Exploring Sparse Visual Prompt for Cross-domain Semantic Segmentation

Senqiao Yang[1,2*], Jiarui Wu[1,3*], Jiaming Liu [1*], Xiaoqi Li[1], Qizhe Zhang[1],
Mingjie Pan[1], Yulu Gan[1], Shanghang Zhang[1†]
[1]Peking University, [2]Harbin Institute of Technology, Shenzhen, [3] Beihang University
{liujiaming.pku, shzhang.pku}@gmail.com

Abstract

Visual Domain Prompts (VDP) have shown promising potential in addressing visual cross-domain problems. Existing methods adopt VDP in classification domain adaptation (DA), such as tuning image-level or feature-level prompts for target domains. Since the previous dense prompts are opaque and mask out continuous spatial details in the prompt regions, it will suffer from inaccurate contextual information extraction and insufficient domain-specific feature transferring when dealing with the dense prediction (i.e. semantic segmentation) DA problems. Therefore, we propose a novel Sparse Visual Domain Prompts (SVDP) approach tailored for addressing domain shift problems in semantic segmentation, which holds minimal discrete trainable parameters (e.g. 10%) of the prompt and reserves more spatial information. To better apply SVDP, we propose Domain Prompt Placement (DPP) method to adaptively distribute several SVDP on regions with large data distribution distance based on uncertainty guidance. It aims to extract more local domain-specific knowledge and realizes efficient cross-domain learning. Furthermore, we design a Domain Prompt Updating (DPU) method to optimize prompt parameters differently for each target domain sample with different degrees of domain shift, which helps SVDP to better fit target domain knowledge. Experiments, which are conducted on the widely-used benchmarks (Cityscapes, Foggy-Cityscapes, and ACDC), show that our proposed method achieves state-of-the-art performances on the source-free adaptations, including six Test Time Adaptation and one Continual Test-Time Adaptation in semantic segmentation. The code will be released.

1. 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 [36], which contain diverse and disparate domains. When applying a pre-trained model in real-world tasks, the domain gap commonly exists [40], 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 [7]. Therefore, the Domain Adaptation (DA) methods (i.e. Test Time Adaptation, Continual Test-Time Adaptation) are introduced and have drawn growing attention in the community.

Recently, motivated by the recent advances of Prompting in NLP [28, 29, 34], visual prompt [22] is introduced in computer vision tasks which fine-tunes a small number of trainable parameters. Specifically, [12, 7, 14, 21] utilize Visual Domain Prompts (VDP) to address the classification domain shift problem by dynamically updating the domain prompts for the target domain. These works randomly set the dense VDP on the input or feature-level and fine-tune them to extract target domain knowledge or maintain the domain-invariant knowledge [12]. However, when these methods are applied in dense prediction (e.g. semantic segmentation) DA problems, the opaque dense prompts will mask out continuous spatial information in the prompt regions. As shown in Fig .1 (a), due to the occlusion brought by prompts, the feature representation suffers from partial semantic knowledge deficiency, leading to a negative impact on semantic segmentation. Meanwhile, the occluded information in the corresponding feature latent space also results in insufficient domain knowledge extraction during cross-domain learning.

To this end, as shown in Fig .1 (b), we propose a novel Sparse Visual Domain Prompts (SVDP) approach to better extract target domain knowledge, which is tailored for addressing domain shift in semantic segmentation. By introducing image-level sparse prompts which are of minimal discrete trainable parameters (e.g. 10%) of the previous prompt, we are able to reserve more spatial information. Furthermore, the corresponding semantic information can be extracted from the prompt regions (line 2 of Fig .1) and the segmentation result will be obviously improved (line 3 of Fig .1). In order to better apply SVDP in the pixel-wise DA task, we propose the Domain Prompt Placement (DPP)

*Equal contribution: liujiaming.pku@gmail.com
†Corresponding author: shzhang.pku@gmail.com

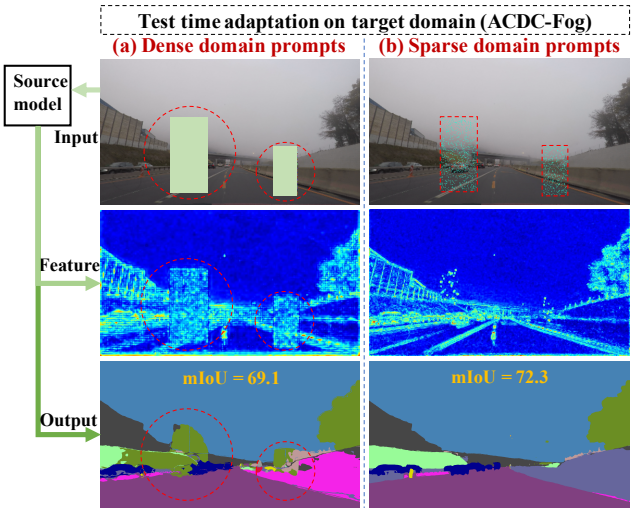

Figure 1. **The motivation and main idea of our method.** (a) Traditional dense visual domain prompts (VDP) are opaque and mask out consecutive spatial details in the prompt regions as shown in red circles. In semantic segmentation DA problems, applying dense VDP will lead to inaccurate contextual information extraction and severe performance degradation. (b) We are the first to explore applying visual prompt learning in pixel-wise prediction DA. As shown in red boxes, we introduce Sparse Visual Domain Prompts (SVDP), which are tailored for addressing the occlusion problem of pixel-wise information and can better extract domain knowledge for cross-domain learning. Though the parameters of SVDP are less than VDP, SVDP achieves better semantic segmentation performance in the Test Time Adaptation.

which tactfully selects locations to place domain prompts. Specifically, it adaptively distributes several SVDP on more target domain-specific regions based on uncertainty guidance [11, 37]. In this way, SVDP can efficiently extract local domain knowledge and thus transfer pixel-wise data distribution from the source to the target domain. Furthermore, we design a Domain Prompt Updating (DPU) which can efficiently optimize prompt parameters to fit limited target data. In detail, we adopt an uncertainty scheme to measure the image-level domain shift between the source and the target domain. According to the degree of domain gap when facing the target domain samples, we adopt different weight Exponential Moving Average to update corresponding prompts. It thus efficiently updates SVDP on each target sample and realizes domain adaptation during test time. Note that, we are the first to design specific placement and optimizing strategies for vision prompt learning, which jointly address the domain shift in semantic segmentation.

We evaluate our method on semantic segmentation DA task. Since data privacy and transmission cost limit access to source domain data in many real-world scenarios, we thus conduct extensive experiments on the source-free adaptation settings, including six online Test Time Adaptation [26, 30] (TTA) and one Continual Test-Time Adaptation [48, 12] (CTTA). In particular, we apply our methods on Cityscapes [10], Foggy-Cityscapes [39], and ACDC [40] datasets. For instance, compared with the previous state-of-the-art (SOTA) methods, our method improves the mIoU by 1.5%, 4.1%, 3.1%, and 2.4% respectively in Cityscapes to ACDC semantic segmentation TTA (from Cityscapes to Fog, Night, Rain, and Snow domain respectively).

The main contributions are summarized as follows:

1) We make the first attempt to introduce the visual prompt approach to the semantic segmentation DA problem. Different from classification DA, Sparse Visual Domain Prompts (SVDP) are tailored for addressing the occlusion of pixel-wise information and can better extract domain knowledge for cross-domain learning.

2) In order to better apply SVDP in pixel-wise DA tasks, we propose Domain Prompt Placement (DPP) to adaptively distributes several SVDP on more target domain-specific regions for extracting more domain knowledge. And Domain Prompt Updating (DPU) is designed to efficiently optimize prompts for adapting to limited target domain data.

3) Our proposed approach outperforms the previous SOTA methods on six online TTA and one CTTA in semantic segmentation. It proves that our method is tailored for semantic segmentation cross-domain learning.

2. Related Work

2.1. Test-time adaptation

Test-time adaptation (TTA), also known as source-free domain adaptation [2, 26, 30, 53], aims to adapt a source model to an unknown target domain distribution without using any source domain data. In many real-world scenarios, data privacy and transmission cost limit access to source domain data, resulting in many traditional domain adaptation (DA) algorithms being inapplicable. Recent research has focused on using self-training or entropy regularization to fine-tune the source model [27, 47, 30, 6]. Specifically, Tent [47] updates the training parameters in the batch normalization layers by entropy minimization. SHOT [30] optimizes only the feature extractor using information maximization and pseudo labeling. AdaContrast [6] also uses pseudo labeling for TTA, but introduces self-supervised contrastive learning to enhance performance. In addition to model-level adaptation, [3] adjusts the output distribution to address this problem. While the aforementioned works primarily focus on classification tasks, there has been a recent surge of interest in performing TTA on dense prediction tasks [42, 43, 57]. And we make the first attempt to introduce sparse visual prompts in the semantic segmentation task, which aim to better extract domain knowledge for cross-domain learning.

Continual Test-Time Adaptation (CTTA) is a scenario in which the target domain is not static, increasing chal-

lenges for traditional TTA methods [48]. [48] serves as the first approach to tackle this task, using a combination of bi-average pseudo labels and stochastic weight reset. While [48] addresses the problem in both classification and segmentation tasks at the model level, [12] 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.

2.2. Prompt learning

Visual prompts are inspired by their counterparts [33], 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 [4]. Besides, the prompt has also been widely applied to vision-language models [24, 35, 55, 59, 60]. Recently, researchers have attempted to discard text encoders and use prompts directly for visual tasks. [1] employs visual prompts to pad input images, enabling pre-trained models to adapt to new tasks. Rather than fine-tuning the entire model, VPT [9, 23, 41, 50] inserts prompts into the image or feature-level patches to adapt Transformer-based models. While these approaches all utilize dense prompts, such prompts can cause performance degradation in dense prediction tasks. Therefore, we propose the use of sparse prompts for the first time to address semantic segmentation.

Domain prompts are first introduced in DAPL [16], 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 [7] proposes multi-prompt alignment for multi-source UDA. DePT [15] combines domain prompts with a hierarchical self-supervised regularization for TTA, which aims to solve the error accumulation problem in self-training. [12] 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 simple classification DA tasks. Our method, for the first time, applies sparse domain prompts to more complex dense prediction DA tasks. Besides, we are the first to design specific placement and updating strategies for the domain prompt method, which help to jointly ease the domain shift.

3. Method

3.1. Preliminaries

Test Time Adaptation (TTA) [12, 48]. TTA 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 unknown 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, [12] 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 ($\mathbf{p}$) are dense learnable parameters added to 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.

3.2. Motivation

Table 1. The comparison of Baseline [48] (without prompt), Dense Domain Prompt (DDP), and Sparse Visual Domain Prompt (SVDP) are conducted in the segmentation TTA (Cityscapes-to-ACDC-Rain). All experiment settings, except the usage of prompt, are the same among all methods. The designed placement and updating methods are not utilized. The metric is mIoU.

Region:	with prompt	without prompt	full image
Baseline	61.4	62.8	62.6
DDP	58.7	62.6	62.3
SVDP	64.1	64.8	64.7

Sparse Visual Domain Prompt. Traditional visual prompts [22] are deployed on the image or feature-level to realize fine-tuning by updating a small number of trainable prompt parameters. Recent works [12, 7, 14, 21] explore dense 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, these works randomly set the locations of domain prompts, masking out continuous spatial details in prompt occluded regions. Different from classification cross-domain learning, semantic segmentation 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 adopts sparse and discrete trainable parameters, thus reserving more spatial information in prompt regions. We further verify our motivation in Tab. 1, in which we place dense and sparse visual prompts in the same region (central of image). SVDP achieves better performance compared with other methods

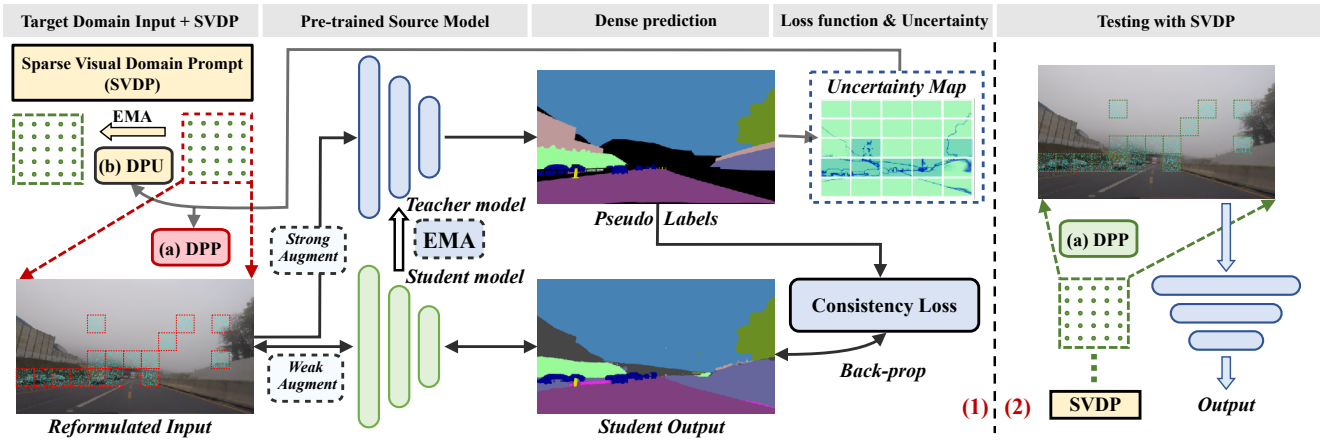

Figure 2. **The framework of Sparse Visual Domain Prompt (SVDP).** (1) **SVDP tuning**. We construct a teacher-student framework to update SVDPs. The SVDPs are placed on the image, which serves as the input of the teacher and student model through strong and weak augmentation. Our student network and prompt adopt consistency loss (Eq. 6) as the optimization objective. We obtain the uncertainty map as described in Eq. 2 through the teacher model. The uncertainty map is used to guide our SVDP placement (DPP in Sec .3.4) on regions with large data distribution distances. For SVDP parameter updating (DPU in Sec .3.4), we propose uncertainty-guided EMA for efficiently updating prompt parameters on limited target domain data. (2) **SVDP testing.** At test time, SVDPs are selectively placed onto the input image with the DPP strategy. Then we feed the reformulated image to the teacher model for semantic segmentation.

on regions with prompt. It shows that SVDP can address domain shift in the prompt region, which further proves SVDP empowers prompt to extract local domain knowledge. Besides, SVDP also achieves competitive results on the region without prompts, which proves that SVDP is of the same ability as traditional visual domain prompt [12] in dealing with global-level domain shift. In addition, along with introducing SVDP, we make the first attempt to design specific placement and updating strategies for SVDP to jointly address the domain shift problem.

Domain Prompt Placement. Previous work [12, 14] randomly put the prompts on the target domain image to extract global domain knowledge. Specifically, it may set prompts on regions with trivial domain shift, hindering the efficiency of cross-domain learning. 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 [11, 17, 38, 13] and tactfully place prompts on the regions with large distribution distance.

Domain Prompt Updating. The amount of prompt parameters is minimal, which brings the challenge of fully learning target domain knowledge. Meanwhile, the degree of domain shift is not only various on regions 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 an uncertainty scheme (same as DPP) 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 weight exponential moving average.

3.3. Sparse visual prompt

In this paper, we introduce a novel Sparse Visual Domain Prompts (SVDP), which alleviates inaccurate contextual information extraction and insufficient domain-specific feature caused by the previous dense visual prompt. In a prompt region of $\mathbf{p} \in \mathbb{R}^{H_P \times W_P \times 3}$, SVDP only masks out original information by minimal discrete trainable parameters (e.g. 10%) on randomly selected pixels. Compared with the previous visual prompt which masks out the whole prompt region, SVDP reverses more local information. The overall framework of our method is shown in Fig .2, and the specially designed prompt Placement and Updating methods are introduced in Sec.3.4 and Sec.3.5 respectively.

3.4. Domain prompt placing

In this section, we propose the Domain Prompt Placement (DPP) strategy of SVDP to efficiently extract more local target domain knowledge. We intend to place SVDP on regions with relatively severe domain shift and better adapt pixel-wise data distribution from source to target domain. Though the confidence score is a straightforward measurement to reflect the reliability of prediction, it is trusting less and fluctuates irregularly in pixel-wise cross-domain scenarios. As shown in the top of Fig .3, we thus adopt Dropout method [11] to realize m times forward propagation and obtain m group probabilities for each pixel. Inspired by [38, 13], we calculate the uncertainty value (Eq.2) of the

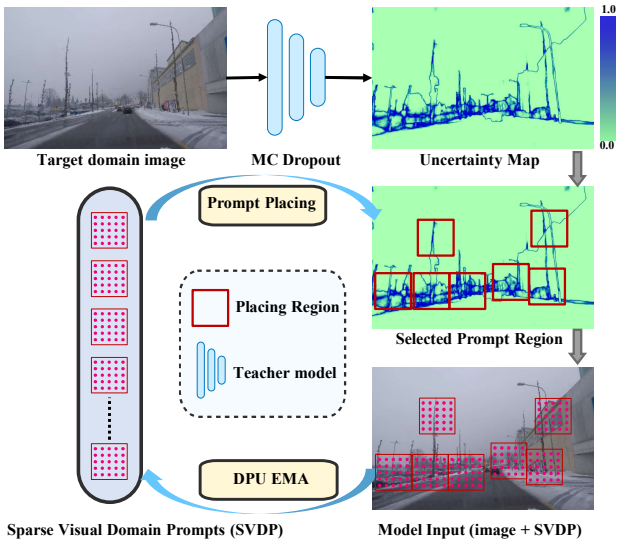

Figure 3. The detailed process of Domain Prompt Placing. The uncertainty map is estimated by MC Dropout [11].

input and figure out the degree of domain shift pixel-wise.

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

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 $\widetilde{x}_j$.

We then split the whole input image ($x \in \mathbb{R}^{H \times W \times 3}$) into $n_H \times n_W$ regions to measure the average uncertainty value of each region, n_H and n_W equal H/H_p and W/W_p respectively. As shown in Eq.3, the uncertainty of a region $\mathcal{U}(x_r)$ is the average value of each pixel uncertainty within the region ($H_p \times W_p$).

$$\mathcal{U}(x_r) = \frac{1}{H_p \times W_p} \sum_{j}^{H_P \times W_P} \mathcal{U}(\widetilde{x}_j) \tag{3}$$

As shown in the bottom of Fig .3, we sort all regions based on their region uncertainty value and place SVDP on the region of high uncertainty score, which represents the large domain shift. In this way, SVDP can extract target domain knowledge efficiently.

3.5. Domain prompt updating

For updating, we adopt the widely-used teacher-student framework and exponential moving average (EMA) to achieve the model and prompt updating [12]. Same as previous works[48], the teacher model ($\mathcal{T}_{mean}$) is updated by EMA from the student model ($\mathcal{S}_{target}$), shown in Eq. 4:

$$\mathcal{T}^t_{mean} = \alpha \mathcal{T}^{t-1}_{mean} + (1-\alpha)\mathcal{S}^t_{target} \tag{4}$$

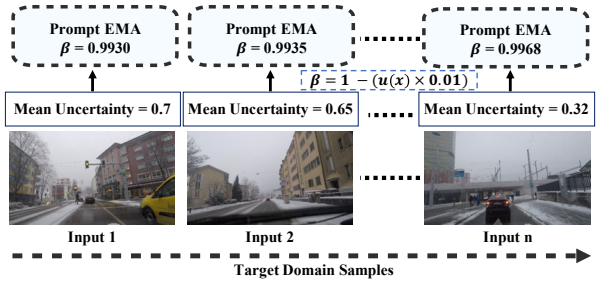

Figure 4. The process of Domain Prompt Updating. We adaptively adjust the prompt EMA updating rate for each target domain sample based on image-level uncertainty value.

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 $\alpha = 0.999$ [44], which is the updating weight of EMA.

Different from traditional model updating, we design a special Domain Prompt Updating (DPU) strategy for SVDP to efficiently 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. Similar to Eq .3, we calculate the average uncertainty value ($\mathcal{U}(x)$) for each pixel in the entire image. According to the image-level uncertainty score, we update prompt parameters for the individual sample with different weight EMA.

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

Note that, p_t represents the parameters of the SVDP that need to be updated. In DPU, we set the prompt EMA updating rate $\beta = 1 - (\mathcal{U}(x) \times 0.01)$. 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. In this way, we can improve the efficiency of target domain adaptation during test time.

3.6. Loss function

Following previous TTA work [48], we utilize the teacher model to generate the pseudo labels ($\widetilde{y}_t$), which is refined by test-time augmentation and confidence filter. Then, we adopt consistency loss (L_{con}) as the optimization objective for SVDP, which is a pixel-wise cross-entropy loss [52].

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

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

Table 2. **Performance comparison of Cityscapes-to-ACDC TTA.** We use Cityscape as the source domain and ACDC as the four target domains in this setting. Source domain data is unavailable, while target domain data is only accessed once during testing.

Test-Time Adaptation		Source2Fog		Source2Night		Source2Rain		Source2Snow		Mean-mIoU
Method	REF	mIoU↑	mAcc↑	mIoU↑	mAcc↑	mIoU↑	mAcc↑	mIoU↑	mAcc↑	
Source	NIPS2021 [52]	69.1	79.4	40.3	55.6	59.7	74.4	57.8	69.9	56.7
TENT	ICLR2021 [46]	69.0	79.5	40.3	55.5	59.9	74.1	57.7	69.7	56.7
CoTTA	CVPR2022[48]	70.9	80.2	41.2	55.5	62.6	75.4	59.8	70.7	58.6
DePT	ICLR2023[14]	71.0	80.2	40.9	55.8	61.3	74.4	59.5	70.0	58.2
VDP	AAAI2023[12]	70.9	80.3	41.2	55.6	62.3	75.5	59.7	70.7	58.5
SVDP	**ours**	**72.5**	**81.4**	**45.3**	**58.9**	**65.7**	**76.7**	**62.2**	**72.4**	**61.4**

4. Experiments

In this section, we conduct extensive experiments to demonstrate the effectiveness of our proposed SVDP for semantic segmentation Domain Adaptation (DA) tasks. Since data privacy and transmission cost limit access to source domain data in many real-world scenarios, we thus conduct extensive semantic segmentation experiments on Test Time Adaptation (TTA) and Continual Test-Time Adaptation (CTTA). In Sec 4.1, we provide the details of the task settings for TTA and CTTA, as well as a description of the datasets. In Sec 4.2, we compare our method with other baselines [52, 46, 48, 14, 14] in six TTA and one CTTA scenarios. Comprehensive ablation studies are conducted in Sec 4.3, which investigate the impact of each component. Besides, we provide qualitative analysis in Appendix A.1.

4.1. Task settings and Datasets

TTA and CTTA are commonly used technology in real-world scenarios in which a pre-trained model adapts to the distribution of an unseen target domain. Both scenarios can only adopt the source domain pre-trained model and can not access source domain data. In the TTA, the data in each target domain is unlabeled and can only be accessed once, which makes the efficiency of domain adapting crucial. Meanwhile, CTTA is of the same setting as TTA but further sets the target domain constantly changing, bringing more difficulties during test time adaptation.

Cityscapes-to-ACDC is a semantic segmentation task designed for 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 [10]. The ACDC dataset [40] contains images collected in four different unseen visual conditions: Fog, Night, Rain, and Snow. For the TTA, we adapt the source domains' 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-life scenarios [48, 12].

Cityscapes-to-Foggy&Rainy Cityscapes. To demonstrate the generalization of our method, we conducted experiments in this scenario, which is a commonly used benchmark for semantic segmentation TTA scenarios [56, 49]. In comparison to the Foggy scenario in ACDC, Foggy Cityscapes [39] has a larger dataset and a higher density of the simulated fog, which is a more challenging TTA scene. Besides, we also evaluate the effectiveness of our method on Cityscapes-to-Rainy Cityscapes [39] TTA scenario.

Implementation Details. We follow the basic implementation details [48, 12] to set up our semantic segmentation TTA experiments. Specifically, we use the Segformer-B5 model [52] pre-trained on Cityscapes datasets 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 the Adam optimizer [25] $(\beta_1, \beta_2) = (0.9, 0.999)$ with a learning rate of 3e-4 and batch size 1 for both TTA and CTTA experiments. We apply horizontal-flip in both the strong and weak augmentation strategies [52]. Additionally, we use a range of image resolution scale factors [0.5, 0.75, 1.0, 1.25, 1.5, 1.75, 2.0] for our strong augmentation method. Besides, we set the prompt size to 30×30 and the prompt number to 25. All experiments are conducted on NVIDIA V100 GPUs.

4.2. The effectiveness

We show quantitative comparisons between our method and the baselines on six TTA and one CTTA segmentation scenarios, which are measured by mean Intersection over Union (mIoU) and mean Accuracy (mAcc). **For baselines,** we compare our method with the Segformer [52], TENT [46], CoTTA [48], DePT [14], and VDP [14], which are cutting-edge approaches in related studies.

Cityscapes-to-ACDC TTA. We evaluate the performance of the proposed SVDP on four scenarios with significant domain gap during TTA. Tab .2 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 2.8%. In the night domain with the largest domain gap, our method increased mIoU by 4.1% and mAcc by 3.3%, compared with the previous State-Of-The-Art (SOTA) method (VDP). These results demonstrate that our method can better address the domain shit problem in test time compared to other methods. Additionally, compared to the VDP that also utilizes domain prompts, our method can avoid the occlusion problem and better extract local target domain knowl-

Table 3. **Performance comparison for Cityscape-to-ACDC CTTA.** We take the Cityscape as the source domain and ACDC as the four target domains. During testing, we sequentially evaluate the four target domains multiple times. Mean is the average score of mIoU for all times. Gain refers to the improvement achieved by the method compared to the Source model.

Time		t ⟶																
Round		1					2					3					Mean↑	Gain
Method	REF	Fog	Night	Rain	Snow	Mean↑	Fog	Night	Rain	Snow	Mean↑	Fog	Night	Rain	Snow	Mean↑		
Source	NIPS2021 [52]	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	ICLR2021 [46]	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	CVPR2022[48]	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	ICLR2023[14]	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	AAAI2023[12]	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	**ours**	**72.5**	**45.9**	**67.0**	**64.1**	**62.4**	**72.2**	**44.8**	**67.3**	**64.1**	**62.1**	**72.0**	**44.5**	**67.6**	**64.2**	**62.1**	**62.2**	**+5.5**

Table 4. **Performance comparison for Cityscape-to-(Foggy&Rainy) Cityscape TTA.**

Foggy	Source	TENT	CoTTA	DePT	VDP	SVDP
mIoU	69.2	69.3	72.1	71.9	71.8	**74.5**
mAcc	79.1	79.0	79.4	80.2	80.0	**82.3**
Rainy	Source	TENT	CoTTA	DePT	VDP	SVDP
mIoU	58.4	58.7	62.4	61.2	63.7	**65.3**
mAcc	71.5	71.4	74.0	72.4	73.1	**75.6**

edge for semantic segmentation TTA.

Cityscapes-to-(Foggy&Rainy) Cityscapes TTA. To further demonstrate the effectiveness of our method, we evaluate the performance of SVDP on the CityScapes-to-Foggy Cityscapes benchmark. As illustrated in Tab. 4, our approach achieved a 2.4% higher mIoU than the previous SOTA model (CoTTA). We observe that the results show the same trend as the above TTA experiments. In the CityScapes-to-Rainy Cityscapes TTA benchmark, compared with CoTTA, our SVDP improved mIoU and mAcc by 2.9% and 1.6%. These results prove that our method is generalized for addressing different domain shifts in the semantic segmentation TTA setting.

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 [48]. As shown in Tab .3, due to error accumulation and 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 found that our method gains 3.6% increase of mIoU more than the previous SOTA CTTA method [48]. The results prove that our method shows the ability to avoid error accumulation and catastrophic forgetting phenomena in semantic segmentation CTTA problems. Specifically, our method introduces Domain Prompt Placement (DPP) scheme to only extract more target domain-specific knowledge and adopt the Domain Prompt Updating (DPU) method to adaptively preserve the original knowledge of the source domain by dynamic adjustment of the prompt updating weight.

Table 5. **Ablation: Contribution of each component.**

	TS	SVDP	DPP	DPU	mIoU↑	mAcc↑
Ex_1					40.3	55.6
Ex_2	✓				41.2	55.5
Ex_3	✓	✓			43.7	56.9
Ex_4	✓	✓	✓		44.5	58.1
Ex_5	✓	✓		✓	44.4	58.2
Ex_6	✓	✓	✓	✓	45.3	58.9

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

4.3. Ablation study

In this subsection, we evaluate the contribution of each component in our method. Since the Night domain is the most challenging scenario for camera-based methods, we conduct the ablation study on the Cityscapes-to-ACDC Night TTA. Due to the limitation of space, we show the ablation study on other scenarios in Appendix A.2.

Effectiveness of each component. As presented in Tab. 7 Ex_2, Teacher-student (TS) structure is a common technique in TTA [32, 48, 12], which is used to generate pseudo label in the target domain and only has 0.9% mIoU improvement without our method. This verifies 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 mIoU and mAcc increase 2.5% and 1.4%, 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 0.8% mIoU and 1.2% mAcc improvement since the specially designed prompt placement strategy can assist SVDP in extracting more target domain-specific knowledge. Compared with Ex_3, DPU (Ex_5) also improves the mIoU and mAcc by 0.7% and 1.3% respectively. The results prove the effectiveness of DPU and

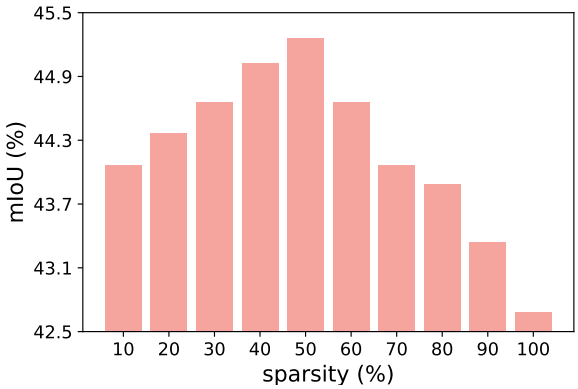

Figure 5. Effect of prompts' sparsity

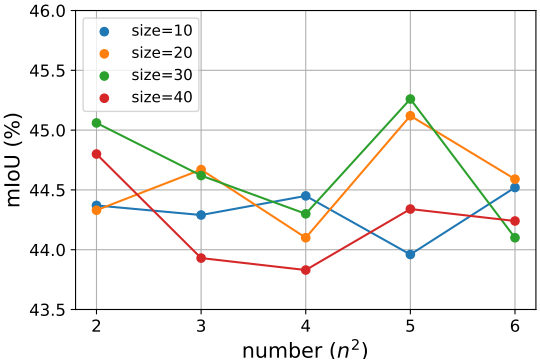

Figure 6. Effect of prompts' size and number

show the importance of adaptively optimizing for different samples during testing. Ex_6 shows the complete combination of all components which achieves 5.0% mIoU improvement in total. It proves that all components compensate each other and jointly address the semantic segmentation domain shift problem in test time.

How does the prompt sparsity affect the performance? As shown in Fig .5, we investigate the performance impact caused by the sparsity of SVDP. Specifically, we gradually increase the density of SVDP pixel-wise parameters and record the corresponding mIoU values. We find that mIoU initially improves along with increasing SVDP density and then starts to decrease when the density exceeds 60%. This observation suggests that when SVDP is 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 continuous spatial details, leading to segmentation 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 50% sparsity.

How do the prompt size and number influence the performance? According to Fig .6, we observe that changing the prompt size has a small effect on performance, with a small variance of 0.2%∼1.0% mIoU. Experimental results presented in [23] also support this finding that the prompt size is not sensitive. This observation provides the opportunity for us to deal with different input sizes with different prompt sizes. Regarding the prompt number, the results show that using 25 prompts can achieve the highest segmentation performance. This suggests that when the number of prompts is too small, the available parameters may not be sufficient to fit in the target domain knowledge. However, when the number of prompts reaches a certain value, increasing the number of prompts may not lead to a significant performance improvement, and may even result in some performance degradation due to occlusion.

How do the prompt placement strategies influence the performance? According to Tab .6, we compare the performance of several commonly used prompt placement methods with our DPP (Domain Prompt Placement). Center and Corners stand for placing sparse prompts on the center and corner of the input image, while the Grid represents that sparse prompts are distributed uniformly on the image with equal distance. Different from DPP, CPP leverage the confidence score to reflect the degree of domain shift and select the prompt placement regions. As we can see, the mIoU of CPP and DPP are obviously higher than other methods, which demonstrate the SVDP should be set on the regions with large domain shift. Meanwhile, compared with CPP, DPP further improves 0.8% mIoU and 0.7% mAcc since the uncertainty scheme is more suitable for measuring domain shift problems in semantic segmentation task. Though the confidence score is a straightforward measurement to reflect the reliability of prediction, it is less trustworthy and fluctuates irregularly in pixel-wise cross-domain scenarios. In contrast, the uncertainty value is relatively more reliable in reflecting the degree of domain shift and more reasonable in guiding the domain prompt placement.

Table 6. **Effects of prompts' placement on the TTA task.**

	Center	Corners	Grid	Random	CPP	DPP
mIoU↑	43.5	43.8	44.1	44.1	44.5	45.3
mAcc↑	56.9	56.6	57.4	57.7	58.2	58.9

5. Conclusion and discussion of limitations

In this paper, we are the first to introduce the Sparse Visual Domain Prompt (SVDP) in semantic segmentation DA tasks, 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 domain shift better. Extensive experiments on multiple TTA and CTTA scenarios

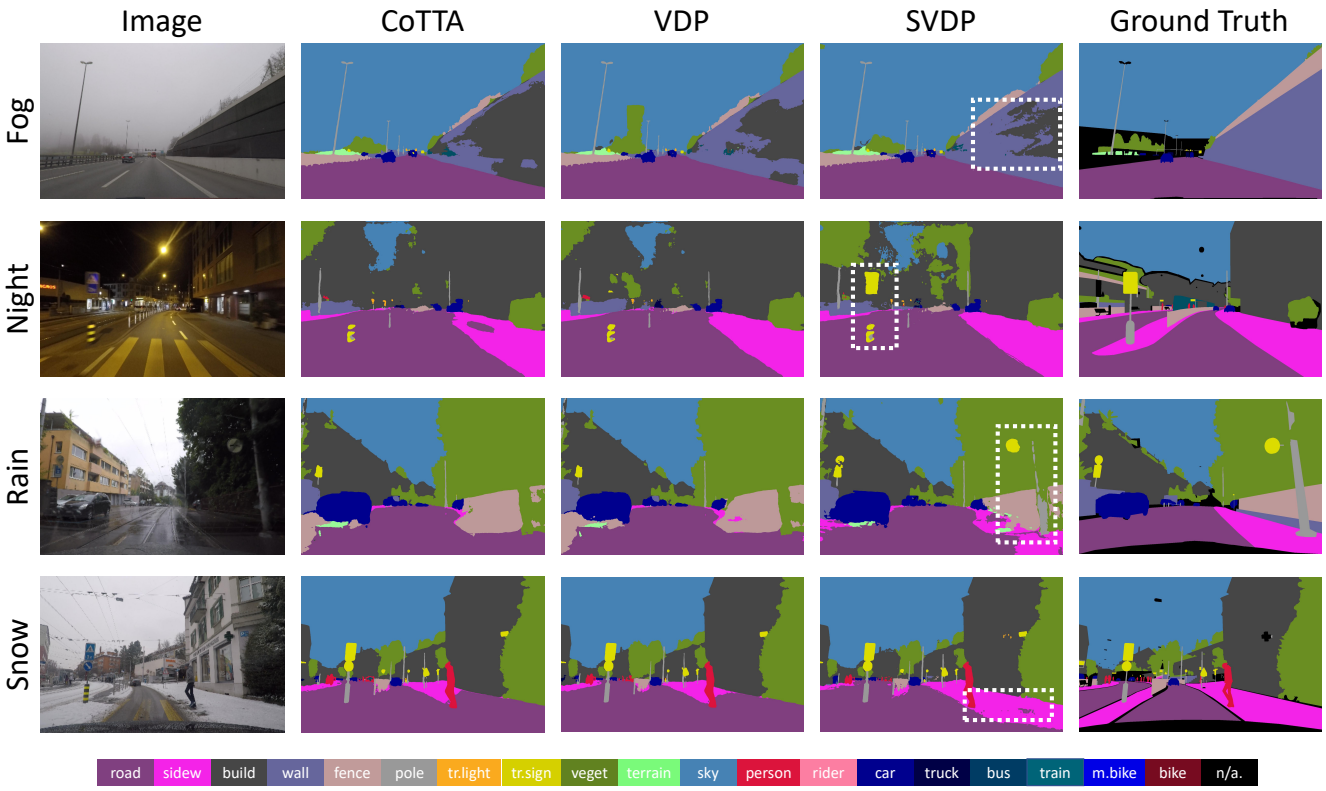

Figure 7. Qualitative comparison of SVDP with previous SOTA method: CoTTA[48], VDP[12] on ACDC Fog, Night, Rain, and Snow four scenarios. SVDP could better segment different pixel-wise classes such as shown in the white box.

demonstrate that our method achieves SOTA performance and efficiently tackles the domain shift. For limitations, the teacher-student framework brings more computational costs during SVDP tuning. However, the forward time and computational costs are the same as the baseline in testing.

Supplementary Material

The supplementary material presented in this paper offers a comprehensive analysis and additional experimental results of Sparse Visual Domain Prompts (SVDP). Specifically, Sec. A.1 presents the qualitative analysis and comparison results of the SVDP. Sec. A.2 provides additional ablation studies of our method in the Continual Test-Time Adaptation (CTTA) scenario. Sec. A.3 investigates the parameter sensitivity of Domain Prompt Updating (DPU). In addition, Sec. A.4 provides more quantitative results of the TTA and CTTA scenarios. Furthermore, in Sec. B, we extend our method to the Unsupervised Domain Adaptation (UDA) scenario, demonstrating its strong generalization capabilities and plug-and-play characteristics. Finally, in Sec. C, we present related work on semantic segmentation. The code will be released.

A. Experiment results and analysis

A.1. Qualitative analysis

To further demonstrate the effectiveness of our proposed method, SVDP, we conduct a qualitative comparison with two current leading methods, CoTTA [48] and VDP [12], on the CTTA scenario (Cityscapes-to-ACDC).

The results of the comparison are presented in Fig .7. In the foggy target-domain, we highlight a white box that contains a tall and dark wall object. This object is difficult to segment as it shares characteristics with the *building* class. Our proposed method, SVDP, has a significant advantage in dealing with such confusing semantic segmentation categories with high uncertainty, thanks to the contribution of the Domain Prompt Placement (DPP) method. Our method also outperforms CoTTA and VDP in the remaining three domains. In the night and rain target-domains, due to blurring and occlusions, CoTTA and VDP fail to identify the traffic sign, while our method successfully segments it. In the snow domain, our proposed method correctly distinguish the sidewalk from the road, avoiding misclassification. Overall, our method can achieve better local segmentation results and neglect the influence of local domain shift. And our method produces finer results than the previous state-of-the-art methods, with clear visual improvement.

Table 7. **Ablation: Contribution of each component (CTTA).**

	TS	SVDP	DPP	DPU	mIoU↑	mAcc↑
Ex_1					56.7	68.8
Ex_2	✓				58.5	70.4
Ex_3	✓	✓			60.5	71.8
Ex_4	✓	✓	✓		61.4	72.3
Ex_5	✓	✓		✓	61.3	72.4
Ex_6	✓	✓	✓	✓	62.4	73.0

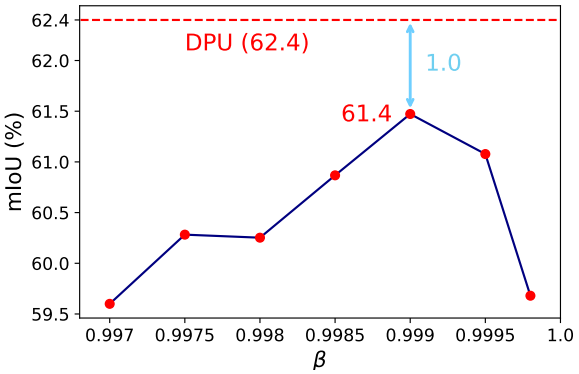

Figure 8. Sensitivity Analysis: The effect of prompt EMA's parameter β on semantic segmentation performance in the CTTA scenario.

A.2. Additional ablation studies

The proposed method comprises a Sparse Visual Prompt (SVDP), a Domain Prompt Placement (DPP) strategy, and a Domain Prompt Updating (DPU) strategy to mitigate domain shifts in semantic segmentation tasks. In this study, we conduct ablation experiments on the most challenging CTTA scenario (Cityscapes-to-ACDC) to evaluate the effectiveness of each component.

To compare the performance of our method with and without using the teacher-student (TS) structure, a common technique in CTTA used to generate pseudo labels in the target domain, we present the results in Tab. 7 Ex_2. The results show that without our method, TS only has a 0.8% mIoU improvement, indicating that our method's improvement does not come from the usage of this prevalent scheme. In Ex_3, we introduce SVDP to extract local target domain knowledge without damaging the original semantic information. The results demonstrate that SVDP achieves a 2.0% mIoU and 1.4% mAcc improvement, effectively addressing the domain shift problem. In Ex_4, DPP achieves a further 0.9% mIoU and 0.5% mAcc improvement by serving as a specially designed prompt placement strategy to assist SVDP in extracting more target domain-specific knowledge. We evaluate the effectiveness of the DPU in Ex_5, which adaptively optimizes for different samples during testing. The results show that DPU improves the mIoU and mAcc by 0.8% and 0.6%, respectively. Finally, in Ex_6, we show the complete combination of all components, which achieves a total of 5.7% mIoU improvement. These results demonstrate that all components of our method effectively address the semantic segmentation domain shift and compensate for each other to achieve superior performance.

A.3. Additional sensitivity analysis

In Sec 3.5, we utilize Eq.5 to update the prompt parameters and conduct an analysis of the sensitivity of the parameter β in the CTTA scenario. As depicted in Fig. 8, we investigate the impact of β values on the performance. Specifically, we gradually increase the value of β and record the corresponding mIoU values. We observe that the mIoU improves with increasing β; however, it starts to decrease once β exceeds 0.999. Compare with the best fixed β value, our proposed DPU (red line) strategy can further achieve 1.0% mIoU improvement. Therefore, due to the different degrees of domain shift, we need to update prompt parameters for the each sample with different EMA weights.

A.4. Additional quantitative results

We present a comprehensive presentation of experimental results on the Test-time adaptation task for Cityscapes-to-ACDC, as shown in Tab . 8 - Tab . 11. Our findings suggest that our proposed approach can better address the domain shift problem and achieve better IoU value in most categories.

B. SVDP on Unsupervised Domain Adaptation

To demonstrate the effectiveness and generalizability of our approach, we apply SVDP in Unsupervised Domain Adaptation (UDA) to evaluate its performance. Our proposed SVDP has a relatively small number of parameters, which does not significantly increase computational cost and can be easily integrated into existing models by directly adding SVDP to the input. Therefore, SVDP serves as a plug-and-play method for any UDA methods. DAFormer [18] is a widely used method in recent UDA semantic segmentation tasks, and many state-of-the-art approaches [5, 19, 20] build upon its foundation for further improvement. To demonstrate the effectiveness of our proposed SVDP, we integrate SVDP into DAFormer with DPP strategy and DPU strategy, which aims to evaluate the UDA segmentation performance of our methods.

In our experiment, we utilize the same model and training strategy as DAFormer, with the addition of SVDP to the original model. The learning rate of the Prompt is set to 0.006. In the UDA setting, we can access labeled source domain data and unlabeled target domain data during training, and the model is tested on the target domain during

Table 8. Performance Comparison for **Cityscapes-to-ACDC Fog domain in TTA scenario**. The IoU score of each class and the mIoU score are reported. The bests results are highlighted in **bold**.

Method	road	side.	buil.	wall	fence	pole	light	sign	veg.	terr.	sky	pers.	rider	car	truck	bus	train	mbike	bike	mIoU
Source	94.0	63.9	79.8	55.7	24.9	45.0	41.5	69.8	86.6	71.0	97.6	64.1	66.2	87.4	73.0	92.6	87.7	50.2	61.7	69.1
TENT	94.0	64.0	79.7	55.4	24.6	44.6	41.4	69.9	86.7	71.1	97.6	64.0	65.9	87.4	73.0	92.6	88.0	50.2	61.9	69.0
CoTTA	93.9	63.6	**80.0**	55.5	25.1	49.0	43.4	73.0	87.0	70.7	97.8	68.6	71.3	87.1	74.8	**93.6**	89.1	58.0	66.7	70.9
DePT	94.0	64.0	79.9	56.1	25.3	48.8	43.5	73.0	87.1	70.6	97.5	67.9	71.5	87.3	75.1	93.5	89.1	57.4	66.5	71.0
VDP	93.9	63.6	**80.0**	55.6	25.1	49.0	43.4	73.0	86.9	70.7	97.7	68.5	71.1	87.2	74.7	93.5	89.2	57.9	66.6	70.9
SVDP	**94.8**	**68.7**	79.6	**59.9**	**25.8**	**50.7**	**44.3**	**73.5**	**87.3**	**71.1**	**97.8**	**70.1**	**72.9**	**87.3**	**76.5**	93.2	**91.2**	**65.4**	**68.6**	**72.6**

Table 9. Performance Comparison for **Cityscapes-to-ACDC Night domain in TTA scenario**. The IoU score of each class and the mIoU score are reported. The bests results are highlighted in **bold**.

Method	road	side.	buil.	wall	fence	pole	light	sign	veg.	terr.	sky	pers.	rider	car	truck	bus	train	mbike	bike	mIoU
Source	87.6	46.3	61.8	27.0	25.3	40.8	38.7	39.4	47.7	26.8	**11.4**	48.6	39.9	76.1	15.9	24.2	52.0	26.5	29.6	40.3
TENT	87.7	46.4	61.9	27.1	25.2	40.8	38.8	39.3	47.0	26.8	9.6	48.7	40.0	76.2	16.1	24.3	51.9	26.6	29.7	40.2
CoTTA	87.6	46.7	**62.3**	27.2	25.0	44.0	42.9	40.8	47.2	26.7	8.8	51.8	41.9	76.6	18.8	22.4	51.7	**27.8**	32.1	41.2
DePT	87.3	46.5	62.0	27.0	25.3	43.5	40.9	41.0	47.2	26.6	8.8	51.0	42.5	77.1	17.5	23.0	51.5	26.4	31.7	40.9
VDP	87.6	46.8	62.2	27.1	25.0	44.0	42.9	41.0	47.3	26.6	9.0	51.7	41.9	76.6	18.7	23.2	51.9	**27.8**	32.0	41.2
SVDP	**91.9**	**64.7**	51.4	**31.2**	**25.9**	**50.2**	**47.1**	**46.4**	**58.8**	**28.3**	0.4	**55.0**	**45.7**	**79.3**	**23.0**	**26.6**	**70.2**	27.7	**36.2**	**45.3**

Table 10. Performance Comparison for **Cityscapes-to-ACDC Rain domain in TTA scenario**. The IoU score of each class and the mIoU score are reported. The bests results are highlighted in **bold**.

Method	road	side.	buil.	wall	fence	pole	light	sign	veg.	terr.	sky	pers.	rider	car	truck	bus	train	mbike	bike	mIoU
Source	82.3	47.1	89.5	36.8	26.6	51.0	64.8	62.9	89.5	60.3	97.8	46.0	53.0	81.1	25.3	65.4	56.7	47.6	51.2	59.7
TENT	82.4	46.7	89.6	37.3	27.0	50.6	64.6	62.9	89.5	60.4	97.7	46.7	54.5	81.2	25.4	65.2	56.6	47.3	51.8	59.9
CoTTA	83.0	48.4	90.2	38.3	28.0	55.5	68.1	67.5	90.3	61.2	98.0	54.1	60.1	82.0	27.4	67.0	59.1	50.9	55.4	62.6
DePT	82.0	47.3	89.8	37.5	26.9	53.0	66.2	65.8	89.6	60.8	97.7	52.8	59.5	81.6	26.5	66.5	57.9	49.5	53.2	61.3
VDP	83.0	48.3	90.2	38.2	28.0	55.5	68.2	67.5	90.2	61.2	97.9	54.1	59.9	82.0	**27.5**	67.0	59.1	50.9	55.3	62.3
SVDP	**85.4**	**55.3**	**91.3**	**43.4**	**30.9**	**58.7**	**70.3**	**70.4**	**91.1**	**64.0**	**98.2**	**55.9**	**60.9**	**82.1**	25.6	**79.4**	**73.3**	**54.6**	**59.1**	**65.7**

Table 11. Performance Comparison for **Cityscapes-to-ACDC Snow domain in TTA scenario**. The IoU score of each class and the mIoU score are reported. The bests results are highlighted in **bold**.

Method	road	side.	buil.	wall	fence	pole	light	sign	veg.	terr.	sky	pers.	rider	car	truck	bus	train	mbike	bike	mIoU
Source	79.8	40.8	86.9	43.6	46.5	56.4	72.3	65.5	82.9	**5.7**	97.1	62.8	40.4	85.4	54.7	44.5	73.1	22.7	36.0	57.8
TENT	79.6	40.0	86.8	43.4	46.5	56.1	72.2	65.6	82.9	**5.7**	97.1	63.0	40.9	85.5	54.7	43.1	72.7	23.0	36.5	57.7
CoTTA	80.1	40.7	87.5	43.9	47.7	59.9	75.3	69.2	84.0	5.1	97.2	67.3	46.9	86.2	56.1	43.4	74.1	**25.7**	43.3	59.8
DePT	79.1	40.6	86.8	43.4	47.5	59.8	75.1	69.4	83.5	5.2	97.1	67.2	46.5	86.3	56.0	44.0	73.9	25.6	43.1	59.5
VDP	80.1	40.8	87.5	43.9	47.8	59.9	75.1	69.4	83.9	5.1	97.2	67.2	46.7	86.2	56.2	43.9	74.0	25.7	42.9	59.7
SVDP	**81.9**	**44.8**	**89.3**	**52.8**	**50.7**	**62.8**	**76.5**	**71.2**	**85.2**	4.0	**97.7**	**68.0**	**47.3**	**87.0**	**59.1**	**53.8**	**74.9**	25.1	**49.5**	**62.2**

Table 12. Performance Comparison for **Cityscapes-to-ACDC UDA**. We follow the setting of DAFormer[18], taking the Cityscape as the source domain and the entire ACDC's dataset as the target domain. The IoU score of each class and the mIoU score are reported on the ACDC validation set. The bests results are highlighted in **bold**.

Method	road	side.	buil.	wall	fence	pole	light	sign	veg.	terr.	sky	pers.	rider	car	truck	bus	train	mbike	bike	mIoU
DAFormer	71.1	**53.8**	77.8	**45.1**	**36.6**	57.1	47.0	**53.4**	70.3	36.6	67.0	53.9	22.3	81.1	67.6	79.8	85.6	37.2	**41.6**	57.1
DAFormer+SVDP	**79.9**	46.5	**78.7**	44.1	35.6	**57.6**	**48.3**	52.1	**71.7**	**37.1**	**78.6**	**56.9**	**29.3**	**83.1**	**71.2**	**85.9**	**86.2**	**37.3**	39.2	**58.9**

the inferrence. We evaluate our model in the CityScape-to-ACDC scenario, and the results in Tab . 12 demonstrate that our method achieves 1.8% mIoU improvement. This finding further confirms the effectiveness of our method, as well as emphasizes its generalizability in pixel-wise DA.

C. Additional related work

Semantic segmentation is a crucial task in many computer vision applications aimed at assigning a categorical label to every pixel in an image. Several representative works in this field include DeepLab [8], PSPNet [58], RefineNet [31], and Segformer [52]. Despite their high performance, these methods usually require extensive amounts of pixel-level annotated data, which can be laborious and time-consuming to collect. Additionally, they may suffer from poor generalization when applied to new domains. Recent research has focused on addressing these challenges through domain adaptation strategies. For instance, [54] proposes a method that swaps the low-frequency spectrum to align the source and target domains. [45] mixes the images from both domains, along with their corresponding la-

bels and pseudo-labels. In contrast, [51] uses adversarial learning to train a domain adaptation network for nighttime semantic segmentation. [18] develops a novel model and training strategies to enhance training stability and avoid overfitting to the source domain. However, these methods often require retraining the model on the source domain, which is inconvenient. Furthermore, they need to be retrained when adapting to a new target domain, incurring additional time and resource costs. Therefore, we propose SVDP to efficiently address the domain shift problem, which leverages a pre-trained model on the source domain and adds only a few parameters to achieve strong generalization capabilities on the target domain.

References

[1] Hyojin Bahng, Ali Jahanian, Swami Sankaranarayanan, and Phillip Isola. Exploring visual prompts for adapting large-scale models. 2022. 3

[2] Malik Boudiaf, Tom Denton, Bart van Merriënboer, Vincent Dumoulin, and Eleni Triantafillou. In search for a generalizable method for source free domain adaptation. 2023. 2

[3] Malik Boudiaf, Romain Mueller, Ismail Ben Ayed, and Luca Bertinetto. Parameter-free online test-time adaptation. *ArXiv*, abs/2201.05718, 2022. 2

[4] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners, 2020. 3

[5] David Brüggemann, Christos Sakaridis, Prune Truong, and Luc Van Gool. Refign: Align and refine for adaptation of semantic segmentation to adverse conditions. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 3174–3184, 2023. 10

[6] Dian Chen, Dequan Wang, Trevor Darrell, and Sayna Ebrahimi. Contrastive test-time adaptation. *ArXiv*, abs/2204.10377, 2022. 2

[7] Haoran Chen, Zuxuan Wu, and Yu-Gang Jiang. Multi-prompt alignment for multi-source unsupervised domain adaptation. *arXiv preprint arXiv:2209.15210*, 2022. 1, 3

[8] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE transactions on pattern analysis and machine intelligence*, 40(4):834–848, 2017. 11

[9] Jonathan Conder, Josephine Jefferson, Nathan Pages, Khurram Jawed, Alireza Nejati, and Mark Sagar. Efficient transfer learning for visual tasks via continuous optimization of prompts. 2022. 3

[10] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3213–3223, 2016. 2, 6

[11] Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning*, pages 1050–1059. PMLR, 2016. 2, 4, 5

[12] 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, 2, 3, 4, 5, 6, 7, 9

[13] Yulu Gan, Mingjie Pan, Rongyu Zhang, Zijian Ling, Lingran Zhao, Jiaming Liu, and Shanghang Zhang. Cloud-device collaborative adaptation to continual changing environments in the real-world. *arXiv preprint arXiv:2212.00972*, 2022. 4

[14] Yunhe Gao, Xingjian Shi, Yi Zhu, Hao Wang, Zhiqiang Tang, Xiong Zhou, Mu Li, and Dimitris N Metaxas. Visual prompt tuning for test-time domain adaptation. *arXiv preprint arXiv:2210.04831*, 2022. 1, 3, 4, 6, 7

[15] Yunhe Gao, Xingjian Shi, Yi Zhu, Hao Wang, Zhiqiang Tang, Xiong Zhou, Mu Li, and Dimitris N. Metaxas. Visual prompt tuning for test-time domain adaptation, 2022. 3

[16] Chunjiang Ge, Rui Huang, Mixue Xie, Zihang Lai, Shiji Song, Shuang Li, and Gao Huang. Domain adaptation via prompt learning. *ArXiv*, abs/2202.06687, 2022. 3

[17] Dayan Guan, Jiaxing Huang, Aoran Xiao, Shijian Lu, and Yanpeng Cao. Uncertainty-aware unsupervised domain adaptation in object detection. *IEEE Transactions on Multimedia*, 24:2502–2514, 2021. 4

[18] Lukas Hoyer, Dengxin Dai, and Luc Van Gool. Daformer: Improving network architectures and training strategies for domain-adaptive semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9924–9935, 2022. 10, 11, 12

[19] Lukas Hoyer, Dengxin Dai, and Luc Van Gool. Hrda: Context-aware high-resolution domain-adaptive semantic segmentation. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXX*, pages 372–391. Springer, 2022. 10

[20] Lukas Hoyer, Dengxin Dai, Haoran Wang, and Luc Van Gool. Mic: Masked image consistency for context-enhanced domain adaptation. *arXiv preprint arXiv:2212.01322*, 2022. 10

[21] Shishuai Hu, Zehui Liao, and Yong Xia. Prosfda: Prompt learning based source-free domain adaptation for medical image segmentation. *arXiv preprint arXiv:2211.11514*, 2022. 1, 3

[22] Menglin Jia, Luming Tang, Bor-Chun Chen, Claire Cardie, Serge Belongie, Bharath Hariharan, and Ser-Nam Lim. Visual prompt tuning. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXIII*, pages 709–727. Springer, 2022. 1, 3

[23] Menglin Jia, Luming Tang, Bor-Chun Chen, Claire Cardie, Serge Belongie, Bharath Hariharan, and Ser-Nam Lim. Visual prompt tuning. 2022. 3, 8

[24] Chen Ju, Tengda Han, Kunhao Zheng, Ya Zhang, and Weidi Xie. Prompting visual-language models for efficient video understanding, 2021. 3

[25] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 6

[26] Jogendra Nath Kundu, Naveen Venkat, Rahul M, and R. Venkatesh Babu. Universal source-free domain adaptation. 2020. 2

[27] Qicheng Lao, Xiang Jiang, and Mohammad Havaei. Hypothesis disparity regularized mutual information maximization, 2020. 2

[28] Brian Lester, Rami Al-Rfou, and Noah Constant. The power of scale for parameter-efficient prompt tuning. *arXiv preprint arXiv:2104.08691*, 2021. 1

[29] Xiang Lisa Li and Percy Liang. Prefix-tuning: Optimizing continuous prompts for generation. *arXiv preprint arXiv:2101.00190*, 2021. 1

[30] Jian Liang, D. Hu, and Jiashi Feng. Do we really need to access the source data? source hypothesis transfer for unsupervised domain adaptation. In *ICML*, 2020. 2

[31] Guosheng Lin, Anton Milan, Chunhua Shen, and Ian Reid. Refinenet: Multi-path refinement networks for high-resolution semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1925–1934, 2017. 11

[32] Jiaming Liu, Qizhe Zhang, Jianing Li, Ming Lu, Tiejun Huang, and Shanghang Zhang. Unsupervised spike depth estimation via cross-modality cross-domain knowledge transfer. *arXiv preprint arXiv:2208.12527*, 2022. 7

[33] Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *arXiv: Computation and Language*, 2021. 3

[34] Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Computing Surveys*, 55(9):1–35, 2023. 1

[35] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision, 2021. 3

[36] Ilija Radosavovic, Tete Xiao, Stephen James, Pieter Abbeel, Jitendra Malik, and Trevor Darrell. Real-world robot learning with masked visual pre-training. *CoRL*, 2022. 1

[37] Mamshad Nayeem Rizve, Kevin Duarte, Yogesh S Rawat, and Mubarak Shah. In defense of pseudo-labeling: An uncertainty-aware pseudo-label selection framework for semi-supervised learning. *arXiv preprint arXiv:2101.06329*, 2021. 2

[38] Subhankar Roy, Martin Trapp, Andrea Pilzer, Juho Kannala, Nicu Sebe, Elisa Ricci, and Arno Solin. Uncertainty-guided source-free domain adaptation. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXV*, pages 537–555. Springer, 2022. 4

[39] Christos Sakaridis, Dengxin Dai, and Luc Van Gool. Semantic foggy scene understanding with synthetic data. *International Journal of Computer Vision*, 126:973–992, 2018. 2, 6

[40] Christos Sakaridis, Dengxin Dai, and Luc Van Gool. Acdc: The adverse conditions dataset with correspondences for semantic driving scene understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10765–10775, 2021. 1, 2, 6

[41] Mark Sandler, Andrey Zhmoginov, Max Vladymyrov, and Andrew Jackson. Fine-tuning image transformers using learnable memory, 2022. 3

[42] Inkyu Shin, Yi-Hsuan Tsai, Bingbing Zhuang, Samuel Schulter, Buyu Liu, Sparsh Garg, In So Kweon, and Kuk-Jin Yoon. Mm-tta: Multi-modal test-time adaptation for 3d semantic segmentation, 2022. 2

[43] Junha Song, Kwanyong Park, Inkyu Shin, Sanghyun Woo, and In So Kweon. Cd-tta: Compound domain test-time adaptation for semantic segmentation, 2022. 2

[44] Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. *Learning*, 2017. 5

[45] Wilhelm Tranheden, Viktor Olsson, Juliano Pinto, and Lennart Svensson. Dacs: Domain adaptation via cross-domain mixed sampling. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1379–1389, 2021. 11

[46] Dequan Wang, Evan Shelhamer, Shaoteng Liu, Bruno Olshausen, and Trevor Darrell. Tent: Fully test-time adaptation by entropy minimization. *arXiv preprint arXiv:2006.10726*, 2020. 6, 7

[47] Dequan Wang, Evan Shelhamer, Shaoteng Liu, Bruno A. Olshausen, and Trevor Darrell. Tent: Fully test-time adaptation by entropy minimization. In *ICLR*, 2021. 2

[48] Qin Wang, Olga Fink, Luc Van Gool, and Dengxin Dai. Continual test-time domain adaptation. *ArXiv*, abs/2203.13591, 2022. 2, 3, 5, 6, 7, 9

[49] Wen Wang, Yang Cao, Jing Zhang, Fengxiang He, Zheng-Jun Zha, Yonggang Wen, and Dacheng Tao. Exploring sequence feature alignment for domain adaptive detection transformers. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 1730–1738, 2021. 6

[50] Zifeng Wang, Zizhao Zhang, Chen-Yu Lee, Han Zhang, Ruoxi Sun, Xiaoqi Ren, Guolong Su, Vincent Perot, Jennifer Dy, and Tomas Pfister. Learning to prompt for continual learning. In *CVPR*, 2022. 3

[51] Xinyi Wu, Zhenyao Wu, Hao Guo, Lili Ju, and Song Wang. Dannet: A one-stage domain adaptation network for unsupervised nighttime semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15769–15778, 2021. 12

[52] 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. 5, 6, 7, 11

[53] Shiqi Yang, Yaxing Wang, Joost van de Weijer, Luis Herranz, and Shangling Jui. Generalized source-free domain adaptation. *international conference on computer vision*, 2021. 2

[54] Yanchao Yang and Stefano Soatto. Fda: Fourier domain adaptation for semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4085–4095, 2020. 11

[55] Yuan Yao, Ao Zhang, Zhengyan Zhang, Zhiyuan Liu, Tat-Seng Chua, and Maosong Sun. Cpt: Colorful prompt tuning for pre-trained vision-language models, 2021. 3

[56] Jinze Yu, Jiaming Liu, Xiaobao Wei, Haoyi Zhou, Yohei Nakata, Denis Gudovskiy, Tomoyuki Okuno, Jianxin Li, Kurt Keutzer, and Shanghang Zhang. Cross-domain object detection with mean-teacher transformer. *arXiv preprint arXiv:2205.01643*, 2022. 6

[57] Yizhe Zhang, Shubhankar Borse, Hong Cai, and Fatih Porikli. Auxadapt: Stable and efficient test-time adaptation for temporally consistent video semantic segmentation, 2021. 2

[58] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2881–2890, 2017. 11

[59] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Conditional prompt learning for vision-language models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 3

[60] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *International Journal of Computer Vision (IJCV)*, 2022. 3