A. Datasets and Adaptation Settings

We evaluate the proposed method, CropAT, on adverse weather adaptation and real-to-artistic adaptation, with dataset details provided below.

Cityscapes → Foggy Cityscapes. Cityscapes [7] collects real-world street scene images under normal weather conditions. It comprises 2,975 and 500 images for training and validation. Foggy Cityscapes [29] is a synthetic dataset derived from Cityscapes, simulating three levels (0.02, 0.01, 0.005) of fog intensity. The most severe level (0.02) split is used as the target domain in our experiment.

Cityscapes → Cityscapes-C. Cityscapes-C [15, 25] is synthetically generated from Cityscapes. It includes 15 types of corruption, each with five levels of severity, to evaluate the impact of various distortions. We select four weather-related corruptions (Brightness, Fog, Frost, and Snow) at their highest severity level as the target domains.

Single-DGOD Dataset. Single-domain generalized object detection (Single-DGOD) dataset [35] is an urban-scene dataset and provides five different weather conditions (Daytime Sunny, Daytime Foggy, Dusk Rainy, Night Sunny, and Night Rainy). In our experiment, the Daytime Sunny scene serves as the source domain, with other scenes used as target domains. Following [35], we employ the test split for evaluation.

PASCAL VOC → Clipart1k. PASCAL VOC [10] provides real-world images across 20 categories of common objects. Clipart1k [18] contains clipart images and shares the same object classes with PASCAL VOC. Following [22, 28, 30], we combine PASCAL VOC 2007 and 2012 with a total of 16,551 images as the source domain and split Clipart1k into training and testing sets, containing 500 images each as the target domain.

B. Implementation Details

We employ InstructPix2Pix [1] as our image editing model. The image classifier free guidance (cfg) scale and text cfg scale are configured to 1.5 and 7.5, respectively. The learnable prompt is optimized in a separate step before UDAOD training, using AdamW [23] with the learning rate of 0.001 and a batch size of 4 over 10,000 iterations. To ensure a fair comparison for UDAOD training, we follow previous UDAOD methods by adopting ResNet-101 [13] as the backbone of Faster R-CNN [27] for the **PASCAL VOC → Clipart1k**. For other experiments, we utilize VGG-16 [31] as the backbone. The Mixup ratio r used in image-level Mixup is sampled from a beta distribution with parameters (0.5, 0.5). For cropped objects, we first uniformly sample $n = 5$ objects and then randomly select one cropped object from each chosen class. The Mixup ratios r' in Equation 9 is set to 0.5. The adversarial loss weight λ_{dis} is set to 0.1. The unsupervised loss weight λ_{unsup} is set to 1.0. The EMA decay rate α is set to 0.9996. The object detector is optimized by Stochastic Gradient Descent (SGD). For the initialization stage, the student model is trained on source data in a supervised manner for 20,000 iterations. Subsequently, the mutual learning stage proceeds for an additional 20,000 iterations. All experiments are conducted on 4 NVIDIA RTX A6000 GPUs and are built upon Detectron2 [36].

C. Analyzing FPP, Recall, and Performance

To better evaluate the effectiveness of our proposed method, we analyze the relationship between False Positive Proportion (FPP), Recall, and Performance (mAP). As shown in Figure S2, our method (CropAT) consistently achieves higher Recall and lower FPP compared to the baseline method AT [22], resulting in improved mAP on the target domain. Additionally, we observe that improvements in Recall naturally introduce some additional false positives. For instance, the model becomes more sensitive to small or ambiguous objects and may attempt to predict hard cases or partially occluded objects based on limited structural cues, as shown in Figure S1. As a result, FPP fluctuates around 14% instead of decreasing monotonically.

D. Image Editing Results

In this supplementary material, we present additional image editing results (Figures S3–S7) and provide the analysis. For side-by-side comparison, each row displays the source image, the edited image without the learnable prompt, the edited image with the learnable prompt, and the corresponding target image.

D.1. More Qualitative Analysis

Cityscapes → Foggy Cityscapes. In the UDAOD setting, using the prompt "an image in the style of fog" without a learnable prompt leads to over-synthesized fog, object occlusion, and undesired artifacts such as hallucinated trees, as shown in Figure S3. The resulting images are also darker than the target, introducing noise during training. In contrast, the learnable prompt mitigates these issues by better capturing fog texture, though it still struggles to model depth-dependent fog density present in the target images.

Cityscapes → Cityscapes-C (Brightness). A prompt "an image in the style of brightness" without learnable prompt is utilized in this UDAOD setting. As illustrated in Figure S4, although all objects are preserved, InstructPix2Pix [1] tends to increase the contrast of the source image and adds a yellowish tint to the bright regions. This makes objects in darker areas appear less visible and results in an overall unnatural appearance. After incorporating a learnable prompt, the brightness, contrast, and color tone become significantly closer to those of the target images.

Cityscapes → Cityscapes-C (Fog). Using the same prompt

Figure S1. **Examples of False Positives.** We visualize the ground truths and pseudo labels generated by CropAT on the Foggy Cityscapes [29] dataset. Bounding boxes are distinguished by color: green indicates ground truths, blue represents correctly predicted pseudo labels, and red denotes false positives. While CropAT demonstrates strong performance overall, it remains prone to errors in challenging scenarios such as distant views, heavily occluded regions, and objects with ambiguous.

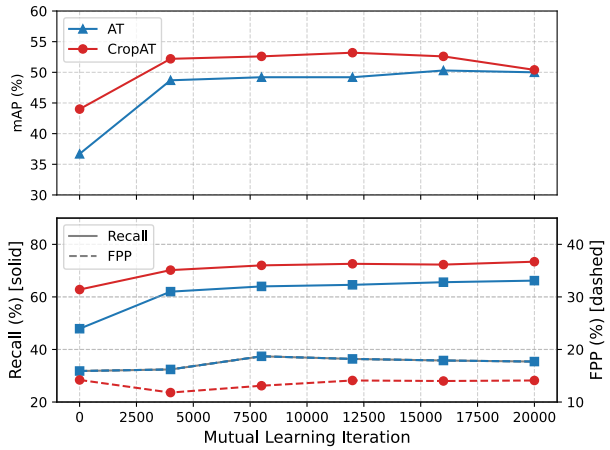

Figure S2. **Comparison of AT and CropAT over Training Iterations on Cityscapes to Foggy Cityscapes Adaptation.** CropAT (Ours) consistently achieves higher recall and lower FPP compared to the baseline AT [22], resulting in improved performance.

as in the Cityscapes → Foggy Cityscapes setting without a learnable prompt causes similar issues, with foreground vehicles and buildings obscured by fog, as shown in Figure S5). Images edited with a learnable prompt improve visibility and adopt a grayish tone closer to the target. However, the learnable prompt still fails to capture fog distribution and texture, which are crucial for preserving scene semantics and improving pseudo-label reliability during adaptation.

Cityscapes → Cityscapes-C (Frost). Without a learnable prompt, using "an image in the style of frost" leads InstructPix2Pix [1] to produce overly bluish images that deviate from the target style, as shown in Figure S6, indicating the limitation of handcrafted prompts. While the learnable prompt captures overall brightness, it fails to model fine-grained frost textures and overlooks substyle variations, which may limit generalization to diverse target appearances.

Cityscapes → Cityscapes-C (Snow). As shown in Figure S7, using InstructPix2Pix [1] with the prompt "an image in the style of snow" results in snow-covered ground, which aligns with the prompt but not the target domain style. This highlights the need for a learnable prompt, which better captures the overall tone but still struggles with fine-grained snow patterns, revealing the difficulty in modeling subtle textures.

D.2. More Quantitative Analysis

We adopt several metrics to measure image similarity, and the results are presented in Table S1. As observed, traditional metrics such as PSNR and SSIM indicate that edited images with a learnable prompt perform better than both source images and edited images without a learnable prompt, suggesting that the learnable prompt effectively captures target style information. For LPIPS, the use of a learnable prompt yields better results compared to not using one. However, compared to source images, the learnable prompt does not achieve higher perceptual similarity. A possible reason is that the learnable prompt fails to capture the fine-grained textures of the target style.

Target Domain	PSNR↑			SSIM↑			LPIPS↓		
	Source images	Edited images w/o a learnable prompt	Edited images w/ a learnable prompt	Source images	Edited images w/o a learnable prompt	Edited images w/ a learnable prompt	Source images	Edited images w/o a learnable prompt	Edited images w/ a learnable prompt
Foggy Cityscapes	11.1	14.6	**16.8**	0.75	0.76	**0.86**	**0.17**	0.54	0.35
Cityscapes-C (Brightness)	7.3	5.7	**17.1**	0.58	0.29	**0.60**	**0.24**	0.56	0.50
Cityscapes-C (Fog)	12.3	11.1	**15.8**	0.69	0.75	**0.85**	0.43	0.56	**0.41**
Cityscapes-C (Frost)	8.9	12.1	**15.9**	0.46	0.49	**0.48**	**0.58**	0.67	0.60
Cityscapes-C (Snow)	7.8	13.1	**15.5**	0.22	0.21	**0.22**	**0.71**	0.73	**0.73**

Table S1. **Image Similarity Results.** We compute the image similarity between source images, edited images with and without a learnable prompt, and each target image separately. Specifically, we average the similarity scores of each paired image in each UDAOD setting. The edited images with a learnable prompt achieve the best results in PSNR and SSIM. However, the learnable prompt enhances the image editing model's ability to generate images with improved perceptual similarity, it still does not surpass the source images in this aspect.

(a) Source image (b) Image edited w/o learnable prompt (c) Image edited w/ learnable prompt (d) Target image

Figure S3. **Image Editing Results for Cityscapes → Foggy Cityscapes**. Without learnable prompt, the generated fog tends to heavily cover most of objects in the scene, or even introduces new objects that are not present in the original images, as shown in **image (b)**. These issues are effectively mitigated through the introduction of a learnable prompt, as illustrated in **image (c)**.

(a) Source image (b) Image edited w/o learnable prompt (c) Image edited w/ learnable prompt (d) Target image

Figure S4. **Image Editing Results for Cityscapes → Cityscapes-C (Brightness)**. A comparison between **images (b)** and **(c)** demonstrates that the introduced learnable prompt can effectively suppress artifacts such as exaggerated contrast and yellowish tints, resulting in a more desirable style.

(a) Source image (b) Image edited w/o learnable prompt (c) Image edited w/ learnable prompt (d) Target image

Figure S5. **Image Editing Results for Cityscapes → Cityscapes-C (Fog)** in Figure S3. As with Cityscapes to Foggy Cityscapes, the learnable prompt resolves fog occlusion but fails to capture fine-grained fog textures in the target images.

(a) Source image (b) Image edited w/o learnable prompt (c) Image edited w/ learnable prompt (d) Target image

Figure S6. **Image Editing Results for Cityscapes → Cityscapes-C (Frost)**. While the learnable prompt helps correct the bluish tone, it is unable to generate the distinctive shapes and textures of frost. Additionally, it tends to produce a single style, lacking the diversity of substyles observed in the target images.

(a) Source image (b) Image edited w/o learnable prompt (c) Image edited w/ learnable prompt (d) Target image

Figure S7. **Image Editing Results for Cityscapes → Cityscapes-C (Snow)**. While handcrafted prompts alone can produce plausible and natural snowy scenes, they do not align with the style of the target images. Introducing a learnable prompt helps adjust the overall style, but it remains unable to generate fine-grained snowflakes as observed in the target domain.