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

Han Jiang[1], Wenfei Yang[1,2,*], Tianzhu Zhang[1,2], Yongdong Zhang[1]
[1]University of Science and Technology of China
[2]National Key Laboratory of Deep Space Exploration, Deep Space Exploration Laboratory
hjiang12@mail.ustc.edu.cn, {yangwf, tzzhang, zhyd73}@ustc.edu.cn

Abstract

Single domain generalized object detection aims to train an object detector on a single source domain and generalize it to any unseen domain. Although existing approaches based on data augmentation exhibit promising results, they overlook domain discrepancies across multiple augmented domains, which limits the performance of object detectors. To tackle these problems, we propose a novel diffusion-based framework, termed SDG-DiffDet, to mitigate the impact of domain gaps on object detectors. The proposed SDG-DiffDet consists of a memory-guided diffusion module and a source-guided denoising module. Specifically, in the memory-guided diffusion module, we design feature statistics memories that mine diverse style information from local parts to augment source features. The augmented features further serve as noise in the diffusion process, enabling the model to capture distribution differences between practical domain distributions. In the source-guided denoising module, we design a text-guided condition to facilitate distribution transfer from any unseen distribution to source distribution in the denoising process. By combining these two designs, our proposed SDG-DiffDet effectively models feature augmentation and target-to-source distribution transfer within a unified diffusion framework, thereby enhancing the detection performance on unseen domains. Extensive experiments demonstrate that the proposed SDG-DiffDet achieves state-of-the-art performance across two challenging scenarios.

1. Introduction

In recent years, the rapid advancement of deep learning [17, 29] has significantly prompted the development of object detection area [7, 22, 32, 47], achieving remarkable performance with large-scale labeled datasets. However, in many real-world scenarios, directly applying trained object detectors to unseen datasets leads to a notable decline in performance due to domain shifts [37], including changes in lighting, weather conditions, and so on. Domain adaptation for object detection (DAOD) is a common approach to mitigate this issue by learning domain-invariant features between labeled source and unlabeled target domains. However, DAOD methods [15, 28] require target data during the training stage, significantly limiting their practical application.

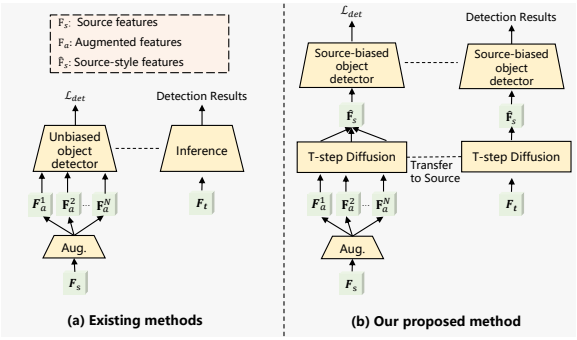


Figure 1. Comparison with existing methods. (a) Existing methods aim to learn an unbiased detector by using data augmentation techniques to improve the data diversity. (b) Our method aims to learn a source-biased detector by transferring distribution from unseen domains to the source domain based on a diffusion model.

To deal with this problem, domain generalization for object detection [21, 24, 31, 39] has been proposed to generalize knowledge from single or multiple source domains to unseen target domains. Early approaches [21, 41, 41] primarily focus on the scenario where multiple source domains are available during training, aiming to learn domain-invariant features to improve the generalization ability of object detectors. Nevertheless, obtaining labeled data from multiple source domains incurs significant annotation costs, thereby hindering the practicability of such techniques. Consequently, single domain generalized object detection has been proposed [3, 23, 42], which aims to generalize a

*Corresponding author

model trained on a single source domain to multiple unseen target domains, providing a more feasible and challenging approach. Existing single domain generalized object detection can be generally divided into two categories, involving feature disentanglement based methods [42] and data augmentation based methods [20, 23, 39]. Feature disentanglement based methods aim to separate domain-invariant representations from domain-specific ones. However, some studies [45] have indicated that completely removing domain-specific features with a single source domain is challenging. On the other hand, data augmentation based methods, including input-level augmentation [23] and feature-level augmentation [39], aim to learn an unbiased object detector by enhancing data diversity as shown in Figure 1 (a). However, these methods still face the following two challenges: (1) Previous methods adopt predefined augmentation techniques to diversify source data, which may fail to cover the distribution of the target domain, especially when a significant distribution shift exists between source and target domains. (2) Training an object detector on multiple augmented domains neglects domain discrepancies across these domains, thereby limiting the object detection performance.

To solve these two problems, we aim to learn a source-biased object detector by modeling feature augmentation and target-to-source distribution transfer in a unified diffusion model as shown in Figure 1 (b). Specifically, the forward process gradually adds noise to the input data over different time steps T during the training stage, thereby simulating various data distributions. The reverse process can facilitate distribution transfer from augmented domains to source domain. In this way, the object detector is trained solely on source-style samples, thereby preventing it from being influenced by domain gaps across multiple augmented domains and improving the performance on both source and target domains. However, directly applying diffusion model to single domain generalized object detection poses the following two challenges: (1) *How to maintain the structure of feature distribution in the forward process?* Traditional diffusion models [12, 16, 48] define a Markovian chain in the forward process by gradually adding Gaussian noise. However, the augmented features generated in this way tend to be a gaussian noise, lacking the structural characteristics of real-world features. As a result, the diffusion denoising process fails to perceive the true distribution differences between unseen target domains and the source domain, which impedes the performance of object detectors. (2) *How to guide the reverse process to source distribution during the inference stage?* In the reverse process, it is essential to guide the diffusion model to generate source-style features that preserve semantic consistency with input features. A straightforward approach is to utilize corresponding source features as conditions to control the reverse process. However, this approach becomes unfeasible since source information is unavailable during inference.

To overcome these two challenges, we propose a novel diffusion-based framework for single domain generalized object detection, namely **SDG-DiffDet**, which consists of an **Memory-guided Diffusion Module** and a **Source-guided Denoising Module**. In the memory-guided diffusion module, we introduce memory modules to store the channel-wise mean and standard deviation of local parts across the entire dataset. Subsequently, we randomly sample feature statistics from these memories to generate augmented features with Adaptive Instance Normalization (AdaIN) [14], which transfers diverse styles to the source domain, while preserving the content structure. Therefore, by using augmented features as noise, we effectively shift the source distribution closer to the augmented distributions, enabling the diffusion model to better capture the distributional differences between practical domains. The source-guided denoising module aims to transfer augmented distributions to the source distribution. Considering that source information is unavailable during the inference stage, we incorporate text-guided conditions as a bridge, generating conditions that maintain the semantic content of input features while aligning them with the style of source domain. Specifically, we create a set of text embeddings that involve source-domain style and class information. These embeddings further interact with input features through a cross attention mechanism to generate conditions with source style, which guide the diffusion model to transfer any unseen domain distributions to source distribution through multiple steps of reverse process. By combining these two designs, our SDG-DiffDet effectively models feature augmentation and target-to-source distribution transfer within a unified diffusion framework, thereby enhancing the detection peerformance on unseen domains.

The major contributions of our work can be summarized as follows: (1) We propose SDG-DiffDet, a novel diffusion-based framework for single domain generalized object detection, which explicitly models the distribution shift between source and augmented domains. To the best of our knowledge, this is the first work to apply diffusion model to single domain generalized object detection. (2) We introduce a memory-guided diffusion module to model feature augmentation in the diffusion process and a source-guided denoising module to perform target-to-source transfer in the denoising process. (3) Experimental results on two domain generalization datasets show the effectiveness and superiority of our method.

2. Related Work

In this section, we provide a brief overview of methods related to single domain generalized object detection and diffusion model for perception task.

2.1. Single Domain Generalized Object Detection

Existing single domain generalized object detection can be broadly grouped into two categories: feature disentanglement based methods [42] and data augmentation based methods [3, 23, 45]. Feature disentanglement based methods focus on disentangling domain-invariant representations from domain-specific representations. CDSD [42] employs a cyclic-disentangled module to extract domain-invariant feature representations within a single domain and design a self-distillation module to further enhance the detection performance on unseen target domains. Nowadays, data augmentation-based methods have achieved superior performance by enhancing the diversity of source domain images. AFDA [3] uses the common off-the-shelf image corruptions to disturb input-level distribution and align predictions across different augmentations of an image. UFR [23] first considers single domain generalized object detection from a casual view and proposes an unbiased Faster-RCNN to reduce the data bias, attention bias and prototype bias. OA-DG [18] introduces object-aware mixing to prevent global data augmentation from damaging object annotation. Besides, some methods [19, 39] utilize a pretrained CLIP [30] to augment source data in the feature space. CLIPGap [39] introduces a semantic augmentation method to augment the source domain to specific target domains with the corresponding textual prompts. PGST [20] further leverages grounded language-image pre-training model (GLIP) to achieve style transfer from source domain to target domain. Different from these methods that only focus on generating diversity augmented domains to cover unseen target domains, our method learns to model distribution transfer between augmented domains and source domain by diffusion model.

2.2. Diffusion Model for Perception Task

Diffusion models [12] have attracted significant attention due to their impressive progress in image generation [25, 25, 33, 48]. Recently, some works have explored its potential application in discriminative tasks. ProtoDiff [4] incorporates probabilistic modeling and task-guided prototypes to enhance the performance of few-shot learning. ODISE [44] employs the diffusion model as a feature extractor and demonstrates the great potential of text-to-image generation models in open vocabulary segmentation tasks In the context of object detection, DiffusionDet [1] formulates object detection as a denoising diffusion process, which generates bounding boxes from random boxes by reversing the diffusion process. Diffusion-SS3D [11] introduces a diffusion process for semi-supervised 3D object detection, aiming to produce high-quality pseudo-labels by denoising random noise to object sizes and label distributions. In this paper, we utilize the diffusion model for single domain generalized object detection, aiming to provide a new perspective of distribution transfer in feature space.

3. Method

3.1. Preliminaries

Problem Formulation. In the context of single domain generalized object detection, we have a labeled source domain and T unseen target domains. For simplicity, we denote the source domain as $\mathcal{D}_s = \{x_i^s, y_i^s\}_{i=1}^{N_s}$, where x_i^s is an image, $y_i^s = \{c_i, b_i\}$ denotes the corresponding labels, including class labels c_i and bounding box coordinates b_i. The T unseen target domains are represented as $\{\mathcal{D}_t\}_{t=1}^{T}$. The source domain and T target domains share the same category label space. Our goal is to train a detector on the source domain and generalize it to unseen target domains.

Adaptive Instance Normalization (AdaIN) [14]. Given a feature map $\mathbf{F}$, AdaIN shows that the channel-wise feature statistics of $\mathbf{F}$ capture style information of the corresponding image, allowing style transfer between different images. Therefore, transferring the style from a source feature $\mathbf{F}_s$ to a target feature $\mathbf{F}_t$ can be expressed as:

$$\textbf{AdaIN}(\mathbf{F}_s, \mathbf{F}_t) = \sigma(\mathbf{F}_t)\left(\frac{\mathbf{F}_s - \mu(\mathbf{F}_s)}{\sigma(\mathbf{F}_s)}\right) + \mu(\mathbf{F}_t), \quad (1)$$

where $\mu \in \mathbb{R}^C$ and $\sigma \in \mathbb{R}^C$ denote channel-wise mean and standard deviation, respectively.

3.2. Overview

The overall architecture of our proposed SDG-DiffDet is shown in Figure 2. The primary objective of our method is to transfer unseen features to the source style, thereby training a source-biased object detector. Our approach is composed of two main modules: an memory-guided diffusion module for feature augmentation (Sec. 3.3) and a source-guided denoising module for target-to-source distribution transfer (Sec. 3.4).

3.3. Memory-guided Diffusion Module

In this section, we aim to generate diverse features that can be used as noise in the diffusion process. Previous works [31] points out that the image itself is a style library for feature augmentation due to the style discrepancy between local parts, such as textures, colors, etc. Therefore, we design feature statistics memories to transfer the style of local parts to the whole image with AdaIN. Specifically, given an image x_s from the source domain, we first employ E_{low} to extract its low-level feature map $\mathbf{F}_s^l \in \mathbb{R}^{H^l \times W^l \times C^l}$, where E_{low} denotes the first three layers of the feature encoder, H^l,W^l and C^l denote the height, weight and channels of $\mathbf{F}_s^l$, respectively. Then, to incorporate diverse style information from the entire dataset, we introduce two memories, including a mean memory $M_\mu = \{\mu_m\}_{m=1}^{M}$ and a standard deviation memory $M_\sigma = \{\sigma_m\}_{m=1}^{M}$. All elements

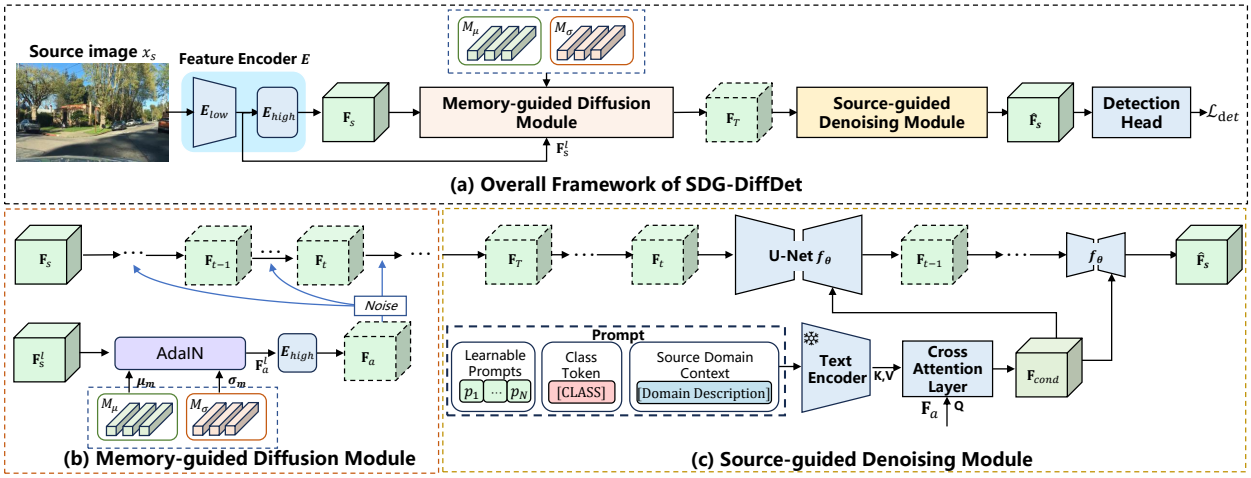


Figure 2. (a) The overall architecture of our proposed SDG-DiffDet. Source images are initially fed into a feature encoder to extract their 2D feature representations. The image features are then processed by the memory-guided diffusion module to generate noised source features, which are subsequently denoised with the source stylization denoising module. The denoised augmented features are input into the detection head for final prediction. (b) Illustration of the memory-guided diffusion module. We utilize feature statistics sampled from memories to augment source features, which act as noise in the diffusion process. (c) Illustration of the source-guided denoising module. The frozen text encoder takes source-style text prompts to generate text embeddings, which interact with augmented features via a cross-attention layer, producing conditions that guide the denoising process.

in these memories are learnable parameters and initialized with Kaiming initialization [9]. During the training stage, we randomly sample a mean value and a standard deviation from the corresponding memory to generate a augmented source feature $\mathbf{F}_a^l$ with Eq (1). Subsequently, we pass $\mathbf{F}_a^l$ and $\mathbf{F}_s^l$ through the remaining layers E_{high} to obtain $\mathbf{F}_a$ and $\mathbf{F}_s$, and $\mathbf{F}_a$ serves as noise to guide the diffusion process in capturing the distribution of the augmented domain. Finally, the generated noise $\mathbf{F}_a$ is gradually added to F_s using a fixed forward process:

$$\begin{aligned} \mathbf{F}_t &= \sqrt{1-\beta_t}\mathbf{F}_{t-1} + \beta_t \mathbf{F}_a, \\ \mathbf{F}_0 &= \mathbf{F}_s, \end{aligned} \tag{2}$$

where T is the overall timesteps and $\{\beta_t\}_{t=1}^T$ [36] is a set of predefined parameters that controls step sizes.

The major challenge is how to ensure these memories involve diverse styles of feature statistics across the entire dataset. In the following, we take mean memory M_μ as an example and describe how to update the memory in detail.

Memory update. We first split $\mathbf{F}_s$ into different parts with the ground truth to achieve $\mathbf{F}_s^{split} = \{f_1^o, \ldots, f_N^o, f^b\}$:

$$\begin{aligned} f_i^o &= \mathbf{F}_s^l \cdot m_i^o, \\ f^b &= \mathbf{F}_s^l \cdot m^b, \end{aligned} \tag{3}$$

where $i \in \{1, 2, \ldots N\}$ and N is the number of objects. m_i^o and m^b represent the mask of the i-th objects and the rest background regions, respectively. For simplicity, we omit the superscript o and denote f^b as f_{N+1}. The channel-wise mean of each part can be calculated as:

$$\mu(f_i) = \frac{1}{H^l W^l} \sum_{w=1}^{W^l} \sum_{h=1}^{H^l} f_i, \tag{4}$$

where $i \in \{1, 2, \ldots N+1\}$. We calculate the normalized similarity metric between each element in M_μ and $\{\mu(f_i)\}_{i=1}^{N+1}$:

$$s^{m,i} = \frac{exp(\mu_m \cdot \mu(f_i))}{\sum_{n=1}^{N+1} exp(\mu_m \cdot \mu(f_n))}, \tag{5}$$

where s is an $M \times (N+1)$ similarity matrix. With the similarity matrix, we update μ_m as follows:

$$\mu_m \leftarrow \mu_m + \sum_{i=1}^{N+1} s^{m,i} \mu(f_i). \tag{6}$$

To prevent each memory element from containing identical style information, we further adopt a diversity loss motivated by [24], which is formulated as:

$$\mathcal{L}_{div}^{\mu} = \frac{1}{M(M-1)} \sum_{i=1}^{M} \sum_{j=1, j\neq i}^{M} \frac{\langle \mu_i, \mu_j \rangle}{\|\mu_i\|_2 \|\mu_j\|_2}. \tag{7}$$

We adopt the same way to update M_σ, where only Eq (4) is replaced by calculated channel-wise standard deviation:

$$\sigma(f_i) = \sqrt{\frac{1}{H^l W^l} \sum_{w=1}^{W^l} \sum_{h=1}^{H^l} (f_i - \mu(f_i))^2}. \tag{8}$$

The total diversity loss L_{div} is formulated as:

$$\mathcal{L}_{div} = \mathcal{L}^{\mu}_{div} + \mathcal{L}^{\sigma}_{div} \tag{9}$$

3.4. Source-guided Denoising Module

After the forward diffusion process, we further perform a denoising process to model distribution transfer between unseen domains and the source domain. To guide the denoising process, we first generate condition $\mathbf{F}_{cond}$ that can be utilized in both training and testing stages while preserving the source-style semantic information. To achieve this, we propose to utilize a pretrained CLIP text encoder to obtain text embeddings $\tilde{\mathbf{t}} = [\hat{t}_1, \ldots, \hat{t}_K, \hat{t}_{K+1}]^T$ with learnable textual contexts $\mathbf{p}$:

$$\hat{t}_k = E_T([\mathbf{p}, \mathbf{e}_k, \mathbf{d}]), \tag{10}$$

where K denotes the total category number, $\mathbf{e}_k$ represents the embedding for the name of the k-th class, $\mathbf{e}_{K+1}$ and $\mathbf{d}$ denote the text embedding for "`background`" and source domain prompts "`in a daytime clear scene`", respectively. The embeddings $\hat{\mathbf{t}}$ are further projected through an MLP layer to produce the final output $\mathbf{t} \in \mathbb{R}^{(K+1)\times C}$. Then, we incorporate a cross-attention layer to generate the conditions for $\mathbf{F}_a$. Specifically, keys and values are derived from $\mathbf{t}$, and queries are derived from the flattened $\mathbf{F}_a$:

$$\mathbf{Q} = \mathbf{F}_a\mathbf{W}^{\mathcal{Q}}, \mathbf{K} = \mathbf{t}\mathbf{W}^{\mathcal{K}}, \ \mathbf{V} = \mathbf{t}\mathbf{W}^{\mathcal{V}}, \tag{11}$$

where $\mathbf{W}^{\mathcal{Q}}, \mathbf{W}^{\mathcal{K}}, \mathbf{W}^{\mathcal{V}} \in \mathbb{R}^{C\times d_k}$. The conditions $\mathbf{F}^a_{cond}$ are obtained with the multi-head attention mechanism:

$$\mathbf{F}^a_{cond} = \mathrm{softmax}(\frac{\mathbf{Q}\mathbf{K}^{\mathrm{T}}}{\sqrt{\mathrm{d_k}}})\mathbf{V}. \tag{12}$$

Additionally, we adopt a pixel-text matching loss [24] to fine-tune the prompts to align source visual features with text embeddings at pixel level. In the context of object detection, ground truth segmentation labels are not available. Therefore, we use ground truth bounding boxes and labels to build binary supervision $y \in \{0,1\}^{HW\times(K+1)}$. The pixel-text matching loss is computed with a binary cross-entropy (BCE) loss:

$$\mathcal{L}_{mat} = \mathrm{BCE}(\mathrm{sigmoid}(\tilde{\mathbf{F}}_{\mathrm{s}}\tilde{\mathbf{t}}^{\mathrm{T}}/\tau), \mathrm{y}), \tag{13}$$

where $\tilde{\mathbf{F}}_s$ and $\tilde{\mathbf{t}}$ represent the l_2 normalized version of $\mathbf{F}_s$ and $\mathbf{t}$ along the channel dimension, $\tau = 0.07$ is a temperature coefficient.

With the condition $\mathbf{F}_{cond}$, we design a diffusion model f_θ to reconstruct source features $\mathbf{F}_s$ in a generative paradigm. The detailed architecture of f_θ with a U-Net is presented in the **Supplementary Materials**. Specifically, to recover $\mathbf{F}_0$ from noised features $\mathbf{F}_t$, the diffusion model f_θ is trained to perform the reverse diffusion process with $\mathbf{F}_{cond}$. The training objective is formulated with the MSE loss:

$$\mathcal{L}_{rec} = \|f_\theta(\mathbf{F}_t, \mathbf{F}^a_{cond}, t) - \mathbf{F}_0\|^2. \tag{14}$$

In addition, we conduct a T-step reverse diffusion to generate the corresponding source-style features of $\hat{\mathbf{F}}^a_s$, which are then input into detector head for prediction:

$$\hat{\mathbf{F}}^a_s = Reverse(\mathbf{F}_a, \mathbf{F}^a_{cond}, T). \tag{15}$$

3.5. Training and Inference

Training Stage. We employ Faster-RCNN as the basic object detector and the total loss of our SDG-DiffTection is represented as follows:

$$\mathcal{L} = \mathcal{L}_{det} + \alpha\mathcal{L}_{div} + \beta\mathcal{L}_{rec} + \gamma\mathcal{L}_{mat}, \tag{16}$$

where $\mathcal{L}_{det}$ is the supervised detection loss. α, β and γ are the hyperparameters used to balance the contribution of different losses.

Inference Stage. Given the unseen target domain feature $\mathbf{F}_u$, we perform a T-step reverse diffusion to generate the associated source-style feature:

$$\hat{\mathbf{F}}^u_s = Reverse(\mathbf{F}_u, \mathbf{F}^u_{cond}, T), \tag{17}$$

where $\mathbf{F}^u_{cond}$ can be obtained by Eq (11) and Eq (12), only replacing $\mathbf{F}_a$ with $\mathbf{F}_u$. We employ $\hat{\mathbf{F}}^u_s$ to predict object labels and corresponding bounding boxes.

4. Experiment

4.1. Experimental Setup

Datasets. **Urban Scene Dataset** [42] provides images captured under five various weather conditions, including Daytime Clear (DC), Night Clear (NC), Night Rainy (NR), Dusk Rainy (DR), and Daytime Foggy (DF). The images are selected from three datasets: Barkeley Deep Drive 100k (BDD-100k) [46], FoggyCityscapes [34] and AdverseWeather [8]. Additionally, rainy images are rendered from BDD-100k dataset [43]. For our experiments, we use Daytime Clear dataset as the source domain, which consists of 19,395 training images and 8,313 test images. The remaining four datasets are employed for testing, consisting of 26,158 images in Night Clear scene, 3,775 images in Daytime Foggy scene, 3,501 images in Dusk Rainy scene and 2,494 images in Night Rainy scene. All these datasets contain bounding box annotations for seven categories, including *person*, *car*, *bike*, *rider*, *motor*, *bus* and *truck*.

Real to Artistic consists of four datasets, including Pascal VOC, Clipart1k, Watercolor2k and Comic2k. Pascal VOC, Clipart1k, Watercolor2k, and Comic2k. Pascal VOC is composed of real images covering 20 object classes,

Table 1. Single domain generalization results on the Urban Scene dataset. C denotes that the image encoder is CLIP-initialized. Avg denotes the average mAP across all out-of-domain scenarios.

Method	C	DC	DF	DR	NC	NR	Avg
FR [32]	✗	51.8	38.9	30.0	15.7	33.1	29.4
IBN-Net[26]	✗	49.7	29.6	26.1	32.1	14.3	25.5
SW [27]	✗	50.6	30.8	26.3	33.4	13.7	26.1
IterNorm [13]	✗	43.9	28.4	22.8	29.6	12.6	23.4
ISW [2]	✗	51.3	31.8	25.9	33.2	14.1	26.3
CDSD [42]	✗	56.1	33.5	28.2	36.6	16.6	28.7
CLIPGap [39]	✓	51.3	38.5	32.3	36.9	18.7	31.6
SRCD [31]	✗	-	35.9	28.8	36.7	17.0	29.6
OA-Mix [18]	✗	55.8	38.3	33.9	38.0	16.8	31.8
PDOC [19]	✓	53.6	38.5	33.7	19.2	39.1	32.6
UFR [23]	✗	58.6	39.6	33.2	40.8	19.2	33.2
AFDA [3]	✗	52.8	37.2	38.1	42.5	24.1	35.5
Ours	✗	60.2	41.1	38.9	43.1	25.4	37.1
Ours	✓	**61.8**	**43.2**	**40.9**	**44.7**	**26.6**	**38.9**

while Clipart1k contains artistic images with the same 20 classes. Additionally, Watercolor2k and Comic2k each consist of 6 classes, which are subsets of the Pascal VOC classes. For our experiments, we follow prior methods by using Pascal VOC as the source domain, which consists of 16,551 training images and 5,000 test images. Clipart1k includes 1,000 images, while both Watercolor2k and Comic2k contain 2,000 images each, and all three datasets are treated as unseen domains.

Implementation Details. We employ Faster-RCNN [32] with ResNet101 [10] as our backbone. In all experiments, we train our model for 100k iterations with an initial learning rate of 0.01, which is reduced by a factor of 10 after 80k iterations. Our model is optimized by Stochastic Gradient Descent (SGD) with a momentum of 0.9, and set the batch size to 4. All experiments are conducted on 4 RTX 3090 GPUs and implemented based on Detectron2. For the hyperparameters, we set $\alpha = \beta = \gamma = 0.1$, $T = 10$, $M = 32$. All results are reported using mean average precision (mAP) metric with a 0.5 threshold for Intersection over Union (IoU).

4.2. Comparison with State-of-the-arts Methods

In this section, we present a comparative analysis of our results against other state-of-the-art methods. Following previous works [39, 42], we compare our method with several feature normalization approaches, including SW [27], IBN-Net [26], IterNorm [13], and ISW [2].

Results on Urban Scene Datasets. Table 1 shows the results on different weather conditions. Following most previous DG methods, we only use **ImageNet pre-trained feature encoder** for object detection. We observe that our method achieves the best results of 60.2 % mAP on the source domain. Compared to other augment-based methods [3, 23], our method demonstrates that training with source-style features alone allows the detector to capture domain-specific knowledge, thus providing supplementary information for supervised learning and enhancing the model's performance on the source domain. Furthermore, for unseen target domains, our method achieves the best 37.1 % mAP, demonstrating the effectiveness of our approach in handling challenging weather conditions. Additionally, using CLIP initialization further improves detection performance by 1.8% mAP. **Class-wise results are provided in the Supplementary Material.**

Table 2. Single domain generalization results on Real to Artist dataset. C denotes that the image encoder is CLIP-initialized. Avg represents the average mAP across all out-of-domain scenarios.

Method	C	Pascal VOC	Clipart	Watercolor	Comic	Avg
FR [32]	✗	82.4	25.7	44.5	18.9	29.7
NP [6]	✗	79.2	35.4	53.3	28.9	39.2
AFDA [3]	✗	80.1	38.9	57.4	33.2	43.2
Ours	✗	84.7	40.7	59.4	35.1	45.1
Ours	✓	**86.2**	**42.1**	**60.9**	**36.3**	**46.4**

Table 3. Extend to zero-shot domain adaptation. 'C' denotes that the image encoder is CLIP-initialized.

Method	C	DF	DR	NS	NR	Avg
PODA [5]	✓	**44.4**	40.2	43.4	20.5	37.1
Ours	✓	44.2	**41.8**	**45.9**	**27.3**	**39.8**

Results on Real to Artistic Datasets. Table 2 presents the results on the challenging real-to-artistic scenarios, where the domain shift is relatively large. Our method consistently achieves 45.1 % mAP on unseen target domains, outperforming AFDA by 1.9%. These results indicate that the proposed memory-guided diffusion module effectively simulates unseen domains, while the source-guided denoising module successfully reverses target domains to source distribution even under significant domain shifts.

4.3. Extend to Zero-shot Domain Adaptation

We extend our method to zero-shot domain adaptation [5], which provides target domain descriptions in natural language. The core idea behind these methods is to transfer the source domain to specific target domains using target prompts, whereas our method focuses on transferring any target domain to the source distribution to improve detection performance across all target domains. For a fair comparison, we follow [5] to use target domain descriptions to augment source features, which are subsequently transferred back to source distribution. Our method achieves 39.8 % mAP, outperforming PODA by 2.7 % mAP.

4.4. Quantitative Analysis

Ablation Studies. We conduct a series of ablation studies to verify the effectiveness of each individual module in our

Table 4. Ablation studies. MD, SD represent our proposed memory-guided diffusion module and source-guided denosing module, respectively. We employ vanilla Faster-RCNN as the baseline. w/ aug. denotes the feature augmentation method proposed in sec. 3.3.

Method	MD	SD	DF	DR	NS	NR	Avg
Baseline			33.1	28.4	35.2	15.4	28.0
Baseline w/ aug.			37.9	35.1	39.0	19.2	32.8
+ Diffusion			35.9	30.1	38.4	16.5	30.2
	✓		38.8	34.1	39.5	20.8	33.3
		✓	38.9	35.8	39.7	22.1	34.1
	✓	✓	**41.1**	**38.9**	**43.1**	**25.4**	**37.1**

Table 5. Effect of Memory-guided Diffusion Module.

Method	DF	DR	NS	NR	Avg
Fourier Aug. [35]	40.8	38.2	42.6	23.7	36.3
Text-guided Aug. [39]	40.6	37.8	41.7	23.9	36.0
NP [6]	40.5	38.3	41.9	24.2	36.2
Image Corruption [3]	40.7	38.2	42.2	24.6	36.4
Ours	**41.1**	**38.9**	**43.1**	**25.4**	**37.1**

proposed methods in Table 4. Specifically, the first and second rows demonstrate that by diversifying the source features, the baseline object detector generalizes effectively to unseen domains, achieving a notable improvement of 4.8% mAP. Besides, simply using a traditional diffusion model results in an average improvement of 2.2 % mAP across four target domains. The third row demonstrates that replacing Gaussian noise with the augmented features in the forward process yields a 3.1 % mAP improvement, indicating that using augmented features as noise allows the diffusion model to better capture practical distribution differences. Besides, the source-guided denoising module leads to a 3.9% performance gain, which proves that training a source-biased object detector significantly enhances the performance of the object detector. By further integrating two designed modules together, we achieve an overall 6.9% mAP improvement, highlighting the effectiveness of these modules in applying a diffusion process to single domain generalized object detection.

Effect of Memory-guided Diffusion Module. The proposed memory-guided diffusion module aims to generate perturbed features that may reflect unseen target domains. To prove the effectiveness of the proposed module, we explore different data augmentation methods in feature space [6, 35, 39] or input space [3] to augment features which subsequently act as input in the denoising diffusion step. Table 5 shows that our proposed method leads to at least 0.7 % performance gain than other augmentation methods, indicating that our proposed augmented method can effectively simulate unseen target distribution.

Effect of Source-guided Denoising Module. To demonstrate the unity of the proposed denoising process, we evaluate alternative baselines for transferring unseen target distributions to the source domain. First, we apply AdaIN using the stored mean and standard deviations of the source data, which results in a 3.0 % mAP performance drop. Additionally, inspired by denoising autoencoders [40], we adopt a similar U-Net framework with source conditions, replacing the iterative denoising process with direct source feature reconstruction, resulting in a 1.5% mAP performance drop. These results suggest that the denoising process more effectively maps diverse features back to the source domain.

Table 6. Effect of Source-guided Denoising Module. DAE denotes denoising autoencoder.

Method	DF	DR	NS	NR	Avg
AdaIN	39.1	35.4	40.8	21.2	34.1
DAE	40.8	36.2	42.1	23.2	35.6
Ours	**41.1**	**38.9**	**43.1**	**25.4**	**37.1**

Table 7. Effect of external CLIP supervison. All results are reproduced by official code.

Method	DC	DF	DR	NC	NR	Avg
Faster-RCNN	56.2	34.5	29.2	35.4	16.2	28.9
OA-Mix [18]	57.7	38.6	34.7	39.0	17.5	32.5
AFDA [3]	56.6	38.3	38.5	42.1	24.9	36.0
Ours	**60.2**	**43.2**	**40.9**	**44.7**	**25.4**	**37.1**

Table 8. Effect of different condition designs. St denotes a source style template "`in a daytime sunny scene`".

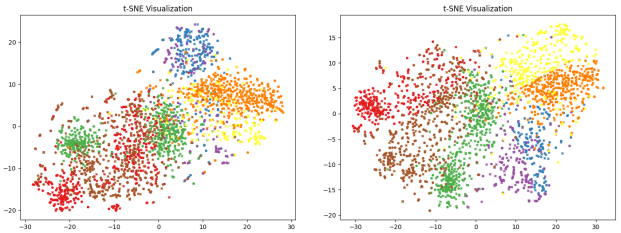
Text Conditions	St	DF	DR	NS	NR	Avg
w/o conditions		39.5	34.7	40.3	21.4	34.0
Hand-crafted prompts		39.9	36.7	41.4	23.1	35.3
	✓	40.5	37.9	42.2	23.5	36.0
Learnable prompts		40.8	38.7	42.4	25.1	36.8
Ours	✓	**41.1**	**38.9**	**43.1**	**25.4**	**37.1**

Effect of CLIP Assisting. Our method integrates an external CLIP model to assist the training process. For a fair comparison, we also utilize pixel-text matching to supervise previous works [3, 18], with the results presented in Table 7. While CLIP supervision with source-style prompts improves detection performance on the source domain for vanilla Faster-RCNN, it struggles on unseen target domains. In contrast, our method achieves an 8.2 % performance gain, highlighting that the performance improvement stems not only from CLIP supervision but also from the effectiveness of the proposed diffusion process.

Effect of Condition Design. We study three different condition designs as shown in Table 8. Specifically, we first employ a handcraft template "`a photo of [class]`" without incorporating source information. Our model achieves a 1.3 % mAP improvement compared to uncondi-

(a) Daytime Foggy (b) Night Clear

Figure 3. Visualization of detection results on (a) Daytime Foggy, (b) Night Clear. **Top**: The predictions of CLIPGap [39]. **Bottom**: The predictions of our SDG-DiffDet.



(a) Baseline w/aug. (b) Ours

Figure 4. t-SNE Visualization. We employ circles to represent Daytime Clear dataset and crosses to represent Daytime Foggy dataset. (a) and (b) represent baseline Faster-RCNN with feature augmentation method proposed in sec. 3.3 and our SDG-DiffDet, respectively.

tional denoising process, indicating that the semantic representations of input features can be preserved by CLIP text embedding during the reverse process. Besides, employing learnable prompts [49] further improves the performance by 1.4 % mAP, suggesting that learnable contexts help the model generalize better to downstream tasks. Moreover, adding source-style prompts results in improvements of 0.7% for handcrafted prompts and 0.3% for learnable prompts. This proves that incorporating source-style prompts helps the model better transfer target distributions to the source distribution.

4.5. Qualitative Analysis

Visualization of Detection Results. We present the visualization of detection results as shown in Figure 3. The proposed method demonstrates improved object classification accuracy with the source-biased object detector. For instance, in the night clear scene, the distinct *car* is misclassified as a *person* by CLIPGap, while our method provides accurate predictions. Furthermore, compared to CLIPGap, our approach better distinguishes foreground and background regions, effectively reducing false positive detections. Please refer to the Supplementary Material for more visualization results.

Table 9. Detection results and model efficiency comparison for different diffusion timestep T. Params denotes model parameter and IT denotes inference time.

Methods	T	Params (M)	DF	DR	NS	NR	Avg	IT (s)
Faster-RCNN	0	89.2	33.1	28.4	35.2	15.4	28.0	**0.151**
Ours	5	97.0	40.5	38.8	43.0	24.9	36.8	0.174
	10		**41.1**	**38.9**	43.1	25.4	37.1	0.198
	15		40.8	38.7	**43.6**	**25.8**	**37.2**	0.225

t-SNE Visualization. We perform t-SNE [38] visualizations to analyze feature representations of different methods. As shown in Figure 4 (a), while utilizing data augmentation to generate multiple augmented domains can align features between the source and unseen target domains, the extracted features lack discriminability, resulting in a significant number of false positives caused by misclassification. In contrast, Figure 4 (b) shows that our method not only aligns features across domains but also achieves better class separation, indicating that our source-biased object detector is more discriminative.

4.6. Limitations

The experimental results demonstrate the remarkable performance of our method. However, due to the iterative nature of the diffusion model, our method requires 198 ms per image at timestep $T = 10$, which is slower than Faster R-CNN, as shown in Table 9. Additionally, our approach incurs approximately 7.8M extra parameters to perform target-to-source distribution transfer. Exploring more efficient techniques to achieve faster processing speeds while reducing the additional model complexity presents an interesting direction for future research.

5. Conclusion

In this paper, we propose SDG-DiffDet, a novel diffusion-based framework for single-domain generalized object detection, which explicitly models target-to-source distribution transfer. SDG-DiffDet consists of two key modules: a memory-guided diffusion module that models feature augmentation during the diffusion process and a source-guided denoising module that facilitates distribution transfer from unseen domains to the source domain during the denoising process. Extensive experiments demonstrate that our method significantly outperforms existing approaches across two challenging domain generalization datasets.

Acknowledgements

This work was partially supported by National Nature Science Foundation of China (12150007, 62121002), Youth Innovation Promotion Association of CAS.

References

[1] Shoufa Chen, Peize Sun, Yibing Song, and Ping Luo. Diffusiondet: Diffusion model for object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 19830–19843, 2023. 3

[2] Sungha Choi, Sanghun Jung, Huiwon Yun, Joanne T Kim, Seungryong Kim, and Jaegul Choo. Robustnet: Improving domain generalization in urban-scene segmentation via instance selective whitening. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11580–11590, 2021. 6

[3] 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, 3, 6, 7

[4] Yingjun Du, Zehao Xiao, Shengcai Liao, and Cees Snoek. Protodiff: learning to learn prototypical networks by task-guided diffusion. *Advances in Neural Information Processing Systems*, 36:46304–46322, 2023. 3

[5] Mohammad Fahes, Tuan-Hung Vu, Andrei Bursuc, Patrick Pérez, and Raoul De Charette. Poda: Prompt-driven zero-shot domain adaptation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 18623–18633, 2023. 6

[6] Qi Fan, Mattia Segu, Yu-Wing Tai, Fisher Yu, Chi-Keung Tang, Bernt Schiele, and Dengxin Dai. Towards robust object detection invariant to real-world domain shifts. In *The Eleventh International Conference on Learning Representations (ICLR 2023)*. OpenReview, 2023. 6, 7

[7] Ross Girshick. Fast r-cnn. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1440–1448, 2015. 1

[8] Mahmoud Hassaballah, Mourad A Kenk, Khan Muhammad, and Shervin Minaee. Vehicle detection and tracking in adverse weather using a deep learning framework. *IEEE Transactions on Intelligent Transportation Systems*, 22(7):4230–4242, 2020. 5

[9] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE international conference on computer vision*, pages 1026–1034, 2015. 4

[10] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016. 6

[11] Cheng-Ju Ho, Chen-Hsuan Tai, Yen-Yu Lin, Ming-Hsuan Yang, and Yi-Hsuan Tsai. Diffusion-ss3d: Diffusion model for semi-supervised 3d object detection. *Advances in Neural Information Processing Systems*, 36:49100–49112, 2023. 3

[12] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020. 2, 3

[13] Lei Huang, Yi Zhou, Fan Zhu, Li Liu, and Ling Shao. Iterative normalization: Beyond standardization towards efficient whitening. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4874–4883, 2019. 6

[14] Xun Huang and Serge Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1501–1510, 2017. 2, 3

[15] 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 (CVPR)*, pages 16541–16550, 2024. 1

[16] Seung Wook Kim, Bradley Brown, Kangxue Yin, Karsten Kreis, Katja Schwarz, Daiqing Li, Robin Rombach, Antonio Torralba, and Sanja Fidler. Neuralfield-ldm: Scene generation with hierarchical latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8496–8506, 2023. 2

[17] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *nature*, 521(7553):436–444, 2015. 1

[18] 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. 3, 6, 7

[19] 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. 3, 6

[20] Hao Li, Wei Wang, Cong Wang, Zhigang Luo, Xinwang Liu, Kenli Li, and Xiaochun Cao. Phrase grounding-based style transfer for single-domain generalized object detection. *arXiv preprint arXiv:2402.01304*, 2024. 2, 3

[21] 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. 1

[22] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. Ssd: Single shot multibox detector. In *European Conference on Computer Vision*, pages 21–37. Springer, 2016. 1

[23] 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. 1, 2, 3, 6

[24] Yuning Lu, Jianzhuang Liu, Yonggang Zhang, Yajing Liu, and Xinmei Tian. Prompt distribution learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5206–5215, 2022. 1, 4, 5

[25] Andreas Lugmayr, Martin Danelljan, Andres Romero, Fisher Yu, Radu Timofte, and Luc Van Gool. Repaint: Inpainting using denoising diffusion probabilistic models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11461–11471, 2022. 3

[26] Xingang Pan, Ping Luo, Jianping Shi, and Xiaoou Tang. Two at once: Enhancing learning and generalization capacities via ibn-net. In *European Conference on Computer Vision*, pages 464–479, 2018. 6

[27] Xingang Pan, Xiaohang Zhan, Jianping Shi, Xiaoou Tang, and Ping Luo. Switchable whitening for deep representation learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1863–1871, 2019. 6

[28] Yuwen Pan, Rui Sun, Naisong Luo, Tianzhu Zhang, and Yongdong Zhang. Exploring reliable matching with phase enhancement for night-time semantic segmentation. In *European Conference on Computer Vision*, pages 408–424. Springer, 2024. 1

[29] Yuwen Pan, Rui Sun, Yuan Wang, Tianzhu Zhang, and Yongdong Zhang. Rethinking the implicit optimization paradigm with dual alignments for referring remote sensing image segmentation. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 2031–2040, 2024. 1

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

[31] 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. *arXiv preprint arXiv:2307.01750*, 2023. 1, 3, 6

[32] 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*, 28:91–99, 2015. 1, 6

[33] 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. 3

[34] 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

[35] Huihui Song, Tiankang Su, Yuhui Zheng, Kaihua Zhang, Bo Liu, and Dong Liu. Generalizable fourier augmentation for unsupervised video object segmentation. In *Proceedings of the AAAI conference on Artificial Intelligence*, pages 4918–4924, 2024. 7

[36] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020. 4

[37] Antonio Torralba and Alexei A Efros. Unbiased look at dataset bias. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1521–1528. IEEE, 2011. 1

[38] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9 (11), 2008. 8

[39] Vidit Vidit, Martin Engilberge, and Mathieu Salzmann. Clip the gap: A single domain generalization approach for object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3219–3229, 2023. 1, 2, 3, 6, 7, 8

[40] Pascal Vincent, Hugo Larochelle, Yoshua Bengio, and Pierre-Antoine Manzagol. Extracting and composing robust features with denoising autoencoders. In *Proceedings of the 25th international conference on Machine learning*, pages 1096–1103, 2008. 7

[41] Kunyu Wang, Xueyang Fu, Yukun Huang, Chengzhi Cao, Gege Shi, and Zheng-Jun Zha. Generalized uav object detection via frequency domain disentanglement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1064–1073, 2023. 1

[42] 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. 1, 2, 3, 5, 6

[43] Aming Wu, Rui Liu, Yahong Han, Linchao Zhu, and Yi Yang. Vector-decomposed disentanglement for domain-invariant object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9342–9351, 2021. 5

[44] 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. 3

[45] 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. 2, 3

[46] 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

[47] 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 *International Conference on Learning Representations*, 2022. 1

[48] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3836–3847, 2023. 2, 3

[49] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *International Journal of Computer Vision*, 130(9):2337–2348, 2022. 8