain-invariant task knowledge and (ii) a *domain self-adaptive expert branch* that grows on demand to capture domain-specific shifts.

Domain-Shared Expert Branch. Given the input token sequence $\mathbf{Z} \in \mathbb{R}^{B \times N \times D}$, where B is the batch size, N is the number of tokens per sample. we feed it into a fixed Mixture-of-Experts (MoE) layer shared across all domains:

$$\mathbf{Y}_{\text{shared}} = \text{MoE}_{\text{shared}}(\mathbf{Z}). \quad (3)$$

This branch facilitates knowledge transfer and supports generalization by preventing the overwriting of task-critical representations during continual adaptation.

Domain Self-Adaptive Expert Branch. To handle domain-specific variations, we introduce an expandable pool of MoE modules in domain self-adaptive expert branch. Input token sequence \mathbf{Z} is processed by one of the domain self-adaptive expert modules: $\{\text{MoE}_{\text{domain}}^{(1)}, \text{MoE}_{\text{domain}}^{(2)}, \dots, \text{MoE}_{\text{domain}}^{(n)}\}$.

Each expert module $\text{MoE}_{\text{domain}}^{(i)}$ is responsible for modeling a previously observed domain and consists of a router $G^{(i)}$ and a set of lightweight experts $\{E_1^{(i)}, \dots, E_M^{(i)}\}$. The expert module is dynamically selected based on the current input using the online Frequency-Aware Domain Discriminator (FDD, see Section 3.3):

$$\mathbf{Y}_{\text{domain}} = \text{MoE}_{\text{domain}}^{(i^*)}(\mathbf{Z}), \quad (4)$$

where i^* denotes the selected expert module. If the current input distribution does not match any existing module, a new module $\text{MoE}_{\text{domain}}^{(n+1)}$ is instantiated and added to the pool.

Fusion of Dual Outputs. To integrate both generalizable and domain-adaptive knowledge, we combine the outputs of the two branches using a residual fusion mechanism:

$$\mathbf{Z}_{\text{out}} = \mathbf{Z} + \lambda \cdot \mathbf{Y}_{\text{shared}} + (1 - \lambda) \cdot \mathbf{Y}_{\text{domain}}, \quad (5)$$

where $\lambda \in [0, 1]$ is a hyperparameter controls the balance between domain-shared and domain-specific representations.

3.3 Online Frequency-Aware Domain Discriminator

Efficient domain-shift detection is critical to our CTTA framework; therefore we introduce online Frequency-Aware Domain Discriminator (FDD), a lightweight and training-free module that leverages frequency statistics to detect shifts and route inputs to the appropriate domain self-adaptive expert.

Bayesian Domain Posterior Estimation and Decision Rule. Given an input image x , we obtain its low-frequency descriptor $z = f(x) \in \mathbb{R}^{d_f}$, where $f(\cdot)$ is the low-frequency feature extractor detailed at the end of this subsection, and d_f denotes the dimension of the resulting frequency descriptor. Suppose the model has so far identified K distinct domains during adaptation, let $y \in \{1, \dots, K\}$ denote the latent domain label. Inspired by classical Gaussian discriminant analysis (Hastie and Tibshirani 1996), we assume that the low-frequency embeddings corresponding to each domain follow a multivariate Gaussian distribution:

$$p(z | y = i) = \mathcal{N}(z | \mu_i, \Sigma_i), \quad (6)$$

where $\mu_i \in \mathbb{R}^{d_f}$ and $\Sigma_i \in \mathbb{R}^{d_f \times d_f}$ are the estimated mean and covariance of domain i , respectively. We further assume a uniform prior over domains, i.e., $P(y = i) = 1/K$. Then, the posterior probability that a sample belongs to domain i is given by Bayes' rule:

$$\begin{aligned} P(y = i | z) &= \frac{p(z | y = i)}{\sum_{j=1}^K p(z | y = j)} \\ &= \frac{\exp(-\frac{1}{2}m_i(z))/\sqrt{\det \Sigma_i}}{\sum_{j=1}^K \exp(-\frac{1}{2}m_j(z))/\sqrt{\det \Sigma_j}}, \end{aligned} \quad (7)$$

where $m_i(z)$ denotes the Mahalanobis distance under a shrinkage-regularized covariance, and $\det \Sigma$ is the determinant of the covariance matrix. More detail can be found in Appendix B.

To assign a domain label to an incoming batch of B samples $\{x_b\}_{b=1}^B$, we compute the average embedding:

$$\bar{z}_t^{(0)} = \frac{1}{B} \sum_{b=1}^B f(x_b), \quad (8)$$

and select the domain with the highest posterior probability, which is equivalent (see Appendix C) to choosing the domain with the smallest Mahalanobis distance:

$$i^* = \arg \min_i m_i(\bar{z}_t^{(0)}). \quad (9)$$

If the minimum distance $\min_i m_i(\bar{z}_t^{(0)})$ exceeds a pre-defined threshold τ , we infer that the current batch likely originates from a previously unseen domain, and a new domain self-adaptive expert module is initialized. Otherwise, the batch is assigned to the expert module corresponding to the closest known domain i^* .

Robust Online Update of Domain Statistics. In CTTA, samples arrive sequentially, making it impossible to observe the full distribution of any domain at once. To address this, we adopt an incremental estimation strategy that is both memory-efficient and noise-robust, enabling online updates of domain statistics based on the current batch.

To this end, we adopt a soft-assignment strategy, assigning each sample x_b a normalized responsibility weight computed from its Gaussian probability density under the selected domain's distribution:

$$w_b = \frac{\exp[-\frac{1}{2}m_{i^*}(f(x_b))]}{\sum_{j=1}^B \exp[-\frac{1}{2}m_{i^*}(f(x_j))]}, \quad \sum_{b=1}^B w_b = 1. \quad (10)$$

These normalized weights effectively down-weight outlier samples with poor domain fit, acting as a robust mechanism to stabilize parameter updates.

Assuming the current cumulative weight (or effective batch count) for domain i^* is c_{i^*} , and its corresponding sample statistics are summarized by the mean μ_{i^*} and covariance Σ_{i^*} . we define the weighted complete-data log-likelihood for the batch as:

$$\mathcal{L}(\mu, \Sigma) = c_{i^*} \log \mathcal{N}(\mu_{i^*} | \mu, \Sigma) + \sum_{b=1}^B w_b \log \mathcal{N}(z_b | \mu, \Sigma), \quad (11)$$

where $z_b = f(x_b)$ is the low-frequency feature of sample x_b . Maximizing \mathcal{L} with respect to μ and Σ yields the following closed-form updates (see Appendix D) for detailed derivation):

$$\mu_{i^*}^{\text{new}} = \frac{c_{i^*} \mu_{i^*} + \sum_{b=1}^B w_b f(x_b)}{c_{i^*} + 1}, \quad (12)$$

$$\Sigma_{i^*}^{\text{new}} = \frac{c_{i^*} \Sigma_{i^*} + \sum_{b=1}^B w_b (f(x_b) - \mu_{i^*})(f(x_b) - \mu_{i^*})^\top}{c_{i^*} + 1}. \quad (13)$$

Finally, the domain's sufficient statistics are updated as:

$$c_{i^*} \leftarrow c_{i^*} + 1, \quad \mu_{i^*} \leftarrow \mu_{i^*}^{\text{new}}, \quad \Sigma_{i^*} \leftarrow \Sigma_{i^*}^{\text{new}}. \quad (14)$$

When a new domain is detected (i.e., no existing domain achieves sufficiently high posterior confidence), we initialize the new domain's statistics from the current batch. Specifically, we use the average embedding $\bar{z}_t^{(0)}$ as the initial mean, and a diagonal covariance scaled by a constant variance σ_0^2 :

$$c_{\text{new}} = 1, \quad \mu_{\text{new}} = \bar{z}_t^{(0)}, \quad \Sigma_{\text{new}} = \sigma_0^2 I. \quad (15)$$

This corresponds to the maximum a posteriori estimate under a Gaussian–Normal–Inverse–Wishart prior with a single observation, ensuring a well-posed initialization for the new domain.

Low-Frequency Feature Extractor. Prior studies (Yin et al. 2019; Wang et al. 2020; Xie et al. 2022) indicate that low-frequency image features, encoding global statistics, are highly sensitive to domain shifts. Thus, we adopt low-frequency spectral features as lightweight indicators for domain discrimination.

Method	REF	Gaussian	Shot	Impulse	Defocus	Glass	Motion	Zoom	Snow	Frost	Fog	Brightness	Contrast	Elastic	Pixelate	JPEG	Mean↓	Gain
Source	ICLR2021	53.0	51.8	52.1	68.5	78.8	58.5	63.3	49.9	54.2	57.7	26.4	91.4	57.5	38.0	36.2	55.8	0.0
Pseudo-label	ICML2013	45.2	40.4	41.6	51.3	53.9	45.6	47.7	40.4	45.7	93.8	98.5	99.9	99.9	98.9	99.6	61.2	-5.4
TENT-continual	ICLR2021	52.2	48.9	49.2	65.8	73.0	54.5	58.4	44.0	47.7	50.3	23.9	72.8	55.7	34.4	33.9	51.0	+4.8
CoTTA	CVPR2022	52.9	51.6	51.4	68.3	78.1	57.1	62.0	48.2	52.7	55.3	25.9	90.0	56.4	36.4	35.2	54.8	+1.0
VDP	AAAI2023	52.7	51.6	50.1	58.1	70.2	56.1	58.1	42.1	46.1	45.8	23.6	70.4	54.9	34.5	36.1	50.0	+5.8
ViDA	ICLR2024	47.7	42.5	42.9	52.2	56.9	45.5	48.9	38.9	42.7	40.7	24.3	52.8	49.1	33.5	33.1	43.4	+12.4
ADMA	CVPR2024	46.3	41.9	42.5	51.4	54.9	43.3	40.7	34.2	35.8	64.3	23.4	60.3	37.5	29.2	31.4	42.5	+13.3
Ours	Proposed	47.7	45.1	42.2	46.6	49.7	42.8	46.5	35.0	38.0	35.2	21.6	51.7	43.5	26.9	31.0	40.2	+15.6

Table 1: Classification error rates (%) for the ImageNet-to-ImageNet-C CTTA task. *Mean* is the average error across 15 corruption types. *Gain* shows accuracy improvement over the source model.

Given an RGB image $x \in \mathbb{R}^{H \times W \times 3}$, we first convert it to a single-channel grayscale image $x_{\text{gray}} \in \mathbb{R}^{H \times W}$. The two-dimensional discrete Fourier transform (2-D DFT) of x_{gray} is then computed as:

$$F(u, v) = \sum_{m=0}^{H-1} \sum_{n=0}^{W-1} x_{\text{gray}}(m, n) e^{-j 2\pi \left(\frac{um}{H} + \frac{vn}{W} \right)}, \quad (16)$$

where $F(u, v) \in \mathbb{C}^{H \times W}$ is the complex-valued spectrum, $j = \sqrt{-1}$ is the imaginary unit, and $0 \leq u < H$, $0 \leq v < W$.

Once the complex spectrum $F(u, v)$ is obtained, we shift the DC component to the center of the frequency plane using a standard frequency-shift operation. The magnitude of the resulting complex spectrum is then taken to obtain a real-valued, centered magnitude map $M \in \mathbb{R}^{H \times W}$. Let the vertical and horizontal coordinates of the map center be $c_r = \lfloor H/2 \rfloor$ and $c_c = \lfloor W/2 \rfloor$, we keep only the low-frequency part by cropping a square patch of side length $L = 2l + 1$ around that centre:

$$f_{\text{low}} = M[c_r - l : c_r + l, c_c - l : c_c + l], \quad (17)$$

where l is a predefined frequency radius that controls the size of the low-frequency region. Finally, the cropped low-frequency patch f_{low} is flattened into a vector $z \in \mathbb{R}^{L^2}$, which serves as the compact low-frequency descriptor.

3.4 Optimization Objective

Following prior work in test-time adaptation, we minimize the entropy of model predictions for target samples, and update our expert modules only when the prediction entropy is below a threshold κ :

$$\mathcal{L}_{\text{TTA}} = -1 \{ \mathcal{H}(\hat{y}) < \kappa \} \sum_{k=1}^C \hat{y}_k \log \hat{y}_k, \quad (18)$$

where $\hat{y} = q_{\theta}(x)$ is the model’s predicted probability vector for input x , C is the number of classes, and $1\{\cdot\}$ denotes the indicator function.

4 Experiment

4.1 Experimental Setting

Dataset. We conduct experiments on both image classification and semantic segmentation tasks. For classifica-

tion, we use CIFAR100-to-CIFAR100C (Krizhevsky, Hinton et al. 2009) and ImageNet-C (Hendrycks and Dietterich 2019). We further evaluate on our proposed Continual Repeated Shifts (CRS) benchmark, with two extended versions: ImageNet+ and ImageNet++, constructed by cyclically combining four ImageNet-derived datasets: ImageNet-V2 (Recht et al. 2019), ImageNet-A (Hendrycks et al. 2021b), ImageNet-R (Hendrycks et al. 2021a), and ImageNet-S (Wang et al. 2019). For segmentation, we evaluate our method on Cityscapes-to-ACDC (Cordts et al. 2016; Sakaridis, Dai, and Van Gool 2021).

CTTA Task Setting. Following CoTTA (Wang et al. 2022a), for CIFAR100C and ImageNet-C, we evaluate on the largest corruption severity (level 5) and process all 15 corruption types sequentially as distinct domains. For ACDC, to reflect realistic temporal changes, we perform continual adaptation by looping through ACDC’s subdomains (Fog \rightarrow Night \rightarrow Rain \rightarrow Snow) in three cycles.

On the CRS benchmark, for ImageNet+, we define four target domains: ImageNet-V2, ImageNet-A, ImageNet-R and ImageNet-S, and adapt in the fixed order (V2 \rightarrow A \rightarrow R \rightarrow S), repeated three times. For ImageNet++, we employ a finer-grained schedule to create 12 domain shifts over three rounds. In each round, we use different subsets of the four datasets. More details are provided in the Appendix E. We further introduce a benchmark-specific metric, *Repeat Forget (RF)*, to quantify the degree of forgetting on repeated domains. For error-based metrics, it is defined as $\text{RF} = \text{Error}_{\text{final}} - \text{Error}_{\text{first}}$, where $\text{Error}_{\text{first}}$ and $\text{Error}_{\text{final}}$ are the average errors during the first and final rounds. A higher RF implies greater forgetting. For accuracy-based metrics, the sign is reversed.

Implementation Details. We adopt standardized CTTA settings for fair comparison. ViT-Base is used as the backbone for classification, and SegFormer-B5 for segmentation. The Adam optimizer and task-specific learning rates are used throughout. More hyperparameter and architecture details can be found in Appendix F.

4.2 Classification CTTA Tasks

We first evaluate our proposed framework on the challenging ImageNet-C and our CRS benchmark. For completeness,

Time		$t \longrightarrow$															Mean \downarrow	RF \downarrow	Gain
Round		1					2					3							
Method	REF	V2	A	S	R	Mean \downarrow	V2	A	S	R	Mean \downarrow	V2	A	S	R	Mean \downarrow			
Source	ICLR2021	44.8	61.3	58.4	43.0	51.2	44.8	61.3	58.4	43.0	51.2	44.8	61.3	58.4	43.0	51.2	51.2	0.0	/
CoTTA	CVPR2022	44.7	61.3	58.0	43.0	51.0	44.6	59.7	57.5	42.3	50.5	44.5	58.6	56.9	42.1	50.1	50.2	-0.9	+1.0
ViDA	ICLR2024	44.0	51.5	51.5	41.2	47.0	44.6	51.4	63.5	44.8	53.2	45.2	52.1	67.9	46.8	55.8	52.0	+5.0	-0.8
Ours	Proposed	43.9	52.2	52.0	38.4	46.5	44.2	50.4	49.2	33.2	44.0	44.1	49.6	47.9	32.1	43.1	44.5	-3.4	+6.7

Table 2: Classification error rates (%) for the ImageNet-to-ImageNet+ CTTA task.

Time		$t \longrightarrow$															Mean \downarrow	RF \downarrow	Gain
Round		1					2					3							
Method	REF	V2	A	S	R	Mean \downarrow	V2'	A'	S'	R'	Mean \downarrow	V2''	A''	S''	R''	Mean \downarrow			
Source	ICLR2021	27.1	61.3	58.3	43.2	48.4	15.8	61.3	58.3	42.6	45.8	19.7	61.3	58.6	43.3	46.9	47.0	-1.5	/
CoTTA	CVPR2022	27.1	61.3	58.1	43.1	48.3	15.6	60.7	57.9	42.8	45.5	19.5	59.8	57.3	41.7	45.8	46.5	-2.5	+0.5
ViDA	ICLR2024	26.0	51.4	52.0	40.5	43.4	15.1	50.8	58.3	41.4	43.5	19.5	50.2	60.5	42.5	45.5	44.2	+2.1	+2.8
Ours	Proposed	25.9	51.6	53.2	40.3	43.9	14.9	50.1	51.4	35.6	39.4	19.1	47.9	48.5	32.5	38.2	40.5	-3.4	+6.5

Table 3: Classification error rates (%) for the ImageNet-to-ImageNet++ CTTA task.

CIFAR100-C results are provided in Appendix G.

Results on ImageNet-C. As shown in Table 1, our method lowers the error to 40.2%, delivering a 15.6% gain over the frozen source model and outperforming the previous best result (ADMA, 42.5%) by 2.3%. It achieves the lowest error on 10 of the 15 corruption types, including the challenging *Defocus*, *Glass Blur*, *Fog*, and *JPEG*, demonstrating that our dual-branch expert architecture, combined with the FDD mechanism, effectively removes corruption-specific noise while preserving task-relevant semantics.

Results on ImageNet+. Table 2 presents results under our CRS benchmark, where four domains (V2, A, R, S) recur over three rounds. Our method achieves a mean error of 44.5%, outperforming the source model by 6.7% and CoTTA by 5.7%. While ViDA performs well initially (error 47.0% vs. 46.5% for ours in round 1), it suffers from severe forgetting in subsequent rounds, yielding a high Repeat Forget (RF) of +5.0. In contrast, our method not only maintains stable performance (44.0%, 43.1%) but also achieves an RF of -3.4, demonstrating strong resistance to forgetting.

Results on ImageNet++. The finer-grained ImageNet++ benchmark (Table 3) presents a more challenging CRS setting with 12 structured shifts. Our method achieves the lowest overall error of 40.5%, outperforming the ViDA by 3.7%. Unlike ViDA, which shows a clear performance drop in later rounds (RF = +2.1), our method consistently improves over time, achieving an RF of -3.4. This highlights our model’s ability to adapt to new domains while retaining and refining knowledge under repeated shifts.

4.3 Semantic Segmentation CTTA Task

On the Cityscapes-to-ACDC CTTA benchmark, as shown in Table 4, our method achieves the best performance with an average mIoU of 62.7%, outperforming all baselines. It

maintains consistent gains across all rounds without performance degradation. Compared to recent methods such as ADMA and SVDP, our approach demonstrates superior robustness and adaptability, highlighting the effectiveness of our dual-expert framework and frequency-aware adaptation.

4.4 Ablation Study

Expert Module Ablation. We ablate the dual-branch expert design by selectively enabling two branches (see Table 5). Without any experts, the backbone struggles on all three benchmarks. Enabling only the shared branch reduces corruption error on ImageNet-C but brings limited gains on the ImageNet+ and ImageNet++. In contrast, Activating only the domain self-adaptive branch cuts error more uniformly, confirming the benefit of per-domain specialization under large shifts. Combining both branches yields the lowest error across all benchmarks, as the shared expert provides stable, reusable semantics, while the self-adaptive experts refine domain-specific residuals. The two are thus complementary, not redundant.

Ablation Study on FDD. We conduct ablations on ImageNet-C to evaluate the effectiveness of FDD, as summarized in Table 6. The *Upper (oracle)* variant assumes known domain boundaries and achieves the lowest error. Our full *FDD* detects 7 domains (Fig. 3(a)) and achieves nearly optimal performance with far fewer experts, showing its efficiency. *Conf-Jump* adds a new expert based on confidence spikes (Gan et al. 2023), leading to 26 experts and 50.2% error, showing its unreliability for domain detection. *FDD w/o covariance* uses Euclidean distance instead of Mahalanobis, which blurs domain boundaries (Fig. 3(b)) and increases error to 41.3%. *FDD w/o discriminator* disables expert selection and routes inputs randomly, resulting in the worst performance among all variants. *FDD w/o Robust Online Update* replaces our weighted update with

Time		$t \longrightarrow$																	
Round		1					2					3					Mean \uparrow	RF \downarrow	Gain
Method	REF	Fog	Night	Rain	Snow	Mean \uparrow	Fog	Night	Rain	Snow	Mean \uparrow	Fog	Night	Rain	Snow	Mean \uparrow			
Source	ICLR2021	69.1	40.3	59.7	57.8	56.7	69.1	40.3	59.7	57.8	56.7	69.1	40.3	59.7	57.8	56.7	56.7	0.0	/
TENT	ICLR2021	69.0	40.2	60.1	57.3	56.7	68.3	39.0	60.1	56.3	55.9	67.5	37.8	59.6	55.0	55.0	55.7	+1.7	-1.0
CoTTA	CVPR2022	70.9	41.2	62.4	59.7	58.6	70.9	41.1	62.6	59.7	58.6	70.9	41.0	62.7	59.7	58.6	58.6	0.0	+1.9
SVDP	AAAI2024	72.1	44.0	65.2	63.0	61.1	72.2	44.5	65.9	63.5	61.5	72.1	44.2	65.6	63.6	61.4	61.3	-0.3	+4.6
BECoTTA	ICML2024	71.5	42.6	63.2	59.1	59.1	71.5	42.6	63.2	59.1	59.1	71.5	42.5	63.2	59.1	59.1	59.1	0.0	+2.4
ADMA	CVPR2024	71.9	44.6	67.4	63.2	61.8	71.7	44.9	66.5	63.1	61.6	72.3	45.4	67.1	63.1	62.0	61.8	-0.2	+5.1
Ours	Proposed	72.6	44.2	67.0	64.4	62.1	73.2	45.5	68.0	64.9	62.9	73.2	45.6	68.2	65.2	63.1	62.7	-1.0	+6.0

Table 4: Performance comparison for Cityscapes-to-ACDC CTTA. “Mean” is the average mIoU.

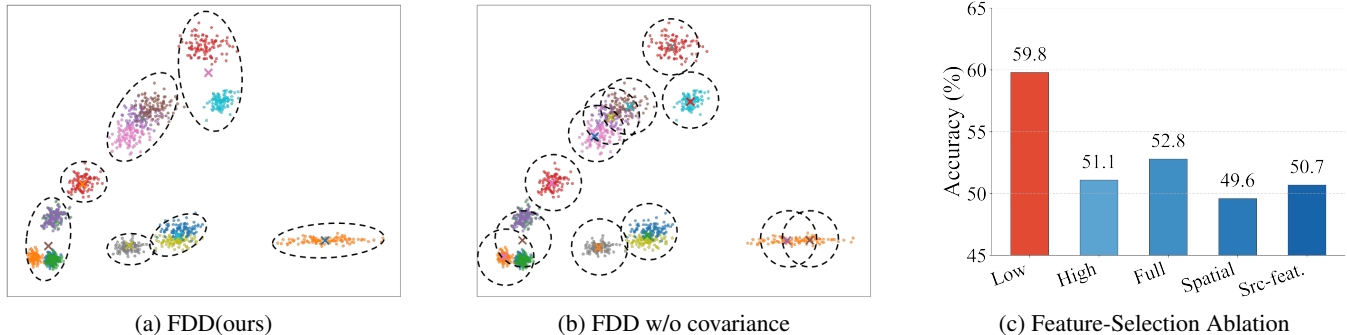


Figure 3: **FDD visualization and feature ablation on ImageNet-C.** (a) PCA projection of low-frequency features with full FDD; six well-separated clusters are observed, “x” denote estimated domain centres μ_i . (b) Same projection with FDD without covariance, showing overlapping clusters and reduced separability. (c) Feature-Selection Ablation showing classification accuracy on ImageNet-C when the default low-frequency descriptor is replaced with four alternatives.

Shared	Self-Adaptive	ImageNet-C	ImageNet+	ImageNet++
		55.8	51.2	47.0
✓		43.1	49.7	45.8
	✓	41.3	46.3	42.7
✓	✓	40.2	44.5	40.5

Table 5: Ablation of the expert branch. Numbers are classification error rates (lower is better).

EMA, increasing error to 41.8%. These results highlight the necessity of each component in FDD to achieve stable and accurate domain shift detection.

Feature-Selection Ablation. Fig. 3(c) reports accuracy after replacing our default low-frequency descriptor with four alternatives. Using high-frequency (High), full-spectrum (Full), spatial RGB histogram (Spatial), or ViT-CLS token representations (Src-feat.) all leads to performance drops. These results reinforce that the raw low-frequency patch achieves the best balance, capturing shift-specific global statistics while suppressing category details for precise and robust discrimination.

5 Conclusion

In this paper, we propose a frequency-discriminative shared & self-adaptive domain expert model that achieves a strong

Variant	K	Error \downarrow	Diff \downarrow
Upper (oracle)	15	39.7	–
Conf-Jump (AAAI2024)	26	50.2	10.5
FDD w/o covariance	12	41.3	1.6
FDD w/o discriminator	7	46.6	3.9
FDD w/o robust online update	7	41.8	2.1
FDD (ours)	7	40.2	0.5

Table 6: Ablation of the FDD; K denotes the number of discovered domains, i.e., the corresponding domain self-adaptive experts created; Diff is the error gap to Upper.

balance between adaptation and forgetting in CTTA. Under the proposed CRS benchmark with repeated and diverse domain shifts, our method demonstrates consistent performance improvement while avoiding catastrophic forgetting. These results highlight its practicality in realistic, non-stationary environments.

Limitation. Our current design adopts a fixed fusion strategy between shared and domain self-adaptive experts. While effective, this static combination may limit adaptability under more complex or dynamic shifts. Future work could explore learnable or input-aware fusion mechanisms to further enhance flexibility and robustness.

Acknowledgments

This work is supported by the National Natural Science Foundation of China under Grant No.U21B2048 and No.62302382, the Shenzhen Key Technical Projects Under Grant CJGJZD20220517141605013, and the China Postdoctoral Science Foundation No.2024M752584.

References

- Aghdam, M. A.; et al. 2024. DA-MoE: Towards Dynamic Expert Allocation for Mixture-of-Experts Models. *arXiv preprint arXiv:2409.06669*.
- Cordts, M.; Omran, M.; Ramos, S.; Rehfeld, T.; Enzweiler, M.; Benenson, R.; Franke, U.; Roth, S.; and Schiele, B. 2016. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 3213–3223.
- Ding, C.; Li, J.; Dong, S.; Gao, X.; He, Y.; and Gong, Y. 2025. SuLoRA: Subspace Low-Rank Adaptation for Parameter-Efficient Fine-Tuning. In *Findings of the Association for Computational Linguistics: ACL 2025*, 5334–5349.
- Döbler, M.; Marsden, R. A.; and Yang, B. 2023. Robust mean teacher for continual and gradual test-time adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7704–7714.
- Dong, S.; Chen, Y.; He, Y.; Jin, Y.; Kot, A. C.; and Gong, Y. 2025. Analogical Augmentation and Significance Analysis for Online Task-Free Continual Learning. *IEEE Transactions on Multimedia*.
- Dong, S.; Gao, X.; He, Y.; Zhou, Z.; Kot, A. C.; and Gong, Y. 2024a. CEAT: Continual Expansion and Absorption Transformer for Non-Exemplar Class-Incremental Learning. *IEEE Transactions on Circuits and Systems for Video Technology*.
- Dong, S.; Tan, C.; Zuo, Z.; He, Y.; Gong, Y.; Zhou, T.; Liu, J.; and Zhang, J. 2024b. Brain-inspired dual-pathway neural network architecture and its generalization analysis. *Science China Technological Sciences*, 67: 2319–2330.
- Gan, Y.; Bai, Y.; Lou, Y.; Ma, X.; Zhang, R.; Shi, N.; and Luo, L. 2023. Decorate the newcomers: Visual domain prompt for continual test time adaptation. In *Proceedings of the AAAI conference on artificial intelligence*, volume 37, 7595–7603.
- Gong, Y.; and Wang, G. 2024. Preface: Brain-Inspired AI Research. *Science China Technological Sciences*, 67: 2281–2281.
- Han, J.; Ding, C.; Dong, S.; He, Y.; Gao, X.; and Gong, Y. 2025. Unleashing the Potential of All Test Samples: Mean-Shift Guided Test-Time Adaptation. *arXiv preprint arXiv:2507.00462*.
- Hastie, T.; and Tibshirani, R. 1996. Discriminant analysis by Gaussian mixtures. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 58(1): 155–176.
- Hendrycks, D.; Basart, S.; Mu, N.; Kadavath, S.; Wang, F.; Dorundo, E.; Desai, R.; Zhu, T.; Parajuli, S.; Guo, M.; et al. 2021a. The many faces of robustness: A critical analysis of out-of-distribution generalization. In *Proceedings of the IEEE/CVF international conference on computer vision*, 8340–8349.
- Hendrycks, D.; and Dietterich, T. 2019. Benchmarking neural network robustness to common corruptions and perturbations. *arXiv preprint arXiv:1903.12261*.
- Hendrycks, D.; Zhao, K.; Basart, S.; Steinhardt, J.; and Song, D. 2021b. Natural adversarial examples. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 15262–15271.
- Houlsby, N.; Giurgiu, A.; Jastrzebski, S.; Morrone, B.; de Laroussilhe, Q.; Gesmundo, A.; Attariyan, M.; and Gelly, S. 2019. Parameter-efficient transfer learning for NLP. In *International Conference on Machine Learning (ICML)*.
- Hu, E. J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, L.; and Chen, W. 2021. LoRA: Low-Rank Adaptation of Large Language Models. *arXiv preprint arXiv:2106.09685*.
- Jia, M.; Tang, L.; Chen, B.-C.; Cardie, C.; Belongie, S.; Hariharan, B.; and Lim, S.-N. 2022. Visual prompt tuning. In *European conference on computer vision*, 709–727. Springer.
- Krizhevsky, A.; Hinton, G.; et al. 2009. Learning multiple layers of features from tiny images.(2009).
- Lee, D.; Yoon, J.; and Hwang, S. J. 2024. Becotta: Input-dependent online blending of experts for continual test-time adaptation. *arXiv preprint arXiv:2402.08712*.
- Liang, J.; He, R.; and Tan, T. 2025. A comprehensive survey on test-time adaptation under distribution shifts. *International Journal of Computer Vision*, 133(1): 31–64.
- Liu, J.; Xu, R.; Yang, S.; Zhang, R.; Zhang, Q.; Chen, Z.; Guo, Y.; and Zhang, S. 2024. Continual-mae: Adaptive distribution masked autoencoders for continual test-time adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 28653–28663.
- Liu, J.; Yang, S.; Jia, P.; Zhang, R.; Lu, M.; Guo, Y.; Xue, W.; and Zhang, S. 2023. Vida: Homeostatic visual domain adapter for continual test time adaptation. *arXiv preprint arXiv:2306.04344*.
- Peng, L.; He, Y.; Wang, S.; Song, X.; Dong, S.; Wei, X.; and Gong, Y. 2024. Global self-sustaining and local inheritance for source-free unsupervised domain adaptation. *Pattern Recognition*, 155: 110679.
- Peng, L.; Wan, C.; Wang, S.; Song, X.; He, Y.; and Gong, Y. 2025. CIA: Class-and Instance-aware Adaptation for Vision-Language Models. In *Proceedings of the 33rd ACM International Conference on Multimedia*, 2870–2879.
- Recht, B.; Roelofs, R.; Schmidt, L.; and Shankar, V. 2019. Do imagenet classifiers generalize to imagenet? In *International conference on machine learning*, 5389–5400. PMLR.
- Sakaridis, C.; Dai, D.; and Van Gool, L. 2021. ACDC: The adverse conditions dataset with correspondences for semantic driving scene understanding. In *Proceedings of the IEEE/CVF international conference on computer vision*, 10765–10775.
- Shazeer, N.; Mirhoseini, A.; Maziarz, K.; Davis, A.; Le, Q.; Hinton, G.; and Dean, J. 2017. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. *arXiv preprint arXiv:1701.06538*.

- Song, J.; Lee, J.; Kweon, I. S.; and Choi, S. 2023. Ecotta: Memory-efficient continual test-time adaptation via self-distilled regularization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11920–11929.
- Wang, H.; Ge, S.; Lipton, Z.; and Xing, E. P. 2019. Learning robust global representations by penalizing local predictive power. *Advances in neural information processing systems*, 32.
- Wang, Q.; Fink, O.; Van Gool, L.; and Dai, D. 2022a. Continual test-time domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7201–7211.
- Wang, Q.; He, Y.; Dong, S.; Gao, X.; Wang, S.; and Gong, Y. 2024. Non-exemplar domain incremental learning via cross-domain concept integration. In *European Conference on Computer Vision*, 144–162. Springer.
- Wang, Q.; He, Y.; Dong, S.; Song, X.; Han, J.; Luo, H.; and Gong, Y. 2025. Dualcp: Rehearsal-free domain-incremental learning via dual-level concept prototype. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 21198–21206.
- Wang, Y.; Agarwal, S.; Mukherjee, S.; Liu, X.; Gao, J.; Awadallah, A. H.; and Gao, J. 2022b. Adamix: Mixture-of-adaptations for parameter-efficient model tuning. *arXiv preprint arXiv:2205.12410*.
- Wang, Z.; Yang, Y.; Shrivastava, A.; Rawal, V.; and Ding, Z. 2020. Towards frequency-based explanation for robust CNN. *arXiv preprint arXiv:2005.03141*.
- Xie, J.; Li, W.; Zhan, X.; Liu, Z.; Ong, Y. S.; and Loy, C. C. 2022. Masked frequency modeling for self-supervised visual pre-training. *arXiv preprint arXiv:2206.07706*.
- Yin, D.; Gontijo Lopes, R.; Shlens, J.; Cubuk, E. D.; and Gilmer, J. 2019. A fourier perspective on model robustness in computer vision. *Advances in Neural Information Processing Systems*, 32.
- Zhang, Z.; et al. 2024. DeepSeekMoE: Towards Ultimate Expert Specialization in Mixture-of-Experts Language Models. *arXiv preprint arXiv:2401.06066*.