
Supplementary Materials - ViDA: Homeostatic Visual Domain Adapter for Continual Test Time Adaptation

Anonymous Author(s)

Affiliation

Address

email

1 Overview

The supplementary materials presented in this paper offer a comprehensive quantitative and qualitative analysis of the proposed method. In Section 2, we provide additional implementation details regarding the evidence of motivation. And we also present an extra continual adaptation experiment for Foundation Models in Section 3.1, which is conducted on ImageNet-to-ImageNet-C. To evaluate the domain generalization ability of our method, we conducted additional experiments directly testing on a different number of unseen domains in Section 3.2. Ablation study on prototype dimension are described in Section 3.3. Furthermore, Section 3.4 presents additional CTTA classification experiments utilizing the convolutional backbone. We offer an additional qualitative analysis in Section 4. In Section 5, we present a more detailed discussion of related work. Moreover, we extend the classification results of our submission to include a fine-grained performance, which examines the error across fifteen corruption types. The checklist is presented in the final section of this report.

2 Supplementary Details for Motivation

The study of Continual Test-Time Adaptation (CTTA) poses significant challenges, particularly in addressing error accumulation and catastrophic forgetting [36, 10]. Notably, the use of adapters with prototypes of varying dimensions has demonstrated promising results in mitigating these challenges in our submission. In this section, we aim to provide comprehensive implementation details regarding the evidence supporting our motivation. Furthermore, we delve deeper the underlying principles behind the effective utilization of domain adapters.

Distribution Qualitative Analysis We employ t-distributed stochastic neighbor embedding (t-SNE) [33] to visualize the distribution of adapters across four continual target domains. This visualization is specifically performed on the Cityscapes-to-ACDC experiment, which represents a scenario with continually changing environments in the real world. To conduct the t-SNE analysis, we analyze the output of the third transformer block in the Segformer-B5 model [37]. The objective is to qualitatively compare the feature distributions of adapters with different dimension prototypes. Furthermore, our findings reveal that the qualitative results obtained from different layers of the Segformer-B5 model exhibit similar distribution representations. Illustrated in Figure 1(b) of our submission, there is a noticeable distribution gap due to the significant domain shift between the night domain and other domains. Interestingly, the low-rank Visual Domain Adapter (ViDA) effectively reduces the distribution distance across different target domains. On the other hand, the high-rank ViDA exhibits notable distribution discrepancies among the various target domains, indicating its focus on extracting domain-specific knowledge.

Distribution Distance To provide clearer evidence for our assumption, we directly calculate the distribution distance to represent different domain representation of adapters. We adopt the domain distance definition proposed by Ben-David [2, 1] and build upon previous domain transfer research [11] by employing the \mathcal{H} -divergence metric to further evaluate the domain representations of

adapters across different target domains. \mathcal{H} -divergence between D_S and D_{T_i} can be calculated as

$$d_{\mathcal{H}}(D_S, D_{T_i}) = 2 \sup_{\mathcal{D} \sim \mathcal{H}} |\Pr_{x \sim D_S} [\mathcal{D}(x) = 1] - \Pr_{x \sim D_{T_i}} [\mathcal{D}(x) = 1]| \quad (1)$$

, where \mathcal{H} denotes hypothetical space and \mathcal{D} denotes discriminator. Similar to [11], due to our model without discriminator architecture, calculating the \mathcal{H} -divergence directly is challenging. For low-rank adapter evaluation, we adopt the *Jensen-Shannon (JS) divergence* between two adjacent domains as an approximation.

$$JS(P_{D_S} || P_{D_{T_i}}) = \frac{1}{2} KL(P_{D_S} || \frac{P_{D_S} + P_{D_{T_i}}}{2}) + \frac{1}{2} KL(P_{D_{T_i}} || \frac{P_{D_S} + P_{D_{T_i}}}{2}) \quad (2)$$

Where *Kullback-Leibler (KL) divergence* between two domain is

$$KL(P_1 || P_2) = \sum_{i=0}^n P_1(x_i) \log(\frac{P_1(x_i)}{P_2(x_i)}) \quad (3)$$

Where P denotes probability distribution of model output features. We split the output feature space into mutually disjoint intervals x_i . n range from 0 to 1000. To investigate the effectiveness of adapters in adapting to continual target domains, we compare the *JS* values obtained by using the source model alone, injecting low-rank adapter, and combining low-high adapters, as illustrated in Figure 3(a) of our submission. For high-rank adapter, we use normalized intra-class divergence to further verify the domain representations of high-rank adapters in CIFAR10C, which is inspired by intra-cluster dissimilarity proposed by k -means [26]. We first calculate the Euclidean distance clustering center for each category $\mu = \frac{1}{|C|} \sum_{e_i \sim C} e_i$, where e_i stands for output feature in class C . Then following [26], we introduce normalized intra-class divergence E by

$$E = \phi(\frac{1}{|C|} \sum_{e_i \sim C} \|e_i - \mu\|_2^2) \quad (4)$$

$\phi(\cdot)$ denotes for normlization function. As illustrated in Figure 3(b) of the submission, the high-rank adapter is found to drive down divergence within almost all domains and can better extract domain-specific knowledge in target domains.

Distribution Quantitative Analysis To provide stronger evidence for our assumption, we have developed two evaluation approaches for both low-rank and high-rank adapters, which directly reflect their ability to extract domain-agnostic and domain-specific knowledge on ImageNet-to-ImageNet-C. After completing the entire process of continual adaptation on fifteen target domains, we employ the model and adapters from the last target domain to directly test on previously seen target domains, thereby evaluating the extent of catastrophic forgetting. As anticipated, we observe a noteworthy overall improvement of 1.0% in the average classification error, as demonstrated in the final row of Table 1 (submission). These findings provide additional support to our assumptions and indicate that utilizing low-rank adapters can mitigate catastrophic forgetting. Secondly, we conduct an ablation study on ImageNet-to-ImageNet-C to demonstrate the effectiveness of the high-rank adapter. In this study, we solely introduce the high-rank adapter into the pre-trained model. The results, presented in Table 6 (Ex_2) of the submission, reveal a sustained reduction (-4.6%) in the classification error rate within the dynamic target domains when employing the high-rank adapter. This finding provides support for our hypothesis that high-rank adapters can extract more dependable domain-specific knowledge. These evaluation approaches serve to strengthen our findings and contribute to a clearer understanding of the capabilities and advantages of both low-rank and high-rank adapters.

3 Additional Experiment

3.1 Additional Continual Adaption Experiments for Foundation Models

To demonstrate the effectiveness of our proposed method in enhancing the continual adaptation ability of foundation models such as DINOv2 [27] and SAM [19], we conducted additional experiments on a more extensive dataset, namely ImageNet-to-ImageNet-C. Our approach involved loading the weight parameters of the foundation model and pre-training it on ImageNet, thus constructing our source

Table 1: Average error rate (%) for the ImageNet-to-ImageNet-C CTTA task. All results are evaluated on the ViT-Base, which uses the pre-trained encoder parameter of DINOv2 [27] and SAM [19].

Backbone	Method	REF	Gaussian	shot	impulse	defocus	glass	motion	zoom	snow	frost	fog	brightness	contrast	elastic_trans	pixelate	jpeg	Mean↓	Gain
DINOv2	Source	ICLR2021 CVPR2022 Proposed	52.3	50.5	51.2	57.3	83.8	60.1	62.6	47.1	56.9	58.1	22.5	88.4	60.3	32.4	35.0	54.6	0.0
	Tent [35]		51.7	43.6	50.4	56.2	74.1	51.7	67.2	46.9	53.2	50.1	25.2	69.6	58.0	29.5	39.4	51.1	+3.5
	CoTTA [36]		51.4	62.1	50.4	78.3	75.2	62.8	60.3	48.4	59.0	58.8	31.6	90.7	49.2	39.1	36.5	56.9	-2.3
	Ours		49.0	49.8	50.7	61.4	60.2	49.7	42.6	47.1	51.9	45.3	27.1	49.7	47.4	32.0	29.4	46.2	+8.4
SAM	Source	ICLR2021 CVPR2022 Proposed	67.9	62.1	51.6	69.7	92.6	65.4	59.8	53.9	61.2	64.1	39.0	91.6	60.1	47.3	67.0	63.6	0.0
	Tent [35]		67.2	59.1	48.8	56.2	72.5	59.4	61.0	49.1	57.9	63.7	33.8	77.0	51.4	39.5	55.2	55.5	+8.1
	CoTTA [36]		68.1	64.5	50.4	67.1	80.1	68.9	67.0	63.1	69.5	61.4	40.6	88.2	58.3	43.5	68.4	63.9	-0.3
	Ours		59.9	55.7	40.2	84.3	49.6	59.7	59.0	47.8	48.3	57.4	26.6	71.8	42.9	41.7	50.3	53.0	+10.6

model. It is important to note that we solely utilized the pre-trained encoder of SAM and incorporated a classification head, which was fine-tuned on the source domain. Subsequently, we adapted the source model to continual target domains (ImageNet-C) comprising fifteen corruption types. The results, as depicted in Table 1, demonstrate that our approach achieved a significant performance improvement of 8.4% on the representative image-level foundation model DINOv2 [27] and 10.6% on the pixel-level foundation model SAM [19]. These outcomes highlight the effectiveness of our method when applied to large-scale models. Our method consistently and reliably enhances the performance of the foundation model across unseen continual target domains.

3.2 Domain Generalization on a Different Number of Unseen Target Domains

Table 2: We performed domain generalization comparisons on ImageNet-C, where the source model was continually adapted on the first 5 domains and directly tested on 10 unseen domains. The evaluation of the results was conducted using ViT-base.

	Directly test on 10 unseen domains										Unseen
Method	motion	zoom	snow	frost	fog	bri.	contrast	elastic_trans	pixelate	jpeg	Mean↓
Source	58.5	63.3	49.9	54.2	57.7	26.4	91.4	57.5	38.0	36.2	53.3
Tent [35]	56.0	61.3	45.7	49.6	56.6	24.8	94.0	55.6	37.1	35.1	51.6
CoTTA [36]	57.3	62.1	49.1	52.0	57.1	26.4	91.9	57.1	37.6	35.3	52.6
Ours	46.4	52.7	39.8	43.7	42.2	23.5	71.5	49.6	33.9	33.3	43.7

Table 3: We performed domain generalization comparisons on ImageNet-C, where the source model was continually adapted on the first 7 domains and directly tested on 8 unseen domains. The evaluation of the results was conducted using ViT-base.

	Directly test on 8 unseen domains								Unseen
Method	snow	frost	fog	bri.	contrast	elastic_trans	pixelate	jpeg	Mean↓
Source	49.9	54.2	57.7	26.4	91.4	57.5	38.0	36.2	51.4
Tent [35]	44.3	48.8	51.8	24.9	83.7	55.2	35.4	34.7	47.4
CoTTA [36]	48.8	52.2	56.7	26.1	91.1	57.0	37.3	35.3	50.6
Ours	39.6	43.7	41.7	23.7	63.7	51.7	33.3	33.6	42.3

Similar to our previous submission, we follow the leave-one-domain-out principle [41, 24], where we utilize a subset of ImageNet-C domains as new source domains for model training, while leaving the remaining domains as target domains without any adaptation. However, in contrast to previous domain generalization experiments, we adopt an unsupervised continual test-time adaptation (CTTA) approach for training the model on these new source domains. We solely utilize the ImageNet pre-trained parameters as the initial weights of the model. In the supplementary material, we utilize 5 out of 15 and 7 out of 15 domains from ImageNet-C as the source domains, leaving the remaining 10 out of 15 and 8 out of 15 domains as unseen target domains. Surprisingly, the results presented in Table 2 and 3 demonstrate that our method achieves a reduction of 9.6% and 9.1% in the average error on these unseen domains, respectively. These promising outcomes validate the DG ability of our method, as it effectively extracts domain-agnostic knowledge and provides a new perspective for enhancing DG performance within an unsupervised paradigm.

3.3 Additional Ablation study

How do the prototype dimension influence the performance?

According to Figure 1, we observe that as the dimensionality decreases, the error rate concurrently drops. This trend suggests that lower-dimension prototypes more effectively extract the domain-shared knowledge, leading to an improved model performance. However, an opposite trend emerges when dimensionality surpasses 16, with performance enhancements accompanying increased dimensionality. This correlation implies that prototypes with a higher dimensionality excel in extracting domain-specific knowledge. And we find that when the dimension is larger than 128, the performance improvement is limited but brings a larger number of parameters. Therefore, we set the dimensionality of the high-dimension prototype to 128 in our study.

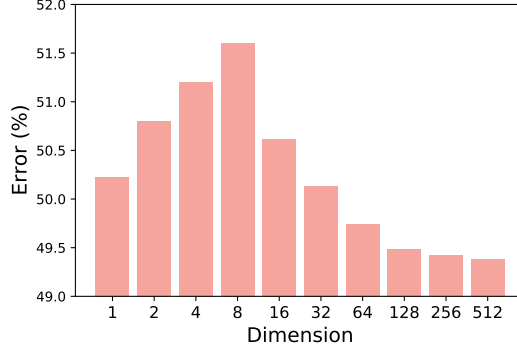


Figure 1: The prototype dimension influence of the performance

3.4 Additional Experiments on Classification CTTA

Table 4: Classification error rate(%) for standard CIFAR10-to-CIFAR10C online CTTA task. Results are evaluated on WideResNet-28. Gain(%) represents the percentage of improvement in model accuracy compared with the source method.

Method	REF	Gaussian	shot	impulse	defocus	glass	motion	zoom	snow	frost	fog	brightness	contrast	elastic_trans	pixelate	jpeg	Mean↓	Gain
Source [39]	BMVC2016	72.3	65.7	72.9	46.9	54.3	34.8	42.0	25.1	41.3	26.0	9.3	46.7	26.6	58.5	30.3	43.5	0.0
BN Stats Adapt [31]	NeurIPS2020	28.1	26.1	36.3	12.8	35.3	14.2	12.1	17.3	17.4	15.3	8.4	12.6	23.8	19.7	27.3	20.4	+23.1
Tent-continual [35]	ICLR2021	24.8	20.6	28.6	14.4	31.1	16.5	14.1	19.1	18.6	18.6	12.2	20.3	25.7	20.8	24.9	20.7	+22.8
CoTTA [36]	CVPR2022	24.3	21.3	26.6	11.6	27.6	12.2	10.3	14.8	14.1	12.4	7.5	10.6	18.3	13.4	17.3	16.2	+27.3
SATA [5]	2023.4.20	23.9	20.1	28.0	11.6	27.4	12.6	10.2	14.1	13.2	12.2	7.4	10.3	19.1	13.3	18.5	16.1	+27.4
Ours	Proposed	24.1	20.6	24.1	11.5	26.5	12.3	10.3	14.7	13.4	12.3	8.3	10.9	17.7	12.9	16.8	15.8	+27.7

CIFAR10-to-CIFAR10C standard task. In contrast to the experiments conducted in our submission, we introduce a change in the backbone of the classification model to WideResNet-28, which is consistent with previous works on CTTA [36]. Specifically, we modify the up-projection layer and down-projection layer to utilize 1×1 convolutions, while the adapters are placed alongside the original 3×3 convolutions. For the adapters, we maintain a low-rank dimension of 1 and a high-rank dimension of 128. Additionally, we provide the pre-trained source model and perform CTTA on CIFAR10C, a dataset that encompasses fifteen corruption types that occur sequentially during the test time. As depicted in Table 5, our method surpasses all previous approaches, achieving a 27.7% and 0.3% improvement over the source model and the previous state-of-the-art (SOTA) method. These findings demonstrate that our method successfully extracts domain-specific knowledge, regardless of the network backbone employed.

4 Additional Qualitative Analysis

To further validate the effectiveness of our proposed method, we present additional qualitative comparisons on the Cityscapes-to-ACDC CTTA scenario. Initially, we pre-train the Segformer-B5 model [37] on the source domain and subsequently adapt it to four target domains in ACDC. In order

to assess the performance of our approach, we conduct a qualitative comparison with two leading methods, namely CoTTA [36] and VDP [10]. The visualizations of the segmentation outputs, obtained through the CTTA process, are depicted in Figure .2. Our method exhibits better segmentation map compared to CoTTA and VDP across all four target domains, as it effectively distinguishes the sidewalk from the road (shown in white box). This demonstrates the capability of our method to achieve more accurate segmentation results while mitigating the impact of dynamic domain shifts. Moreover, in the other categories, our method’s segmentation maps closely resemble the Ground Truth, leading to a visual enhancements. Lastly, we have included a video visualization in the supplementary material that showcases a comprehensive comparison of segmentation performance. This video provides a dynamic and visual representation of the results obtained from our experiments.

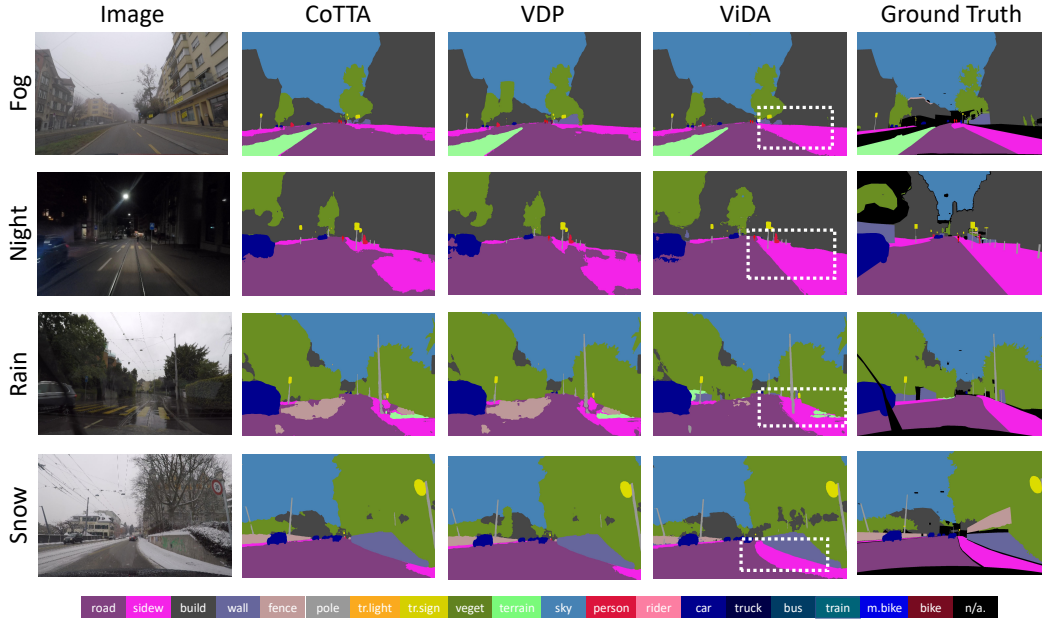


Figure 2: Qualitative comparison of our method with previous SOTA methods on the ACDC dataset. Our method could better segment different pixel-wise classes such as shown in the white box.

5 Additional Related Work

5.1 Continual Test-Time Adaptation

Test-time adaptation (TTA), also known as source-free domain adaptation [3, 20, 25, 38], aims to adapt a source model to an unknown target domain distribution without relying on any source domain data. In practical scenarios where data privacy and transmission costs are a concern, access to source domain data may be limited, rendering traditional unsupervised domain adaptation (UDA) algorithms ineffective. Recent research has explored various techniques such as self-training and entropy regularization to fine-tune the source model [21, 35, 25, 6]. Specifically, Tent [35] updates the training parameters in batch normalization layers by minimizing entropy. SHOT [25] focuses on optimizing only the feature extractor using information maximization and pseudo-labeling. AdaContrast [6] combines pseudo-labeling with self-supervised contrastive learning to improve performance. Additionally, some studies have approached the problem from an output distribution adjustment perspective [4]. Recently, SVDP introduces the sparse domain prompt to address the pixel-level domain shift for segmentation TTA task. While the works mentioned above primarily concentrate on convolutional neural networks, there has been a recent surge of interest in applying Transformer-based models [36, 13, 13].

Continual Test-Time Adaptation (CTTA) refers to a scenario where the target domain is dynamic, presenting additional challenges for traditional TTA methods. The pioneering work by [36] introduced a comprehensive approach that combines bi-average pseudo labels and stochastic weight reset to address this complex task. While [36] focuses on tackling the problem at the model level for

both classification and segmentation tasks, [10] proposes the use of visual domain prompts at the input level, specifically for the classification task. In a similar vein, inspired by Tent [35], SATA [5] modifies the batch-norm affine parameters through a source anchoring-based self-distillation scheme. Building upon these advancements, our paper takes a pioneering step by introducing visual domain adapters to address the challenges of error accumulation and catastrophic forgetting in CTTA. By simultaneously tackling both classification tasks and dense prediction tasks, our approach provides a holistic solution for CTTA.

5.2 Parameter-Efficient Fine-Tuning

Selective updating or introducing a small subset of parameters instead of updating all parameters in a pre-trained model during standard fine-tuning has been shown to be effective. In the field of natural language processing (NLP), Parameter-Efficient Fine-Tuning (PEFT) has gained considerable traction, with notable studies including [18, 16, 15, 40, 23, 17, 12, 14, 34, 28]. Adapter-based models, a type of PEFT, have emerged as popular techniques in NLP. These models utilize bottleneck architecture adapter modules that are inserted between layers of pre-trained models, and during fine-tuning, only these adapter modules are updated. Adapter-based models have demonstrated superior performance in certain tasks, sometimes surpassing standard fine-tuning approaches [8]. The success of adapters in NLP has also led to widespread interest in applying adapter techniques to visual tasks. During the early stages of adapter development, residual adapter modules were introduced as a means to facilitate effective adaptation of convolutional neural networks across multiple downstream tasks [29, 30]. Building on this foundation, AdaptFormer [7] improves the ViT [9] model by replacing the original multi-layer perceptron (MLP) block with AdaptMLP. AdaptMLP introduces a trainable down-to-up bottleneck module in a parallel manner, effectively mitigating catastrophic interference between tasks. Another notable approach, VL-Adapter [32], enhances the efficiency and performance of adapters by leveraging shared low-dimensional layer weights to transfer knowledge across tasks. Despite these advancements, existing methods have not adequately addressed the challenges of long-term preservation of domain-agnostic knowledge and the timely exploration of domain-specific knowledge in the face of continuous unknown domain variations. Consequently, there is an urgent demand for an adapter that can simultaneously address the challenges of error accumulation and catastrophic forgetting through specific adapters with different domain representations.

6 Fine-grained Performance

Table 5: A fine-grained Classification error rate(%) for standard CIFAR10-to-CIAFAR10C online CTTA task. Results are evaluated on ViT-base.

Method	gaussian	shot	impulse	defocus	glass	motion	zoom	snow	frost	fog	bri.	contrast	elastic_trans	pixelate	jpeg	Mean↓	Gain
Source	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
Pseudo-label [22]	59.8	52.5	37.2	19.8	35.2	21.8	17.6	11.6	12.3	20.7	5.0	41.7	21.5	25.2	22.1	26.9	+1.3
Tent-continual [35]	58.7	51.8	34.2	18.9	33.5	21.6	16.4	10.8	11.7	18.6	4.7	38.5	20.6	22.1	20.8	25.5	+2.7
CoTTA [36]	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[10]	57.5	49.5	31.7	21.3	35.1	19.6	15.1	10.8	10.3	18.1	4	27.5	18.4	22.5	19.9	24.1	+4.1
Ours (proposed)	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

Table 6: A fine-grained Classification error rate(%) for standard CIFAR100-to-CIAFAR100C online CTTA task. Results are evaluated on ViT-base.

Method	gaussian	shot	impulse	defocus	glass	motion	zoom	snow	frost	fog	bri.	contrast	elastic_trans	pixelate	jpeg	Mean↓	Gain
Source	55.0	51.5	26.9	24.0	60.5	29.0	21.4	21.1	25.0	35.2	11.8	34.8	43.2	56.0	35.9	35.4	0.0
Pseudo-label [22]	53.8	48.9	25.4	23.0	58.7	27.3	19.6	20.6	23.4	31.3	11.8	28.4	39.6	52.3	33.9	33.2	+2.2
Tent-continual [35]	53.8	48.7	25.5	23	59.1	27.4	19.7	20.9	23.5	31.8	11.8	27.9	39.9	50.9	33.8	33.2	+2.2
CoTTA [36]	55.0	51.3	25.8	24.1	59.2	28.9	21.4	21.0	24.7	34.9	11.7	31.7	40.4	55.7	35.6	34.8	+0.6
VDP [10]	54.8	51.2	25.6	24.2	59.1	28.8	21.2	20.5	23.3	33.8	7.5	11.7	32.0	51.7	35.2	32.0	+3.4
Ours (proposed)	50.1	40.7	22.0	21.2	45.2	21.6	16.5	17.9	16.6	25.6	11.5	29.0	29.6	34.7	27.1	27.3	+8.1

In this section, we expand upon the classification results presented in our submission by providing a details of fine-grained performance. We assess the error rates across fifteen corruption types to gain deeper insights. To be specific, we augment the information provided in Table 2 of our submission with the additional details presented in Table 5 and 6. These tables offer a comprehensive view of the performance of our approach in addressing the CIFAR-10-to-CIFAR-10C and CIFAR-100-to-CIFAR-100C CTTA scenarios, respectively.

7 Main Checklist

Claims: The abstract, introduction, and conclusion effectively convey the main claims of the paper, accurately reflecting its contributions and scope. The claims put forth in the paper align with both theoretical and experimental findings, ensuring that the extent of generalization expected from the results is appropriately conveyed. Furthermore, the introduction and conclusion clearly outline the paper’s contributions, important assumptions, and limitations, leaving no room for ambiguity.

Code Of Ethics: After thoroughly reviewing the NeurIPS Code of Ethics (<https://neurips.cc/public/EthicsGuidelines>), I have ensure that our research adheres to its principles and guidelines.

Broader Impacts: I have conscientiously considered the potential societal impacts of our work, including any potential risks associated with malicious or unintended uses, as well as negative applications. I strive to minimize any adverse consequences and ensure that our research is conducted responsibly.

Limitations: In the last section of our submission, I have thoroughly documented the limitations inherent in our work.

Experiments: I will promptly provide the necessary code, data, and comprehensive instructions to ensure the reproduction of our main experimental results.

Training Detail and Compute: I provide explicit descriptions of all the training details in implementation details that are unique to this work. For other aspects, we closely adhere to previously published work, following their established methodologies and specifications.

References

- [1] Shai Ben-David, John Blitzer, Koby Crammer, Alex Kulesza, Fernando Pereira, and Jennifer Wortman Vaughan. A theory of learning from different domains. *Machine learning*, 79(1):151–175, 2010.
- [2] Shai Ben-David, John Blitzer, Koby Crammer, and Fernando Pereira. Analysis of representations for domain adaptation. *Advances in neural information processing systems*, 19, 2006.
- [3] Malik Boudiaf, Tom Denton, Bart van Merriënboer, Vincent Dumoulin, and Eleni Triantafillou. In search for a generalizable method for source free domain adaptation. 2023.
- [4] Malik Boudiaf, Romain Mueller, Ismail Ben Ayed, and Luca Bertinetto. Parameter-free online test-time adaptation. *ArXiv*, abs/2201.05718, 2022.
- [5] Goirik Chakrabarty, Manogna Sreenivas, and Soma Biswas. Sata: Source anchoring and target alignment network for continual test time adaptation. *arXiv preprint arXiv:2304.10113*, 2023.
- [6] Dian Chen, Dequan Wang, Trevor Darrell, and Sayna Ebrahimi. Contrastive test-time adaptation. *ArXiv*, abs/2204.10377, 2022.
- [7] Shoufa Chen, GE Chongjian, Zhan Tong, Jiangliu Wang, Yibing Song, Jue Wang, and Ping Luo. Adaptformer: Adapting vision transformers for scalable visual recognition. In *Advances in Neural Information Processing Systems*.
- [8] Ning Ding, Yujia Qin, Guang Yang, Fuchao Wei, Zonghan Yang, Yusheng Su, Shengding Hu, Yulin Chen, Chi-Min Chan, Weize Chen, et al. Parameter-efficient fine-tuning of large-scale pre-trained language models. *Nature Machine Intelligence*, pages 1–16, 2023.
- [9] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.
- [10] Yulu Gan, Xianzheng Ma, Yihang Lou, Yan Bai, Renrui Zhang, Nian Shi, and Lin Luo. Decorate the newcomers: Visual domain prompt for continual test time adaptation. *arXiv preprint arXiv:2212.04145*, 2022.

- [11] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. Domain-adversarial training of neural networks. *The journal of machine learning research*, 17(1):2096–2030, 2016.
- [12] Tianyu Gao, Adam Fisch, and Danqi Chen. Making pre-trained language models better few-shot learners. In *Joint Conference of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL-IJCNLP 2021*, pages 3816–3830. Association for Computational Linguistics (ACL), 2021.
- [13] Yunhe Gao, Xingjian Shi, Yi Zhu, Hao Wang, Zhiqiang Tang, Xiong Zhou, Mu Li, and Dimitris N Metaxas. Visual prompt tuning for test-time domain adaptation. *arXiv preprint arXiv:2210.04831*, 2022.
- [14] Ruidan He, Linlin Liu, Hai Ye, Qingyu Tan, Bosheng Ding, Liying Cheng, Jia-Wei Low, Lidong Bing, and Luo Si. On the effectiveness of adapter-based tuning for pretrained language model adaptation. *arXiv preprint arXiv:2106.03164*, 2021.
- [15] Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. Parameter-efficient transfer learning for nlp. In *International Conference on Machine Learning*, pages 2790–2799. PMLR, 2019.
- [16] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.
- [17] Shengding Hu, Ning Ding, Huadong Wang, Zhiyuan Liu, Jingang Wang, Juanzi Li, Wei Wu, and Maosong Sun. Knowledgeable prompt-tuning: Incorporating knowledge into prompt verbalizer for text classification. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2225–2240, 2022.
- [18] Rabeeh Karimi Mahabadi, James Henderson, and Sebastian Ruder. Compacter: Efficient low-rank hypercomplex adapter layers. *Advances in Neural Information Processing Systems*, 34:1022–1035, 2021.
- [19] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. *arXiv preprint arXiv:2304.02643*, 2023.
- [20] Jogendra Nath Kundu, Naveen Venkat, Rahul M, and R. Venkatesh Babu. Universal source-free domain adaptation. 2020.
- [21] Qicheng Lao, Xiang Jiang, and Mohammad Havaei. Hypothesis disparity regularized mutual information maximization, 2020.
- [22] Dong-Hyun Lee. Pseudo-label : The simple and efficient semi-supervised learning method for deep neural networks. 2013.
- [23] Brian Lester, Rami Al-Rfou, and Noah Constant. The power of scale for parameter-efficient prompt tuning. *arXiv preprint arXiv:2104.08691*, 2021.
- [24] Da Li, Yongxin Yang, Yi-Zhe Song, and Timothy M Hospedales. Deeper, broader and artier domain generalization. In *Proceedings of the IEEE international conference on computer vision*, pages 5542–5550, 2017.
- [25] Jian Liang, D. Hu, and Jiashi Feng. Do we really need to access the source data? source hypothesis transfer for unsupervised domain adaptation. In *ICML*, 2020.
- [26] J MacQueen. Classification and analysis of multivariate observations. In *5th Berkeley Symp. Math. Statist. Probability*, pages 281–297. University of California Los Angeles LA USA, 1967.
- [27] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

- [28] Yujia Qin, Xiaozhi Wang, Yusheng Su, Yankai Lin, Ning Ding, Zhiyuan Liu, Juanzi Li, Lei Hou, Peng Li, Maosong Sun, et al. Exploring low-dimensional intrinsic task subspace via prompt tuning. *arXiv preprint arXiv:2110.07867*, 2021.
- [29] Sylvestre-Alvise Rebuffi, Hakan Bilen, and Andrea Vedaldi. Learning multiple visual domains with residual adapters. *Advances in neural information processing systems*, 30, 2017.
- [30] Sylvestre-Alvise Rebuffi, Hakan Bilen, and Andrea Vedaldi. Efficient parametrization of multi-domain deep neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8119–8127, 2018.
- [31] Steffen Schneider, Evgenia Rusak, Luisa Eck, Oliver Bringmann, Wieland Brendel, and Matthias Bethge. Improving robustness against common corruptions by covariate shift adaptation. *ArXiv*, abs/2006.16971, 2020.
- [32] Yi-Lin Sung, Jaemin Cho, and Mohit Bansal. VI-adapter: Parameter-efficient transfer learning for vision-and-language tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5227–5237, 2022.
- [33] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.
- [34] Tu Vu, Brian Lester, Noah Constant, Rami Al-Rfou, and Daniel Cer. Spot: Better frozen model adaptation through soft prompt transfer. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5039–5059, 2022.
- [35] Dequan Wang, Evan Shelhamer, Shaoteng Liu, Bruno A. Olshausen, and Trevor Darrell. Tent: Fully test-time adaptation by entropy minimization. In *ICLR*, 2021.
- [36] Qin Wang, Olga Fink, Luc Van Gool, and Dengxin Dai. Continual test-time domain adaptation. *ArXiv*, abs/2203.13591, 2022.
- [37] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M Alvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. *Advances in Neural Information Processing Systems*, 34:12077–12090, 2021.
- [38] Shiqi Yang, Yaxing Wang, Joost van de Weijer, Luis Herranz, and Shangling Jui. Generalized source-free domain adaptation. *international conference on computer vision*, 2021.
- [39] Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. *british machine vision conference*, 2016.
- [40] Elad Ben Zaken, Shauli Ravfogel, and Yoav Goldberg. Bitfit: Simple parameter-efficient fine-tuning for transformer-based masked language-models. *arXiv preprint arXiv:2106.10199*, 2021.
- [41] Kaiyang Zhou, Ziwei Liu, Yu Qiao, Tao Xiang, and Chen Change Loy. Domain generalization in vision: A survey. *arXiv preprint arXiv:2103.02503*, 2021.