# Unified Interaction Consistency Learning for Single-Source Domain-Generalized Object Detection in Urban Scene

**Peng Zhang, Xiang Yuan, Gong Cheng***

School of Automation, Northwestern Polytechnical University, Xi'an, China
{zhangpeng2001, shaunyuan}@mail.nwpu.edu.cn, gcheng@nwpu.edu.cn

## Abstract

Domain generalization remains a critical challenge for deploying neural networks, particularly in out-of-distribution object detection. The distributional discrepancy between training (*e.g.,* daytime-sunny) and the realistic condition (*e.g.,* night-rainy) inevitably produces imprecise localization and wrong classification. To address these issues, we propose a unified interaction consistency learning (UICL) framework, a novel single-source domain-generalized method designed to learn intra-class domain-invariant representations. Specifically, we put forth a cross-domain interaction mechanism to exchange region proposals between original and augmented pipelines, enriching the diversity of instance-level representations. Building upon this, we propose prediction-guided consistency learning to unify the interaction mechanism and harmonize the cross-domain representations, contributing to a discriminative prediction distribution under domain shift. In addition, we devise a cyclic interaction resilient detection strategy, which mitigates inaccurate predictions suffering from partial occlusion and ambiguous boundaries among different domains. Extensive experiments evidence that UICL significantly improves the robustness of detectors over several target domains, achieving state-of-the-art generalization performance on the diverse weather benchmark.

**Code** — https://github.com/zhangpeng2001/uicl.git

## Introduction

Mainstream detectors (Tian et al. 2019; Zhu et al. 2020; Zhang et al. 2022a; Yuan et al. 2023; Cheng et al. 2023; Wang et al. 2024a; Cheng et al. 2025) deliver strong results with consistent training and testing distributions, but suffer degraded performance in the presence of a domain gap (Chen et al. 2018; Wu and Deng 2022; Cao et al. 2023; Vidit, Engilberge, and Salzmann 2023; Zhang et al. 2025). For example, it is arduous to recognize the instances under adverse weather when the detectors are trained on clean images, as shown in the top left of Figure 1.

Recent advances for alleviating the domain shift have focused on two strategies: *domain adaptation* and *domain generalization*. Typically, domain adaptation methods (Zhao and Wang 2022; Cao et al. 2023; Gao et al. 2023; VS, Oza, and Patel 2023; Zhang et al. 2023; Weng and Yuan 2024) transfer domain-invariant knowledge from the source domain to a specific target domain, while the demand for unlabeled target images severely hinders their development. In contrast, the latter set of approaches (Zhou et al. 2022; Zhang et al. 2022b; Chen et al. 2023; Tan, Yang, and Huang 2024; Qiao, Zhao, and Peng 2020; Wang et al. 2021b; Qu et al. 2023; Cheng, Gokhale, and Yang 2023) presents a promising solution in this challenging scenario, owing to their target domain-free setup. Furthermore, improving generalization ability from a single source-domain training set poses an even greater challenge.

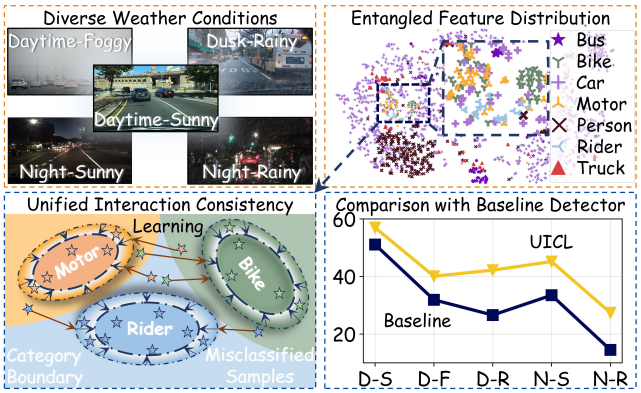


Figure 1: A novel framework for single-source domain generalized object detection. **Top-Left:** Detectors trained on the source domain (daytime-sunny) generalize to four target domains: *daytime-foggy, dusk-rainy, night-sunny, and night-rainy*. **Top-Right:** The baseline detector (Ren et al. 2015) produces overlapping category-specific distributions across four target domains, struggling to yield separable category boundaries. **Bottom-Left:** The diagram of our proposed *unified interaction consistency learning (UICL)*, exemplified by the enlarged figure from the top-right corner, narrows category boundaries to generate a distinct feature distribution. **Bottom-Right:** Quantitative comparison with the baseline verifies the superior generalization performance of UICL on the diverse weather benchmark (Wu and Deng 2022). D-S: Daytime-Sunny, D-F: Daytime-Foggy, D-R: Dusk-Rainy, N-S: Night-Sunny, N-R: Night-Rainy.

*Corresponding author.

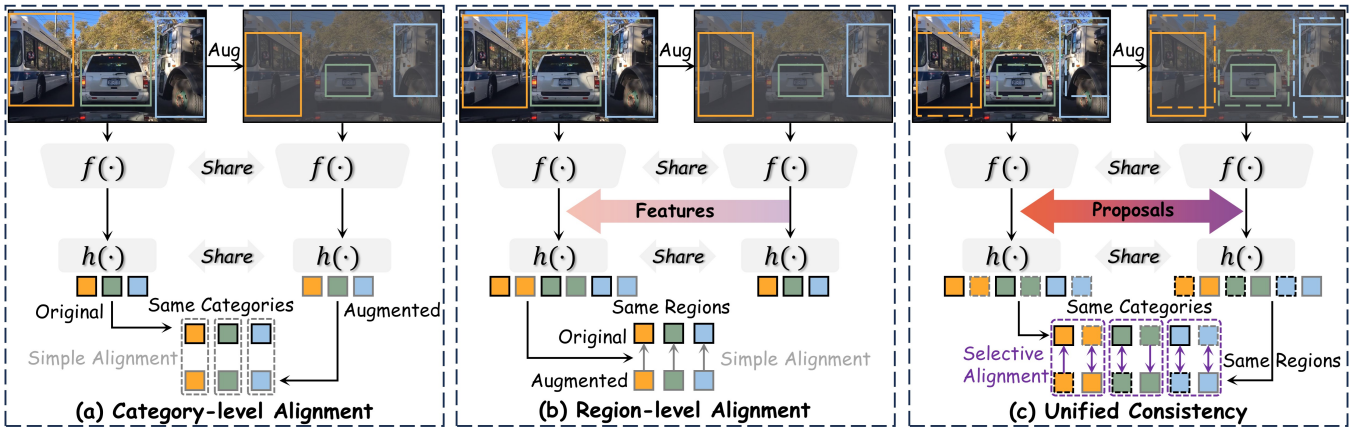


Figure 2: Comparison of alignment strategies in single-source domain-generalized object detection: **(a)** Category-level methods align samples within the same category using vanilla knowledge distillation or contrastive loss. **(b)** Region-level methods employ standard distillation loss to align samples from augmented images with their original counterparts by leveraging regions generated by the original data. **(c)** Our proposed unified consistency learning employs prediction-guided distillation loss for bidirectional alignment between corresponding regions of original and augmented images, while prediction-guided contrastive loss is leveraged to align samples with their correctly predicted counterparts.

The primary challenge lies in the entanglement of semantic features across different categories in unseen domains, even when their visual appearances differ significantly (*e.g.*, bike, motor, and rider), as illustrated in the top right of Figure 1. It indicates that when presented with images from the unseen target domain, the detector fails to capture sufficiently discriminative representations, making it challenging to establish clear classification boundaries. Consequently, the model exhibits extremely limited generalization ability, struggling to adapt effectively to domain shifts.

To this end, previous efforts mainly resort to augmentation in image space (Danish et al. 2024; Liu et al. 2024; Ahn et al. 2024) or semantic space (Li et al. 2024a; Chen et al. 2024; Addepalli et al. 2024), leveraging either textual descriptions or visual transformations to provide diverse distributions. Despite partially alleviating the aforementioned barriers, they still fail to deliver consistent predictions. Worse still, detectors are commonly prone to producing trivial solutions, generating multiple highly overlapped predictions with different categories for the same instance. We argue that the core issue is primarily from two folds: **the mismatch of proposals** and **the entangled feature distribution**, both of which suffer from domain shift. It inspires us to ask: *can we narrow the authentic category boundaries to generate a more consolidated distribution?*

With this goal, we propose a unified interaction consistency learning (UICL) framework that unifies a cross-domain interaction mechanism (CDIM) and a prediction-guided consistency learning (PGCL) strategy. The CDIM exchanges the region proposals between original and augmented pipelines to generate adequate instance-level features, and the PGCL further leverages classification scores to guide the alignment. As shown in Figure 1 (bottom-left), UICL selects true positive samples to shrink the category boundaries and aligns others toward their semantic centers. These two operations support networks to obtain unified consistency instead of solely category-level (Lee et al. 2024) or region-level alignment (Danish et al. 2024), in which the former seeks to learn the invariant distribution of the same category, while the latter maintains similar predictions for the high-quality regions, as shown in Figure 2. Moreover, we observe that detectors produce erroneous predictions in target domains characterized by partial occlusion or ambiguous boundaries. To alleviate this problem, we propose a cyclic interaction resilient detection (CIRD) strategy that randomly perturbs the coordinates of predicted bounding boxes and refines them using cross-domain features. Building upon the unified alignment, UICL significantly enhances generalization (see Figure 1, bottom-right). Ultimately, our primary contributions are summarized as follows:

- We devise a cross-domain interaction mechanism, which provides diverse instance-level distributions for consistency learning and exchanges the region proposals across original and augmented pipelines.
- We propose a prediction-guided consistency loss that shrinks the category boundaries and pulls misclassified samples toward their corresponding centers.
- We design a cyclic resilient detection strategy to refine positive instances and maintain spatial coherence under distributional discrepancy introduced by domain shift.
- Experimental results validate that our proposed UICL framework significantly enhances the generalization of detectors across diverse weather conditions.

## Related Works

### Out-of-Distribution Object Detection

Many domain adaptation methods (Li et al. 2022; Hoyer, Dai, and Van Gool 2022a,b; Li et al. 2024b; Kennerley et al. 2024; Khanh et al. 2024) train networks utilizing labeled source-domain data and unlabeled target-domain images to

address accuracy degradation influenced by the distribution discrepancy. Chen *et al.* (2018) employed an adversarial training strategy to mitigate the domain shifts at both the image and instance levels. Cai *et al.* (2019) enhanced the mean teacher paradigm to bridge the domain gap by integrating object relations into the measure of consistency cost. However, the assumption of accessible target-domain images restricts their practical application and generalization.

Moreover, several works have sought to develop a robust detection network against domain shift. Wu *et al.* (2024) designed a generalizable neural architecture search method to prevent over-fitting and learn predictive representations. Liu *et al.* (2024) formulated single-source domain generalization in object detection from a causal perspective and proposed an unbiased Faster R-CNN to learn generalizable features. These methods incorporate auxiliary modules to learn domain-invariant feature representations, which limits their extensive deployment in real-time applications.

Conversely, we propose a unified interaction consistency learning (UICL) framework to enhance the generalization performance of detectors without increasing the computational complexity of inference.

### Cross-Domain Consistency Learning

In recent years, domain generalization methods (Shu et al. 2021; Choi et al. 2021; Yang, Gu, and Sun 2023; Cheng, Gokhale, and Yang 2023; Guo, Qi, and Shi 2023; Li et al. 2023; Lee et al. 2024; Danish et al. 2024; Liu et al. 2024) have utilized augmented data to enhance robustness. A theory proposed in previous work (Muandet, Balduzzi, and Schölkopf 2013) suggests that the domain shift primarily affects the marginal distribution $P(X)$ while the conditional distribution $P(Y|X)$ remains relatively stable. Based on this, Ahn *et al.* (2024) combined covariance alignment and semantic consistency contrastive learning to generate style-invariant features. Wang *et al.* (2021a) proposed a pixel-wise contrastive learning method that maximizes the global similarities among labeled pixels to construct a well-structured latent space for semantic segmentation. Both these methods require a substantial amount of negative samples for effective consistency learning.

Considering the potential deterioration arising from data augmentation, Lee *et al.* (2024) proposed an object-aware mixing and learning strategy for object detection. As illustrated in Figure 2 (a), this approach only aligns features from two independent sources, leaving the generation of numerous and diverse instance-level samples a challenge. For another solution, Danish *et al.* (2024) improved the generalization of detectors by aligning the augmented predictions to the original domain and leveraging the same regions generated from the original images, as shown in Figure 2 (b). These methods neglect the prediction scores during feature alignment, limiting their effectiveness in the cross-domain consistency learning framework.

In contrast, we design a cross-domain interaction mechanism to enrich the diversity of instance-level representations and a cyclic interaction mechanism to synthesize challenging samples. Moreover, a prediction-guided consistency loss is devised to maximize the similarities among diverse representations within the same category and minimize the discrepancy between predictions from the same region.

## Method

### Cross-Domain Interaction Mechanism

The overall framework of our proposed unified interaction consistency learning (UICL) is illustrated in Figure 3. We denote the feature extractor and proposal generator as $f(\cdot)$, while the region of interest (RoI) alignment and detection head are defined as $h(\cdot)$. Formally, an original image $\mathbf{X}$ is augmented online using a randomly selected operation to generate an image $\mathbf{Y}$ with a different distribution. Both $\mathbf{X}$ (via pipeline $\boldsymbol{\alpha}$) and $\mathbf{Y}$ (via pipeline $\boldsymbol{\beta}$) are processed by a shared $f(\cdot)$, producing original features $\mathbf{F}_{\boldsymbol{\alpha}}$ and proposals $\mathbf{P}_{\boldsymbol{\alpha}}$, as well as augmented features $\mathbf{F}_{\boldsymbol{\beta}}$ and proposals $\mathbf{P}_{\boldsymbol{\beta}}$.

Influenced by the distribution discrepancies in cross-domain scenarios, detectors commonly produce ambiguous representations in semantic space, resulting in numerous redundant detections with erroneous classifications and imprecise bounding box regressions. To alleviate these challenges, we propose a cross-domain interaction mechanism that exchanges region proposals between original and augmented pipelines. As demonstrated in Figure 3, both pipelines process their respective and exchanged proposals together. For both pipelines, positive samples are assigned following the standard strategy used in Faster R-CNN, which helps ensure the quality of the selected proposals. With spatial alignment preserved by photometric augmentations, identical predictions are expected from a robust detector for different feature representations of the same regions. Moreover, the predicted coordinates of positive samples are randomly shifted with small perturbations and then cyclically fed back into the prediction head $h(\cdot)$. Consequently, the detection loss $\mathcal{L}_{\text{det}}$ is defined as:

$$\mathcal{L}_{\text{det}} = \mathcal{L}^{\text{f}}_{\text{det}} + \mathcal{L}^{\text{i}}_{\text{det}} + \mathcal{L}^{\text{c}}_{\text{det}}, \tag{1}$$

where $\mathcal{L}^{\text{f}}_{\text{det}}, \mathcal{L}^{\text{i}}_{\text{det}}, \mathcal{L}^{\text{c}}_{\text{det}}$ denote forward detection loss, interactive detection loss, and cyclic detection loss, respectively. Benefiting from cross-domain RoI alignment, UICL supplies a diverse set of instance-level samples for consistency loss ($\mathcal{L}_{\text{cl}}$), consequently enabling robust predictive alignment across original and augmented pipelines. This process produces domain-invariant features by leveraging correctly predicted samples as positive group samples. Concretely, it aligns classification predictions between the two pipelines, thereby effectively reducing the discrepancy brought by the distribution shift. Finally, the overall objective in UICL is written as:

$$\mathcal{L} = \mathcal{L}_{\text{det}} + \mathcal{L}_{\text{cl}}. \tag{2}$$

### Cyclic Interaction Resilient Detection

Domain-specific features often lead to low-quality proposals in region proposal networks. To mitigate this problem, we propose a cyclic interaction resilient detection (CIRD) strategy that iteratively predicts positive samples to produce stable regression results, as depicted in Figure 4. Specifically, the coordinates of corners (top-left and bottom-right) are modified by adding random perturbations before being

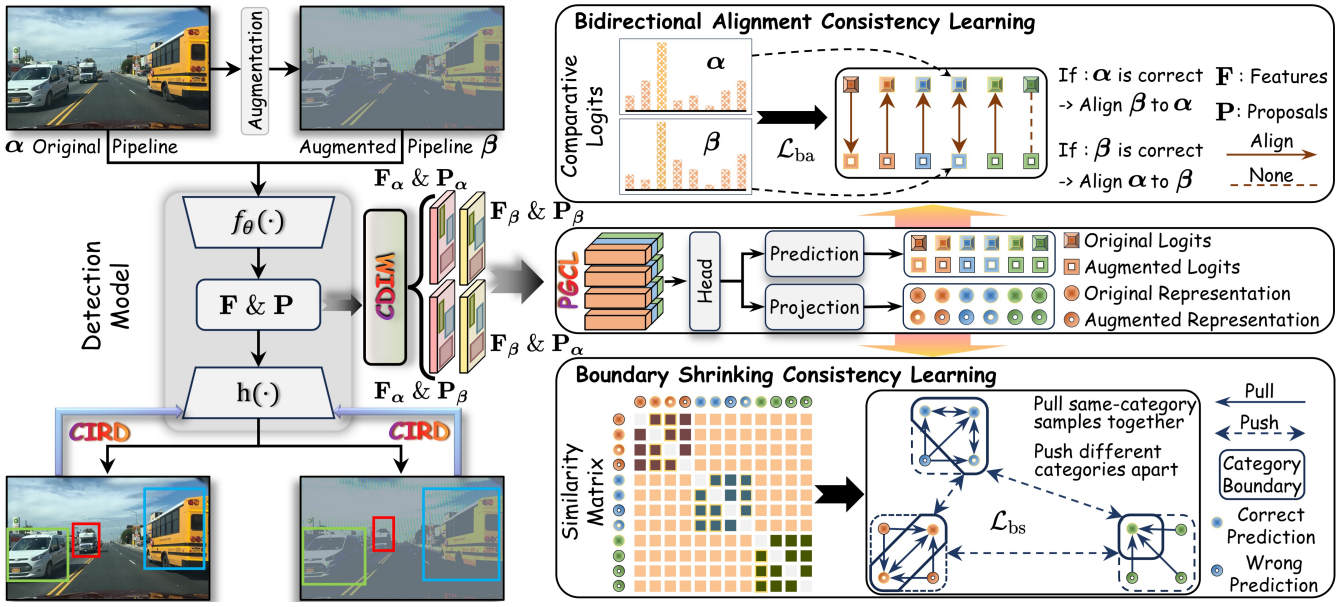


Figure 3: The *unified interaction consistency learning (UICL)* framework comprises three core components: **Cross-domain interaction mechanism (CDIM)** exchanges region proposals between the original and augmented pipelines, generating four feature-proposal combinations ($\mathbf{F}_{\boldsymbol{\alpha}}\&\mathbf{P}_{\boldsymbol{\alpha}}$, $\mathbf{F}_{\boldsymbol{\alpha}}\&\mathbf{P}_{\boldsymbol{\beta}}$, $\mathbf{F}_{\boldsymbol{\beta}}\&\mathbf{P}_{\boldsymbol{\alpha}}$, $\mathbf{F}_{\boldsymbol{\beta}}\&\mathbf{P}_{\boldsymbol{\beta}}$) for the detection head. **Cyclic interaction resilient detection (CIRD)** reprocesses positive samples by shifting the coordinates of bounding box corners (top-left and bottom-right) within a narrow range to generate challenging proposals for consistency learning. **Prediction-guided consistency learning (PGCL)** simultaneously aligns samples at the object and region levels to enhance cross-domain generalization. Specifically, *bidirectional alignment consistency learning* aligns samples with their counterparts if correctly predicted, ensuring consistency between original and augmented pipelines. *Boundary shrinking consistency learning* maximizes the similarity among correctly predicted samples within the same category, and minimizes the distances between misclassified samples and their semantic centers, thereby guaranteeing intra-class alignment.

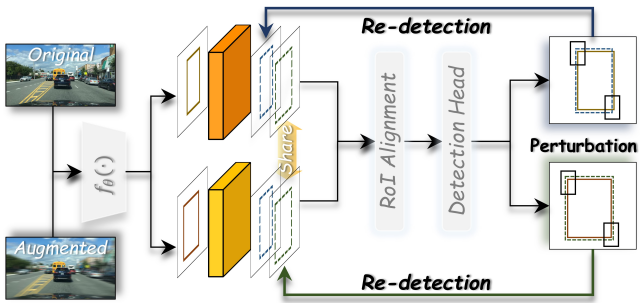


Figure 4: *Cyclic interaction resilient detection (CIRD)* refines predictions by adding random perturbations to the output coordinates of positive samples before reprocessing.

fed back into the network, and the shift is formulated as:

$$\begin{cases} x^{'} = x + \delta \cdot w \\ y^{'} = y + \delta \cdot h \end{cases}, \tag{3}$$

where $w$ and $h$ are the width and height of the predicted bounding box, while $\delta$ is a small random perturbation factor sampled independently for each coordinate. To maintain the spatial overlap, $\delta$ is constrained within a narrow range $[-\Delta, \Delta]$, where $\Delta$ is a hyperparameter controlling the maximum shift magnitude. Therefore, the minimum intersection over union (IoU) between these two bounding boxes is bounded by:

$$\text{IoU} \geq \min\left[(1-2\Delta)^2, \frac{1}{(1+2\Delta)^2}\right]. \tag{4}$$

Additionally, the perturbed coordinates (resilient positional shifts) are fed into the counterpart pipeline, generating complex synthetic samples by leveraging noisy bounding boxes and cross-domain features. This operation enhances the robustness of detectors against partial occlusion and improves generalization under ambiguous boundaries commonly encountered in target domains.

## Prediction-guided Consistency Learning

To achieve robust performance under the domain shift, the instance-level predictions that maintain common input characteristics should remain consistent in classification outputs. To be more specific, a detector should generate consistent classification predictions in the same spatial regions regardless of visual transformations. Toward this goal, we design a bidirectional alignment loss ($\mathcal{L}_{\text{ba}}$) by employing knowledge distillation (Wang et al. 2024b) to align the outputs across different detection pipelines. It forces the network to generate highly confident results even when faced with different visual transformations. Secondly, a detector should maintain

| Methods Venues | D-S | D-F | D-R | N-S | N-R | Avg. |
|---|---|---|---|---|---|---|
| FR-CNN *(NeurIPS15)* | 51.1 | 31.9 | 26.6 | 33.5 | 14.5 | 26.7 |
| IBN-Net *(ECCV18)* | 49.7 | 29.6 | 26.1 | 32.1 | 14.3 | 25.5 |
| SW *(ICCV19)* | 50.6 | 30.8 | 26.3 | 33.4 | 13.7 | 26.1 |
| IterNorm *(CVPR19)* | 43.9 | 28.4 | 22.8 | 29.6 | 12.6 | 23.4 |
| ISW *(CVPR21)* | 51.3 | 31.8 | 25.9 | 33.2 | 14.1 | 26.3 |
| CDSD *(CVPR22)* | 56.1 | 33.5 | 28.2 | 36.6 | 16.6 | 28.7 |
| CLIP-Gap *(CVPR23)* | 51.3 | 38.5 | 32.3 | 36.9 | 18.7 | 31.6 |
| OA-DG *(AAAI24)* | 55.8 | 38.3 | 33.9 | 38.0 | 16.8 | 31.8 |
| DOCP *(CVPR24)* | 53.6 | 39.1 | 33.7 | 38.5 | 19.2 | 32.6 |
| UFR *(CVPR24)* | 58.6 | 39.6 | 33.2 | 40.8 | 19.2 | 33.2 |
| G-NAS *(AAAI24)* | 58.4 | 36.4 | 35.1 | 45.0 | 17.4 | 33.5 |
| DivAlign *(CVPR24)* | 52.8 | 37.2 | 38.1 | 42.5 | 24.1 | 35.5 |
| SECT *(CVPR25)* | 55.4 | 40.6 | 39.2 | 42.0 | 24.5 | 36.6 |
| **UICL** | 57.0 | 40.1 | 42.2 | 45.2 | 27.4 | **38.7** |

Table 1: Comparison results with SOTA methods in domain generalized object detection, where the best results are highlighted in **bold**. D-S: Daytime-Sunny, D-R: Dusk-Rainy, N-S: Night-Sunny, N-R: Night-Rainy. The UICL framework achieves the highest average mAP over four target domains on the diverse weather dataset (Wu and Deng 2022).

| Methods | Person | Rider | Car | Truck | Bus | Train | Motor | Bicycle | mAP |
|---|---|---|---|---|---|---|---|---|---|
| ***Domain Adaptation*** | | | | | | | | | |
| OADA | 47.8 | 46.5 | 62.9 | 32.1 | 48.5 | 50.9 | 34.3 | 39.8 | 45.4 |
| CIGAR | 46.1 | 47.3 | 62.1 | 27.8 | 56.6 | 44.3 | 33.7 | 41.3 | 44.9 |
| CMT | 45.9 | 55.7 | 63.7 | 39.6 | 66.0 | 38.8 | 41.4 | 51.2 | 50.3 |
| HT | 52.1 | 55.8 | 67.5 | 32.7 | 55.9 | 49.1 | 40.1 | 50.3 | 50.4 |
| ***Source-Free Domain Adaptation*** | | | | | | | | | |
| IRG | 37.4 | 45.2 | 51.9 | 24.4 | 39.6 | 25.2 | 31.5 | 41.6 | 37.1 |
| LPLD | 38.3 | 42.9 | 52.5 | 28.4 | 42.1 | 43.9 | 33.4 | 41.8 | 40.4 |
| ***Domain Generalization*** | | | | | | | | | |
| FACT | 26.2 | 41.2 | 35.9 | 13.6 | 27.7 | 3.0 | 23.3 | 31.3 | 25.3 |
| MAD | 34.2 | 47.4 | 45.0 | 25.6 | 44.0 | 42.4 | 30.3 | 40.1 | 38.6 |
| **UICL** | 58.2 | 62.0 | 69.7 | 34.5 | 53.1 | 40.5 | 46.3 | 61.1 | **53.2** |

Table 2: Comparison results with various domain-related methods on the Foggy-Cityscapes dataset, where all models adopt ResNet-50 as the backbone. In *unsupervised domain adaptation*, both Cityscapes and Foggy-Cityscapes images are accessible during training. *Source-free domain adaptation* leverages a model trained on Cityscapes along with unlabeled Foggy-Cityscapes images. In contrast, *domain generalization* uses only Cityscapes data for training.

consistent and discriminative representations for the same category despite diverse appearances. To this end, we design a boundary shrinking loss ($\mathcal{L}_{\text{bs}}$) by utilizing supervised contrastive learning (Khosla et al. 2020) to enhance intra-class compactness and inter-class variations in the latent space. It maximizes the similarities among correctly predicted samples, while minimizing the distances between misclassified samples and their corresponding semantic centers. As shown in Figure 3, the detection head generates instance representations. Subsequently, a linear classifier yields predictions for bidirectional alignment, while a non-linear projection head produces latent features for boundary shrinking.

The *bidirectional alignment loss* ($\mathcal{L}_{\text{ba}}$) treats predictions from different pipelines equally, where only correctly classified predictions are registered as positive group samples. Notably, no distinction is made between foreground and background regions during alignment. This loss is formulated as:

$$\mathcal{L}_{\text{ba}} = \frac{1}{N}\sum_{i=1}^{N}\left[\sigma(\boldsymbol{p}_i^{\boldsymbol{\beta}})\text{KL}(\boldsymbol{p}_i^{\boldsymbol{\alpha}}||\boldsymbol{p}_i^{\boldsymbol{\beta}}) + \sigma(\boldsymbol{p}_i^{\boldsymbol{\alpha}})\text{KL}(\boldsymbol{p}_i^{\boldsymbol{\beta}}||\boldsymbol{p}_i^{\boldsymbol{\alpha}})\right], \tag{5}$$

where $i$ indexes the $N$ samples, $\sigma(\cdot)$ is the selection function to identify correctly classified samples, and $\text{KL}(\cdot)$ denotes the Kullback-Leibler divergence (Hinton, Vinyals, and Dean 2015), measuring the discrepancy between predictions $\boldsymbol{p}^{\boldsymbol{\alpha}}$ and $\boldsymbol{p}^{\boldsymbol{\beta}}$ from pipelines $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$, respectively.

The *boundary shrinking loss* ($\mathcal{L}_{\text{bs}}$) flattens all samples and calculates the similarity matrix using a dot product. It first selects correctly predicted instances from foreground categories as positive group samples for supervised contrastive loss and maximizes their pairwise similarities. Meanwhile, the misclassified instances are further aligned with true positive samples to minimize the distances to their semantic centers. In contrast, it only chooses the corresponding (respective or exchanged) counterparts in the same regions for background instances. Finally, the loss is formulated as:

$$\mathcal{L}_{\text{bs}} = \sum_{i\in I}\frac{-1}{|P|}\sum_{p\in P}\log\left[\frac{\exp\left(\text{sim}(\boldsymbol{z}_i,\boldsymbol{z}_p)/\tau\right)}{\sum_{a\in I\setminus\{i\}}\exp\left(\text{sim}(\boldsymbol{z}_i,\boldsymbol{z}_a)/\tau\right)}\right], \tag{6}$$

where $I$ denotes the set of predicted samples, while $P$ represents the set of correctly predicted samples that share the same category as the sample $i$. Moreover, $\text{sim}(\cdot)$ computes dot-product similarity, and $\tau$ is the temperature parameter. $\boldsymbol{z}_i, \boldsymbol{z}_p, \boldsymbol{z}_a$ represent anchor, positive, and contrasting samples, respectively. Finally, the consistency loss $\mathcal{L}_{\text{cl}}$ is formally defined as:

$$\mathcal{L}_{\text{cl}} = \mathcal{L}_{\text{ba}} + \lambda \cdot \mathcal{L}_{\text{bs}}, \tag{7}$$

where $\lambda$ is a hyperparameter that balances the bidirectional alignment loss $\mathcal{L}_{\text{ba}}$ and the boundary-shrinking loss $\mathcal{L}_{\text{bs}}$.

# Experiment

## Dataset

We utilize an urban-scene diverse weather dataset (Wu and Deng 2022) to evaluate the generalization performance of detectors trained with the unified interaction consistency learning (UICL) framework and other competitors. It comprises images from several well-known datasets (Yu et al. 2020; Sakaridis, Dai, and Van Gool 2018; Hassaballah et al. 2020). All models are trained on the *daytime-sunny* condition and tested on four target domains: *daytime-foggy, dusk-rainy, night-rainy, and night-sunny*. Following the prior work (Wu and Deng 2022), we focus on the common categories in driving scenes to evaluate the accuracy, including *bus, bike, car, motor, person, and truck*.

Beyond the primary evaluation on the diverse weather benchmark, we further assess the generalization of UICL on Foggy-Cityscapes (Sakaridis, Dai, and Van Gool 2018),

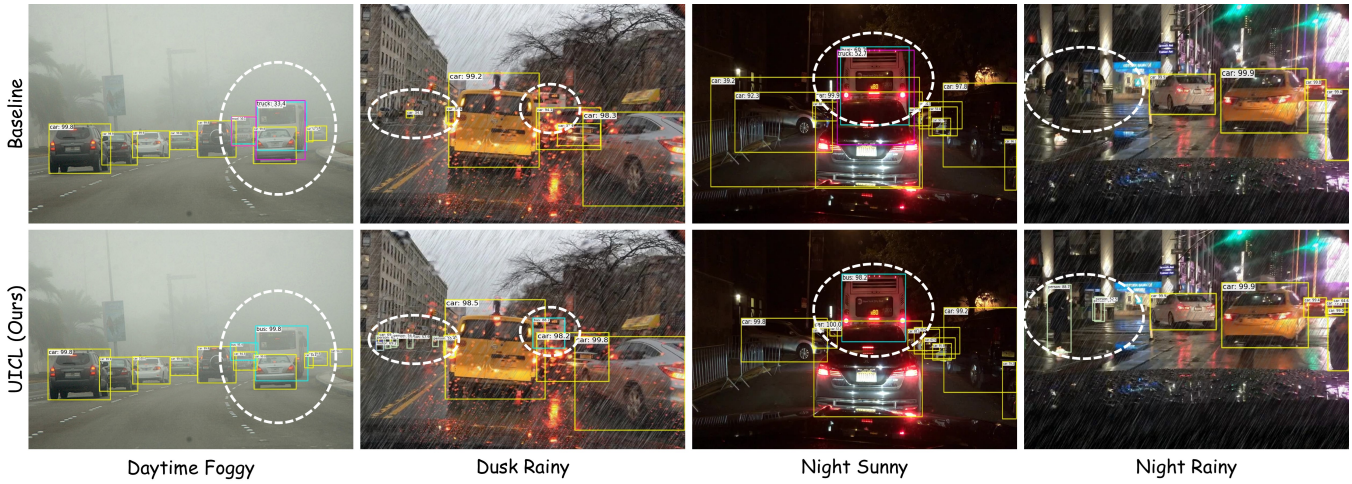


Figure 5: Visual comparison in four target domains. **Top row**: Detection results from the baseline. **Bottom row**: Results from our proposed *unified interaction consistency learning (UICL)* framework.

| | Noise | | | Blur | | | | Weather | | | Digital | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Methods** | Gauss. | Shot | Impulse | Defocus | Glass | Motion | Zoom | Snow | Frost | Fog | Bright | Contrast | Elastic | JPEG' | Pixel | mPC |
| FSCE | 7.4 | 10.2 | 8.2 | 23.3 | 20.3 | 21.5 | 4.8 | 5.6 | 23.6 | 37.1 | 38.0 | 31.9 | 40.0 | 20.4 | 23.2 | 21.0 |
| OA-Mix | 7.2 | 9.6 | 7.7 | 22.8 | 18.8 | 21.9 | 5.4 | 5.2 | 23.6 | 37.3 | 38.7 | 31.9 | 40.2 | 20.2 | 22.2 | 20.8 |
| OA-DG | 8.2 | 10.6 | 8.4 | 24.6 | 20.5 | 22.3 | 4.8 | 6.1 | 25.0 | 38.4 | 39.7 | 32.8 | 40.2 | 22.0 | 23.8 | 21.8 |
| **UICL** | 28.5 | 30.9 | 29.2 | 31.9 | 21.6 | 26.0 | 14.6 | 13.6 | 37.2 | 44.5 | 40.5 | 38.2 | 41.6 | 25.8 | 38.5 | **31.5** |

Table 3: Generalization results on the Cityscapes-Corruption benchmark, where all models are trained on Cityscapes. For each type of corruption, the average mAP is reported over five severity levels, and mPC denotes the mean value over all types.

with comparisons to different out-of-distribution methods. Additionally, we validate its robustness against diverse corruptions on Cityscapes-Corruption (Lee et al. 2024).

## Implementation Details

For a fair comparison, we adopt the augmentation strategy of DivAlign (Danish et al. 2024), applying universal techniques such as ImageNet-C (Hendrycks and Dietterich 2019) corruptions and Fourier-based (Vaish, Wang, and Strisciuglio 2024) methods to source-domain images. For the diverse weather dataset, we exclude weather-related corruption augmentations during the training phase. The widely used Faster R-CNN (Ren et al. 2015) in out-of-distribution object detection serves as the baseline detector, which employs a ResNet-101 (He et al. 2016) backbone and feature pyramid networks (Lin et al. 2017) for feature extraction. All models are optimized using the stochastic gradient descent optimizer with a learning rate of 0.002, momentum of 0.9, and weight decay of 0.0001. Unless otherwise specified, we utilize a batch size of 4 and train the models for 12 epochs. Following SimCLR (Chen et al. 2020), we adopt a non-linear projection head to generate features in the latent space, with a hidden layer size of 128 and an output dimension of 32, and the temperature is set to 0.07 by default.

## Generalization Analysis

We compare the generalization performance of our UICL method with previous state-of-the-art (SOTA) single-source domain-generalized object detectors (Zhang, Wu, and Han 2025) and feature normalization approaches (Pan et al. 2018, 2019; Huang et al. 2019; Choi et al. 2021). As listed in Table 1, UICL generalizes to diverse weather conditions effectively, achieving the highest average mAP of 38.7% on the diverse weather generalization benchmark (Wu and Deng 2022). Compared to previous alignment-based methods like OA-DG (Lee et al. 2024) and DivAlign (Danish et al. 2024), UICL outperforms them in all conditions, highlighting its effective consistency learning. While previous methods (OA-DG, DivAlign, and UFR) also employ data augmentation strategies, UICL leverages them more effectively to substantially improve prediction consistency across views, ultimately attaining the best performance on all target domains.

The promising generalization of UICL is further demonstrated through comparisons with other out-of-distribution methods, including Fourier augmentation (Xu et al. 2021) and causal mechanism (Xu et al. 2023), as summarized in Table 2. Without access to any target-domain images during training, UICL outperforms unsupervised domain adaptation methods (Liu et al. 2023; Cao et al. 2023; Deng et al. 2023) by at least 2.9% mAP. UICL significantly surpasses source-free domain adaptation methods (VS, Oza, and Patel 2023; Yoon et al. 2024) by over 10% mAP with a straightforward training setup. Furthermore, Table 3 provides clear evidence that UICL exhibits superior robustness against diverse corruption conditions, significantly improving detection accuracy across multiple severity levels.

| DA | DP | CDIM | CIRD | $\mathcal{L}_{ba}$ | $\mathcal{L}_{bs}$ | D-F | D-R | N-S | N-R |
|---|---|---|---|---|---|---|---|---|---|
| | | | | | | 31.9 | 26.6 | 33.5 | 14.5 |
| ✓ | | | | | | 36.4 | 37.1 | 41.5 | 22.8 |
| ✓ | ✓ | | | | | 37.7 | 38.3 | 43.9 | 24.7 |
| ✓ | ✓ | ✓ | | | | 38.3 | 39.7 | 44.6 | 26.5 |
| ✓ | ✓ | ✓ | ✓ | | | 39.4 | 40.6 | 44.2 | 26.4 |
| ✓ | ✓ | ✓ | ✓ | ✗ | ✓ | 39.7 | 41.9 | 45.0 | 26.8 |
| | | | | ✓ | ✗ | 39.8 | 42.2 | 45.2 | 27.3 |
| | | | | ✓ | ✓ | 40.1 | 42.2 | 45.2 | 27.4 |

Table 4: Ablation studies systematically evaluate the individual contributions of the three core components in **UICL**, where the *first* row denotes the baseline detector. The *second* row represents the baseline detector retrained solely with data augmentation (DA). Subsequently, the *third* row corresponds to the results of a model trained jointly on both the original and augmented images (Dual Pipeline, DP), with the detector architecture and loss remaining unchanged.

## Ablation Studies

In this section, we conduct experiments to evaluate the individual contributions in UICL, as detailed in Table 4. Although the improvement yielded by augmentations is evident, a substantial gap still exists among different domains. Moreover, the gains achieved by the strategy of training a detector on the dual pipeline deliver the potential of interaction consistency learning. Visually, our UICL enables the detector to produce reliable results, as shown in Figure 5.

**Analysis of CDIM.** We first evaluate the cross-domain interaction mechanism (CDIM), which exchanges proposals between original and augmented pipelines, encouraging networks to generate consistent predictions across different domains and providing abundant negative samples for consistency loss. As summarized in Table 4, employing CDIM significantly improves detector generalization, especially in the most challenging scenario (night-rainy).

**Analysis of CIRD.** To mitigate misclassified results from coarse region proposals and partial occlusions, we develop the CIRD strategy, which iteratively refines positive samples. To maintain overlap, the maximum perturbation $\Delta$ of the bounding boxes in CIRD is set to 0.1. As reported in Table 4, CIRD enhances accuracy in daytime-foggy and dusk-rainy scenes but leads to a slight decline in night conditions. This observation underscores the necessity of aligning cross-domain samples with their positive counterparts.

**Analysis of PGCL.** In prediction-guided consistency learning (PGCL), we integrate bidirectional alignment and boundary shrinking consistency learning. We utilize bidirectional alignment consistency learning to align multi-domain samples with their correctly classified counterparts. For boundary shrinking consistency learning, we select correctly predicted samples as the positive group to shrink category boundaries and further align misclassified samples toward their semantic centers. As reported in Table 4, both losses enhance generalization, which suggests that the region-level alignment for object detection plays a critical role in mitigating domain shift. In PGCL, we set $\lambda$ in Eq. (7) to 0.1 during training to balance the two losses.

| Method | D-S | D-F | D-R | N-S | N-R |
|---|---|---|---|---|---|
| VCL | 56.5 | 39.5 | 41.8 | 44.7 | 27.2 |
| **PGCL** | **57.0** | **40.1** | **42.2** | **45.2** | **27.4** |

Table 5: Comparison of two consistency learning variants (vanilla consistency learning, VCL *vs.* prediction-guided consistency learning, PGCL).

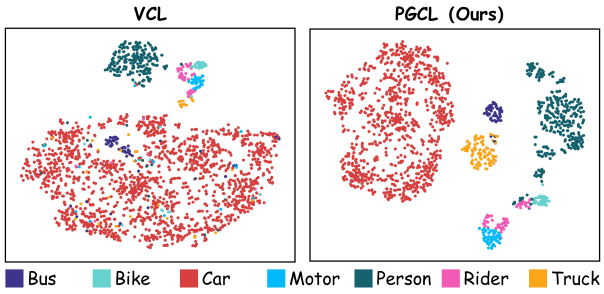


Figure 6: The t-SNE visual comparison between VCL and our proposed PGCL, using randomly selected images from the daytime-foggy condition.

We conduct experiments to evaluate prediction-guided consistency learning. Compared to VCL, which aligns instances without considering classification scores, PGCL employs correct predictions to accomplish this goal. As listed in Table 5, aligning with samples predicted accurately enhances detection accuracy across all domains. Moreover, we visualize the instance-level feature distribution from the target domain (daytime-foggy). As shown in Figure 6, PGCL generates a distinct distribution by shrinking the category boundaries and aligning the wrong predictions toward their semantic centers, *e.g.*, the bus is far from the similar car.

## Conclusion

In this paper, we designed a unified interaction consistency learning (UICL) framework for single-source domain-generalized object detection. Firstly, the interaction between the original and augmented pipelines allows the network to generate consistent predictions across different distributions. Secondly, prediction-guided consistency learning advances knowledge distillation and contrastive learning in a unified manner to better align samples with their authentic semantic centers. Finally, reprocessing resilient coordinates via cyclic detection prevents overfitting to easy-to-learn samples, improving the performance under occlusions and ambiguous boundaries. Extensive empirical results validate its robustness against diverse adverse weather conditions, achieving SOTA generalization performance across all target domains.

## Ackonwledgements

This work was supported in part by the National Natural Science Foundation of China under Grant 62376223 and in part by the Fundamental Research Funds for the Central Universities.

## References


Addepalli, S.; Asokan, A. R.; Sharma, L.; and Babu, R. V. 2024. Leveraging vision-language models for improving domain generalization in image classification. In *CVPR*, 23922–23932.

Ahn, W.-J.; Yang, G.-Y.; Choi, H.-D.; and Lim, M.-T. 2024. Style Blind Domain Generalized Semantic Segmentation via Covariance Alignment and Semantic Consistence Contrastive Learning. In *CVPR*, 3616–3626.

Cai, Q.; Pan, Y.; Ngo, C.-W.; Tian, X.; Duan, L.; and Yao, T. 2019. Exploring object relation in mean teacher for cross-domain detection. In *CVPR*, 11457–11466.

Cao, S.; Joshi, D.; Gui, L.-Y.; and Wang, Y.-X. 2023. Contrastive mean teacher for domain adaptive object detectors. In *CVPR*, 23839–23848.

Chen, L.; Zhang, Y.; Song, Y.; Shan, Y.; and Liu, L. 2023. Improved test-time adaptation for domain generalization. In *CVPR*, 24172–24182.

Chen, T.; Kornblith, S.; Norouzi, M.; and Hinton, G. 2020. A simple framework for contrastive learning of visual representations. In *ICML*, 1597–1607.

Chen, Y.; Li, W.; Sakaridis, C.; Dai, D.; and Van Gool, L. 2018. Domain adaptive faster r-cnn for object detection in the wild. In *CVPR*, 3339–3348.

Chen, Z.; Wang, W.; Zhao, Z.; Su, F.; Men, A.; and Meng, H. 2024. Practicaldg: Perturbation distillation on vision-language models for hybrid domain generalization. In *CVPR*, 23501–23511.

Cheng, G.; Yuan, X.; Yao, X.; Yan, K.; Zeng, Q.; Xie, X.; and Han, J. 2023. Towards large-scale small object detection: Survey and benchmarks. *IEEE TPAMI*, 45(11): 13467–13488.

Cheng, J.; Yao, X.; Yuan, X.; and Han, J. 2025. Not All Tokens Matter All The Time: Dynamic Token Aggregation Towards Efficient Detection Transformers. In *ICML*, 1–11.

Cheng, S.; Gokhale, T.; and Yang, Y. 2023. Adversarial bayesian augmentation for single-source domain generalization. In *ICCV*, 11400–11410.

Choi, S.; Jung, S.; Yun, H.; Kim, J. T.; Kim, S.; and Choo, J. 2021. Robustnet: Improving domain generalization in urban-scene segmentation via instance selective whitening. In *CVPR*, 11580–11590.

Danish, M. S.; Khan, M. H.; Munir, M. A.; Sarfraz, M. S.; and Ali, M. 2024. Improving Single Domain-Generalized Object Detection: A Focus on Diversification and Alignment. In *CVPR*, 17732–17742.

Deng, J.; Xu, D.; Li, W.; and Duan, L. 2023. Harmonious teacher for cross-domain object detection. In *CVPR*, 23829–23838.

Gao, C.; Liu, C.; Dun, Y.; and Qian, X. 2023. Csda: Learning category-scale joint feature for domain adaptive object detection. In *ICCV*, 11421–11430.

Guo, J.; Qi, L.; and Shi, Y. 2023. Domaindrop: Suppressing domain-sensitive channels for domain generalization. In *ICCV*, 19114–19124.

Hassaballah, M.; Kenk, M. A.; Muhammad, K.; and Minaee, S. 2020. Vehicle detection and tracking in adverse weather using a deep learning framework. *IEEE TITS*, 22(7): 4230–4242.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *CVPR*, 770–778.

Hendrycks, D.; and Dietterich, T. 2019. Benchmarking neural network robustness to common corruptions and perturbations. *ICLR*, 1–11.

Hinton, G.; Vinyals, O.; and Dean, J. 2015. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*.

Hoyer, L.; Dai, D.; and Van Gool, L. 2022a. Daformer: Improving network architectures and training strategies for domain-adaptive semantic segmentation. In *CVPR*, 9924–9935.

Hoyer, L.; Dai, D.; and Van Gool, L. 2022b. Hrda: Context-aware high-resolution domain-adaptive semantic segmentation. In *ECCV*, 372–391. Springer.

Huang, L.; Zhou, Y.; Zhu, F.; Liu, L.; and Shao, L. 2019. Iterative normalization: Beyond standardization towards efficient whitening. In *CVPR*, 4874–4883.

Kennerley, M.; Wang, J.-G.; Veeravalli, B.; and Tan, R. T. 2024. Cat: Exploiting inter-class dynamics for domain adaptive object detection. In *CVPR*, 16541–16550.

Khanh, T. L. B.; Nguyen, H.-H.; Pham, L. H.; Tran, D. N.-N.; and Jeon, J. W. 2024. Dynamic retraining-updating mean teacher for source-free object detection. In *ECCV*, 328–344. Springer.

Khosla, P.; Teterwak, P.; Wang, C.; Sarna, A.; Tian, Y.; Isola, P.; Maschinot, A.; Liu, C.; and Krishnan, D. 2020. Supervised contrastive learning. *NeurIPS*, 33: 18661–18673.

Lee, W.; Hong, D.; Lim, H.; and Myung, H. 2024. Object-Aware Domain Generalization for Object Detection. In *AAAI*, 2947–2955.

Li, C.; Zhang, D.; Huang, W.; and Zhang, J. 2023. Cross contrasting feature perturbation for domain generalization. In *ICCV*, 1327–1337.

Li, D.; Wu, A.; Wang, Y.; and Han, Y. 2024a. Prompt-Driven Dynamic Object-Centric Learning for Single Domain Generalization. In *CVPR*, 17606–17615.

Li, H.; Zhang, R.; Yao, H.; Zhang, X.; Hao, Y.; Song, X.; Li, X.; Zhao, Y.; Chen, Y.; and Li, L. 2024b. Da-ada: Learning domain-aware adapter for domain adaptive object detection. *NeurIPS*, 37: 103574–103598.

Li, Y.-J.; Dai, X.; Ma, C.-Y.; Liu, Y.-C.; Chen, K.; Wu, B.; He, Z.; Kitani, K.; and Vajda, P. 2022. Cross-domain adaptive teacher for object detection. In *CVPR*, 7581–7590.

Lin, T.-Y.; Dollár, P.; Girshick, R.; He, K.; Hariharan, B.; and Belongie, S. 2017. Feature pyramid networks for object detection. In *CVPR*, 2117–2125.

Liu, Y.; Wang, J.; Huang, C.; Wang, Y.; and Xu, Y. 2023. Cigar: Cross-modality graph reasoning for domain adaptive object detection. In *CVPR*, 23776–23786.

Liu, Y.; Zhou, S.; Liu, X.; Hao, C.; Fan, B.; and Tian, J. 2024. Unbiased Faster R-CNN for Single-source Domain Generalized Object Detection. In *CVPR*, 28838–28847.

Muandet, K.; Balduzzi, D.; and Schölkopf, B. 2013. Domain generalization via invariant feature representation. In *ICML*, 10–18. PMLR.

Pan, X.; Luo, P.; Shi, J.; and Tang, X. 2018. Two at once: Enhancing learning and generalization capacities via ibn-net. In *ECCV*, 464–479.

Pan, X.; Zhan, X.; Shi, J.; Tang, X.; and Luo, P. 2019. Switchable whitening for deep representation learning. In *ICCV*, 1863–1871.

Qiao, F.; Zhao, L.; and Peng, X. 2020. Learning to learn single domain generalization. In *CVPR*, 12556–12565.

Qu, S.; Pan, Y.; Chen, G.; Yao, T.; Jiang, C.; and Mei, T. 2023. Modality-agnostic debiasing for single domain generalization. In *CVPR*, 24142–24151.

Ren, S.; He, K.; Girshick, R.; and Sun, J. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. In *NeurIPS*, 91–99.

Sakaridis, C.; Dai, D.; and Van Gool, L. 2018. Semantic foggy scene understanding with synthetic data. *IJCV*, 126: 973–992.

Shu, Y.; Cao, Z.; Wang, C.; Wang, J.; and Long, M. 2021. Open domain generalization with domain-augmented meta-learning. In *CVPR*, 9624–9633.

Tan, Z.; Yang, X.; and Huang, K. 2024. Rethinking multi-domain generalization with a general learning objective. In *CVPR*, 23512–23522.

Tian, Z.; Shen, C.; Chen, H.; and He, T. 2019. Fcos: Fully convolutional one-stage object detection. In *ICCV*, 9627–9636.

Vaish, P.; Wang, S.; and Strisciuglio, N. 2024. Fourier-basis functions to bridge augmentation gap: Rethinking frequency augmentation in image classification. In *CVPR*, 17763–17772.

Vidit, V.; Engilberge, M.; and Salzmann, M. 2023. CLIP the Gap: A Single Domain Generalization Approach for Object Detection. In *CVPR*, 3219–3229.

VS, V.; Oza, P.; and Patel, V. M. 2023. Instance relation graph guided source-free domain adaptive object detection. In *CVPR*, 3520–3530.

Wang, A.; Chen, H.; Liu, L.; Chen, K.; Lin, Z.; Han, J.; et al. 2024a. Yolov10: Real-time end-to-end object detection. *NeurIPS*, 37: 107984–108011.

Wang, J.; Chen, Y.; Zheng, Z.; Li, X.; Cheng, M.-M.; and Hou, Q. 2024b. CrossKD: Cross-head knowledge distillation for object detection. In *CVPR*, 16520–16530.

Wang, W.; Zhou, T.; Yu, F.; Dai, J.; Konukoglu, E.; and Van Gool, L. 2021a. Exploring cross-image pixel contrast for semantic segmentation. In *ICCV*, 7303–7313.

Wang, Z.; Luo, Y.; Qiu, R.; Huang, Z.; and Baktashmotlagh, M. 2021b. Learning to diversify for single domain generalization. In *ICCV*, 834–843.

Weng, W.; and Yuan, C. 2024. Mean Teacher DETR with Masked Feature Alignment: A Robust Domain Adaptive Detection Transformer Framework. In *AAAI*, 5912–5920.

Wu, A.; and Deng, C. 2022. Single-domain generalized object detection in urban scene via cyclic-disentangled self-distillation. In *CVPR*, 847–856.

Wu, F.; Gao, J.; Hong, L.; Wang, X.; Zhou, C.; and Ye, N. 2024. G-NAS: Generalizable Neural Architecture Search for Single Domain Generalization Object Detection. In *AAAI*, 5958–5966.

Xu, M.; Qin, L.; Chen, W.; Pu, S.; and Zhang, L. 2023. Multi-view adversarial discriminator: Mine the non-causal factors for object detection in unseen domains. In *CVPR*, 8103–8112.

Xu, Q.; Zhang, R.; Zhang, Y.; Wang, Y.; and Tian, Q. 2021. A fourier-based framework for domain generalization. In *CVPR*, 14383–14392.

Yang, L.; Gu, X.; and Sun, J. 2023. Generalized semantic segmentation by self-supervised source domain projection and multi-level contrastive learning. In *AAAI*, 10789–10797.

Yoon, I.; Kwon, H.; Kim, J.; Park, J.; Jang, H.; and Sohn, K. 2024. Enhancing Source-Free Domain Adaptive Object Detection with Low-Confidence Pseudo Label Distillation. In *ECCV*, 337–353. Springer.

Yu, F.; Chen, H.; Wang, X.; Xian, W.; Chen, Y.; Liu, F.; Madhavan, V.; and Darrell, T. 2020. Bdd100k: A diverse driving dataset for heterogeneous multitask learning. In *CVPR*, 2636–2645.

Yuan, X.; Cheng, G.; Yan, K.; Zeng, Q.; and Han, J. 2023. Small object detection via coarse-to-fine proposal generation and imitation learning. In *ICCV*, 6317–6327.

Zhang, H.; Li, F.; Liu, S.; Zhang, L.; Su, H.; Zhu, J.; Ni, L. M.; and Shum, H.-Y. 2022a. Dino: Detr with improved denoising anchor boxes for end-to-end object detection. *arXiv preprint arXiv:2203.03605*.

Zhang, H.; Zhang, Y.-F.; Liu, W.; Weller, A.; Schölkopf, B.; and Xing, E. P. 2022b. Towards principled disentanglement for domain generalization. In *CVPR*, 8024–8034.

Zhang, J.; Huang, J.; Luo, Z.; Zhang, G.; Zhang, X.; and Lu, S. 2023. Da-detr: Domain adaptive detection transformer with information fusion. In *CVPR*, 23787–23798.

Zhang, P.; Cheng, G.; Lang, C.; Xie, X.; and Han, J. 2025. Nirnet: Noise incentive robust network in remote sensing object detection under cloud corruption. *IEEE TGRS*, 63: 1–13.

Zhang, Z.; Wu, A.; and Han, Y. 2025. Style Evolving along Chain-of-Thought for Unknown-Domain Object Detection. In *CVPR*, 14225–14234.

Zhao, L.; and Wang, L. 2022. Task-specific inconsistency alignment for domain adaptive object detection. In *CVPR*, 14217–14226.

Zhou, K.; Liu, Z.; Qiao, Y.; Xiang, T.; and Loy, C. C. 2022. Domain generalization: A survey. *IEEE TPAMI*, 45(4): 4396–4415.

Zhu, X.; Su, W.; Lu, L.; Li, B.; Wang, X.; and Dai, J. 2020. Deformable detr: Deformable transformers for end-to-end object detection. *ICLR*, 1–11.