Exploring Sparse Visual Prompt for Domain Adaptive Dense Prediction

Senqiao Yang[1*], Jiarui Wu[1*], Jiaming Liu[1*], Xiaoqi Li[1], Qizhe Zhang[1],
Mingjie Pan[1], Yulu Gan[1], Zehui Chen[2], Shanghang Zhang[1†]
[1]National Key Laboratory for Multimedia Information Processing, School of Computer Science, Peking University
[2]University of Science and Technology of China
jiamingliu@stu.pku.edu.cn, shanghang@pku.edu.cn

Abstract

The visual prompts have provided an efficient manner in addressing visual cross-domain problems. In previous works, (Gan et al. 2022a) first introduces domain prompts to tackle the classification Test-Time Adaptation (TTA) problem by placing image-level prompts on the input and fine-tuning prompts for each target domain. However, since the image-level prompts mask out continuous spatial details in the prompt-allocated region, it will suffer from inaccurate contextual information and limited domain knowledge extraction, particularly when dealing with dense prediction TTA problems. To overcome these challenges, we propose a novel Sparse Visual Domain Prompts (SVDP) approach, which applies minimal trainable parameters (e.g., 0.1%) to pixels across the entire image and reserves more spatial information of the input. To better apply SVDP in extracting domain-specific knowledge, we introduce the Domain Prompt Placement (DPP) method to adaptively allocates trainable parameters of SVDP on the pixels with large distribution shifts. Furthermore, recognizing that each target domain sample exhibits a unique domain shift, we design Domain Prompt Updating (DPU) strategy to optimize prompt parameters differently for each sample, facilitating efficient adaptation to the target domain. Extensive experiments were conducted on widely-used TTA and continual TTA benchmarks, and our proposed method achieves state-of-the-art performance in both semantic segmentation and depth estimation tasks.

Introduction

Deep neural networks can achieve promising performance if test data is of the same distribution as the training data. However, it is not the common case in real-world scenarios (Radosavovic et al. 2022), which contain diverse and disparate domains. When applying a pre-trained model in real-world tasks, the domain gap commonly exists (Sakaridis, Dai, and Van Gool 2021), leading to significant performance degradation on target data. Though we can manually collect labeled data for each real-world target domain, it is laborious and time-consuming (Chen, Wu, and Jiang 2022). To this end, the domain adaptation (DA) methods are introduced and have drawn growing attention in the community.

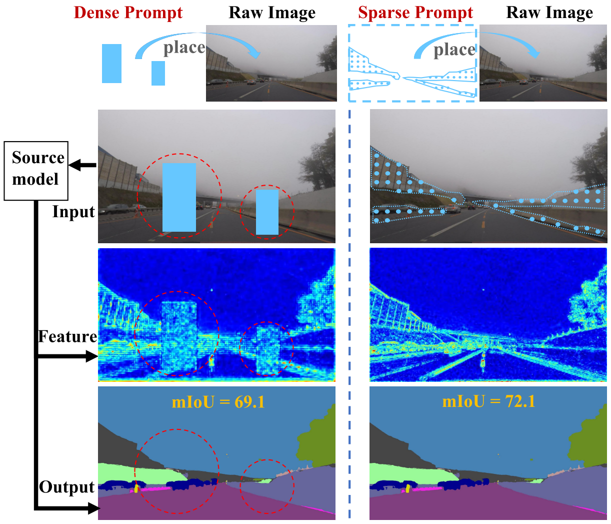

Figure 1: The motivation and main idea of our method. (a) Previous dense visual domain prompts (VDP) mask out consecutive spatial details in the placed regions as shown in red circles. In dense prediction DA problems, applying dense VDP will lead to inaccurate context information extraction and severe performance degradation. (b) We introduce Sparse Visual Domain Prompts (SVDP), which are tailored for addressing the occlusion problem of pixel-wise information and can better extract local domain knowledge for cross-domain learning.

While DA extensively investigates to address distribution shifts, its typical assumption involves access to raw source data. However, in real-world scenarios, raw data often cannot be publicly accessible due to data protection regulations. Meanwhile, traditional DA methods present resource-intensive backward computation, leading to high training costs (Ganin and Lempitsky 2015). To address this, Test-time adaptation (TTA) (Liang, He, and Tan 2023) is gained significant attention, which tackles distribution shifts at test time with only unlabeled test data streams. Prior TTA studies (Wang et al. 2021, 2022a; Chen et al. 2022a; Goyal et al. 2022) predominantly focus on model-based adaptation, utilizing model parameters to fit target domain knowledge.

To better solve the TTA problem, motivated by the re-

[*]Equal contribution
[†]Corresponding author

cent advances of prompting in NLP (Li and Liang 2021; Liu et al. 2023), VDP (Gan et al. 2022a) first introduces a prompt-based method to tackle the classification TTA problem. It employs image-level prompts to enhance domain transfer efficiency and effectiveness. Specifically, it randomly places the dense prompt on the input image and finetunes them to extract target domain knowledge. However, this prompt-based technique encounters limitations when applied to dense prediction tasks such as semantic segmentation and depth estimation TTA. Specifically, the dense prompts obscure continuous spatial information in the allocated regions, as illustrated in Figure 1 (a). This occlusion introduced by prompts leads to incomplete semantic knowledge representation, thereby negatively impacting the quality of segmentation maps. Simultaneously, the occluded details within corresponding features impede the extraction of adequate domain knowledge during cross-domain learning.

To this end, as shown in Fig.1 (b), we propose a novel Sparse Visual Domain Prompts (SVDP) approach for effectively extracting target domain knowledge, specially designed to combat domain shifts in dense prediction tasks. By introducing sparse prompts, which applies minimal trainable parameters (e.g., 0.1%) to pixels across the entire image, more spatial information from the input is preserved. Furthermore, the semantic information can be extracted sufficiently (shown in line 2), leading to noticeable improvements in segmentation outcomes (shown in line 3). In order to better apply SVDP in the pixel-wise TTA task, we propose the Domain Prompt Placement (DPP) to adaptively allocates trainable parameters of SVDP on the pixel with large distribution shifts. In this way, SVDP excels at extracting local domain knowledge, facilitating the transfer of pixel-wise data distribution from the source to the target domain. Furthermore, recognizing that each target domain sample exhibits a unique domain shift, we design a Domain Prompt Updating (DPU) method to optimize prompt parameters efficiently during the TTA process. Specifically, based on the extent of the domain gap observed in target domain samples, we employ varying weights to update the visual prompts. It's worth noting that we are the pioneers in designing specific strategies for pixel-level placement and image-level optimization in vision prompt learning, which work in synergy to address domain shifts in dense prediction TTA tasks.

Since data privacy and transmission limit access to source data in the real world, we evaluate our method on semantic segmentation and depth estimation source-free adaptation settings, including online TTA (Liang, Hu, and Feng 2020) and Continual Test-Time Adaptation (Wang et al. 2022a) (CTTA). Our proposed approach demonstrates superior performance compared to most state-of-the-art (SOTA) methods across three benchmarks, covering Cityscapes to ACDC (Sakaridis, Dai, and Van Gool 2021) and KITTI (Geiger, Lenz, and Urtasun 2012) to Drivingstereo (Yang et al. 2019). The main contributions are shown as follows:

1) We are the first for introducing the visual prompt approach to the dense prediction TTA problem. We propose a novel Sparse Visual Domain Prompts (SVDP) approach to better extract local domain knowledge and transfer pixel-wise data distribution from the source to the target domain.

2) In order to efficiently apply SVDP in pixel-wise TTA tasks, we propose Domain Prompt Placement (DPP) method to adaptively allocates trainable parameters in SVDP based on the degree of distribution shift at the pixel level. And Domain Prompt Updating (DPU) is designed to optimize prompt parameters differently for each sample, facilitating efficient adaptation on target domains.

3) We conduct extensive experiments to evaluate the effectiveness of our method, which outperforms most SOTA methods on four TTA and two CTTA benchmarks, covering semantic segmentation and depth estimation tasks.

Related Work

Test-Time Adaptation

Test-time adaptation (TTA), (Boudiaf et al. 2023; Kundu et al. 2020; Liang, Hu, and Feng 2020; Yang et al. 2021), aims to adapt a source model to an unknown target domain distribution without using any source domain data. Recent research has focused on using self-training or entropy regularization to fine-tune the source model. Specifically, SHOT (Liang, Hu, and Feng 2020) optimizes only the feature extractor using information maximization and pseudo labeling. AdaContrast (Chen et al. 2022b) also uses pseudo labeling for TTA, but introduces self-supervised contrastive learning to enhance performance. In addition to model-level adaptation, (Boudiaf et al. 2022) adjusts the output distribution to address this problem. Tent (Wang et al. 2021) updates the training parameters in the batch normalization layers by entropy minimization. Recent works (Niu et al. 2023; Yuan, Xie, and Li 2023) follow Tent to continually explore the robustness of normalization layers in the TTA process. While the aforementioned works primarily focus on classification tasks, there has been a recent surge of interest in performing TTA on dense prediction tasks (Shin et al. 2022; Song et al. 2022; Zhang et al. 2021). **Continual Test-Time Adaptation (CTTA)** is a scenario in which the target domain is not static, increasing challenges for traditional TTA methods (Wang et al. 2022a). (Wang et al. 2022a) serves as the first approach to tackle this task, using a combination of bi-average pseudo labels and stochastic weight reset. While (Wang et al. 2022a; Song et al. 2023) addresses the continual shifts at the model level, (Gan et al. 2022a) leverages visual domain prompts to address the problem in the classification task at the input level for the first time. In this paper, we evaluate our approach on both TTA and CTTA with a specific focus on the dense prediction task.

Prompt Learning

Visual prompts are inspired by their counterparts (Liu et al. 2021) which are used in natural language processing (NLP). Language prompts are presented as text instructions to improve the pre-trained language model's understanding of downstream tasks (Brown et al. 2020). Recently, researchers have attempted to discard text encoders and use prompts directly for visual tasks. (Bahng et al. 2022) employs visual prompts to pad input images, enabling pre-trained models to adapt to new tasks. Rather than fine-tuning the entire model, VPT (Conder et al. 2022; Jia et al. 2022a; Sandler et al. 2022; Wang et al. 2022b) inserts prompts into image or feature-level patches to adapt Transformer-based models. While these approaches all utilize opaque-block prompts, such prompts can cause performance degradation in dense prediction tasks. **Domain prompts** are first introduced in DAPL (Ge et al. 2022), which proposes a novel prompt learning paradigm for unsupervised domain adaptation (UDA). Embedding domain information using prompts can minimize the cost of fine-tuning and enable efficient domain adaptation. Recognizing the potential of prompt learning for UDA, MPA (Chen, Wu, and Jiang 2022) proposes multi-prompt alignment for multi-source UDA. DePT (Gao et al. 2022a) combines domain prompts with a hierarchical self-supervised regularization for TTA, which aims to solve the error accumulation problem in self-training. (Gan et al. 2022a) further divides domain prompts into domain-specific ones and domain-agnostic ones to address the more challenging CTTA task. However, these studies mainly focus on classification DA tasks. Our method, for the first time, applies sparse domain prompts to dense prediction DA tasks.

Method

Preliminaries

Test Time Adaptation (TTA) (Liang, He, and Tan 2023) aims at adapting a pre-trained model with parameters trained on the source data $(\mathcal{X}_S, \mathcal{Y}_S)$ to multiple unlabeled target data distribution $\mathcal{X}_{T_1}, \mathcal{X}_{T_2}, \ldots, \mathcal{X}_{T_n}$ at inference time. The entire process can not access any source domain data and can only access target domain data once. $\mathcal{X}_{T_i} = \{x_i^T\}_{i=1}^{N_t}$, where N_t denotes the scale of the target domain. The upcoming target domain can be a single one (TTA) or multiple continually changing distributions (CTTA), the latter of which is a more realistic setting that requires the model to achieve stability while preserving plasticity.

Domain Prompt. Inspired by language prompt in NLP, (Gan et al. 2022a) first introduces visual domain prompt (VDP) serving as a reminder to continually adapt to the target domain for the classification task, which aims to extract target domain-specific knowledge. Specifically, VDP (**p**) are dense learnable parameters added on the input image.

$$\widetilde{\mathbf{x}} = \mathbf{x} + \mathbf{p} \tag{1}$$

where x represent the original input image. The reformulated image $\widetilde{\mathbf{x}}$ will serve as the input for our model instead.

Motivation

Sparse Visual Domain Prompt. Traditional visual prompts (Jia et al. 2022b) are deployed on the image or feature level to realize fine-tuning by updating a small number of prompt parameters. Recent works (Gan et al. 2022a; Gao et al. 2022b) explore visual prompts in classification DA problems, which extract domain knowledge for the target domain and transfer data distribution from the source to the target domain. However, DePT (Gao et al. 2022b) concatenate the domain prompts with class token and image tokens to the input of transformer layers, which neglect the local domain knowledge extraction. Meanwhile, VDP (Gan et al. 2022a) randomly set the locations of dense prompts on the input image, masking out continuous spatial details in prompt allocated regions. Different from classification cross-domain learning, dense prediction DA not only requires global domain knowledge but also relies on extracting intact local domain knowledge. As shown in Fig.1(a), partial spatial information deficiency caused by dense prompts will lead to inaccurate contextual information and negative effects on target domain knowledge extraction. This observation motivates us to propose a novel Sparse Visual Domain Prompts (SVDP), which is tailored for pixel-wise prediction DA tasks. It inserts minimal trainable parameters into pixels across the entire image and reserves more spatial information.

Domain Prompt Placement. Previous work (Gan et al. 2022a; Gao et al. 2022b) randomly put the prompts on the target domain image to extract global domain knowledge. Specifically, it may set prompts on regions with trivial domain shifts, hindering the extraction of local domain knowledge. Especially in the source-free TTA setting, we can only access target domain data once, which makes the efficiency of transfer learning crucial. Therefore, we propose Domain Prompt Placement (DPP) which efficiently extracts more domain-specific knowledge and addresses local domain shift. Specifically, we measure the degree of domain gap by general uncertainty scheme (Gal and Ghahramani 2016; Guan et al. 2021; Roy et al. 2022; Gan et al. 2022b) and tactfully place trainable parameters of SVDP on the pixel with large distribution shifts.

Domain Prompt Updating. The amount of prompt parameters is minimal which brings the challenge of fully learning target domain knowledge during TTA process. Meanwhile, the degree of domain shift is not only various on pixels within the image but also on each target domain test sample. It thus motivates us to update prompt parameters differently for each target sample. Therefore, we design a Domain Prompt Updating (DPU) which efficiently optimizes prompt parameters to fit in target domain distribution. Specifically, we adopt the same uncertainty scheme to measure the degree of domain shift for each target sample. According to the degree, we update prompt parameters for the individual sample with different updating weights.

Sparse Visual Domain Prompt

SVDP maintains the same resolution as the input image ($\mathbf{p} \in \mathbb{R}^{H \times W \times 3}$), it only masks out original information by minimal discrete trainable parameters (e.g. 0.1%) on the pixels with large domain shifts. Compared with the previous dense visual prompt, SVDP preserves more contextual information and possesses the capacity to capture local domain knowledge through pixel-wise prompt parameters. The overall framework of our method is shown in Fig .2, and the specially designed prompt Placement and Updating methods are introduced in the following.

Domain Prompt Placing

We propose the Domain Prompt Placement (DPP) strategy of SVDP to efficiently extract local domain knowledge in pixel-wise. We intend to place trainable parameters of SVDP on the pixel with large distribution shifts and adapt pixel-wise data distribution from source to target domain.

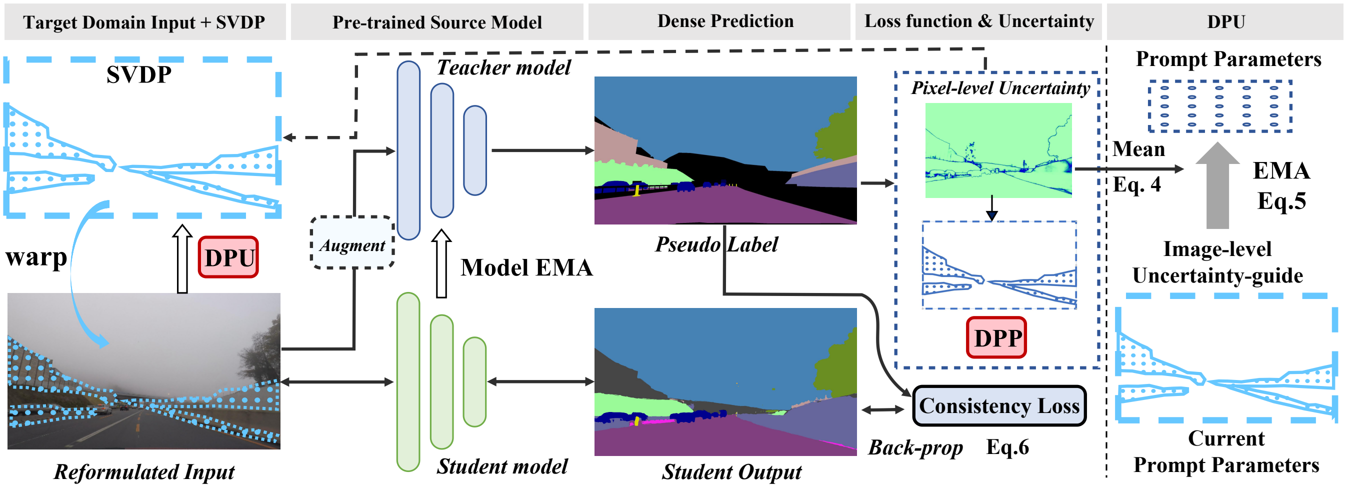

Figure 2: The overall framework. Left: We warp the SVDP into the image and place prompt parameters on the selected pixel by Domain Prompt Placement (DPP) method. The reformulated image serves as the input of the teacher and student model. We obtain the uncertainty map as described in Eq. (2) through the teacher model. The uncertainty map is used to evaluate the degree of pixel-level distribution shift. SVDP adopts consistency loss (Eq. (6)) and exponential moving average (EMA) as the optimization strategies. Right: Domain Prompt Updating (DPU). Based on the image-level uncertainty value, we adopt different EMA weights to realize stable updating of SVDP parameters, facilitating efficient adaptation to the target domain.

As shown in Fig. 3, we employ the MC Dropout method (Gal and Ghahramani 2016) to perform m forward propagations ($m = 10$) and obtain m group probabilities for each pixel. Specifically, dropout operation is only applied to the linear layer within the prediction head. Calculating the uncertainty value does not significantly increase the computational cost. Meanwhile, we can also obtain m sets of probabilities through the simpler method of input image resolution augmentation. Inspired by (Roy et al. 2022; Gan et al. 2022b), we calculate the uncertainty value (Eq.(2)) of the input and figure out the pixel-wise degree of domain shift.

$$\mathcal{U}(\widetilde{x}_j) = \left(\frac{1}{m} \sum_{i=1}^{m} \| p_i(\widetilde{y}_j | \widetilde{x}_j) - \mu \|^2 \right)^{\frac{1}{2}} \tag{2}$$

, where $p_i(\widetilde{y}_j|\widetilde{x}_j)$ is the predicted probability of input pixel $\widetilde{x}_j$ in the i^{th} forward propagation, and μ is the mean prediction (m rounds) of $\widetilde{x}_j$. $\mathcal{U}(\widetilde{x}_j)$ thus represents the uncertainty of the source model for pixel-wise target input $\widetilde{x}_j$. As shown in the bottom of Fig .3, we sort all pixels based on their pixel-wise uncertainty value and place prompt parameters on the pixels with high uncertainty score.

Domain Prompt Updating

Motivated by the fact that the mean teacher predictions have a higher quality than the standard model (Tarvainen and Valpola 2017a), we utilize a mean-teacher model to provide more accurate predictions in the TTA process. To be specific, we adopt the widely-used exponential moving average (EMA) to achieve the model and prompt updating. Same as previous works (Wang et al. 2022a), the teacher model ($\mathcal{T}_{mean}$) is updated by EMA from the student model ($\mathcal{S}_{target}$), shown in Eq. (3):

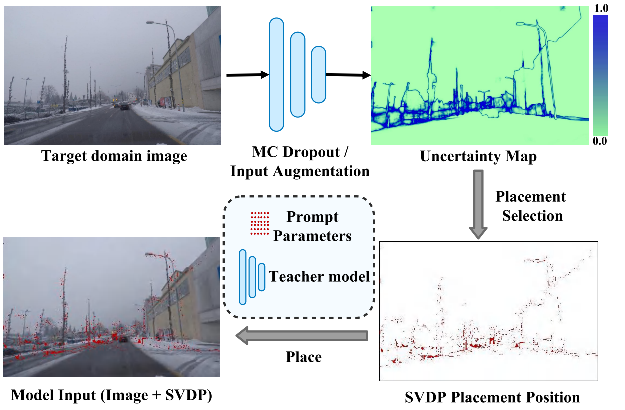

Figure 3: The detailed process of Domain Prompt Placing. The uncertainty map is estimated by MC Dropout (Gal and Ghahramani 2016). The SVDP parameters are placed on the pixels with high uncertainty, then warp into the raw image.

$$\mathcal{T}_{mean}^{t} = \alpha \mathcal{T}_{mean}^{t-1} + (1 - \alpha) \mathcal{S}_{target}^{t} \tag{3}$$

When $t = 0$ (t is the time step), we utilize the source domain pre-trained model to initialize the weight of the teacher and student model. And we set $\alpha = 0.999$ (Tarvainen and Valpola 2017b), which is the updating weight of EMA.

Different from traditional model updating, we design a special Domain Prompt Updating (DPU) strategy for SVDP to stably fit in target domain distribution. As shown in Fig .4, we adopt image-level uncertainty value to reflect the degree of domain shift for each target domain sample. We calculate the image-level uncertainty value $\mathcal{U}(x)$ by average the pixel-

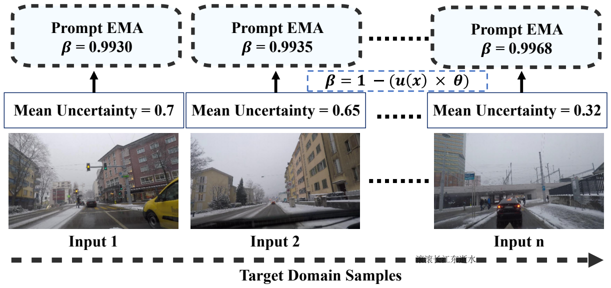

Figure 4: The process of Domain Prompt Updating. We adaptively adjust the prompt EMA updating rate for each target domain sample based on image-level uncertainty value.

wise uncertainty, shown in Eq. (4):

$$\mathcal{U}(x) = \frac{1}{H \times W} \sum_{j}^{H \times W} \mathcal{U}(\widetilde{x}_j) \tag{4}$$

Based on the image-level uncertainty score, we update prompt parameters for each sample with different weight.

$$p_t = \beta p_{t-1} + (1 - \beta) p_t, \tag{5}$$

Note that, p_t represents the parameters of the SVDP that is updated by Eq. (6). In DPU, we set the prompt EMA updating rate $\beta = 1 - (\mathcal{U}(x) \times \theta)$. θ is intended to bring the value of uncertainty up to the same order of magnitude as the value of the common EMA update parameter (e.g., $\theta = 0.01$). As shown in the top of Fig .4, the prompt EMA weight is set to a large value when the input is of high uncertainty score since the large weight can efficiently adapt to the sample with the large data distribution shift.

Loss Function

We utilize teacher model to generate the pseudo labels ($\widetilde{y}_t$), which is refined by test-time augmentation and confidence filter (Wang et al. 2022a). Then, we adopt consistency loss (L_{con}) as the optimization objective for segmentation task, which is a pixel-wise cross-entropy loss (Xie et al. 2021).

$$\mathcal{L}_{con}(\widetilde{x}) = -\frac{1}{H \times W} \sum_{w,h}^{W,H} \sum_{c}^{C} \widetilde{y}_t(w,h,c) \log \hat{y}_t(w,h,c) \tag{6}$$

Where $\hat{y}_t$ is the output of our student model, C means the amount of categories.

Experiments

In the first subsection, we provide the details of the task settings for test-time adaptation (TTA) and continual TTA (CTTA), as well as a description of the datasets. In the second and third subsections, we compare our method with other baselines (Xie et al. 2021; Wang et al. 2021, 2022a; Gao et al. 2022b,b) in four TTA and two CTTA scenarios. In the final subsection, comprehensive ablation study explore the impact of each component.

Task Settings and Datasets

TTA and CTTA are commonly used source-free technology in real-world scenarios in which a source pre-trained model adapts to the distribution of an unseen target domain (Liang, He, and Tan 2023). CTTA is of the same setting as TTA but further sets the target domain constantly changing, bringing more difficulties during the continual adaptation process.

Cityscapes-to-ACDC is designed for semantic segmentation cross-domain learning. And we conduct four TTA and one CTTA experiment on the scenario. The source model is an off-the-shelf pre-trained segmentation model that was trained on the Cityscapes dataset (Cordts et al. 2016). The ACDC dataset (Sakaridis, Dai, and Van Gool 2021) contains images collected in four different unseen visual conditions: Fog, Night, Rain, and Snow. For the TTA, we adapt the source pre-trained model to each of the four ACDC target domains separately. For the CTTA, we repeat the same sequence of target domains (Fog→Night→Rain→Snow) multiple times to simulate environment changes in real-world scenarios (Wang et al. 2022a).

KITTI-to-Driving Stereo. To demonstrate the generalization of our method, we also conduct experiments in depth estimation CTTA scenario. The source model employed is an off-the-shelf, pre-trained model, initially trained on the KITTI dataset (Geiger, Lenz, and Urtasun 2012). The Driving Stereo(Yang et al. 2019) comprises images collected under four disparate, unseen visual conditions: foggy, rainy, sunny, and cloudy. For the CTTA, we repeat the same sequence of target domains (Foggy→Rainy→Sunny→Cloudy) multiple times.

Implementation Details. We follow the implementation details (Wang et al. 2022a) to set up our semantic segmentation TTA experiments. Specifically, we use the Segformer-B5 (Xie et al. 2021) pre-trained on Cityscapes as our off-the-shelf source model. We down-sample the original image size of 1920x1080 of the ACDC dataset to 960x540, which serves as network input. We evaluate our predictions under the original resolution. We use a range of image resolution scale factors [0.5, 0.75, 1.0, 1.25, 1.5, 1.75, 2.0] for the augmentation method in teacher model. For depth estimation CTTA, we follow the implementation details in previous work (Liu et al. 2022) and adopt the pre-trained DPT (Ranftl, Bochkovskiy, and Koltun 2021) on the KITTI as the source model. The optimizer is performed using Adam optimizer (Kingma and Ba 2014) with $(\beta_1, \beta_2) = (0.9, 0.999)$. We set the learning rate specific values for each backbone, such as 3e-4 for Segformer and 1e-4 for DPT, and batch size 1 for both TTA and CTTA experiments. All experiments are conducted on NVIDIA A100 GPUs.

The Effectiveness on Semantic Segmentation

Cityscapes-to-ACDC TTA. We evaluate the performance of the proposed SVDP on four scenarios with significant domain gap during TTA. Tab .1 shows that the Mean-mIoU for the four domains using the source domain model alone is only 56.7%. Recent advanced methods CoTTA increases it to 58.6% while our method further increases it by 1.5%. These results demonstrate that our method can better address the domain shit problem in test time compared to other

Test-Time Adaptation	Source2Fog		Source2Night		Source2Rain		Source2Snow		Mean-mIoU
Method	mIoU↑	mAcc↑	mIoU↑	mAcc↑	mIoU↑	mAcc↑	mIoU↑	mAcc↑	
Source	69.1	79.4	40.3	55.6	59.7	74.4	57.8	69.9	56.7
TENT	69.0	79.5	40.3	55.5	59.9	74.1	57.7	69.7	56.7
CoTTA	70.9	80.2	41.2	55.5	62.6	75.4	59.8	70.7	58.6
DePT	71.0	80.2	40.9	**55.8**	61.3	74.4	59.5	70.0	58.2
VDP	70.9	80.3	41.2	55.6	62.3	75.5	59.7	70.7	58.5
SVDP	**72.1**	**81.2**	**42.0**	54.9	**64.4**	**76.7**	**62.2**	**72.8**	**60.1±0.2**

Table 1: Performance comparison of Cityscapes-to-ACDC TTA. We use Cityscape as the source domain and ACDC as the four target domains in this setting. Mean-mIoU represents the average mIoU value in four TTA experiments.

Time	$t \longrightarrow$																
Round	1					2					3					Mean↑	Gain
Method	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	/
TENT	69.0	40.2	60.1	57.3	56.7	68.3	39.0	60.1	56.3	55.9	67.5	37.8	59.6	55.0	55.0	55.7	-1.0
CoTTA	70.9	41.2	62.4	59.7	58.6	70.9	41.1	62.6	59.7	58.6	70.9	41.0	62.7	59.7	58.6	58.6	+1.9
DePT	71.0	40.8	58.2	56.8	56.5	68.2	40.0	55.4	53.7	54.3	66.4	38.0	47.3	47.2	49.7	53.4	-3.3
VDP	70.5	41.1	62.1	59.5	58.3	70.4	41.1	62.2	59.4	58.2	70.4	41.0	62.2	59.4	58.2	58.2	+1.5
SVDP	**72.1**	**44.0**	**65.2**	**63.0**	**61.1**	**72.2**	**44.5**	**65.9**	**63.5**	**61.5**	**72.1**	**44.2**	**65.6**	**63.6**	**61.4**	**61.1±0.3**	**+4.4±0.3**

Table 2: Performance comparison for Cityscape-to-ACDC CTTA. We take the Cityscape as the source domain and ACDC as the continual target domains. During testing, we sequentially evaluate the four target domains three times. Mean is the average score of mIoU. Gain refers to the improvement achieved by the method compared to the Source model.

methods. Furthermore, in contrast to VDP, which employs dense prompts, our method successfully circumvents the occlusion issue, leading to improved extraction of both semantic and domain knowledge for TTA. In comparison to DePT, which introduces prompts at the token level, our SVDP approach operates at the image-level. This aspect enables the extraction of local domain knowledge, thereby resulting in substantial performance enhancements.

Cityscapes-to-ACDC CTTA. To demonstrate that our method can also address continuously changing domain shifts, we deal with the four domain data during test time periodically. As shown in Tab .2, due to catastrophic forgetting, the performance of TENT and DePT gradually decreases over time. These methods only focus on acquiring new domain-specific knowledge from the target domain, resulting in a neglect of the original knowledge from the source domain. And we find that our method gains 2.5% increase of mIoU more than the previous SOTA CTTA method (Wang et al. 2022a). The results prove that our method can continuously extract target domain knowledge via sparse prompt and preserve previous domain knowledge via model parameters, showing the ability to address dynamic domain shifts. In term of qualitative analysis, shown in Fig .5, our method correctly distinguish the sidewalk from the road, avoiding mis-classification in target domains.

Overall, our method outperforms several previous SOTA methods on all semantic segmentation TTA and CTTA tasks and shows promising potential for real-world applications.

The Effectiveness on Depth Estimation

KITTI-to-Driving Stereo CTTA. To demonstrate the effectiveness of our approach in addressing CTTA problem in depth estimation task, we conducted a series of evaluations on four distinct target domains from the Driving Stereo dataset at regular intervals during the testing phase. As shown in Table **??**, our method consistently outperforms the state-of-the-art (SOTA) technique across all four evaluation metrics. Particularly noteworthy is the significant enhancement in the mean $\delta > 1.25$, achieving a remarkable improvement of 65.9% when compared to the Source model, and an impressive 7.5% improvement over the previous SOTA method. This result underscores the robust continual adaptation ability of our method in the context of depth estimation. Given that CTTA has access to the data only once, as opposed to CoTTA, our approach leverages sparse prompt to effectively adapt to the target domain, resulting in significant performance gains. Overall, these results show that our SVDP consistently attains superior outcomes in the depth estimation tasks.

Ablation Study

In this subsection, we evaluate the contribution of each component in our method. Since the CTTA is the most challenging and realistic scenario, we conduct the ablation study on the KITTI-to-Driving Stero CTTA. **Effectiveness of each component.** As presented in Tab. 3 Ex_2, Teacher-student (TS) structure is a common technique in CTTA (Wang et al. 2022a; Gan et al. 2022a), which is used to generate pseudo label in the target domain and only has 0.063 Abs Rel reduces without our method. This verifies

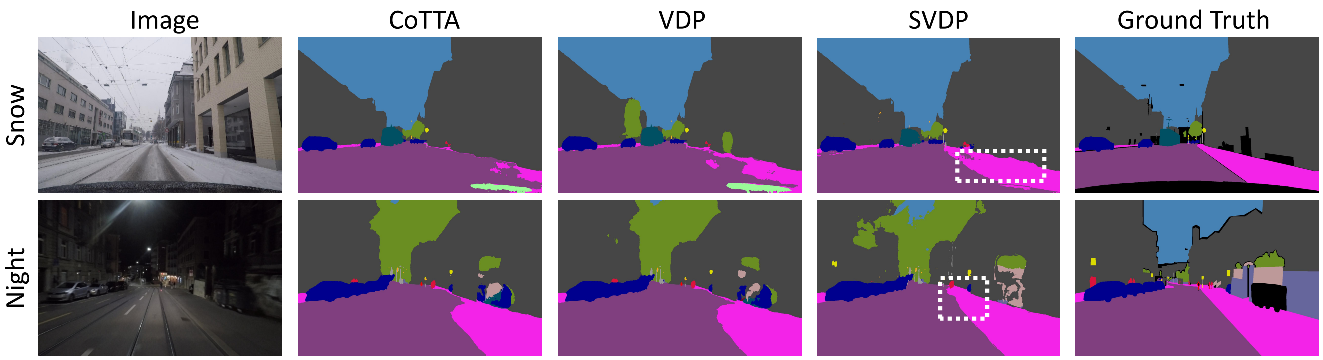

Figure 5: Qualitative comparison of our method with previous SOTA methods on the ACDC dataset. Our method could better segment different pixel-wise classes such as shown in the white box.

	TS	SVDP	DPP	DPU	Abs Rel↓	$\delta > 1.25\uparrow$
Ex_1					0.312	0.093
Ex_2	✓				0.249	0.503
Ex_3	✓	✓			0.187	0.705
Ex_4	✓	✓	✓		0.169	0.737
Ex_5	✓	✓		✓	0.177	0.723
Ex_6	✓	✓	✓	✓	0.162	0.741

Table 3: Ablation: Contribution of each component.

the improvement of our method does not come from the usage of this prevalent scheme. In Ex_3, by introducing sparse prompts (SVDP), we observe that the Abs Rel reduces 0.062 and $\delta > 1.25$ increases 20.2%, respectively. The result demonstrates that SVDP facilitates addressing the domain shift problem, since it can extract local target domain knowledge without damaging the original semantic information. As shown in Ex_4, DPP achieves further 0.018 Abs Rel reduces and 3.2% $\delta > 1.25$ improvement since the specially designed prompt placement strategy can assist SVDP in extracting target domain-specific knowledge more efficiently. Compared with Ex_3, DPU (Ex_5) also reduces the Abs Rel 0.01 and improves 1.8% $\delta > 1.25$, respectively. The results prove the effectiveness of DPU and show the importance of adaptively optimizing for different samples during TTA process. Ex_6 shows the complete combination of all components which achieves 64.8% $\delta > 1.25$ improvement and 0.150 Abs Rel reduction in total. It proves that all components compensate each other and jointly address the depth estimation domain shift problem in test time.

How does the prompt sparsity affect the performance? As shown in Fig .6, we investigate the performance impact caused by the sparsity of SVDP. Specifically, we gradually increase the density of SVDP pixel-wise parameters and place it into more pixel. We find that $\delta > 1.25$ initially improves along with increasing SVDP density and then starts to decrease when the density exceeds 1e-3. This observation suggests that when SVDP is excessively sparse, it fails to capture the domain-specific knowledge effectively due to the limited number of parameters. In contrast, if the SVDP becomes too dense, the prompt will occlude many spatial details, leading to depth estimation performance degradation. Therefore, it is crucial to strike a balance on the degree of prompt sparsity and we consider that SVDP can achieve optimal potential in 1e-3 sparsity.

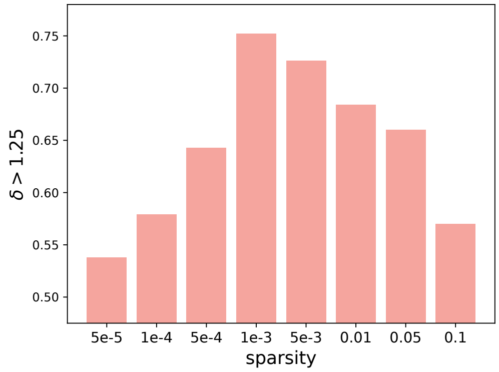

Figure 6: Effect of prompts' sparsity

Conclusion

In this paper, we are the first to introduce the Sparse Visual Domain Prompt (SVDP) in dense prediction TTA tasks (i.e., semantic segmentation, depth estimation), which address the problem of inaccurate contextual information extraction and insufficient domain-specific feature transferring caused by dense prompt occlusion. Moreover, the Domain Prompt Placement (DPP) and Domain Prompt Updating (DPU) strategies are specially designed for applying SVDP to ease the pixel wise domain shift better. Our method demonstrates state-of-the-art performance and effectively addresses domain shift through extensive experimentation across various TTA and CTTA scenarios.

Acknowledgments

Shanghang Zhang is supported by the National Key Research and Development Project of China (No.2022ZD0117801). This work was supported by Emerging Engincering interdisciplinar Youth Program at

Peking University under Grant 7100604372

References

Bahng, H.; Jahanian, A.; Sankaranarayanan, S.; and Isola, P. 2022. Exploring Visual Prompts for Adapting Large-Scale Models.

Boudiaf, M.; Denton, T.; van Merriënboer, B.; Dumoulin, V.; and Triantafillou, E. 2023. In Search for a Generalizable Method for Source Free Domain Adaptation.

Boudiaf, M.; Mueller, R.; Ayed, I. B.; and Bertinetto, L. 2022. Parameter-free Online Test-time Adaptation. *ArXiv*, abs/2201.05718.

Brown, T. B.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; Agarwal, S.; Herbert-Voss, A.; Krueger, G.; Henighan, T.; Child, R.; Ramesh, A.; Ziegler, D. M.; Wu, J.; Winter, C.; Hesse, C.; Chen, M.; Sigler, E.; Litwin, M.; Gray, S.; Chess, B.; Clark, J.; Berner, C.; McCandlish, S.; Radford, A.; Sutskever, I.; and Amodei, D. 2020. Language Models are Few-Shot Learners.

Chen, D.; Wang, D.; Darrell, T.; and Ebrahimi, S. 2022a. Contrastive test-time adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 295–305.

Chen, D.; Wang, D.; Darrell, T.; and Ebrahimi, S. 2022b. Contrastive Test-Time Adaptation. *ArXiv*, abs/2204.10377.

Chen, H.; Wu, Z.; and Jiang, Y.-G. 2022. Multi-Prompt Alignment for Multi-source Unsupervised Domain Adaptation. *arXiv preprint arXiv:2209.15210*.

Conder, J.; Jefferson, J.; Pages, N.; Jawed, K.; Nejati, A.; and Sagar, M. 2022. Efficient Transfer Learning for Visual Tasks via Continuous Optimization of Prompts.

Cordts, M.; Omran, M.; Ramos, S.; Rehfeld, T.; Enzweiler, M.; Benenson, R.; Franke, U.; Roth, S.; and Schiele, B. 2016. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 3213–3223.

Gal, Y.; and Ghahramani, Z. 2016. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning*, 1050–1059. PMLR.

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

Gan, Y.; Pan, M.; Zhang, R.; Ling, Z.; Zhao, L.; Liu, J.; and Zhang, S. 2022b. Cloud-Device Collaborative Adaptation to Continual Changing Environments in the Real-world. *arXiv preprint arXiv:2212.00972*.

Ganin, Y.; and Lempitsky, V. 2015. Unsupervised domain adaptation by backpropagation. In *International conference on machine learning*, 1180–1189. PMLR.

Gao, Y.; Shi, X.; Zhu, Y.; Wang, H.; Tang, Z.; Zhou, X.; Li, M.; and Metaxas, D. N. 2022a. Visual Prompt Tuning for Test-time Domain Adaptation.

Gao, Y.; Shi, X.; Zhu, Y.; Wang, H.; Tang, Z.; Zhou, X.; Li, M.; and Metaxas, D. N. 2022b. Visual Prompt Tuning for Test-time Domain Adaptation. *arXiv preprint arXiv:2210.04831*.

Ge, C.; Huang, R.; Xie, M.; Lai, Z.; Song, S.; Li, S.; and Huang, G. 2022. Domain Adaptation via Prompt Learning. *ArXiv*, abs/2202.06687.

Geiger, A.; Lenz, P.; and Urtasun, R. 2012. Are we ready for autonomous driving? the kitti vision benchmark suite. In *2012 IEEE conference on computer vision and pattern recognition*, 3354–3361. IEEE.

Goyal, S.; Sun, M.; Raghunathan, A.; and Kolter, J. Z. 2022. Test time adaptation via conjugate pseudo-labels. *Advances in Neural Information Processing Systems*, 35: 6204–6218.

Guan, D.; Huang, J.; Xiao, A.; Lu, S.; and Cao, Y. 2021. Uncertainty-aware unsupervised domain adaptation in object detection. *IEEE Transactions on Multimedia*, 24: 2502–2514.

Jia, M.; Tang, L.; Chen, B.-C.; Cardie, C.; Belongie, S.; Hariharan, B.; and Lim, S.-N. 2022a. Visual Prompt Tuning.

Jia, M.; Tang, L.; Chen, B.-C.; Cardie, C.; Belongie, S.; Hariharan, B.; and Lim, S.-N. 2022b. Visual prompt tuning. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXIII*, 709–727. Springer.

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

Kundu, J. N.; Venkat, N.; M, R.; and Babu, R. V. 2020. Universal Source-Free Domain Adaptation.

Li, X. L.; and Liang, P. 2021. Prefix-tuning: Optimizing continuous prompts for generation. *arXiv preprint arXiv:2101.00190*.

Liang, J.; He, R.; and Tan, T. 2023. A comprehensive survey on test-time adaptation under distribution shifts. *arXiv preprint arXiv:2303.15361*.

Liang, J.; Hu, D.; and Feng, J. 2020. Do We Really Need to Access the Source Data? Source Hypothesis Transfer for Unsupervised Domain Adaptation. In *ICML*.

Liu, J.; Zhang, Q.; Li, J.; Lu, M.; Huang, T.; and Zhang, S. 2022. Unsupervised Spike Depth Estimation via Cross-modality Cross-domain Knowledge Transfer. *arXiv preprint arXiv:2208.12527*.

Liu, P.; Yuan, W.; Fu, J.; Jiang, Z.; Hayashi, H.; and Neubig, G. 2021. Pre-train, Prompt, and Predict: A Systematic Survey of Prompting Methods in Natural Language Processing. *arXiv: Computation and Language*.

Liu, P.; Yuan, W.; Fu, J.; Jiang, Z.; Hayashi, H.; and Neubig, G. 2023. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Computing Surveys*, 55(9): 1–35.

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

Radosavovic, I.; Xiao, T.; James, S.; Abbeel, P.; Malik, J.; and Darrell, T. 2022. Real-World Robot Learning with Masked Visual Pre-training. *CoRL*.

Ranftl, R.; Bochkovskiy, A.; and Koltun, V. 2021. Vision Transformers for Dense Prediction. *ArXiv preprint*.

Roy, S.; Trapp, M.; Pilzer, A.; Kannala, J.; Sebe, N.; Ricci, E.; and Solin, A. 2022. Uncertainty-guided source-free domain adaptation. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXV*, 537–555. Springer.

Sakaridis, C.; Dai, D.; and Van Gool, L. 2021. ACDC: The adverse conditions dataset with correspondences for semantic driving scene understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 10765–10775.

Sandler, M.; Zhmoginov, A.; Vladymyrov, M.; and Jackson, A. 2022. Fine-tuning Image Transformers using Learnable Memory.

Shin, I.; Tsai, Y.-H.; Zhuang, B.; Schulter, S.; Liu, B.; Garg, S.; Kweon, I. S.; and Yoon, K.-J. 2022. MM-TTA: Multi-Modal Test-Time Adaptation for 3D Semantic Segmentation.

Song, J.; Lee, J.; Kweon, I. S.; and Choi, S. 2023. EcoTTA: Memory-Efficient Continual Test-time Adaptation via Self-distilled Regularization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11920–11929.

Song, J.; Park, K.; Shin, I.; Woo, S.; and Kweon, I. S. 2022. CD-TTA: Compound Domain Test-time Adaptation for Semantic Segmentation.

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

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

Wang, D.; Shelhamer, E.; Liu, S.; Olshausen, B. A.; and Darrell, T. 2021. Tent: Fully Test-Time Adaptation by Entropy Minimization. In *ICLR*.

Wang, Q.; Fink, O.; Gool, L. V.; and Dai, D. 2022a. Continual Test-Time Domain Adaptation. *ArXiv*, abs/2203.13591.

Wang, Z.; Zhang, Z.; Lee, C.-Y.; Zhang, H.; Sun, R.; Ren, X.; Su, G.; Perot, V.; Dy, J.; and Pfister, T. 2022b. Learning to Prompt for Continual Learning. In *CVPR*.

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

Yang, G.; Song, X.; Huang, C.; Deng, Z.; Shi, J.; and Zhou, B. 2019. DrivingStereo: A Large-Scale Dataset for Stereo Matching in Autonomous Driving Scenarios. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Yang, S.; Wang, Y.; van de Weijer, J.; Herranz, L.; and Jui, S. 2021. Generalized Source-Free Domain Adaptation. *international conference on computer vision*.

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

Zhang, Y.; Borse, S.; Cai, H.; and Porikli, F. 2021. Aux-Adapt: Stable and Efficient Test-Time Adaptation for Temporally Consistent Video Semantic Segmentation.