UPRE: Zero-Shot Domain Adaptation for Object Detection via Unified Prompt and Representation Enhancement

Supplementary Material

Appendix

This supplementary material is organized as follows.

- In Section A, we provide additional ablation studies to further analyze the effectiveness of the proposed components in UPRE. In particular, we conduct a comprehensive analysis of each proposed modules, test the performance of various enhancement strategies, evaluate the effectiveness of different prompt priors, and examine the impact of training strategies.
- In Section B, we list the prompt templates used for zero-shot domain adaptation in object detection.
- In Section C, we describe more implementation details of UPRE.
- In Section D, we provide additional qualitative visualization results under cross-city scenarios and virtual-to-real world transitions.

A. Additional Ablation Studies

In this section, we provide additional quantitative experiments to further analyze the effectiveness of components in UPRE.

The Effect of Each Proposed Components. To comprehensively evaluate the effectiveness of our proposed method, we conducted extensive ablation studies across its key components. The ablation study results are systematically organized in Tab. 1, where *Prompt* denotes the proposed domain adaptation prompt, *Enhance* represents unified representation enhancement, *Img* indicates relative domain distance strategy, and *Ins* signifies positive-negative separation strategy. First, we investigate the limitations of addressing only one aspect bias. Rows 1-2 reveal that focusing solely on either detection bias (*Prompt*) or domain bias (*Enhance*) leads to suboptimal performance. For instance, Row 1 achieves an mAP of 37.8 on the Daytime Foggy scenario, while Row 2 achieves 38.5. These results underscore the necessity of jointly addressing both biases for effective adaptation.

	Prompt	*Enhance*	*Img*	*Ins*	Daytime Foggy	Night Clear	Night Rainy	Dusk Rainy
1.	-	✓	✓	✓	37.8	38.3	17.1	32.2
2.	✓	-	✓	✓	38.5	38.7	16.9	32.8
3.	✓	✓	-	✓	39.2	39.8	18.5	33.1
4.	✓	✓	✓	-	38.9	39.5	18.3	32.8
5.	✓	✓	-	-	35.1	36.2	16.7	30.5
6.	-	✓	✓	-	35.6	36.1	16.3	29.9
7.	-	✓	-	✓	35.9	36.5	16.8	30.5
8.	✓	-	-	✓	32.9	34.8	14.2	27.5
9.	✓	-	✓	-	32.8	34.3	14.1	27.2
10.	✓	-	-	-	32.1	34.0	13.5	26.5
11.	✓	✓	✓	✓	**40.0**	**41.5**	**19.8**	**34.5**

Table 1. Ablation study of internal modules. *Prompt* denotes the proposed domain adaptation prompt, *Enhance* represents the unified representation enhancement, *Img* indicates the relative domain distance strategy, and *Ins* signifies the positive-negative separation strategy.

Next, we validate the efficacy of our proposed RDD (relative domain distance) and PND (positive-negative separation) strategies. Rows 3-4 confirm our theoretical analysis, showing significant improvements when these strategies are applied. Specifically, Row 3 achieves an mAP of 39.2 on Daytime Foggy, compared to Row 4's 38.9, demonstrating the complementary benefits of *Img* and *Ins* in enhancing detection performance. We further analyze the impact of static prompts and image-level alignment. Row 6 simulates the approach used in [10], revealing that image-level methods fail to effectively fine-tune CLIP, resulting in degraded performance (e.g., mAP of 35.6 on Daytime Foggy). Similarly, Row 7 highlights the importance of instance-level contextual knowledge, showing that learnable prompts are essential as detection knowledge priors for VLMs (mAP of 35.9).

To approximate the effect of proposal-based training in DetPro [1], we remove *Img* and *Enhance* to train learnable prompts, as shown in Row 10. The results indicate that prompt learning alone is insufficient to overcome domain bias, which achieves an mAP of only 32.1 on Daytime Foggy. In contrast, Row 5 demonstrates that adding trainable prompts to DetPro achieves a consistent mAP improvement of +3.1 across all test scenarios, highlighting the effectiveness of our enhancement approach.

Finally, Rows 8-9 focus on learning image- and instance-level knowledge but fail to handle cross-domain knowledge effectively, achieving mAP values of 32.9 and 32.8, respectively. This limitation underscores the need for our proposed method's comprehensive design, which integrates multiple components to achieve robust zero-shot domain adaptation. The overall effectiveness of our method is demonstrated in Row 11, where all components (*Prompt*, *Enhance*, *Img*, *Ins*) are combined. These results validate the synergy of our proposed components and their ability to address both detection and domain biases effectively.

Choice of Enhancements. One of the key factors of our method is the alteration of feature style through enhancement, enabling the acquisition of pseudo target domain fea-

Size	$\mathcal{E}_\mu$	$\mathcal{E}_\sigma$	Daytime Foggy	Night Clear	Night Rainy	Dusk Rainy
1×1	-	✓	36.0	36.7	16.5	30.9
1×1	✓	✓	36.7	37.7	17.1	31.5
M × N	-	✓	38.0	38.5	18.1	33.3
M × N	✓	✓	**40.0**	**41.5**	**19.8**	**34.5**

Table 2. Ablation study of enhancements. To fit the feature size of $\mathcal{I}$, The sizes of M, N are set to 7, 7 under diverse weather conditions.

Method	Prior	Daytime Foggy	Night Rainy
Gaussian	noise physics	34.5	16.3
CLIP-GAP[10]	static prompt	36.9	18.7
PODA* [3]	static prompt	39.2	19.0
DAI-Net*[2]	darkness physics	36.7	18.9
PDD[7]	static prompt	39.1	19.2
UPRE	unbias prompt	**40.0**	**19.8**

Table 3. Comparison of various prompt priors

Iterative train	Run Steps	Daytime Foggy	Night Clear	Night Rainy	Dusk Rainy
-	-	**40.0**	**41.5**	**19.8**	**34.5**
✓	100	39.5	38.8	18.6	33.9
✓	500	37.1	36.9	17.5	32.5
✓	1000	36.6	36.3	16.9	31.4

Table 4. Training schedule of prompt and enhancement

Methods	Daytime Foggy	Night Clear	Night Rainy	Dusk Rainy
$\mathcal{L}_{ce}$	37.0	37.3	16.5	30.8
$\mathcal{L}_{bg}$ only	36.6	35.7	15.5	29.6
$\mathcal{L}_c$ only	37.6	37.8	17.1	31.4
$\mathcal{L}_{bg} + \mathcal{L}_c$	**38.7**	**38.9**	**17.8**	**32.8**

Table 5. Influence study of the PNS. $\mathcal{L}_{ce}$ is cross-entropy loss.

Method	Prompt Design	Category Space	Label of Negatives	Daytime Foggy	Night Clear	Night Rainy	Dusk Rainy
DetPro	Negative	$\mathcal{C} \cup \mathcal{C}_{bg}$	0 or 1	38.1	38.4	19.0	32.8
DetPro	Shared	$\mathcal{C}$	$\frac{1}{\lvert\mathcal{C}\rvert}$	38.5	37.9	18.8	33.3
PNS (Ours)	Negative	$\mathcal{C} \cup \mathcal{C}_{bg}$	$\frac{1}{\lvert\mathcal{C} \cup \mathcal{C}_{bg}\rvert}$	**40.0**	**41.5**	**19.8**	**34.5**

Table 6. DetPro vs. PNS: comparative analysis and ablation study.

tures. In Tab. 2, we analyze various enhancement selections. Compared to previous method[3] that create pseudo target domain features at image level (Size 1×1), our region-level design achieves superior results, with average 3.2% mAP improvement. Furthermore, compared with only applying $\mathcal{E}_\sigma$ [10] to features, our approach achieves mAP improvements of 0.7% and 2.1% on 1×1 and M × N settings, respectively.

Different Prompt Priors. To investigate the effectiveness of different prompt priors, we studied four prior methods in Tab. 3, including noise physics, static prompt, dark physics, and our proposed unbiased prompt. We use Gaussian noise as the noise physics with our framework. For the static prompt-driven results, we report the performance of CLIP-GAP [10] , PODA*[3] and PDD [7]. Our approach achieves improvement of 4.5% mAP over noise physics methods and 1.2% over static prompt methods. We also report the results of the darkness physics prior method DAI-Net [2]. DAI-Net*, a day-to-night adaptation method based on dark physics prompts, demonstrates excellent performance for Night Clear conditions but performs poorly in other scenarios. In comparison to DAI-Net, our method achieves an average 2.1% mAP gain, demonstrating that our method is applicable to various scenarios.

Training Schedule of Prompt and Enhancement. We investigate the impact of different training strategies, as illustrated in Tab. 4. In the unified prompt and representation enhancement training stage, jointly training prompt representations and enhancement features demonstrates the best performance. As the interval step increases, a noticeable decline in model performance is observed. The unified training of prompt and enhancement representations provides positive interaction. Freezing one of these variables disrupts the unified nature of the training process, resulting in a suboptimal approach that is insufficient for mitigating either detection bias or domain bias.

Influence of Instance-level Enhancement. Tab. 5 highlights the effectiveness of our PNS strategy in improving performance. By separating positive and negative proposals and computing foreground ($\mathcal{L}_c$) and background ($\mathcal{L}_{bg}$) losses independently, our method achieves a substantial mAP gain, e.g., +1.6% on Night Clear. Notably, $\mathcal{L}_{bg}$ performs competitively, reducing mAP by average 1.0% compared to the vanilla cross-entropy loss ($\mathcal{L}_{ce}$). Moreover, $\mathcal{L}_c$ alone achieves comparable performance to the combined loss ($\mathcal{L}_{bg} + \mathcal{L}_c$), underscoring the importance of effectively modeling background context.

Comparison with DetPro. In Table 6, we present ablation studies that highlight differences between PNS and DetPro. As shown in line 1, DetPro struggles with training negative prompt, because it only learns a prompt embedding that draws all negative proposals close with a simple label of 0 or 1 (Detpro's Eq.(8)). Therefore, as depicted in line 2, DetPro opts to train the shared prompt by forcing negative proposals to be equally unlike any **foreground classes** in category space $\mathcal{C}$ (Detpro's Eq.(5)). However, selecting from foreground classes with a probability of $\frac{1}{|\mathcal{C}|}$ overlooks the utilization of the background category.

In contrast to DetPro, PNS trains negative prompt in category space $\mathcal{C} \cup \mathcal{C}_{bg}$, using $\frac{1}{|\mathcal{C} \cup \mathcal{C}_{bg}|}$ for labeling negative proposals in Eq.(12). This approach allows negative proposals to be equally unlike any **classes** in Eq.(11), which aids the negative prompt in learning diverse context, as negative proposals variably encompass either pure backgrounds or parts of objects. Contrasted with DetPro's shared and negative prompts, our method excels in performance across all target domains, achieving notable improvement by 3.6% and 3.1% in Night Clear, respectively. DetPro insufficiently explores negative prompt, while PNS effectively trains the negative prompt. PNS is not a direct application of DetPro; it innovatively explores proposal separation in training neg-

Method	Category	Inference Time	Computation Cost	Total Parameters	Daytime Clear	Daytime Foggy	Night Clear	Night Rainy	Dusk Rainy
Faster-RCNN	Traditional	67ms	183G	52.7M	48.1	32.0	34.4	12.4	26.0
OA-DG		69ms	183G	52.7M	55.8	38.3	38.0	16.8	33.9
DAI-Net*		124ms	297G	78.1M	54.4	36.7	41.0	18.9	33.0
UFR		-	-	-	**58.6**	39.6	40.8	19.2	33.2
CLIP-GAP	VLM-based	98ms	526G	131.4M	51.3	38.5	36.9	18.7	32.3
PODA*		72ms	185G	129.3M	51.8	39.2	38.7	19.0	33.4
PDD		101ms	531G	134.3M	53.6	39.1	38.5	19.2	33.7
UPRE(Ours)	VLM-based	111ms	528G	129.8M	53.9	**40.0**	**41.5**	**19.8**	**34.5**

Table 7. The comparison of efficiency and effectiveness. All test settings are same. CLIP-GAP, PDD and UPRE are all based on the Detectron2 framework, using CLIP's ResNet101 backbone. UFR's - denotes code and model remain unavailable. PODA only use visual encoder during inference.

Daytime Clear to Daytime Foggy:

foggy night
foggy day
foggy evening
foggy dusk
foggy dawn
foggy afternoon
foggy sunset
foggy morning
foggy sunrise
foggy midnight

Daytime Clear to Night Clear:

clear night
clear evening
clear dusk
clear dawn
clear midnight
fine night
night
evening
dusk
dawn
midnight

Daytime Clear to Dusk Rainy:

rainy night
rainy day
rainy evening
rainy dusk
rainy dawn
rainy afternoon
rainy sunset
rainy morning
rainy sunrise
rainy midnight

Daytime Clear to Night Rainy:

rainy night
rainy evening
rainy dusk
rainy dawn
rainy midnight
fine night
night
evening
dusk
dawn
midnight

Virtual to Real World:

foggy day
foggy evening
foggy dusk
foggy dawn
foggy afternoon
foggy sunset
foggy morning
foggy sunrise
rainy day
rainy evening
rainy dusk
rainy dawn
rainy afternoon
rainy sunset
rainy morning
rainy sunrise
clear night
clear evening
fine night
night
evening
dusk
dawn
midnight

Cross-City Scenarios:

city street
highway
residential
parking lot
sidewalk
crosswalk
rainy city street
rainy highway
rainy residential
rainy parking lot
rainy sidewalk
rainy crosswalk
foggy city street
foggy highway
foggy residential
foggy parking lot
foggy sidewalk
foggy crosswalk
dark city street
dark highway
dark residential
dark parking lot
dark sidewalk
dark crosswalk

Figure 1. The domain prompt templates used for zero-shot domain adaptation object detection.

ative prompt. Moreover, proposal separation is a commonly used trick in detection tasks.

Efficiency analysis. Applying advanced large models [8, 9, 11] to ZSDA is promising but inevitably faces higher computational and memory costs. As illustrated in Table 7, UPRE outshines VLM-based methods, notably PDD, with 4.5M fewer parameters and a reduction of 3 GFLOPs in computational cost. Although UPRE is 10ms slower than PDD, this latency is acceptable given its superior performance. To ensure efficiency, we only leverage the MDP module during inference, which uses less than 0.1M parameters.

Evidence of improvement attributed to domain adaptation. To assess whether improvements stem from better domain adaptation [4, 5] or just a stronger detector, we compare different methods on the source domain to directly reflect detector performance. As shown in the Daytime clear column of Table 7, UPRE performs moderately but shows significant improvement in target domains.

B. Prompt Templates

For a fair comparison, we adopt the same prompts used in CLIP [9]. Following previous works [7, 10], we extend the category names and domain characteristics into sentences using multi-view prompts at both the image and instance levels. Specifically, we use ”A photo taken in a

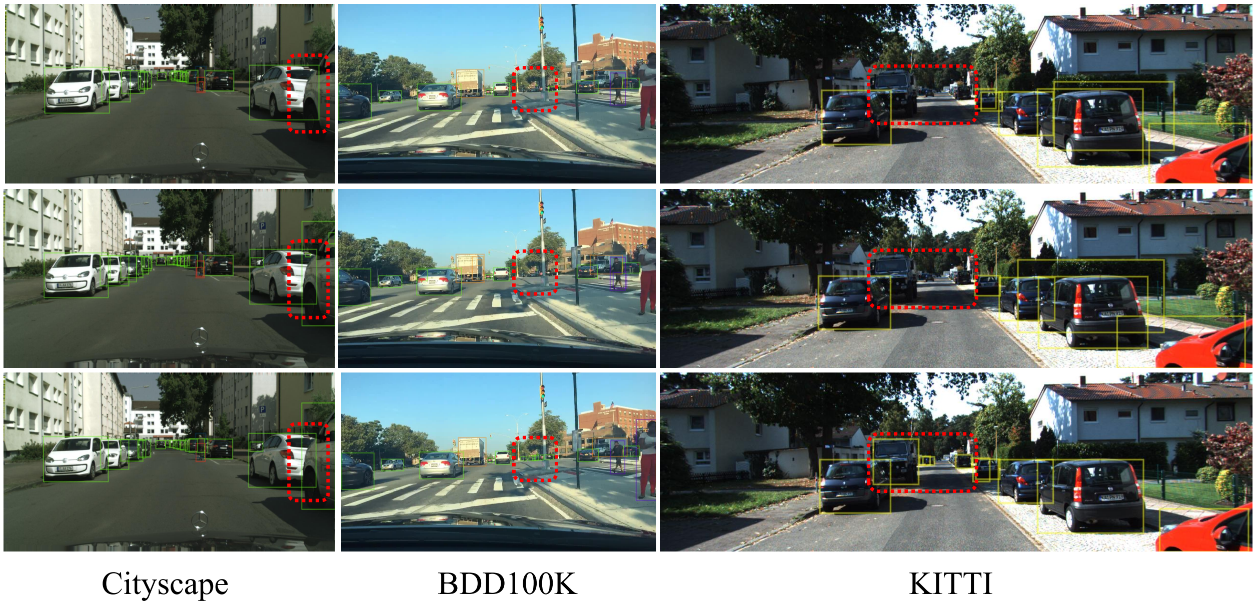

Figure 2. Visualization results under cross-city scenarios. The top to bottom rows show results from CLIP-GAP [10], OA-DG [6], and UPRE, respectively.

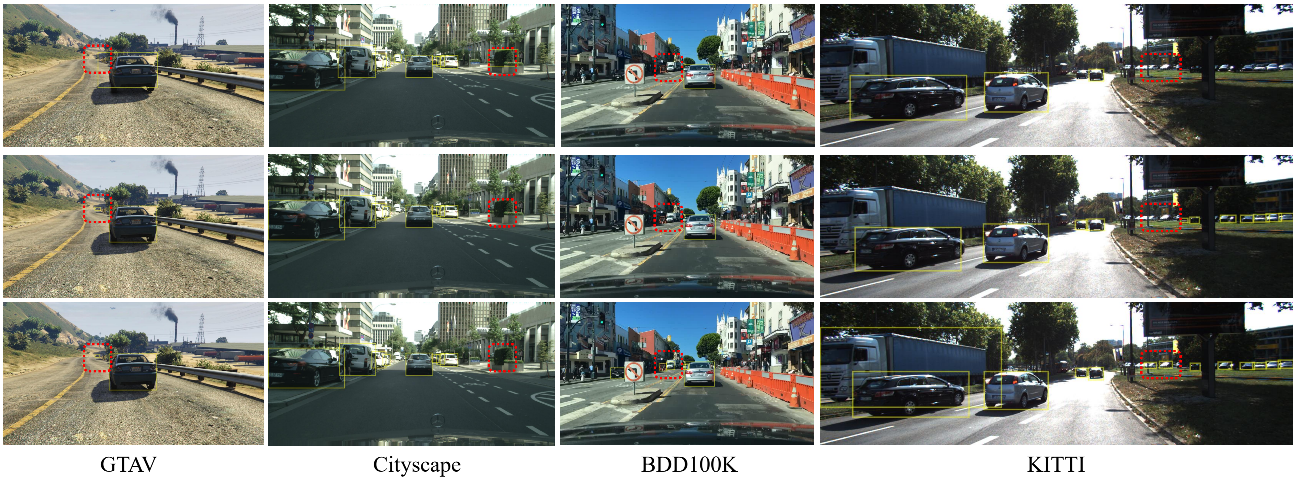

Figure 3. Visualization results of virtual-to-real world transitions. The top to bottom rows show results from CLIP-GAP [10], OA-DG [6], and UPRE, respectively.

$[domain]$." as the image-level prompt input for the text encoder. Then, we define "A $[domain]$ photo of a $[class]$." as the instance-level positive prompt and "A $[domain]$ photo of an $[unknown\ class]$." as the instance-level negative prompt. For the three cross-domain scenarios, we employ 90 $[domain]$-specific prompt templates, as illustrated in Figure 1. We apply L2 normalization to obtain the final multi-view prompt representations.

C. More Implementation Details

In CLIP, the input is a 224×224 image, and the final Attention Pool processes a 7×7 feature map. Current approaches [10] primarily rely on image cropping for data augmentation, resizing images to 224×224. However, this method conflicts with the nature of object detection, as images often contain multiple object instances. To ensure that training images retain more object instances, we use the original 1067×600 images as input. In the relative domain distance strategy, to align the feature size from the third layer of the CLIP image encoder with the input size of the CLIP Attention Pool, we first downsample the third-layer features to 21×21. Subsequently, a 3×3 average pooling layer is applied to produce a 7×7 feature map. In the positive-negative separation strategy, the third-layer features from the CLIP image encoder serve as input to the RPN head. These features are then processed by ROI-Align to extract

14×14 region features. Next, the 14×14 region features are passed through the fourth layer of the image encoder, resulting in 7×7 detection features. Finally, these features are fed into the classification and bounding box regression heads to generate the detection results.

D. Additional Visualizations

In this section, we provide more visualization results under cross-city scenarios and virtual-to-real world transitions. As shown in Fig. 2, our method achieves the best performance, while other methods exhibit issues with both duplicate detections and miss detections, demonstrating the effectiveness of our approach. Compared to cross-city scenarios, the virtual-to-real scenario is more challenging due to the significant domain gap between the synthetic GTAV game world and the real world. Despite this challenge, our method achieves satisfactory detection accuracy compared to other approaches (see Fig. 3). This validates the theory that incorporating real-world weather styles into a clear virtual environment can effectively transform it into a realistic representation under various weather conditions.

References

[1] Yu Du, Fangyun Wei, Zihe Zhang, Miaojing Shi, Yue Gao, and Guoqi Li. Learning to prompt for open-vocabulary object detection with vision-language model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14084–14093, 2022. 1

[2] Zhipeng Du, Miaojing Shi, and Jiankang Deng. Boosting object detection with zero-shot day-night domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12666–12676, 2024. 2

[3] Mohammad Fahes, Tuan-Hung Vu, Andrei Bursuc, Patrick Pérez, and Raoul De Charette. Poda: Prompt-driven zero-shot domain adaptation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 18623–18633, 2023. 2

[4] Boyong He, Yuxiang Ji, Qianwen Ye, Zhuoyue Tan, and Liaoni Wu. Generalized diffusion detector: Mining robust features from diffusion models for domain-generalized detection. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 9921–9932, 2025. 3

[5] Yuxiang Ji, Boyong He, Chenyuan Qu, Zhuoyue Tan, Chuan Qin, and Liaoni Wu. Diffusion features to bridge domain gap for semantic segmentation. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2025. 3

[6] Wooju Lee, Dasol Hong, Hyungtae Lim, and Hyun Myung. Object-aware domain generalization for object detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 2947–2955, 2024. 4

[7] Deng Li, Aming Wu, Yaowei Wang, and Yahong Han. Prompt-driven dynamic object-centric learning for single domain generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17606–17615, 2024. 2, 3

[8] Liunian Harold Li, Pengchuan Zhang, Haotian Zhang, Jianwei Yang, Chunyuan Li, Yiwu Zhong, Lijuan Wang, Lu Yuan, Lei Zhang, Jenq-Neng Hwang, et al. Grounded language-image pre-training. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10965–10975, 2022. 3

[9] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 3

[10] Vidit Vidit, Martin Engilberge, and Mathieu Salzmann. Clip the gap: A single domain generalization approach for object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3219–3229, 2023. 1, 2, 3, 4

[11] Feng Xiong, Hongling Xu, Yifei Wang, Runxi Cheng, Yong Wang, and Xiangxiang Chu. Hs-star: Hierarchical sampling for self-taught reasoners via difficulty estimation and budget reallocation, 2025. 3