

Boosting Domain Generalized and Adaptive Detection with Diffusion Models: Fitness, Generalization, and Transferability

Boyong He^{1*}, Yuxiang Ji^{1*}, Zhuoyue Tan^{1*}, Liaoni Wu^{1,2†},

¹Institute of Artificial Intelligence, Xiamen University

²School of Aerospace Engineering, Xiamen University

{boyonghe, yuxiangji, tanzhuoyue}@stu.xmu.edu.cn

wuliaoni@xmu.edu.cn[†]

Abstract

*Detectors often suffer from performance drop due to domain gap between training and testing data. Recent methods explore diffusion models applied to domain generalization (DG) and adaptation (DA) tasks, but still struggle with large inference costs and have not yet fully leveraged the capabilities of diffusion models. We propose to tackle these problems by extracting intermediate features from a single-step diffusion process, improving feature collection and fusion to reduce inference time by 75% while enhancing performance on source domains (i.e., **Fitness**). Then, we construct an object-centered auxiliary branch by applying box-masked images with class prompts to extract robust and domain-invariant features that focus on object. We also apply consistency loss to align the auxiliary and ordinary branch, balancing fitness and generalization while preventing overfitting and improving performance on target domains (i.e., **Generalization**). Furthermore, within a unified framework, standard detectors are guided by diffusion detectors through feature-level and object-level alignment on source domains (for DG) and unlabeled target domains (for DA), thereby improving cross-domain detection performance (i.e., **Transferability**). Our method achieves competitive results on 3 DA benchmarks and 5 DG benchmarks. Additionally, experiments on COCO generalization benchmark demonstrate that our method maintains significant advantages and show remarkable efficiency in large domain shifts and low-data scenarios. Our work shows the superiority of applying diffusion models to domain generalized and adaptive detection tasks and offers valuable insights for visual perception tasks across diverse domains. The code is available at [Fitness-Generalization-Transferability](#).*

1. Introduction

Distribution discrepancies between training data and real-world environments are inevitable in practical applications of object detection. Factors such as weather variations, sensor differences, diverse lighting conditions, and environmental noise contribute to domain shift between training and testing datasets. This domain gap causes even the state-of-the-art (SOTA) detectors [55, 56, 64, 81] to suffer from significant performance degradation when applied to diverse and unseen domains [66, 86].

To address domain gaps in object detection, researchers have developed two main approaches: Domain Adaptation (DA) and Domain Generalization (DG). DA leverages unlabeled target domain data through feature alignment [9, 44, 59], style transfer [21, 41, 89], and self-training with pseudo-labels [3, 4, 80]. However, since target data is often unavailable beforehand, DG methods focus on data augmentation [30, 85], domain-invariant feature extraction [13, 37], and adversarial training [39, 90] to build robust detectors using only source domain data that perform well across unseen domains.

Vision foundation models offer fresh perspectives on domain challenges. DDT [23] and GDD [24] utilize diffusion models [19, 28, 61] to build robust detectors for DA and DG tasks, outperforming previous methods. However, these approaches suffer from computational inefficiency due to large parameters and multi-step denoising processes, limiting their practical deployment. Moreover, existing frameworks fail to fully exploit the multimodal capabilities of diffusion models or develop specialized architectures for DG and DA tasks, indicating untapped potential for improving generalization and adaptation performance.

In this paper, we address these limitations through three key strategies. First, unlike DDT [23] and GDD [24] which rely on computationally expensive multi-step diffusion processes, we implement single-step feature extraction with optimized collection and fusion structures. Second, we de-

*Equal contribution.

†Corresponding author.

sign an object-centered auxiliary branch using box-masked images and class prompts, leveraging diffusion models’ multi-modal capabilities. Finally, we align regular and auxiliary branch through consistency loss, enabling domain-invariant feature learning without affecting inference speed. Compared to SOTA method GDD [24], our approach reduces inference time by **75%** while achieving **{2.7, 0.6, 3.8, 4.8, 3.3}%** mAP improvements across 5 DG benchmarks.

We also explore transferring the powerful generalization capabilities of diffusion detectors to standard detectors. Building upon the DDT [23] and GDD [24] approaches, we propose a unified transfer framework that aligns at both feature and object levels, applicable to both DA and DG tasks. Benefiting from our improved diffusion detector, diffusion-guided detectors achieve **{0.4, 1.6, 0.8, 1.8, 1.6}%** mAP improvements across 5 DG benchmarks compared to GDD [24], and **{7.9, 6.6, 1.7}%** improvements on 3 DA benchmarks compared to DDT [23].

Furthermore, we evaluate diffusion detectors on larger-scale benchmarks, training on different proportions (1% and 100%) of COCO [46] and testing across 11 cross-domain datasets. Compared to advanced architectures (ConvNeXt [50], Swin [49], ViT [16]) and pre-trained models (GLIP [40]), our method shows significant advantages, particularly in scenarios with large domain shifts or limited data. Our results on multiple DG and DA benchmarks demonstrate our approach offers an efficient solution to addressing domain gap challenges in detection tasks.

2. Related Work

2.1. DG and DA for Object Detection

Detectors suffer performance drop when deployed in environments different from training data. DA approaches address this by aligning feature distributions through adversarial learning [9, 44, 59, 73] or consistency-based learning with pseudo-labels [3, 4, 15, 44]. However, these methods require target domain data during training. To overcome this limitation, DG methods develop robust models without accessing target data, primarily through data augmentation [12, 13, 30, 85], adversarial training [39, 90], meta-learning [1, 17], and style transfer [21, 89]. Recent works extend these to detection tasks via multi-view learning [87], feature disentanglement [72], augmentation [13, 37] and causal inference [48]. Despite these advances, current methods still struggle to effectively handle the complex domain shifts present in real-world environments.

2.2. Diffusion-Based Applications

Diffusion models [19, 29, 53, 57, 58, 61] have demonstrated exceptional capabilities in image generation and representation learning [2, 74]. Their noise-adding and denoising mechanism provide natural robustness against vi-

sual perturbations [51, 63], making them promising for domain generalization tasks. Recent works have successfully applied diffusion-derived features to semantic segmentation [2], panoptic segmentation [74], and image correspondence [51, 63]. Specifically, DDT [23] and GDD [24] explored diffusion models for DA and DG detection task respectively, achieving significant improvements that demonstrate the potential of building robust diffusion-based detectors. Our approach builds upon the work [23, 24], substantially optimizing inference efficiency while enhancing both fitness and generalization capabilities, showing strong performance across complex cross-domain scenarios and various data scales.

3. Method

3.1. Preliminaries

Object Detection: Object detection aims to locate and classify objects within images. Given an input image I , a detector outputs bounding boxes $\mathcal{B} = \{b_i\}_{i=1}^N$ with corresponding class labels $\mathcal{C} = \{c_i\}_{i=1}^N$, where $c_i \in \{1, 2, \dots, K\}$ for K categories. We adopt Faster R-CNN [56] as our default detector, which first generates region proposals via a Region Proposal Network (RPN), then uses Region of Interest (ROI) head to extract features for classification and bounding box regression.

DG and DA for Object Detection: Domain Generalization (DG) and Domain Adaptation (DA) address distributional shift between domains. With source domain $\mathcal{D}_S = \{(x_i^S, y_i^S)\}_{i=1}^{N_S}$ and target domain $\mathcal{D}_T = \{(x_j^T, y_j^T)\}_{j=1}^{N_T}$, the goal is optimal performance on \mathcal{D}_T . DG works without accessing target data during training, while DA leverages unlabeled target data to adapt the model.

Diffusion Process of Diffusion Models: Diffusion models define a forward process that gradually adds Gaussian noise to data samples. This forward process transforms a data sample \mathbf{x}_0 into noise \mathbf{x}_T through a Markov chain. The transition probability is given by: $q(\mathbf{x}_t | \mathbf{x}_{t-1}) = \mathcal{N}(\mathbf{x}_t; \sqrt{1 - \beta_t} \mathbf{x}_{t-1}, \beta_t \mathbf{I})$, where $\beta_t \in (0, 1)$ controls the noise added at each step. Remarkably, we can sample \mathbf{x}_t directly from \mathbf{x}_0 using: $\mathbf{x}_t = \sqrt{\bar{\alpha}_t} \mathbf{x}_0 + \sqrt{1 - \bar{\alpha}_t} \epsilon$, where $\bar{\alpha}_t = \prod_{i=1}^t (1 - \beta_i)$ accumulates the noise schedule effects and $\epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ is standard Gaussian noise.

3.2. Feature Collection and Fusion

Collection: To optimize inference efficiency, we propose to extract rich and robust features from a single-step diffusion process rather than across multiple steps in DDT [23]. Given an input image \mathbf{x}_0 , we apply the forward diffusion process to obtain a noisy sample \mathbf{x}_t at a specific timestep t . From the noise predictor \mathcal{F}_θ , we extract two comprehensive sets of features: 12 feature groups from ResNet blocks in the UNet upsampling structure, denoted

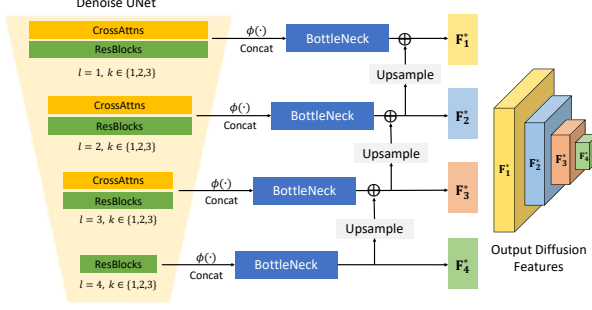


Figure 1. Feature collection and fusion from UNet on single-step diffusion process .

as $\mathbf{s}_{res}^{l,k} \in \mathbb{R}^{C_{l,k}^{res} \times H_{l,k}^{res} \times W_{l,k}^{res}}$, where $l \in \{1, 2, 3, 4\}$ indicates the layer and $k \in \{1, 2, 3\}$ specifies the block position within each layer; and 9 feature groups from cross-attention blocks, denoted as $\mathbf{s}_{att}^{l,k} \in \mathbb{R}^{C_{l,k}^{att} \times H_{l,k}^{att} \times W_{l,k}^{att}}$, where $l \in \{1, 2, 3\}$ and $k \in \{1, 2, 3\}$. This focused extraction strategy captures the multi-scale semantic information while maintaining computational efficiency as shown in the left part of Fig. 1.

Fusion: Our feature fusion approach leverages the hierarchical structure as shown in Fig. 1. From upsampling stages, we obtain features at four different spatial resolutions \mathbf{F}_l for $l \in \{1, 2, 3, 4\}$. Different from processing each layer separately as in GDD [24], we concatenate all features of the same scale into a unified representation: $\mathbf{F}_l^\oplus = \phi(\mathbf{s}_{res}^{l,1}, \mathbf{s}_{res}^{l,2}, \mathbf{s}_{res}^{l,3}, \mathbf{s}_{att}^{l,1}, \mathbf{s}_{att}^{l,2}, \mathbf{s}_{att}^{l,3})$ for each layer l , where $\phi(\cdot)$ denotes channel-wise concatenation. Each concatenated feature group then undergoes a single bottleneck projection: $\mathbf{F}_l^p = \mathcal{B}(\mathbf{F}_l^\oplus)$, where $\mathcal{B}(\cdot)$ represents the bottleneck operation that projects features to dimensions $C_l = 256 \times 2^{l-1}$. Then we implement skip connections by adding features of matching resolutions: $\mathbf{F}_{l-1}^* = \mathcal{U}(\mathbf{F}_l^p) + \mathbf{F}_{l-1}^p$, where $\mathcal{U}(\cdot)$ denotes the upsampling operation, preserving fine-grained details that might otherwise be lost with direct upsampling in GDD [24]. The final feature pyramid $\mathbf{F}^{final} = \{\mathbf{F}_l^*\}_{l=1}^4$ with $\mathbf{F}_l^* \in \mathbb{R}^{C_l \times H/2^{l+1} \times W/2^{l+1}}$ (where $C_l = 256 \times 2^{l-1}$ for $l \in \{1, 2, 3, 4\}$) aligns structurally with standard ResNet [25] outputs, ensuring compatibility with detection heads while effectively compensating for the performance limitations of single-step feature extraction.

3.3. Dual-branch of Diff. Detector

Ordinary Branch: Our diffusion backbone extracts features as described above to construct the *Diff. Detector* (\mathcal{F}_{diff}). Given source domain images x^S and labels y^S , we process inputs through the frozen diffusion structure and trainable feature extraction components to generate the feature pyramid \mathbf{F}_{ord}^{final} . These features feed into RPN and ROI heads, the training objective is:

$$\mathcal{L}_{ord} = \mathcal{L}(\mathcal{F}_{diff}(x^S), y^S) \quad (1)$$

where \mathcal{L} combines the classification, bounding box regression, and region proposal losses of Faster R-CNN, as illustrated in the bottom panel of Fig. 2.

Object-centered Auxiliary Branch: Diffusion models demonstrate powerful multimodal understanding capabilities through text-conditioned image generation, yet this advantage remains unexplored in DDT [23] and GDD [24]. We propose an object-centered auxiliary branch that leverages the cross-modal capabilities of diffusion to enhance feature representation and generalization for detection.

Specifically, our Object-centered Auxiliary Branch takes source domain images x^S along with their corresponding bounding box masks m^S and class prompts p^S as inputs. First, we generate object-centered images via $x_{mask}^S = x^S \odot m^S$, where \odot represents element-wise multiplication and m^S is a binary mask derived from ground-truth bounding boxes, with non-object regions set to zero. We then feed x_{mask}^S and class prompts p^S into the diffusion model’s conditioning process, where $\mathbf{z} = \mathcal{E}(x_{mask}^S)$ is the latent representation from the image encoder and $\mathbf{c} = \mathcal{T}(p^S)$ is the semantic embedding from the text encoder. The cross-attention mechanism facilitates interaction between textual conditions and image features, focusing attention on regions relevant to the prompted class.

This process enables the diffusion model to generate features with enhanced focus on specified object categories. We extract features \mathbf{F}_{aux}^{final} from the same positions as in the ordinary branch and apply identical detection heads as shown in the top part of Fig. 2. The objective function for this auxiliary branch is:

$$\mathcal{L}_{aux} = \mathcal{L}(\mathcal{F}_{diff}(x_{mask}^S, p^S), y^S) \quad (2)$$

By fully leveraging ground-truth labels during training, this branch provides additional supervision that better exploits the multimodal capabilities of diffusion models, leading to more domain-invariant and object-centered representations.

3.4. Dual-branch Consistency Loss

To promote stronger domain-invariant feature learning in the ordinary branch, we align its features and ROI outputs with those from the Object-centered branch. This alignment enables the inference-time model, which relies solely on the ordinary branch, to benefit from domain-invariant features, thereby enhancing generalization on unseen domains. Specifically, for features \mathbf{F}_{ord}^{final} and \mathbf{F}_{aux}^{final} from the two branches, we align them using mean squared error loss:

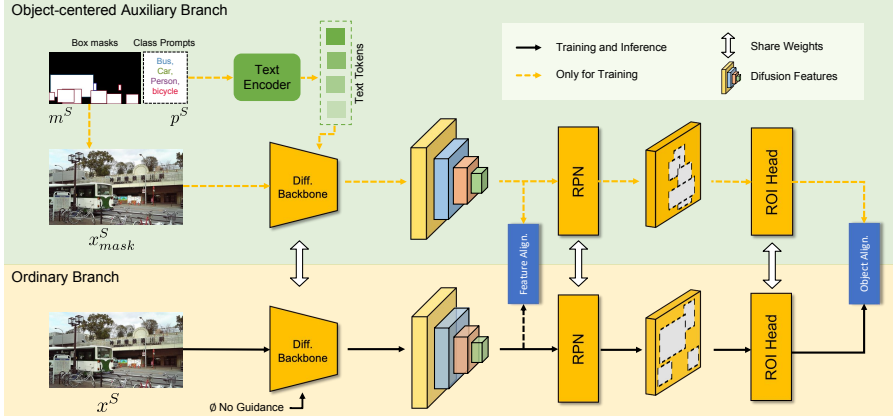


Figure 2. Our proposed dual-branch detection framework: Object-centered Auxiliary Branch (**top**) and Ordinary Detection Branch (**bottom**), unified through novel feature-level and object-level consistency alignments.

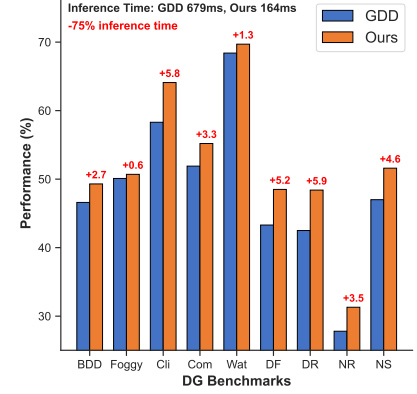


Figure 3. Performance comparison with GDD [24] across DG benchmarks. Our method shows improvements with 75% less inference time.

$$\mathcal{L}_{feature} = \|\mathbf{F}_{ord}^{final} - \mathbf{F}_{aux}^{final}\|_2^2 \quad (3)$$

Furthermore, referencing CrossKD [67], we design cross-head alignment to align the ROI outputs from both branches, including bounding box alignment and class alignment. The ROI outputs consist of \mathbf{B}_{ord} and \mathbf{B}_{aux} for bounding box predictions, and \mathbf{C}_{ord} and \mathbf{C}_{aux} for classification logits from the ordinary and auxiliary branches respectively:

$$\mathcal{L}_{box} = |\mathbf{B}_{ord} - \mathbf{B}_{aux}| \quad (4)$$

$$\mathcal{L}_{cat} = \sum_i P_{ord}^\tau(i) \log \frac{P_{ord}^\tau(i)}{P_{aux}^\tau(i)} \quad (5)$$

where P_{ord}^τ and P_{aux}^τ represent the softened probability distributions obtained by applying softmax with temperature τ to the classification logits \mathbf{C}_{ord} and \mathbf{C}_{aux} respectively.

The full alignment objective is formulated as a weighted combination of feature and output consistency:

$$\mathcal{L}_{cons} = \mathcal{L}_{feature} + \gamma \cdot (\mathcal{L}_{box} + \mathcal{L}_{cat}) \quad (6)$$

where γ is a weighting parameter that balances the importance between feature-level and output-level alignments.

3.5. Full Objective

Combining the ordinary branch detection loss, auxiliary branch detection loss, and consistency loss, the full training objective of our diff. detector is formulated as:

$$\mathcal{L}_{total} = \mathcal{L}_{ord} + \mathcal{L}_{aux} + \lambda \mathcal{L}_{cons} \quad (7)$$

where λ is a weighting parameter that controls the contribution of the consistency regularization between the two branches.

3.6. Dual-branch and Consistency Loss for Generalization

From a representation learning perspective, achieving domain generalization requires disentangling task-relevant domain-invariant features from domain-specific features. We can formally decompose image features as:

$$\Phi(x) = \alpha \Phi_{inv}(x) + \Phi_{spe}(x) \quad (8)$$

where $\Phi_{inv}(x)$ represents domain-invariant features essential for detection tasks (e.g., object geometry, semantic attributes), $\Phi_{spe}(x)$ represents domain-specific features (e.g., lighting conditions, background environments), and α is a coefficient indicating the contribution weight of invariant features.

Dual-branch Mechanism for Feature Disentanglement:

Our dual-branch architecture implements this disentanglement principle through complementary learning paths. The Ordinary Branch processes complete images, initially capturing both $\Phi_{inv}(x)$ and $\Phi_{spe}(x)$ with a smaller α , while the Object-centered Branch emphasizes $\Phi_{inv}(x)$ with a larger α while suppressing $\Phi_{spe}(x)$ through object masks m^S and class prompts p^S , such that $\Phi_{aux}(x_{mask}^S, p^S)$ primarily contains $\Phi_{inv}(x)$ with enhanced α .

Consistency Loss as Knowledge Distillation: The consistency loss functions as a knowledge transfer mechanism that guides the ordinary branch toward domain-invariant representations. By aligning features and detection outputs between branches, we guide the ordinary branch to amplify its domain-invariant features by increasing α :

$$\mathcal{F}_{diff}(x^S) \xrightarrow{\mathcal{L}_{cons}} \mathcal{F}_{diff}(x_{mask}^S, p^S) \quad (9)$$

From a risk minimization perspective, minimizing these consistency losses helps reduce the domain generalization

error bound:

$$\mathcal{R}_T(h) \leq \mathcal{R}_S(h) + d_{\mathcal{H}}(\mathcal{D}_S, \mathcal{D}_T) + \delta \quad (10)$$

The consistency loss reduces $d_{\mathcal{H}}(\mathcal{D}_S, \mathcal{D}_T)$ by encouraging features with domain-invariant components, ensuring predictions become invariant across domains d_A and d_B :

$$\min_{\mathcal{F}_{\text{diff}}} \|\mathbb{E}[Y|\mathcal{F}_{\text{diff}}(X), D = d_A] - \mathbb{E}[Y|\mathcal{F}_{\text{diff}}(X), D = d_B]\| \quad (11)$$

Additionally, this loss serves as a regularization term that mitigates overfitting on source domains:

$$\min_{\mathcal{F}_{\text{diff}} \in \mathcal{F}} \mathcal{L}_{\text{task}}(\mathcal{F}_{\text{diff}}(x^S)) + \lambda \cdot \mathcal{L}_{\text{cons}}(\mathcal{F}_{\text{diff}}(x^S), \mathcal{F}_{\text{diff}}(x_{\text{mask}}^S; p^S)) \quad (12)$$

This constraint effectively shrinks the hypothesis space, with both branches sharing the same feature extractor of $\mathcal{F}_{\text{diff}}$ but operating on different inputs.

3.7. Unified Transfer Framework for DG and DA

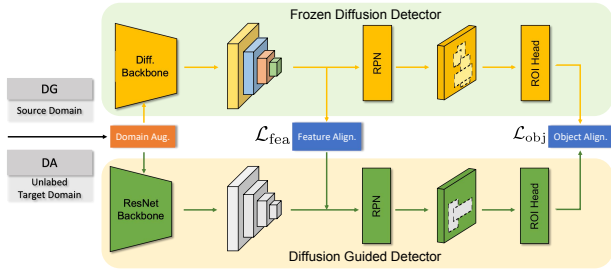


Figure 4. Unified transfer framework for DG and DA with feature- and object-level alignment.

Given the strong generalization capabilities demonstrated by the diffusion detector, we aim to transfer these capabilities to ordinary detectors as shown in Fig. 4. We present a unified transfer framework that consolidates approaches from both DDT [23] and GDD [24] and follow their settings, establishing alignment at feature (\mathcal{L}_{fea}) and object levels (\mathcal{L}_{obj}), as:

$$\mathcal{L}_{\text{transfer}} = \mathcal{L}_{\text{det}} + \mathcal{L}_{\text{fea}} + \mathcal{L}_{\text{obj}} \quad (13)$$

where \mathcal{L}_{det} is detection loss from supervised learning on source domain.

This framework adapts to each task differently. For DA, it generates pseudo-labels in the target domain for object-level alignment while aligning representations for feature-level alignment. For DG, both alignments occur only within source domains, creating a knowledge distillation process that transfers domain-invariant qualities from diffusion to ordinary detector.

4. Experiments

4.1. DG and DA Benchmarks

(1) **Cross Camera:** Cityscapes [11] (2,975 training images from 50 cities) to BDD100K [79] day-clear split with 7 shared categories following SWDA [59].

(2) **Adverse Weather:** Cityscapes to FoggyCityscapes [60] with the most challenging 0.02 split to evaluate robustness under degraded visibility conditions.

(3) **Real to Artistic:** VOC [20] (16,551 real-world images) to Clipart (1K images, 20 categories), Comic (2K images, 6 categories), and Watercolor (2K images, 6 categories) [34] following AT [44].

(4) **Diverse Weather Datasets:** Daytime-Sunny (26,518 images) to Night-Sunny (26,158 images), Night-Rainy (2,494 images), Dusk-Rainy (3,501 images), and Daytime-Foggy (3,775 images) following [71] and [37].

(5) **Corruption Benchmark:** Cityscapes-C [52] with 15 corruption types (noise, blur, weather, and digital perturbations) at 5 severity levels following OADG [37].

We test our DG and DA methods on benchmarks (1)-(5) and (1)-(3) respectively, while comparing with existing works. Additionally, we propose a larger-scale DG benchmark: **COCO Generalization Benchmark**, trained on COCO2017 [46] dataset and tested on 11 various datasets as shown in Tab. 6.

4.2. Implementation Details

Experimental Settings: Our experimental settings generally align with DDT [23] and GDD [24]. Specifically, all experiments are implemented on MMDetection [6]. We consistently use SGD optimizer with a learning rate of 0.02 and a total batch size of 16, training for 20K iterations on two 4090 GPUs. For diff. detector (*Diff. Detector*), we apply frozen official weights from Stable Diffusion v1.5 (*SD-1.5*) and v2.1 (*SD-2.1*) provided by StabilityAI. In our diffusion guided DG and DA experiments, we employ Faster R-CNN [56] with R101 [25] as our baseline detector (*Diff. Guided, R101*). We report AP_{50} for each category and mAP across all categories. For Cityscapes-C [52], we report mPC (average $\text{AP}_{50:95}$ across 15 corruptions with 5 levels) following [37] as shown in Tab. 5.

Data Augmentations: We follow GDD’s [24] domain augmentation strategy (*Domain Aug.*) in all experiments, using image-level (color and spatial transformations) and domain-level augmentations (*FDA* [78], *Histogram Matching*, and *Pixel Distribution Matching*). For DG experiments, we apply these augmentations only on source domains, while for DA experiments, we implement them between source and target domains. *Code and more detailed settings are provided in supplementary materials, along with additional experimental analyses, class-wise results, and qualitative visualizations.*

4.3. Results and Comparisons

Table 1. DG and DA Results (%) on BDD100K.

Methods	Bike	Bus	Car	Motor	Psn.	Rider	Truck	mAP
<i>DG methods (without target data)</i>								
CDS [71] (CVPR'22)	22.9	20.5	33.8	14.7	18.5	23.6	18.2	21.7
SHADE [84] (ECCV'22)	25.1	19.0	36.8	18.4	24.1	24.9	19.8	24.0
MAD [75] (CVPR'23)	-	-	-	-	-	-	-	28.0
SRCD [54] (TNNLS'24)	24.8	21.5	38.7	19.0	25.7	28.4	23.1	25.9
DDT (SD-1.5) [23] (MM'24)	-	-	-	-	-	-	-	32.7
GDD (SD-1.5) [24] (CVPR'25)	38.9	31.0	71.5	37.6	61.5	47.0	38.5	46.6
GDD (R101) [24] (CVPR'25)	38.4	33.4	72.0	38.3	60.3	47.0	35.0	46.3
Ours (Diff. Detector, SD-1.5)	41.2	41.7	72.7	37.2	62.8	48.7	40.5	49.3
Ours (Diff. Guided, R101)	39.4	34.1	72.2	37.4	61.3	46.9	35.7	46.7
<i>DA methods (with unlabeled target data)</i>								
EPM [32] (ECCV'20)	20.1	19.1	55.8	14.5	39.6	26.8	18.8	27.8
TDD [26] (CVPR'22)	28.8	25.5	53.9	24.5	39.6	38.9	24.1	33.6
PT [8] (ICML'22)	28.8	33.8	52.7	23.0	40.5	39.9	25.8	34.9
SIGMA [42] (CVPR'22)	26.3	23.6	64.1	17.9	46.9	29.6	20.2	32.7
SIGMA++ [43] (TPAMI'23)	27.1	26.3	65.6	17.8	47.5	30.4	21.1	33.7
NSA [88] (ICCV'23)	-	-	-	-	-	-	-	35.5
HT [15] (CVPR'23)	38.0	30.6	63.5	28.2	53.4	40.4	27.4	40.2
MTM [70] (AAAI'24)	28.0	28.8	68.8	23.8	53.7	35.1	23.0	37.3
CAT [36] (CVPR'24)	34.6	31.7	61.2	24.4	44.6	41.5	31.4	38.5
DDT (R101) [23] (MM'24)	40.3	32.3	66.7	31.8	59.1	41.6	31.8	43.4
Ours (Diff. Guided, R101)	43.6	42.9	75.2	40.5	64.6	49.6	42.9	51.3

Table 2. DG and DA Results (%) on FoggyCityscapes.

Methods	Bus	Bike	Car	Motor	Psn.	Rider	Train	Truck	mAP
<i>DG methods</i>									
DIDN [45] (CVPR'21)	35.7	33.1	49.3	24.8	31.8	38.4	26.5	27.7	33.4
FACT [76] (CVPR'21)	27.7	31.3	35.9	23.3	26.2	41.2	3.0	13.6	25.3
FSDR [33] (CVPR'22)	36.6	34.1	43.3	27.1	31.2	44.4	11.9	19.3	31.0
MAD [75] (CVPR'23)	44.0	40.1	45.0	30.3	34.2	47.4	42.4	25.6	38.6
DDT (SD-1.5) [23] (MM'24)	-	-	-	-	-	-	-	-	36.1
GDD (SD-1.5) [24] (CVPR'25)	56.2	50.4	66.7	39.9	50.2	59.5	39.9	38.0	50.1
GDD (R101) [24] (CVPR'25)	53.8	54.2	67.5	45.6	52.1	60.8	53.9	32.4	52.5
Ours (Diff. Detector, SD-1.5)	53.0	55.4	68.1	42.1	51.0	59.9	39.4	36.4	50.7
Ours (Diff. Guided, R101)	55.3	62.7	68.2	45.5	52.9	62.0	48.4	37.4	54.1
<i>DA methods</i>									
MIC [31] (CVPR'23)	52.4	47.5	67.0	40.6	50.9	55.3	33.7	33.9	47.6
SIGMA++ [43] (TPAMI'23)	52.2	39.9	61.0	34.8	46.4	45.1	44.6	32.1	44.5
CIGAR [47] (CVPR'23)	56.6	41.3	62.1	33.7	46.1	47.3	44.3	27.8	44.9
CMT [3] (CVPR'23)	66.0	51.2	63.7	41.4	45.9	55.7	38.8	39.6	50.3
HT [15] (CVPR'23)	55.9	50.3	67.5	40.1	52.1	55.8	49.1	32.7	50.4
DSCA [22] (PR'24)	54.7	54.9	68.4	40.1	55.4	61.0	35.9	35.1	50.7
UMGA [82] (TPAMI'24)	58.0	43.7	64.9	38.3	47.9	50.1	45.6	34.8	47.9
CAT [36] (CVPR'24)	66.0	53.0	63.7	44.9	44.6	57.1	49.7	40.8	52.5
MTM [70] (AAAI'24)	54.4	47.7	67.2	38.4	51.0	53.4	41.6	37.2	48.9
DDT (R101) [23] (MM'24)	53.5	52.2	64.2	43.5	50.9	60.0	42.4	33.6	50.0
Ours (Diff. Guided, R101)	56.9	66.2	71.7	49.1	55.6	63.1	48.9	41.1	56.6

We present our DA results in Tab. 1, 2, and 3, while DG results are shown in Tab. 1, 2, 3, 4, and 5, with comparisons to SOTA methods. The **bold** values indicate the best results, and **Yellow Background** highlights methods achieving the best average performance. To save space in tables, we use the following dataset abbreviations: **City.** (Cityscapes), **BDD.** (BDD100K), **Foggy.** (FoggyCityscapes), **Cli.** (Clipart), **Com.** (Comic), **Wat.** (Watercolor), **DF** (Daytime-Foggy), **DR** (Dusk-Rainy), **NR** (Night-Rainy), and **NS** (Night-Sunny).

Table 3. DG and DA Results (%) on Clipart, Comic, and Watercolor.

DG Methods			DA Methods		
Methods	Cli.	Com. Wat.	Methods	Cli.	Com. Wat.
Div. [13] (CVPR'24)	33.7	25.5 52.5	SWDA [59] (CVPR'19)	38.1	29.4 53.3
DivAlign [13] (CVPR'24)	38.9	33.2 57.4	UMT [14] (CVPR'21)	44.1	- 58.1
DDT (SD-1.5) [23] (MM'24)	47.4	44.4 58.7	SADA [10] (IJCV'21)	43.3	- 56.0
GDD (SD-1.5) [24] (CVPR'25)	58.3	51.9 68.4	DBGL [5] (ICCV'21)	41.6	29.7 53.8
GDD (R101) [24] (CVPR'25)	40.8	29.7 54.2	AT [44] (CVPR'22)	49.3	- 59.9
Ours (Diff. Detector, SD-1.5)	64.1	55.2 69.7	D-ADAPT [35] (ICLR'22)	49.0	40.5 -
Ours (Diff. Guided, R101)	40.5	30.0 56.6	TIA [83] (CVPR'22)	46.3	- -
			LODS [41] (CVPR'22)	45.2	- 58.2
			CIGAR [47] (CVPR'23)	46.2	- -
			CMT [3] (CVPR'23)	47.0	- -
			DAVimNet [18] (ArXiv'24)	43.8	- 54.8
			UMGA [82] (TPAMI'24)	49.9	- 62.1
			CAT [36] (CVPR'24)	49.1	- -
			DDT (R101) [23] (MM'24)	55.6	50.2 63.7
			Ours (Diff. Guided, R101)	58.2	50.5 68.0

Table 4. DG Results (%) on Diverse Weather Datasets.

Methods	DF	DR	NR	NS	Average
CDS [71] (CVPR'22)	33.5	28.2	16.6	36.6	28.7
SHADE [84] (ECCV'22)	33.4	29.5	16.8	33.9	28.4
CLIPgap [65] (CVPR'23)	32.0	26.0	12.4	34.4	26.2
SRCD [54] (TNNLS'24)	35.9	28.8	17.0	36.7	29.6
G-NAS [72] (AAAI'24)	36.4	35.1	17.4	45.0	33.5
PhysAug [77] (ArXiv'24)	40.8	41.2	23.1	44.9	37.5
OA-DG [37] (AAAI'24)	38.3	33.9	16.8	38.0	31.8
DivAlign [13] (CVPR'24)	37.2	38.1	24.1	42.5	35.5
UFR [48] (CVPR'24)	39.6	33.2	19.2	40.8	33.2
Prompt-D [38] (CVPR'24)	39.1	33.7	19.2	38.5	32.6
DIDM [27] (ArXiv'25)	39.3	35.4	19.2	42.0	34.0
GDD (SD-1.5) [24] (CVPR'25)	43.3	42.5	27.8	47.0	40.2
GDD (R101) [24] (CVPR'25)	44.7	37.4	21.7	48.7	38.1
Ours (Diff. Detector, SD-1.5)	48.5	48.4	31.3	51.6	45.0
Ours (Diff. Guided, R101)	46.7	39.1	22.4	50.5	39.7

4.3.1. DG Results of Diff. Detector

Tab. 1, 2, 3, 4, and 5 show diff. detector outperforms SOTA GDD [24] across all benchmarks. Both approaches significantly surpass other recent methods [13, 37, 38, 48, 54, 65, 72, 75], confirming diffusion models' effectiveness for DG. Notably, our method achieves **3.0%** average mAP improvement while reducing inference time by **75%** compared to GDD [24] as shown in Fig. 3.

4.3.2. DG Results of Diff. Guided Detector

Tab. 1, 2, 3, 4, and 5 show the results of ordinary detectors (Faster R-CNN with R101) guided by diff. detector. Following the settings in GDD [24], our methods achieve improvements **{0.4, 1.6, 0.8, 1.8, 1.8}**% mAP improvements on 5 DG benchmarks compared to GDD, indicating that a stronger diff. detector typically provides better guidance.

4.3.3. DA Results of Diff. Guided Detector

Similarly, we adopt the diff. detector as a teacher model to generate pseudo-labels on unlabeled target domains and guides the student model through semi-supervised learn-

Table 5. Generalization Detection Results (%) on Cityscapes-Corruption Benchmark.

Methods	Clean	Noise			Blur				Weather			Digital					mPC \uparrow
		Gauss.	Shot	Impulse	Defocus	Glass	Motion	Zoom	Snow	Frost	Fog	Bright	Contrast	Elastic	JPEG	Pixel	
FSCE [62] (CVPR'21)	43.1	7.4	10.2	8.2	23.3	20.3	21.5	4.8	5.6	23.6	37.1	38.0	31.9	40.2	20.4	23.2	21.0
OA-Mix [37] (AAAI'24)	42.7	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 [37] (AAAI'24)	43.4	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
GDD (SD-1.5) [24] (CVPR'25)	34.7	20.3	23.2	17.2	26.8	21.7	23.7	3.4	16.6	24.2	32.5	34.4	30.6	33.7	29.1	24.4	24.1
GDD (R50) [24] (CVPR'25)	42.1	11.0	13.6	10.8	25.0	14.2	21.4	3.4	5.4	24.0	39.6	40.3	36.3	39.2	18.9	16.0	21.3
Ours (Diff. Detector, SD-1.5)	40.1	23.1	26.6	20.6	29.7	24.5	25.5	4.1	18.3	28.2	37.2	39.2	35.5	37.4	32.7	28.8	27.4
Ours (Diff. Guided, R50)	41.5	13.9	16.6	13.5	24.8	14.2	22.0	3.7	5.9	24.7	40.0	40.4	35.9	39.5	18.9	17.1	22.1

Table 6. Comparison of DG Performance on COCO Generalization Benchmark Under Different Training Data Scales. * indicates we don't apply object alignment in consistency loss in FCOS and DINO detectors.

Settings	Models	COCO Val	VOC	City.	BDD.	Foggy.	Cli.	Com.	Wat.	DF	DR	NR	NS	Avg. on 11	Inf. time (ms)
1% COCO, Faster R-CNN	ResNet-R50 [25]	20.4	36.4	26.6	18.6	8.2	10.1	6.6	14.3	10.6	5.9	1.1	6.5	13.2	-
	ConvNeXt-Base [50]	32.5	53.9	34.9	28.2	29.8	18.8	14.9	26.9	25.6	16.2	5.9	18.6	24.9	-
	Swin-Base [49]	27.7	46.0	29.0	23.1	16.3	12.9	8.1	19.8	16.8	10.3	2.1	9.0	17.6	-
	VIT-Base [16]	28.1	53.3	18.0	18.0	8.8	11.1	6.1	11.4	13.8	11.1	2.6	6.9	14.6	-
	GLIP (Swin-Tiny) [40]	32.5	53.9	36.1	27.5	19.6	18.6	13.0	16.1	20.1	14.5	4.1	13.7	21.6	-
	Ours (SD-1.5)	33.4	70.4	36.5	27.9	31.0	44.0	37.8	50.1	26.6	19.8	10.7	20.8	34.1	-
100% COCO, Faster R-CNN	ResNet-R50 [25]	58.1	84.0	49.5	45.8	35.2	32.6	23.7	40.8	33.3	24.8	10.7	30.3	37.3	27
	ConvNeXt-Base [50]	64.5	83.4	53.4	50.9	43.1	43.5	35.5	47.9	40.5	34.4	17.6	36.0	44.2	54
	Swin-Base [49]	61.5	79.6	52.9	48.0	38.9	37.8	29.4	41.7	37.6	32.1	14.4	33.0	40.5	55
	VIT-Base [16]	62.7	86.3	35.1	40.4	22.9	38.3	24.1	48.9	31.2	29.1	12.4	26.3	35.9	78
	GLIP (Swin-Tiny) [40]	62.0	79.9	54.0	48.6	41.4	38.9	29.6	40.3	37.9	31.5	15.0	34.9	41.1	31
	Ours (SD-1.5)	67.0	86.6	54.1	51.2	45.8	64.6	51.9	66.6	41.7	39.2	21.7	40.7	51.3	164
100% COCO, SD-1.5	FCOS* [64]	64.3	85.0	50.3	48.3	42.0	62.9	52.1	65.9	39.6	36.4	19.7	39.0	49.2	171
	DINO* [81]	68.3	88.2	51.2	51.1	45.9	61.7	50.8	63.6	42.0	39.3	22.0	40.2	50.5	232

ing. Following the same settings as DDT [23], our method achieves {7.9, 6.6, 1.7}% improvements on 3 DA benchmarks.

4.3.4. COCO Benchmark Evaluation

As shown in Tab. 6, our diff. detector outperforms [16, 25, 40, 49, 50] on the **COCO Generalization Benchmark** using both 1% and 100% training data. Our method excels particularly in **Data-Scarce Scenarios, Extreme Domain Shifts** (Clipart, Comic, Watercolor [34], and Night-Rainy), demonstrating its suitability for limited data and substantial domain gaps. Results remain consistent across different detectors (**FCOS** [64] and **DINO** [81]), confirming our framework's broad applicability.

4.4. Failure Cases Analysis

Although diff. detector and diff. guided detector achieved significant improvements on DG and DA benchmarks, the ordinary detector guided by diff. detector showed limited gains in *Real to Artistic* scenarios, underperforming compared to GDD [24] (by -0.3% on Clipart) and DivAlign [13] (by -3.2% on Comic and -0.8% on Watercolor). While generating pseudo-labels works efficiently for DA tasks, frameworks that align solely on source domains may fail when target domains remain unseen with enormous domain gaps.

5. Ablation Studies

5.1. Studies on Diffusion Detector

Components of Diff. Detector: We present ablation studies in Tab. 7. *Domain Augmentation, Feature Collection, and Feature Fusion* improve both source domain fitness and target domain generalization. The *Auxiliary Branch* builds a more robust diffusion detector, while the *Consistency Loss* enhances cross-domain generalization with minimal impact on source domain accuracy, validating its role in balancing fitness and generalization. The Inference Time results show *Feature Collection* and *Feature Fusion* mitigate accuracy drops from single-step diffusion efficiently. *Domain Augmentation, Auxiliary Branch, and Consistency Loss* only apply during training with no impact on inference speed.

Studies of Loss Weights γ and λ : As shown in Tab. 8, both γ and λ demonstrate stable performance within [0.5, 1.5], because L_{cons} gradually decreases during the later training steps, minimally affecting the final results.

Results and Comparisons of SD-2.1: We present results using SD-2.1 weights in Tab. 9. Similar to SD-1.5, our method demonstrates consistent improvements across all settings compared to DDT [23] (average **12.3%** on diff. detector and **4.1%** on diff. guided DA) and GDD [24] (average **5.4%** on diff. detector and **3.1%** on diff. guided DG), confirming the generality of our approach.

Table 7. Ablation Studies of Proposed Components on *Diff. Detector*. * indicates we test the inference time by applying same settings on 4090 GPU with scale (1333, 800), which may differ from GDD [24] reported.

Detector	Domain Aug.	Feature Coll.	Feature Fusion	Aux. Branch	Consist. Loss	In-Domain		Cross-Domain			Inf. time (ms)
						City.	VOC	Foggy.	BDD.	Cli.	
w/o Noise Adding	-	-	-	-	-	37.4	74.9	30.8	29.1	38.2	132
w/ Noise Adding	-	-	-	-	-	38.5	74.7	34.9	34.3	43.6	132
Ours (T = 1) w/ Noise Adding	✓	-	-	-	-	40.9+2.4	75.2+0.5	39.4+4.5	36.1+1.8	47.2+3.6	132
	-	✓	-	-	-	44.3+5.8	77.3+2.6	37.8+2.9	35.2+0.9	46.6+3.0	144
	-	-	✓	-	-	46.1+7.6	79.4+4.7	38.6+3.7	36.1+1.8	49.4+5.8	151
	-	✓	✓	-	-	52.4+13.9	81.3+6.6	40.1+5.2	37.9+3.6	51.2+7.6	164
	-	-	-	✓	-	46.9+8.4	78.9+4.2	40.4+5.5	39.0+4.7	53.3+9.7	164
	-	-	-	✓	✓	46.0+7.5	78.8+4.1	43.2+8.3	41.3+7.0	57.6+14.0	164
GDD [24]	T = 5, w/ Noise Adding					59.8	84.8	50.1	46.6	58.3	789 / 679*
	T = 1, w/o Noise Adding					36.8-23.0	74.2-10.6	30.8-19.3	28.6-18.0	37.4-20.9	270 / 194*

Table 9. Comprehensive Results (%) of *SD-2.1* Version.

Methods	BDD.	Foggy.	Cli.	Com.	Wat.	DF	DR	NR	NS
<i>Diff. Detector, DG Settings</i>									
DDT (<i>SD-2.1</i>) [23]	34.6	-	45.4	42.8	58.7	-	-	-	-
GDD (<i>SD-2.1</i>) [24]	45.8	48.3	51.7	46.6	62.1	44.6	41.6	23.2	46.4
Ours (<i>SD-2.1</i>)	48.0	50.3	59.7	54.5	68.6	48.7	47.3	29.8	51.8
+Gain	+2.2	+2.0	+8.0	+7.9	+6.5	+4.1	+5.7	+6.6	+5.4
<i>Diff. Guided Detector, DG Settings</i>									
GDD (<i>SD-2.1</i>) [24]	46.1	51.0	32.7	24.9	50.6	44.7	37.1	20.0	49.3
Ours (<i>SD-2.1</i>)	46.5	54.4	38.4	30.0	56.3	46.7	39.1	22.4	50.5
+Gain	+0.4	+3.4	+5.7	+5.1	+5.7	+2.0	+2.0	+2.4	+1.2
<i>Diff. Guided Detector, DA Settings</i>									
DDT (<i>SD-2.1</i>) [23]	42.3	-	53.7	48.9	63.3	-	-	-	-
Ours (<i>SD-2.1</i>)	51.7	56.6	56.3	50.2	66.2	-	-	-	-
+Gain	+9.4	-	+2.6	+1.3	+2.9	-	-	-	-

Table 10. Ablation Study on Diff. Guided Detector for *DG*.

Domain Aug.	Feature Alignment	Object Alignment	Unseen Target Domain					
			BDD.	Foggy.	DF	DR	NR	NS
-	-	-	25.4	30.7	28.8	24.1	12.4	31.4
✓	-	-	36.2+10.8	47.9+17.2	39.9+11.1	34.8+10.7	16.1+3.7	41.2+9.8
✓	✓	-	39.5+3.3	48.9+1.0	42.1+2.2	35.2+0.4	18.2+2.1	43.0+1.8
✓	✓	✓	46.7+7.2	54.1+5.2	46.7+4.6	39.1+3.9	22.4+4.2	50.5+7.5

Table 11. Ablation Study on Diff. Guided Detector for *DA*.

Domain Aug.	Feature Alignment	Object Alignment	Unlabeled Target Domain				
			BDD.	Foggy.	Cli.	Com.	Wat.
-	-	-	25.4	30.7	27.2	18.1	41.5
✓	-	-	38.1+12.7	48.9+18.2	39.4+12.2	27.4+9.3	50.9+9.4
✓	✓	-	40.2+2.1	49.6+0.7	42.1+2.7	33.9+6.5	52.8+1.9
✓	✓	✓	51.3+11.1	56.6+7.0	58.2+16.1	50.5+16.6	68.0+15.2

5.2. Diff. Guided Detector for DG and DA

We present the Diff. guided detector’s performance in DG and DA tasks (Tab. 10, 11). *Domain Augmentation* brings significant improvements (average **11.4%**), while the diff. detector enhancement is relatively minor (average **3.3%**, Tab. 7). With unlabeled target domain data, pseudo-label generation (*Object Alignment*) is crucial, bringing aver-

Table 8. Studies of loss weights γ in \mathcal{L}_{cons} (Equ. 6) and λ in \mathcal{L}_{total} (Equ. 7).

γ	Foggy.	BDD.	Cli.
0.0	47.6	45.9	61.2
0.5	50.1	49.0	63.7
1.0	50.7	49.3	64.1
1.5	50.8	49.3	64.2
2.0	49.1	49.1	62.7

λ	Foggy.	BDD.	Cli.
0.0	45.8	44.7	59.8
0.5	50.3	48.9	63.1
1.0	50.7	49.3	64.1
1.5	50.6	49.6	64.0
2.0	50.1	48.1	63.8

age **13.1% mAP**. Without target domain data, both *Object Alignment* and *Feature Alignment* are essential, contributing average **5.43%** and **1.8%** mAP respectively.

5.3. Limitations

Although our work improves diffusion detector inference efficiency, we do not implement engineering techniques like model distillation or TensorRT acceleration due to time constraints. Our method is only tested on official Stable-Diffusion weights, without exploring specialized instance-related models [7, 68, 69] that might better fit detection tasks. Additionally, our experiments in Tab. 3 reveal that in DG tasks where target domain data is inaccessible, transferring generalization from diffusion detectors to ordinary detectors still faces limited improvement, indicating that more efficient transfer structures deserve further investigation.

6. Conclusion

In this paper, we present a framework for domain generalized and adaptive detection using diffusion models. Our approach optimizes single-step feature collection and fusion structures to reduce inference time by **75%**, incorporates an object-centered auxiliary branch and consistency loss to enhance generalization for diff. detector. Then we transfer the generalization capabilities to standard detectors through feature- and object-level alignment strategies. Through comprehensive experiments across multiple benchmarks, we demonstrate average performance improvements of up to **3.0%** (diff. detector), **1.3%** mAP (diff. guided detector) for DG tasks and **5.4%** mAP for DA. When scaled to larger datasets, our approach maintains substantial advantages, particularly in scenarios with large domain shifts and limited training data. This work offers practical solutions for generalized and adaptive detection and provides valuable insights for visual perception tasks requiring strong generalization and adaptation capabilities.

References

- [1] Yogesh Balaji, Swami Sankaranarayanan, and Rama Chellappa. Metareg: Towards domain generalization using meta-regularization. *Advances in neural information processing systems*, 31, 2018. 2
- [2] Dmitry Baranchuk, Andrey Voynov, Ivan Rubachev, Valentin Khruikov, and Artem Babenko. Label-efficient semantic segmentation with diffusion models. In *International Conference on Learning Representations*, 2022. 2
- [3] 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. 1, 2, 6
- [4] Chaoqi Chen, Zebiao Zheng, Xinghao Ding, Yue Huang, and Qi Dou. Harmonizing transferability and discriminability for adapting object detectors. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8869–8878, 2020. 1, 2
- [5] Chaoqi Chen, Jiongheng 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. 6
- [6] Kai Chen, Jiaqi Wang, Jiangmiao Pang, Yuhang Cao, Yu Xiong, Xiaoxiao Li, Shuyang Sun, Wansen Feng, Ziwei Liu, Jiarui Xu, et al. Mmdetection: Open mmlab detection toolbox and benchmark. *arXiv preprint arXiv:1906.07155*, 2019. 5
- [7] Kai Chen, Enze Xie, Zhe Chen, Yibo Wang, Lanqing Hong, Zhenguo Li, and Dit-Yan Yeung. Geodiffusion: Text-prompted geometric control for object detection data generation. In *ICLR*, 2024. 8
- [8] Meilin Chen, Weijie Chen, Shicai Yang, Jie Song, Xinchao Wang, Lei Zhang, Yunfeng Yan, Donglian Qi, Yueting Zhuang, Di Xie, et al. Learning domain adaptive object detection with probabilistic teacher. In *International Conference on Machine Learning*, pages 3040–3055. PMLR, 2022. 6
- [9] 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. 1, 2
- [10] 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. 6
- [11] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3213–3223, 2016. 5
- [12] Ekin D Cubuk, Barret Zoph, Jonathon Shlens, and Quoc V Le. Randaugment: Practical automated data augmentation with a reduced search space. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 702–703, 2020. 2
- [13] 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. 1, 2, 6, 7
- [14] 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. 6
- [15] Jinhong Deng, Dongli Xu, Wen Li, and Lixin Duan. Harmonious teacher for cross-domain object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23829–23838, 2023. 2, 6
- [16] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 2, 7
- [17] Yingjun Du, Jun Xu, Huan Xiong, Qiang Qiu, Xiantong Zhen, Cees GM Snoek, and Ling Shao. Learning to learn with variational information bottleneck for domain generalization. In *Computer Vision—ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part X 16*, pages 200–216. Springer, 2020. 2
- [18] A Enes Doruk and Hasan F Ates. Davimnet: Ssms-based domain adaptive object detection. *arXiv e-prints*, pages arXiv–2502, 2025. 6
- [19] 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. 1, 2
- [20] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88:303–338, 2010. 5
- [21] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020. 1, 2
- [22] Yinsai Guo, Hang Yu, Shaorong Xie, Liyan Ma, Xinzhi Cao, and Xiangfeng Luo. Dsca: A dual semantic correlation alignment method for domain adaptation object detection. *Pattern Recognition*, 150:110329, 2024. 6
- [23] Boyong He, Yuxiang Ji, Zhuoyue Tan, and Liaoni Wu. Diffusion domain teacher: Diffusion guided domain adaptive object detector. In *ACM Multimedia 2024*, 2024. 1, 2, 3, 5, 6, 7, 8
- [24] Boyong He, Yuxiang Ji, Qianwen Ye, Zhuoyue Tan, and Liaoni Wu. Generalized diffusion detector: Mining robust features from diffusion models for domain-generalized detection. In *Proceedings of the Computer Vision and Pattern*

- Recognition Conference (CVPR)*, pages 9921–9932, 2025. 1, 2, 3, 4, 5, 6, 7, 8
- [25] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 3, 5, 7
- [26] Mengzhe He, Yali Wang, Jiayi Wu, Yiru Wang, Hanqing Li, Bo Li, Weihao Gan, Wei Wu, and Yu Qiao. Cross domain object detection by target-perceived dual branch distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9570–9580, 2022. 6
- [27] Zhenwei He and Hongsu Ni. Single-domain generalized object detection by balancing domain diversity and invariance. *arXiv preprint arXiv:2502.03835*, 2025. 6
- [28] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 1
- [29] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 2
- [30] Minui Hong, Jinwoo Choi, and Gunhee Kim. Stylemix: Separating content and style for enhanced data augmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14862–14870, 2021. 1, 2
- [31] Lukas Hoyer, Dengxin Dai, Haoran Wang, and Luc Van Gool. Mic: Masked image consistency for context-enhanced domain adaptation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11721–11732, 2023. 6
- [32] Cheng-Chun Hsu, Yi-Hsuan Tsai, Yen-Yu Lin, and Ming-Hsuan Yang. Every pixel matters: Center-aware feature alignment for domain adaptive object detector. In *Computer Vision—ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IX 16*, pages 733–748. Springer, 2020. 6
- [33] Jiaying Huang, Dayan Guan, Aoran Xiao, and Shijian Lu. Fsd: Frequency space domain randomization for domain generalization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6891–6902, 2021. 6
- [34] Naoto Inoue, Ryosuke Furuta, Toshihiko Yamasaki, and Kiyoharu Aizawa. Cross-domain weakly-supervised object detection through progressive domain adaptation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5001–5009, 2018. 5, 7
- [35] Junguang Jiang, Baixu Chen, Jianmin Wang, and Mingsheng Long. Decoupled adaptation for cross-domain object detection. In *International Conference on Learning Representations*, 2021. 6
- [36] Mikhail Kennerley, Jian-Gang Wang, Bharadwaj Veeravalli, and Robby T Tan. Cat: Exploiting inter-class dynamics for domain adaptive object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16541–16550, 2024. 6
- [37] 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. 1, 2, 5, 6, 7
- [38] Deng Li, Aming Wu, Yaowei Wang, and Yahong Han. Prompt-driven dynamic object-centric learning for single domain generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17606–17615, 2024. 6
- [39] Haoliang Li, Sinno Jialin Pan, Shiqi Wang, and Alex C Kot. Domain generalization with adversarial feature learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5400–5409, 2018. 1, 2
- [40] Liunian Harold Li, Pengchuan Zhang, Haotian Zhang, Jianwei Yang, Chunyuan Li, Yiwu Zhong, Lijuan Wang, Lu Yuan, Lei Zhang, Jenq-Neng Hwang, et al. Grounded language-image pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10965–10975, 2022. 2, 7
- [41] 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. 1, 6
- [42] Wuyang Li, Xinyu Liu, and Yixuan Yuan. Sigma: Semantic-complete graph matching for domain adaptive object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5291–5300, 2022. 6
- [43] Wuyang Li, Xinyu Liu, and Yixuan Yuan. Sigma++: Improved semantic-complete graph matching for domain adaptive object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023. 6
- [44] 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. 1, 2, 5, 6
- [45] Chuang Lin, Zehuan Yuan, Sicheng Zhao, Peize Sun, Changhu Wang, and Jianfei Cai. Domain-invariant disentangled network for generalizable object detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 8771–8780, 2021. 6
- [46] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision—ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014. 2, 5
- [47] 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. 6
- [48] 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. 2, 6

- [49] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021. 2, 7
- [50] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11976–11986, 2022. 2, 7
- [51] 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. 2
- [52] Claudio Michaelis, Benjamin Mitzkus, Robert Geirhos, Evgenia Rusak, Oliver Bringmann, Alexander S Ecker, Matthias Bethge, and Wieland Brendel. Benchmarking robustness in object detection: Autonomous driving when winter is coming. *arXiv preprint arXiv:1907.07484*, 2019. 5
- [53] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *International conference on machine learning*, pages 8821–8831. Pmlr, 2021. 2
- [54] 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. 6
- [55] Joseph Redmon and Ali Farhadi. Yolov3: An incremental improvement. *arXiv preprint arXiv:1804.02767*, 2018. 1
- [56] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28, 2015. 1, 2, 5
- [57] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 2
- [58] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in neural information processing systems*, 35:36479–36494, 2022. 2
- [59] 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. 1, 2, 5, 6
- [60] Christos Sakaridis, Dengxin Dai, and Luc Van Gool. Semantic foggy scene understanding with synthetic data. *International Journal of Computer Vision*, 126:973–992, 2018. 5
- [61] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020. 1, 2
- [62] Bo Sun, Banghuai Li, Shengcai Cai, Ye Yuan, and Chi Zhang. Fsce: Few-shot object detection via contrastive proposal encoding. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7352–7362, 2021. 7
- [63] 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. 2
- [64] Zhi Tian, Chunhua Shen, Hao Chen, and Tong He. Fcos: A simple and strong anchor-free object detector. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(4): 1922–1933, 2020. 1, 7
- [65] Vidit Vedit, Martin Engilberge, and Mathieu Salzmann. Clip the gap: A single domain generalization approach for object detection. In *CVPR*, pages 3219–3229, 2023. 6
- [66] Jindong Wang, Cuiling Lan, Chang Liu, Yidong Ouyang, Tao Qin, Wang Lu, Yiqiang Chen, Wenjun Zeng, and S Yu Philip. Generalizing to unseen domains: A survey on domain generalization. *IEEE transactions on knowledge and data engineering*, 35(8):8052–8072, 2022. 1
- [67] 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. 4
- [68] Xudong Wang, Trevor Darrell, Sai Saketh Rambhatla, Rohit Girdhar, and Ishan Misra. Instancediffusion: Instance-level control for image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6232–6242, 2024. 8
- [69] Yibo Wang, Ruiyuan Gao, Kai Chen, Kaiqiang Zhou, Yingjie Cai, Lanqing Hong, Zhenguo Li, Lihui Jiang, Dit-Yan Yeung, Qiang Xu, et al. Detdiffusion: Synergizing generative and perceptive models for enhanced data generation and perception. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7246–7255, 2024. 8
- [70] Weixi Weng and Chun Yuan. Mean teacher detr with masked feature alignment: a robust domain adaptive detection transformer framework. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 5912–5920, 2024. 6
- [71] 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. 5, 6
- [72] 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. 2, 6
- [73] Chang-Dong Xu, Xing-Ran Zhao, Xin Jin, and Xiu-Shen Wei. Exploring categorical regularization for domain adaptive object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11724–11733, 2020. 2

- [74] Jiarui Xu, Sifei Liu, Arash Vahdat, Wonmin Byeon, Xiaolong Wang, and Shalini De Mello. Open-vocabulary panoptic segmentation with text-to-image diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2955–2966, 2023. 2
- [75] Mingjun Xu, Lingyun Qin, Weijie Chen, Shiliang Pu, and Lei Zhang. Multi-view adversarial discriminator: Mine the non-causal factors for object detection in unseen domains. In *Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition*, pages 8103–8112, 2023. 6
- [76] Qinwei Xu, Ruipeng Zhang, Ya Zhang, Yanfeng Wang, and Qi Tian. A fourier-based framework for domain generalization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14383–14392, 2021. 6
- [77] Xiaoran Xu, Jiangan Yang, Wenhui Shi, Siyuan Ding, Luqing Luo, and Jian Liu. Physaug: A physical-guided and frequency-based data augmentation for single-domain generalized object detection. *arXiv preprint arXiv:2412.11807*, 2024. 6
- [78] Yanchao Yang and Stefano Soatto. Fda: Fourier domain adaptation for semantic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4085–4095, 2020. 5
- [79] Fisher Yu, Haofeng Chen, Xin Wang, Wenqi Xian, Yingying Chen, Fangchen Liu, Vashisht Madhavan, and Trevor Darrell. Bdd100k: A diverse driving dataset for heterogeneous multitask learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2636–2645, 2020. 5
- [80] Jinze Yu, Jiaming Liu, Xiaobao Wei, Haoyi Zhou, Yohei Nakata, Denis Gudovskiy, Tomoyuki Okuno, Jianxin Li, Kurt Keutzer, and Shanghang Zhang. Mtrtrans: Cross-domain object detection with mean teacher transformer. In *European Conference on Computer Vision*, pages 629–645. Springer, 2022. 1
- [81] Hao Zhang, Feng Li, Shilong Liu, Lei Zhang, Hang Su, Jun Zhu, Lionel Ni, and Heung-Yeung Shum. Dino: Detr with improved denoising anchor boxes for end-to-end object detection. In *The Eleventh International Conference on Learning Representations*, 2023. 1, 7
- [82] Libo Zhang, Wenzhang Zhou, Heng Fan, Tiejian Luo, and Haibin Ling. Robust domain adaptive object detection with unified multi-granularity alignment. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024. 6
- [83] 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. 6
- [84] Yuyang Zhao, Zhun Zhong, Na Zhao, Nicu Sebe, and Gim Hee Lee. Style-hallucinated dual consistency learning for domain generalized semantic segmentation. In *European conference on computer vision*, pages 535–552. Springer, 2022. 6
- [85] Kaiyang Zhou, Yongxin Yang, Yu Qiao, and Tao Xiang. Domain generalization with mixstyle. *arXiv preprint arXiv:2104.02008*, 2021. 1, 2
- [86] 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
- [87] Wenzhang Zhou, Dawei Du, Libo Zhang, Tiejian Luo, and Yanjun Wu. Multi-granularity alignment domain adaptation for object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9581–9590, 2022. 2
- [88] Wenzhang Zhou, Heng Fan, Tiejian Luo, and Libo Zhang. Unsupervised domain adaptive detection with network stability analysis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6986–6995, 2023. 6
- [89] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2223–2232, 2017. 1, 2
- [90] Wei Zhu, Le Lu, Jing Xiao, Mei Han, Jiebo Luo, and Adam P Harrison. Localized adversarial domain generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7108–7118, 2022. 1, 2