Generalized Diffusion Detector: Mining Robust Features from Diffusion Models for Domain-Generalized Detection

Supplementary Material

A. Overview of Supplementary Material

This supplementary material provides additional experimental results, implementation details, and analysis to support our main paper. The contents are organized as follows:

- **Additional Description for Methods**: Detailed motivations and derivations for our proposed approach, including diffusion features for DG detection, two-level guidance framework, feature alignment, and object-level alignment mechanism (in Sec. B).
- **Class-wise Results Analysis**: Detailed per-category performance analysis on Real to Artistic and Diverse Weather benchmarks (in Sec. C).
- **Additional Results of Different SD Versions**: Comprehensive comparison and analysis of different Stable Diffusion versions (SD-1.5, SD-2.1, SD-3-M) across various benchmarks (in Sec. D).
- **Additional Analysis for Results**: In-depth analysis of domain distribution differences and confusion matrix patterns to validate our method's effectiveness (in Sec. E).
- **Visualization of Detection Results**: Qualitative results demonstrating our method's superior performance across different domain generalization scenarios (in Sec. F).

List of Tables:

List of Figures:

B. Additional Description for Methods

B.1. Motivations

Diffusion features for DG detection. Domain generalization for object detection requires learning domain-invariant representations without accessing target domain data, which remains challenging due to complex real-world variations. While existing detectors struggle under domain shifts [36], diffusion models have demonstrated unique advantages in handling diverse variations through their progressive denoising process. These models naturally distinguish intrinsic semantic structures from domain-specific variations [21, 28], building robustness against various perturbations. We leverage these properties for domain-generalized detection: the denoising mechanism filters out domain-specific variations while preserving essential object characteristics [11], and the multi-scale features provide robust semantic representations that generalize across domains.

Two-level guidance from frozen diffusion detector. While directly using diffusion features provides strong generalization capability, it incurs substantial computational overhead. This motivates us to transfer the generalization ability from a frozen diffusion detector to lightweight detectors. We propose a two-level guidance framework to capture both semantic understanding and detection knowledge. At the feature level, we align global feature distributions between diffusion and conventional detectors to learn domain-invariant representations. At the object level, we facilitate task-specific knowledge transfer through shared detection heads following [30], enabling precise localization and classification learning from the diffusion teacher.

Feature alignment for heterogeneous detectors. Direct feature alignment with MSE leads to suboptimal results due to different magnitude distributions and feature dominance issues [2]. We leverage Pearson Correlation Coefficient (PCC) for feature alignment, which captures relational patterns while being invariant to magnitude differences. By normalizing features before alignment, PCC effectively handles discrepancies between diffusion and conventional detectors, enabling stable knowledge transfer between heterogeneous detector pairs.

Object-level alignment for task-specific knowledge transfer. While feature-level alignment helps learn domain-invariant representations, detection-specific knowledge transfer remains challenging due to architectural differences. We propose an object-level alignment scheme that shares detection heads between student and teacher [30], providing task-oriented supervision through classification and regression branches. This complementary guidance enables effective knowledge transfer from the diffusion teacher to conventional detectors.

B.2. Detailed Derivations and Analysis

Relationship between KL divergence, MSE and PCC: Let $\mathcal{M}^l_{\text{comm}}$ and $\mathcal{M}^l_{\text{diff}}$ denote the l-th layer feature maps from student and teacher networks respectively. After standardization, we obtain their normalized versions $\hat{\mathcal{M}}^l_{\text{comm}}$ and $\hat{\mathcal{M}}^l_{\text{diff}}$ with zero mean and unit variance:

$$\begin{aligned} \mathbb{E}[\hat{\mathcal{M}}^l_{\text{comm}}] &= \mathbb{E}[\hat{\mathcal{M}}^l_{\text{diff}}] = 0, \\ \text{Var}[\hat{\mathcal{M}}^l_{\text{comm}}] &= \text{Var}[\hat{\mathcal{M}}^l_{\text{diff}}] = 1 \end{aligned} \tag{1}$$

According to [2], when using KL divergence with temperature scaling for feature distillation, in the high-temperature limit ($T \to \infty$), the gradient of KL divergence between normalized features can be approximated as:

$$\frac{\partial \mathcal{L}_{KL}}{\partial \hat{\mathcal{M}}^l_{\text{comm}}} \approx \frac{1}{N_l}(\hat{\mathcal{M}}^l_{\text{comm}} - \hat{\mathcal{M}}^l_{\text{diff}}) \tag{2}$$

This is proportional to the gradient of MSE loss. For standardized features, their Pearson correlation coefficient (PCC) simplifies to their covariance:

$$\begin{aligned} r_l &= \frac{\text{Cov}(\hat{\mathcal{M}}^l_{\text{comm}}, \hat{\mathcal{M}}^l_{\text{diff}})}{\sqrt{\text{Var}[\hat{\mathcal{M}}^l_{\text{comm}}]\text{Var}[\hat{\mathcal{M}}^l_{\text{diff}}]}} \\ &= \mathbb{E}[\hat{\mathcal{M}}^l_{\text{comm}} \hat{\mathcal{M}}^l_{\text{diff}}] \end{aligned} \tag{3}$$

The feature alignment loss can be expressed in terms of PCC:

$$\begin{aligned} \mathcal{L}_{\text{align}} &= \sum_{l=1}^{L} \frac{1}{N_l} \|\hat{\mathcal{M}}^l_{\text{comm}} - \hat{\mathcal{M}}^l_{\text{diff}}\|_2^2 \\ &= 2\sum_{l=1}^{L} \frac{1}{N_l}(1 - r_l) \end{aligned} \tag{4}$$

Therefore, assuming positive correlation ($-1 < r_l < 1$), we have established the following equivalence chain:

$$\begin{aligned} \min \mathcal{L}_{KL}(P^l_{\text{comm}} \| P^l_{\text{diff}}) &\Leftrightarrow \max r_l \\ &\Leftrightarrow \min \|\hat{\mathcal{M}}^l_{\text{comm}} - \hat{\mathcal{M}}^l_{\text{diff}}\|_2^2 \end{aligned} \tag{5}$$

This equivalence chain demonstrates that through standardization, minimizing KL divergence, maximizing PCC, and minimizing MSE become equivalent objectives, providing both computational simplicity and theoretical guarantees for feature alignment.

Details of Classification Knowledge Transfer: To effectively transfer classification knowledge between diffusion and common detectors, we employ knowledge distillation with temperature scaling. Given the logits $\mathbf{z}^i_{\text{diff}}, \mathbf{z}^i_{\text{comm}} \in \mathbb{R}^{C+1}$ from both feature sources for the i-th proposal, where C is the number of object categories, we first convert them to probability distributions.

The temperature-scaled softmax converts logits to probabilities as:

$$\mathbf{P}^i_{\text{cat}} = \text{softmax}(\mathbf{z}^i_{\text{diff}}/\tau), \quad \mathbf{Q}^i_{\text{cat}} = \text{softmax}(\mathbf{z}^i_{\text{comm}}/\tau) \tag{6}$$

where $\mathbf{P}^i_{\text{cat}}, \mathbf{Q}^i_{\text{cat}} \in \mathbb{R}^{C+1}$ represent the predicted probability distributions over all classes including background, and τ is the temperature parameter that produces softer distributions.

The knowledge distillation loss is computed using KL divergence between these distributions:

$$\mathcal{L}_{\text{cls}} = \frac{1}{N}\sum_{i=1}^{N} \tau^2 D_{KL}(\mathbf{Q}^i_{\text{cat}} \| \mathbf{P}^i_{\text{cat}}) \tag{7}$$

where N is the total number of proposals and:

$$D_{KL}(\mathbf{Q}^i_{\text{cat}} \| \mathbf{P}^i_{\text{cat}}) = \sum_{c=1}^{C+1} Q^i_c \log \frac{Q^i_c}{P^i_c} \tag{8}$$

with Q^i_c and P^i_c denoting the predicted probabilities for class c.

C. Classwise Results

Results on Real to Artistic: As shown in Tab. 4, 1, and 2, our diffusion detector demonstrates remarkable generalization capability on artistic-style datasets, surpassing both DG and DA methods significantly. On Clipart, our diffusion detector achieves 58.3% mAP, leading to substantial improvements of 9.0% and 19.4% over the previous best DA method AT [18] and DG method DivAlign [8], respectively. On Comic dataset, our method reaches 51.9% mAP, exhibiting clear advantages compared to the best DA approach D-ADAPT [13] at 40.5% and DG method DivAlign [8] at 33.2%. For Watercolor, we achieve 68.4% mAP, which significantly surpasses the previous best results of 59.9% from AT [18] and 57.4% from DivAlign [8].

However, the diffusion-guided detector shows limited success in bridging extreme domain gaps. On Clipart, Comic, and Watercolor, the diffusion-guided detector (40.8%, 29.7%, and 54.2% respectively) underperforms compared to both recent DG methods (DivAlign: 38.9%, 33.2%, and 57.4%) and DA approaches (AT and D-ADAPT: 49.3%, 40.5%, and 59.9%). While the improvements over baseline remain notable (+13.6%, +11.6%, and +12.7% respectively), the performance gap suggests that transferring the strong generalization capability from diffusion models to conventional detectors remains challenging when facing significant stylistic variations, likely due to the extreme domain shifts in artistic styles that make feature alignment particularly difficult.

Results on Diverse Weather benchmark: As shown in Tab. 3, our method demonstrates strong robustness across various weather and lighting conditions. For Daytime-Foggy scenarios, our diffusion guided detector achieves 44.7% mAP, exceeding the previous best result from UFR [20] by 5.1%. In Night-Sunny conditions, we obtain 49.1% mAP, surpassing G-NAS [33] which achieves 45.0%. The improvement becomes more pronounced in challenging Night-Rainy scenarios, where our diffusion detector reaches 27.8% mAP, considerably outperforming the previous best of 24.1% from DivAlign [8]. Under Dusk-Rainy conditions, we achieve 42.5% mAP, marking a clear advancement over DivAlign [8] at 38.1%. Most notably, the diffusion-guided detector demonstrates consistent improvements over the baseline across all four scenarios, with remarkable margins of +15.9%, +17.7%, +9.3%, and +13.3%. These comprehensive results not only validate the effectiveness of our knowledge transfer framework in handling natural environmental variations but also confirm our approach's strong capability in enhancing detection generalization across diverse real-world conditions.

Table 1. Real to Artistic DG and DA Results (%) on Comic (Classwise).

Methods	Bike	Bird	Car	Cat	Dog	Person	mAP
***DG methods** (without target data)*							
Div. [8] *(CVPR'24)*	41.7	12.3	29.0	13.2	20.6	36.5	25.5
DivAlign [8] *(CVPR'24)*	54.1	16.9	30.1	25.0	27.4	45.9	33.2
***DA methods** (with unlabeled target data)*							
DA-Faster [5] *(CVPR'18)*	31.1	10.3	15.5	12.4	19.3	39.0	21.2
SWDA [27] *(CVPR'19)*	36.4	21.8	29.8	15.1	23.5	49.6	29.4
STABR [14] *(ICCV'19)*	50.6	13.6	31.0	7.5	16.4	41.4	26.8
MCRA [35] *(ECCV'20)*	47.9	20.5	37.4	20.6	24.5	50.2	33.5
I3Net [4] *(CVPR'21)*	47.5	19.9	33.2	11.4	19.4	49.1	30.1
DBGL [3] *(ICCV'21)*	35.6	20.3	33.9	16.4	26.6	45.3	29.7
D-ADAPT [13] *(ICLR'22)*	52.4	25.4	42.3	**43.7**	25.7	53.5	40.5
Ours (DG settings)							
Diff. Detector (SD-1.5)	**63.3**	**41.7**	**58.2**	31.8	**40.9**	**75.3**	**51.9**
Diff. Detector (SD-2.1)	61.1	35.7	53.6	23.2	35.0	71.2	46.6
Diff. Guided (SD-1.5)	47.6	21.0	35.3	9.1	21.6	43.5	29.7+11.6
Diff. Guided (SD-2.1)	46.4	13.2	24.2	7.5	12.3	35.8	24.9+6.8

Table 2. Real to Artistic DG and DA Results (%) on Watercolor (Classwise).

Methods	Bike	Bird	Car	Cat	Dog	Person	mAP
***DG methods** (without target data)*							
Div. [8] *(CVPR'24)*	87.1	51.7	53.6	35.1	23.6	63.6	52.5
DivAlign [8] *(CVPR'24)*	90.4	51.8	51.9	43.9	35.9	70.2	57.4
***DA methods** (with unlabeled target data)*							
SWDA [27] *(CVPR'19)*	82.3	55.9	46.5	32.7	35.5	66.7	53.3
MCRA [35] *(ECCV'20)*	87.9	52.1	51.8	41.6	33.8	68.8	56.0
UMT [9] *(CVPR'21)*	88.2	55.3	51.7	39.8	43.6	69.9	58.1
IIOD [32] *(TPAMI'21)*	95.8	54.3	48.3	42.4	35.1	65.8	56.9
I3Net [4] *(CVPR'21)*	81.1	49.3	46.2	35.0	31.9	65.7	51.5
SADA [6] *(IJCV'21)*	82.9	54.6	52.3	40.5	37.7	68.2	56.0
CDG [16] *(CVPR'19)*	97.7	53.1	52.1	47.3	38.7	68.9	59.7
VDD [27] *(AAAI'21)*	90.0	56.6	49.2	39.5	38.8	65.3	56.6
DBGL [3] *(ICCV'21)*	83.1	49.3	50.6	39.8	38.7	61.3	53.8
AT [18] *(CVPR'22)*	93.6	56.1	58.9	37.3	39.6	73.8	59.9
LODS [17] *(CVPR'22)*	95.2	53.1	46.9	37.2	47.6	69.3	58.2
Ours (DG settings)							
Diff. Detector (SD-1.5)	**99.8**	**70.3**	**57.5**	**49.8**	**51.0**	**82.0**	**68.4**
Diff. Detector (SD-2.1)	91.1	65.9	55.7	47.6	39.1	73.4	62.1
Diff. Guided (SD-1.5)	90.1	51.0	48.5	40.2	28.9	66.7	54.2+12.7
Diff. Guided (SD-2.1)	99.6	48.4	49.1	28.4	23.4	54.2	50.6+9.1

Table 3. Generalization detection Results (%) on Diverse Weather benchmark (Classwise).

	Daytime-Foggy								Night-Sunny							
Methods	**Bus**	**Bike**	**Car**	**Motor**	**Person**	**Rider**	**Truck**	**mAP**	**Bus**	**Bike**	**Car**	**Motor**	**Person**	**Rider**	**Truck**	**mAP**
IBN-Net [23] *(CVPR'18)*	29.9	26.1	44.5	24.4	26.2	33.5	22.4	29.6	37.8	27.3	49.6	15.1	29.2	27.1	38.9	32.1
SW [24] *(ICCV'19)*	30.6	26.2	44.6	25.1	30.7	34.6	23.6	30.8	38.7	29.2	49.8	16.6	31.5	28.0	40.2	33.4
IterNorm [12] *(CVPR'19)*	29.7	21.8	42.4	24.4	26.0	33.3	21.6	28.5	38.5	23.5	38.9	15.8	26.6	25.9	38.1	29.6
ISW [7] *(CVPR'21)*	29.5	26.4	49.2	27.9	30.7	34.8	24.0	31.8	38.5	28.5	49.6	15.4	31.9	27.5	41.3	33.2
CDSD [31] *(CVPR'22)*	32.9	28.0	48.8	29.8	32.5	38.2	24.1	33.5	40.6	35.1	50.7	19.7	34.7	32.1	43.4	36.6
CLIPGap [29] *(CVPR'23)*	36.1	34.3	58.0	33.1	39.0	43.9	25.1	38.5	37.7	34.3	58.0	19.2	37.6	28.5	42.9	36.9
SRCD [25] *(TNNLS'24)*	36.4	30.1	52.4	31.3	33.4	40.1	27.7	35.9	43.1	32.5	52.3	20.1	34.8	31.5	42.9	36.7
G-NAS [33] *(AAAI'24)*	32.4	31.2	57.7	31.9	38.6	38.5	24.5	36.4	46.9	40.5	67.5	26.5	50.7	35.4	47.8	45.0
OA-DG [15] *(AAAI'24)*	-	-	-	-	-	-	-	38.3	-	-	-	-	-	-	-	38.0
DivAlign [8] *(CVPR'24)*	-	-	-	-	-	-	-	37.2	-	-	-	-	-	-	-	42.5
UFR [20] *(CVPR'24)*	36.9	35.8	61.7	33.7	39.5	42.2	27.5	39.6	43.6	38.1	66.1	14.7	49.1	26.4	47.5	40.8
Diff. Detector (SD-1.5)	37.5	32.4	67.9	35.6	48.3	44.6	**37.1**	43.3	49.6	42.1	70.5	21.4	54.5	38.2	52.6	47.0
Diff. Detector (SD-2.1)	36.4	**36.7**	68.8	36.6	**51.5**	49.1	32.9	44.6	48.2	39.6	69.2	22.8	55.4	37.7	51.6	46.4
Diff. Guided (SD-1.5)	**39.3**	35.8	**69.4**	**37.7**	48.8	**49.7**	32.3	44.7+15.9	51.0	42.8	72.2	27.5	55.9	39.5	52.0	48.6+17.2
Diff. Guided (SD-2.1)	38.8	36.4	68.9	37.4	48.6	49.6	33.4	**44.7**+15.9	**51.3**	**43.6**	**72.3**	**27.6**	**56.2**	**40.2**	**53.7**	**49.1**+17.7
	Night-Rainy								Dusk-Rainy							
IBN-Net [23] *(CVPR'18)*	24.6	10.0	28.4	0.9	8.3	9.8	18.1	14.3	37.0	14.8	50.3	11.4	17.3	13.3	38.4	26.1
SW [24] *(ICCV'19)*	22.3	7.8	27.6	0.2	10.3	10.0	17.7	13.7	35.2	16.7	50.1	10.4	20.1	13.0	38.8	26.3
IterNorm [12] *(CVPR'19)*	21.4	6.7	22.0	0.9	9.1	10.6	17.6	12.6	32.9	14.1	38.9	11.0	15.5	11.6	35.7	22.8
ISW [7] *(CVPR'21)*	22.5	11.4	26.9	0.4	9.9	9.8	17.5	14.1	34.7	16.0	50.0	11.1	17.8	12.6	38.8	25.9
CDSD [31] *(CVPR'22)*	24.4	11.6	29.5	0.4	10.5	11.4	19.2	15.3	37.1	19.6	50.9	13.4	19.7	16.3	40.7	28.2
CLIPGap [29] *(CVPR'23)*	28.6	12.1	36.1	9.2	12.3	9.6	22.9	18.7	37.8	22.8	60.7	16.8	26.8	18.7	42.4	32.3
SRCD [25] *(TNNLS'24)*	26.5	12.9	32.4	0.8	10.2	12.5	24.0	17.0	39.5	21.4	50.6	11.9	20.1	17.6	40.5	28.8
G-NAS [33] *(AAAI'24)*	28.6	9.8	38.4	0.1	13.8	9.8	21.4	17.4	44.6	22.3	66.4	14.7	32.1	19.6	45.8	35.1
OA-DG [15] *(AAAI'24)*	-	-	-	-	-	-	-	16.8	-	-	-	-	-	-	-	33.9
DivAlign [8] *(CVPR'24)*	-	-	-	-	-	-	-	24.1	-	-	-	-	-	-	-	38.1
UFR [20] *(CVPR'24)*	29.9	11.8	36.1	9.4	13.1	10.5	23.3	19.2	37.1	21.8	67.9	16.4	27.4	17.9	43.9	33.2
Diff. Detector (SD-1.5)	**42.0**	**15.0**	**53.6**	6.5	**26.2**	**13.8**	**37.5**	**27.8**	**49.7**	27.9	**74.9**	18.2	**45.5**	**24.5**	**56.8**	**42.5**
Diff. Detector (SD-2.1)	30.1	11.3	46.1	**10.2**	24.1	9.2	31.5	23.2	44.6	**30.6**	73.5	**22.1**	44.4	20.1	55.6	41.6
Diff. Guided (SD-1.5)	35.4	12.7	46.2	3.2	13.8	10.7	29.7	21.7+9.3	43.1	23.9	73.6	13.4	33.2	22.1	52.3	37.4+13.3
Diff. Guided (SD-2.1)	34.4	7.8	43.3	2.2	14.3	7.5	30.3	20.8+8.4	44.6	22.5	73.1	15.7	31.7	19.3	52.6	37.3+13.2

Table 4. Real to Artistic DG and DA Results (%) on Clipart (Classwise).

Methods	**aero.**	**bike**	**bird**	**boat**	**bottle**	**bus**	**car**	**cat**	**chair**	**cow**	**table**	**dog**	**horse**	**bike**	**psn.**	**plant.**	**sheep**	**sofa**	**train**	**tv**	**mAP**
DG methods *(without target data)*																					
Div. [8] *(CVPR'24)*	29.3	50.9	23.4	35.3	45.3	49.8	33.4	10.6	43.3	22.3	31.6	4.5	32.9	51.9	40.2	51.1	18.2	29.6	42.3	28.5	33.7
DivAlign [8] *(CVPR'24)*	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
DA methods *(with unlabeled target data)*																					
AT [18] *(CVPR'22)*	33.8	60.9	38.6	49.4	52.4	53.9	56.7	7.5	52.8	**63.5**	34.0	25.0	62.2	72.1	77.2	57.7	27.2	52.0	55.7	54.1	49.3
D-ADAPT [13] *(ICLR'22)*	56.4	63.2	42.3	40.9	45.3	77.0	48.7	25.4	44.3	58.4	31.4	24.5	47.1	75.3	69.3	43.5	27.9	34.1	60.7	**64.0**	49.0
TIA [34] *(CVPR'22)*	42.2	66.0	36.9	37.3	43.7	71.8	49.7	18.2	44.9	58.9	18.2	29.1	40.7	87.8	67.4	49.7	27.4	27.8	57.1	50.6	46.3
CIGAR [19] *(CVPR'23)*	35.2	55.0	39.2	30.7	**60.1**	58.1	46.9	**31.8**	47.0	61.0	21.8	26.7	44.6	52.4	68.5	54.4	**31.3**	38.8	56.5	63.5	46.2
CMT [1] *(CVPR'23)*	39.8	56.3	38.7	39.7	60.0	35.0	56.0	7.1	60.1	60.4	35.8	28.1	**67.8**	84.5	80.1	55.5	20.3	32.8	42.3	38.2	47.0
Ours (DG settings)																					
Diff. Detector (SD-1.5)	63.7	**86.1**	**49.8**	56.5	52.9	50.9	**67.3**	19.7	**74.7**	34.3	**57.7**	**41.9**	63.2	**89.4**	**89.6**	59.8	23.5	**64.9**	**65.9**	55.2	**58.3**
Diff. Detector (SD-2.1)	**65.5**	61.7	49.5	**58.7**	59.8	34.2	63.6	20.4	72.9	22.2	47.1	28.5	51.2	82.3	87.0	**61.7**	20.6	57.9	44.6	44.2	51.7
Diff. Guided (SD-1.5)	19.3	57.8	28.4	37.4	57.8	**81.3**	46.3	3.8	57.8	27.2	28.3	19.6	42.5	50.9	57.8	59.8	15.6	36.0	37.7	50.5	40.8+13.6
Diff. Guided (SD-2.1)	25.6	40.2	26.2	25.7	44.8	72.9	34.8	3.8	46.3	14.0	26.6	7.5	27.2	57.1	48.4	56.4	6.8	25.3	24.5	39.2	32.7+5.5

Table 5. Testing Results of Diffusion Detector with Different Stable Diffusion versions. **SD-1.5**: Stable Diffusion v1.5, **SD-2.1**: Stable Diffusion v2.1, **SD-3-M**: Stable Diffusion v3 Medium. **Foggy**: FoggyCityscapes, **Rainy**: RainCityscapes. In Diverse Weather benchmark: **DF** (Daytime-Foggy), **DR** (Dusk-Rainy), **NR** (Night-Rainy), **NS** (Night-Sunny).

	Cross Camera	**Adverse Weather**		**Synthetic to Real**		**Real to Artistic**			**Diverse Weather benchmark**			
Version	BDD100K	Foggy	Rainy	Cityscapes (car)	BDD100K (car)	Clipart	Comic	Watercolor	DF	DR	NR	NS
SD-1.5	**46.6**	**50.1**	58.2	62.8	**64.4**	**58.3**	**51.9**	**68.4**	43.3	**42.5**	**27.8**	**47.0**
SD-2.1	45.8	48.3	56.1	**64.5**	64.1	51.7	46.6	62.1	**44.6**	41.6	23.2	46.4
SD-3-M	40.4	46.1	**59.1**	59.7	54.2	28.7	24.1	45.0	36.0	30.5	15.9	32.8

D. Additional Results of Different Stable Diffusion Versions

Performance Comparison of Different SD Versions: Experimental results in Tab. 5 demonstrate varying performance among Stable Diffusion versions across different scenarios. SD-1.5 consistently achieves superior performance, particularly in adverse weather (50.1% for foggy, 58.2% for rainy) and artistic style transfer (58.3%, 51.9%, 68.4% for Clipart, Comic, Watercolor). While SD-2.1 maintains competitive performance and achieves 64.5% accuracy in Cityscapes car detection, it shows performance gaps of 6.6%, 5.3%, and 6.3% compared to SD-1.5 in artistic style transfer. SD-3-M shows significantly lower performance, with substantial degradation in artistic style transfer (28.7%, 24.1%, 45.0%) and diverse weather conditions (10.4% lower than SD-1.5).

Analysis of Architecture Differences: The inferior performance of SD-3-M primarily stems from its architectural differences. Unlike SD-1.5 and SD-2.1 with UNet [26] architecture that produces multi-scale hierarchical features, SD-3-M with transformer-based structure [10] outputs fixed-dimensional feature maps. This limitation affects its ability to capture fine-grained spatial information crucial for object detection, particularly impacting performance across diverse domains.

Ongoing Research: We are currently conducting extensive experiments to improve the cross-domain detection performance of SD-3-M. Our ongoing research focuses on developing effective methods to leverage the intermediate features of SD-3-M, aiming to fully utilize its strong semantic understanding capabilities while addressing the challenges in dense prediction tasks. The experimental results and detailed analysis will be reported in future work.

E. Additional Analysis for Results

E.1. Visualization of Domain Distribution Differences

Distribution Analysis: As visualized in Fig. 1, significant distribution gaps exist between source and target domains across different benchmarks. The diverse scenarios including cross-camera, adverse weather, synthetic-to-real transfer, artistic style transfer, and various weather conditions all demonstrate distinct distribution separations between source and target domains. These distribution discrepancies explain the challenges faced by conventional detectors when deploying across domains, highlighting the necessity of robust domain-generalized detection approaches.

E.2. Confusion Matrix Error Analysis

Analysis of Confusion Matrices: As shown in Fig. 2 and 3, the confusion matrices reveal that false negatives (missed detections) are the primary factor affecting detection performance in the baseline detector. Our proposed Diff. Detector significantly reduces the probability of missed detections, as evidenced by the stronger diagonal patterns in both FoggyCityscapes and Clipart scenarios.

Through our designed feature and object alignment mechanism, the Diff. Guided Detector successfully inherits the robust detection capability from Diff. Detector, showing similar improvements in reducing missed detections. The enhanced diagonal patterns in confusion matrices validate the effectiveness of our knowledge transfer framework in improving cross-domain generalization performance.

F. Visualization of Detection Results

Visualization Results: As shown in Fig. 4, 5, 6, 7, 8, and 9, our proposed methods demonstrate superior detection performance across various challenging scenarios. Compared to the baseline detector, both Diff. Detector and Diff. Guided Detector achieve more comprehensive detection results, successfully identifying objects under different conditions such as varying scales, weather conditions, lighting variations, and artistic styles. These qualitative results consistently validate the effectiveness of our proposed diffusion-based framework in improving detection generalization across different domains.

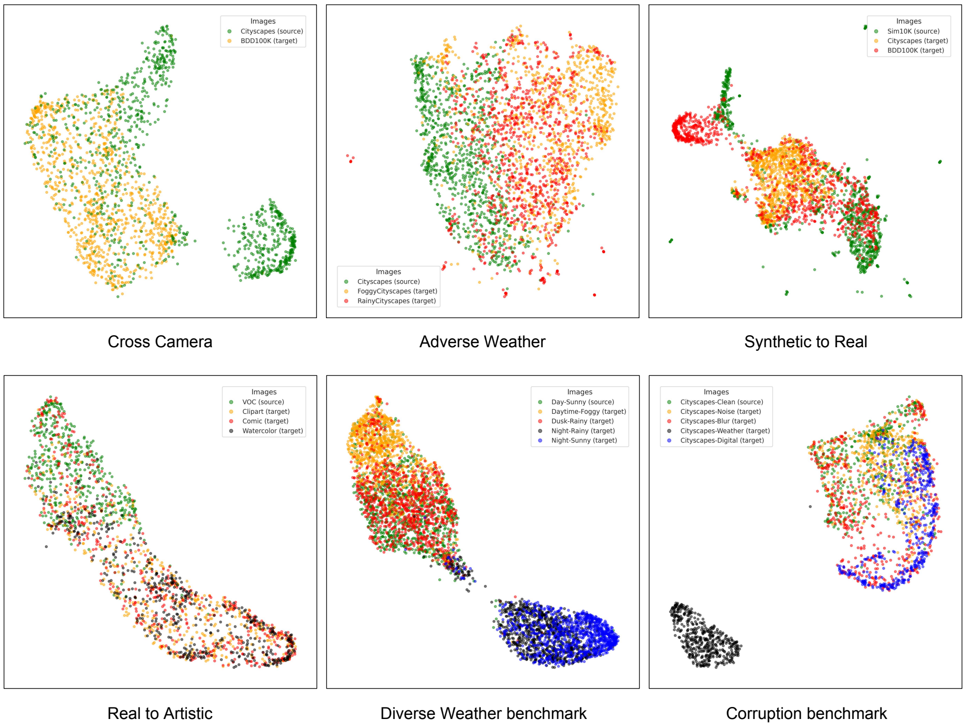

Figure 1. Image-level distribution visualization using UMAP [22] on six domain generalization benchmarks.

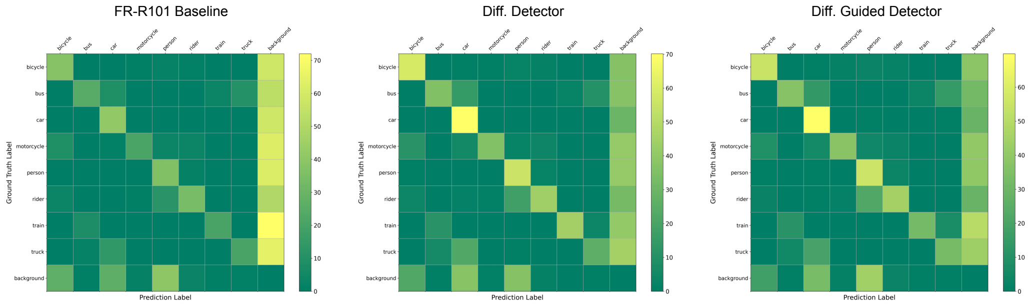

Figure 2. Confusion matrix of **Baseline** (left), **Diff. Detector** (middle), and **Diff. Guided Detector** (right) on FoggyCityscapes.

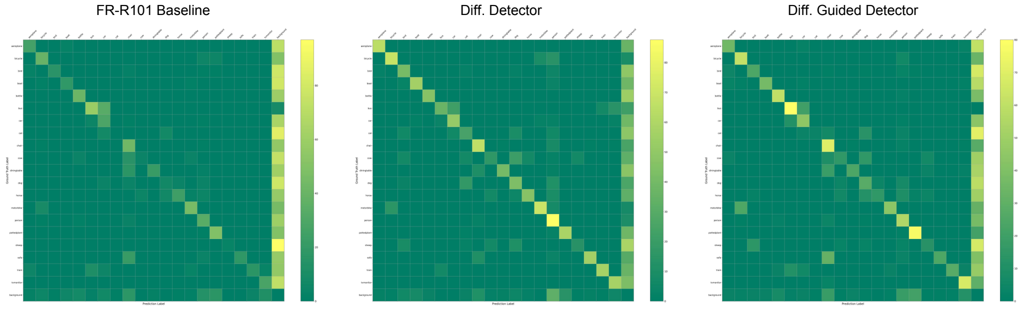

Figure 3. Confusion matrix of **Baseline** (left), **Diff. Detector** (middle), and **Diff. Guided Detector** (right) on Clipart.

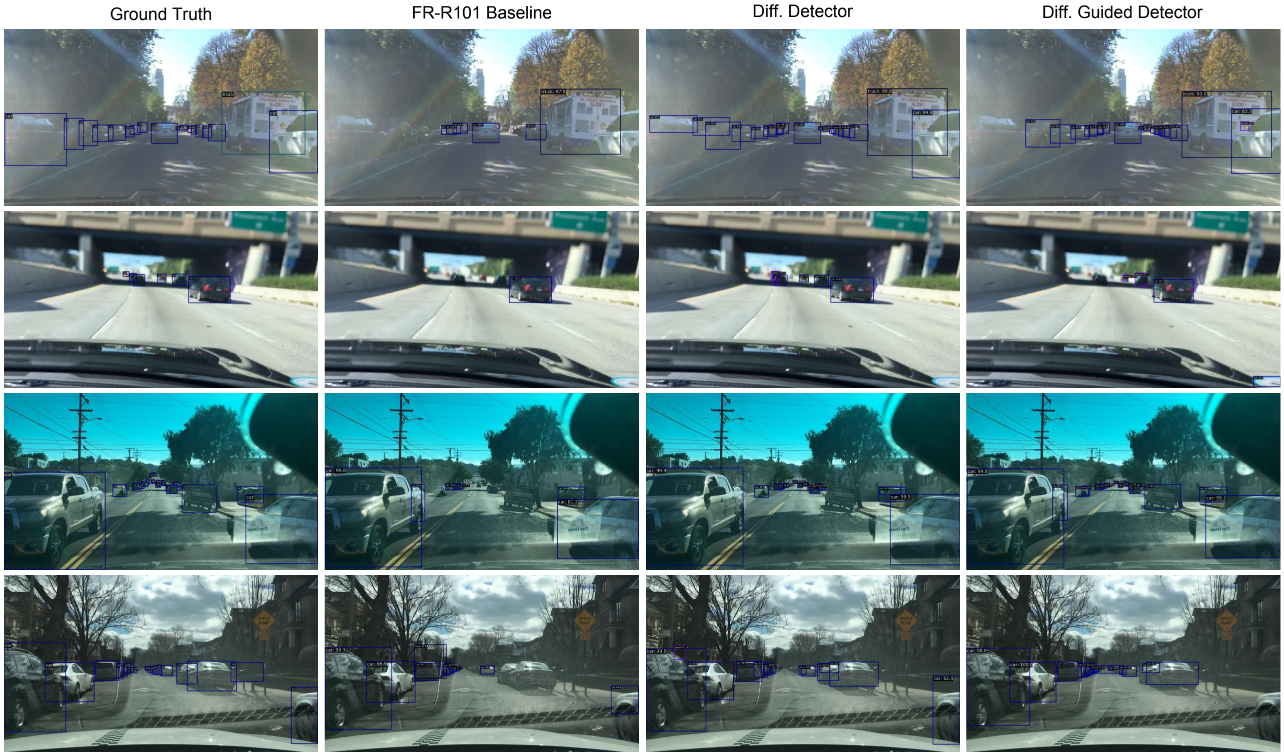

Figure 4. Qualitative prediction results on BDD100K.

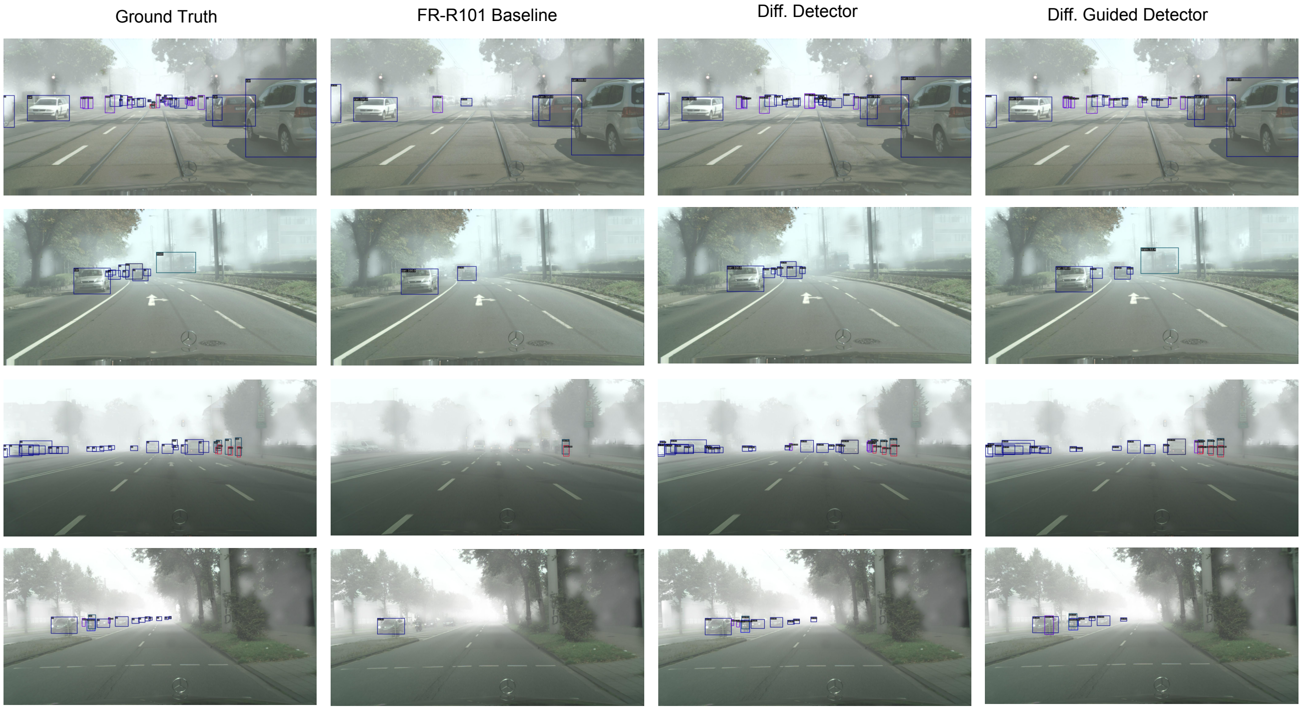

Figure 5. Qualitative prediction results on FoggyCityscapes.

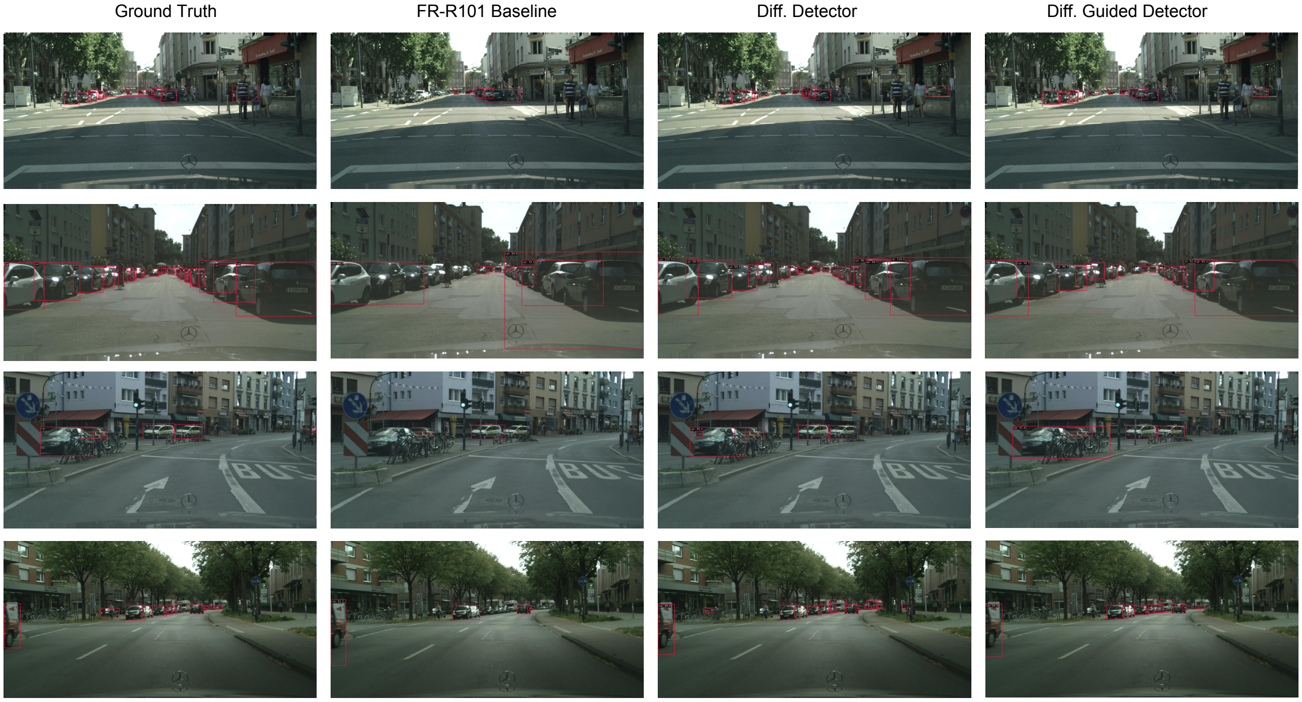

Figure 6. Qualitative prediction results on Cityscapes (Car).

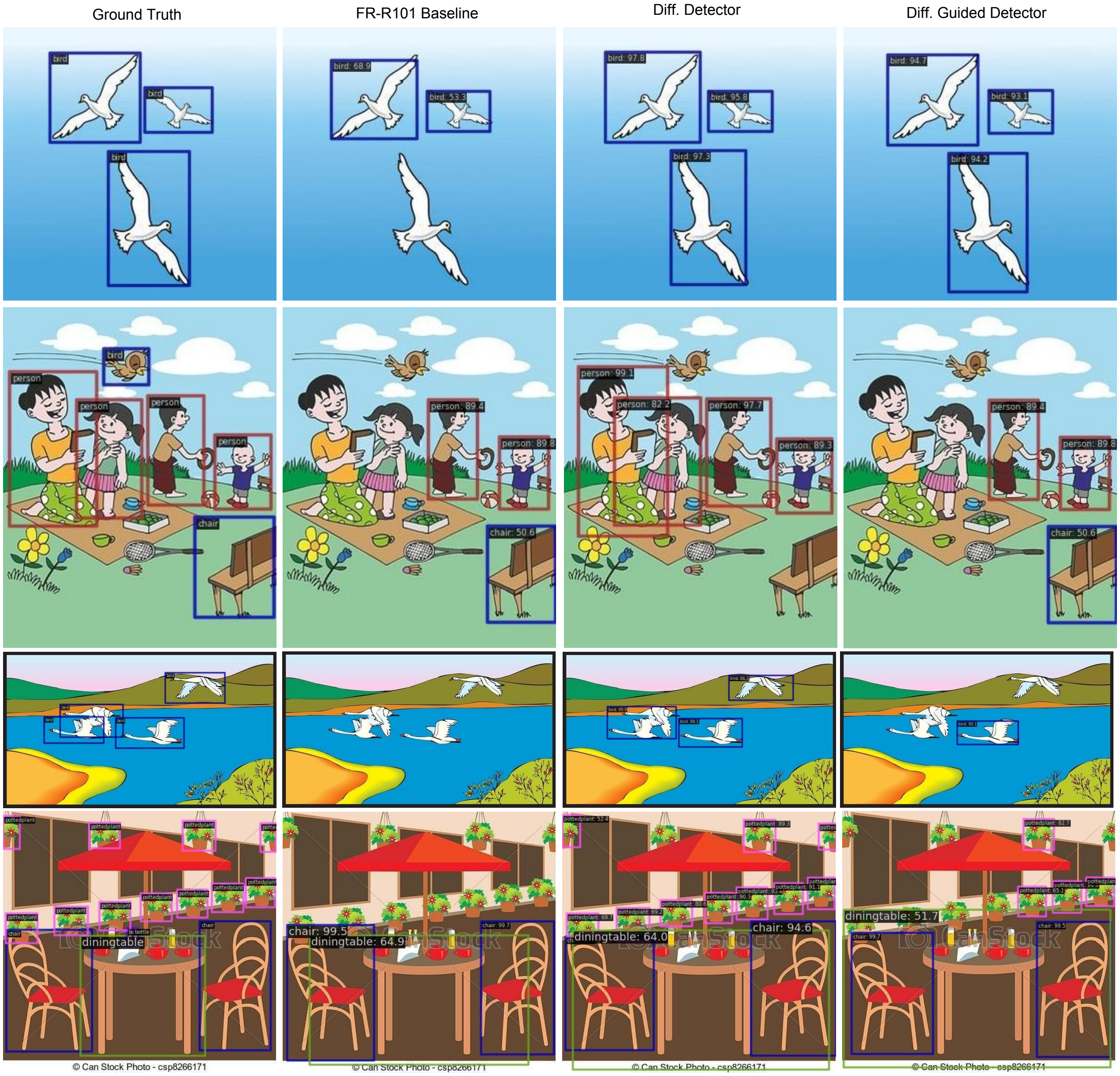

Figure 7. Qualitative prediction results on Clipart.

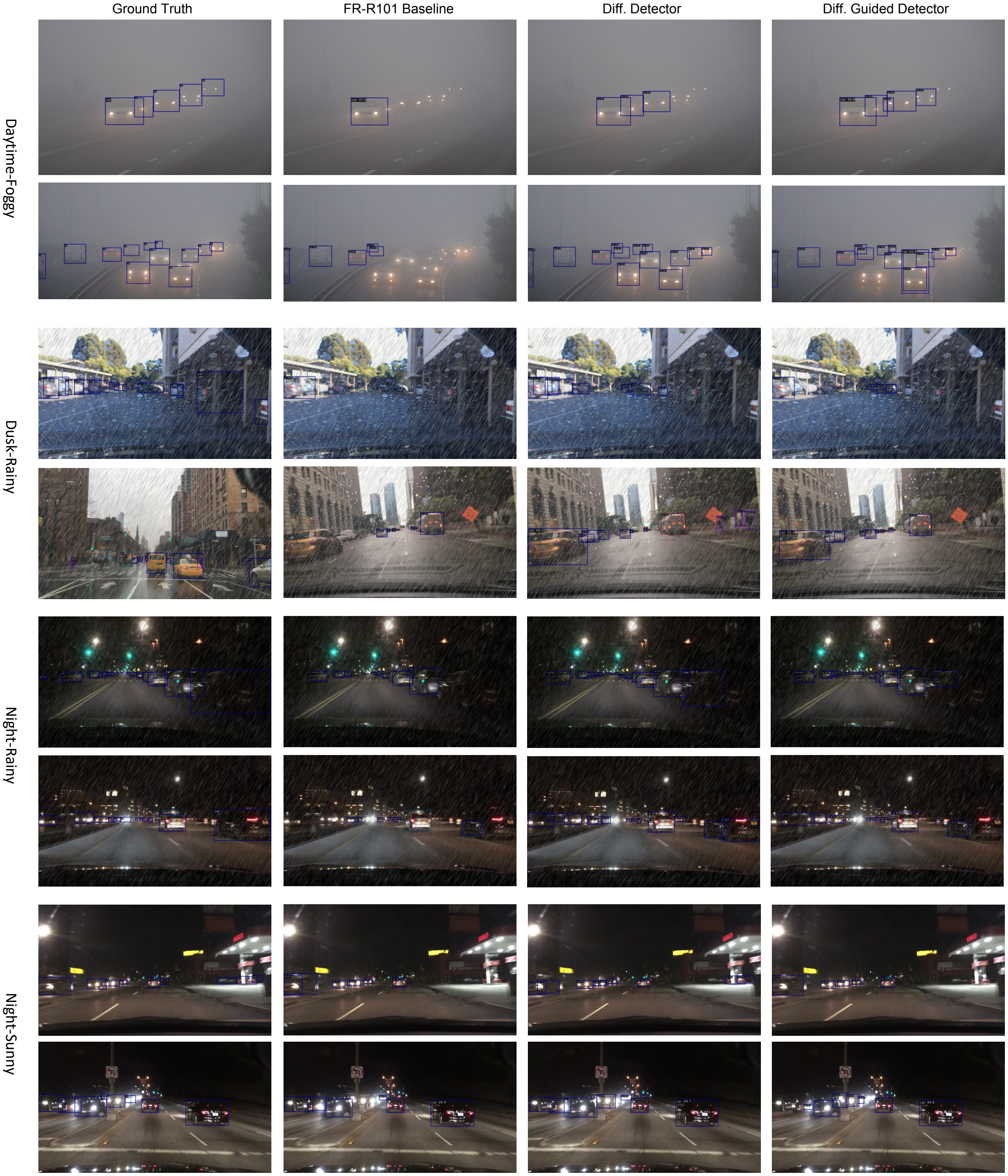

Figure 8. Qualitative prediction results on Diverse Weather Benchmark.

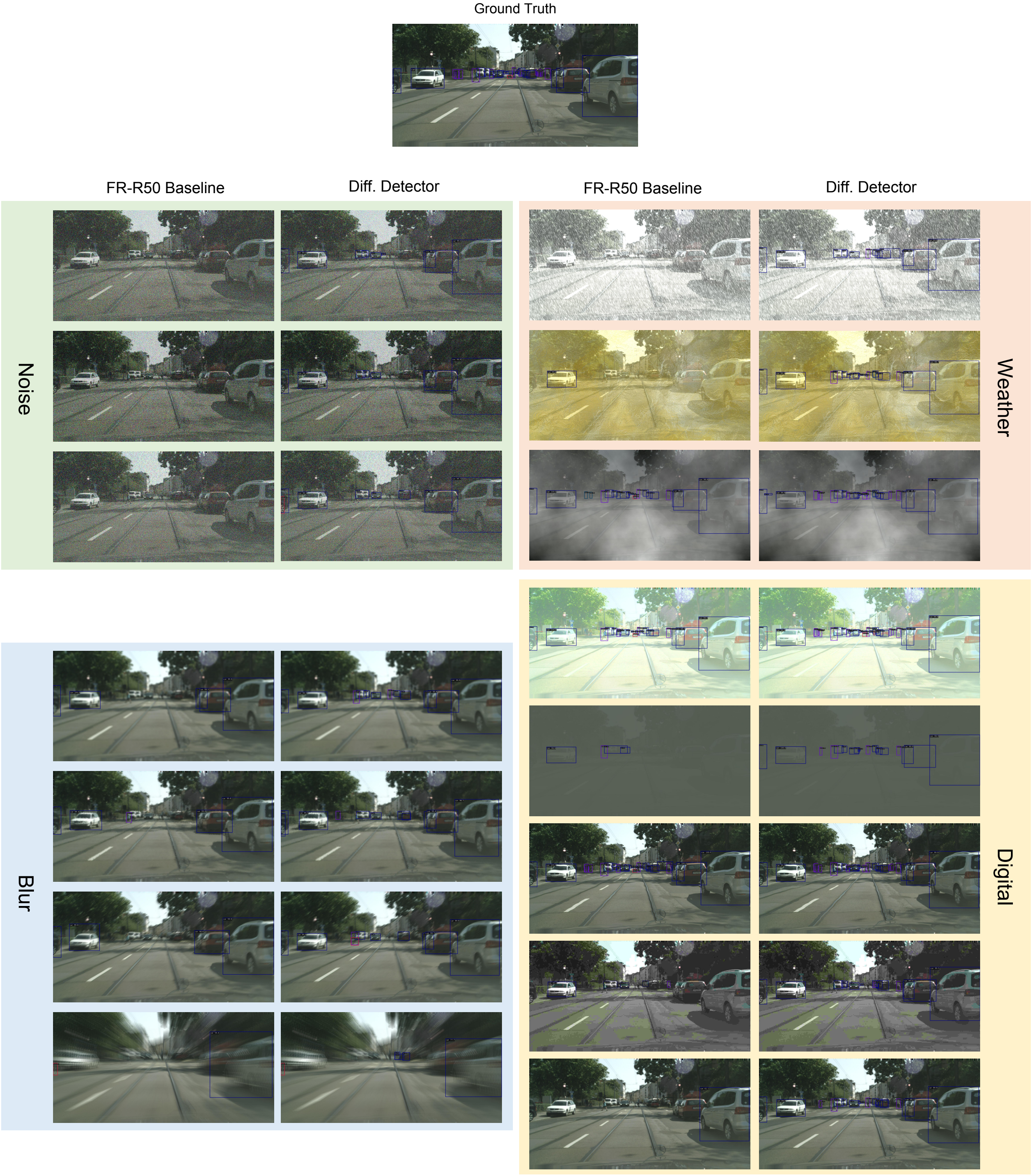

Figure 9. Qualitative prediction results on Corruption benchmark, showing detection results under 15 different corruption types (noise, blur, weather, and digital) at maximum severity level.

References

[1] Shengcao Cao, Dhiraj Joshi, Liang-Yan Gui, and Yu-Xiong Wang. Contrastive mean teacher for domain adaptive object detectors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23839–23848, 2023. 4

[2] Weihan Cao, Yifan Zhang, Jianfei Gao, Anda Cheng, Ke Cheng, and Jian Cheng. Pkd: General distillation framework for object detectors via pearson correlation coefficient. *Advances in Neural Information Processing Systems*, 35: 15394–15406, 2022. 1, 2

[3] Chaoqi Chen, Jiongcheng Li, Zebiao Zheng, Yue Huang, Xinghao Ding, and Yizhou Yu. Dual bipartite graph learning: A general approach for domain adaptive object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2703–2712, 2021. 3

[4] Chaoqi Chen, Zebiao Zheng, Yue Huang, Xinghao Ding, and Yizhou Yu. I3net: Implicit instance-invariant network for adapting one-stage object detectors. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12576–12585, 2021. 3

[5] Yuhua Chen, Wen Li, Christos Sakaridis, Dengxin Dai, and Luc Van Gool. Domain adaptive faster r-cnn for object detection in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3339–3348, 2018. 3

[6] Yuhua Chen, Haoran Wang, Wen Li, Christos Sakaridis, Dengxin Dai, and Luc Van Gool. Scale-aware domain adaptive faster r-cnn. *International Journal of Computer Vision*, 129(7):2223–2243, 2021. 3

[7] Sungha Choi, Sanghun Jung, Huiwon Yun, Joanne T Kim, Seungryong Kim, and Jaegul Choo. Robustnet: Improving domain generalization in urban-scene segmentation via instance selective whitening. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11580–11590, 2021. 4

[8] Muhammad Sohail Danish, Muhammad Haris Khan, Muhammad Akhtar Munir, M Saquib Sarfraz, and Mohsen Ali. Improving single domain-generalized object detection: A focus on diversification and alignment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17732–17742, 2024. 3, 4

[9] Jinhong Deng, Wen Li, Yuhua Chen, and Lixin Duan. Unbiased mean teacher for cross-domain object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4091–4101, 2021. 3

[10] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first International Conference on Machine Learning*, 2024. 5

[11] Boyong He, Yuxiang Ji, Zhuoyue Tan, and Liaoni Wu. Diffusion domain teacher: Diffusion guided domain adaptive object detector. In *ACM Multimedia 2024*, 2024. 1

[12] Lei Huang, Yi Zhou, Fan Zhu, Li Liu, and Ling Shao. Iterative normalization: Beyond standardization towards efficient whitening. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4874–4883, 2019. 4

[13] Junguang Jiang, Baixu Chen, Jianmin Wang, and Mingsheng Long. Decoupled adaptation for cross-domain object detection. In *International Conference on Learning Representations*, 2021. 3, 4

[14] Seunghyeon Kim, Jaehoon Choi, Taekyung Kim, and Changick Kim. Self-training and adversarial background regularization for unsupervised domain adaptive one-stage object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6092–6101, 2019. 3

[15] 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

[16] Shuai Li, Jianqiang Huang, Xian-Sheng Hua, and Lei Zhang. Category dictionary guided unsupervised domain adaptation for object detection. In *Proceedings of the AAAI conference on artificial intelligence*, pages 1949–1957, 2021. 3

[17] Shuaifeng Li, Mao Ye, Xiatian Zhu, Lihua Zhou, and Lin Xiong. Source-free object detection by learning to overlook domain style. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8014–8023, 2022. 3

[18] Yu-Jhe Li, Xiaoliang Dai, Chih-Yao Ma, Yen-Cheng Liu, Kan Chen, Bichen Wu, Zijian He, Kris Kitani, and Peter Vajda. Cross-domain adaptive teacher for object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7581–7590, 2022. 3, 4

[19] Yabo Liu, Jinghua Wang, Chao Huang, Yaowei Wang, and Yong Xu. Cigar: Cross-modality graph reasoning for domain adaptive object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23776–23786, 2023. 4

[20] Yajing Liu, Shijun Zhou, Xiyao Liu, Chunhui Hao, Baojie Fan, and Jiandong Tian. Unbiased faster r-cnn for single-source domain generalized object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 28838–28847, 2024. 3, 4

[21] Grace Luo, Lisa Dunlap, Dong Huk Park, Aleksander Holynski, and Trevor Darrell. Diffusion hyperfeatures: Searching through time and space for semantic correspondence. *Advances in Neural Information Processing Systems*, 36, 2024. 1

[22] Leland McInnes, John Healy, Nathaniel Saul, and Lukas Großberger. Umap: Uniform manifold approximation and projection. *Journal of Open Source Software*, 3(29), 2018. 6

[23] Xingang Pan, Ping Luo, Jianping Shi, and Xiaoou Tang. Two at once: Enhancing learning and generalization capacities via ibn-net. In *Proceedings of the european conference on computer vision (ECCV)*, pages 464–479, 2018. 4

[24] Xingang Pan, Xiaohang Zhan, Jianping Shi, Xiaoou Tang, and Ping Luo. Switchable whitening for deep representation learning. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1863–1871, 2019. 4

[25] Zhijie Rao, Jingcai Guo, Luyao Tang, Yue Huang, Xinghao Ding, and Song Guo. Srcd: Semantic reasoning with compound domains for single-domain generalized object detection. *IEEE Transactions on Neural Networks and Learning Systems*, 2024. 4

[26] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*, pages 234–241. Springer, 2015. 5

[27] Kuniaki Saito, Yoshitaka Ushiku, Tatsuya Harada, and Kate Saenko. Strong-weak distribution alignment for adaptive object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6956–6965, 2019. 3

[28] Luming Tang, Menglin Jia, Qianqian Wang, Cheng Perng Phoo, and Bharath Hariharan. Emergent correspondence from image diffusion. *Advances in Neural Information Processing Systems*, 36:1363–1389, 2023. 1

[29] Vidit Vidit, Martin Engilberge, and Mathieu Salzmann. Clip the gap: A single domain generalization approach for object detection. In *CVPR*, pages 3219–3229, 2023. 4

[30] Jiabao Wang, Yuming Chen, Zhaohui Zheng, Xiang Li, Ming-Ming Cheng, and Qibin Hou. Crosskd: Cross-head knowledge distillation for object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16520–16530, 2024. 1

[31] Aming Wu and Cheng Deng. Single-domain generalized object detection in urban scene via cyclic-disentangled self-distillation. In *Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition*, pages 847–856, 2022. 4

[32] Aming Wu, Yahong Han, Linchao Zhu, and Yi Yang. Instance-invariant domain adaptive object detection via progressive disentanglement. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(8):4178–4193, 2021. 3

[33] Fan Wu, Jinling Gao, Lanqing Hong, Xinbing Wang, Chenghu Zhou, and Nanyang Ye. G-nas: Generalizable neural architecture search for single domain generalization object detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 5958–5966, 2024. 3, 4

[34] Liang Zhao and Limin Wang. Task-specific inconsistency alignment for domain adaptive object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14217–14226, 2022. 4

[35] Zhen Zhao, Yuhong Guo, Haifeng Shen, and Jieping Ye. Adaptive object detection with dual multi-label prediction. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXVIII 16*, pages 54–69. Springer, 2020. 3

[36] Kaiyang Zhou, Ziwei Liu, Yu Qiao, Tao Xiang, and Chen Change Loy. Domain generalization: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(4):4396–4415, 2022. 1