Decomposing the Neurons: Activation Sparsity via Mixture of Experts for Continual Test Time Adaptation

Rongyu Zhang[1,2,3*], **Aosong Cheng**[3*], **Yulin Luo**[3*], **Gaole Dai**[3*], **Huanrui Yang**[5], **Jiaming Liu**[3], **Ran Xu**[3], **Li Du**[1], **Dan Wang**[4†], **Yuan Du**[1†]

[1]Nanjing University
[2]The Hong Kong Polytechnic University
[3]State Key Laboratory of Multimedia Information Processing, School of Computer Science, Peking University
[4]The Hong Kong University of Science and Technology
[5]Arizona University
danwang@ust.hk, yuandu@nju.edu.cn

Abstract

Continual Test-Time Adaptation (CTTA), which aims to adapt the pre-trained model to ever-evolving target domains, emerges as an important task for vision models. As current vision models appear to be heavily biased towards texture, continuously adapting the model from one domain distribution to another can result in serious catastrophic forgetting. Drawing inspiration from the the encoding characteristics of neuron activation in neural networks, we propose the Mixture-of-Activation-Sparsity-Experts (MoASE) for the CTTA task. Given the distinct reaction of neurons with low and high activation to domain-specific and agnostic features, MoASE decomposes the neural activation into high-activation and low-activation components in each expert with a Spatial Differentiable Dropout (SDD). Based on the decomposition, we devise a Domain-Aware Router (DAR) that utilizes domain information to adaptively weight experts that process the post-SDD sparse activations, and the Activation Sparsity Gate (ASG) that adaptively assigns feature selection thresholds of the SDD for different experts for more precise feature decomposition. Finally, we introduce a Homeostatic-Proximal (HP) loss to maintain update consistency between the teacher and student experts to prevent error accumulation. Extensive experiments substantiate that MoASE achieves state-of-the-art performance in both classification and segmentation tasks.

Introduction

The rapid advancement of deep learning in autonomous driving (Yurtsever et al. 2020; Yang et al. 2023a) and robotics (Chen and Liu 2023; Li et al. 2023) has highlighted significant challenges in adapting to continuously changing test-time scenarios. Traditional stationary machine perception systems (Zhang et al. 2024b; Xie et al. 2021), which assume that test-time data distributions mirror training data, often suffer from severe error accumulation and catastrophic forgetting when confronted with distribution shifts. To address this, Continual Test-Time Adaptation (CTTA) (Wang et al. 2022; Song et al. 2023) has been proposed, extending the conventional Test-Time Adaptation (TTA) framework (Wang et al. 2020; Liang, He, and Tan 2023) to handle sequences of evolving distribution shifts over time.

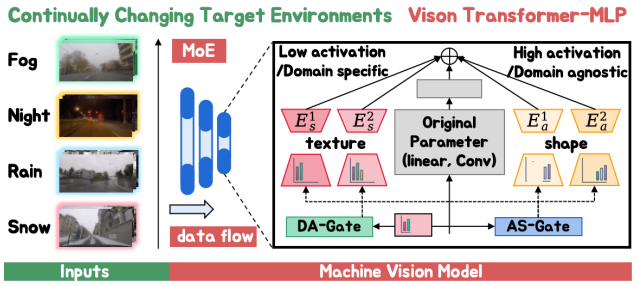

Figure 1: **The problem and motivation.** Our goal is to adapt pre-trained models to evolving target domains.

Current CTTA methods (Liu et al. 2023b; Yang et al. 2023b; Gan et al. 2022; Liu et al. 2023a) predominantly rely on teacher-student frameworks with pseudo-labeling to extract domain knowledge. For instance, Liu et al.(Liu et al. 2023b) demonstrate that processing feature domains separately can enhance model resilience by reducing domain-specific disturbances. However, these approaches often depend on `implicit` mechanisms, such as self-training visual prompts and adapters (Yang et al. 2023b; Liu et al. 2023b), which lack interpretability and limit control over the adaptation process. Moreover, in contrast to implicit models, vision science reveals that the human visual system employs a clearly defined, explicit mechanism with an absolute threshold (Barlow 1956; Koenig and Hofer 2011) to process visual signals separately. Therefore, we aim to explore the solution for CTTA tasks from an `explicit` perspective to decompose the features for better perception.

Recent studies (Li et al. 2024; Zhang et al. 2024a) on activation sparsity have revealed a unique characteristic of neu-

*These authors contributed equally.
†Corresponding authors.

ral networks: strongly activated neurons primarily encode foreground shapes and structures that demand the most attention, whereas weak activation corresponds to background texture. To demystify the role of activation sparsity in CTTA, we manually decompose strongly and weakly activated neurons in a pretrained model and visualize the input features from varying domains as shown in Fig.2. We observe a clear distinction in the neuron attention, where high activation neurons focus on domain-agnostic outlines and intersection of edges, while low activation neurons attend to domain-specific features of styles and noises. Based on the the prior researches, we propose an intriguing hypothesis: **Can we explicitly decompose neurons according to activation degree to encode shapes and textures separately for better differentiating the continuously changing environments?**

This motivates our study on the explicit decomposition of neural activation with the plug-in Mixture-of-Activation-Sparsity-Experts (MoASE) module for various pretrained vision backbones. MoASE introduces a Spatial Differentiable Dropout (SDD) to each expert that only preserves highest or lowest neuron activation to generate the domain-agnostic and domain-specific experts. MoASE enhances the extraction of domain-agnostic structures and identifies domain-specific textures from a spatial-wise perspective, which allows the MoASE to dynamically adjust the balance between general and specific knowledge, enhancing its adaptability and accommodating different knowledge.

However, a significant challenge lies in defining the thresholds for high and low activation values, as the distribution of activation values varies greatly. Therefore, we develop the input-aware Activation Sparsity Gate (ASG) that adaptively assigns distinct thresholds in SDD module for different experts to enhance the model perception ability at various levels. In addition, we further take advantage of the SDD module and devise a novel Domain-Aware Router (DAR) that further enhances expert allocation with guidance of only low activation domain-specific information to better differentiate domain knowledge. Moreover, indiscriminately enabling two distinct types of experts to independently learn divergent features may cause gradient vectors to misalign, thereby increasing the risk of converging to suboptimal local minima. To mitigate this issue, we introduce the Homeostatic-Proximal (HP) loss to maintain the updates coherence of the student and teacher. HP loss ensures that the optimization trajectory adheres to a consistent and accurate objective and further mitigates the error accumulation caused by the random initialization of the injected MoASE.

Extensive experiments demonstrate the superiority of our proposed MoASE across two image classification benchmarks (Krizhevsky, Hinton et al. 2009; Hendrycks and Dietterich 2019) and one segmentation benchmark (Cordts et al. 2016; Sakaridis, Dai, and Van Gool 2021) on CTTA scenarios with improvements exceeding 15.3% in classification accuracy and 5.5% in segmentation mIoU. These results underscore MoASE's robust capability to adapt reliably across diverse environments. The major contributions include:

- We develop a MoASE model, which addresses the issues of error accumulation and catastrophic forgetting to face the continuously changing distribution.
- We decompose the activation into domain-specific and domain-agnostic features, using distinct expert models to encode texture and shape independently with SDD.
- We leverage a multi-gate module featuring the DAR and ASG, leveraging domain information to generate adaptive routing strategies and activation thresholds.
- We propose a tailored HP loss to ensure optimization objective consistency between the experts and enhance performance within the teacher-student framework.

Related Works

Continual Test-Time Adaptation addresses the challenge of adapting to a non-static target domain, which complicates traditional TTA methods. The pioneering work CoTTA (Wang et al. 2022) combined bi-average pseudo labels with stochastic weight resets to tackle this issue. To mitigate error accumulation, ECoTTA (Song et al. 2023) employs a meta-network for output regularization. While these approaches focus on model-level solutions, other studies (Gan et al. 2023; Ni et al. 2023; Liu et al. 2023b) explore the use of visual domain prompts or minimal parameter adjustments for continual learning. Liu (Liu et al. 2023a) introduced reconstruction techniques for continual adaptation, and BECoTTA (Lee, Yoon, and Hwang 2024) implemented a Mixture of Experts strategy in CTTA, promoting effective domain-specific knowledge retention with data augmentation. Unlike previous `implicit` methods, our MoASE adopts an `explicit` approach to CTTA.

Activation Sparsity refers to the presence of numerous weakly-contributing elements in activation outputs (Chen et al. 2023; Yang et al. 2019). SparseViT (Song et al. 2024) revisits this concept for modern window-based vision transformers, aiming to increase speed and reduce computation. Grimaldi (Grimaldi et al. 2023) introduces semi-structured activation sparsity that can be leveraged with minor runtime adjustments to significantly enhance speed on embedded devices. Meanwhile, Mirzadeh (Mirzadeh et al. 2023) explores the reuse of activated neurons in LLMs, proposing strategies to reduce computation. However, all the previous works aim to improve model efficiency until (Li et al. 2024) reveal the contribution of shape bias for model performance.

Mixture-of-Experts is initially introduced by Jacobs and Jordan (Jacobs et al. 1991; Jordan and Jacobs 1994), uses independent modules to boost expressiveness and cut computational costs. Eigen and Ma (Eigen, Ranzato, and Sutskever 2013; Ma et al. 2018) evolved it into the MoE layer. In natural language processing, GShard (Lepikhin et al. 2021) and Switch Transformer (Fedus, Zoph, and Shazeer 2021) incorporated MoE with top-1/2 routing to enhance capacity. Fixed routing (Dai et al. 2022) and ST-MoE (Zoph 2022) aimed to stabilize training. Recent developments (Zhu et al. 2023) introduced efficient adapters within MoE and (Zhao et al. 2023) combined MoE with implicit neural network for image compression. In computer vision, M^3ViT (Liang et al. 2022) selectively engages experts, while MoFME (Zhang et al. 2024b) merged feature modulation with MoE for better image restoration. MoASE leverages a multi-router MoE to manage diverse neuron activation.

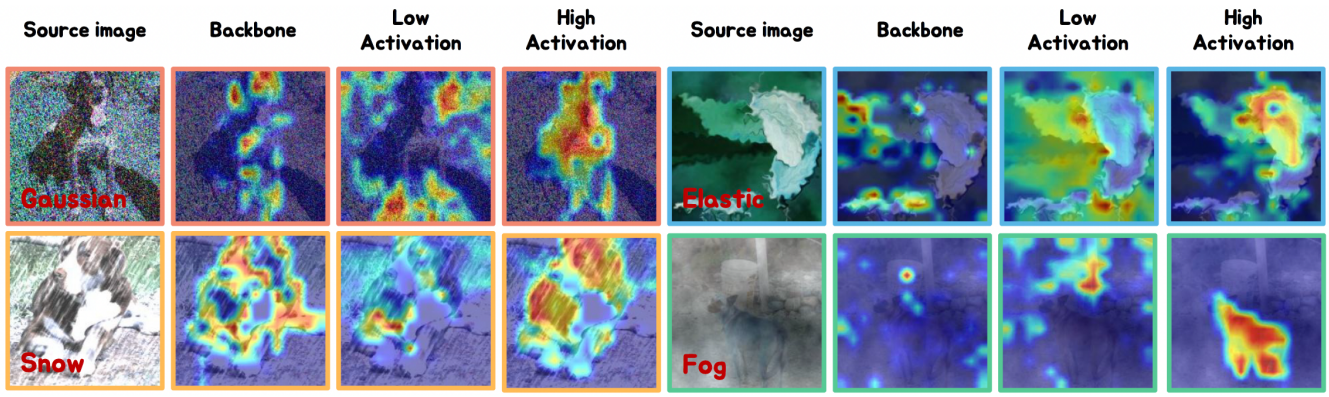

Figure 2: **The visualization analysis of the Class Activation Map (CAM).** We adopt CAM to compare the attention of the low-activation, high-activation MoASE, and the original model during the continual adaptation process.

Motivation

Drawing inspiration from the complexities of the human visual system (Von Helmholtz 1867; Mustafi, Engel, and Palczewski 2009; Bringmann et al. 2018) and reinforced by seminal findings in recent research (Li et al. 2024; Zhang et al. 2024a; Yang and Soatto 2020; Xu et al. 2021), we come up with the hypothesis of leveraging activation sparsity via the innovative integration of parallel activation sparsity experts to effectively decompose activation neurons for continual test adaptation tasks. Such configuration not only ensures domain-agnostic components can be reused for new tasks to mitigate catastrophic forgetting, thereby minimizing the need for extensive retraining, but also ensures that inaccuracies in domain-specific learning do not compromise the overall integrity and performance of the model across various tasks, minimizing error accumulation in dynamic environments (Liu et al. 2023b)..

To provide empirical support for our hypothesis, we extended our analysis to include a qualitative evaluation using Class Activation Mapping (CAM) (Zhou et al. 2016) within the ImageNet-to-ImageNet-C CTTA scenario. Specifically, we selectively retained only the top 50% high or low activation values within the ViT-base model under the CTTA scenario (Wang et al. 2022). As depicted in Fig .2, analysis of two canine samples on the left reveals that while high activation distinctly outlines the dogs, the distribution of low activation features significantly differs across domains. Our findings verify that models maintaining only weakly activated neurons primarily accentuate fluctuations in background noise while overlooking the foreground object. This implies that weakly activated neurons are adept at capturing domain-specific features, which motivates us to design a low-activation-only DAG to ensure token assignment completely based on domain-relevant information. Conversely, models with strongly activated neurons exhibit a contrasting pattern and focus more intensively on object shapes and structures even under globally applied corruption. This observation supports the notion that strongly activated neurons are sensitive to domain-agnostic structures, validating previous visual science principles and our hypothesis.

Methods

Preliminary

Continual Test-Time Adaptation. We pre-train the model $\theta(y_s|x_s)$ on source domain $D_S = (\mathcal{Y}_S, \mathcal{X}_S)$ and adapt it to multiple target domains $D_{T_i} = \{\mathcal{X}_{\mathcal{T}_i}\}_{i=1}^{n}$, where n indicates the number of target datasets. Utilizing the robustness of mean teacher predictions (Tarvainen and Valpola 2017a), we implement a teacher-student framework ($\theta^{\mathcal{T}}$ and $\theta^{\mathcal{S}}$) to maintain stability during adaptation (Wang et al. 2022; Gan et al. 2022). This adaptation process is unsupervised and single-pass for target domain data $x \in \mathcal{X}_T$, does not require access to source domain data, as shown in Fig 3.

Mixture-of-Experts. The MoE model is fundamentally composed of E expert functions $e_i : \mathcal{X}_{\mathcal{T}} \to \mathbb{R}^p$ for $i \in E$, alongside a trainable gating mechanism $g : \mathcal{X}_{\mathcal{T}} \to \mathbb{R}^q$ which allocates inputs to experts by outputting a probability vector. For an input sample x, the MoE's output is the aggregate of expert contributions, each weighted by the router's assigned probabilities, which is represented as:

$$y = \sum_{i=1}^{E} e_i(x) g_i(x), \quad g(x) = \sigma(\boldsymbol{A}x + \boldsymbol{b}) \\ s.t. \;\; g(x) \geq 0 \;\text{ and }\; \sum_{i=1}^{E} g_i(x) = 1 \tag{1}$$

where $\sigma(\cdot)$ signifies the softmax for **soft routing**, $\boldsymbol{A} \in \mathbb{R}^{n \times d}$ represents trainable weights, and $\boldsymbol{b} \in \mathbb{R}^n$ is the bias. It operates densely, allocating nonzero probabilities to all experts.

Mixture-of-Activation-Sparsity-Experts

Spatial Differentiable Dropout. We have adopted the sparse coding principle by incorporating a spatially oriented Top-K dropout operation $\mathcal{T}(\cdot)$ in our framework. However, unlike traditional dropout layer between the linear layers that randomly discards $p(\%)$ of neuron activation on channel dimension, our MoASE incorporates the innovative Spatial Differentiable Dropout (SDD) that selectively retains the top/bottom $K_i(\%)$ significant responses from

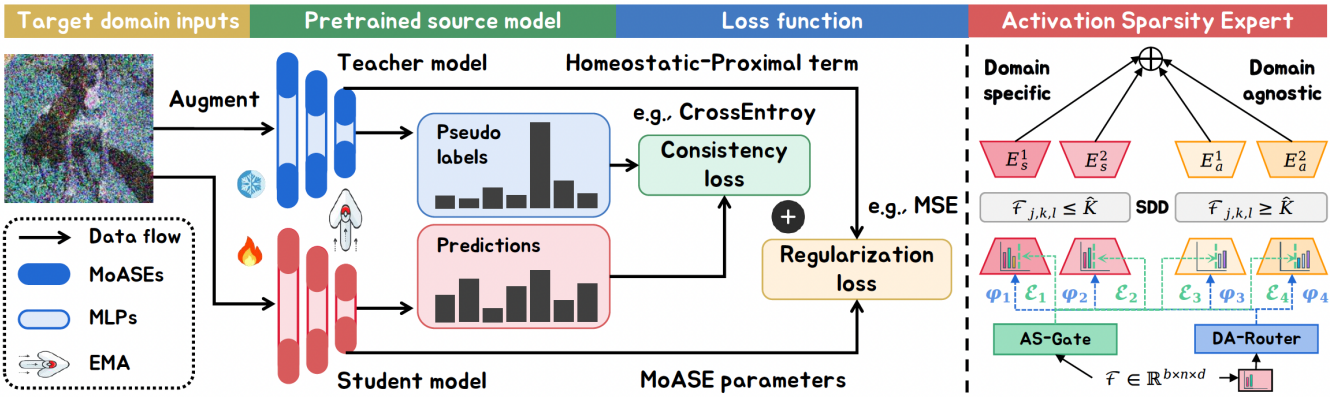

Figure 3: **The framework of Mixture-of-Activation-Sparsity-Experts.** (Left) We integrate the MoASE into the linear layers of a pre-trained source model with a teacher-student framework. (Right) Detailed illustration of activation sparsity expert.

the spatial-wise token dimension between the linear layers within the experts. Given a total of E experts, where E is an even number ($E = 2k, k \in \mathbb{N}^+$), each expert is indexed by $i \in \{1, 2, \ldots, E\}$. The Spatial Differentiable Dropout (SDD) mechanism decomposes the input feature tensor $F \in \mathbb{R}^{b\times n\times d}$ into domain-specific and domain-agnostic components. The output of the SDD layer is defined as:

$$\hat{\mathcal{F}}[j,k,l] := \begin{cases} \mathcal{F}[j,k,l], \text{if } \mathcal{F}[j,k,l] \geq \mathcal{T}(K_i, \mathcal{F}[j,:,:], L) \\ \qquad \text{and } L = True, \\ \mathcal{F}[j,k,l], \text{if } \mathcal{F}[j,k,l] \leq \mathcal{T}(K_i, \mathcal{F}[j,:,:], L) \\ \qquad \text{and } L = False. \end{cases} \tag{2}$$

where $K = \lfloor nd \times q_i \rfloor$, $q \in \mathbb{R}^E$ is a vector of adaptive thresholds, and L corresponds to the largest argument in **torch.topk** ($L = False$ for domain-specific experts and vice versa). q is manually set as $q = \{E_a^i/E, ..., 1/2, E_s^i/E, ..., 1/2\}$ to make each of the two types experts account for half of the features, where E_a^i and E_s^i suggests the expert ID for domain-agnostic experts and domain-specific experts (e.g., $E_a^2 = 2$, $E_s^3 = 3$) only to ensure different experts precept different levels of activation.

Domain-Aware Router and Activation Sparsity Gate. To better suit the dynamic nature of the CTTA task, we propose two different input-aware modules to provide domain-specific information and differentiate domain-agnostic objects for each activation sparsification experts. Both modules employ the function of $g(\cdot)$ described in Eq.1.

The DAR acts as a common router in an MoE architecture, where it dynamically selects the experts to utilize and combines their outputs. One unique design of DAR is that we want it to specifically focus on the domain-specific information captured by the low activations, so that it can better adapt to the diverse domains in the CTTA scenario. Therefore, we simultaneously apply the SDD layer with $q = \frac{1}{2}$ to decomposing low activations from the input feature $\mathcal{F} \in \mathbb{R}^{b\times n\times d}$ in the DAR which can be formulated as:

$$\varphi = g_{DAR}(\delta(\mathcal{F}, \lfloor \frac{nd}{2} \rfloor, False)) \tag{3}$$

ASG, on the other hand, directly adjusts each expert's behavior by generating a dynamic threshold $\varepsilon \in \mathbb{R}^{1\times E}$ for SDD in each expert to facilitate adaptive activation decomposition. As ASG controls both domain-specific and domain-agnostic experts, we use the full input features as the module input to better respond to change in both domains and input information. The ASG is formulated as:

$$\varepsilon = g_{ASG}(\mathcal{F}) \tag{4}$$

The DAR output $\varphi \in \mathbb{R}^{b\times n\times E}$ is used as the expert weight g as in Eq.1. The thresholds generated by ASG are combined with predefined thresholds in Eq.2 to compute the $\hat{K}_i = \lfloor nd \times (q + \eta \cdot \varepsilon_i) \rfloor$ for each expert, where η is a scaling factor set to 0.1 that ensures $\hat{K}$ does not exceed the upper limit. This integration supports a balanced adaptation.

Optimization Objective

Homeostatic-Proximal Loss. Building on prior CTTA research (Wang et al. 2022), we employ the teacher model $\theta^{\mathcal{T}}$ to generate pseudo labels y_{pd} to minimize the task-specific training objective $\mathcal{L}_{TS}(\cdot)$, which are used to update MoASE. Despite the successful implementation of multiple experts to perceive various degrees of activation, we also strive to maintain statistical homeostasis amid continuous domain shifts by constraining student updates $\theta^{\mathcal{S}}$ to stay closely aligned with the initial (teacher) model $\theta^{\mathcal{T}}$, thus eliminating the need for manual intervention. In particular, instead of just minimizing the $\mathcal{L}_{TS}(\cdot)$, we further introduce the Homeostatic-Proximal (HP) term to the original loss to approximately minimize the optimization objective $\mathcal{L}_{Overall}$:

$$\mathcal{L}_{Overall} = \mathcal{L}_{TS}(\theta^{\mathcal{S}}) + \frac{\mu}{2}\sum_{e=1}^{E} ||\theta_e^{\mathcal{S}} - \theta_e^{\mathcal{T}}||^2. \tag{5}$$

where μ is set to 1 following (Li et al. 2020), $\theta_e^{\mathcal{S}}$ and $\theta_e^{\mathcal{T}}$ represent the parameters of each expert in the MoASE teacher-student framework, respectively. The teacher model is updated according to the Exponential Moving Average (EMA)

$$\theta_t^{\mathcal{T}} = \alpha \times \theta_{t-1}^{\mathcal{T}} + (1-\alpha) \times \theta_t^{\mathcal{S}} \tag{6}$$

with the weight $\alpha = 0.999$ (Tarvainen and Valpola 2017b).

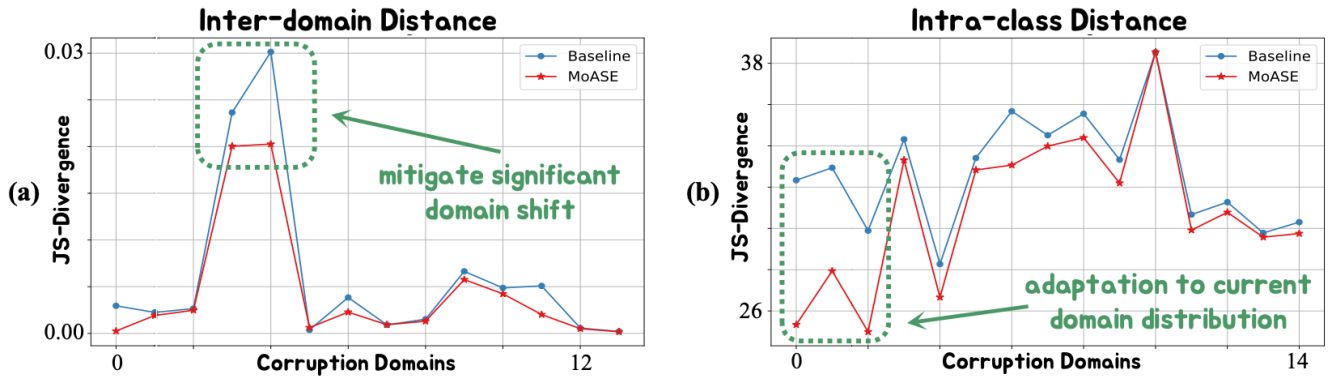

Figure 4: **Inter-domain and intra-class distance in CIFAR10-C.** (a) MoASE more effectively reduces inter-domain divergence than the source model. (b) MoASE significantly improves intra-class feature aggregation.

Justification

To substantiate our hypothesis, we measure the domain representation of MoASE by calculating the distribution distance using Ben-David's domain distance (Ben-David et al. 2006, 2010) and $\mathcal{H}$-divergence, building on previous studies (Ganin et al. 2016). The $\mathcal{H}$-divergence between source domain D_S and target domain D_{T_i} is given as follows:

$$d_{\mathcal{H}}(D_S, D_{T_i}) = 2 \sup_{\mathcal{D}\in\mathcal{H}} \left| \Pr_{D_S}[\mathcal{D}(x) = 1] - \Pr_{D_{T_i}}[\mathcal{D}(x) = 1] \right|. \tag{7}$$

where $\mathcal{H}$ is the hypothesis space and $\mathcal{D}$ the discriminator. Consistent with the methodologies in (Ruder and Plank 2017; Allaway, Srikanth, and McKeown 2021), we employ the $Jensen\text{-}Shannon\ (JS)\ divergence$ between two adjacent domains as an approximation of $\mathcal{H}$-divergence, owing to its demonstrated efficacy in differentiating between domains. A relatively small inter-domain divergence suggests the feature representation is robust and exhibits reduced susceptibility to cross-domain shifts:

$$\begin{aligned} JS(P_{D_S}||P_{D_{T_i}}) = &\frac{1}{2} KL(P_{D_S}||\frac{P_{D_S} + P_{D_{T_i}}}{2}) \\ &+ \frac{1}{2} KL(P_{D_{T_i}}||\frac{P_{D_S} + P_{D_{T_i}}}{2}), \end{aligned} \tag{8}$$

where P denotes probability distribution of model output features and $KL(\cdot||\cdot)$ denotes the $Kullback - Leibler\ (KL)\ divergence$. As shown in Fig.4(a), the MoASE shows significantly lower inter-domain divergence, indicating robust feature representation across domains. Moreover, we evaluate domain representation based on intra-class divergence, inspired by k-means clustering (MacQueen 1967). This can be formulated as:

$$IC = \frac{1}{|C|} \sum_{f_i \sim C} ||f_i - \frac{1}{|C|} \sum_{f_i \sim C} f_i||_2^2 \tag{9}$$

where f_i is the encoder output feature from each domain. Smaller intra-class divergence indicates a superior understanding of the domain, as depicted for the in Fig.4(b). The substantial and consistent margin between MoASE and the baseline model, as shown in Fig.4, demonstrates the effectiveness of our proposed method in mitigating domain shifts.

Experiments

CTTA Task setting. Following (Wang et al. 2022), in classification CTTA tasks, we adapt the pre-trained source model sequentially across fifteen target domains on CIFAR10-C, CIFAR100C, and ImageNet-C exhibiting the highest level of corruption severity (level 5). In the context of segmentation CTTA, as per (Yang et al. 2023b; Liu et al. 2023b), we employ an off-the-shelf model pre-trained on the Cityscapes dataset. For the continual adaptation to target domains, we utilize the ACDC dataset, which comprises images captured under four distinct adverse weather conditions: Fog→Night→Rain→Snow.

Implementation Details. For the backbone architectures in classification CTTA, we employ ViT-base (Dosovitskiy et al. 2020), where we standardize input image sizes to 384×384 for CIFAR and 224×224 pixels for ImageNet. In segmentation CTTA, the Segformer-B5 model (Xie et al. 2021), pre-trained, serves as our source model, with input dimensions reduced from 1920x1080 to 960x540 for processing in target domains. η is set to 0.1. Optimization utilizes the Adam algorithm (Kingma and Ba 2014) with $(\beta_1, \beta_2) = (0.9, 0.99)$. Specific learning rates are assigned to each task: 1e-4 for classification, and 2e-4 for segmentation.

Baselines. We compare our model with sota CTTA methods including the entropy-based method TENT (Wang et al. 2020), the landmark CTTA work with mean-teacher framework CoTTA (Wang et al. 2022), visual prompt-based method VDP (Gan et al. 2023) and SVDP (Yang et al. 2023b), multi-rank adapter-based method ViDA (Liu et al. 2023b), and MoE-based method BECoTTA (Lee, Yoon, and Hwang 2024). Noted that we report the BECoTTA results *w/o* Source Domain Augmentation for a fair comparison.

Quantitative analysis

The effectiveness on classification CTTA. As Tab.1 validate the effectiveness of our method as we conduct experiments on **CIFAR10-to-CIFAR10-C** and **ImageNet-to-ImageNet-C**, which consists of fifteen corruption types that occur sequentially during the test time. For MoASE, the average classification error is up to 55.8% when we directly test the source model on target domains with ImageNet-C.

Method	Venue	*Gaussian*	*shot*	*impulse*	*defocus*	*glass*	*motion*	*zoom*	*snow*	*frost*	*fog*	*brightness*	*contrast*	*elastic_trans*	*pixelate*	*jpeg*	Mean↓	Gain↑
CIFAR10 ⇒ CIFAR10-C																		
Source	ICLR2021	60.1	53.2	38.3	19.9	35.5	22.6	18.6	12.1	12.7	22.8	5.3	49.7	23.6	24.7	23.1	28.2	0.0
TENT	CVPR2021	57.7	56.3	29.4	16.2	35.3	16.2	12.4	11.0	11.6	14.9	4.7	22.5	15.9	29.1	19.5	23.5	+4.7
CoTTA	CVPR2022	58.7	51.3	33.0	20.1	34.8	20	15.2	11.1	11.3	18.5	4.0	34.7	18.8	*19.0*	17.9	24.6	+3.6
VDP	AAAI2023	57.5	49.5	31.7	21.3	35.1	19.6	15.1	10.8	10.3	18.1	4.0	27.5	18.4	22.5	19.9	24.1	+4.1
BECoTTA	ICML2024	54.6	48.1	26.5	22.1	32.8	19.7	14.9	10.1	10.2	16.3	3.9	27.2	16.4	25.7	*15.4*	22.9	+5.3
ViDA	ICLR2024	*52.9*	*47.9*	**19.4**	**11.4**	*31.3*	**13.3**	**7.6**	**7.6**	*9.9*	*12.5*	*3.8*	*26.3*	*14.4*	33.9	18.2	*20.7*	*+7.5*
Ours	**Proposed**	**43.7**	**31.3**	*25.1*	*16.5*	**28.1**	*13.8*	*9.7*	*8.3*	**7.1**	**10.1**	**3.0**	**12.9**	**12.0**	**16.3**	**13.5**	**16.8**	**+11.4**
ImageNet ⇒ ImageNet-C																		
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
TENT	CVPR2021	52.2	48.9	49.2	65.8	73	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
BECoTTA	ICML2024	50.1	46.6	42.3	57.1	65.8	51.3	51.7	42.0	*41.4*	42.5	25.0	67.3	50.3	*31.6*	34.4	44.0	+11.8
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*
Ours	**Proposed**	**43.1**	**38.4**	**36.8**	*54.7*	**52.2**	**41.2**	**48.3**	**37.7**	**35.6**	*41.1*	25.2	*63.5*	**34.7**	**27.7**	**28.3**	**40.5**	**+15.3**

Table 1: **Classification error rate(%) for CTTA task.** Gain(%) represents the percentage of improvement in accuracy.

Backbone	Method	Fog	Night	Rain	Snow	Mean↑
Segformer-B0	ViDA	57.9	27.8	53.1	51.6	47.6
	Ours	**58.2**	**28.7**	**53.6**	**52.2**	**48.2**
SAM-SETR	ViDA	76.5	47.2	68.1	70.7	65.6
	Ours	**76.8**	**47.6**	**68.7**	**71.0**	**66.0**

Table 2: Different size of backbone and foundation model.

Method	bri.	contrast	elastic	pixelate	jpeg	Mean↓
Source	26.4	91.4	57.5	38.0	36.2	49.9
CoTTA	25.3	88.1	55.7	36.4	34.6	48.0
ViDA	24.6	68.2	49.8	34.7	34.1	42.3
Ours	**25.4**	**65.5**	**37.3**	**29.5**	**29.6**	**37.5**

Table 3: The DG comparisons on ImageNet-C.

Our method can outperform all previous methods, achieving a 15.3% and 2.9% improvement over the source model and previous SOTA method, respectively. Moreover, our method showcases remarkable performance across the majority of corruption types, highlighting its effective mitigation of error accumulation and catastrophic forgetting.

The effectiveness on segmentation CTTA. As presented in Tab.5 as we conduct experiments on the four scenarios of Cityscapes-to-ACDC and repeat for three times, we observed a gradual decrease in the mIoUs of TENT over time, indicating the occurrence of catastrophic forgetting. In contrast, our method has a continual improvement of average mIoU (61.8→62.3→62.3) when the same sequence of target domains is repeated. Significantly, the proposed method surpasses the previous state-of-the-art CTTA method by achieving a 0.3% increase. This notable improvement showcases our method's ability to adapt continuously to target domains.

Adaptation across various model backbone. We evaluate the flexibility of MoASE with Segformer-B0(Xie et al. 2021) and introduce the foundation model SAM(Kirillov et al. 2023) as the pre-trained model to continual target domains in the Cityscapes-to-ACDC following (Liu et al. 2023b). Our method significantly enhanced performance in dynamic target domains, as shown in Tab.2, achieving improvements of 0.6% and 0.4% for Segformer-B0 and SAM-SETR, respectively. These findings confirm that the MoASE supports effective transfer learning across model sizes.

num. E.	Fog	Night	Rain	Snow	Mean↑
$E = 2$	71.6	44.0	**66.5**	63.7	61.5
$E = 4$	**72.4**	**44.5**	66.4	**63.8**	**61.8**
$E = 8$	71.4	44.0	65.0	61.7	60.5
$E = 16$	71.5	44.1	65.9	63.3	61.2

Table 4: Different number of experts within the 1^{st} round.

Exploration on domain generalization (DG). To evaluate the DG capabilities of our method, we employed a leave-one-domain-out approach (Zhou et al. 2021; Li et al. 2017), training on 10 of the 15 ImageNet-C domains and using the remaining 5 as unsupervised target domains. Our method adapts a pre-trained model to these 10 domains, then tests directly on the five unseen domains, as shown in Tab.3. Impressively, it reduces average error in these domains by 12.4%, outperforming ViDA by over 4.8%. These results validate the effectiveness of MoASE in DG.

Ablation study

Different number of experts. We evaluate the impact of expert counts in MoASE on mIoU under adverse weather in Cityscape-to-ACDC in Tab. 4. Configurations with 2 to 16 experts were tested, with four experts achieving the best performance, as mean mIoU 61.8%, in Fog and Snow. The results indicate that performance does not scale linearly with expert count; instead, precise activation decomposition is key, as it balances general and specific knowledge by determining each expert's perceptive field.

Time		t ⟶																
Round		1					2					3					Mean↑	Gain↑
Method	Venue	Fog	Night	Rain	Snow	Mean↑	Fog	Night	Rain	Snow	Mean↑	Fog	Night	Rain	Snow	Mean↑		
Source	NIPS2021	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	-1.0
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	+1.9
VDP	AAAI2023	70.5	41.1	62.1	59.5	58.3	70.4	41.1	62.2	59.4	58.2	70.4	41.0	62.2	59.4	58.2	58.2	+1.5
BECoTTA	ICML2024	72.3	42.0	63.5	60.1	59.5	72.4	41.9	63.5	60.2	59.5	72.3	41.9	63.6	60.2	59.5	59.5	+2.8
ViDA	ICLR2024	71.6	*43.2*	*66.0*	*63.4*	*61.1*	**73.2**	*44.5*	*67.0*	**63.9**	*62.2*	*73.2*	**44.6**	*67.2*	**64.2**	**62.3**	*61.9*	+5.2
Ours	**Proposed**	**72.4**	**44.5**	**66.4**	**63.8**	**61.8**	*73.0*	**45.1**	**67.5**	*63.5*	**62.3**	**73.5**	*44.5*	**67.4**	*63.5*	**62.3**	**62.2**	**+5.5**

Table 5: **Performance comparison for Cityscape-to-ACDC CTTA.** We sequentially repeat the target domains three times.

	MoE	SDD	DAR	ASG	HP	Mean↑
Ex_0	✗	✗	✗	✗	✗	58.6
Ex_1	✓	✗	✗	✗	✗	57.9
Ex_2	✓	✓	✗	✗	✗	61.1
Ex_3	✓	✓	✓	✗	✗	61.5
Ex_4	✓	✓	✓	✓	✗	62.0
Ex_5	✓	✓	✓	✓	✓	**62.2**

Table 6: Ablation study on segmentation task.

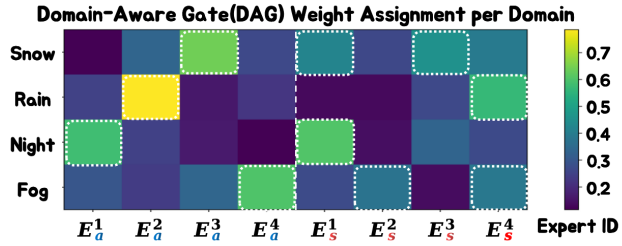

Figure 5: **The expert activation of MoASE.** E_s^1 and E_a^3 stands for the 1^{st} DS and the 3^{rd} DA experts.

Effectiveness of each proposed module. We conducted an ablation study in the Cityscape-to-ACDC CTTA. Tab.6 shows Ex_0 as the baseline using the CoTTA and Ex_1 adding a 4-expert MoE architecture to Ex_0. However, simply adding MoE did not enhance performance as it instead decreased by 0.7%. In contrast, implementing our SDD in Ex_2 improved segmentation results by 3.2%. Further introductions of our modules from Ex_3 and Ex_5 increased mIoU from 61.5% to 62.2%, confirming the effectiveness of MoASE. It is noteworthy that Ex_4, employing ASG, significantly outperforms Ex_3 with full token inputs, indicating that a low activation domain-specific feature alone is sufficient to effectively determine the weights of various experts.

Qualitative analysis

CAM visualization. We conducted a qualitative analysis of the CAM on ImageNet-C, as illustrated at the top left of Fig.6. MoASE effectively concentrates on regions relevant to the target categories, such as dogs and cars, during classification decisions. In contrast, the original model's attention is dispersed due to continuous domain shifts. These findings

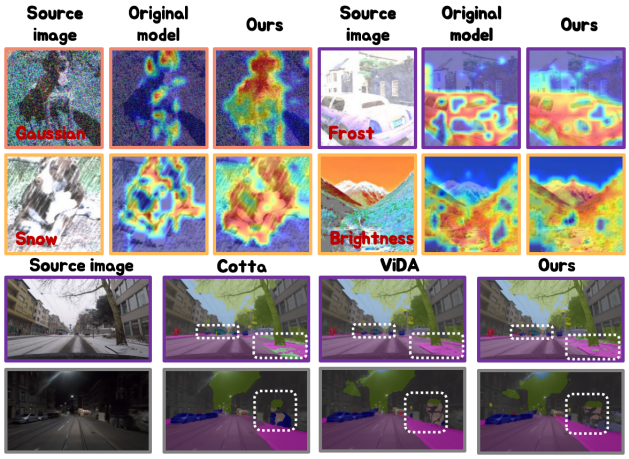

Figure 6: The qualitative analysis of the CAM, routing strategy, and the segmentation qualitative comparison.

underscore MoASE's superiority.

Routing strategy. We sum and normalized the output of the router on ACDC with 8 experts for better illustration, as shown in top right of Fig.5. DAR assigns varying weights to experts based on domain type. For instance, E_s^1 and E_s^3 primarily address snow, while E_s^2 and E_s^4 are geared towards Fog. Moreover, each domain-agnostic expert is also sensitive to different types of feature, suggesting a low-activation-only DAR can also manage characteristic token assignment.

Segmentation visualization. For further validation, we provide additional qualitative comparisons in the Cityscapes-to-ACDC CTTA scenario as shown in the bottom of Fig.6. Our method outperforms other baselines in all scenarios, precisely differentiating sidewalks from roads and identifying small objects like people, vegetation, and fences.

Conclusion

We introduce a novel architecture MoASE for CTTA, which design decouples neural activation into high-activation and low-activation components using SDD and enhances domain-specific feature extraction while improving the perception of domain-agnostic objects through expert modules. The HP loss is also designed to reduce error accumulation. Enhanced by a multi-gate module and HP loss, MoASE surpasses state-of-the-art baselines in four benchmarks.

Acknowledgments

This work was supported by the National Key Research and Development Program of China under Grant 2021YFA0717700, the National Natural Science Foundation of China under Grant W2442028, and also in part by RGC GRF 15200321, 15201322, 15230624, 15239925, ITC ITF-ITS/056/22MX, ITS/052/23MX, and PolyU 1-CDKK, G-SAC8, K-ZYAP.

References

Allaway, E.; Srikanth, M.; and McKeown, K. 2021. Adversarial learning for zero-shot stance detection on social media. *arXiv preprint arXiv:2105.06603*.

Barlow, H. B. 1956. Retinal noise and absolute threshold. *Josa*, 46(8): 634–639.

Ben-David, S.; Blitzer, J.; Crammer, K.; Kulesza, A.; Pereira, F.; and Vaughan, J. W. 2010. A theory of learning from different domains. *Machine learning*, 79(1): 151–175.

Ben-David, S.; Blitzer, J.; Crammer, K.; and Pereira, F. 2006. Analysis of representations for domain adaptation. *NeurIPS*, 19.

Bringmann, A.; Syrbe, S.; Görner, K.; Kacza, J.; Francke, M.; Wiedemann, P.; and Reichenbach, A. 2018. The primate fovea: structure, function and development. *Progress in retinal and eye research*, 66: 49–84.

Chen, P. J.; and Liu, D. R. 2023. Prime editing for precise and highly versatile genome manipulation. *Nature Reviews Genetics*, 24(3): 161–177.

Chen, X.; Liu, Z.; Tang, H.; Yi, L.; Zhao, H.; and Han, S. 2023. SparseViT: Revisiting Activation Sparsity for Efficient High-Resolution Vision Transformer. In *CVPR*, 2061–2070.

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 *CVPR*, 3213–3223.

Dai, D.; Dong, L.; Ma, S.; Zheng, B.; Sui, Z.; Chang, B.; and Wei, F. 2022. StableMoE: Stable Routing Strategy for Mixture of Experts. In Muresan, S.; Nakov, P.; and Villavicencio, A., eds., *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, 7085–7095. Association for Computational Linguistics.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Eigen, D.; Ranzato, M.; and Sutskever, I. 2013. Learning factored representations in a deep mixture of experts. *arXiv:1312.4314*.

Fedus, W.; Zoph, B.; and Shazeer, N. 2021. Switch Transformers: Scaling to Trillion Parameter Models with Simple and Efficient Sparsity. *CoRR*, abs/2101.03961.

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 *AAAI*, volume 37, 7595–7603.

Gan, Y.; Ma, X.; Lou, Y.; Bai, Y.; Zhang, R.; Shi, N.; and Luo, L. 2022. Decorate the Newcomers: Visual Domain Prompt for Continual Test Time Adaptation. *arXiv preprint arXiv:2212.04145*.

Ganin, Y.; Ustinova, E.; Ajakan, H.; Germain, P.; Larochelle, H.; Laviolette, F.; Marchand, M.; and Lempitsky, V. 2016. Domain-adversarial training of neural networks. *JMLR*, 17(1): 2096–2030.

Grimaldi, M.; Ganji, D. C.; Lazarevich, I.; and Sah, S. 2023. Accelerating Deep Neural Networks via Semi-Structured Activation Sparsity. In *ICCV Workshops*, 1179–1188.

Hendrycks, D.; and Dietterich, T. 2019. Benchmarking neural network robustness to common corruptions and perturbations. *arXiv preprint arXiv:1903.12261*.

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.

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

Kingma, D. P.; and Ba, J. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Kirillov, A.; Mintun, E.; Ravi, N.; Mao, H.; Rolland, C.; Gustafson, L.; Xiao, T.; Whitehead, S.; Berg, A. C.; Lo, W.-Y.; et al. 2023. Segment anything. *arXiv preprint arXiv:2304.02643*.

Koenig, D.; and Hofer, H. 2011. The absolute threshold of cone vision. *Journal of vision*, 11(1): 21–21.

Krizhevsky, A.; Hinton, G.; et al. 2009. Learning multiple layers of features from tiny images.

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*.

Lepikhin, D.; Lee, H.; Xu, Y.; Chen, D.; Firat, O.; Huang, Y.; Krikun, M.; Shazeer, N.; and Chen, Z. 2021. GShard: Scaling Giant Models with Conditional Computation and Automatic Sharding. In *9th ICLR, ICLR 2021*. OpenReview.net.

Li, D.; Yang, Y.; Song, Y.-Z.; and Hospedales, T. M. 2017. Deeper, broader and artier domain generalization. In *ICCV*, 5542–5550.

Li, T.; Sahu, A. K.; Zaheer, M.; Sanjabi, M.; Talwalkar, A.; and Smith, V. 2020. Federated optimization in heterogeneous networks. *Proceedings of Machine learning and systems*, 2: 429–450.

Li, T.; Wen, Z.; Li, Y.; and Lee, T. S. 2024. Emergence of Shape Bias in Convolutional Neural Networks through Activation Sparsity. *NeurIPS*, 36.

Li, X.; Zhang, M.; Geng, Y.; Geng, H.; Long, Y.; Shen, Y.; Zhang, R.; Liu, J.; and Dong, H. 2023. Manipllm: Embodied multimodal large language model for object-centric robotic manipulation. *arXiv preprint arXiv:2312.16217*.

Liang, H.; Fan, Z.; Sarkar, R.; Jiang, Z.; Chen, T.; Zou, K.; Cheng, Y.; Hao, C.; and Wang, Z. 2022. M^3 ViT: Mixture-of-Experts Vision Transformer for Efficient Multi-task Learning with Model-Accelerator Co-design. In *NeurIPS*.

Liang, J.; He, R.; and Tan, T. 2023. A Comprehensive Survey on Test-Time Adaptation under Distribution Shifts. *arXiv preprint arXiv:2303.15361*.

Liu, J.; Xu, R.; Yang, S.; Zhang, R.; Zhang, Q.; Chen, Z.; Guo, Y.; and Zhang, S. 2023a. Adaptive Distribution Masked Autoencoders for Continual Test-Time Adaptation. *arXiv preprint arXiv:2312.12480*.

Liu, J.; Yang, S.; Jia, P.; Lu, M.; Guo, Y.; Xue, W.; and Zhang, S. 2023b. Vida: Homeostatic visual domain adapter for continual test time adaptation. *arXiv preprint arXiv:2306.04344*.

Ma, J.; Zhao, Z.; Yi, X.; Chen, J.; Hong, L.; and Chi, E. H. 2018. Modeling task relationships in multi-task learning with multi-gate mixture-of-experts. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*.

MacQueen, J. 1967. Classification and analysis of multivariate observations. In *5th Berkeley Symp. Math. Statist. Probability*, 281–297. UCLA.

Mirzadeh, I.; Alizadeh, K.; Mehta, S.; Mundo, C. C. D.; Tuzel, O.; Samei, G.; Rastegari, M.; and Farajtabar, M. 2023. ReLU Strikes Back: Exploiting Activation Sparsity in Large Language Models. arXiv:2310.04564.

Mustafi, D.; Engel, A. H.; and Palczewski, K. 2009. Structure of cone photoreceptors. *Progress in retinal and eye research*, 28(4): 289–302.

Ni, J.; Yang, S.; Liu, J.; Li, X.; Jiao, W.; Xu, R.; Chen, Z.; Liu, Y.; and Zhang, S. 2023. Distribution-aware continual test time adaptation for semantic segmentation. *arXiv preprint arXiv:2309.13604*.

Ruder, S.; and Plank, B. 2017. Learning to select data for transfer learning with bayesian optimization. *arXiv preprint arXiv:1707.05246*.

Sakaridis, C.; Dai, D.; and Van Gool, L. 2021. ACDC: The adverse conditions dataset with correspondences for semantic driving scene understanding. In *CVPR*, 10765–10775.

Song, C.; Han, X.; Zhang, Z.; Hu, S.; Shi, X.; Li, K.; Chen, C.; Liu, Z.; Li, G.; Yang, T.; and Sun, M. 2024. ProSparse: Introducing and Enhancing Intrinsic Activation Sparsity within Large Language Models. arXiv:2402.13516.

Song, J.; Lee, J.; Kweon, I. S.; and Choi, S. 2023. EcoTTA: Memory-Efficient Continual Test-time Adaptation via Self-distilled Regularization. In *CVPR*, 11920–11929.

Tarvainen, A.; and Valpola, H. 2017a. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. *NeurIPS*, 30.

Tarvainen, A.; and Valpola, H. 2017b. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. *Learning*.

Von Helmholtz, H. 1867. *Handbuch der physiologischen Optik*, volume 9. Voss.

Wang, D.; Shelhamer, E.; Liu, S.; Olshausen, B.; and Darrell, T. 2020. Tent: Fully test-time adaptation by entropy minimization. *arXiv preprint arXiv:2006.10726*.

Wang, Q.; Fink, O.; Van Gool, L.; and Dai, D. 2022. Continual test-time domain adaptation. In *CVPR*, 7201–7211.

Xie, E.; Wang, W.; Yu, Z.; Anandkumar, A.; Alvarez, J. M.; and Luo, P. 2021. SegFormer: Simple and efficient design for semantic segmentation with transformers. *NeurIPS*, 34: 12077–12090.

Xu, Q.; Zhang, R.; Zhang, Y.; Wang, Y.; and Tian, Q. 2021. A fourier-based framework for domain generalization. In *CVPR*, 14383–14392.

Yang, C.; Chen, Y.; Tian, H.; Tao, C.; Zhu, X.; Zhang, Z.; Huang, G.; Li, H.; Qiao, Y.; Lu, L.; et al. 2023a. BEVFormer v2: Adapting modern image backbones to bird's-eye-view recognition via perspective supervision. In *CVPR*, 17830–17839.

Yang, S.; Wu, J.; Liu, J.; Li, X.; Zhang, Q.; Pan, M.; and Zhang, S. 2023b. Exploring sparse visual prompt for cross-domain semantic segmentation. *arXiv preprint arXiv:2303.09792*.

Yang, T.-H.; Cheng, H.-Y.; Yang, C.-L.; Tseng, I.-C.; Hu, H.-W.; Chang, H.-S.; and Li, H.-P. 2019. Sparse reram engine: Joint exploration of activation and weight sparsity in compressed neural networks. In *Proceedings of the 46th International Symposium on Computer Architecture*, 236–249.

Yang, Y.; and Soatto, S. 2020. Fda: Fourier domain adaptation for semantic segmentation. In *CVPR*, 4085–4095.

Yurtsever, E.; Lambert, J.; Carballo, A.; and Takeda, K. 2020. A survey of autonomous driving: Common practices and emerging technologies. *IEEE Access*, 8: 58443–58469.

Zhang, R.; Chen, Y.; Wu, C.; Wang, F.; and Li, B. 2024a. Multi-level Personalized Federated Learning on Heterogeneous and Long-Tailed Data. *arXiv preprint arXiv:2405.06413*.

Zhang, R.; Luo, Y.; Liu, J.; Yang, H.; Dong, Z.; Gudovskiy, D.; Okuno, T.; Nakata, Y.; Keutzer, K.; Du, Y.; et al. 2024b. Efficient Deweahter Mixture-of-Experts with Uncertainty-Aware Feature-Wise Linear Modulation. In *AAAI*, volume 38, 16812–16820.

Zhao, J.; Tseng, C.-C.; Lu, M.; An, R.; Wei, X.; Sun, H.; and Zhang, S. 2023. MoEC: Mixture of Experts Implicit Neural Compression. *arXiv preprint arXiv:2312.01361*.

Zhou, B.; Khosla, A.; Lapedriza, A.; Oliva, A.; and Torralba, A. 2016. Learning deep features for discriminative localization. In *CVPR*, 2921–2929.

Zhou, K.; Liu, Z.; Qiao, Y.; Xiang, T.; and Loy, C. C. 2021. Domain generalization in vision: A survey. *arXiv preprint arXiv:2103.02503*.

Zhu, Y.; Wichers, N.; Lin, C.-C.; Wang, X.; Chen, T.; Shu, L.; Lu, H.; Liu, C.; Luo, L.; Chen, J.; et al. 2023. Sira: Sparse mixture of low rank adaptation. *arXiv preprint arXiv:2311.09179*.

Zoph, B. 2022. Designing Effective Sparse Expert Models. In *IEEE IPDPS Workshops 2022, Lyon, France, May 30 - June 3, 2022*, 1044. IEEE.