MoE-TTA: Enhancing Continual Test-Time Adaptation for Vision-Language Models through Mixture of Experts

Md. Akil Raihan Iftee[1], Wahida Mahjabin[1], Anik Ekka[1], and Sunanda Das[1, 2]
[1] *Department of CSE, Khulna University of Engineering & Technology, Khulna-9203, Bangladesh.*
[2] *Department of EECS, University of Arkansas, Fayetteville, AR 72701, USA.*
{iftee1807002, mahjabin1807047, ekka2007121}@stud.kuet.ac.bd, and sunandad@uark.edu

***Abstract*—Continual learning enables vision-language models to incrementally acquire new knowledge without relying on access to the entire historical dataset. This capability is crucial for adapting to evolving data distributions in real-world scenarios, where models must handle domain shifts and incorporate new information while retaining previously learned knowledge. However, maintaining performance in large-scale models remains challenging due to parameter shifts during learning and the significant computational costs associated with full-model updates. To address these challenges, this paper introduces a novel method for Continual Test-Time Domain Adaptation (CTTDA) on vision-language datasets, leveraging a MoE and adapter modules to optimize domain-specific adaptation while maintaining zero-shot classification capabilities. Utilising LoRA in MoE framework within transformer layers, the model efficiently updates a small set of parameters by dynamically selecting experts. This approach minimizes computational costs by activating only a portion of the model, avoiding the need for full-model updates during domain adaptation. During test-time adaptation, entropy loss is calculated without access to labels, improving the model's fine-tuning and guiding towards confident predictions across domains. A contrastive warm-up phase further optimizes the adapter blocks by enhancing the differentiation of domain-specific and domain-invariant features, thereby establishing a strong foundation for effective test-time adaptation. The proposed MoE-TTA model achieves an average accuracy of 36.43% across diverse ImageNet datasets and 32.93% in fine-grained classification, demonstrating promising results, especially in datasets like EuroSAT (51.67%), while being lower than several competitors, with no doubt in its ability to capture domain shifts effectively.**

***Index Terms*—Continual Test Time Adaptation, Mixture of Expert, Domain Specific Feature, Domain-Independent Feature**

I. Introduction

In real-world environments, machine learning models often face domain shifts, where testing data significantly differs from the training data, such as autonomous vehicles handling unexpected weather conditions or medical systems dealing with new patient demographics. Retraining models in these scenarios is impractical, making Test-Time Domain Adaptation (TTDA) [1] [2] critical for dynamic adaptation without source data or target labels. CTTDA [3] [4] builds on TTDA by allowing models to continuously adapt across multiple evolving domains, while preserving prior knowledge. This scenario presents the challenge of balancing adaptability and generalization under evolving, unlabeled data.

Traditional TTDA methods, which focus on vision-based approaches, often struggle without source data or labels, relying on techniques like updating batch-normalization layers [5] or pseudo-labeling [6], which can introduce errors and hinder adaptation. These methods face challenges in separating domain-specific from generalizable features. Vision-language models like CLIP [7] leverage this synergy by utilizing both modalities, enables robust zero-shot classification and domain adaptation with prompts such as "An image of class," making it more effective in real-world applications.

In this work, the target is CTTDA using the CLIP model on a vision-language dataset where the vision encoder adapts to new, unseen domains while the text encoder remains frozen. To the best of our knowledge, this is the first work to leverage Mixture of Experts (MoE) [8] for CTTDA on a vision-language dataset. Here, adapter blocks in the vision encoder have been used with the MoE framework ensuring significant computational benefits: the dynamic selection of specialized experts allows the model to handle domain-specific features efficiently without activating the entire network. This reduces the computational overhead typically associated with full-model adaptation while still capturing domain-specific nuances. In the MoE framework, adapter modules such as LoRA [9] act as experts, accelerating the adaptation process during training. In addition, the separation of domain-invariant and domain-specific features is crucial and this approach ensures that domain-specific features are adapted for each new target domain, while domain-invariant features remain untouched, preserving the model's ability to generalize. A feature-gating mechanism that merges these feature types, ensures that both adaptability and generalization are retained. Moreover, we have employed an entropy loss to guide confident predictions in new domains during the inference time and before deployment we have integrated a contrastive warm-up phase to help the model distinguish between domain-invariant and domain-specific modules effectively.

In summary, the key contributions of this work are:

- Unique application of MoE for CTTDA on a vision-language dataset.
- Enhancement of computational efficiency by using MoE to manage parameters in CTTDA with minimal overhead,

979-8-3315-1909-4/24/$31.00 ©2024 IEEE

making it ideal for adaptive and scalable real-world applications.

II. Related Works

Continual Learning (CL) [10] addresses challenges from incremental data and domain shifts but struggles with scalability and zero-shot transfer to unseen domains. These challenges become more complex in CTTDA, where models must continually adapt to evolving target domains without access to labeled source data. Building on the foundations of CL, CTTDA requires balancing adaptability to new domains with the retention of generalizable knowledge.

Early works like TENT [5] introduced batch normalization updates for affine transformations. CoTTA [6] employed a teacher-student framework to generate and refine pseudo labels using consistency loss. EcoTTA [3] utilized meta-networks and self-distillation for memory-efficient adaptation. DePT [11] incorporated visual prompts to adapt target domains and enhance source representation. However, these methods often struggle with convergence issues due to their reliance on shared architectures. BECoTTA [12] introduces a modularized Mixture of Domain Low-Rank Experts (MoDE) architecture, where each expert captures domain-specific knowledge and enhances mutual synergy to maximize the dependency between domains and experts. However, it is designed solely as a vision-based model. Vision-language dataset on the other hand for CTTDA can enhance adaptability by leveraging both visual and textual modalities, allowing the model to use semantic cues from text prompts for more robust zero-shot classification, improving its ability to generalize across unseen domains.

The MoE framework enables efficient domain adaptation by activating only a subset of specialized experts, reducing computational costs while maintaining performance [13]. It has been employed in models like Adamix for efficient fine-tuning [14] and Meta DMoE for knowledge distillation across unlabeled domains [15]. However, these models did not address the need for continual adaptation in evolving target domains, as required in CTTDA. The potential of MoE in CTTDA, particularly on vision-language datasets, remains underexplored, despite its success in handling large-scale models across diverse domains. Recently, continual learning frameworks with MoE, such as dynamic expansion in CLIP models through MoE adapters, have shown promising results by preserving zero-shot recognition and reducing parameter tuning burdens by 60%, highlighting their potential in vision-language CTTDA scenarios [16].

In the context of CTTDA, various approaches utilize the zero-shot capabilities of CLIP [17]. Vision language models, with their multimodal nature and ability to generalize without prior training, offer promising avenues for test-time adaptation. Test-time prompt tuning (TPT) [18] dynamically optimizes prompts using a single test sample by minimizing entropy across augmented views, enhancing CLIP's zero-shot accuracy by 3.6% on average and improving generalization to unseen domains without requiring additional training data. CoOp [19] adapts CLIP-like models by turning context words into learnable vectors, achieving significant improvements over manual prompts with minimal labeled data, but struggles with generalization to unseen classes, which CoCoOp [20] addresses by introducing dynamic, instance-specific prompts to enhance adaptability. ProDA [21] improves domain adaptation by leveraging class prototypes and feature distances to correct noisy pseudo labels and compact target feature space during training, while PromptAlign [22] focuses on minimizing feature distribution shift between source and out-of-distribution (OOD) test samples through prompt tuning, improving zero-shot accuracy. Additionally, VTE Ensemble [23] employs a vision-text-space ensemble strategy, enhancing robustness under distribution shifts by combining predictions from multiple models, surpassing text-space-only ensembles in handling real-world scenarios.

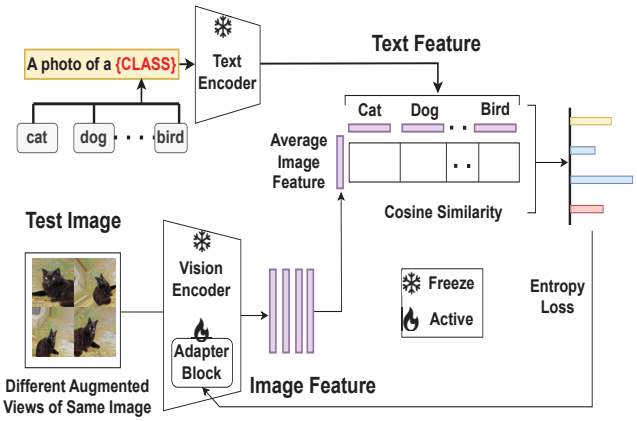

Fig. 1. Architecture for continual test-time adaptation: Test images are processed through a Clip model with an adapter block, using cosine similarity and entropy loss to adapt to domain shifts.

III. Methodology

This section describes the detailed methodology used in this study for continual test time domain adaptation using mixture of domain experts in the vision language model, Clip.

A. Problem Statement

In this work, we tackle the challenge of test-time domain adaptation using the CLIP model, where the vision encoder needs to adapt to new, unseen domains during inference. The task is to extract and adapt features that vary across domains while maintaining those that are invariant and applicable to general classification. In Figure 1, the source model, consisting of a pre-trained CLIP model, is initially given, but during test time, the model is exposed to images from various target domains that differ significantly from the source domain. The text encoder is kept frozen and provides text prompts, such as "An image of {class}," for zero-shot classification. We denote the source domain as $D_S = \{(x_S, y_S)\}$, where x_S are images and y_S are labels. The target domain is represented as $D_T = \{x_T\}$, where x_T are images encountered during

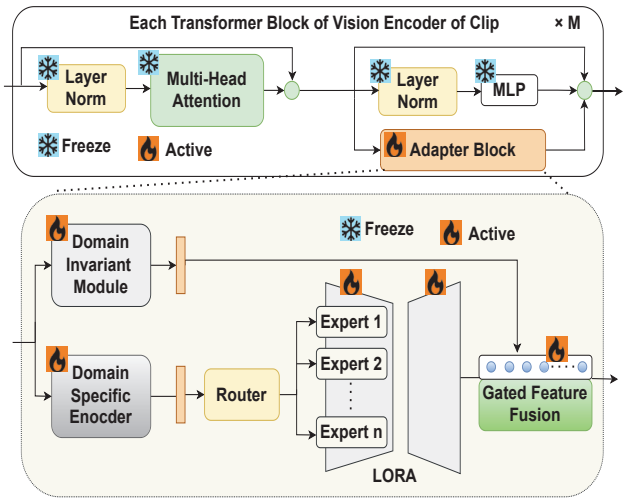

Fig. 2. Test-time adaptation architecture with a mixture of experts (MoE): after disentangling Domain-specific features and domain-invariant features, the Domain-specific features are passed to MoE mechanism to apply selective freezing for efficient adaptation across domains in inference time.

test time without labels. The vision encoder is denoted as $E_v(\cdot)$, which extracts visual features, and the text encoder as $E_t(\cdot)$. Our objective is to adapt E_v in the target domain by disentangling domain-invariant features, $\mathbf{f}_{\text{invariant}}$, and domain-specific features, $\mathbf{f}_{\text{specific}}$, in order to facilitate robust classification during test time.

B. Vision Encoder Fine-Tuning

For domain adaptation in Figure 2, only the vision encoder E_v is fine-tuned while the text encoder E_t remains frozen. During the test phase, images from the new domains are passed through the vision encoder. To handle domain shifts effectively, the vision encoder is equipped with Adapter blocks that enable the model to learn and represent both domain-specific features and domain-invariant ones. This fine-tuning process ensures that the model not only adapts to specific characteristics of the target domain but also retains the generalizable features necessary for accurate classification using the textual descriptions from the text encoder.

C. Adapter Block

The Adapter block is a critical component of our model, designed to handle the unique challenges of test-time domain adaptation. It consists of multiple sub-modules that perform various operations to disentangle and process domain-specific and domain-invariant features. The Adapter block first performs feature disentanglement, separating input features into domain-specific and domain-invariant components. The domain-invariant module extracts features common across domains, ensuring the model retains generalizable information. In contrast, the domain-specific encoder focuses on the characteristics unique to the domain currently being processed, resulting in $\mathbf{f}_{\text{specific}}$ features that capture the new domain's nuances.

After disentanglement, the domain-specific features $\mathbf{f}_{\text{specific}}$ are passed through a MoE block. A router mechanism [8] is employed to select the appropriate experts based on the domain-specific features, allowing the model to handle diverse characteristics across domains. Each expert in the MoE is trained using low-rank adaptation (LoRA) techniques to capture unique aspects of the target domains effectively. This dynamic selection and adaptation process produce an updated set of domain-specific features, $\mathbf{f}'_{\text{specific}}$, which are then ready to be merged with the domain-invariant features.

To combine the domain-specific and domain-invariant features, we introduce a feature-gating mechanism. This mechanism learns a gate vector through a small neural network, determining the appropriate contribution of each feature type to the final representation. The final representation, $\mathbf{x}_{\text{final}}$, is computed as a weighted sum of the domain-specific and domain-invariant features, ensuring both adaptability to new domains and retention of generalizable information.

D. Adaptation Through Entropy Loss

During test-time adaptation, we employ an entropy loss function to guide the vision encoder. This approach is aimed at refining the model's confidence in its predictions on target domain images. The entropy loss is defined as

$$\mathcal{L}_{\text{entropy}} = -\sum p(y|x_T) \log p(y|x_T),$$

where $p(y|x_T)$ is the probability distribution over the possible classes for the target image x_T. By minimizing this entropy, the model is encouraged to produce more confident predictions, reducing uncertainty and aligning its predictions with the domain-specific characteristics it encounters. This process effectively facilitates domain adaptation, helping the model to focus on the relevant features specific to the target domain while avoiding overfitting to the noise inherent in domain shifts.

E. Warm-Up of the Adapter Block Using Contrastive Loss

Before test-time adaptation, the Adapter block undergoes a warm-up process using contrastive loss, which helps fine-tune the modules for disentangling domain-specific and domain-invariant features. During this warm-up phase, images from the target domain are augmented to simulate the presence of multiple domains. The domain-invariant module is trained using a contrastive loss that brings the domain-invariant features of these different augmentations closer in the feature space. This encourages the module to learn a representation that remains consistent across various domains. Mathematically, this is expressed as

$$\mathcal{L}_{\text{invariant}} = -\log \frac{\exp(\text{sim}(\mathbf{f}^{(1)}_{\text{invariant}}, \mathbf{f}^{(2)}_{\text{invariant}}))}{\sum_{i=1}^{N} \exp(\text{sim}(\mathbf{f}^{(1)}_{\text{invariant}}, \mathbf{f}^{(i)}_{\text{invariant}}))}.$$

Simultaneously, the domain-specific module is trained to differentiate between augmentations, ensuring that features from

different augmentations remain distinct. The contrastive loss for this module is defined as

$$\mathcal{L}_{\text{specific}} = \log \frac{\sum_{i \neq j} \exp(-\text{sim}(\mathbf{f}_{\text{specific}}^{(i)}, \mathbf{f}_{\text{specific}}^{(j)}))}{\sum_{i=1}^{N} \sum_{j=1}^{N} \exp(-\text{sim}(\mathbf{f}_{\text{specific}}^{(i)}, \mathbf{f}_{\text{specific}}^{(j)}))}.$$

This dual contrastive training ensures that the domain-invariant module focuses on generalizable features while the domain-specific module becomes sensitive to variations introduced by domain shifts. This warm-up phase prepares the model for test-time adaptation by establishing a solid baseline for distinguishing between invariant and specific features.

IV. Results and Discussion

A. Datasets

We evaluate the proposed method in a continual test-time adaptation setting following [24] using various domain-shifted versions of ImageNet (set 1) and other fine-grained datasets (set 2). Specifically, the performance is assessed on ImageNet-C, which contains 15 types of corruptions applied to ImageNet validation images, as well as natural domain shifts from datasets such as ImageNet-R, ImageNet-Sketch, and ImageNet-D109. To evaluate adversarial robustness, ImageNet-A is included, and results on the independent test set ImageNet-V2 are also reported. To evaluate performance beyond the ImageNet domain, the proposed method is tested on several fine-grained classification datasets. The Aircraft dataset includes various airplane models, while Caltech101 covers general object categories. For transportation-related classes, Stanford Cars is used. DTD focuses on textures, and EuroSAT on satellite imagery. Flowers102 and Food101 represent flowers and food items, respectively, while Oxford-Pets features pet breeds, UCF101 targets human actions, and SUN397 captures diverse scenes. These ten datasets(set 2) enable a comprehensive assessment of the model's robustness across different domain shifts.

B. Performance comparision

The CTTDA performance, as shown in I, evaluates various methods, including the proposed MoE-TTA model, across several datasets with significant domain shifts, such as ImageNet, ImageNet-C, ImageNet-A, ImageNet-V2, ImageNet-R, ImageNet-S, and ImageNet-D109. MoE-TTA achieves an average accuracy of 36.43%, which, while lower than several competitors, demonstrates competitive performance on specific datasets. Notably, MoE-TTA performs best on ImageNet-V2 with 30.9% accuracy and lowest on ImageNet at 26.1%, highlighting its strengths and weaknesses in adapting to varying domain conditions. For datasets of ImageNet variations, MoE-TTA leverages selective freezing and low-rank adaptation to efficiently fine-tune domain-specific parameters, outperforming traditional methods in scalability

The comparative analysis shows that Zero-shot CLIP leads with an average accuracy of 43.64%, followed by TPT and CoOp, while MoE-TTA lags behind but exhibits potential for improvement through continual adaptation. MoE-TTA achieves an average accuracy of 32.93% in fine-grained classification tasks, as summarized in II, with notable results on datasets like EuroSAT (51.67%) but lower performance on more challenging datasets like Cars (29.50%) and UCF101 (30.07%). Despite Zero-shot CLIP's average accuracy of 36.42%, MoE-TTA demonstrates resilience in adapting to domain shifts and shows promise for handling unseen categories beyond the ImageNet domain.

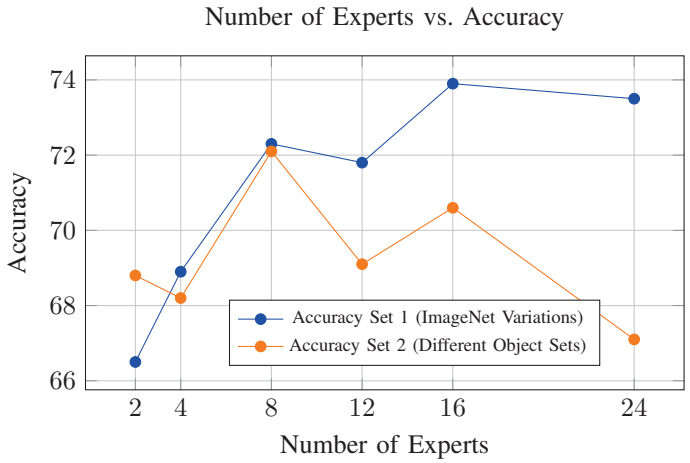

Fig. 3. The Accuracy in set 1 and set 2 of the dataset sequence fluctuates with variations of the Number of Experts of moe-tta method

Lastly, MoE-TTA achieves significant computational efficiency by activating only 61.54 million parameters during adaptation, compared to the 149.62 million parameters required by the current best-performing VTE Ensemble [23].

C. Analysis of Mixure of Different Domain Experts in CTTA

The figure 3 provides insights into how the number of domain-specific experts affects accuracy in a multi-domain task using the MoE-TTA (Mixture of Experts - Test-Time Adaptation) method, where each expert specializes in a particular domain feature. For Accuracy Set 1 (ImageNet Variations, blue), performance increases steadily up to 16 experts, achieving a peak accuracy of 73.9%, indicating that using 16 experts offers the best performance for this dataset. In contrast, Accuracy Set 2 (Different Object Sets, orange) sees its highest accuracy at 8 experts (72.1%), after which performance declines, suggesting that fewer experts may be more suitable for this dataset. Considering the fluctuations in both datasets, 8 experts would be preferable for balancing performance across different domain-specific tasks. However, for tasks focused solely on domains similar to Accuracy Set 1, 16 experts may be optimal. The results also underline the risk of overfitting when increasing experts for smaller datasets, as performance deteriorates beyond 8 experts, whereas larger datasets like ImageNet benefit from a higher number of experts.

Understanding the contribution of individual experts in the MoE-TTA model is crucial for optimizing its performance. Figure 4 illustrates the percentage contribution of four experts across various transformer encoder layers. The results reveal

TABLE I

EVALUATION OF MOE-TTA IN A TTDA SCENARIO. DIFFERENT VARIATIONS OF ADAPTATION IN THE CLIP MODEL FROM EXISTING RESEARCH ARE ASSESSED ON DATASETS OF IMAGENET VARIATIONS (SET 1) EXPERIENCING DOMAIN SHIFTS AND SHOWN THE AVERAGE ERROR RATES(%)

Methods	ImageNet	ImageNet-C	ImageNet-A	ImageNet-V2	ImageNet-R	ImageNet-S	ImageNet-D109	Average
Zero-shot CLIP [17]	33.1	75.2	49.3	39.5	26.4	53.7	27.3	43.64
TPT [18]	29.9	74.8	45.8	35.9	23.7	52.4	26.9	41.34
CoOp [19]	29.2	73.1	48.6	35.7	23.4	48.2	26.1	40.63
CoCoOp [20]	29.5	73.8	43.9	35.6	**18.1**	49.9	25.7	39.36
ProDA [21]	30.7	**72.4**	41.2	36.4	22.1	50.3	25.8	39.84
PromtAlign [22]	28.3	74.6	41.1	35.1	21.8	**43.2**	26.5	38.65
VTE Ensemble [23]	28.2	72.9	36.4	34.6	20.1	49.7	24.5	38.06
MoE-TTA	**26.1**	73.4	**36.1**	**30.9**	18.4	44.1	**24.0**	**36.43**

TABLE II

TO ASSESS CATEGORIES BEYOND THE IMAGENET DOMAIN, VARIOUS EXISTING RESEARCHES OF THE ADAPTIVE CLIP MODEL ARE EVALUATED ACROSS TEN DATASETS, EACH INVOLVING CLASSIFICATIONS (SET 2) UNDERGOING DOMAIN SHIFTS AND SHOWN THE AVERAGE ERROR RATES(%)

Methods	Aircraft	Caltech	Cars	DTD	EuroSAT	Flowers	Food101	Pets	SUN397	UCF101	Average
Zero-shot CLIP [17]	76.33	6.65	34.52	55.73	57.99	32.56	16.35	11.75	37.41	34.87	36.42
TPT [18]	75.22	**5.84**	33.13	**52.25**	55.00	29.65	15.70	10.98	35.69	32.44	35.14
CoOp [19]	74.99	5.98	32.52	54.31	54.76	29.23	15.22	11.01	32.78	29.32	34.10
CoCoOp [20]	73.56	5.66	30.75	53.01	51.67	26.79	14.52	10.33	29.22	27.34	32.23
ProDA [21]	72.40	5.99	31.47	53.45	53.33	25.46	14.36	10.92	28.56	26.87	31.92
PromtAlign [22]	69.12	5.52	30.13	49.98	52.98	25.01	14.22	10.77	27.41	25.30	30.16
VTE Ensemble [23]	68.87	5.85	29.72	48.54	52.54	24.12	14.67	10.45	26.33	24.87	29.94
MoE-TTA	**66.22**	5.81	**29.50**	50.25	**51.67**	**27.80**	**15.67**	**10.96**	**29.82**	**30.07**	**32.93**

that different experts contribute varying degrees across layers, with Expert 1 demonstrating a consistent contribution of around 24-30%, particularly in earlier layers, while Expert 4 shows a more pronounced influence in later layers (35% at Layer 10). This variability indicates that different experts may be specialized for certain tasks or features, thus enhancing the overall model performance through a collective learning mechanism.

To further analyze the effectiveness of the MoE-TTA model, we examined the feature representations through t-SNE visualizations. Figure 5 presents these representations across domains (Fog, Gaussian, Shot, Elastic) and selected ImageNet-C classes (Cat, Giraffe, Airplane). The results highlight that domain-invariant features exhibit a blended representation across different domains, with clear class separation, while domain-specific features show more distinct separations along domain boundaries. This distinction confirms the effectiveness of the MoE-TTA model in separating domain-invariant from domain-specific features, thereby enhancing the robustness of the model against varying domain conditions.

Overall, the results indicate that the MoE-TTA model demonstrates a viable approach for continual test-time adaptation in the face of domain shifts. While its performance metrics are competitive, they also underscore the need for further refinement to enhance accuracy, particularly in challenging datasets.

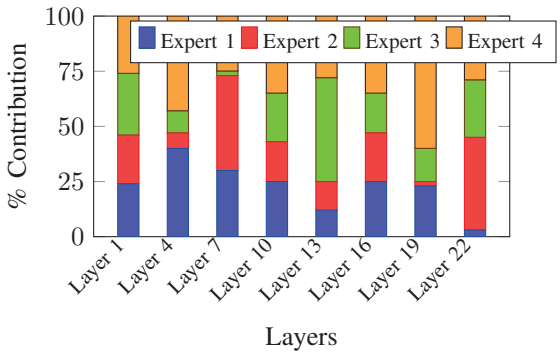

Fig. 4. Activate the expert ratio in MoE-TTA for four experts across various transformer encoder layers in the Vision Encoder of the CLIP model.

V. CONCLUSION

This paper presents MoE-TTA, to the best of our knowledge, a novel framework for CTTDA that integrates the MoE architecture with LoRA for efficient domain adaptation on vision-language datasets. The model achieves competitive performance, particularly in specific datasets such as EuroSAT (51.67%), demonstrating its potential to handle diverse domain shifts with reduced computational complexity. However, its performance, while promising, is lower than several competitors in certain datasets, indicating areas for improvement. The results highlight MoE-TTA's ability to efficiently separate domain-specific and domain-invariant features, which is crucial for adapting to unseen domains. Despite these strengths, limitations remain, particularly in adapting to more challenging datasets such as Cars and UCF101. The reliance on labeled data for the Adapter block's warm-up process using contrastive loss means MoE-TTA is not entirely unsupervised, limiting its adaptability in fully unsupervised scenarios. Future work will focus on refining the MoE-TTA framework to enhance its accuracy, exploring advanced strategies for more robust feature disentanglement and optimizing expert selection in dynamic

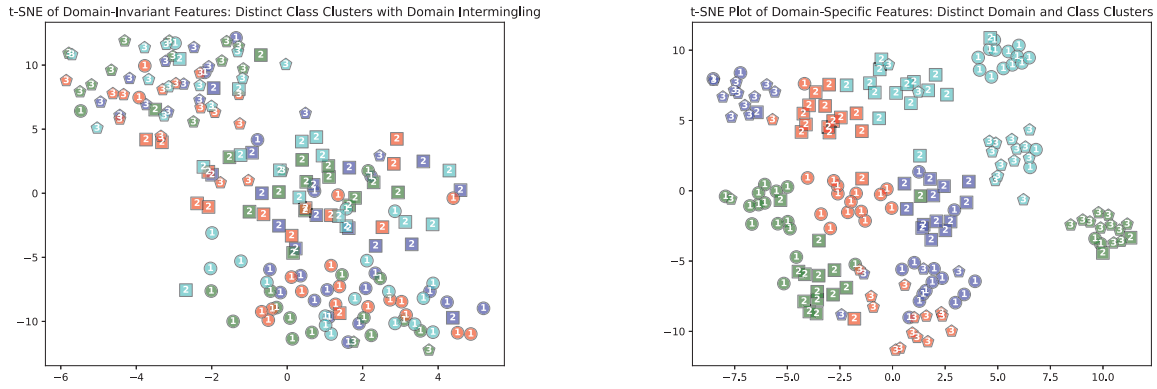

Fig. 5. Feature representations (t-SNE) of four domains (Fog, Gaussian, Shot, Elastic) and three ImageNet-C classes (Cat, Giraffe, Airplane). Domain-invariant features show intermingled domains, while domain-specific features exhibit clear separation between domains along with distinct class separation.

environments. Moreover, extending the model's application to a broader range of datasets and investigating its scalability across larger architectures will further solidify its applicability in real-world, evolving domain settings.

REFERENCES

[1] S. Niu, J. Wu, Y. Zhang, Z. Wen, Y. Chen, P. Zhao, and M. Tan, "Towards stable test-time adaptation in dynamic wild world," *arXiv preprint arXiv:2302.12400*, 2023.

[2] H. Lim, B. Kim, J. Choo, and S. Choi, "Ttn: A domain-shift aware batch normalization in test-time adaptation," *arXiv preprint arXiv:2302.05155*, 2023.

[3] J. Song, J. Lee, I. S. Kweon, and S. Choi, "Ecotta: Memory-efficient continual test-time adaptation via self-distilled regularization," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11920–11929, 2023.

[4] T. Lee, J. Tremblay, V. Blukis, B. Wen, B.-U. Lee, I. Shin, S. Birchfield, I. S. Kweon, and K.-J. Yoon, "Tta-cope: Test-time adaptation for category-level object pose estimation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 21285–21295, 2023.

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

[6] Q. Wang, O. Fink, L. Van Gool, and D. Dai, "Continual test-time domain adaptation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7201–7211, 2022.

[7] P. Gao, S. Geng, R. Zhang, T. Ma, R. Fang, Y. Zhang, H. Li, and Y. Qiao, "Clip-adapter: Better vision-language models with feature adapters," *International Journal of Computer Vision*, vol. 132, no. 2, pp. 581–595, 2024.

[8] S. Pavlitska, C. Hubschneider, L. Struppek, and J. M. Zöllner, "Sparsely-gated mixture-of-expert layers for cnn interpretability," in *2023 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–10, IEEE, 2023.

[9] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen, "Lora: Low-rank adaptation of large language models," *arXiv preprint arXiv:2106.09685*, 2021.

[10] R. Aljundi, P. Chakravarty, and T. Tuytelaars, "Expert gate: Lifelong learning with a network of experts," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3366–3375, 2017.

[11] Y. Gao, X. Shi, Y. Zhu, H. Wang, Z. Tang, X. Zhou, M. Li, and D. N. Metaxas, "Visual prompt tuning for test-time domain adaptation," *arXiv preprint arXiv:2210.04831*, 2022.

[12] D. Lee, J. Yoon, and S. J. Hwang, "Becotta: Input-dependent online blending of experts for continual test-time adaptation," *arXiv preprint arXiv:2402.08712*, 2024.

[13] B. Mustafa, C. Riquelme, J. Puigcerver, R. Jenatton, and N. Houlsby, "Multimodal contrastive learning with limoe: the language-image mixture of experts," *Advances in Neural Information Processing Systems*, vol. 35, pp. 9564–9576, 2022.

[14] Y. Wang, S. Mukherjee, X. Liu, J. Gao, A. H. Awadallah, and J. Gao, "Adamix: Mixture-of-adapter for parameter-efficient tuning of large language models," *arXiv preprint arXiv:2205.12410*, vol. 1, no. 2, p. 4, 2022.

[15] T. Zhong, Z. Chi, L. Gu, Y. Wang, Y. Yu, and J. Tang, "Meta-dmoe: Adapting to domain shift by meta-distillation from mixture-of-experts," *Advances in Neural Information Processing Systems*, vol. 35, pp. 22243–22257, 2022.

[16] J. Yu, Y. Zhuge, L. Zhang, P. Hu, D. Wang, H. Lu, and Y. He, "Boosting continual learning of vision-language models via mixture-of-experts adapters," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 23219–23230, 2024.

[17] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, *et al.*, "Learning transferable visual models from natural language supervision," in *International conference on machine learning*, pp. 8748–8763, PMLR, 2021.

[18] M. Shu, W. Nie, D.-A. Huang, Z. Yu, T. Goldstein, A. Anandkumar, and C. Xiao, "Test-time prompt tuning for zero-shot generalization in vision-language models," *Advances in Neural Information Processing Systems*, vol. 35, pp. 14274–14289, 2022.

[19] K. Zhou, J. Yang, C. C. Loy, and Z. Liu, "Learning to prompt for vision-language models," *International Journal of Computer Vision*, vol. 130, no. 9, pp. 2337–2348, 2022.

[20] K. Zhou, J. Yang, C. C. Loy, and Z. Liu, "Conditional prompt learning for vision-language models," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 16816–16825, 2022.

[21] P. Zhang, B. Zhang, T. Zhang, D. Chen, Y. Wang, and F. Wen, "Prototypical pseudo label denoising and target structure learning for domain adaptive semantic segmentation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 12414–12424, 2021.

[22] J. Abdul Samadh, M. H. Gani, N. Hussein, M. U. Khattak, M. M. Naseer, F. Shahbaz Khan, and S. H. Khan, "Align your prompts: Test-time prompting with distribution alignment for zero-shot generalization," *Advances in Neural Information Processing Systems*, vol. 36, 2024.

[23] M. Döbler, R. A. Marsden, T. Raichle, and B. Yang, "A lost opportunity for vision-language models: A comparative study of online test-time adaptation for vision-language models," *arXiv preprint arXiv:2405.14977*, 2024.

[24] R. A. Marsden, M. Döbler, and B. Yang, "Universal test-time adaptation through weight ensembling, diversity weighting, and prior correction," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 2555–2565, 2024.