# *Supplementary Materials.* ViDA: Homeostatic Visual Domain Adapter for Continual Test Time Adaptation

**Jiaming Liu**[1], **Senqiao Yang**[1*], **Peidong Jia** [1†], **Renrui Zhang** [3],
Ming Lu[1], Yandong Guo[2], Wei Xue[4], Shanghang Zhang[1] ✉
[1]National Key Laboratory for Multimedia Information Processing,
School of Computer Science, Peking University [2]AI[2]Robotics
[3] The Chinese University of Hong Kong [4] Hong Kong University of Science and Technology
jiamingliu@stu.pku.edu.cn, yangsenqiao.ai@gmail.com, shanghang@pku.edu.cn

## A Appendix

The supplementary materials presented in this paper offer a comprehensive quantitative and qualitative analysis of the proposed method. In Appendix B, we provide additional empirical observations and justifications for our motivation, including specifically designed quantitative analysis, qualitative analysis of the distribution, and justification for distribution divergence. Additionally, we present extra continual adaptation experiments for Foundation Models in Appendices C.1 and C.2, which are conducted on ImageNet-to-ImageNet-C and Cityscape-to-ACDC scenarios. To assess the domain generalization ability of our method, we conducted additional experiments directly testing a varying number of unseen domains in Appendix C.3. The ablation study on middle-layer dimension is described in Appendix C.4. Furthermore, Appendix C.5 presents additional CTTA classification experiments utilizing the convolutional backbone, while Appendix C.6 outlines 10 rounds of semantic segmentation CTTA experiments. We provide an additional qualitative analysis in Appendix D. Moreover, we extend the classification results of our submission to include fine-grained performance in Appendix E, showcasing the error rates across fifteen corruption types.

## B Supplementary Justifications for Motivation

The study of Continual Test-Time Adaptation (CTTA) poses significant challenges, particularly in addressing error accumulation and catastrophic forgetting (Wang et al., 2022; Gan et al., 2023). Notably, the use of adapters with low-rank and high-rank features have 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 have introduced two new specially designed quantitative experiments in Section B.1. The first one is a 10-round CTTA experiment aimed at investigating the different domain representations of low-rank and high-rank ViDA during the long-term adaptation process. The second experiment explores the performance when all adapters adopt the same structures, such as using two high-rank adapters or two low-rank adapters. This experiment is conducted to validate that low-rank ViDA and high-rank ViDA complement each other in adapting to continually changing environments.

### B.1 Specially Designed 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-shared and domain-specific knowledge on ImageNet-to-ImageNet-C.

First, as shown in Figure 1 (b), we execute a 10 rounds CTTA experiment on ImageNet-to-ImageNet-C. In this comprehensive experiment, we simulate a long-term adaptation scenario by repeating 10

[*]Equal contribution, [†] Equal technical contribution,
✉ Corresponding author. Project page: https://sites.google.com/view/iclr2024-vida/home

Table 1: Classification error rate(%) for ImageNet-to-ImageNet-C online CTTA task. Gain(%) represents the percentage of improvement in model accuracy compared with the source method. 2× means using two same structures of adapters.

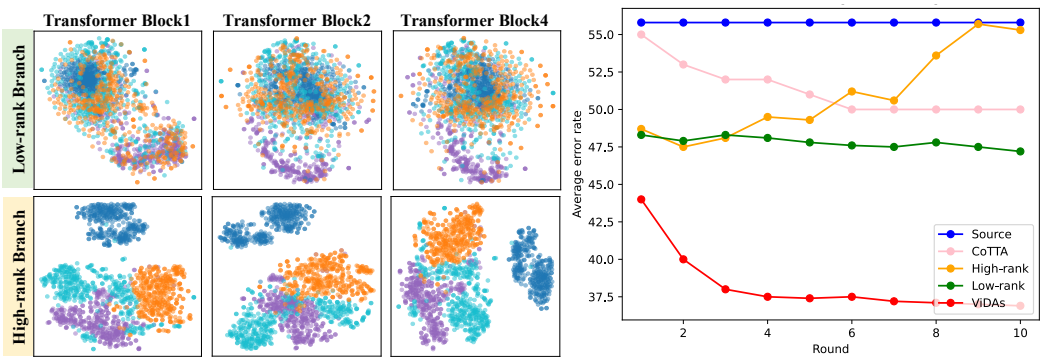
| | Method | *Gaussian* | *shot* | *impulse* | *defocus* | *glass* | *motion* | *zoom* | *snow* | *frost* | *fog* | *brightness* | *contrast* | *elastic* | *pixelate* | *jpeg* | Mean↓ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Ex1 | Source[58] | 53.0 | 51.8 | 52.1 | 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 |
| Ex2 | 2×Low-rank | 51.2 | 48.3 | 47.8 | 56.9 | 66.5 | 49.3 | 54.4 | 42.1 | 47.0 | 45.2 | 23.2 | 65.6 | 52.0 | **33.4** | 33.5 | 47.7 |
| Ex3 | 2×High-rank | **50.1** | 47.9 | **45.3** | **54.8** | 66.7 | 51.4 | 56.1 | 44.0 | 49.2 | 48.3 | 25.7 | 69.7 | 56.3 | 34.6 | 33.7 | 48.9 |
| Ex4 | Ours | 50.3 | **45.9** | 45.5 | 55.1 | **62.3** | **46.6** | **51.7** | **39.7** | **44.0** | **42.2** | **23.0** | **62.4** | **50.1** | 33.4 | **32.5** | **45.6** |

rounds of 15 corruption sequences in the ImageNet-C. Remarkably, the high-rank ViDA achieves competitive results over other methods during the initial 1 to 3 rounds. This result demonstrates the high-rank feature's capacity to efficiently learn target domain-specific knowledge. However, an increment in error rates becomes obvious during the later rounds (rounds 5 to 10). The results validate the potential for encountering catastrophic forgetting when focusing exclusively on domain-specific knowledge. In contrast, the performance of the low-rank ViDA remains consistently robust throughout the continual adaptation process, verifying it concentrates more on extracting task-relevant knowledge and effectively prevents the catastrophic forgetting problem. And our proposed method consistently improves over time, demonstrating its robustness in the long-term adaptation process.

Second, we execute an ImageNet-to-ImageNet-C CTTA experiment using a combination of two high-rank adapters or two low-rank adapters, as shown in Table 1. To ensure fairness, we conducted these experiments without implementing the homeostatic knowledge allotment (HKA) strategy. Notably, the two low-rank adapters (Ex2) demonstrated consistently lower long-term error rates compared to the source model and two high-rank adapters. The above results can be attributed to the fact that the two low-rank ViDAs tend to learn general information and domain-shared knowledge during continual adaptation. However, our method outperforms the two low-rank adapters across 14 out of 15 corruption types. This indicates that solely relying on low-rank adapters without the involvement of high-rank adapters is insufficient to fit target domains and match their data distribution. On the other hand, the performance of the two high-rank adapters initially surpasses our approach (Ex4) in the early stages, covering the first few target domains. Nevertheless, a noticeable performance degradation becomes apparent in later target domains. This observation underscores a crucial finding: while increasing the number of high-rank ViDAs might enhance domain-specific knowledge acquisition during the initial phases of CTTA, it simultaneously exacerbates catastrophic forgetting throughout the entire adaptation process. In contrast, the fusion of both low-rank and high-rank ViDAs (Ex4) yields the most substantial improvement when compared to other configurations. Our collaborative approach leverages the distinct domain representations of these adapters to compensate for each other's advantages and achieve a more robust and effective continual adaptation.

### B.2 Additional Distribution Qualitative Analysis

We employed t-distributed stochastic neighbor embedding (t-SNE) (Van der Maaten & Hinton, 2008) to visualize the distribution of adapters across four continual target domains. This visualization was specifically conducted in the context of the Cityscapes-to-ACDC experiment, representing a scenario with continually changing real-world environments.In our submission, we perform t-SNE analysis on the outputs of the third transformer block in the Segformer-B5 model (Xie et al., 2021). The objective was to qualitatively compare the feature distributions of ViDAs with different dimension features. Furthermore, our findings revealed that the qualitative results obtained from different layers (i.e., transformer block 1, 2, and 4) of the Segformer-B5 model exhibited similar distribution representations. As illustrated in Figure 1 (a), there is a noticeable distribution gap due to the significant domain shift between the night domain and other domains. Interestingly, the low-rank ViDA effectively reduces the distribution distance across different target domains, indicating its focus on extracting task-relevant knowledge. 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.




**(a) t-SNE distribution analysis** **(b) 10 Rounds CTTA on ImageNet-to-ImageNet-C**

Figure 1: (a) We conduct more t-SNE results for the low-rank adapter and high-rank adapter on the ACDC dataset. The first to third columns illustrate the feature distributions of transformer blocks 1, 2, and 4, respectively. (b) The 10 rounds CTTA experiment on ImageNet-to-ImageNet-C, repeating 10 rounds of 15 corruption sequences.

### B.3 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 (Ben-David et al., 2006; 2010) and build upon previous domain transfer research (Ganin et al., 2016) 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

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

, where $\mathcal{H}$ denotes hypothetical space and $\mathcal{D}$ denotes discriminator. Similar to (Ruder & Plank, 2017; Allaway et al., 2021), we adopt the $Jensen$-$Shannon$ $(JS)$ $divergence$ between two adjacent domains as an approximation of $\mathcal{H}$-divergence because it has been shown to successfully distinguish domains. If the inter-domain divergence is relatively small, it can be demonstrated that the feature representation is consistent and less influenced by cross-domain shifts (Ganin et al., 2016).

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

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

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

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, injecting high-rank adapter, and combining low-high adapters, as illustrated in Figure 3(a) of our submission. The low-rank adapter exhibits notably lower divergence values compared to the others, demonstrating robust task-relevant feature representation in various cross-domain phases. 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 (MacQueen, 1967). We first calculate the Euclidean distance clustering center for each category:

$$\mu = \frac{1}{|C|} \sum_{e_i \sim C} e_i \tag{4}$$

, where $e_i$ stands for output feature in class $C$. Then following (MacQueen, 1967), we introduce normalized intra-class divergence $E$ by

$$E = \phi(\frac{1}{|C|} \sum_{e_i \sim C} ||e_i - \mu||_2^2) \tag{5}$$

$\phi(\cdot)$ denotes for normlization function. In a given domain, if the intra-class divergence for each category is smaller, it demonstrates that the model has a better understanding of the current distribution (Li et al., 2020). 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.

# C Additional Experiment

## C.1 Additional Classification CTTA Experiments for Foundation Models

Table 2: 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 and SAM.

| Backbone | Method | REF | *Gaussian* | *shot* | *impulse* | *defocus* | *glass* | *motion* | *zoom* | *snow* | *frost* | *fog* | *brightness* | *contrast* | *elastic_trans* | *pixelate* | *jpeg* | Mean↓ | Gain |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| DINOv2 | Source | | 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 (Wang et al., 2021) | ICLR2021 | 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 (Wang et al., 2022) | CVPR2022 | 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** | Proposed | **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 | | 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 (Wang et al., 2021) | ICLR2021 | 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 (Wang et al., 2022) | CVPR2022 | 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** | Proposed | **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** |

To demonstrate the effectiveness of our proposed method in enhancing the continual adaptation ability of foundation models such as DINOv2 (Oquab et al., 2023) and SAM (Kirillov et al., 2023), we conduct additional experiments on a more extensive dataset, namely ImagNet-to-ImageNet-C. Our approach involve loading the weight parameters of the foundation model and fine-tuning it on ImagNet, thus constructing our source model. It is important to note that we solely utilize the pre-trained encoder of SAM and incorporated a classification head, which is fine-tuned on the source domain. Subsequently, we adapt the source model to continual target domains (ImageNet-C) comprising fifteen corruption types. The results, as depicted in Table 2, demonstrate that our approach achieved a significant performance improvement of 8.4% on the representative image-level foundation model DINOv2 and 10.6% on the pixel-level foundation model SAM. These outcomes underscore the effectiveness of our method for large-scale models, consistently and reliably improving performance across target domains. Combining Table 1-3 from the submission, we were surprised to discover a significant decrease in model performance for the classification CTTA task when using the pre-trained encoder parameters of SAM. As SAM is a pixel-level foundation model, we then attempted to investigate the effectiveness of SAM's pretrained parameters in the segmentation CTTA task.

## C.2 Additional Segmentation CTTA Experiments for Foundation Models

As shown in Table 3, we conducted segmentation CTTA using SAM's pre-trained parameters on the Cityscapes-to-ACDC scenario. However, it's worth noting that the Segformer model (Xie et al., 2021), which we employed in our main experiments, does not incorporate positional encoding. Therefore, we adopted the SETR model (Zheng et al., 2021) as our new baseline for loading SAM's pre-trained parameters. As shown in the table, our approach with SAM's pre-trained parameters outperforms others on the ACDC target domains. This aligns with the assumption that SAM, being a pixel-level foundational model, excels in capturing fine-grained feature representations in dense CTTA tasks.

## C.3 Domain Generalization on a Different Number of Unseen Target Domains

Similar to our previous submission, we follow the leave-one-domain-out principle (Zhou et al., 2021; Li et al., 2017), where we utilize a subset of ImageNet-C domains as new source domains for model

Table 3: Performance comparison for Cityscape-to-ACDC CTTA. All results are evaluated on the SETR, which uses the pre-trained parameter of source model or SAM.

| Method | Pre-trained | Fog | Night | Rain | Snow | Mean mIoU |
|---|---|---|---|---|---|---|
| Source | Source model | 72.6 | 43.1 | 63.0 | 64.3 | 60.8 |
| Source | SAM | 74.8 | 44.1 | 66.7 | 66.6 | 63.0 |
| CoTTA | SAM | 75.4 | 45.9 | 67.3 | 68.7 | 64.3 |
| **Ours** | SAM | **76.5** | **47.2** | **68.1** | **70.7** | **65.6** |

Table 4: The domain generalization experiments 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* | *brightness* | *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 (Wang et al., 2021) | 56.0 | 61.3 | 45.7 | 49.6 | 56.6 | 24.8 | 94.0 | 55.6 | 37.1 | 35.1 | 51.6 |
| CoTTA (Wang et al., 2022) | 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** |

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 unlabeled 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 4 and 5 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-shared knowledge and provides a new perspective for enhancing DG performance within an unsupervised paradigm.

### C.4 Additional Ablation study

**How does the middle-layer dimension influence the performance?**

According to Figure 2, we observe that as the dimension decreases, the error rate concurrently drops. This trend suggests that lower-dimension middle layer more effectively extract the domain-shared knowledge, leading to an improved model performance. However, an opposite trend emerges when dimension surpasses 16, with performance enhancements accompanying increased dimension. This

Table 5: The domain generalization experiments 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* | *brightness* | *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 (Wang et al., 2021) | 44.3 | 48.8 | 51.8 | 24.9 | 83.7 | 55.2 | 35.4 | 34.7 | 47.4 |
| CoTTA (Wang et al., 2022) | 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** |

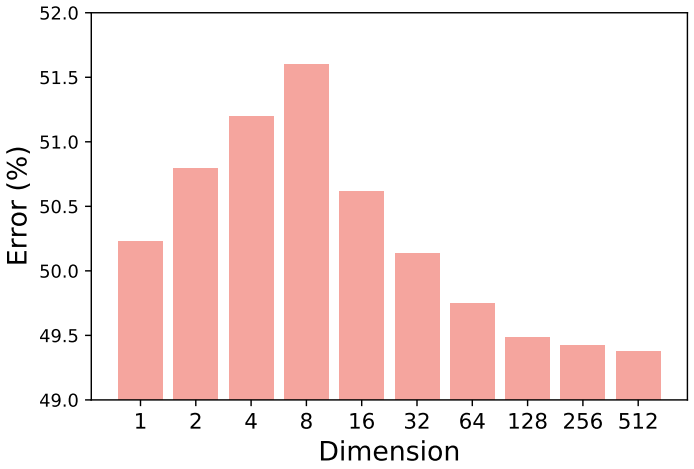



Figure 2: The middle-layer dimension influence of the performance

correlation implies that middle layers with a higher dimension 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 dimension of the high-dimension middle layer to 128 in our study.

**How do different adapter initialization methods impact ViDA performance?**

Pre-training the low-rank and high-rank ViDAs using source data is an unnecessary step and does not compromise the effectiveness of our approach. ViDAs can demonstrate comparable CTTA performance when they have a relatively stable initial parameter. As illustrated in the Table 6, we conduct an additional experiment on the Cityscape-to-ACDC scenario. ViDAs with random initial parameters and ViDAs with parameters pre-trained on ImageNet achieved 60.5 and 61.4 mIoU in target domains, respectively, exhibiting notable improvements compared to previous methods.

Table 6: The ablation study examines adapter initialization methods on the Cityscape-to-ACDC CTTA scenario.

| | Adapter pre-train | Fog | Night | Rain | Snow | Mean (IoU) |
|---|---|---|---|---|---|---|
| Source | - | 69.1 | 40.3 | 59.7 | 57.8 | 56.7 |
| CoTTA | - | 70.9 | 41.2 | 62.4 | 59.7 | 58.6 |
| Ours | Source | 71.6 | 43.2 | 66.0 | 63.4 | 61.1 |
| Ours | Random initial | 71.6 | 43.6 | 64.9 | 61.9 | 60.5 |
| Ours | ImageNet | 71.6 | 44.3 | 66.0 | 63.5 | 61.4 |

### C.5 Experiments on Classification CTTA with Convolutional Backbones

**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 (Wang et al., 2022). Specifically, we modify the up-projection layer and down-projection layer to utilize $1 \times 1$ convolutions, while the adapters are placed alongside the original $3 \times 3$ convolutions. For ViDA, we maintain a low-rank dimension of 1 and a high-rank dimension of 128. As depicted in Table 7, our method achieves a 27.7% improvement over the source model. These findings demonstrate that our method successfully address error accumulation and catastrophic forgetting problem, regardless of the network backbone employed.

### C.6 Additional Experiments on Segmentation CTTA

We further present the segmentation CTTA experiment with 10 rounds on Table 8. Notably, it demonstrates a consistent enhancement in mean mIoU during the initial rounds (rounds 1-3) while maintaining stable performance in subsequent rounds (rounds 4-10). After averaging over 10 rounds ,

Table 7: Classification error rate(%) for standard CIFAR10-to-CIAFAR10C online CTTA task. Results are evaluated on WideResNet-28. Mean is the average value of the error rate. Gain(%) represents the percentage of improvement in model accuracy compared with the source method.

| Method | REF | Conference | Mean↓ | Gain |
|---|---|---|---|---|
| Source | (Zagoruyko & Komodakis, 2016) | BMVC2016 | 43.5 | 0.0 |
| BN Stats Adapt | (Schneider et al., 2020) | NeurIPS2020 | 20.4 | +23.1 |
| TENT | (Wang et al., 2021) | ICLR2021 | 20.7 | +22.8 |
| CoTTA | (Wang et al., 2022) | CVPR2022 | 16.2 | +27.3 |
| RoTTA | (Yuan et al., 2023) | CVPR2023 | 17.5 | +26.0 |
| NOTE | (Gong et al., 2022) | NeurIPS2022 | 20.2 | +23.3 |
| EcoTTA | (Song et al., 2023) | ICCV2023 | 16.8 | +26.7 |
| SATA | (Chakrabarty et al., 2023) | 2023.4.20 | 16.1 | +27.4 |
| **Ours** | Proposed | 2023.5.18 | **15.8** | **+27.7** |

Table 8: **10 rounds segmentation CTTA on Cityscape-to-ACDC.** We sequentially repeat the same sequence of target domains 10 times. Mean is the average score of mIoU.

| Round | 1 | | | | | 2 | | | | | 3 | | | | | 4 | | | | | 5 | | | | | Mean |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Method | Fog | Night | Rain | Snow | Mean | Fog | Night | Rain | Snow | Mean | Fog | Night | Rain | Snow | Mean | Fog | Night | Rain | Snow | Mean | Fog | Night | Rain | Snow | Mean | |
| Source | 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 | 40.3 | 59.7 | 57.8 | 56.7 | 56.7 | 40.3 | 59.7 | 57.8 | 56.7 | cont. |
| CoTTA | 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 | 70.9 | 41.0 | 62.7 | 59.7 | 58.6 | 70.9 | 41.0 | 62.8 | 59.7 | 58.6 | cont. |
| CoTTA* | 71.9 | 45.0 | 67.1 | 63.1 | 61.8 | 71.9 | 43.6 | 65.6 | 61.8 | 60.7 | 69.6 | 39.7 | 63.5 | 60.4 | 58.3 | 68.3 | 39.6 | 61.8 | 59.4 | 57.3 | 67.8 | 38.9 | 62.1 | 59.7 | 57.1 | cont. |
| Ours | 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** | 70.9 | 44.0 | 66.0 | 63.2 | **61.0** | 72.0 | 43.7 | 66.3 | 63.1 | **61.3** | cont. |
| Round | 6 | | | | | 7 | | | | | 8 | | | | | 9 | | | | | 10 | | | | | Mean |
| Method | Fog | Night | Rain | Snow | Mean | Fog | Night | Rain | Snow | Mean | Fog | Night | Rain | Snow | Mean | Fog | Night | Rain | Snow | Mean | Fog | Night | Rain | Snow | Mean | |
| Source | 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 | 40.3 | 59.7 | 57.8 | 56.7 | 56.7 | 40.3 | 59.7 | 57.8 | 56.7 | 56.7 |
| CoTTA | 70.9 | 41.0 | 62.8 | 59.7 | 58.6 | 70.9 | 41.1 | 62.6 | 59.7 | 58.6 | 70.9 | 41.1 | 62.6 | 59.7 | 58.6 | 70.8 | 41.1 | 62.6 | 59.7 | 58.6 | 70.8 | 41.1 | 62.6 | 59.7 | 58.6 | 58.6 |
| CoTTA* | 67.7 | 39.8 | 62.7 | 59.7 | 57.5 | 67.3 | 39.7 | 63.2 | 59.6 | 57.7 | 67.6 | 40.1 | 63.2 | 58.0 | 57.2 | 65.0 | 38.8 | 60.7 | 58.5 | 55.8 | 66.9 | 38.9 | 62.7 | 58.7 | 56.8 | 58.0 |
| Ours | 72.2 | 44.0 | 66.6 | 62.9 | **61.4** | 72.3 | 44.8 | 66.5 | 62.9 | **61.6** | 72.1 | 45.1 | 66.2 | 62.9 | **61.5** | 71.9 | 45.3 | 66.3 | 62.9 | **61.5** | 72.2 | 45.2 | 66.5 | 62.9 | **61.6** | **61.6** |

our method achieved a 3.0% mIoU improvement compared to the previous SOTA method. As shown in Table 8 (CoTTA*), we adjust the hyperparameters of the CoTTA method by raising the learning rate to 3e-4, which aligns with our implementation details. The impact of this adjustment is evident in the initial three rounds of segmentation, where performance notably improves. However, as we progress to subsequent CTTA rounds, we observe a noticeable decline in segmentation accuracy and encounter the problem of catastrophic forgetting.

## D 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 (Xie et al., 2021) 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 (Wang et al., 2022) and VDP (Gan et al., 2023). The visualizations of the segmentation outputs, obtained through the CTTA process, are depicted in Figure 3. 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.

## E Fine-grained Performance

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

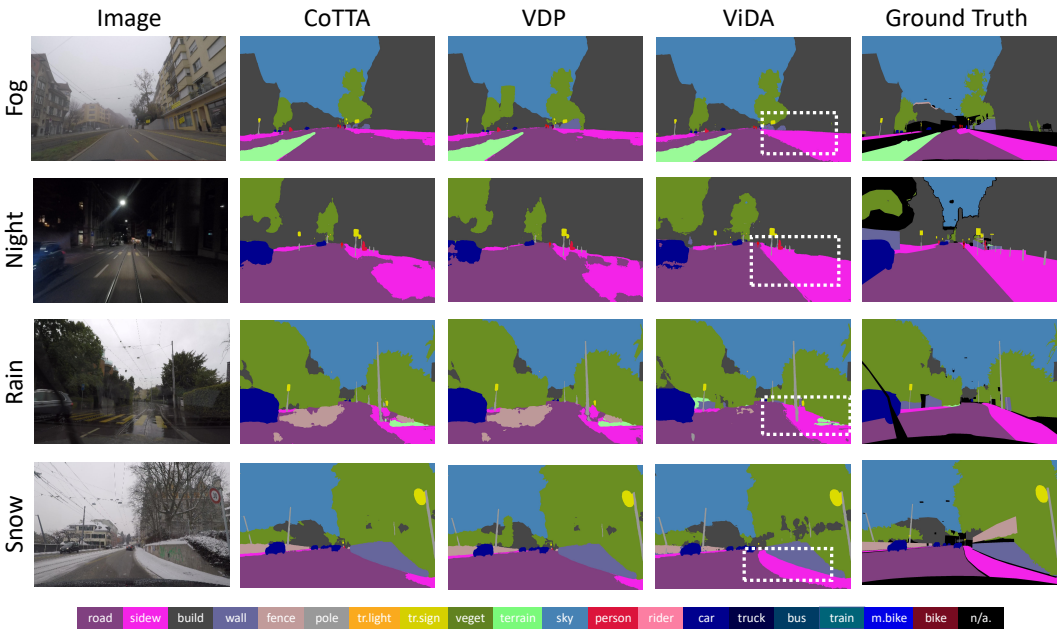



Figure 3: Qualitative comparison of our method with previous SOTA methods on the ACDC dataset. Our method could better segment different pixel-wise classes such as shown in the white box.

Table 9: A fine-grained Classification error rate(%) for standard CIFAR10-to-CIAFAR10C online CTTA task. Results are evaluated on ViT-base.

| Method | *gaussion* | *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 (Lee, 2013) | 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 (Wang et al., 2021) | 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 (Wang et al., 2022) | 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(Gan et al., 2023) | 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** |

with the additional details presented in Table 9 and 10. 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.

Table 10: A fine-grained Classification error rate(%) for standard CIFAR100-to-CIAFAR100C online CTTA task. Results are evaluated on ViT-base.

| Method | *gaussion* | *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 (Lee, 2013) | 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 (Wang et al., 2021) | 53.0 | 47.0 | 24.6 | 22.3 | 58.5 | 26.5 | 19.0 | 21.0 | 23.0 | 30.1 | 11.8 | 25.2 | 39.0 | 47.1 | 33.3 | 32.1 | +3.3 |
| CoTTA (Wang et al., 2022) | 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 (Gan et al., 2023) | 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** |

## REFERENCES


Emily Allaway, Malavika Srikanth, and Kathleen McKeown. Adversarial learning for zero-shot stance detection on social media. *arXiv preprint arXiv:2105.06603*, 2021.

Shai Ben-David, John Blitzer, Koby Crammer, and Fernando Pereira. Analysis of representations for domain adaptation. *Advances in neural information processing systems*, 19, 2006.

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.

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.

Yulu Gan, Yan Bai, Yihang Lou, Xianzheng Ma, Renrui Zhang, Nian Shi, and Lin Luo. Decorate the newcomers: Visual domain prompt for continual test time adaptation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 7595–7603, 2023.

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.

Taesik Gong, Jongheon Jeong, Taewon Kim, Yewon Kim, Jinwoo Shin, and Sung-Ju Lee. Note: Robust continual test-time adaptation against temporal correlation. *Advances in Neural Information Processing Systems*, 35:27253–27266, 2022.

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.

Dong-Hyun Lee. Pseudo-label : The simple and efficient semi-supervised learning method for deep neural networks. 2013.

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*, pp. 5542–5550, 2017.

Jingjing Li, Erpeng Chen, Zhengming Ding, Lei Zhu, Ke Lu, and Heng Tao Shen. Maximum density divergence for domain adaptation. *IEEE transactions on pattern analysis and machine intelligence*, 43(11):3918–3930, 2020.

J MacQueen. Classification and analysis of multivariate observations. In *5th Berkeley Symp. Math. Statist. Probability*, pp. 281–297. University of California Los Angeles LA USA, 1967.

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.

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

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.

Junha Song, Jungsoo Lee, In So Kweon, and Sungha 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.

Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.

Dequan Wang, Evan Shelhamer, Shaoteng Liu, Bruno A. Olshausen, and Trevor Darrell. Tent: Fully test-time adaptation by entropy minimization. In *ICLR*, 2021.

Qin Wang, Olga Fink, Luc Van Gool, and Dengxin Dai. Continual test-time domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7201–7211, 2022.

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.

Longhui Yuan, Binhui Xie, and Shuang Li. Robust test-time adaptation in dynamic scenarios. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15922–15932, 2023.

Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. *british machine vision conference*, 2016.

Sixiao Zheng, Jiachen Lu, Hengshuang Zhao, Xiatian Zhu, Zekun Luo, Yabiao Wang, Yanwei Fu, Jianfeng Feng, Tao Xiang, Philip HS Torr, et al. Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 6881–6890, 2021.

Kaiyang Zhou, Ziwei Liu, Yu Qiao, Tao Xiang, and Chen Change Loy. Domain generalization in vision: A survey. *arXiv preprint arXiv:2103.02503*, 2021.