Diffusion-based Source-biased Model for Single Domain Generalized Object Detection

Supplementary Material

1. Overview

In the supplementary material, we first show class-wise results on Urban Scene dataset and Real to Artist dataset. (Sec. 2). We then conduct additional ablation studies (Sec. 3) and extend our method to FCOS backbone [35] to further validate the effectiveness of the proposed method (Sec. 4). Consequently, we show model calibration (Sec. 5) and present more detailed designs of our method (Sec. 6). Finally, we provide more visualization results (Sec. 7).

2. The Class-wise Results

2.1. Urban Scene Dataset

Results on Daytime Clear Scene. Table 1 shows the model performance on the source domain. Our method achieves the best results of 60.2 % mAP, outperforming the vanilla FR by 4.0 %. Furthermore, our method improves upon UFR by 1.6%, demonstrating that training with source-style features alone allows the detector to capture domain-specific knowledge, thus providing supplementary information for supervised learning and enhancing the model's performance on the source domain.

Table 1. Quantitative results on the Daytime Clear scene.

Methods	Bus	Bike	Car	Motor	Person	Rider	Truck	mAP
FR [29]	66.9	45.9	69.8	46.5	50.6	49.4	64.0	56.2
SW [29]	62.3	42.9	53.3	49.9	39.2	46.2	60.6	50.6
IBN-Net [26]	63.6	40.7	53.2	45.9	38.6	45.3	60.7	49.7
IterNorm [14]	58.4	34.2	42.4	44.1	31.6	40.8	55.5	43.9
ISW [2]	62.9	44.6	53.5	49.2	39.9	48.3	60.9	51.3
CDSD [41]	68.8	50.9	53.9	56.2	41.8	52.4	68.7	56.1
CLIPGap [38]	55.0	47.8	67.5	46.7	49.4	46.7	54.7	52.5
UFR [4]	66.8	51.0	70.6	55.8	**49.8**	48.5	67.4	58.6
AFDA [3]	-	-	-	-	-	-	-	52.8
Ours	**69.7**	**52.8**	69.4	**56.2**	47.7	**53.9**	**71.4**	**60.2**

Results on Daytime Foggy Scene. Compared to the day-clear scene, objects in foggy scene images appear blurred due to the scattering of light, reducing the images' visibility and detail. As shown in Table 2, the proposed SDG-DiffTection achieves 41.1 % mAP, outperforming previous methods by at least 1.5 %. Additionally, compared to vanilla Faster-RCNN, our method demonstrates a significant improvement in average precision (AP) for the *bike*, *car* and *Rider* categories, achieving gains of 12.8%, 9.2%, and 9.1%, respectively. These results suggest the superiority of our method.

Table 2. Quantitative results on the Daytime Foggy scene.

Methods	Bus	Bike	Car	Motor	Person	Rider	Truck	mAP
FR [29]	34.5	29.6	49.3	26.2	33.0	35.1	26.7	33.5
SW [26]	30.6	36.2	44.6	25.1	30.7	34.6	23.6	30.8
IBN-Net [25]	29.9	26.1	44.5	24.4	26.2	33.5	22.4	29.6
IterNorm [14]	29.7	21.8	42.4	24.4	26	33.3	21.6	28.4
ISW [2]	29.5	26.4	49.2	27.9	30.7	34.8	24.0	31.8
CDSD [41]	32.9	28	48.8	29.8	32.5	38.2	24.1	33.5
CLIPGap [38]	36.2	34.2	57.9	**34.0**	38.7	43.8	25.1	38.5
SRCD [38]	36.4	30.1	52.4	31.3	33.4	40.1	27.7	35.9
PDOC [18]	36.1	34.5	58.4	33.3	40.5	44.2	26.2	39.1
UFR [4]	36.9	35.8	**61.7**	33.7	39.5	42.2	27.5	39.6
AFDA [3]	-	-	-	-	-	-		37.2
Ours	**38.9**	**42.4**	58.5	31.9	**42.7**	**44.2**	**28.9**	**41.1**

Table 3. Quantitative results on the Dusk Rainy scene.

Methods	Bus	Bike	Car	Motor	Person	Rider	Truck	mAP
FR [29]	34.2	21.8	47.9	16.0	22.9	18.5	34.9	28.0
SW [26]	35.2	16.7	50.1	10.4	20.1	13.0	38.8	26.3
IBN-Net [25]	37	14.8	50.3	11.4	17.3	13.3	38.4	26.1
IterNorm [14]	32.9	14.1	38.9	11.0	15.5	11.6	35.7	22.8
ISW [2]	34.7	16.0	50.0	11.1	17.8	12.6	38.8	25.9
CDSD [41]	37.1	19.6	50.9	13.4	19.7	16.3	40.7	28.2
CLIPGap [38]	37.8	22.8	60.7	16.8	26.8	18.7	42.4	32.3
SRCD [38]	39.5	21.4	50.6	11.9	20.1	17.6	40.5	28.8
PDOC [18]	39.4	25.2	60.9	20.4	29.9	16.5	43.9	33.7
UFR [4]	37.1	21.8	**67.9**	16.4	27.4	17.9	43.9	33.2
AFDA [3]	-	-	-	-	-	-	-	38.1
Ours	37.4	**29.7**	67.4	**30.8**	**32.7**	**24.9**	**49.5**	**38.9**

Table 4. Quantitative results on the Night Rainy scene.

Methods	Bus	Bike	Car	Motor	Person	Rider	Truck	mAP
FR [29]	21.3	7.7	28.8	6.1	8.9	10.3	16.0	14.2
SW [26]	22.3	7.8	27.6	0.2	10.3	10.0	17.7	13.7
IBN-Net [25]	24.6	10.0	28.4	0.9	8.3	9.8	18.1	14.3
IterNorm [14]	21.4	6.7	22.0	0.9	9.1	10.6	17.6	12.6
ISW [2]	22.5	11.4	26.9	0.4	9.9	9.8	17.5	14.1
CDSD [41]	24.4	11.6	29.5	9.8	10.5	11.4	19.2	16.6
CLIPGap [38]	28.6	12.1	36.1	9.2	12.3	9.6	22.9	18.7
UFR [4]	29.9	11.8	36.1	9.4	13.1	10.5	23.3	19.2
SRCD [38]	26.5	12.9	32.4	0.8	10.2	12.5	24.0	17.0
PDOC [18]	25.6	12.1	35.8	10.1	14.2	12.9	22.9	19.2
AFDA [3]	-	-	-	-	-	-	-	24.1
Ours	**29.9**	**18.6**	**39.9**	**19.4**	**20.2**	**22.6**	**27.4**	**25.4**

Results on Dusk Rainy Scene. The Dusk Rainy scene is affected by both low light conditions and rainy weather, leading to a substantial domain shift from source daytime

Table 5. Quantitative results on the Clipart scene.

Method	place	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	motor.	person	plant	sheep	sofa	train	tv	mAP
FR [29]	19.9	51.6	17	21.9	27.2	49.6	25.5	9.1	35.1	9.1	25.4	3	29.2	48.9	30.1	40.3	9.1	6.7	35.2	21	25.7
AFDA [3]	34.4	64.4	22.7	**27.0**	45.6	**59.2**	32.9	7.0	46.8	**55.8**	28.9	14.5	44.4	**58.0**	55.2	**52.1**	14.8	38.4	**42.5**	33.9	38.9
Ours	**37.8**	**70.9**	**32.1**	25.4	**55.3**	53.5	**34.3**	**11.8**	**49.4**	53.2	**33.3**	**16.8**	**49.4**	51.3	**56.8**	44.7	**19.2**	**41.7**	40.0	**36.8**	**40.7**

Table 6. Quantitative results on the Night Clear scene.

Methods	Bus	Bike	Car	Motor	Person	Rider	Truck	mAP
FR [29]	43.5	31.2	49.8	17.5	36.3	29.2	43.1	35.8
SW [26]	38.7	29.2	49.8	16.6	31.5	28.0	40.2	33.4
IBN-Net [25]	37.8	27.3	49.6	15.1	29.2	27.1	38.9	32.1
IterNorm [14]	38.5	23.5	38.9	15.8	26.6	25.9	38.1	29.6
ISW [2]	38.5	28.5	49.6	15.4	31.9	27.5	41.3	33.2
CDSD [41]	40.6	35.1	50.7	19.7	34.7	32.1	43.4	36.6
SRCD [28]	43.1	32.5	52.3	20.1	34.8	31.5	42.9	36.7
CLIPGap [38]	37.7	34.3	58.0	19.2	37.6	28.5	42.9	36.9
PDOC [18]	40.9	35.0	59.0	21.3	40.4	29.9	42.9	38.5
UFR [4]	43.6	38.1	**66.1**	14.7	49.1	26.4	47.5	40.8
AFDA [3]	-	-	-	-	-	-	-	42.5
Ours	**45.2**	**40.8**	60.6	**20.9**	**50.2**	**33.4**	**50.3**	**43.1**

clear images. Table 3 presents that our method achieves the best 38.9 % mAP, showing a 10.7% improvement over CDSD. Notably, our method consistently outperforms CDSD across all categories, highlighting that our source-biased object detector can learn more discriminative features.

Results on Night Rainy Scene. Similar to the dusk rainy scene, images captured in night rainy scene also exhibit the combined challenges of low-light conditions and rainy weather, compromising the generalization capability of object detectors. The performance comparison is shown in Table 4. Compared to the two leading methods, AFDA and PDOC, our method achieves improvements of 1.3 % and 6.2 % mAP, respectively. Additionally, we observe that feature normalization-based methods have a poor performance in *motor* category, while our method achieves 19.1 % mAP, demonstrating the effectiveness of our approach in handling challenging weather conditions.

2.2. Real to Artist Dataset

Tables 5, Table 7 and Table 8 present the class-wise results on Pascal VOC to Clipart, Watercolor, and Comic datasets, respectively. Our proposed method consistently outperforms AFDA by at least 1.6 % mAP. Furthermore, on the Comic dataset, our method surpasses AFDA across all categories, demonstrating the effectiveness and robustness of our approach.

Table 7. Quantitative results on the Watercolor scene.

Method	bike	bird	car	cat	dog	person	mAP
FR [29]	85.7	42.5	36.4	29	18.7	54.5	44.5
AFDA [3]	90.4	51.8	51.9	43.9	35.9	**70.2**	57.4
Ours	**95.2**	**56.4**	**57.2**	**45.0**	**44.9**	66.6	**60.9**

Table 8. Quantitative results on the Comic scene.

Method	bike	bird	car	cat	dog	person	mAP
FR	39.7	9.1	23.9	9.1	9.1	22.2	18.9
AFDA [3]	54.1	16.9	30.1	25	27.4	45.9	33.2
Ours	**57.6**	**24.2**	**33.2**	**27.4**	**29.2**	**46.4**	**36.3**

Table 9. Effectiveness of different losses.

$\mathcal{L}_{div}$	$\mathcal{L}_{mat}$	DF	DR	NS	NR	Avg
	✓	39.7	37.8	41.9	23	35.6
✓		40.4	38.2	42.5	24.7	36.5
✓	✓	**41.1**	**38.9**	**43.1**	**25.4**	**37.1**

3. More Ablation Studies

3.1. Effectiveness of Different Losses

As shown in Table 9, we further conduct experiments to demonstrates the effectiveness of $\mathcal{L}_{div}$ and $\mathcal{L}_{mat}$. We observe that without $\mathcal{L}_{div}$, the performance demonstrates a significant drop of 1.5 % mAP, highlighting the importance of storing diverse style information in the proposed memory modules. Additionally, removing $\mathcal{L}_{mat}$ leads to a 0.6 % mAP reduction, which implies the critical role of maintaining vision-language alignment at the pixel level.

3.2. Hyperparameters Analysis

We first analyze the sensitivity of hyperparameters α and β, which control relative importance of reconstruction loss and diversity loss, respectively. As shown in Table 10, Table 11 and Table 12, our model achieves optimal performance when $\alpha = 0.1$, $\beta = 0.1$ and $\gamma = 0.1$. We then explore the effect of memory length M in Table 13. The performance continues to grow until $M = 32$, which suggests that too many elements stored in memory may lead to undesirable redundancy, while too few elements may not adequately represent diverse style information.

Table 10. Hyperparameters analysis of α.

α	DF	DR	NS	NR	Avg
0.01	40.4	38.5	42.4	24.4	36.4
0.05	40.8	38.6	42.7	24.9	36.8
0.1	**41.1**	**38.9**	**43.1**	**25.4**	**37.1**
0.5	40.6	38.8	42.5	24.7	36.7
1.0	40.1	38.1	41.9	23.6	35.9

Table 11. Hyperparameters analysis of β.

β	DF	DR	NS	NR	Avg
0.01	41.0	38.6	42.8	24.9	36.8
0.05	**41.2**	38.7	42.9	25.3	37.0
0.1	41.1	**38.9**	**43.1**	**25.4**	**37.1**
0.5	41.0	38.6	42.9	25.2	36.9
1.0	40.7	38.6	42.6	25.1	36.8

Table 12. Hyperparameters analysis of γ.

	DF	DR	NS	NR	Avg
0.01	41.1	38.5	42.9	24.7	36.8
0.05	41.0	38.8	43.0	25.1	37.0
0.1	**41.1**	**38.9**	**43.1**	**25.4**	**37.1**
0.5	41.3	38.5	42.9	25.3	37.0
1	40.9	38.9	42.2	24.9	36.7

Table 13. Hyperparameters analysis of M.

M	DF	DR	NS	NR	Avg
8	40.3	37.9	42.3	24.7	36.3
16	40.8	38.6	42.7	25.3	36.9
32	**41.1**	**38.9**	**43.1**	**25.4**	**37.1**
64	41.0	38.7	42.9	25.1	36.9

3.3. Effect of CLIP Initialization.

We present the effect of CLIP-initialization on different methods in Table 14. The results show that all methods consistently improve detection performance, highlighting the crucial role of model weight initialization in domain generalization tasks. Furthermore, compared to previous methods, our approach achieves the best performance with 40.9 % mAP, demonstrating the effectiveness of our diffusion-based framework combined with CLIP-initialized features.

3.4. Effect of Different Noise.

We further give a detailed analysis of different noise in Table 8. All augmented feature-based noise consistently outperforms Gaussian noise, enabling the diffusion model to fully understand practical distribution differences. Additionally, we explore some other common data augmentation methods, such as normalization perturbation (NP) [7] in feature space and image corruption [3] in input space, as alternative noise sources, which lead to 0.8 % and 0.5 % performance degradation, indicating the effectiveness of our proposed augmented method. Note that compared to Table 5 in our paper, we use these augmented features as noise but not directly input into denoising step. The performance gap the diffusion process can help enhance data diversity.

Table 14. Effect of external CLIP Initialization. All results are reproduced by official code.

Methods	C	DF	DR	NC	NR	Avg
FR	✓	36.5	29.4	36.3	16.8	29.8
OA-Mix [17]	✓	39.5	35.6	39.1	18.4	33.2
AFDA [3]	✓	39.8	39.9	43.6	25.8	37.1
Ours	✓	**43.2**	**40.9**	**44.7**	**26.6**	**38.9**

Table 15. Effect of using different augmentation methods as noise.

	DF	DR	NS	NR	Avg
Gaussian Noise	38.9	35.8	39.7	22.1	34.1
NP [7]	40.6	38.7	42.5	25.5	36.8
Image Corruption [3]	40.9	38.2	42.9	25.1	36.8
Ours	**41.1**	**38.9**	**43.1**	**25.4**	**37.1**

4. Extend to FCOS

To further prove the effectiveness of our proposed method, we extend our method to a one-stage detector, namely FCOS [35]. Following [3], we adopt FCOS with ResNet50 as the feature backbone, and the results are shown in Table 16 Compared to the baseline FCOS, our method achieves performance gains of 12.8 %, 15.5 %, 14.4 %, 9.9 % and 7.2 % on Daytime Clear, Daytime Foggy, Dusk Rainy, Night Sunny and Night Rainy scenes, respectively.

Table 16. Results of our method with FCOS baseline on the Urban Scene Detection dataset, where our model is trained on Daytime Clear (DC) and tested on Daytime Foggy (DF), Dusk Rainy (DR), Night Sunny (NS) and Night Rainy (NR).

	DC	DF	DR	NS	NR
FCOS	42.1	24.2	23.9	28.6	14.1
Ours	**54.9**	**39.7**	**38.3**	**38.5**	**21.3**

5. Model Calibration

Figure 1 shows the calibration curves of Faster R-CNN and our proposed method across multiple unseen domains. We observe that our calibration curves are closer to the diagonal line, indicating that our method achieves better model calibration and more reliable model predictions.

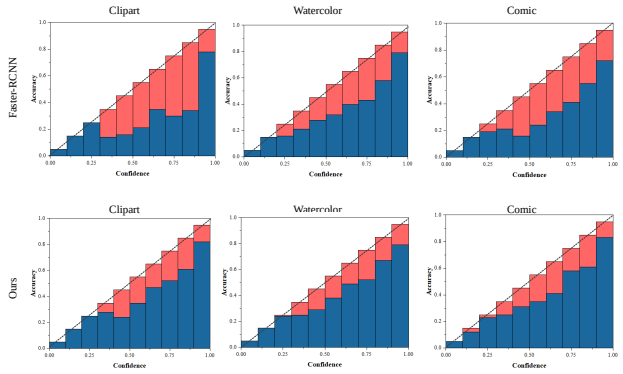

Figure 1. Reliability Diagram for different target domains.

6. The Designs of the Proposed Method

To use a diffusion model in latent space, we adopt a U-Net structure [32] as shown in Figure 2. We simply adopt linear attention to interact input features and condition embeddings to reduce the calculation cost. Besides, we take mean memory as an example to illustrate the memory update process in Figure 3.

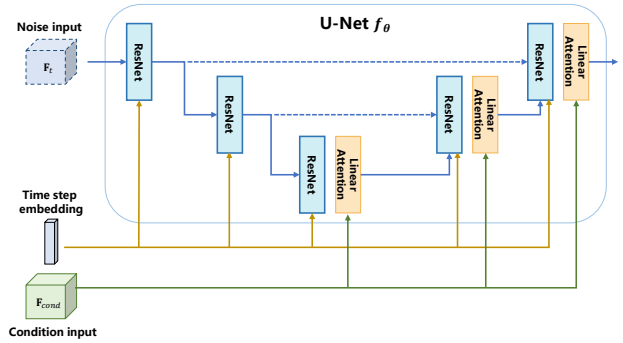

Figure 2. The detailed structure of U-Net.

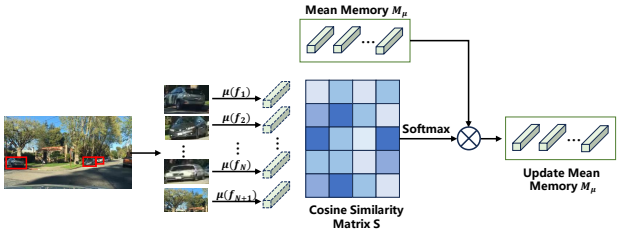

Figure 3. Illustration of updating mean memory module.

7. More Visualization

We provide a visualization of the detection results obtained by CLIPGap [38] and our method on five weather conditions in Figure 4, Figure 5 and Figure 6. Specifically, as shown in the third column of Figure 4, we observe that our method provides accurate label for the *Truck*, while CLIP-Gap misclassifies it as *Car*. This suggests that our method can generate more discriminative features, effectively reducing the false positives caused by misclassification. Furthermore, our approach demonstrates superior capability in distinguishing foreground from background regions, particularly under challenging weather conditions (e.g., the first column in Figure 4(a)).

Figure 4. Visualization of detection result on Daytime Clear. Top row: The predictions of CLIPGap [38]. Bottom row: The predictions of our proposed method.

Figure 5. Visualization of detection result on (a) Daytime Foggy and (b) Dusk Rainy. Top row: The predictions of CLIPGap [38]. Bottom row: The predictions of our proposed method.

Figure 6. Visualization of detection result on (a) Night Rainy and (b) Night Sunny. Top row: The predictions of CLIPGap [38]. Bottom row: The predictions of our proposed method.