On the Importance of Dual-Space Augmentation for Domain Generalized Object Detection

Hayoung Park Choongsang Cho Guisik Kim
Korea Electronics Technology Institute (KETI)
{hyformal,ideafisher}@keti.re.kr, specialre@naver.com

Abstract

The distribution gap between training data and real-world data often causes significant performance drops in networks trained via naive supervised learning. To address this, domain generalization methods have been developed to gain robust performance in unseen domains. In this paper, we propose a single-domain generalized object detection (S-DGOD) method. Unlike previous works, we utilize both image-level and feature-level augmentations and experimentally demonstrate their synergistic effects. Image-level augmentations expand the source domain, while feature-level augmentations leverage CLIP to incorporate potential domain descriptions. Our method achieves superior performance, with 29.2% mAP on the Cityscapes-C and 37.1% mAP on the Diverse-Weather dataset.

1. Introduction

Object detection plays a crucial role in various deep learning fields, including surveillance, self-navigation, crowd counting, and manufacturing. A detection model typically performs well when the input data shares the same domain as the training data. However, performance can significantly degrade when the target domain differs from the training source [5, 11, 26]. A straightforward solution is to create a new dataset for the target domain, but this approach is time-consuming and costly, especially since instance-wise labeling for detection is more complex than image-wise labeling for classification.

Domain generalization (DG) aims to develop models that generalize well to unseen domains without requiring target domain annotations or images. By relying solely on source domain data, DG can achieve robust performance across a variety of applications, including urban scene segmentation [2, 8, 20, 53] and medical image processing [14, 23, 30].

Single-domain generalized object detection (S-DGOD) focuses on object detection using training data from a single domain. Most research in this area emphasizes designing image [4, 7, 10, 12, 16, 22, 24, 39, 43, 52] or feature [17, 19, 31, 37, 38, 40, 41] augmentations to expand data distribution, allowing models to better handle target domain data. In contrast to prior works, we incorporate both image-level and feature-level augmentations and demonstrate their effectiveness when used simultaneously. Fig. 1 illustrates how our augmentations work, showcasing sample images, their augmented versions, intermediate features, and augmented features.

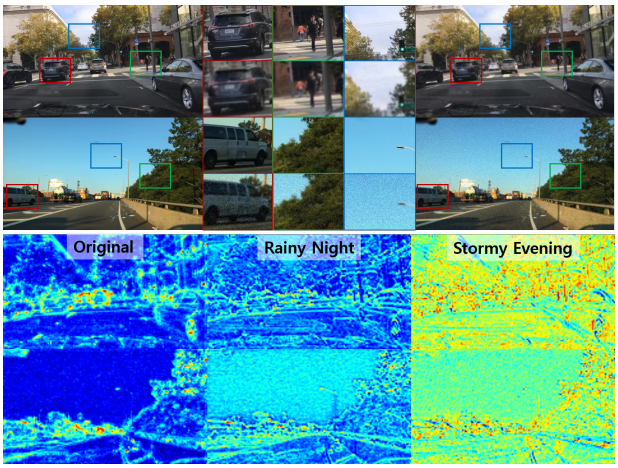

Figure 1. Visualization of our image-level augmentation (top) and feature-level augmentation (bottom). For the image-level augmentation, we show original images of the source domain on the left, and results of our image-level augmentation on the right. Cropped parts are shown at the middle to help visualizing the difference. For the feature-level augmentation, we denote mean images of intermediate features that has extracted from augmented images as *original*, and show their augmented versions with augmentations corresponding to *rainy night* and *stormy evening*.

Our image-level augmentation is designed to be robust enough to push the boundaries of the source data distribution, thereby improving the model's generalizability while is easy to implement and apply, and consumes low computational resources. We found that augmentations used in low-level vision problems [27, 34, 51] not only meet this requirement but are also ideal for training detection models,

as they preserve the spatial information of instances. We experimentally show the effectiveness of our image-level augmentation on single-DGOD despite of its simplicity.

For feature-level augmentation, we adopt CLIP [32] for extracting text features to enhance source image features. Neural style injection [15] and algebraic operations [29] rooted from text descriptions of potential unseen target domain are used for reflecting new characteristics to source image features. Text descriptions used in our work refer from [38] with expansion of target domain numbers. Even though their approach has similar aspect with ours, our work differs from theirs for some points; we did not used image encoder of CLIP trained jointly with text encoder; we propose a new strategy for projecting target domain characteristics that can effectively reflect target statistics compared to addition-based transformation.

Our contributions can be summarized as follows:

- We propose a single-DGOD framework based on a modern detector architecture that leverages multi-scale features, deformable operations, and attention mechanisms. By replacing the outdated backbone network with a state-of-the-art model, our approach establishes a new benchmark for future studies in the field.

- We propose a dual-space augmentation framework where image-level augmentations enhance model robustness, and text-driven feature-level augmentations enable the model to learn domain-specific characteristics. The synergy between the two methods significantly improves generalization.

- Our proposed training framework demonstrates substantial improvements in generalization ability, outperforming previous state-of-the-art methods by a significant margin.

2. Related Work

2.1. Domain Generalization for Object Detection

Domain generalization for object detection has been less explored compared to other fields such as classification [1, 21, 54] and segmentation [8, 22, 36, 47] but has recently gained attention due to its potential to reduce the high annotation costs associated with object detection. Wu et al. [44] proposed domain generalization for urban scenes using cyclic disentangled self-distillation. Vidit et al. [38] leveraged CLIP for single-DGOD problem by learning additional augmentations to mimic target features. Lee et al. [17] employed object-aware image transformation and contrastive loss to improve generalizability and detection performance. Liu et al. [24] utilized causal learning and global-local transformation to address DGOD. Danish et al. [10] highlighted the importance of instance-wise alignment by using various image degradation techniques to enhance DGOD performance. G-NAS [46] proposed an algorithm utilizing neural architecture search. Li et al. [18] introduced a prompt-based object-centric gating module.

2.2. Representation of Image and Text

Many studies have aimed to improve image processing performance by leveraging multi-modal inputs, such as combining images with audio, text, or optical flow. In our approach, we use text features alongside visual representations to achieve better domain generalization.

UNITER [6] is a pretrained model designed to learn joint image-text representations and perform well across various vision-language tasks. CLIP [32], trained on 400 million image-text pairs, minimizes the distance between image and text features and excels in tasks like zero-shot classification and domain generalization.

Mikolov et al. [29] demonstrated that algebraic operations could be applied to text features. For instance, by subtracting the feature vector of *Man* from that of *Groom*, and adding the feature of *Woman*, the resultant vector approximates the feature of *Bride*. We leverage this property to project potential target domain characteristics into a feature extracted from source domain image.

2.3. Augmentation for Domain Generalization

Image augmentation plays a crucial role in enhancing a model's domain generalization by expanding the source data distribution. Numerous studies have proposed image-level augmentation techniques for domain-generalized object detection. Volpi et al. [39] introduced iterative adversarial augmentation, while Cheng et al. [7] combined adversarial learning with Bayesian neural networks. Li et al. [22] utilized generative models to transfer image styles for improved domain generalization. Chen et al. [4] proposed augmentation that pushes samples away from the center of distribution. Jiang et al. [16] constructed data augmentation based on contrastive relationship between samples. Wang et al. [43] augmented images to be complementary to the style of source domain samples. Gokhale et al. [12] learned image transformations in an adversarial manner. Zheng et al. [52] applied learnable semantic transformation on image level.

Feature augmentation has also been extensively explored. Shu et al. [37] mixed domains in the feature space, and Li et al. [19] demonstrated that adding noise to features improves generalizability. Wang et al. [40, 41] investigated the impact of semantic feature augmentation, and Qi et al. [31] introduced a novel batch normalization technique for feature diversification. Zeng et al. [49] mimic unobserved target domain feature by forecasting direction of the domain shift. Liu et al. [25] transformed domain-specific

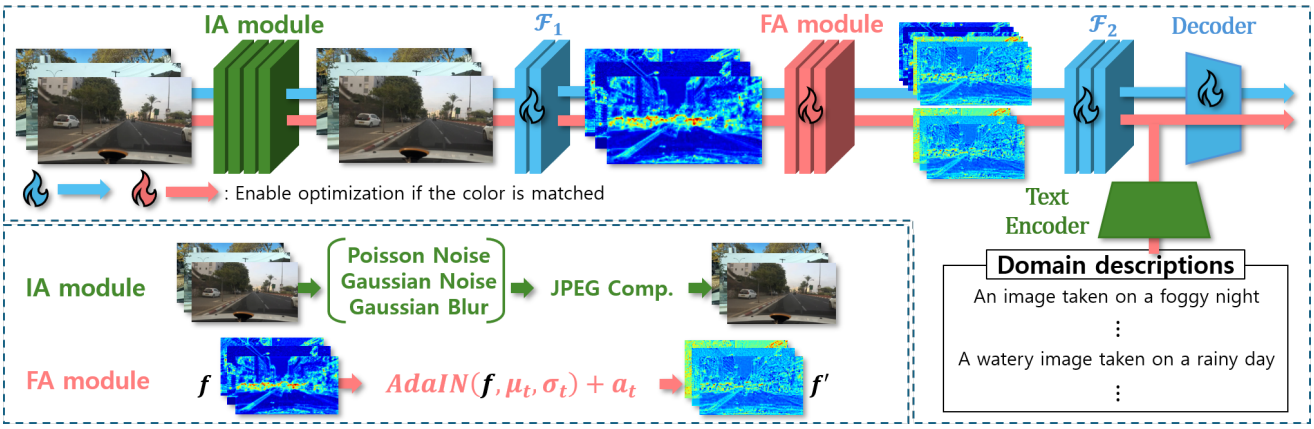

Figure 2. **Overview of the proposed method.** The proposed method consists of an Image-level Augmentation (IA) module and a text-driven Feature Augmentation (FA) module, where the FA and concatenation with the original features are applied at the end of the second layer of the feature extractor. For more details, refer to the supplementary materials.

features while class-specific features are preserved. Wang et al. [42] randomized style of features to achieve diversity.

In contrast to prior works, our approach integrates both image-level and feature-level augmentation, leveraging the strengths of each to maximize domain generalization.

3. Method

In this section, we describe our framework for single-DGOD. Our training process consists of two stages: one for training dual-space augmentations and another for training the detector with the learned augmentations. Fig. 2 shows overall procedure of our framework. In Section 3.1, we briefly introduce the advantages of using a Transformer-based backbone model, which has been shown to improve generalization and performance. Then, in Section 3.2, we present our dual-space augmentation strategy. Finally, in Section 3.3, we describe the training procedure that incorporates both image-level and feature-level augmentations to improve generalization performance.

3.1. Toward Transformer-based Model

Most prior works on single-DGOD have utilized Faster R-CNN [33], which was introduced in the mid-2010s. However, Faster R-CNN lacks the ability to leverage multi-scale features, leading to suboptimal performance for small objects. Additionally, it heavily relies on proposals generated by predefined anchor boxes, which limits flexibility. Moreover, due to its reliance on a region proposal network, RoI pooling, and bounding box regression, the learning process is known to be complex.

In contrast, transformer-based approaches, which have gained popularity since the introduction of DETR [3], offer a more streamlined, end-to-end structure and deliver superior performance. These models benefit from multi-scale features, deformable operations, and attention mechanisms. As a result, many recent object detection studies have shifted toward transformer-based architectures.

Following this trend and considering the flexibility of transformers, we replaced the baseline model with a transformer-based architecture [50], which leverages these advanced features and mechanisms. This straightforward modification to the baseline model can surpass previous state-of-the-art (SOTA) performance on certain datasets, as demonstrated in the experimental section. By incorporating these enhancements, we set a new performance benchmark.

3.2. Dual-Space Augmentation Framework

Since we only have access to source domain data during the training phase, expanding the input distribution is crucial for improving the model's generalizability. However, because the target domain's context (e.g., tone, texture, brightness, contrast) is unknown, image-level augmentation alone cannot fully capture the diversity of the target domain. Therefore, while general augmentation techniques are applied to enhance model robustness and maintain performance in the source domain, we introduce text-driven feature augmentation, leveraging domain-specific descriptions to capture relevant characteristics from the target domain.

3.2.1 Image-level Augmentation

For image-level augmentation, we design transformations that are expressive enough to expand the data distribution and improve the model's generalizability, while preserving the key information necessary for accurate detection. We employ augmentations such as noise, blur, and JPEG compression to enhance model robustness across various domains. The bottom of Fig. 2 illustrates the image-level augmentation process. These augmentations minimally affect source domain performance, with domain-specific charac-

teristics addressed through text-driven augmentation at the feature level. Experimentally, we demonstrate that combining image-level and feature-level augmentations significantly improves generalizability.

3.2.2 Text-driven Feature-level Augmentation

Although target domain data is unavailable, we can hypothesize potential conditions of the target domain and describe them in a sentence, such as *a watery image taken on a stormy day*. We slightly modified the descriptions from [38] by adding new possible conditions for our work. To leverage these descriptions, we use CLIP's text encoder to incorporate target characteristics derived from the domain descriptions into the source image features. This is possible because CLIP's text encoder is jointly trained with images.

Importantly, this approach does not limit generalizability by the number of descriptions, as it is combined with image-level augmentations. In our experiments, we found that using feature-level augmentation alone could introduce domain bias, negatively affecting generalization. However, when used in conjunction with image-level augmentation, it significantly enhances generalizability. Furthermore, adding new domain information is straightforward, requiring only a single descriptive sentence.

Let $\mathcal{S} = \{x_s, y_s | 1 \leq s \leq N_{src}\}$ denote source domain data with N_{src} image-annotation pairs, d_s is source domain description, and $\mathcal{D}_{tgt} = \{d_t | 1 \leq t \leq N_{tgt}\}$ be target domain descriptions where N_{tgt} is the number of possible domains. We first augment input image x_s from source domain using augmenter $\mathcal{T}$ and pass through a module $\mathcal{F}_1$ which projects image x_s to the feature space where feature-level learnable augmentation will be applied. The augmentation process in this space can be described as follows:

$$f_s = \mathcal{F}_1(\mathcal{T}(x_s)), \tag{1}$$

$$f' = AdaIN(f_s, \mu_t, \sigma_t) + a_t, \tag{2}$$

where μ_t and σ_t are learnable style augmentation parameters, a_t is learnable addition augmentation, and $AdaIN(x, \mu, \sigma) = \mu + \sigma \frac{x - x_\mu}{x_\sigma}$.

The augmented feature f' is projected to image-text joint hyperspace through the second module $\mathcal{F}_2$ to obtain final feature $\tilde{f}$ that is expected to have target domain characteristics and share the space with text representation. The target domain characteristics can be learned by minimizing cosine loss between $\tilde{f}$ and f^*, which is computed via combining image feature and text features. Note that our feature extractors $\mathcal{F}_1$ and $\mathcal{F}_2$ is not jointly learned with text encoder, they can still used for training learnable augmentations that improves model generalizability. The procedure for training learnable augmentation parameters can be written as:

$$\tilde{f} = \mathcal{F}_2(f'), \tag{3}$$

$$f^* = \mathcal{F}_2(f_s) + \mathcal{E}_{txt}(d_t) - \mathcal{E}_{txt}(d_s), \tag{4}$$

$$\mathcal{L}_{aug} = cos(\tilde{f}, f^*) + \alpha \cdot l_1(\tilde{f}, \mathcal{F}_2(f_s)), \tag{5}$$

where $\mathcal{E}_{txt}$ is the CLIP text encoder, and α is a weight parameter which is set to 1 in our work. Algorithm 1 briefly shows the overall procedure of learning text-driven feature augmentation parameters.

Algorithm 1 Training Text-driven Feature-level Augmentation

Input: Source image x_s, domain description set $\mathcal{D}_t$, augmentation parameters $\{\mu_t, \sigma_t, a_t\}$, image feature extractors $\mathcal{F}_1$, $\mathcal{F}_2$, text encoder $\mathcal{E}_{txt}$, image augmentation module $\mathcal{T}$, loss weight parameter α
Output: Learned parameters $\{\hat{\mu}_t, \hat{\sigma}_t, \hat{a}_t\}$
Initialize variables: $\{\mu_t\} \leftarrow \mathbb{0}$, $\{\sigma_t\} \leftarrow \mathbb{1}$, $\{a_t\} \leftarrow \mathbb{0}$
Extract image feature: $f_s = \mathcal{F}_1(\mathcal{T}(x_s))$
Source domain text features: $\bar{d_s} = \mathcal{E}_{txt}(d_s)$
for t **in** N_{tgt} **do**
    Feat. augmentation: $f'_s = AdaIN(f_s, \mu_t, \sigma_t) + a_t$
    Feat. projection: $\tilde{f} = \mathcal{F}_2(f'_s)$
    Target domain text features: $\bar{d_t} = \mathcal{E}_{txt}(d_t)$
    Algebraic operations: $f^* = \mathcal{F}_2(f_s) + \bar{d_t} - \bar{d_s}$
    Compute loss: $cos(\tilde{f}, f^*) + \alpha \cdot l_1(\tilde{f}, \mathcal{F}_2(f_s))$
    Update μ_t, σ_t, a_t
end for
Return results: $\{\hat{\mu}_t, \hat{\sigma}_t, \hat{a}_t\}$

3.3. Training Object Detection Model

After training dual-space augmentations that reflect potential target domain characteristics, the detector can learn generalizable representations. For image-level augmentation, we follow the same process but apply additional techniques for feature-level augmentation to enhance feature representation learning. Feature augmentation is applied after the second layer of the backbone network by injecting neural features and incorporating additional augmentation parameters. The augmented features are then concatenated with the pre-augmentation features and passed into the subsequent layers. By concatenating original and augmented features along the batch dimension, the model learns both feature types simultaneously and enforces consistency between them. This combined mini-batch introduces diversity, allowing the model to generalize across domains while maintaining prediction consistency. We experimentally show that feature-level augmentation alone may decrease performance, but when combined with image-level augmentation, it significantly improves generalizability.

Since our learned feature augmentations are guided by the CLIP text encoder, we adopt a loss function $\mathcal{L}_{clip\text{-}t}$ from

Method	Noise			Blur				Weather				Digital				Avg.
	Gauss.	Shot	Impulse	Defocus	Glass	Motion	Zoom	Snow	Frost	Fog	Bright	Cont.	Elastic	Pixel	JPEG	
OA-DG [17]	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	23.8	22.0	21.8
Baseline	11.4	15.3	9.8	39.1	29.8	31.3	**6.9**	6.0	19.2	41.0	**45.8**	33.4	52.0	44.9	39.4	28.4
Ours	**12.1**	**15.6**	**11.2**	**40.4**	**30.6**	**32.7**	6.6	**6.4**	20.1	**41.7**	**45.8**	**35.6**	**52.9**	**45.6**	**41.4**	**29.2**

Table 1. **Quantitative Comparison.** Domain-wise mAP@0.5 performances on test set of Cityscapes-C [28] dataset.

[38], which minimizes cross-entropy between features extracted by the detector and text features of the corresponding instance descriptions. Although our dual-space augmentation is designed to retain spatial and content integrity, it may still confuses the model. Therefore we apply foreground contrastive loss $\mathcal{L}_{cnt}$ which focuses on object features rather than domain differences to mitigate this effect. This loss minimizes Euclidean distance between features of the same class, regardless of domain, while maximizing the distance for different classes. Our experimental results show the effectiveness of $\mathcal{L}_{cnt}$, despite its simplicity in counting distances between foreground objects only, differing from complicated losses that consider background objects. However, minimizing or maximizing feature distances based on class labels can be harmful when predictions are noisy, particularly in the early training stages. To address this, we restrict the contribution to $\mathcal{L}_{cnt}$ to features with a confidence score above a certain threshold and whose predicted class matches the ground truth label. This restriction initially limits the number of contributing features but gradually allows more features to contribute as training progresses.

The overall training loss is described as:

$$\mathcal{L} = \mathcal{L}_{det} + w_1 \cdot \mathcal{L}_{clip\text{-}t} + w_2 \cdot \mathcal{L}_{cnt}, \tag{6}$$

where $\mathcal{L}_{det}$ represents the conventional detection loss including box regression and classification losses, $\mathcal{L}_{clip\text{-}t}$ and $\mathcal{L}_{cnt}$ denote the clip-t loss and contrastive loss, respectively, with w_1 and w_2 as their corresponding weight parameters which we used 1 for both of them in our work.

After the training process, no further augmentations are required to perform object detection on unseen domains. Despite having no access to images from those domains, the model demonstrates high performance due to the generalizability learned through dual-space augmentations.

4. Experiments

In this section, we describe the setting used for our experiments and show the effectiveness of proposed training scheme.

4.1. Setup

Dataset. Following prior works [10, 17, 24, 38], we report mAP under a 0.5 IoU threshold using the Diverse-Weather Dataset (DWD) [44], the Cityscapes-C dataset [28], and the Foggy Driving dataset [35]. DWD consists of five sets of images collected under different environmental conditions: daytime sunny (27,708 images), night sunny (26,158 images), dusk rainy (3,501 images), night rainy (2,494 images), and daytime foggy (3,775 images). The dataset was derived from BDD-100k [48], synthetic rainy versions [45], Foggy Cityscapes [35], and Adverse-Weather dataset [13]. The Cityscapes-C dataset is a benchmark built upon Cityscapes dataset [9], featuring 19 different transformations across 5 severity levels, resulting in 95 types of manipulations and a total of 47,500 images. The dataset is divided into a validation and a test set, where we evaluate our model on the test set following [17]. The Foggy Driving dataset is a real-world dataset that contains 101 foggy images, with 51 images collected by a phone camera and 50 images from web.

Method	Day sunny	Night sunny	Dusk rainy	Night rainy	Day foggy	Avg.
Cyclic-SD [44]	56.1	36.6	28.2	16.6	33.5	28.7
CLIP-G [38]	52.6†	36.9	32.3	18.7	38.5	31.6
OA-DG [17]	55.8	38.0	33.9	16.8	38.3	31.8
G-NAS [46]	58.4	45.0	35.1	17.4	36.4	33.5
U-FRCNN [24]	58.6	40.8	33.2	19.2	**39.6**	33.2
DivAlign [10]	52.8	42.5	38.1	**24.1**	37.2	35.5
Prompt-D [18]	53.6	38.5	33.7	19.1	39.0	32.6
Ours	**64.1**	**47.3**	**39.7**	22.4	38.9	**37.1**

†The performance of official weight tested on local environment.

Table 2. **Quantitative Comparison.** Domain-wise performance on DWD [44].

Implementation Details. We utilized the PyTorch library and two Nvidia A5000 GPUs for our training steps. The learnable augmentations were optimized over 200 epochs with a batch size of 16 and a learning rate of 4e-4 using the AdamW optimizer. The detection model was trained for 20 epochs with a batch size of 2, while all other settings remained unchanged. Contrastive loss is computed based on Euclidean distance between two vectors from the output of transformer encoder, maximizing inter-class distance and minimizing intra-class distance regardless of augmentation, considering only foregrounds. We applied margin of 0.5 for maximizing inter-class distance.

Augmentation Details. For image-level augmentations, we applied Gaussian blur, Gaussian noise, Poisson noise, and JPEG compression. The order of augmentations was randomized for each iteration, except for JPEG compression,

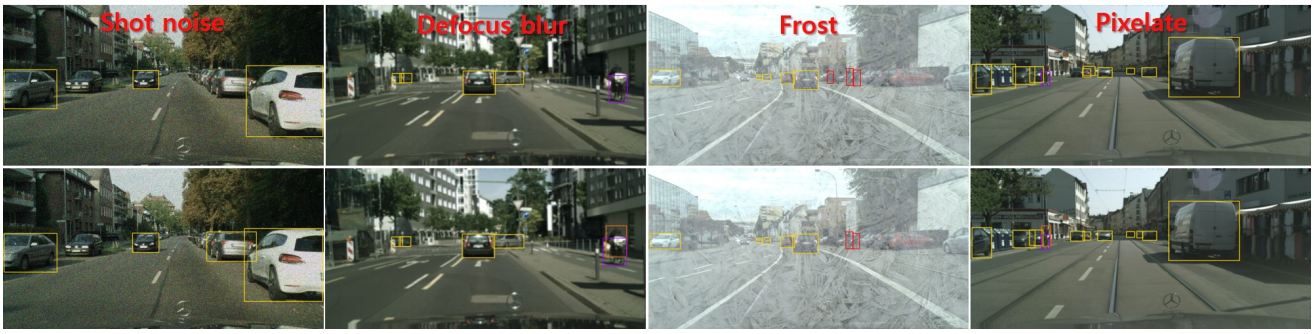

Figure 3. **Qualitative Comparison.** To evaluate performance on unseen domains, our model was trained using only Cityscapes images. The top row shows results from the baseline detector, while the bottom row displays the results from proposed method on Cityscapes-C [28].

Method	bus	bike	car	motor	person	rider	truck	mAP
Cyclic-SD [44]	**68.8**	50.9	53.9	**56.2**	41.8	52.4	**68.7**	56.1
CLIP-G† [38]	55.0	47.8	67.5	46.7	49.4	46.8	54.7	52.6
U-FRCNN [24]	66.8	51.0	70.6	55.8	49.8	48.5	67.4	58.6
Ours	63.2	**53.6**	**84.9**	55.0	**67.8**	**57.7**	66.2	**64.1**

†The performance of official weight tested on local environment.

Table 3. **Quantitative Comparison.** Class-wise performances on daytime-sunny domain of DWD [44].

which was always applied last. The kernel size for Gaussian blur was randomly selected between 7 and 25, with the σ value chosen between 0.0 and 2.8. Gaussian noise was randomly generated as either RGB or grayscale, with a zero mean and a standard deviation randomly selected between $\frac{2}{255}$ and $\frac{25}{255}$. Poisson noise is applied with a probability of 0.1. For feature-level augmentation, we used CLIP to extract text features from a total of 33 domain descriptions, which included weather, time of day, and noise conditions, to maximize generalization ability.

4.2. Performance

4.2.1 Comparison with Other Models.

We compared our work with several state-of-the-art single-DGOD methods, which demonstrated the effectiveness of our proposed framework. Results of other papers are directly imported from reported numbers except noticed explicitly. Table 1 shows comparison of previous state-of-the-art [17], baseline model, and our proposed model performance using Cityscapes-C dataset [28]. Our method outperforms comparing methods at 13 types of corruptions, achieving an average mAP of 29.2%. Noticeable point is that our baseline model shows better performance compared to the previous state-of-the-art [17] which strongly supports our argument on necessity of replacing backbone network to a recent model. Fig. 3 visualizes the test results of the baseline and the proposed method across various corrupted domains.

Table 2 shows a comparison of our model and single-SGOD methods including Cyclic-SD [44], CLIP-G [38], G-NAS [46], U-FRCNN [24], and Prompt-D [18] on domains provided in DWD. Our method shows strong performance on mAP, achieving an average of 37.1% which outperforms the previous state-of-the-art by 1.6% mAP. We also compare class-wise performances across each domains at following tables, note that the comparison is done only if per-class results exist. As shown in Table 3 which depicts results on daytime-sunny (source) domain, our model outperforms previous works with a large margin on *car* and *person* class which occupy 82% and 10% of all annotations, respectively, while shows relatively weak performance on *motor* (motorcycle) class which occupies only 0.45%. This tendency is due to the class-imbalance characteristic of specific data and the fact that transformers are data-hungry, which can be resolved by applying additional techniques that mitigates class imbalance in the training dataset.

Method	bus	bike	car	motor	person	rider	truck	mAP
Cyclic-SD [44]	40.6	35.1	50.7	19.7	34.7	32.1	43.4	36.6
CLIP-G [38]	37.7	34.3	58.0	19.2	37.6	28.5	42.9	36.9
G-NAS [46]	46.9	**40.5**	67.5	**26.5**	50.7	35.4	47.8	45.0
U-FRCNN [24]	43.6	38.1	66.1	14.7	49.1	26.4	47.5	40.8
Prompt-D [18]	40.9	35.0	59.0	21.3	40.4	29.9	42.9	38.5
Ours	**49.2**	39.2	**73.3**	24.4	**57.2**	**35.7**	**52.2**	**47.3**

Table 4. **Quantitative Comparison.** Class-wise performances on night-clear domain of DWD [44].

Method	bus	bike	car	motor	person	rider	truck	mAP
Cyclic-SD [44]	37.1	19.6	50.9	13.4	19.7	16.3	40.7	28.2
CLIP-G [38]	37.8	22.8	60.7	16.8	26.8	18.7	42.4	32.3
G-NAS [46]	44.6	22.3	66.4	14.7	32.1	19.6	45.8	35.1
U-FRCNN [24]	37.1	21.8	67.9	16.4	27.4	17.9	43.9	33.2
Prompt-D [18]	39.4	25.2	60.9	**20.4**	29.9	16.5	43.9	33.7
Ours	**45.6**	**26.8**	**75.8**	16.0	**40.3**	**22.3**	**51.1**	**39.7**

Table 5. **Quantitative Comparison.** Class-wise performances on dusk-rainy domain of DWD [44].

Night-clear domain. Table 4 shows results on the night-sunny domain of DWD. Our method shows better performance compared to previous works for 5 classes. As a re-

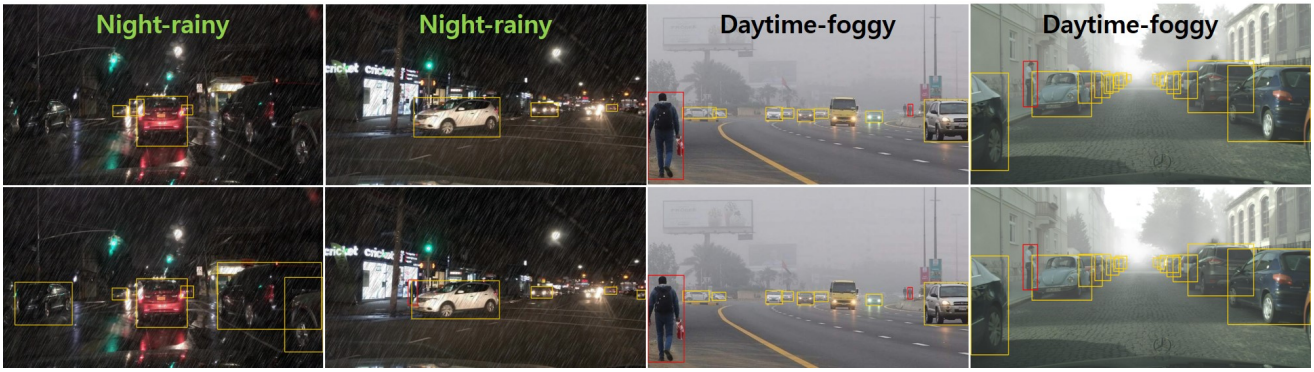

Figure 4. **Qualitative Comparison.** To evaluate performance on unseen domains, the model was trained using only daytime-sunny images. The top row shows results from the baseline model, while the bottom row displays the results from proposed method on DWD [44].

Method	bus	bike	car	motor	person	rider	truck	mAP
Cyclic-SD [44]	24.4	11.6	29.5	9.8	10.5	11.4	19.2	16.6
CLIP-G [38]	28.6	12.1	36.1	9.2	12.3	9.6	22.9	18.7
G-NAS [46]	28.6	9.8	38.4	0.1	13.8	9.8	21.4	17.4
U-FRCNN [24]	29.9	11.8	36.1	9.4	13.1	10.5	23.3	19.2
Prompt-D [18]	25.6	12.1	35.8	**10.1**	14.2	**12.9**	22.9	19.1
Ours	**33.1**	**18.6**	**47.0**	3.9	**19.8**	8.5	**25.6**	**22.4**

Table 6. **Quantitative Comparison.** Class-wise performances on night-rainy domain of DWD [44].

Method	bus	bike	car	motor	person	rider	truck	mAP
Cyclic-SD [44]	32.9	28.0	48.8	29.8	32.5	38.2	24.1	33.5
CLIP-G [38]	36.1	34.3	58.0	33.1	39.0	43.9	25.1	38.5
G-NAS [46]	32.4	31.2	57.7	31.9	38.6	38.5	24.5	36.4
U-FRCNN [24]	**36.9**	**35.8**	61.7	**33.7**	39.5	42.2	27.5	**39.6**
Prompt-D [18]	36.1	34.5	58.4	33.3	40.5	**44.2**	26.2	39.0
Ours	35.0	30.6	**64.2**	28.6	**43.3**	42.7	**27.6**	38.9

Table 7. **Quantitative Comparison.** Class-wise performances on daytime-foggy domain of DWD [44].

sult, our framework achieved 46.9% mAP gaining 2.3% mAP compared to the previous state-of-the-art. The left column of Fig. 5 shows that the proposed method has better detection performance compared to the baseline.

Dusk-rainy domain. Table 5 shows results on the dusk-rainy domain of DWD. All classes exceed compared methods except *motor* which is the rarest class, resulting in 39.7% mAP. In the right column of Fig. 5, the proposed method also shows superior detection, particularly for people and buses, when compared to the baseline.

Night-rainy domain. Table 6 shows results on the night-rainy domain, which is the most challenging set in DWD where all previous single-DGOD methods consistently showed poor performance. Even though the results of *motor* and *rider* classes show low scores, our method still demonstrates strong performance, achieving 22.4% mAP. The left half of Fig. 4 shows that the proposed method demonstrates better detection performance compared to the baseline.

Daytime-foggy domain. Table 7 shows results on the daytime-foggy domain of DWD. In this domain, our method showed slightly lower performance than the previous state-of-the-art, achieving 38.9% mAP which is 0.7% mAP behind. Classes with a minor number of annotations, such as *bus*, *bike*, *motor*, and *rider*, where their combined portion is 3.6% of all boxes, lowered the overall result. In the right column of Fig. 4, both the baseline and the proposed method show a certain level of performance, but challenges remain in areas with very dense fog.

Real-world foggy domain. We also evaluate our model in a real-world setting using Foggy Driving [35] dataset. Fig. 6 visualizes results of baseline and proposed method. Table 8 shows quantitative results, indicating that our method improves detection performance under challenging real-world condition.

4.2.2 Ablation Study

We conducted an ablation study to validate the effectiveness of our proposed training framework, particularly focusing on how the combination of two types of augmentation significantly enhances generalization performance, even though one component may degrade performance when applied alone. Table 9 presents the results, where IA, FA, and cont refer to image-level augmentation, feature-level augmentation, and contrastive loss, respectively.

Analysis of augmentations. As shown in the first two rows of Table 9, our simple image-level augmentations, comprising blur, noise, and JPEG compression, improved model generalizability by an average of 1.1% mAP across four unseen domains. This result demonstrates that even lightweight transformations of input images can yield noticeable improvements in the single-DGOD problem.

Analysis of training design. As seen in the fourth row of Table 9, combining image-level and feature-level augmentations improved performance by 1.8% compared to the baseline, despite the decrease observed when using feature-

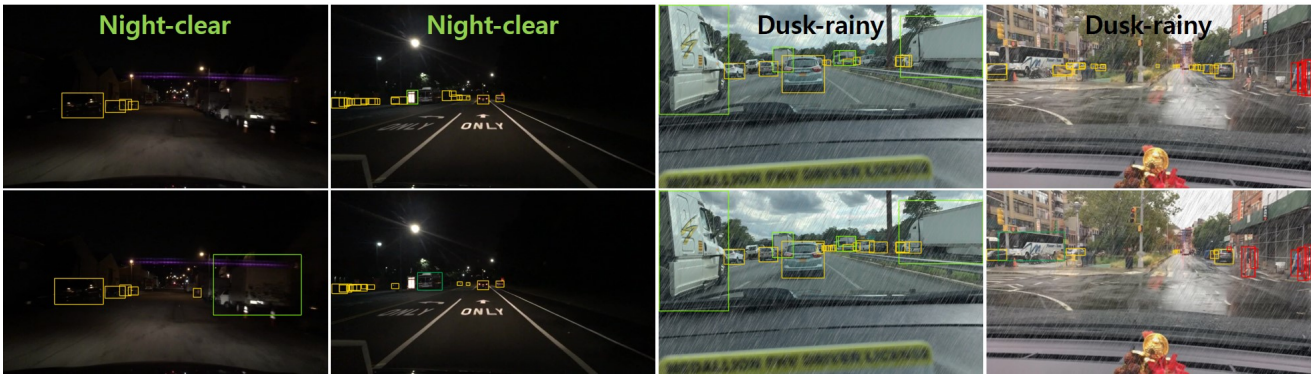

Figure 5. **Qualitative Comparison.** To evaluate performance on unseen domains, our model was trained using only daytime-sunny images. The top row shows results from the baseline detector, while the bottom row displays the results from proposed method on DWD [44].

Figure 6. **Qualitative Comparison.** Results of baseline (left) and our model (right) on Foggy Driving [35] dataset.

Method	bus	bike	car	motor	person	rider	train	truck	mAP
Baseline	45.8	**47.4**	58.1	9.2	**28.3**	18.8	1.1	**49.5**	32.3
Ours	**46.1**	42.6	**58.3**	**10.1**	22.3	**25.1**	**5.9**	49.4	**32.5**

Table 8. **Quantitative Comparison.** Class-wise performances on Foggy Driving [35] dataset.

level augmentation alone. This result indicates that our proposed framework, which leverages both types of augmentation simultaneously, compensates for the limitations of each method. Notably, the result in the fourth row is 0.7% higher than that of image augmentation alone. The last row shows the complete model, where the contrastive loss is added to both image and feature augmentations, achieving an average mAP of 37.1%, a significant improvement of 2.6% over the baseline.

5. Discussion and Limitation

There is potential for further expanding the proposed image-level augmentation techniques. While our current approach focuses on enhancing model robustness, more advanced augmentation methods could be developed to model specific target domain conditions, such as various weather scenarios. Additionally, there is room to expand the range of text domain descriptions.

A limitation of our approach, inherent to Transformer-based models, is their reliance on large datasets. This issue becomes evident when dealing with class imbalance, where performance on minority classes remains suboptimal. This limitation is clearly reflected in the class-wise performance (Table 5 and 6). It can be viewed as trade-off introduced by using transformer-based architecture as our baseline model.

Components			Night	Dusk	Night	Day	Avg.
IA	FA	cont	sunny	rainy	rainy	foggy	
✗	✗	✗	45.1	35.4	16.7	**40.7**	34.5
✓	✗	✗	46.1	37.8	19.4	39.2	35.6
✗	✓	✗	45.7	35.0	16.8	39.2	34.2
✓	✓	✗	**47.4**	38.9	20.7	38.3	36.3
✓	✓	✓	47.3	**39.7**	**22.4**	38.9	**37.1**

Table 9. **Ablation study** of the proposed training components on DWD [44]. *IA*, *FA*, and *cont* refers to image-level augmentation, feature-level augmentation, and contrastive learning, respectively.

6. Conclusion

In this paper, we propose a novel framework that combines both image-level and feature-level augmentations, distinguishing our approach from previous studies. We utilize the CLIP text encoder to project unseen target domain characteristics into the feature space through neural style injection and algebraic operations, while image-level augmentations compensate for the bias introduced by feature-level modifications. To mitigate potential disruptions from these augmentations, we incorporate a foreground contrastive loss and demonstrate its effectiveness. Experimental results show that our method significantly improves generalizability in the object detection field. For future work, we aim to develop more effective ways to represent unseen domain characteristics.

Acknowledgments

This work was partly supported by Institute of Information communications Technology Planning Evaluation (IITP) grant funded by the Korea government (MSIT) (No. RS-2022-II220926, RS-2024-00456709)

References

[1] Sravanti Addepalli, Ashish Ramayee Asokan, Lakshay Sharma, and R Venkatesh Babu. Leveraging vision-language models for improving domain generalization in image classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23922–23932, 2024. 2

[2] Qi Bi, Shaodi You, and Theo Gevers. Learning content-enhanced mask transformer for domain generalized urban-scene segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 819–827, 2024. 1

[3] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European conference on computer vision*, pages 213–229. Springer, 2020. 3

[4] Tianle Chen, Mahsa Baktashmotlagh, Zijian Wang, and Mathieu Salzmann. Center-aware adversarial augmentation for single domain generalization. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 4157–4165, January 2023. 1, 2

[5] Yuhua Chen, Wen Li, Christos Sakaridis, Dengxin Dai, and Luc Van Gool. Domain adaptive faster r-cnn for object detection in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3339–3348, 2018. 1

[6] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. Uniter: Universal image-text representation learning. In *European conference on computer vision*, pages 104–120. Springer, 2020. 2

[7] Sheng Cheng, Tejas Gokhale, and Yezhou Yang. Adversarial bayesian augmentation for single-source domain generalization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11400–11410, 2023. 1, 2

[8] 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 (CVPR)*, pages 11580–11590, June 2021. 1, 2

[9] 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 *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 5

[10] 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, 5

[11] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario March, and Victor Lempitsky. Domain-adversarial training of neural networks. *Journal of machine learning research*, 17(59):1–35, 2016. 1

[12] Tejas Gokhale, Rushil Anirudh, Jayaraman J Thiagarajan, Bhavya Kailkhura, Chitta Baral, and Yezhou Yang. Improving diversity with adversarially learned transformations for domain generalization. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 434–443, 2023. 1, 2

[13] M. 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, 2021. 5

[14] Shishuai Hu, Zehui Liao, Jianpeng Zhang, and Yong Xia. Domain and content adaptive convolution based multi-source domain generalization for medical image segmentation. *IEEE Transactions on Medical Imaging*, 42(1):233–244, 2022. 1

[15] Xun Huang and Serge Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *Proceedings of the IEEE international conference on computer vision*, pages 1501–1510, 2017. 2

[16] Ziyi Jiang, Liwen Zhang, Xiaoxuan Liang, and Zhenghan Chen. Cbda: Contrastive-based data augmentation for domain generalization. *IEEE Transactions on Computational Social Systems*, 2024. 1, 2

[17] 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*, volume 38, pages 2947–2955, 2024. 1, 2, 5, 6

[18] 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. 2, 5, 6, 7

[19] Pan Li, Da Li, Wei Li, Shaogang Gong, Yanwei Fu, and Timothy M Hospedales. A simple feature augmentation for domain generalization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8886–8895, 2021. 1, 2

[20] Xinhui Li, Mingjia Li, Xiaopeng Li, and Xiaojie Guo. Learning generalized knowledge from a single domain on urban-scene segmentation. *IEEE Transactions on Multimedia*, 25:7635–7646, 2022. 1

[21] Ya Li, Xinmei Tian, Mingming Gong, Yajing Liu, Tongliang Liu, Kun Zhang, and Dacheng Tao. Deep domain generalization via conditional invariant adversarial networks. In *Proceedings of the European conference on computer vision (ECCV)*, pages 624–639, 2018. 2

[22] Yumeng Li, Dan Zhang, Margret Keuper, and Anna Khoreva. Intra-source style augmentation for improved domain generalization. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 509–519, 2023. 1, 2

[23] Quande Liu, Cheng Chen, Jing Qin, Qi Dou, and Pheng-Ann Heng. Feddg: Federated domain generalization on medical image segmentation via episodic learning in continuous frequency space. In *Proceedings of the IEEE/CVF Conference*

on Computer Vision and Pattern Recognition (CVPR), pages 1013–1023, June 2021. 1

[24] 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, 5, 6, 7

[25] Yingnan Liu, Yingtian Zou, Rui Qiao, Fusheng Liu, Mong Li Lee, and Wynne Hsu. Cross-domain feature augmentation for domain generalization. *arXiv preprint arXiv:2405.08586*, 2024. 2

[26] Yawei Luo, Liang Zheng, Tao Guan, Junqing Yu, and Yi Yang. Taking a closer look at domain shift: Category-level adversaries for semantics consistent domain adaptation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2507–2516, 2019. 1

[27] Zhengxiong Luo, Yan Huang, Shang Li, Liang Wang, and Tieniu Tan. End-to-end alternating optimization for real-world blind super resolution. *International Journal of Computer Vision (IJCV)*, 2023. 1

[28] 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, 6

[29] Tomáš Mikolov, Wen-tau Yih, and Geoffrey Zweig. Linguistic regularities in continuous space word representations. In *Proceedings of the 2013 conference of the north american chapter of the association for computational linguistics: Human language technologies*, pages 746–751, 2013. 2

[30] Cheng Ouyang, Chen Chen, Surui Li, Zeju Li, Chen Qin, Wenjia Bai, and Daniel Rueckert. Causality-inspired single-source domain generalization for medical image segmentation. *IEEE Transactions on Medical Imaging*, 42(4):1095–1106, 2022. 1

[31] Lei Qi, Hongpeng Yang, Yinghuan Shi, and Xin Geng. Normaug: Normalization-guided augmentation for domain generalization. *IEEE Transactions on Image Processing*, 33:1419–1431, 2024. 1, 2

[32] 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. 2

[33] 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. 3

[34] Hshmat Sahak, Daniel Watson, Chitwan Saharia, and David Fleet. Denoising diffusion probabilistic models for robust image super-resolution in the wild. *arXiv preprint arXiv:2302.07864*, 2023. 1

[35] Christos Sakaridis, Dengxin Dai, and Luc Van Gool. Semantic foggy scene understanding with synthetic data. *International Journal of Computer Vision*, 126(9):973–992, Sep 2018. 5, 7, 8

[36] Jules Sanchez, Jean-Emmanuel Deschaud, and François Goulette. Domain generalization of 3d semantic segmentation in autonomous driving. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 18077–18087, October 2023. 2

[37] Yang Shu, Zhangjie Cao, Chenyu Wang, Jianmin Wang, and Mingsheng Long. Open domain generalization with domain-augmented meta-learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9624–9633, 2021. 1, 2

[38] 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, 4, 5, 6, 7

[39] Riccardo Volpi, Hongseok Namkoong, Ozan Sener, John C Duchi, Vittorio Murino, and Silvio Savarese. Generalizing to unseen domains via adversarial data augmentation. *Advances in neural information processing systems*, 31, 2018. 1, 2

[40] Mengzhu Wang, Yuehua Liu, Jianlong Yuan, Shanshan Wang, Zhibin Wang, and Wei Wang. Inter-class and inter-domain semantic augmentation for domain generalization. *IEEE Transactions on Image Processing*, 2024. 1, 2

[41] Mengzhu Wang, Jianlong Yuan, Qi Qian, Zhibin Wang, and Hao Li. Semantic data augmentation based distance metric learning for domain generalization. In *Proceedings of the 30th ACM international conference on multimedia*, pages 3214–3223, 2022. 1, 2

[42] Yue Wang, Lei Qi, Yinghuan Shi, and Yang Gao. Feature-based style randomization for domain generalization. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(8):5495–5509, 2022. 3

[43] Zijian Wang, Yadan Luo, Ruihong Qiu, Zi Huang, and Mahsa Baktashmotlagh. Learning to diversify for single domain generalization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 834–843, October 2021. 1, 2

[44] 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. 2, 5, 6, 7, 8

[45] 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 (ICCV)*, pages 9342–9351, October 2021. 5

[46] 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*, volume 38, pages 5958–5966, 2024. 2, 5, 6, 7

[47] Qi Xu, Liang Yao, Zhengkai Jiang, Guannan Jiang, Wenqing Chu, Wenhui Han, Wei Zhang, Chengjie Wang, and Ying Tai. Dirl: Domain-invariant representation learning for

generalizable semantic segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 2884–2892, 2022. 2

[48] 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 *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 5

[49] Qiuhao Zeng, Wei Wang, Fan Zhou, Charles Ling, and Boyu Wang. Foresee what you will learn: data augmentation for domain generalization in non-stationary environment. In *Proceedings of the AAAI conference on artificial intelligence*, volume 37, pages 11147–11155, 2023. 2

[50] Hao Zhang, Feng Li, Shilong Liu, Lei Zhang, Hang Su, Jun Zhu, Lionel M Ni, and Heung-Yeung Shum. Dino: Detr with improved denoising anchor boxes for end-to-end object detection. *arXiv preprint arXiv:2203.03605*, 2022. 3

[51] Kai Zhang, Jingyun Liang, Luc Van Gool, and Radu Timofte. Designing a practical degradation model for deep blind image super-resolution. In *IEEE International Conference on Computer Vision*, pages 4791–4800, 2021. 1

[52] Guangtao Zheng, Mengdi Huai, and Aidong Zhang. Advst: Revisiting data augmentations for single domain generalization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 21832–21840, 2024. 1, 2

[53] Zhun Zhong, Yuyang Zhao, Gim Hee Lee, and Nicu Sebe. Adversarial style augmentation for domain generalized urban-scene segmentation. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 338–350. Curran Associates, Inc., 2022. 1

[54] Kaiyang Zhou, Yongxin Yang, Yu Qiao, and Tao Xiang. Domain generalization with mixstyle. In *ICLR*, 2021. 2