Robust Continual Test Time Adaptation via Visual Domain Adapter

Ran Xu[1], Peidong Jia[1], Senqiao Yang[1], Jiayi Ni[1], Jiaming Liu[1], Zehui Chen[2],
Feng Zhao[2], Yandong Guo[3], Shanghang Zhang[1],
[1]National Key Laboratory for Multimedia Information Processing, Peking University,
[2]University of Science and Technology of China, [3]AI2Robot

Abstract

In this report, we introduce our method, Visual Domain Adapter (ViDA), for the Continuous Test-time Adaptation (CTTA) in the Semantic Segmentation track of the 1st Visual Continual Learning (VCL) challenge at ICCV 2023 Workshop. Existing CTTA methods primarily rely on model-based adaptation, utilizing self-training to extract ongoing domain knowledge. However, noisy pseudo labels and unstable model parameters in changing data distributions lead to error accumulation and catastrophic forgetting problems. To tackle these problems, we propose a robust CTTA framework that employs ViDA to enhance the continual adaptation capabilities of the foundational model while maintaining its plasticity. Specifically, we opt for SETR as our segmentation model and integrate a pre-trained foundation model, SAM (Segment Any Thing), to improve initial generalization ability. During the continuous testing phase, we introduce low-rank and high-rank ViDAs to mitigate error accumulation and catastrophic forgetting. Our method achieved an impressive 74.6 overall score (84.6% mIoU with only a 5.0% mIoU drop) on the SHIFT test set, securing 1st place on the online leaderboard.

1. Introduction

The Visual Continual Learning Challenge, hosted at the ICCV Workshop, marks a significant milestone as the first of its kind. We are actively participating in the ”Continuous Test-time Adaptation for Semantic Segmentation” track, which centers around utilizing video sequences to guide models in adapting to continual target domains. This task addresses critical, unresolved issues in Visual Continual Learning (VCL), including error accumulation and catastrophic forgetting, especially in dynamic real-world scenarios where domain shifts occur continuously. Our track is based on SHIFT [8], an extensive synthetic driving video dataset. It comprises both a discrete set with 4,250 sequences and a continuous set with an additional 600 sequences. The continuous set showcases gradual transitions between domains, simulating real-world scenarios such as the shift from clear daytime to nighttime.

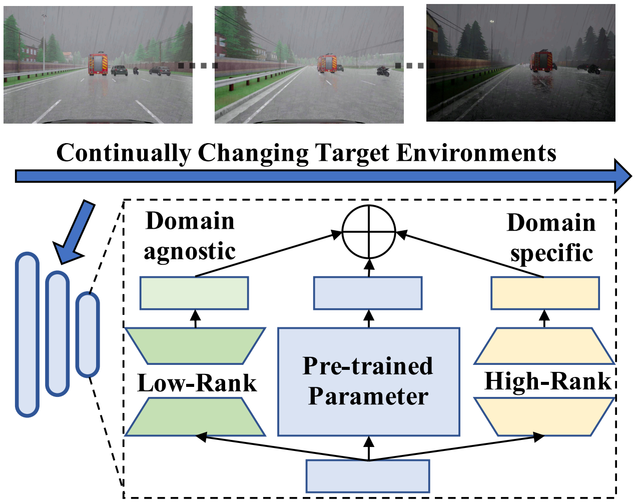

Figure 1. **Visual Domain Adapter**. Our objective is to efficiently adapt the source pre-trained model to the continually changing target environments. We employ ViDAs with high-rank and low-rank prototypes to extract domain-specific and domain-agnostic knowledge, respectively.

Existing CTTA techniques [3,7,10,12] primarily employ model-based or prompt-based approaches to extract target domain-specific knowledge while maintaining task-relevant knowledge through continual updates of partial model parameters or soft prompts. However, for model-based methods [10], the noisy pseudo labels are still unreliable and play a limited role in avoiding error accumulation, particularly in scenarios with significant distribution gaps. Meanwhile, prompt-based methods [3, 12] face difficulties in leveraging only a small number of trainable parameters on input images to learn long-term domain-shared knowledge and prevent catastrophic forgetting.

In this report, we introduce our method, Visual Domain Adapter (ViDA) [5], for the Semantic Segmentation CTTA problem. To address error accumulation and catas-

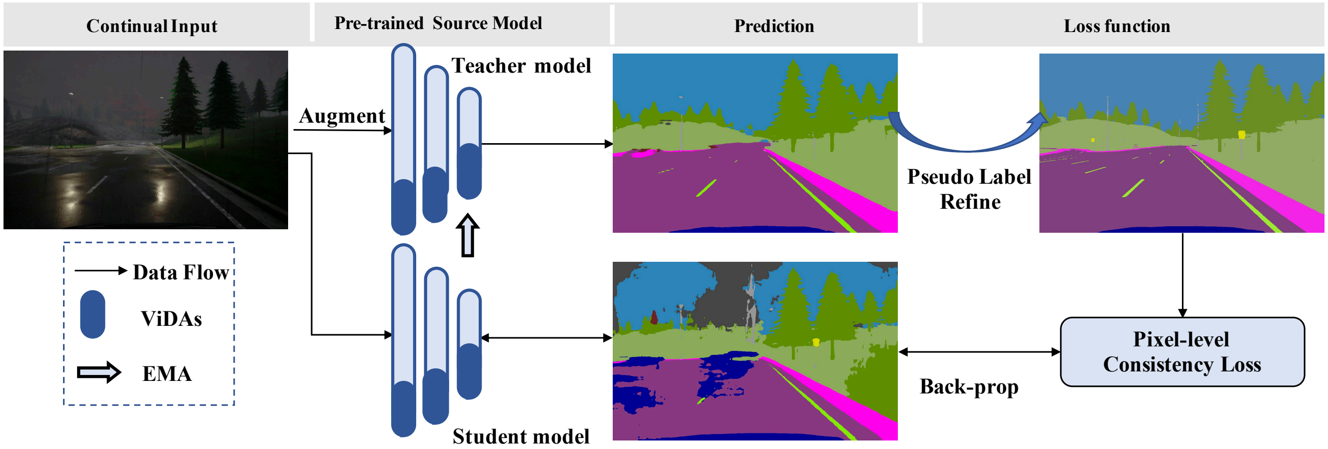

Figure 2. **An overview of our framework.** We inject a variety of Visual Domain Adaptation (ViDA) models, each representing distinct domains, into the linear layers of a pre-trained source model. ViDA updates are achieved through a teacher-student framework utilizing a pixel-level consistency loss (defined as Eq. 3) as the optimization target. Specifically, the student model operates on the original image, while the teacher model works with an augmented version of the same image. Additionally, the teacher model is updated using the EMA.

trophic forgetting in a continuously changing environment, we present a robust CTTA framework. ViDA enhances the continual adaptation capabilities of the large-scale model while preserving its flexibility. We select SETR [13] as our segmentation model and incorporate a pre-trained foundation model, SAM (Segment Any Thing) [4], to improve initial generalization. During continuous testing, we introduce both low-rank and high-rank ViDAs to explicitly manage domain-specific and domain-agnostic knowledge during continual adaptation. Our solution achieves outstanding performance, boasting an impressive 74.6 overall score on the SHIFT test set. Specifically, it maintains an 84.6% mean Intersection over Union (mIoU) with only a minimal 5.0% mIoU drop. These remarkable results secure our position at the top of the online leaderboard.

2. Our Solution

2.1. Overview

In Section 2.2, we will introduce our notation and preliminary details. In Section 2.3, we enhance our model's ability to generalize in the target domain by utilizing the vision foundation model. In Section 2.4, we introduce the state-of-the-art CTTA method, Visual Domain Adapter (ViDA), which incorporates both high and low-dimensional prototypes. This approach effectively enables our model to adapt to diverse distribution shifts. The overall robust framework is illustrated in Figure 2. In Section 2.5, we introduce our optimization objective.

2.2. Preliminaries

In Continual Test-Time Adaptation (CTTA), we initially pre-train the model $q_\theta(y|x)$ using data from the source domain (specified discrete datasets), denoted as $D_S = (Y_S, X_S)$. Subsequently, we adapt this model to the continually changing target domains (continual datasets), represented as $D_{T_1}, D_{T_2}, ..., D_{T_n}$, where 'n' signifies the frames of continual target datasets. It's important to note that during this entire process, there is no access to any source domain data, and access to target domain data is limited to a single instance. Additionally, the distributions of the target domains (specifically, $D_{T_1}, D_{T_2}, ..., D_{T_n}$) continually evolve over time. Our primary objective is to adapt the pre-trained model to these target domains while preserving the model's ability to perceive data from the initially observed domain distribution.

2.3. Generalized Backbone

In recent years, there has been remarkable progress in the field of foundational models [4, 6], which greatly improves the performance and generalization of the model. Recently, Meta AI released a promptable Segment Anything Model (SAM [4]). By incorporating a single user interface as a prompt, SAM is capable of segmenting any object in any image or without additional training, which is a state-of-the-art foundational model specifically tailored for semantic segmentation tasks.

One critical factor influencing the performance of segmentation models is the generalization capability of their models' backbone. The ability of a model to adapt efficiently to various domains during continual test-time domain adaptation relies on the generalizability and diversity of its model backbone. Recognizing the significance of this aspect, our aim is to enhance our model's generalization potential. To achieve this objective, we employ the SETR architecture [13], which uses the VIT-large [1] as its backbone. Additionally, we integrate the pre-trained parameters from SAM [4] into the encoder. This not only enhances

Table 1. **Performance comparison for SHIFT dataset's continuous validation set CTTA.** Mean is the average score of mIoU. Gain refers to the improvement achieved by the method compared to the Source model.

Scenarios			$Daytime - Night$		$Clear - Foggy$		$Clear - Rainy$		Mean↑
Method	Backbone	Resolution	mIoU	ACC	mIoU	ACC	mIoU	ACC	
Deeplab V3+	R50	1280 × 800	55.3	65.3	43.9	56.0	46.3	59.6	49.1
SETR	ViT-L	768 × 768	50.1	56.9	44.9	52.2	49.6	57.2	48.4
Our SETR	SAM-L	1024 × 1024	71.1	76.4	68.6	74.2	73.1	78.4	70.9
Ours	SAM-L	1024 × 1024	72.7	77.6	66.2	72.1	78.4	82.8	72.8

our model's segmentation performance in the source domain but also improves its overall generalization. However, though the foundation model presents a strong zero-shot transfer ability, it shows a significant performance degradation in continually changing environments.

2.4. Visual Domain Adapter

We follow the implementation of ViDA [5]. As shown in Figure 1, there are three sub-branches, with the linear layer in the middle branch remaining identical to the original network. On the other hand, the right branch and left branch represent bottleneck structures and indicate the high-rank ViDA and low-rank ViDA, respectively. Specifically, the right branch (high-rank) consists of an up-projection layer with parameters $W^h_{up} \in R^{d \times d_h}$ and a down-projection layer with parameters $W^h_{down} \in R^{d_h \times d}$, where d_h represents the middle dimension of high-rank prototypes and satisfies the condition $d_h \geq d$.

In contrast, the left branch (low-rank) first employs a down-projection layer with parameters $W^l_{down} \in R^{d \times d^l}$ and then incorporates an up-projection layer with parameters $W^l_{up} \in R^{d_l \times d}$, where d_l represents the middle dimension of the low-rank prototype ($d_l \ll d$).

For an input feature f, the generated features of the high-rank ViDA (f_h) and low-rank ViDA (f_l) are formulated as:

$$f_h = W^h_{down} \cdot (W^h_{up} \cdot f); \quad f_l = W^l_{up} \cdot (W^l_{down} \cdot f) \quad (1)$$

The two-branch bottleneck is connected to the output feature of the original network (f_o) through the residual connection via scale factors (λ_h and λ_l). The fusion knowledge (f_f) can be described as:

$$f_f = f_o + \lambda_h \times f_h + \lambda_l \times f_l \quad (2)$$

2.5. Optimization Objective

Following the previous Continual Test-Time Adaptation (CTTA) research [2, 5, 11, 12], we utilize the teacher model $\mathcal{T}$ to generate pseudo labels $\widetilde{y}$ for updating ViDAs. Our optimization objective involves the adoption of a consistency loss, denoted as L_{ce}, expressed as:

$$\mathcal{L}_{ce}(\widetilde{x}) = -\frac{1}{H \times W} \sum_{w,h}^{W,H} \sum_{c}^{C} \widetilde{y}(w,h,c) \log \hat{y}(w,h,c) \quad (3)$$

where $\hat{y}$ represents the output of our student model $\mathcal{S}$, C represents the number of categories, and H, W represent the height and width of the image, respectively. Additionally, we employ exponential moving average (EMA) to update the teacher model using ViDAs, as shown in the equation:

$$\mathcal{T}^t = \alpha \mathcal{T}^{t-1} + (1 - \alpha)\mathcal{S}^t \quad (4)$$

where t represents the time step, and we set $\alpha = 0.999$ [9] as the updating weight for EMA.

3. Experiments

3.1. Implementation Details

We follow the implementation details of cotta [11] to set up our semantic segmentation CTTA experiments. Specifically, we use the SETR [13] as our segmentation model and integrate a pre-trained foundation model, SAM [4]. We resize the original image size of 1280 × 1280 of the SHIFT dataset to 1024 × 1024, which serves as network input. We evaluate our predictions under the 1024 × 1024 resolution. We use a range of image resolution scale factors [1.0, 1.25, 1.5, 1.75, 2.0] for the augmentation method in the teacher model. The optimizer is SGD, the learning rate is 0.01, the batch size is 2, and all experiments are conducted on NVIDIA A100 GPUs.

3.2. Evaluation metrics

We use mean intersection-over-union (mIoU) and mean accuracy (mAcc) for evaluation on SHIFT dataset's continuous val set, and additional mIoU_drop (mIoU_drop = mIoU_target − mIoU_source, where mIoU_target represents the combined predictions and GTs from 0-20th frames and mIoU_source represents the combined predictions and GTs from 180-220th frames) and overall (overall = mIoU − 2 × mIoU_drop) for online evaluation on SHIFT test set.

Table 2. **Performance comparison for SHIFT dataset's continous test set CTTA.**

Method	Backbone	mIoU	mIoU_drop	mIoU_source	mIoU_target	mIoU_loopback	*overall*
Deeplab V3+	R50	68.1	29.3	77.2	47.9	78.2	9.5
Ours	SAM-L	84.6	5.0	83.9	78.9	87.4	74.6

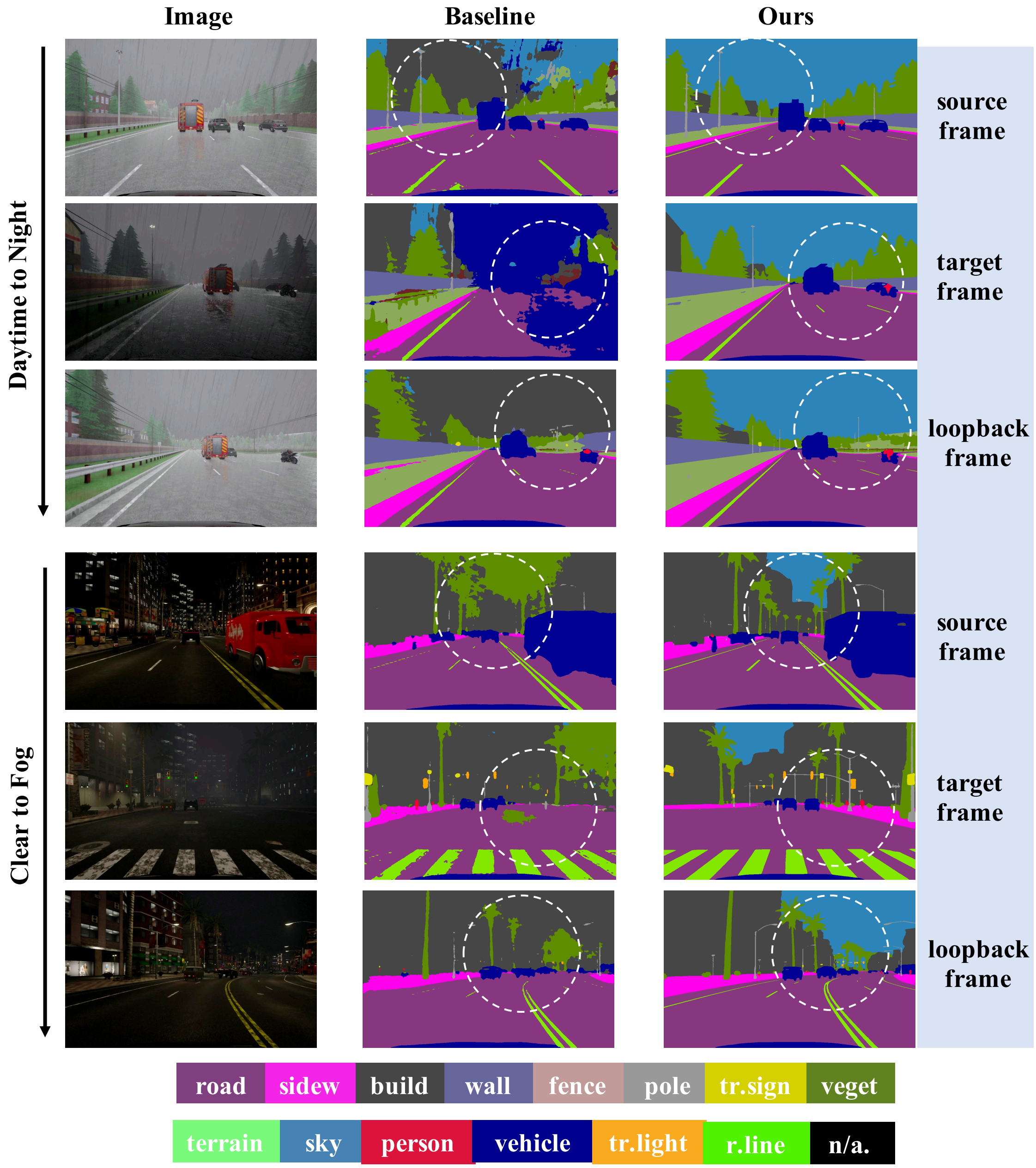

Figure 3. **Visualisation results of baseline (Deeplab V3+) and our method on the continous val set of the SHIFT dataset.**

3.3. Results & Analysis

Table. 1 shows the performance of the different methods on the SHIFT dataset's continuous validation set CTTA. We evaluate three domain shifts: daytime to night, clear to foggy, and clear to rainy. We chose Deeplab V3+ and our method for online testing on the SHIFT dataset's continuous test set. Table. 2 shows the evaluation results.

Effectiveness of each component. From the results in Table. 1, it can prove our method has better performance on

CTTA tasks. The foundation model SETR achieves a similar mIoU to the baseline model Deeplab V3+, but when we integrate a pre-trained foundation model, SAM (Segment Any Thing) to SETR, the initial generalization ability of our SETR gained a significant improvement. This gives the model a more stable semantic segmentation over a constantly changing domain. In further experiments, during continuous test time, we add low-rank and high-rank ViDAs to our SETR and end up getting a gain of 23.7% on mIoU. This excellent result shows that ViDAs can mitigate error accumulation and catastrophic forgetting. The results of the online test phase, Deeplab V3+ overall score is only 9.5, and our method overall score is up to 74.6. In this continuous domain shifts scenario, the difference between DeeplabV3 in the segmentation effect of the source frame and target frame is obvious, and it is apparent that the model does not learn enough domain-specific knowledge, which leads to its high mIoU_drop. On the contrary, our method owes the SAM pretrain's strong generalization ability, and the addition of ViDAs allows the model to continuously learn domain-specific and domain-agnostic knowledge. Hence, our model can better adapt domain shifts in test-time, and its mIoU_drop is only 9.5, which also indicates our method maintains the model's plasticity better on CTTA tasks.

4. Visualization

In order to show the ability of domain shift adaptation of different models more intuitively, we visualize the test results of all the experiments on the SHIFT continuous validation set, as shown in Figure 3.

5. Conclusion

In this report, our proposed method achieves very competitive results on the Continuous Test-time Adaptation for Semantic Segmentation track. We are very grateful to ICCV Workshop for organizing this visual continual learning challenge and hope that our work will inspire more ctta methods to achieve better performance on the SHIFT dataset.

References

[1] 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. 2

[2] 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, 2023. 3

[3] 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. 1

[4] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment anything, 2023. 2, 3

[5] Jiaming Liu, Senqiao Yang, Peidong Jia, Ming Lu, Yandong Guo, Wei Xue, and Shanghang Zhang. Vida: Homeostatic visual domain adapter for continual test time adaptation. *arXiv preprint arXiv:2306.04344*, 2023. 1, 3

[6] 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. 2

[7] 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*, pages 11920–11929, 2023. 1

[8] Tao Sun, Mattia Segu, Janis Postels, Yuxuan Wang, Luc Van Gool, Bernt Schiele, Federico Tombari, and Fisher Yu. Shift: a synthetic driving dataset for continuous multi-task domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21371–21382, 2022. 1

[9] Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. *Advances in neural information processing systems*, 30, 2017. 3

[10] Qin Wang, Olga Fink, Luc Van Gool, and Dengxin Dai. Continual test-time domain adaptation. *ArXiv*, abs/2203.13591, 2022. 1

[11] Qin Wang, Olga Fink, Luc Van Gool, and Dengxin Dai. Continual test-time domain adaptation, 2022. 3

[12] Senqiao Yang, Jiarui Wu, Jiaming Liu, Xiaoqi Li, Qizhe Zhang, Mingjie Pan, Mingjie Pan, and Shanghang Zhang. Exploring sparse visual prompt for cross-domain semantic segmentation, 2023. 1, 3

[13] Sixiao Zheng, Jiachen Lu, Hengshuang Zhao, Xiatian Zhu, Zekun Luo, Yabiao Wang, Yanwei Fu, Jianfeng Feng, Tao Xiang, Philip H. S. Torr, and Li Zhang. Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers, 2021. 2, 3