CoDA: Instructive Chain-of-Domain Adaptation with Severity-Aware Visual Prompt Tuning

Ziyang Gong[1], Fuhao Li[2], Yupeng Deng[3], Deblina Bhattacharjee[4], Xianzheng Ma[5(✉)], Xiangwei Zhu[1(✉)], and Zhenming Ji[1(✉)]

[1] Sun Yat-sen University, Guangzhou, China
[2] Wuhan University of Science and Technology, Wuhan, China
[3] National University of Singapore, Singapore, Singapore
[4] EPFL, Lausanne, Switzerland
[5] Wuhan, China
maxianzheng97@gmail.com

Abstract. Unsupervised Domain Adaptation (UDA) aims to adapt models from labeled source domains to unlabeled target domains. When adapting to adverse scenes, existing UDA methods fail to perform well due to the lack of instructions, leading their models to overlook discrepancies within all adverse scenes. To tackle this, we propose CoDA which instructs models to distinguish, focus, and learn from these discrepancies at scene and image levels. Specifically, CoDA consists of a Chain-of-Domain (CoD) strategy and a Severity-Aware Visual Prompt Tuning (SAVPT) mechanism. CoD focuses on scene-level instructions to divide all adverse scenes into *easy* and *hard* scenes, guiding models to adapt from source to easy domains with easy scene images, and then to hard domains with hard scene images, thereby laying a solid foundation for whole adaptations. Building upon this foundation, we employ SAVPT to dive into more detailed image-level instructions to boost performance. SAVPT features a novel metric *Severity* that divides all adverse scene images into *low-severity* and *high-severity* images. Then Severity directs visual prompts and adapters, instructing models to concentrate on unified severity features instead of scene-specific features, without adding complexity to the model architecture. CoDA achieves SOTA performances on widely-used semantic segmentation benchmarks under all adverse scenes. Notably, CoDA outperforms the existing ones by 4.6%, and 10.3% mIoU on the Foggy Driving, and Foggy Zurich benchmarks, respectively. Our code is available at https://github.com/Cuzyoung/CoDA.

Keywords: Adverse Scenes · Discrepancy · Semantic Segmentation · Chain-of-Domain · Severity · Visual Prompt Tuning

Z. Gong, F. Li, Y. Deng and D. Bhattacharjee—These authors contributed equally.
X. Ma—Independent Researcher.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/978-3-031-72980-5_8.

A. Leonardis et al. (Eds.): ECCV 2024, LNCS 15135, pp. 130–148, 2025.
https://doi.org/10.1007/978-3-031-72980-5_8

1 Introduction

Understanding all adverse scenes including fog, rain, snow, and night has become the common goal for existing UDA methods [19,23,25,32,36]. Although current methods show a good understanding of fog, rain, and snow scenes, when encountering the most challenging night scene which is full of unpredictable noise due to the extremely low light conditions, their performance drops a lot. Our findings in Fig. 1(a) illustrate that current SOTA methods [24,25] train models on all adverse scenes achieve precise segmentation on details of the image shown in white circles, but they struggle to recognize the sky in night scenes shown in yellow circles. On the contrary, their models trained on a single night scene can clearly segment the sky but show worse results in other details. The results reveal that current methods are trapped in dilemmas that adapting to all adverse scenes leads models to hallucinations while adapting to a single adverse scene causes models to underfit.

Given the essence of UDA is knowledge distillation allowing the student network to learn *instructive* knowledge from the teacher network [51], we argue that for existing methods, the adaptation to all adverse scenes requires scene-level instructions and the adaptation of a single adverse scene needs image-level instructions to overcome hallucinations and underfitting, respectively. Addressing these issues, we propose CoDA (**C**hain-**o**f-**D**omain **A**daptation) methodology focusing on these two levels to instruct and enhance models to learn domain-invariant features through the Chain-of-Domain (CoD) strategy with Severity-Aware Visual Prompt Tuning (SAVPT) mechanism.

The design of CoD motivated by Chain-of-Thought (CoT) is focused on providing scene-level instruction. It divides all adverse scenes into *easy* and *hard* categories of scenes and instructs models to adapt from source to target domains, in a step-by-step fashion [30], through extra intermediate domains as shown in Fig. 1(b). As the adage says "*Well begun is half done*", we maintain that acquiring high-quality prior knowledge at the start of training is critical for iterative learning. Thus, CoD starts by instructing models to train on *easy* scene images to build a solid foundation, and then, CoD adapts to *hard* scene images for further knowledge transfer. Additionally, as we argue that a large number of the original target images are somewhat hard for models to learn, we propose a tailored mini-dataset of adverse scenes generated by combining three Large Multimodal Models (LMMs) to serve as a part of the easy scene images for good starts to the training. More details are illustrated in the Sect. 3.2. In summary, CoD constructs intermediate domains according to scene-level difficulties of adverse scenes, instructing models to adapt to mixed gaps from easy to hard.

One might ask "Why does the direct adaptation to all adverse scenes not instruct models with an easy start?" We denote this kind of direct adaptation as the *traditional strategy* which randomly samples all adverse scene images, potentially treating disparate scenes as equivalent, overlooking their various difficulties. It might sample challenging night images at the onset of training causing initial errors. Concurrently, since pseudo-labels within UDA are inherently uncertain, the initial errors easily accumulate during iterative training. Thus, the traditional strategy is not suitable for models at the start of training. Compared

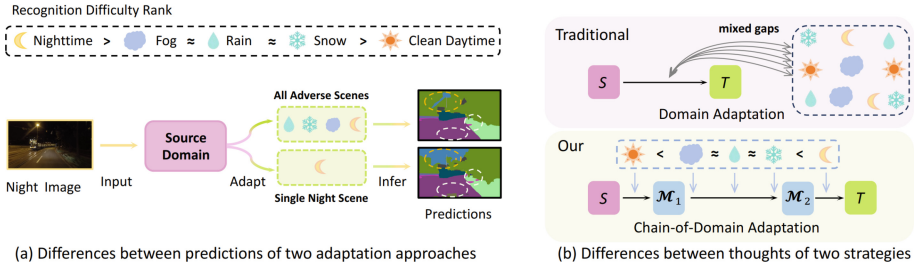


Fig. 1. (a) Current SOTA models [24,25] trained on all adverse scenes within a target domain can achieve good performance on other details but struggle to recognize the sky under night scenes. These models, typically, trained on a single night scene show the contrary results. Yellow circles in predictions denote the sky recognition and white ones indicate other classes' recognition. (b) Traditional strategy directly adapts from source to target domains with chaotic gaps. Our Chain-of-Domain (CoD) strategy *instructs* models to adapt from source to target domains according to the difficulties of scenes through introducing intermediate domains. (Color figure online)

to CoD, however, the traditional strategy is less instructive but more diverse. Therefore, to improve scene diversity, we employ CoD to guide models during the preliminary training and activate the traditional strategy later.

Based on the solid foundation brought by the scene-level CoD strategy, we hope to focus on a more detailed image-level perspective to further enhance the inherent abilities of models for extracting domain-invariant features. Since the complex architectures of existing models overfit scene-specific features, we aim to present a method that can enhance models' abilities without complicating the network architecture. The seminal work of Darcet et al. [11] has significantly inspired our approach, particularly their remarkable finding that *visual prompts, when incorporated into the input data of Vision Transformer (ViT), can be omitted during inference, thereby enhancing the ViT's performance.* This finding validates that visual prompts can enhance models' inherent ability without being a part of network structures, which aligns well with our motivation. Thus, we present the Severity-Aware Visual Prompt Tuning (SAVPT) with an image-level metric *Severity* to measure the severity differences within each image.

SAVPT contains a Severity Perception Trigger (SPT), Meta-Visual Prompts, and Meta-Adapters. SPT classifies adverse images to *low-severity* and *high-severity* images. The Meta-Visual Prompts and Meta-Adapters modules are learnable components that build upon the SPT mechanism and are divided into two severity branches. When encountering a low-severity image, the low-severity branch will be activated. Then the Meta-Visual Prompts and Meta-Adapters in this branch will enhance the low-severity image and optimize its features respectively. Concurrently, another branch will be frozen without training. This important mechanism ensures that the two branches will gradually demonstrate different severity emphases, leading Meta-Visual Prompts and Meta-Adapters to help models focus on severity features rather than scene-specific features. In later experiments, we will validate that SAVPT trained with models can also be discarded during inference.

The whole CoDA method offers three key contributions: (1) CoDA is the first to propose CoT-based variants, Chain-of-Domain, in UDA for adverse scene understanding; (2) CoDA validates the findings in [11] through experiments on the application of SAVPT during inference. This further confirms that SAVPT empowers models to learn domain-invariant features. (3) CoDA achieves state-of-the-art performance on multiple widely used adverse scene benchmarks. In particular, CoDA outperforms SOTA methods by large margins with **4.6%** and **10.3%** mIoU on Foggy Driving and Foggy Zurich benchmarks.

2 Related Work

2.1 Semantic Segmentation Under Adverse Scenes Within UDA

Since existing works [56,57,59,60,68] predominately focus on four adverse scenes (fog, rain, snow, and night), we divide them into three branches by the scene benchmarks that they evaluate. The first branch comprises methods focusing on parsing foggy scenes [9,42,43] trying to generate synthetic fog images to narrow the real-to-foggy gaps. Later works tried other paths like CuDA-Net [36] that reported disentangling gaps by introducing an intermediate domain, and FIFO [31] that presented an auxiliary network to help the segmentation model to learn fog-invariant features. Since the above works generalize well to rainy and snowy scenes, no methods are specializing in rainy and snowy scenes. Night-specialized methods [58] are contained in the second branch and they too focus on narrowing the gaps by introducing extra domain images like twilight images or extra information [10,41,44]. The final branch comprises models trying to solve multiple scenes including normal and adverse scenes. DAFormer [23] was the first to introduce the SegFormer-based [62] architecture to this field. Subsequently, some DAFormer-based methods [4,61] emerged like STA [19] which achieves domain generalization in UDA, and HRDA [24] that proposes multi-scale features fusion network. Some works also take HRDA as the baseline model, such as MIC [25] using masks to help models learn context-level features. Different from previous methods, CoDA is the first UDA method to focus models on differences within adverse scenes to learn domain-invariant features.

2.2 CoT and its Variants

The essential of CoT is a series of intermediate reasoning steps. CoT is proposed by [55] as a simple and effective prompting strategy to enhance the complex reasoning ability of Large Language Models (LLMs) to accomplish reasoning tasks including arithmetic, commonsense, and symbolic reasoning tasks. Recent years have witnessed an explosive growth of CoT works in LMMs [6,18,21,27,37,40], where they arouse LLMs' and LMMs' potential and boost alignment between multimodality. Concurrently, there have been a series of works focusing on CoT variants. Specifically, Chain-of-Thought Self-Consistency (CoT-SC) [52] samples the CoT output multiple times and chooses the most consistent one, Tree-of-Thought (ToT) [63] proposes tree-structured CoT allowing LLMs to consider

different reasoning paths and self-evaluate to determine the action, Graph-of-Thought (GoT) [3] creates graph-based CoT closing LLMs and human thinking, and Chain-of-Reasoning (CoR) [49] trains models to question themselves according to an uncertainty factor, further enhancing the answer's confidence. However, there are no CoT variants in UDA for adverse scene understanding currently. Thus, we hope our Chain-of-Domain (CoD) method will provide an initial contribution in this field.

2.3 Visual Prompt Tuning in UDA

Prompt-based learning [34] initially serving as Parameter-Efficient Fine-Tuning (PEFT) [7,22,67] is used to finetune large pre-trained models to downstream tasks in Natural Language Processing (NLP). Subsequently, the prompt has begun to be transferred to the field of Computer Vision (CV) bringing about a significant academic trend. Many methods [29,38,64,69] first tried to migrate prompt to Vision Language Models (VLM) in text form, and then VPT [28] first introduced "visual prompt" into CV as learnable vectors (soft prompt). [17] followed by other works [15,16,46] is the first work focusing on domain adaptation and employing prompt. Most relevant to our work are the methods that focus on UDA or DG for adverse scene understanding, such as [14] employs text features to strengthen features achieving a novel style augmentation implicitly, and [50] combines clip [38] to conduct semantic augmentation.

3 Approach

3.1 Preliminary

We denote a student network as f_θ and a teacher network as f_ϕ. Unlike traditional UDA methods, CoDA utilizes multiple domain images including source domain images $X_S \in \mathbb{R}^{H_S \times W_S \times C}$, target domain images $X_T \in \mathbb{R}^{H_T \times W_T \times C}$, and intermediate domain images $X_{\mathcal{M}_n} \in \mathbb{R}^{H_T \times W_T \times C}$, where H, W, and C represent the height, width, and channel of images. In our experiments, we set n to 2 and $X_\mathcal{M}$ to share the same shape with X_T. Thus the union of intermediate and target images is represented as $X_{T'} = X_\mathcal{M} \cup X_T$. Notably, only source domain labels $Y_S \in \mathbb{R}^{H_S \times W_S}$ are available. All domain images are combined to calculate the Cross-Entropy (CE) loss to iteratively update f_θ. Then the parameters of f_θ will be synchronized by EMA [47] to f_ϕ which has no gradient backpropagation.

3.2 Training-Free Pipeline to Generate Mini-Dataset

Our mini-dataset contains 1200 (3×400) images including fog, rain, and snow images aiming to serve as the easiest images with slightly adverse weather factors to construct a part of M_1. Since text prompts Φ are decisive for the quality of images generated by Stable Diffusion 2 (SD2) [39], we propose a training-free pipeline to generate and optimize Φ step by step through collaborating with CLIP [38], GPT-4V [1], and Human Feedback. The pipeline consists of 4 stages:

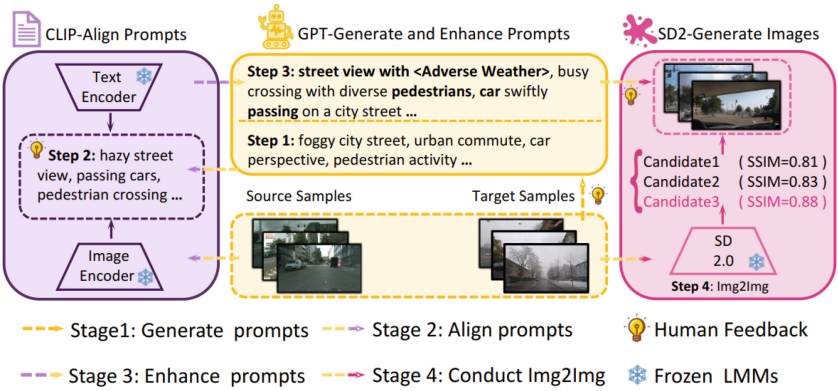


Fig. 2. Four Stages Training-free Pipeline generates the mini-dataset to serve as intermediate steps providing diverse difficulty levels. All images are adverse scene images with slight weather factors generated based on ACDC adverse scene images and possess the common features of target and source images. Notably, stages 1, 2, and 4 contain manual processes based on human feedback.

(1) The first stage uses GPT-4V to generate the basic phrasal descriptions Φ_1 of X_T. Notably, we aim to obtain Φ_1 to represent the basic features of X_T. Since all X_T are captured in the same city Zurich, the generated Φ_1 of different images are similar to each other, except for the weather factors. Thus, we only randomly feed 6 target domain images X_Φ (fog, rain, and snow respectively contain 2 images) to GPT-4V with text instruction Ψ_1: "*Please generate some phrases to describe the scene, style, and content of this picture, such as driving in Europe.*" Then, every image will generate dozens of output phrases like "Foggy city street", and "Urban commute" as shown in stage 1 of Fig. 2.

$$\Phi_1 = GPT - 4V(\, X_\Phi, \Psi_1 \,) \tag{1}$$

(2) Since the dataset is the intermediate step between source and target, the second stage aims to select the most appropriate text phrases Φ_2 that represent both features of X_S and X_T by aligning Φ_1 and X_S. We input Φ_1 and all of X_S into the CLIP text encoder $\mathcal{T}$ and image encoder $\mathcal{V}$ respectively to calculate cosine similarity. Then we select Φ_1 with high similarity.

$$\Phi_2 = argmax_{C'}\left(\frac{\mathcal{T}(\Phi_1) \cdot \mathcal{V}(X_S)}{\|\mathcal{T}(\Phi_1)\| \times \|\mathcal{V}(X_S)\|}\right) \tag{2}$$

where C' is the number of Φ_2 for each image in X_Φ and we set it as 1 or 2. Since many phrasal prompts have similar semantic information, we manually filter the redundant based on human feedback to choose 5 texts to build Φ_2: "hazy street view, passing cars, pedestrian crossing, cyclists on city lane, car Perspective."

(3) The third stage recurs to GPT-4V to enhance Φ_2. We employ GPT-4V to enrich the semantic representation of Φ_2 by this instruction Ψ_2: "Please

enrich the description of $<\Phi_2>$, requiring to maintain the simplicity and the semantic information." We denote enhanced Φ_2 as Φ_3: "street view with *fog*, car swiftly passing on a city street, busy crossing with diverse pedestrians, cars and cyclists parked on a quiet road, inside view of a car on a busy road."

(4) Before conducting image-to-image at stage 4, we first finetune the Φ_3 based on human feedback. In detail, we adjust the $<fog>$ in Φ_3 to "slight <fog/rain/snow>" to generate different scene images. Besides, we found that SD2 is sensitive for cyclists and it always generates a huge number of cyclists, so we changed the cyclists to motorbikes and controlled their number. Thus, the final Φ_4 used to prompt SD2 consists of: "street view with *slight* $<AdverseWeather>$, car swiftly passing on a city street, busy crossing with diverse pedestrians, *single car and motorbike* parked on a quiet road, inside view of a car on a busy road." We feed Φ_4 and X_T to SD2 to generate 3 images as candidates and choose the one with the highest SSIM [53] with X_T to be final images X_G. More details are attached to the supplementary material.

$$X_G = argmax(\, SSIM(\, SD2(\, X_T, \Phi_4\,)\,)\,) \tag{3}$$

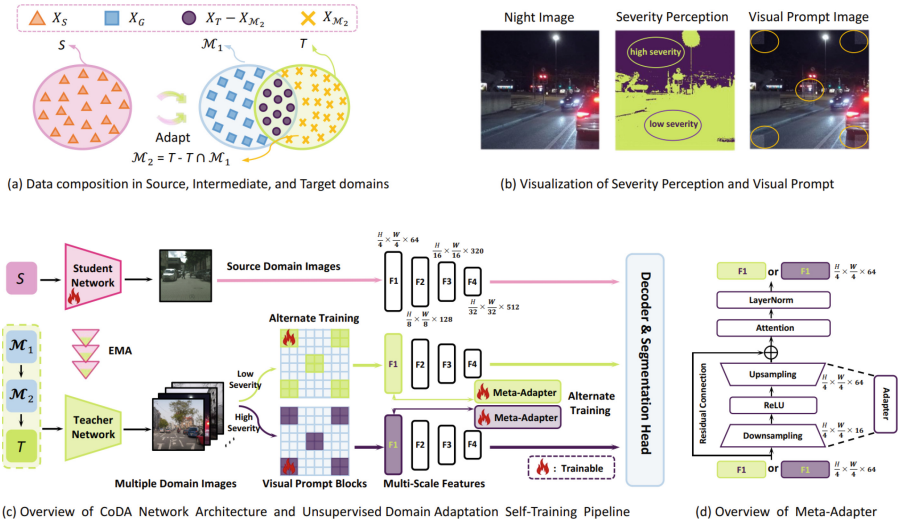
3.3 Instructive Chain-of-Domain Adaptation

The first key point in scene-level CoD aims to simply divide all adverse scenes into **Easy** and **Hard** two-level scenes to construct two intermediate domains, $\mathcal{M}_1$ and $\mathcal{M}_2$. Take Cityscapes to ACDC experiment as an example, ACDC representing the target domain T consists of fog, rain, snow, and night scenes, so we directly choose daytime scene images (fog, rain, snow, and X_G) as easy scene images to construct $X_{\mathcal{M}_1}$ and use night images to build $X_{\mathcal{M}_2}$. The data composition of Cityscapes to ACDC can be represented as Fig. 3(a). Thus, the CoD adaptation process transforms from $S \rightarrow T$ into $S \rightarrow \mathcal{M}_1 \rightarrow \mathcal{M}_2 \rightarrow T'$.

The second pivot is how to train models during the CoD adaptation process. As our motivation in the Intro that "easy before difficult", we first train models on $\mathcal{M}_1$ with 1.2k iterations and then we train models on $\mathcal{M}_2$ with 4k iterations. Since $\mathcal{M}_2$ is a tougher scene than $\mathcal{M}_1$, we advocate models to spend more time on $\mathcal{M}_2$. The first 1.2k iterations can instruct good prior knowledge to models, which makes models robust enough when they first encounter tough images. Then, we repeat this process until iterations come to 12k. We also emphasize that although the traditional strategy is not suitable for the start of training, it is still advantageous as scene diversity to enhance models' performance upper bound. Thus, from 12k to the end of the training, we activate the traditional strategy to conduct random sampling from $X_{T'}$ formulated as Eq. 4 and 5.

$$Iterations = \begin{cases} 1.2k, & when \quad S \rightarrow \mathcal{M}_1 \\ 4k, & when \quad \mathcal{M}_1 \rightarrow \mathcal{M}_2 \\ \infty, & when \quad \mathcal{M}_2 \rightarrow T' \end{cases} \tag{4}$$

$$CoD = \begin{cases} S \rightarrow \mathcal{M}_1 \rightarrow \mathcal{M}_2, & when \quad Iter \leqslant 12k \\ \mathcal{M}_2 \rightarrow T', & until \ training \ EOF \end{cases} \tag{5}$$



(a) Data composition in Source, Intermediate, and Target domains

(b) Visualization of Severity Perception and Visual Prompt

(c) Overview of CoDA Network Architecture and Unsupervised Domain Adaptation Self-Training Pipeline

(d) Overview of Meta-Adapter

Fig. 3. (a) shows the data composition in experiments of Cityscapes to ACDC. (b) demonstrates the visualization of SPT and Meta-Visual Prompts. The purple pixels are severe pixels and the green pixels are the nonsevere. (c) and (d) shows the details of CoDA's architecture and pipeline. (Color figure online)

3.4 Severity-Aware Visual Prompt Tuning

SAVPT consists of a Severity Perception Trigger (SPT) mechanism, Meta-Visual Prompts, and Meta-Adapters. According to SPT, Meta-Visual Prompts, and Meta-Adapters are divided into two branches to focus on images with two severity levels. Both of these two branches share the same structure and initial parameters but train alternately.

Severity Perception Trigger. Different from CoD which conducts scene-level classification, the SPT mechanism focuses on image-level dividing and guides the whole SAVPT by measuring the severities of all adverse scene images. First of all, images are input to SPT will be transformed into gray-scale maps $X_g \in \mathbb{R}^{H \times W}$, and then all pixels will be divided into severe pixels and nonsevere pixels. In SPT, the pixel values lower than the gray-scale value σ are severe pixels. If the ratio of severe pixels in an image passes the severity threshold τ, this image will be classified as a high-severity image; otherwise, it will be recognized as a low-severity image. We show the visualization in Fig. 3(b) that the purple regions are severe pixels with high severity and the green pixels are nonsevere pixels with low severity.

$$Severity = \begin{cases} high, & if \ \frac{X_g < \sigma}{H \times W} > \tau \\ low\ , & else \end{cases} \tag{6}$$

Meta-visual Prompts and Adapters. We denote Meta-Visual Prompts as $\delta_v \in \mathbb{R}^{H_v \times W_v \times C}$ which is explicitly added to the images similar to the adversarial

reprogramming [13]. We set default $C = 3$, $H_v = W_v = 64$, and the number of δ_v for every image is 5. To avoid δ_v overly masking useful information of images, the 5 δ_v are added to the four corners and the center of images visualized in Fig. 3(b). Notably, since we aim to instruct models understanding adverse scenes, δ_v is set to be activated only on $X_{T'}$ rather than X_S. Thus, we denote the augmented images as $\hat{X}_{T'}$ and augment equation as below:

$$\hat{X}_{T'} = X^H_{T'} \cup X^L_{T'}, \quad X^H_{T'} = X^h_{T'} + 5 \cdot \delta^h_v, \quad X^L_{T'} = X^l_{T'} + 5 \cdot \delta^l_v \tag{7}$$

where X^h, X^l, δ^h_v, δ^l_v, X^H, and X^L respectively means that high-severity images, low-severity images, high-severity visual prompts, low-severity visual prompts, augmented X^h, and augmented X^l. The pipeline of Meta-Visual Prompts adding to images is shown in Fig. 3(c).

The design of Meta-Adapters shown in Fig. 3(d) is lightweight consisting of a vanilla adapter [22] with residual connection [20], and common attention layer. The adapter layer will absorb the input feature F_1 of $\hat{X}_{T'}$ by downsampling and output an optimized feature by upsampling. Then, the feature will pass an attention layer Att again with layer normalization LN to obtain a new robust F_1. During this process, Meta-Adapters aim to absorb the prompt information from Meta-Visual Prompts. The pipeline of using Meta-Adapter can be formalized as below:

$$F_1^{H/L} = LN(\, Att^{H/L}(\, Adapter^{H/L}(\, F_1^{H/L}\,)\,)\,) \tag{8}$$

where H/L denotes two severity branches and F_1 denote one of the multi-scale features shown in Fig. 3(c) and (d).

3.5 Loss for Training

As we said above, only student network f_θ is trainable and the function of teacher network f_ϕ is generating pseudo-labels $\hat{Y}_{T'}$. The whole CoDA architecture is updated by calculating CE Loss. The student loss L_S as following:

$$L_S = -\sum_{H_S, W_S} \sum_{C} Y_S \, log f_\theta(X_S) \tag{9}$$

$$\hat{Y}_{T'} = [\, c = argmax_{c' \in C} f_\phi(X_{T'})\,] \tag{10}$$

where $[\cdot]$ denotes the Iverson bracket. Before calculating the teacher loss L_T, we still need to obtain a confidence estimate q [23–25] to limit the pseudo-labels which are not completely believable. Expect for L_S and L_T, we follow our baseline setting of feature distance loss L_{FD}. The summary training loss function L is as below:

$$L_T = -\sum_{H_T, W_T} \sum_{C} q\, \hat{Y}_{T'} \, log f_\theta(X_{T'}) \tag{11}$$

$$L = L_S + L_T + L_{FD} \tag{12}$$

Table 1. Comparision with existing methods on all adverse scene benchmarks. Bold denotes the best mIoU, underline means the second-best mIoU, and italics denotes the performance tested by ourselves.

Models	Backbone	Adverse Scenes (mIoU)									
		Foggy			Nighttime				Rainy	Snowy	All
		ACDC-Fog	FZ	FD	ACDC-Night	DZ	ND	BD	ACDC-Rain	ACDC-Snow	ACDC-All
Scene-Specialized Models											
SFSU [43]	RefineNet [33]	-	35.7	46.3	-	-	-	-	-	-	-
CMAda3 [9]	RefineNet	-	46.8	49.8	-	-	-	-	-	-	-
CuDA-Net [36]	DeepLab-v2 [5]	55.6	49.1	53.5	-	-	-	-	-	-	-
FIFO [31]	RefineNet	-	48.4	50.7	-	-	-	-	-	-	-
FogAdapt [26]	ResNet-38	-	50.6	53.4	-	-	-	-	-	-	-
SAM-EDA [54]	ViT-H [12]	-	-	56.4	-	-	-	-	-	-	-
GCMA [41]	RefineNet	-	-	-	-	42.0	45.6	33.2	-	-	-
MCGDA [44]	RefineNet	-	-	-	-	42.5	49.4	34.9	-	-	-
DANNet [58]	RefineNet	-	-	-	-	44.3	47.7	-	-	-	50.0
Scene-Agnostic Models											
AdaptSeg [48]	DeepLab-v2	-	26.1	37.6	-	30.4	34.5	22.0	-	-	-
DAFormer [23]	SegFormer [62]	48.9	40.8	-	44.7	48.5	51.8	33.9	59.9	53.7	55.4
SePiCo [61]	SegFormer	58.5	-	-	50.5	54.2	56.9	40.6	66.1	57.9	59.1
STA [19]	SegFormer	60.2	46.9	54.9	48.4	-	-	-	61.3	58.0	60.9
HRDA [24]	SegFormer	69.9	46.0	-	53.1	55.9	-	-	73.6	69.5	68.0
MIC [25](baseline)	SegFormer	*67.0*	*53.3*	*56.6*	*57.2*	60.2	*58.6*	*41.3*	*72.3*	*66.6*	70.4
CoDA (Ours)	SegFormer	**71.8**	**60.9**	**61**	**66.4**	**61.2**	**59.2**	**41.9**	**75.3**	**70.9**	**72.6**

4 Experiments

4.1 Datasets and Benchmarks

Cityscapes (CS) [8] is a real-world dataset captured with normal driving scenes in 50 urban areas. **Foggy Zurich** (FZ) [42] and **Foggy Driving** (FD) [43] are two foggy real-world datasets. FZ contains 3808 images with light and medium fog. Notably, only 40 images are for testing in FZ. Unlike FZ, FD is a dataset purely for testing as it contains 101 labeled images. **Dark Zurich** (DZ) [41] is captured during nighttime, twilight, and daytime including 50 validation images and 151 test images. **Nighttime Driving** (ND) [10] contains 50 nighttime images with coarsely annotated ground truth. **BDD100K-Night** (BD) [65] is a subset of the BDD100K, we only use 87 test images. ND is also used for testing. **ACDC** [45] is a dataset under fog, rain, snow, and night scenes. There are 400 training images, 100 validation images (106 in night), and 500 testing images in every scene. Besides, ACDC also contains 1600 clean images called ACDC-ref.

4.2 State-of-the-Art Performance Comparison

Implement Details. Our performance experiments contain four aspects: foggy, nighttime, all-scenes, and other benchmarks. In experiments, we divide existing methods into two classes, the first is Scene-Specific methods and the second is Scene-Agnostic methods. For details of network settings, we follow the default settings of our baseline MIC [25], including the optimizer AdamW [35], encoder

Table 2. The results of MIC and MIC tained with CoDA from CS to ACDC. w/ SAVPT indicates that SAVPT is activated during inference, while w/o SAVPT indicates the opposite. Bold means the best performance.

Method	Road	S.walk	Build	Wall	Fence	Pole	T.Light	Sign	Veget	Terrain	Sky	Person	Rider	Car	Truck	Bus	Train	M.bike	Bike	mIoU
MIC	90.8	67.1	89.2	54.5	40.5	57.2	62.0	68.4	76.3	**61.8**	87.0	71.3	**49.4**	89.7	75.7	**86.8**	89.1	**56.9**	63.0	70.4
w/ SAVPT	93.1	**72.8**	90.7	57.3	47.4	**59.8**	69.8	69.9	87.2	59.7	95.4	71.1	47.3	90.2	76.7	82.9	**89.8**	55.0	63.7	72.5
w/o SAVPT	**93.1**	72.7	**90.7**	**57.3**	**47.4**	56.8	**69.9**	**70.0**	**87.3**	59.8	**95.4**	**71.4**	47.6	**90.3**	**77.1**	83.8	89.1	54.7	**64.1**	**72.6**

Table 3. Range comparison of DAFormer, HRDA. and MIC with different strategies and seeds. Bold denotes the best range and all scores are mIoU at 40k iterations. The results show that CoD-included strategies can enhance the stability of models.

Models	Strategy	Average mIoU	Seed = 1	Seed = 2	Seed = 38	Range (max-min)
DAFormer [23]	Tradition	56.3	55.5 (−0.8)	57.2 (+0.9)	56.3 (+0.0)	1.7
	CoD	57.7	58.6 (+0.9)	58.5 (+0.8)	56.1 (−1.5)	1.7 (↓ 0%)
	CoD+Tra.	57.7	57.2 (−0.5)	58.5 (+0.8)	57.4 (−0.3)	**1.3(↓ 24%)**
HRDA [24]	Tradition	64.6	64.8 (+0.2)	62.8 (−1.8)	66.2 (+1.6)	3.4
	CoD	65.4	65.0 (−0.4)	64.2 (−1.2)	67.0 (+1.6)	2.8 (↓18%)
	CoD+Tra.	65.7	65.8 (+0.1)	65.9 (+0.2)	65.4 (−0.3)	**0.5(↓ 85%)**
MIC [25]	Tradition	65.9	67.7 (+1.8)	63.0 (−2.9)	67.1 (+1.2)	4.7
	CoD	68.3	67.9 (−0.4)	68.4 (+0.1)	68.7 (+0.4)	0.8 (↓ 83%)
	CoD+Tra.	70.0	70.2 (+0.2)	69.6 (−0.4)	70.3 (+0.3)	**0.7(↓ 85%)**

with a learning rate of 6×10^{-5}, decoder with 6×10^{-4} learning rate, and a linear learning rate warm-up. We mainly adjust one hyper-parameter that severity threshold τ in SPT. All experiments are conducted on a 32G Tesla V100 GPU.

Cityscapes to Foggy Scenes. We conduct experiments on CS to ACDC-Fog, FZ, and FD. For the CS to ACDC-Fog, we first train CS to ACDC-All (fog, rain, snow, night) and test on ACDC-Fog. For CS to FZ and FD, we train CS to FZ and test on FZ and FD validation. The results in Table 1 show that CoDA achieves SOTA performance FogAdapt and SAM-EDA with improvements of 10.3% and 4.6% mIoU in FZ and FD respectively, and outperforms our baseline MIC with 4.8% in ACDC-Fog, 7.6% in FZ, and 4.4% in FD. Notably, we set τ to 0.38 in CS to ACDC and 0.05 to FZ and FD. We contend that the significant discrepancies in the values of τ can be attributed to the diverse scene composition of ACDC requires clear distinction, unlike the singular foggy scene in FZ and FD. For the data composition of CS to FZ, we distribute the light fog images in FZ as the $X_{\mathcal{M}_1}$ and medium fog images as the $X_{\mathcal{M}_2}$ without using X_G.

Cityscapes to Nighttime Scenes. The nighttime experiments contain CS to ACDC-Night, DZ, ND, and BD. Similar to the above, we also train CS to ACDC-All and test on ACDC-Night. For the rest experiments, we train CS to DZ and generalize to ND and BD. The results in Nighttime of Table 1 demonstrate that CoDA outperforms the SOTA methods with improvements of 1% mIoU in DZ, 2.3% in ND, and 1.3% in BD, and outperforms MIC with 9.2%, 0.6%, and 0.6% mIoU in ACDC-Night, ND, and BD respectively. From CS to DZ, ND, and BD,

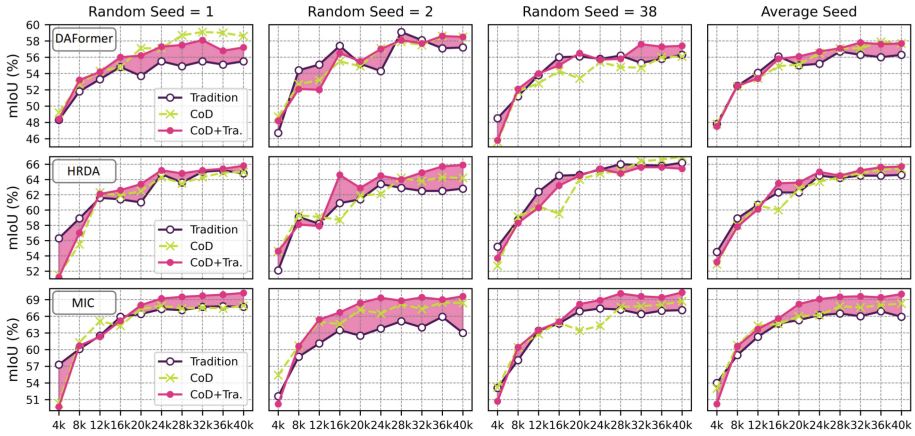


Fig. 4. Ablation studies on CS to ACDC val. The x and y axes respectively mean the mIoU and Iteration. The purple, green, and red lines respectively mean original models with traditional strategy, CoD strategy, and CoD+traditional strategy that we implement in CoDA. (Color figure online)

we designate the threshold τ of 0.05. Additionally, we allocated 400 night images in the ACDC-ref to establish $\mathcal{M}_1$, while utilizing the original night images from DZ to form $\mathcal{M}_2$.

Cityscapes to Rainy, Snowy, and All Scenes. We conduct CS to ACDC-All and test on ACDC-Rain, Snow, and All for these experiments. As shown in Table 1, CoDA achieves 72.6% mIoU on the ACDC-All benchmark outperforming the MIC 2.2%, and demonstrates strong generalizability on ACDC-Rain and Snow with improvements of 3.0% and 4.3% respectively compared to MIC. For more details, we show the results of every class of CS to ACDC-All in Table 2. As illustrated in the Intro section, the previous methods show weakness in recognizing the sky under night scenes, which our CoDA alleviates with 8.4% improvements in the sky class compared to MIC.

4.3 Ablation Study Analysis

Well Begun Enhances Models Stability. We defend that CoD is instructive for models to acquire high-quality prior knowledge which is critical for models' robustness. To validate it, we conduct CS to ACDC (val) experiments on 3 SegFormer [62]-based UDA models, DAFormer [23], HRDA [24], and MIC [25] with 3 random seeds shown in Table 3. Every model is trained with three different strategies, traditional strategy, Chain-of-Domain (CoD), and Chain-of-Domain+traditional strategy (CoD+Tra.).

Notably, we mainly compare the *Range* in Table 3 to show the stability of models. We first focus on the range of models with traditional strategy. The ranges escalate from 1.7 of DAFormer to 4.1 of MIC which means that with

Table 4. Ablation studies of positions and initialization of Meta-Visual Prompts on CS to ACDC val. All the mIoU score is the highest score during the training. $\delta_v \sim \mathcal{U}(0,1)$ and $\delta_v \sim \mathcal{N}(0,1)$ mean δ_v samples from uniform and normal distributions respectively.

Position of Prompts	Initial Parameter			
	$\delta_v = 1$	$\delta_v \sim \mathcal{U}(0,1)$	$\delta_v \sim \mathcal{N}(0,1)$	$\delta_v = 0$
Best mIoU results during 0∼40k iteration				
Random Patch	**70.8**	68.6	69.0	71.1
Padding Patch	67.2	**69.4**	**69.9**	**71.4**
Corner+Center	67.6	**69.4**	69.2	71.3
Best mIoU results during 40∼60k iteration				
Random Patch	**70.8**	69.6	**70.2**	71.8
Padding Patch	67.4	69.5	69.2 ↓	71.6
Corner+Center	67.7	**69.7**	69.6	**72.1**

the performance and the complexity of models improving, the instability is worse. CoD can alleviate it by stabilizing the ranges with improvements of 18% range in HRDA and 83% range in MIC. The results also show that the range of DAFormer with CoD is not optimized. We argue the reason is that DAFormer with low complexity is relatively stable. But CoD optimizes DAFormer performances with 1.4% improvements in the average mIoU. Notably, CoD with the traditional strategy brings huge improvement to every model with 24%, 85%, and 85% ranges in DAFormer, HRDA, and MIC respectively, which means that the traditional strategy possessing diversity can also stabilize the models when it is employed with CoD. To observe the tendency of Table 3, we also show the line charts of performance from 0 to 40k iterations in Fig. 4. The red shadows demonstrate models trained CoD with the traditional strategy outperform the single traditional strategy significantly.

SAVPT Enhances Models' Inherent Abilities. We create SAVPT to enhance models' inherent abilities to learn domain-invariant features without complicating original models' architectures. To validate it, we conduct ablation studies of the activation of SAVPT during inference time. As illustrated in Table 2, inference w/ SAVPT and w/o SAVPT achieve similar performance, and w/o activating SAVPT even shows superiority over w/ SAVPT by 0.1% mIoU, which shows SAVPT indeed strengthens the models themselves rather than serve as a part of models. Concurrently, compared to the original MIC model, MIC trained with CoDA achieves obvious improvement in road, sidewalk, wall, fence, vegetation, and sky classes which empirically represent domain-invariant classes shown in Fig. 5. Thus, these phenomena effectively validate that SAVPT is a knock-down composition that can lead models to learn domain-invariant features, which aligns well with our motivation. Remarkably, our findings based on [11] unveil another interesting phenomenon: not only can visual prompts be discarded at the inference time without negatively impacting models' perfor-

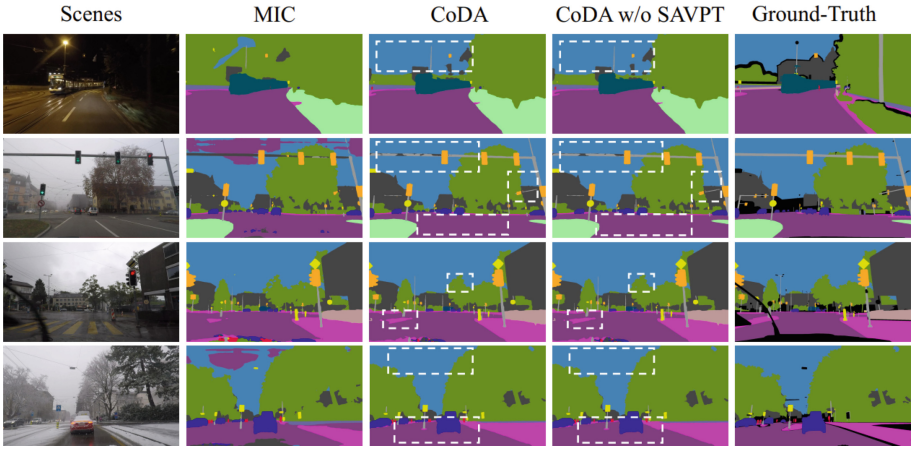


Fig. 5. Quantitative experiments between MIC, MIC trained with CoDA, and Mic trained with CoDA but without SAVPT during inference time. The results reveal that CoDA understands all scenes better and SAVPT enhances models' abilities.

mance, but adapters also exhibit attributes akin to that of visual prompts. This observation can probably be ascribed to the efficient yet potent design of both the original adapters and our Meta-Adapters, allowing the implementation of SAVPT without requiring additional computational costs during inference.

Although the adding position of visual prompts is discussed in [2,15,66], the discussion of the initial parameters of visual prompts is relatively blank. Thus, we conduct joint experiments of positions with different initializations shown in Table 4. During 40∼60k, almost each experiment improves compared to their 0∼40k scores and the Corner+Center shows the best performance with 72.1% mIoU. When initial $\delta_v = 1$, Padding Patch and Corner+Center performances reduce a lot contrary to Random Patch with 70.8% mIoU. We argue that $\delta_v = 1$ gives models a noisy begin and enough randomness can alleviate the noise influences. Also, the initialization with uniform and normal distribution achieves similar mIoU, which means MIC is not sensitive to initial parameter distributions and results of $\delta_v = 0$ demonstrate that starting from scratch is the best choice for Meta-Visual Prompts.

5 Conclusion

In this paper, we propose a method named CoDA, which consists of a CoD strategy with a tailored dataset and SAVPT mechanism. Extensive experiments validate that CoD provides scene-level instructions to models for eliminating hallucinations on tough scenes, while SAVPT serves as a plug-in mechanism enhancing models' inherent abilities without complicating networks through image-level instructions. In this paper, however, we do not detailedly discuss why the Meta-

adapter emerges attributes akin to visual prompts that can be discarded during training. This important phenomenon will be one of our future investigations.

Acknowledgement. This study was funded by the National Natural Science Foundation of China (Grants No. U21A6001) and the Innovation Group Project of Southern Marine Science and Engineering Guangdong Laboratory (Zhuhai) (Grants No. SML2023SP208). We also acknowledge the high-performance computing support from School of Atmospheric Science at Sun Yat-sen University.

References

1. Achiam, J., et al.: GPT-4 technical report. arXiv preprint arXiv:2303.08774 (2023)
2. Bahng, H., Jahanian, A., Sankaranarayanan, S., Isola, P.: Exploring visual prompts for adapting large-scale models. arXiv preprint arXiv:2203.17274 (2022)
3. Besta, M., et al.: Graph of thoughts: solving elaborate problems with large language models. arXiv preprint arXiv:2308.09687 (2023)
4. Brüggemann, D., Sakaridis, C., Truong, P., Van Gool, L.: 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, pp. 3174–3184 (2023)
5. Chen, L.C., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A.L.: Deeplab: semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs. IEEE Trans. Pattern Anal. Mach. Intell. **40**(4), 834–848 (2017)
6. Chen, Y., Sikka, K., Cogswell, M., Ji, H., Divakaran, A.: Measuring and improving chain-of-thought reasoning in vision-language models. arXiv preprint arXiv:2309.04461 (2023)
7. Chen, Z., et al.: Vision transformer adapter for dense predictions. arXiv preprint arXiv:2205.08534 (2022)
8. Cordts, M., et al.: The cityscapes dataset for semantic urban scene understanding. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3213–3223 (2016)
9. Dai, D., Sakaridis, C., Hecker, S., Van Gool, L.: Curriculum model adaptation with synthetic and real data for semantic foggy scene understanding. Int. J. Comput. Vis. **128**, 1182–1204 (2020)
10. Dai, D., Van Gool, L.: Dark model adaptation: semantic image segmentation from daytime to nighttime. In: 2018 21st International Conference on Intelligent Transportation Systems (ITSC), pp. 3819–3824. IEEE (2018)
11. Darcet, T., Oquab, M., Mairal, J., Bojanowski, P.: Vision transformers need registers. arXiv preprint arXiv:2309.16588 (2023)
12. Dosovitskiy, A., et al.: An image is worth 16×16 words: transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
13. Elsayed, G.F., Goodfellow, I., Sohl-Dickstein, J.: Adversarial reprogramming of neural networks. arXiv preprint arXiv:1806.11146 (2018)
14. Fahes, M., Vu, T.H., Bursuc, A., Pérez, P., De Charette, R.: PODA: prompt-driven zero-shot domain adaptation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 18623–18633 (2023)
15. Gan, Y., et al.: Decorate the newcomers: visual domain prompt for continual test time adaptation. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 37, pp. 7595–7603 (2023)

16. Gao, Y., et al.: Visual prompt tuning for test-time domain adaptation. arXiv preprint arXiv:2210.04831 (2022)
17. Ge, C., et al.: Domain adaptation via prompt learning. IEEE Trans. Neural Netw. Learn. Syst. (2023)
18. Ge, J., Luo, H., Qian, S., Gan, Y., Fu, J., Zhan, S.: Chain of thought prompt tuning in vision language models. arXiv preprint arXiv:2304.07919 (2023)
19. Gong, Z., et al.: Train one, generalize to all: generalizable semantic segmentation from single-scene to all adverse scenes. In: Proceedings of the 31st ACM International Conference on Multimedia, pp. 2275–2284 (2023)
20. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778 (2016)
21. Himakunthala, V., et al.: Let's think frame by frame with VIP: a video infilling and prediction dataset for evaluating video chain-of-thought. In: Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, pp. 204–219 (2023)
22. Houlsby, N., et al.: Parameter-efficient transfer learning for NLP. In: International Conference on Machine Learning, pp. 2790–2799. PMLR (2019)
23. Hoyer, L., Dai, D., Van Gool, L.: 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, pp. 9924–9935 (2022)
24. Hoyer, L., Dai, D., Van Gool, L.: HRDA: context-aware high-resolution domain-adaptive semantic segmentation. In: Avidan, S., Brostow, G., Cissé, M., Farinella, G.M., Hassner, T. (eds.) ECCV 2022. LNCS, vol. 13690, pp. 372–391. Springer, Cham (2022). https://doi.org/10.1007/978-3-031-20056-4_22
25. Hoyer, L., Dai, D., Wang, H., Van Gool, L.: MIC: masked image consistency for context-enhanced domain adaptation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 11721–11732 (2023)
26. Iqbal, J., Hafiz, R., Ali, M.: FogAdapt: self-supervised domain adaptation for semantic segmentation of foggy images. Neurocomputing **501**, 844–856 (2022)
27. Jacovi, A., et al.: A chain-of-thought is as strong as its weakest link: a benchmark for verifiers of reasoning chains. arXiv preprint arXiv:2402.00559 (2024)
28. Jia, M., et al.: Visual prompt tuning. In: Avidan, S., Brostow, G., Cissé, M., Farinella, G.M., Hassner, T. (eds.) ECCV 2022. LNCS, vol. 13693, pp. 709–727. Springer, Cham (2022). https://doi.org/10.1007/978-3-031-19827-4_41
29. Ju, C., Han, T., Zheng, K., Zhang, Y., Xie, W.: Prompting visual-language models for efficient video understanding. In: Avidan, S., Brostow, G., Cissé, M., Farinella, G.M., Hassner, T. (eds.) ECCV 2022. LNCS, vol. 13695, pp. 105–124. Springer, Cham (2022). https://doi.org/10.1007/978-3-031-19833-5_7
30. Kojima, T., Gu, S.S., Reid, M., Matsuo, Y., Iwasawa, Y.: Large language models are zero-shot reasoners. In: Advances in Neural Information Processing Systems, vol. 35, pp. 22199–22213 (2022)
31. Lee, S., Son, T., Kwak, S.: FIFO: learning fog-invariant features for foggy scene segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 18911–18921 (2022)
32. Li, F., et al.: Parsing all adverse scenes: severity-aware semantic segmentation with mask-enhanced cross-domain consistency. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 38, pp. 13483–13491 (2024)

33. Lin, G., Milan, A., Shen, C., Reid, I.: RefineNet: multi-path refinement networks for high-resolution semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1925–1934 (2017)
34. Liu, P., Yuan, W., Fu, J., Jiang, Z., Hayashi, H., Neubig, G.: Pre-train, prompt, and predict: a systematic survey of prompting methods in natural language processing. ACM Comput. Surv. **55**(9), 1–35 (2023)
35. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101 (2017)
36. Ma, X., et al.: Both style and fog matter: cumulative domain adaptation for semantic foggy scene understanding. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 18922–18931 (2022)
37. Mitra, C., Huang, B., Darrell, T., Herzig, R.: Compositional chain-of-thought prompting for large multimodal models. arXiv preprint arXiv:2311.17076 (2023)
38. Radford, A., et al.: Learning transferable visual models from natural language supervision. In: International Conference on Machine Learning, pp. 8748–8763. PMLR (2021)
39. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 10684–10695 (2022)
40. Rose, D., et al.: Visual chain of thought: bridging logical gaps with multimodal infillings. arXiv preprint arXiv:2305.02317 (2023)
41. Sakaridis, C., Dai, D., Gool, L.V.: Guided curriculum model adaptation and uncertainty-aware evaluation for semantic nighttime image segmentation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 7374–7383 (2019)
42. Sakaridis, C., Dai, D., Hecker, S., Van Gool, L.: Model adaptation with synthetic and real data for semantic dense foggy scene understanding. In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (eds.) ECCV 2018. LNCS, vol. 11217, pp. 707–724. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-01261-8_42
43. Sakaridis, C., Dai, D., Van Gool, L.: Semantic foggy scene understanding with synthetic data. Int. J. Comput. Vis. **126**, 973–992 (2018)
44. Sakaridis, C., Dai, D., Van Gool, L.: Map-guided curriculum domain adaptation and uncertainty-aware evaluation for semantic nighttime image segmentation. IEEE Trans. Pattern Anal. Mach. Intell. **44**(6), 3139–3153 (2020)
45. Sakaridis, C., Dai, D., Van Gool, L.: ACDC: the adverse conditions dataset with correspondences for semantic driving scene understanding. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 10765–10775 (2021)
46. Sun, J., et al.: VPA: fully test-time visual prompt adaptation. In: Proceedings of the 31st ACM International Conference on Multimedia, pp. 5796–5806 (2023)
47. Tarvainen, A., Valpola, H.: Mean teachers are better role models: weight-averaged consistency targets improve semi-supervised deep learning results. In: Advances in Neural Information Processing Systems, vol. 30 (2017)
48. Tsai, Y.H., Hung, W.C., Schulter, S., Sohn, K., Yang, M.H., Chandraker, M.: Learning to adapt structured output space for semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 7472–7481 (2018)
49. Uehara, K., et al.: Advancing large multi-modal models with explicit chain-of-reasoning and visual question generation. arXiv preprint arXiv:2401.10005 (2024)
50. Vidit, V., Engilberge, M., Salzmann, M.: Clip the gap: a single domain generalization approach for object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 3219–3229 (2023)

51. Wang, L., Yoon, K.J.: Knowledge distillation and student-teacher learning for visual intelligence: a review and new outlooks. IEEE Trans. Pattern Anal. Mach. Intell. **44**(6), 3048–3068 (2021)
52. Wang, X., et al.: Self-consistency improves chain of thought reasoning in language models. arXiv preprint arXiv:2203.11171 (2022)
53. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: from error visibility to structural similarity. IEEE Trans. Image Process. **13**(4), 600–612 (2004)
54. Wang, Z., et al.: Exploring semantic prompts in the segment anything model for domain adaptation. Remote Sens. **16**(5), 758 (2024)
55. Wei, J., et al.: Chain-of-thought prompting elicits reasoning in large language models. In: Advances in Neural Information Processing Systems, vol. 35, pp. 24824–24837 (2022)
56. Wei, Z., et al.: Stronger fewer & superior: harnessing vision foundation models for domain generalized semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 28619–28630 (2024)
57. Wei, Z., Chen, L., Tu, T., Ling, P., Chen, H., Jin, Y.: Disentangle then parse: night-time semantic segmentation with illumination disentanglement. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), pp. 21593–21603 (2023)
58. Wu, X., Wu, Z., Guo, H., Ju, L., Wang, S.: 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, pp. 15769–15778 (2021)
59. Xiao, A., et al.: 3D semantic segmentation in the wild: learning generalized models for adverse-condition point clouds. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 9382–9392 (2023)
60. Xiao, A., et al.: CAT-SAM: conditional tuning network for few-shot adaptation of segmentation anything model. arXiv preprint arXiv:2402.03631 (2024)
61. Xie, B., Li, S., Li, M., Liu, C.H., Huang, G., Wang, G.: SePiCo: semantic-guided pixel contrast for domain adaptive semantic segmentation. IEEE Trans. Pattern Anal. Mach. Intell. **45**, 9004–9021 (2023)
62. Xie, E., Wang, W., Yu, Z., Anandkumar, A., Alvarez, J.M., Luo, P.: SegFormer: simple and efficient design for semantic segmentation with transformers. In: Advances in Neural Information Processing Systems, vol. 34, pp. 12077–12090 (2021)
63. Yao, S., et al.: Tree of thoughts: deliberate problem solving with large language models. arXiv preprint arXiv:2305.10601 (2023)
64. Yao, Y., Zhang, A., Zhang, Z., Liu, Z., Chua, T.S., Sun, M.: CPT: colorful prompt tuning for pre-trained vision-language models. arXiv preprint arXiv:2109.11797 (2021)
65. Yu, F., et al.: Bdd100k: a diverse driving dataset for heterogeneous multitask learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 2636–2645 (2020)
66. Zhang, J., Wang, B., Li, L., Nakashima, Y., Nagahara, H.: Instruct me more! random prompting for visual in-context learning. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pp. 2597–2606 (2024)
67. Zhang, R., et al.: Tip-adapter: training-free clip-adapter for better vision-language modeling. arXiv preprint arXiv:2111.03930 (2021)

68. Zhong, X., Tu, S., Ma, X., Jiang, K., Huang, W., Wang, Z.: Rainy WCity: a real rainfall dataset with diverse conditions for semantic driving scene understanding. In: IJCAI, pp. 1743–1749 (2022)
69. Zhou, K., Yang, J., Loy, C.C., Liu, Z.: Learning to prompt for vision-language models. Int. J. Comput. Vis. **130**(9), 2337–2348 (2022)