This CVPR paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

CLIP the Gap: A Single Domain Generalization Approach for Object Detection

Vidit Vidit[1] Martin Engilberge[1] Mathieu Salzmann[1,2]
CVLab, EPFL[1], ClearSpace SA[2]
firstname.lastname@epfl.ch

Abstract

Single Domain Generalization (SDG) tackles the problem of training a model on a single source domain so that it generalizes to any unseen target domain. While this has been well studied for image classification, the literature on SDG object detection remains almost non-existent. To address the challenges of simultaneously learning robust object localization and representation, we propose to leverage a pre-trained vision-language model to introduce semantic domain concepts via textual prompts. We achieve this via a semantic augmentation strategy acting on the features extracted by the detector backbone, as well as a text-based classification loss. Our experiments evidence the benefits of our approach, outperforming by 10% the only existing SDG object detection method, Single-DGOD [52], on their own diverse weather-driving benchmark.

1. Introduction

As for most machine learning models, the performance of object detectors degrades when the test data distribution deviates from the training data one. Domain adaptation techniques [3, 5, 8, 33, 44, 46] try to alleviate this problem by learning domain invariant features between a source and a known target domain. In practice, however, it is not always possible to obtain target data, even unlabeled, precluding the use of such techniques. Domain generalization tackles this by seeking to learn representations that generalize to any target domain. While early approaches [1, 10, 28, 29, 31, 50, 60] focused on the scenario where multiple source domains are available during training, many recent methods tackle the more challenging, yet more realistic, case of Single Domain Generalization (SDG), aiming to learn to generalize from a single source dataset. While this has been well studied for image classification [14, 38, 48, 51, 59], it remains a nascent topic in object detection. To the best of our knowledge, a single existing approach, Single-DGOD [52], uses disentanglement and self-distillation [25] to learn domain-invariant features.

In this paper, we introduce a fundamentally different approach to SDG for object detection. To this end, we build on two observations: (i) Unsupervised/self-supervised pretraining facilitates the transfer of a model to new tasks [2,

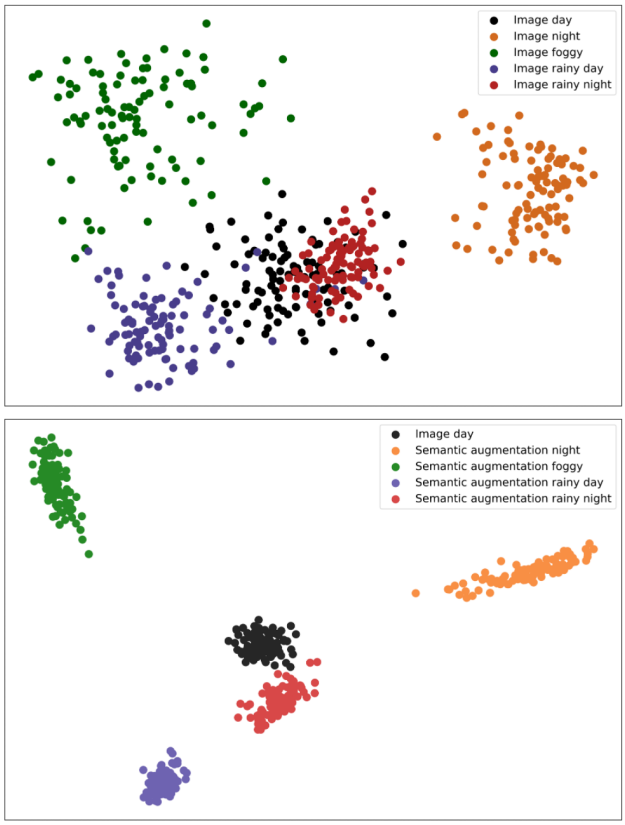

Figure 1. **Semantic Augmentation:** We compare the PCA projections of CLIP [39] image embeddings obtained in two different manners: (Top) The embeddings were directly obtained from the real images from 5 domains corresponding to different weather conditions. (Bottom) The embeddings were obtained from the *day* images only and modified with our semantic augmentation strategy based on text prompts to reflect the other 4 domains. Note that the relative positions of the clusters in the bottom plot resembles that of the top one, showing that our augmentations let us generalize to different target domains. The principal components used are the same for both the figures.

4, 20]; (ii) Exploiting language supervision to train vision models allows them to generalize more easily to new categories and concepts [9, 39]. Inspired by this, we therefore propose to leverage a self-supervised vision-language model, CLIP [39], to guide the training of an object detector so that it generalizes to unseen target domains. Since the visual CLIP representation has been jointly learned with the textual one, we transfer text-based domain variations to the image representation during training, thus increasing the diversity of the source data.

Specifically, we define textual prompts describing potential target domain concepts, such as weather and daytime variations for road scene understanding, and use these prompts to perform semantic augmentations of the images. These augmentations, however, are done in feature space, not in image space, which is facilitated by the joint image-text CLIP latent space. This is illustrated in Fig. 1, which shows that, even though we did not use any target data for semantic augmentation, the resulting augmented embeddings reflect the distributions of the true image embeddings from different target domains.

We show the effectiveness of our method on the SDG driving dataset of [52], which reflects a practical scenario where the training (source) images were captured on a clear day whereas the test (target) ones were acquired in rainy, foggy, night, and dusk conditions. Our experiments demonstrate the benefits of our approach over the Single-DGOD [52] one.

To summarize our contributions, we employ a vision-language model to improve the generalizability of an object detector; during training, we introduce domain concepts via text-prompts to augment the diversity of the learned image features and make them more robust to an unseen target domain. This enables us to achieve state-of-the-art results on the diverse weather SDG driving benchmark of [52]. Our implementation can be accessed through the following url: https://github.com/vidit09/domaingen.

2. Related Work

Domain Adaptation for Object Detection. Domain adaptation methods seek to align the source domain distribution to a particular target domain. To bridge the global and instance-level domain gaps, [3, 5, 44, 46] learn feature alignment via [16] adversarial training; [61] and [49] utilize category-level centroids and attention maps, respectively, to better align instances in the two domains; [8, 33] generate pseudo-labels in the target domain and use them for target-aware training. Domain adaptation, however, assumes that images from the target domain are available during training. In contrast, domain generalization aims to learn models that generalize to domains that were not seen at all during training. Below, we focus on the domain generalization methods that, as us, use a single source domain to do so.

Single Domain Generalization (SDG). Several image classification works [14, 38, 48, 51, 59] have proposed strategies to improve the performance on *unseen* domains while training on a single source domain. In particular, [38, 48, 51] introduce data augmentation strategies where diverse input images are generated via adversarial training; [14, 59] propose normalization techniques to adapt the feature distribution to unseen domains. While SDG has been reasonably well studied for image classification, the case of object detection remains largely unexplored, and poses additional challenges related to the need to further localize the objects of interest. This was recently tackled by Single-DGOD [52] with an approach relying on learning domain-specific and domain-invariant features. Specifically, this was achieved by exploiting contrastive learning to disentangle the features and self-distillation [25] to further improve the network's generalizability. Here, we introduce a fundamentally different approach that leverages the CLIP [39] pre-trained model and semantically augments the data using textual prompts. As will be shown by our results, our method outperforms the state-of-the-art Single-DGOD [52].

Vision-Language Models. Jointly learning a representation of images and text has been studied in many works [9, 11, 13, 15, 27, 30, 39, 58]. They use image-text pairs to train visual-semantic embeddings which can be used not only for image classification, captioning or retrieval but also for zero-shot prediction on unseen labels. VirTex [9] relies on image-caption-based pre-training to learn a rich visual embedding from a small amount of data. CLIP [39] proposes a scalable contrastive pre-training method for joint text and image feature learning. CLIP leverages a corpus of 400 million image-text pairs and a large language model [40] to learn a joint embedding space, which was shown to have superior zero-shot learning ability on classification tasks. The image-text-based training is also useful for Open Vocabulary Detection (OVD) [56], where the objects are detected using arbitrary textual descriptions. To address this task, [56] train their own visual-semantic representation, whereas [17, 42] employ CLIP embeddings. Recently, [32, 57] introduced a phrase-grounding-based pre-training for better OVD and zero-shot object detection. In contrast to these works, whose objective is to generalize to novel *categories or objects*, we seek to generalize to new *domains* depicting the same object categories as the source one.

3. Method

Let us now introduce our approach to exploiting a vision-language model for single-domain generalization in object detection. Below, we first present our semantic augmentation strategy aiming to facilitate generalization to new domains. We then describe the architecture and training strategy for our object detector.

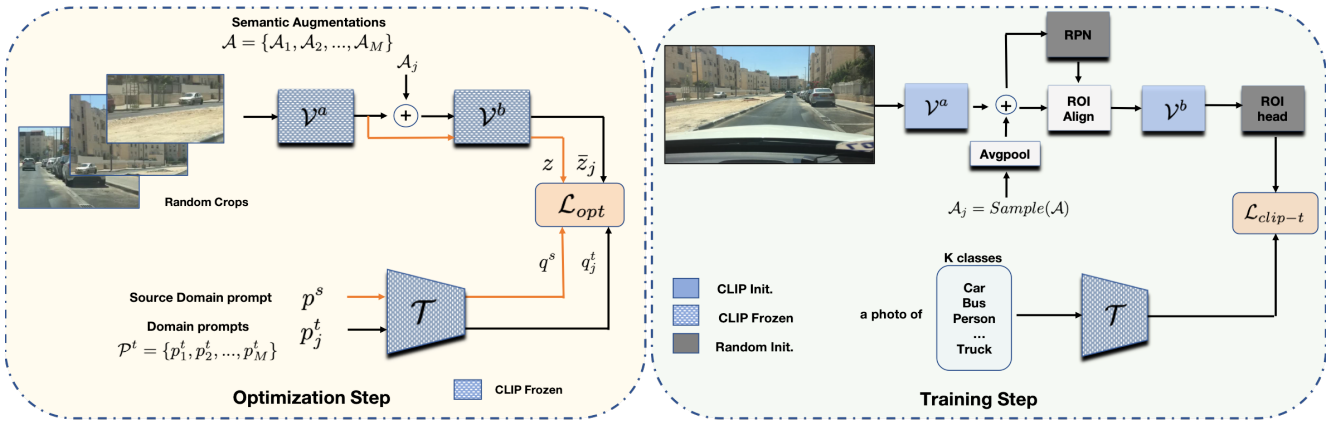

Figure 2. **Our Approach: (Left)** We first estimate a set of semantic augmentations $\mathcal{A}$ using a set of textual domain prompts $\{\mathcal{P}^t, p^s\}$ and source domain images. The goal of these semantic augmentations is to translate source domain image embeddings to the domain specified by the prompts. We can do this because of the CLIP's joint embedding space and its ability to encode semantic relationships via algebraic operations. $\mathcal{L}_{opt}$ is minimized w.r.t $\mathcal{A}$ over random image crops of the same size as CLIP [39]. **(Right)** The optimized semantic augmentations are used to train our modified detector which minimizes a text-based classification loss $\mathcal{L}_{clip-t}$. Here, we train with the full image and add a randomly sampled $\mathcal{A}_j$ after average pooling. This pooling operation allows us to use $\mathcal{A}$ on extracted feature maps of the arbitrary-sized image. We initialize the detector with the pre-trained CLIP [39] $\mathcal{V}$ and $\mathcal{T}$ encoders to leverage their general representations.

3.1. Semantic Augmentation

In SDG, we have access to images from only a single domain. To enable generalization, we seek to learn object representations that are robust to domain shifts. Here, we do so by introducing such shifts while training the model on the source data. Specifically, we exploit CLIP's joint representation to estimate shifts in the visual domain using textual prompts, as illustrated in Fig. 1. This corresponds to the optimization step shown in the left portion of Fig. 2.

Formally, let $\mathcal{T}$ denote CLIP's text encoder and $\mathcal{V}$ its image one. For reasons that will become clear later, we further split $\mathcal{V}$ into a feature extractor $\mathcal{V}^a$ and a projector to the embedding space $\mathcal{V}^b$. The CLIP [39] model is trained to bring image features closer to their textual captions. In essence, this means that, for an image $\mathcal{I}$ and a corresponding prompt p, it seeks to minimize the distance between $\mathcal{V}^b(\mathcal{V}^a(I))$ and $\mathcal{T}(p)$.

A useful property of the text embedding space is that algebraic operations can be used to estimate semantically related concepts. Word2Vec [34] had demonstrated such a learned relationship (e.g. *king-man+woman* approaches the word representation of *queen*). Such a relationship exists with CLIP embeddings as well [41].

To exploit this for SDG, we define a generic textual prompt p^s related to the source domain, such as `An image taken during the day`, and a set of prompts $\mathcal{P}^t = \{p^t_j\}^M_1$ encompassing variations that can be expected to occur in different target domains, e.g, describing different weather conditions or times of the day. Our objective then is to define augmentations $\{\mathcal{A}_j\}$ of the features extracted from a source image such that the shift incurred by $\mathcal{A}_j$ corresponds to the semantic difference between p^s and p^t_j.

To achieve this, we first compute the embeddings $q^s = \mathcal{T}(p^s)$ and $q^t_j = \mathcal{T}(p^t_j)$ of the textual prompt. We then take multiple random crops from a source image. For each such crop $\mathcal{I}_{crop}$, we create a target image embedding

$$z^*_j = z + \frac{q^t_j - q^s}{\|q^t_j - q^s\|_2}\,, \tag{1}$$

where $z = \mathcal{V}(\mathcal{I}_{crop})$. We then search for an augmentation $\mathcal{A}_j \in \mathbb{R}^{H\times W\times C}$ such that

$$\bar{z}_j = \mathcal{V}^b(\mathcal{V}^a(\mathcal{I}_{crop}) + \mathcal{A}_j) \tag{2}$$

is as similar as possible to z^*_j, which we measure with the cosine similarity. Ultimately, we estimate the augmentations $\{\mathcal{A}_j\}^M_1$ through an optimization process using only source domain images. Specifically, we minimize the loss function

$$\mathcal{L}_{opt} = \sum_{\mathcal{I}_{crop}} \sum_j \mathcal{D}(z^*_j, \bar{z}_j) + \|\bar{z}_j - z\|_1\,, \tag{3}$$

where

$$\mathcal{D}(a, b) = 1 - \frac{a - b}{\|a - b\|_2} \tag{4}$$

is the cosine distance. The loss also includes an l_1 regularizer that prevents the embeddings from deviating too far from their initial values, so as to preserve the image content.

As the objective is to estimate the meaningful feature augmentation while preserving the original CLIP pretraining, we keep the image crop size the same as the original CLIP training. Note that the optimization of the augmentations is done once in an offline stage, and we then use the resulting augmentations to train our detector.

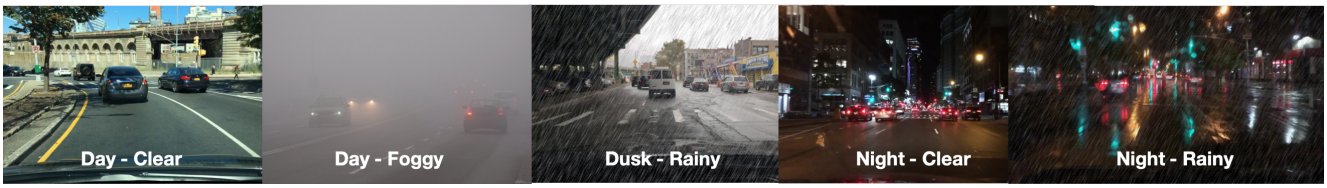

Figure 3. **Diverse Weather Dataset [52]:** Day-Clear acts as our source domain while the other weather condition are our target domains. In these domains, the objects' appearance drastically changes from the Day-Clear scenario. As we do not utilize any target domain images, learning generalizable features on source images is crucial for the SDG task.

3.2. Architecture

Let us now describe our detector architecture. As shown in the right portion of Fig. 2, it follows a standard Faster-RCNN [43] structure but departs from it in two ways. First, to exploit the augmentations optimized as discussed in the previous section, we initialize the blocks before and after the ROI align one with the corresponding $\mathcal{V}^a$ and $\mathcal{V}^b$ modules of the ResNet-based trained CLIP model. Second, to further leverage the vision-language model, we incorporate a text-based classifier in our model's head. Note that, in contrast to OVD [17, 42] where a text-based classifier is used to handle novel categories, we employ it to keep the image features close to the pre-trained joint embedding space.

Specifically, we define textual prompts that represent the individual categories we seek to detect, and extract corresponding embeddings $\mathcal{Q} \in \mathbb{R}^{(K+1)\times D_{clip}}$, for K categories and the background class, using the text encoder $\mathcal{T}$. For a candidate image region r proposed by the Region Proposal Network(RPN) [43], we then compute the cosine similarities between the text embeddings $\mathcal{Q}$ and the features $\mathcal{F}_r \in \mathbb{R}^{D_{clip}}$ obtained by projection to the embedding space using $\mathcal{V}^b$ after ROI-Align [21] and the text embeddings $\mathcal{Q}$. These cosine similarities, $sim(\mathcal{F}_r, \mathcal{Q}) \in \mathbb{R}^{K+1}$, act as logits to the softmax based cross-entropy loss

$$\mathcal{L}_{clip\text{-}t} = \sum_r \mathcal{L}_{CE}\left(\frac{e^{\gamma \cdot sim(\mathcal{F}_r, \mathcal{Q}_k)}}{\sum_{k=0}^{K} e^{\gamma \cdot sim(\mathcal{F}_r, \mathcal{Q}_k)}}\right). \quad (5)$$

where γ is a temperature factor. Similarly to [39], we formulate prompts of the form `a photo of a {category name}` to obtain our text embeddings.

3.3. Training with Augmentation

Following the standard detector training [43], we use the full image as our input. This subsequently increases the output feature map size of $\mathcal{V}^a$, hence we use average pooling operation and obtain channel-wise augmentations which can work for arbitrary-sized feature maps. The training of our modified object detector with the semantic augmentations is as follows, first, we randomly sample an augmentation $\mathcal{A}_j$ from the full set and collapse its spatial dimension using average pooling. We then add the resulting vector to every element in the feature map extracted by $\mathcal{V}^a$. In practice, we apply augmentations to a batch with a probability θ.

The detector is then trained with the loss

$$\mathcal{L}_{det} = \mathcal{L}_{rpn} + \mathcal{L}_{reg} + \mathcal{L}_{clip\text{-}t}\ , \quad (6)$$

which combines the $\mathcal{L}_{clip\text{-}t}$ loss of Eq. (5) with the standard RPN and regression losses [43]. During inference, we use the detector without any augmentation of the feature maps.

4. Experiments

4.1. Experimental setup

Datasets. To evaluate our model, we use the same datasets as [52]. They include five sets, each containing images with different weather conditions: daytime sunny, night clear, dusk rainy, night rainy, and daytime foggy. The images have been selected from three primary datasets, Berkeley Deep Drive 100K (BBD-100K) [55], Cityscapes [7] and Adverse-Weather [19]. Additionally, rainy images are rendered by [53], and some of the foggy images are synthetically generated from [45]. Our model is trained on the daytime sunny scenes, consisting of 19,395 training images, the remaining 8,313 daytime sunny images are used for validation and model selection. The four other weather conditions are only used during testing. They consist of 26,158 images of clear night scenes, 3501 images of rainy scenes at dusk, 2494 images of rainy scenes at night, and 3775 images of foggy scenes during daytime. All the datasets contain bounding box annotations for the objects *bus, bike, car, motorbike, person, rider* and *truck*. Fig. 3 shows examples from this dataset.

Metric. In all our experiments, we use the Mean Average Precision (mAP) as our metric. Specifically, following [52], we report the mAP@0.5, which considers a prediction as a true positive if it matches the ground-truth label and has an intersection over union (IOU) score of more than 0.5 with the ground-truth bounding box.

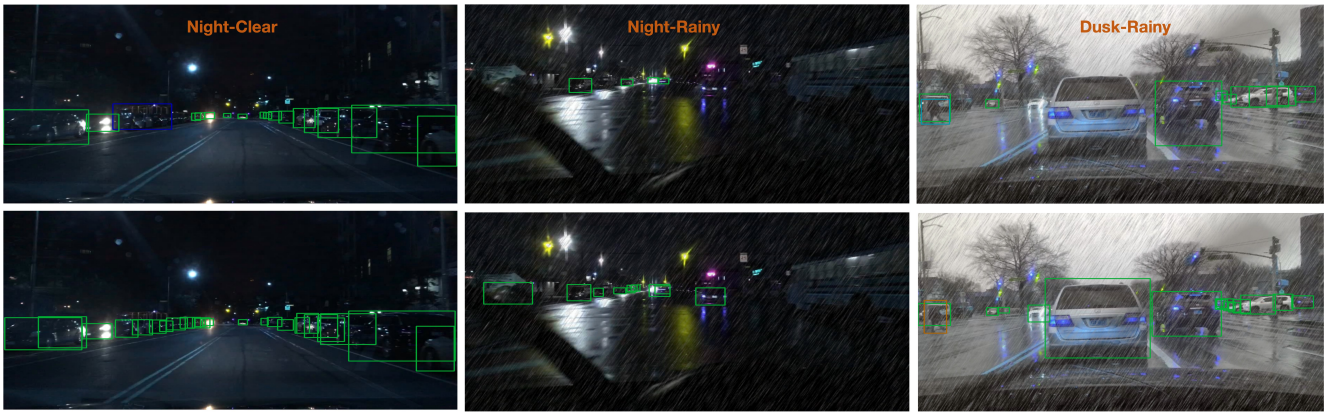

Figure 4. **Qualitative Results.** We visualize the predictions of the detectors trained only with day-clear images. (Top) FasterRCNN [43] predictions. (Bottom) The predictions with our approach. Night-Clear and Night-Rainy contain scenes that are taken under low light conditions. Due to this, the appearance of the object is obscure and deviates from the daytime case. FasterRCNN fails to detect most of the objects. As shown in the Night-Clear, it misclassifies a car to bus. By contrast, we can still detect car under such a big shift. For Dusk-Rainy scenes, the rain pattern on the windscreen and the wet ground causes an appearance shift. As shown FasterRCNN fails to detect several cars and misclassifies person on the bottom-left.

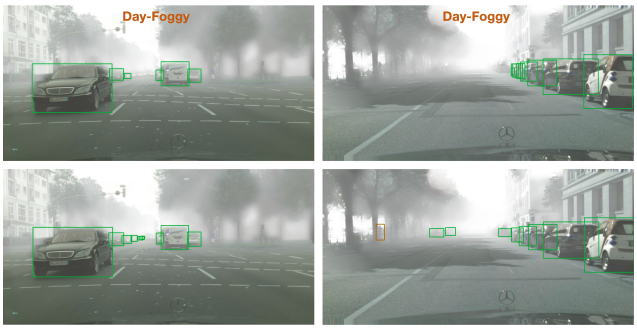

Figure 5. **Qualitative Results.** In the foggy scenes, the objects further away w.r.t the camera are more obscure than the near ones. Due to this FasterRCNN (Top) struggles to detect them. car and person missed by FasterRCNN are successfully recovered by our approach (Bottom).

4.2. Implementation Details

We use the Detectron2 [54] implementation of FasterRCNN with a ResNet101 [22] backbone. We initialize the detector with CLIP [39] pre-trained weights, where ResNet convolution blocks 1-3 act as $\mathcal{V}^a$, and block-4 along with the CLIP attention pooling act as $\mathcal{V}^b$. This follows from the standard FasterRCNN implementation with ResNet backbone. We set γ as 100, similar to CLIP [39].

Optimization Step. As the benchmark dataset evaluates the method on different weather conditions, we curated a list of domain prompts $\mathcal{P}^t$ matching the concept *weather*. To this end, we take all the *hyponyms* of the term *weather* from WordNet [47] and generate their text embeddings using the CLIP text encoder $\mathcal{T}$. We prune away the words whose cosine similarity with the term *weather* is lower than 0.5. Additionally, we filter out the words that are not in the top 10k frequent words in GloVe wordlist [37]. After combining the synonyms, we get to a list of six words: *snow, fog, cloudy, rain, stormy, sunshine*. We remove *sunshine* as it corresponds to our source domain concept. Furthermore, we consider three times of the day: *day, night, evening*. This lets us generate $M = 15$ prompts using the template `an image taken on a {weather} {time of the day}`. We use `an image taken during the day` as the source domain prompt p^s. We provide more details in our supplementary material.

To optimize the augmentations with these prompts, we generated random crops from the source images and resized them to 224×224 pixels. The resulting output feature map of $\mathcal{V}^a$ and $\mathcal{A}_j$ are in $\mathbb{R}^{14\times14\times1024}$. We initialize $\mathcal{A}_j \ \forall \ 1 \geq j \geq M$ with zeros and train it using the Adam [26] optimizer while keeping the CLIP encoder, $\mathcal{V}$ and $\mathcal{T}$, frozen. Optimization was done for 1000 iterations with a learning rate of 0.01.

Detector Training with Augmentation. When training the detector, the input image is resized to 600×1067 and $\mathcal{V}$ and $\mathcal{T}$ are initialized with CLIP pre-trained weights. While $\mathcal{T}$ is kept frozen during the training, the ResNet blocks 3-4 and attention pooling of $\mathcal{V}$, along with the other FasterRCNN learnable blocks, are trained with Stochastic Gradient Descent (SGD) for 100k iterations. We train with a learning rate of $1e^{-3}$, scaled down by a factor of 0.1 after 40k iterations. We use a batch size of 4 and apply $\mathcal{A}_j$ to the features with probability $\theta = 0.5$. We also use random horizontal flipping augmentation as in Single-DGOD [52].

Method	mAP				
	Day Clear	Night Clear	Dusk Rainy	Night Rainy	Day Foggy
FR [43]	48.1	34.4	26.0	12.4	32.0
SW [36]	50.6	33.4	26.3	13.7	30.8
IBN-Net [35]	49.7	32.1	26.1	14.3	29.6
IterNorm [23]	43.9	29.6	22.8	12.6	28.4
ISW [6]	51.3	33.2	25.9	14.1	31.8
S-DGOD [52]	**56.1**	36.6	28.2	16.6	33.5
Ours	51.3	**36.9**	**32.3**	**18.7**	**38.5**

Table 1. **Single domain generalization results.** We show consistent improvements across all the target domains. S-DGOD boosts the source domain results, but at the cost of reduced generalization ability. By contrast, our approach is robust to domain changes. The numbers for S-DGOD, SW, IBN-Net, IterNorm, ISW are taken from [52].

D_{clip} is set to 512 as in [39] and background class is initialized by zeros in $\mathcal{Q}$. All of our training was done on a single NVIDIA A100 GPU.

4.3. Comparison with the State of the Art

We compare our method trained with semantic augmentations against the state-of-the-art Single-DGOD [52]. Similar to them, we also show comparisons with feature normalization methods, SW [36], IBN-Net [35], IterNorm [23], and ISW [6]. These methods improve network generalization by using better feature normalization. We additionally report the performance of FasterRCNN (FR) initialized with ImageNet pre-trained weights. For the SDG task, we evaluate the generalization performance on unseen target domains, hence we compare the mAP scores on the out-of-domain datasets: day-foggy, night-rainy, dusk-rainy, and night-clear. Following Single-DGOD, we adopt training-domain validation strategy [18] for the model selection.

Our approach of combining CLIP pre-training and semantic augmentation outperforms the baselines on all of the target domains. Tab. 1 shows a consistent improvement in all domains with close to 15% improvement on day-foggy and dusk-rainy compared to Single-DGOD. In the challenging scenario with Night conditions, we improve by 12.6% on night-rainy while being comparable with Single-DGOD on night-clear. On the source domain, both our method and Single-DGOD are better than the FR baseline. However, while Single-DGOD gains improvement at the cost of losing out for domain generalization, we improve on both the source and target domains. The failure of feature normalization baselines suggests a large domain gap between the source and target domains. Fig. 4 and Fig. 5 provide a qualitative results on different weather-datasets.

In the remainder of this section, we discuss the per-class results on the individual target domains.

Method	AP							mAP
	Bus	Bike	Car	Motor	Person	Rider	Truck	All
FR [43]	28.1	29.7	49.7	26.3	33.2	35.5	21.5	32.0
S-DGOD [52]	32.9	28.0	48.8	29.8	32.5	38.2	24.1	33.5
Ours	**36.1**	**34.3**	**58.0**	**33.1**	**39.0**	**43.9**	**25.1**	**38.5**

Table 2. **Per-class results on Daytime Clear to Day Foggy.** Our method consistently performs better on all categories for the difficult foggy domain. This shows that CLIP initialization and our semantic augmentations improve the detector's generalizability.

Method	AP							mAP
	Bus	Bike	Car	Motor	Person	Rider	Truck	All
FR [43]	28.5	20.3	58.2	6.5	23.4	11.3	33.9	26.0
S-DGOD [52]	37.1	19.6	50.9	13.4	19.7	16.3	40.7	28.2
Ours	**37.8**	**22.8**	**60.7**	**16.8**	**26.8**	**18.7**	**42.4**	**32.3**

Table 3. **Per-class results on Daytime Clear to Dusk Rainy.** Our approach generalizes to rainy road conditions along with the low light conditions of the dusk hours. The *car* category sees the biggest improvement, but we nonetheless also boost the performance of all the other classes.

Daytime Clear to Day Foggy. The object appearance drastically changes in the foggy images compared to the day-clear scenario. As shown in Tab. 2, our method brings in a large improvement for the *car*, *person*, and *bike* categories, while still being consistently better than Single-DGOD and FR on the others.

Daytime Clear to Dusk Rainy. Dusk Rainy scenes reflect a low light condition and along with the rainy pattern. The image distribution is thus further away from the daytime clear images. As shown in Tab. 3, our method improves the AP of each class, with the biggest improvement in the *car* and *person* categories. Since we leverage CLIP pre-training and bring in concepts such as rain/cloudy/stormy and evening/night hours through our semantic augmentation, the learned detector generalizes better.

Method	AP							mAP
	Bus	Bike	Car	Motor	Person	Rider	Truck	All
FR [43]	34.7	32.0	56.6	13.6	37.4	27.6	38.6	34.4
S-DGOD [52]	**40.6**	**35.1**	50.7	**19.7**	34.7	**32.1**	**43.4**	36.6
Ours	37.7	34.3	**58.0**	19.2	**37.6**	28.5	42.9	**36.9**

Table 4. **Per-class results on Daytime Clear to Night Clear.** While being comparable to S-DGOD on most of the categories, we improve on *car* and *person*.

Method	AP							mAP
	Bus	Bike	Car	Motor	Person	Rider	Truck	All
FR [43]	16.8	6.9	26.3	0.6	11.6	9.4	15.4	12.4
S-DGOD [52]	24.4	11.6	29.5	**9.8**	10.5	**11.4**	19.2	16.6
Ours	**28.6**	**12.1**	**36.1**	9.2	**12.3**	9.6	**22.9**	**18.7**

Table 5. **Per-class results on Daytime Clear to Night Rainy.** This dataset presents the most challenging scenario, where the low light and rainy conditions obscure the objects. We still perform better than the baseline on most of the categories.

Daytime Clear to Night Clear. The Night Clear dataset shows a challenging night driving scene under severe low-light conditions. In Tab. 4, we show that while being comparable to Single-DGOD, we bring in a larger improvement in the *car* and *person* categories. Night scenes are particularly challenging as the low light condition leads to more confusion among visually closer categories such as *bus* and *truck*.

Daytime Clear to Night Rainy. This is the most challenging scenario where dark night conditions are exacerbated by patterns occurring due to rain. Tab. 5 shows consistent improvement by our approach for most of the classes. The *car* class sees the biggest improvement with an increase in AP of more than 22% compared to Single-DGOD. The lower performance of the class *rider* can be attributed to an increase in the confusion between the visually similar *person* and *rider* classes under adverse conditions.

4.4. Ablation Study

To understand how each element of the proposed method contributes to the overall performance, we conduct an ablation study. We test five individual components of our model. Specifically, we remove semantic augmentation, replace CLIP attention pooling in $\mathcal{V}^b$ with average pooling, replace $\mathcal{L}_{clip\text{-}t}$ with the FasterRCNN classification loss, and change the weight initialization from the CLIP model to an ImageNet classification model. Removing those five components turns our model back into the standard Faster-RCNN. The ablation study results are provided in Tab. 6 and discussed below.

CLIP initialization. When the FasterRCNN backbone $\mathcal{V}$ is initialized with CLIP pre-trained weights, the model performance consistently increases both in the in-domain and out-of-domain scenarios, as shown in the second row of Tab. 6. This setting itself already outperforms Single-DGOD (penultimate row of Tab. 1). This goes to show that, for the generalization task, model weight initialization plays a crucial role. We further improve this performance with semantic augmentations.

Attention pooling and $\mathcal{L}_{clip\text{-}t}$. Next we test the impact of the text-embedding-based loss $\mathcal{L}_{clip\text{-}t}$ for classification. As visible in the third row of Tab. 6, when combined with CLIP initialization, it improves the generalization performance for the rainy scenarios, but degrades it for the other ones. Replacing average pooling in $\mathcal{V}^b$ with CLIP attention pooling helps to mitigate the detrimental effect of $\mathcal{L}_{clip\text{-}t}$ and exhibits consistent improvement on all datasets.

Semantic augmentation. Finally, adding semantic augmentation gives us the best results, as shown in the last row of Tab. 6. Exposing the visual encoder $\mathcal{V}$ to targeted semantic augmentations helps the overall model to better generalize when exposed to new domains sharing similarity with the augmentations.

4.5. Additional Analyses

Study of semantic augmentation. Our proposed method involves translating feature maps by semantic augmentations learned using plausible domain prompts. To further study the utility of our approach, we replace the augmentation strategy in our training pipeline with (a) **no-aug**: no augmentation; (b) **random**: $\mathcal{A}$ is initialized with a normal distribution; (c) **clip-random**: we define $\mathcal{P}^t$ with concepts that are not specific to *weather*. We generate prompts with a template `an image of {word}`, where the words are *desert*, *ocean*, *forest*, and *mountain*. Tab. 7 illustrates the importance of the semantics in our augmentation strategy. The **random** augmentation performs worse than the **no-aug** strategy. **clip-random** is comparable to **no-aug** and doesn't show any consistent trend but is mostly better than **random**. Our semantic augmentation strategy provides a consistent improvement over **no-aug** because the translations are performed with prompts from the relevant *weather* concept.

In supp. material Sec. A.2, we present additional generalization results with training on the natural images, Pascal-VOC [12] and testing on styled images, Comics, and Watercolor datasets [24].

Model Component				mAP				
				Source	Target			
CLIP init	$\mathcal{L}_{clip\text{-}t}$	Attn. Pool	Sem. Aug	Day Clear	Night Clear	Dusk Rainy	Night Rainy	Day Foggy
				48.1	34.4	26.0	12.4	32.0
✓				51.2	**37.0**	31.0	15.7	37.5
✓	✓			50.7	36.0	31.3	16.3	36.9
✓	✓	✓		51.0	35.9	31.3	16.7	37.7
✓	✓	✓	✓	51.3	36.9	**32.3**	**18.7**	**38.5**

Table 6. **Ablation study.** We study the influence of five different components of our approach: the backbone weight initialization strategy, the classification loss, the attention pooling, and the semantic augmentation. When those five components are removed (first row of the table) the model is equivalent to the standard FasterRCNN. Initializing the detector with CLIP weights (second row) largely improves the generalization performance; on its own it already outperforms Single-DGOD (penultimate row of Tab. 1) on most of the datasets, hence suggesting that CLIP has better generalizability than ImageNet pre-trained weights. Combining this with the text embedding-based loss $\mathcal{L}_{clip\text{-}t}$ (third row) improves the results on the challenging scenarios of dusk rainy and night rainy, but has a detrimental effect for the other weather conditions. Adding attention pooling to the architecture (fourth row) helps to mitigate these detrimental effects as it brings the visual features closer to the joint embedding space. Finally, the best results are obtained when the semantic augmentation is added (last row), greatly helping with adverse weather, rainy and foggy, scenarios.

Aug. Type	mAP				
	Day Clear	Night Clear	Dusk Rainy	Night Rainy	Day Foggy
no-aug.	51.0	35.9	31.3	16.7	37.7
random	51.2	36.0	30.4	15.3	37.3
clip-random	51.5	36.4	30.2	15.9	37.9
Ours w/ seg.aug	51.3	**36.9**	**32.3**	**18.7**	**38.5**

Table 7. **Semantic Augmentation.** Our semantic augmentation consistently outperforms other augmentation strategies. While *random* augmentations are worse than *no-aug.*, *clip-random* is comparable to *no-aug.*. Only when we give relevant prompts, there is a consistent improvement across datasets.

5. Limitations

Our method augments visual features using textual prompts. To generate these prompts, it is assumed that some information about the domain gap is known. In our experiments, we assumed that the domain gap was due to changes in weather and daytime conditions. In practice, we only used the word *weather* and *time of the day* to derive all the prompts used in our augmentation; nonetheless, some extra information was used. In most applications, however, the domain gap can be known in advance, and providing a few keywords characterizing it shouldn't be an issue. In the rare cases where no information can be known, our approach still has the potential to be used by using multiple broad concept keywords such as weather, ambiance, or location.

6. Conclusion

We have proposed an approach to improving the generalization of object detectors on *unseen* target domains. Our approach fundamentally departs from existing method by leveraging a pre-trained vision-language model, CLIP, to help the detector to generalize. Specifically, we have exploited textual prompts to develop a semantic augmentation strategy that alters image embeddings so that they reflect potential target domains, and to design a text-based image classifier. We have shown that our approach outperforms the state of the art on four adverse-weather target datasets. In future work, we plan to extend our approach to learning the prompts to further improve generalization.

Acknowledgment: This work was funded in part by the Swiss National Science Foundation and the Swiss Innovation Agency (Innosuisse) via the BRIDGE Discovery grant 40B2-0 194729.

References

[1] Yogesh Balaji, Swami Sankaranarayanan, and Rama Chellappa. Metareg: Towards domain generalization using meta-regularization. *Advances in neural information processing systems*, 31, 2018. 1

[2] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9650–9660, 2021. 1

[3] Chaoqi Chen, Zebiao Zheng, Xinghao Ding, Yue Huang, and Qi Dou. Harmonizing transferability and discriminability for adapting object detectors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8869–8878, 2020. 1, 2

[4] Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15750–15758, 2021. 1

[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, 2

[6] 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

[7] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3213–3223, 2016. 4

[8] Jinhong Deng, Wen Li, Yuhua Chen, and Lixin Duan. Unbiased mean teacher for cross-domain object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4091–4101, 2021. 1, 2

[9] Karan Desai and Justin Johnson. Virtex: Learning visual representations from textual annotations. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11162–11173, 2021. 2

[10] Qi Dou, Daniel Coelho de Castro, Konstantinos Kamnitsas, and Ben Glocker. Domain generalization via model-agnostic learning of semantic features. *Advances in Neural Information Processing Systems*, 32, 2019. 1

[11] Martin Engilberge, Louis Chevallier, Patrick Pérez, and Matthieu Cord. Finding beans in burgers: Deep semantic-visual embedding with localization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3984–3993, 2018. 2

[12] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge 2012 (VOC2012) Results. http://www.pascal-network.org/challenges/VOC/voc2012/workshop/index.html. 7

[13] Fartash Faghri, David J Fleet, Jamie Ryan Kiros, and Sanja Fidler. Vse++: Improving visual-semantic embeddings with hard negatives. *arXiv preprint arXiv:1707.05612*, 2017. 2

[14] Xinjie Fan, Qifei Wang, Junjie Ke, Feng Yang, Boqing Gong, and Mingyuan Zhou. Adversarially adaptive normalization for single domain generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8208–8217, 2021. 1, 2

[15] Andrea Frome, Greg S Corrado, Jon Shlens, Samy Bengio, Jeff Dean, Marc'Aurelio Ranzato, and Tomas Mikolov. Devise: A deep visual-semantic embedding model. *Advances in neural information processing systems*, 26, 2013. 2

[16] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. Domain-adversarial training of neural networks. *The journal of machine learning research*, 17(1):2096–2030, 2016. 2

[17] Xiuye Gu, Tsung-Yi Lin, Weicheng Kuo, and Yin Cui. Open-vocabulary object detection via vision and language knowledge distillation. *arXiv preprint arXiv:2104.13921*, 2021. 2, 4

[18] Ishaan Gulrajani and David Lopez-Paz. In search of lost domain generalization. *arXiv preprint arXiv:2007.01434*, 2020. 6

[19] 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. 4

[20] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9729–9738, 2020. 1

[21] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017. 4

[22] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 5

[23] 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

[24] Naoto Inoue, Ryosuke Furuta, Toshihiko Yamasaki, and Kiyoharu Aizawa. Cross-domain weakly-supervised object detection through progressive domain adaptation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5001–5009, 2018. 7

[25] Kyungyul Kim, ByeongMoon Ji, Doyoung Yoon, and Sangheum Hwang. Self-knowledge distillation: A simple way for better generalization. *arXiv preprint arXiv:2006.12000*, 3, 2020. 1, 2

[26] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 5

[27] Ryan Kiros, Ruslan Salakhutdinov, and Richard S Zemel. Unifying visual-semantic embeddings with multimodal neural language models. *arXiv preprint arXiv:1411.2539*, 2014. 2

[28] Da Li, Yongxin Yang, Yi-Zhe Song, and Timothy Hospedales. Learning to generalize: Meta-learning for domain generalization. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018. 1

[29] Da Li, Jianshu Zhang, Yongxin Yang, Cong Liu, Yi-Zhe Song, and Timothy M Hospedales. Episodic training for domain generalization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1446–1455, 2019. 1

[30] Gen Li, Nan Duan, Yuejian Fang, Ming Gong, Daxin Jiang, and Ming Zhou. Unicoder-vl: A universal encoder for vision and language by cross-modal pre-training. arxiv e-prints, page. *arXiv preprint arXiv:1908.06066*, 2019. 2

[31] Haoliang Li, Sinno Jialin Pan, Shiqi Wang, and Alex C Kot. Domain generalization with adversarial feature learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5400–5409, 2018. 1

[32] Liunian Harold Li, Pengchuan Zhang, Haotian Zhang, Jianwei Yang, Chunyuan Li, Yiwu Zhong, Lijuan Wang, Lu Yuan, Lei Zhang, Jenq-Neng Hwang, et al. Grounded language-image pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10965–10975, 2022. 2

[33] Yu-Jhe Li, Xiaoliang Dai, Chih-Yao Ma, Yen-Cheng Liu, Kan Chen, Bichen Wu, Zijian He, Kris Kitani, and Peter Vajda. Cross-domain adaptive teacher for object detection. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 1, 2

[34] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013. 3

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

[36] 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

[37] Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014. 5

[38] Fengchun Qiao, Long Zhao, and Xi Peng. Learning to learn single domain generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12556–12565, 2020. 1, 2

[39] 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. 1, 2, 3, 4, 5, 6

[40] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019. 2

[41] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 2022. 3

[42] Hanoona Rasheed, Muhammad Maaz, Muhammad Uzair Khattak, Salman Khan, and Fahad Shahbaz Khan. Bridging the gap between object and image-level representations for open-vocabulary detection. *arXiv preprint arXiv:2207.03482*, 2022. 2, 4

[43] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: towards real-time object detection with region proposal networks. *IEEE transactions on pattern analysis and machine intelligence*, 39(6):1137–1149, 2016. 4, 5, 6, 7

[44] Kuniaki Saito, Yoshitaka Ushiku, Tatsuya Harada, and Kate Saenko. Strong-weak distribution alignment for adaptive object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6956–6965, 2019. 1, 2

[45] Christos Sakaridis, Dengxin Dai, and Luc Van Gool. Semantic foggy scene understanding with synthetic data. *International Journal of Computer Vision*, 126(9):973–992, 2018. 4

[46] Zhiqiang Shen, Harsh Maheshwari, Weichen Yao, and Marios Savvides. Scl: Towards accurate domain adaptive object detection via gradient detach based stacked complementary losses. *arXiv preprint arXiv:1911.02559*, 2019. 1, 2

[47] Princeton University. About wordnet. https://wordnet.princeton.edu, 2010. 5

[48] 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

[49] Vibashan VS, Vikram Gupta, Poojan Oza, Vishwanath A Sindagi, and Vishal M Patel. Mega-cda: Memory guided attention for category-aware unsupervised domain adaptive object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4516–4526, 2021. 2

[50] Xudong Wang, Zhaowei Cai, Dashan Gao, and Nuno Vasconcelos. Towards universal object detection by domain attention. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7289–7298, 2019. 1

[51] 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*, pages 834–843, 2021. 1, 2

[52] 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, 4, 5, 6, 7

[53] 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. 4

[54] Yuxin Wu, Alexander Kirillov, Francisco Massa, Wan-Yen Lo, and Ross Girshick. Detectron2. https://github.com/facebookresearch/detectron2, 2019. 5

[55] 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. 4

[56] Alireza Zareian, Kevin Dela Rosa, Derek Hao Hu, and Shih-Fu Chang. Open-vocabulary object detection using captions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14393–14402, 2021. 2

[57] Haotian Zhang, Pengchuan Zhang, Xiaowei Hu, Yen-Chun Chen, Liunian Harold Li, Xiyang Dai, Lijuan Wang, Lu Yuan, Jenq-Neng Hwang, and Jianfeng Gao. Glipv2: Unifying localization and vision-language understanding. *arXiv preprint arXiv:2206.05836*, 2022. 2

[58] Yuhao Zhang, Hang Jiang, Yasuhide Miura, Christopher D Manning, and Curtis P Langlotz. Contrastive learning of medical visual representations from paired images and text. *arXiv preprint arXiv:2010.00747*, 2020. 2

[59] Yabin Zhang, Minghan Li, Ruihuang Li, Kui Jia, and Lei Zhang. Exact feature distribution matching for arbitrary style transfer and domain generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8035–8045, 2022. 1, 2

[60] Shanshan Zhao, Mingming Gong, Tongliang Liu, Huan Fu, and Dacheng Tao. Domain generalization via entropy regularization. *Advances in Neural Information Processing Systems*, 33:16096–16107, 2020. 1

[61] Xinge Zhu, Jiangmiao Pang, Ceyuan Yang, Jianping Shi, and Dahua Lin. Adapting object detectors via selective cross-domain alignment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 687–696, 2019. 2