Prompt-Driven Dynamic Object-Centric Learning for Single Domain Generalization

Deng Li[1], Aming Wu[2], Yaowei Wang[3], Yahong Han[1]*
[1]College of Intelligence and Computing, Tianjin University, Tianjin, China
[2]School of Electronic Engineering, Xidian University, Xi'an, China
[3]Peng Cheng Laboratory, Shenzhen, China
lideng@tju.edu.cn, amwu@xidian.edu.cn, wangyw@pcl.ac.cn, yahong@tju.edu.cn

Abstract

Single-domain generalization aims to learn a model from single source domain data attaining generalized performance on other unseen target domains. Existing works primarily focus on improving the generalization ability of static networks. However, static networks are unable to dynamically adapt to the diverse variations in different image scenes, leading to limited generalization capability. Different scenes exhibit varying levels of complexity, and the complexity of images further varies significantly in cross-domain scenarios. In this paper, we propose a dynamic object-centric perception network based on prompt learning, aiming to adapt to the variations in image complexity. Specifically, we propose an object-centric gating module based on prompt learning to focus attention on the object-centric features guided by the various scene prompts. Then, with the object-centric gating masks, the dynamic selective module dynamically selects highly correlated feature regions in both spatial and channel dimensions enabling the model to adaptively perceive object-centric relevant features, thereby enhancing the generalization capability. Extensive experiments were conducted on single-domain generalization tasks in image classification and object detection. The experimental results demonstrate that our approach outperforms state-of-the-art methods, which validates the effectiveness and versatility of our proposed method.

1. Introduction

Recently, deep learning visual models have achieved rapid development [2, 17, 52]. These methods are based on the assumption that the training and testing data share a similar distribution. However, in practical applications, the training and testing data are often not drawn from the same distribution. Due to the domain shift [39], models often exhibit poor generalization performance when tested on out-of-distribution datasets. To mitigate the impact of domain shift, several approaches have been proposed, such as domain adaptation [31, 38, 46] and domain generalization [3, 15, 30] methods. Domain adaptation methods typically require the inclusion of unlabeled target domain images during the model training phase. Multiple domain generalization methods aim to mitigate the domain shift by combining data from multiple training domains to some extent. However, both of these approaches have limitations due to the expensive data acquisition and data privacy.

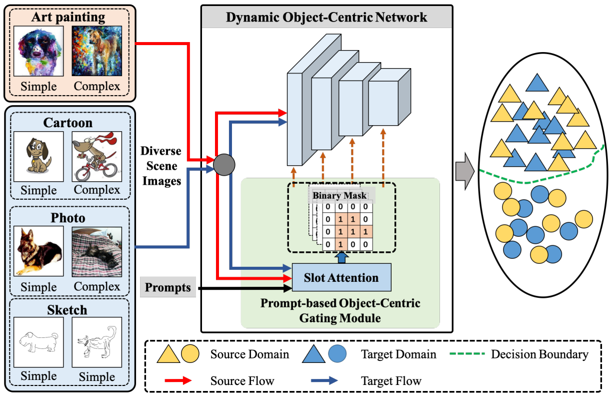


Figure 1. Illustration of dynamic object-centric learning via prompts for single domain generalization. Object-centric features capture the essential information related to individual objects. Incorporating the given scene prompts to dynamically optimize the extraction of object-centric features is beneficial for improving the generalization performance of models.

Single-domain generalization aims to train a model on a single domain and generalize its performance to diverse unseen target domains [43]. This learning paradigm poses significant challenges due to the model being trained only on a single source domain and the target domains being unavailable during the training process. Existing approaches for

*Corresponding author.

single-domain generalization primarily focus on two main methods: data augmentation [42, 49] and feature disentanglement [45]. Although the aforementioned methods have contributed positively towards mitigating domain shift in single-domain generalization tasks, they mainly focus on static networks. Static networks lack the capability to dynamically adapt to the diverse variations in different visual scenes, which limits the representation power of the models. Dynamic networks [16] dynamically adjust the structure or parameters to adapt the characteristics of the input data, expanding the parameter space, and improving the generalization performance.

In the visual tasks, each image may have its unique characteristics, such as variations in lighting conditions, object appearances, or scene structures, which result in variations in image complexity. Object-centric representations are robust to variations in appearance, context, or scene complexity, which enables the model to generalize well to unseen or novel samples. Considering the above factors, we propose a dynamic object-centric learning approach for single-domain generalization as shown in Figure 1. Specifically, a prompt-based object-centric gating module is designed to perceive object-centric features of objects, leveraging the multi-modal feature representation capabilities of the visual-language pre-trained CLIP [35] model, and the prompts that describe different domain scenes guide the learning of the dynamic gating decision for different domains. Furthermore, we proposed a Slot-Attention multimodal fusion module to fuse the linguistic features and visual features and then extract effective object-centric representations. With learned object-centric gating decisions, we selectively connect the features of the network in both spatial and channel dimensions. We validated the effectiveness of our proposed method on image classification and object detection tasks.

The main contributions of our method can be summarized as follows:

(1) To address the issue of insufficient generalization ability of single-domain generalization tasks, we propose a dynamic object-centric learning framework to enhance the generalization capability.

(2) We propose an object-centric gating module based on prompt learning which leverages the textual descriptions of various scenes to guide the learning of the gating decision for different domains. Additionally, we introduce a Slot-Attention multi-modal fusion module to extract effective object-centric representations.

(3) Extensive experiments conducted on image classification and object detection tasks of varying complexities validate the effectiveness and generality of the proposed method.

2. Related Works

2.1. Single Domain Generalization

Existing single-domain generalization methods can be divided into two categories: data or feature augmentation and learning domain-invariant features. The data augmentation method aims to generate some out-of-distribution samples at the data level or feature level. In particular, some works [42, 49] show that the method of adversarial domain augmentation can effectively improve the generalization ability and robustness of the model by synthesizing virtual images during the training process. CLIP-Gap [41] utilizes the joint representation space of visual and textual features in the pre-trained multi-modal CLIP model to learn the feature shift between the visual and textual descriptions of the target domain. L2D [44] explores improving generalization capabilities by alternating diverse sample generation and discriminative style-invariant representation learning. Wu et al. [45] proposed a method that disentanglements features into domain-specific and domain-invariant components, and then uses the domain-invariant features as teacher feature representations to enhance the generalization capability of the detection model through self-distillation.

Different from the above methods, considering that the dynamic network dynamically adjusts the network structure according to the input data, expanding the parameter space of the model and improving the representation capacity. We propose a prompt-based dynamic network single-domain generalization method, which guides the learning of the dynamic gating decision with various domain descriptive prompts.

2.2. Dynamic Networks

Dynamic networks adaptively adjust their network structure based on input data to perform inference on different input data. These methods can make decisions based on different criteria to select different sub-networks for computational execution. The prevailing dynamic network methods can be divided into two categories: early exit and gating function-based methods. MSDNet [19] employs confidence-based criteria to explore early exit methods, which divide the model into multiple stages and handle simpler inputs that require fewer complex stages in the network. GaterNet [6] and SBNet [36] utilize strategy networks or learn dynamic decisions based on gate functions. CGNet [18] and PGNet [48] take advantage of the sparsity of spatial features to achieve different output activations for the input feature maps. There are also some domain generalization methods based on dynamic networks that have been proposed. DDG [11], PE [10], and DFRL [12] are the methods of decoupling the parameters or the features into static and dynamic parts. We propose a dynamic network based on prompt learning that leverages the textual descriptions

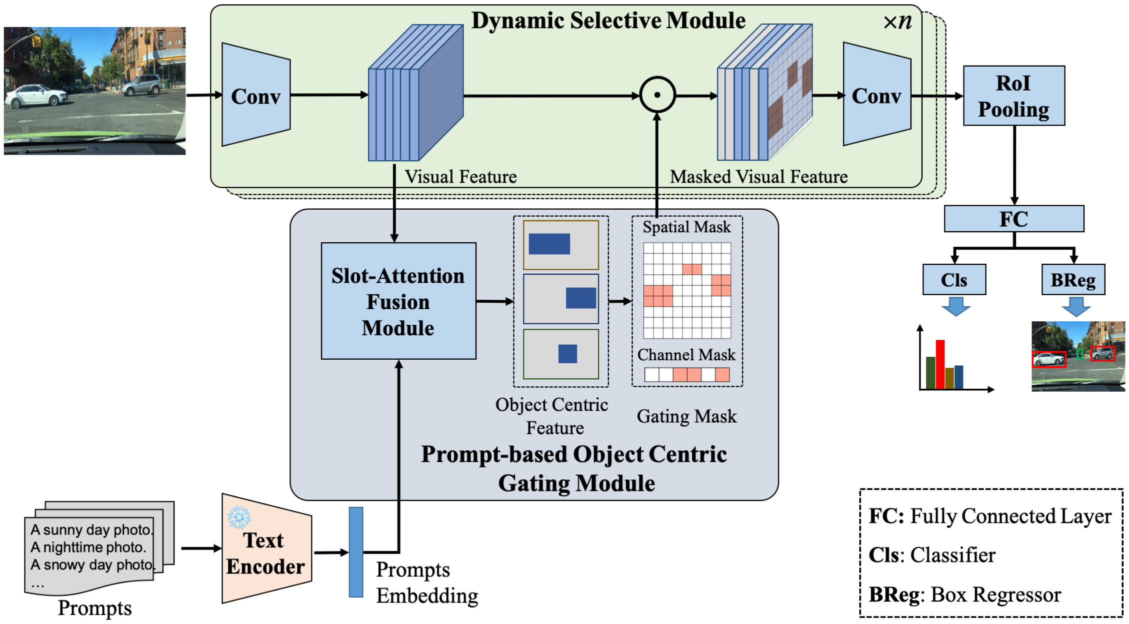


Figure 2. Illustration of our proposed prompt-based dynamic object-centric learning network for single domain generalization. This method mainly includes a prompt-based object-centric gating module and a dynamic selective module. First, the Slot-Attention multi-modal fusion module extracts object-centric features and leverages the various scene prompts to guide the object-centric gating mask learning for the input from different scenes. Next, the gating mask is used to dynamically select the relevant object-centric features to improve the generalization ability.

of various scenes to guide the learning of the gating decision for different samples. Additionally, we introduce a Slot-Attention multi-modal fusion module to extract effective object-centric representations.

2.3. Prompt Learning

Prompt learning was first studied in the NLP field as a method for fine-tuning Pre-trained Language Models (PLMs) to downstream tasks. The effectiveness of prompt learning and its advantage of only updating a small portion of parameters have recently attracted widespread attention. CoOp [13] fine-tuning CLIP [35] by optimizing a set of continuous prompt vectors in its language branch for few-shot image recognition. CoCoOp [51] addresses the overfitting problem in CoOp and proposes a dynamic prompt based on visual features to improve the performance of generalization tasks. MaPLE [23] proposed a multi-modal prompt learning method that combines the visual and linguistic branches of CLIP to learn hierarchical prompts. To incorporate the prompt description information from different scenes and dynamically adjust network structures for images of varying complexities in different scene domains. We construct a gating module based on prompt learning, which enhances the representation power of the features and guides the learning of the gating module for the inputs from different scenarios.

3. Methodology

3.1. Framework

Given a source domain $\mathcal{D}^s = \{(x_i^s, y_i^s)\}_{i=1}^{N_s}$ containing N_s samples. Single-domain generalization aims to learn a model that can generalize to many unseen target domains $\mathcal{D}^t = \{(x_i^t)\}_{i=1}^{N_t}$ using only the source domain data without prior knowledge about the target domains $\mathcal{D}^t$. To improve the generalization ability of the model, we propose a prompt-based dynamic object-centric learning network for single-domain generalization as shown in Figure 2. It contains two key components, the prompt-based object-centric gating module and the dynamic selective module. The prompt-based object-centric gating module fuses the text prompt embeddings with the visual features to learn enhanced scene information and extract object-centric representation from the fusion feature via the Slot-Attention mechanism. The dynamic selective module is used to dynamically activate the components of the network. With the gating masks output by the prompt-based object-centric gating module, we dynamically select feature maps from the blocks of the model backbone in both spatial and channel dimensions. In the spatial dimension, it identifies the spatial regions that contain significant object-centric information by the gating masks. Similarly, in the channel dimension, the gating masks help us select the most relevant

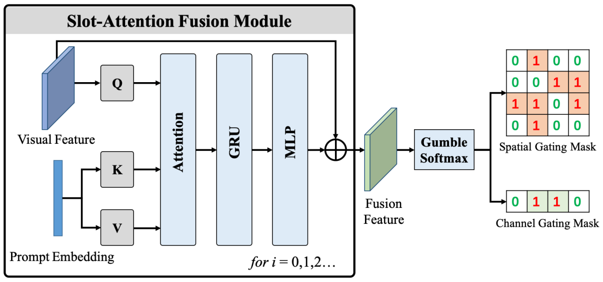


Figure 3. Illustration of our proposed prompt-based object-centric gating module.

channels that capture object-centric features.

The prompt learning module is based on the CLIP [35] which combines an image encoder and a text encoder and bridges the representation of visual and textual in joint space. We designed text description prompts in different image scenes and got the prompt embedding with the frozen text encoder of CLIP [35]. Guided by the prompts, the dynamic object-centric network can learn and extract more valuable information from the scene, and improve the performance in various scene-related tasks.

3.2. Prompt-based Object-Centric Gating Module

The dynamic network method based on gate function shows remarkable versatility and applicability and can be applied to different networks. Most works only utilize visual features for learning gating modules. These methods may involve biases in scene information and lead to overfitting problems to scenes, thereby hindering generalization capabilities to new scenes. To alleviate this issue, we utilize specifically designed scene prompts as compensation to obtain diverse information from various scenes. Object-centric representations can improve generalization capabilities by capturing essential visual attributes. By explicitly modeling objects, the learned representations can capture meaningful and transferable object-centric features that are robust to variations in appearance, context, or scene complexity. In order to fully leverage both textual prompt information and visual features and then extract meaningful object-centric representation, we have developed a multi-modal fusion module based on Slot-Attention [26], as shown in Figure 3. Slot-Attention is an attention mechanism that focuses attention on different slots, where each slot corresponds to a specific object or concept. We use the visual features as the initial slot and set the linearly transformed features as the Query Q in the Slot-Attention mechanism. The linear transformations are applied to the prompt embedding to obtain the Key K, and Value V. The attention score A is obtained by calculating the dot product between Query Q and Key K and followed it with softmax function:

$$\mathcal{A} = \text{Softmax}\left(\frac{Q \cdot K^{\top}}{\sqrt{d_Q}}\right), \tag{1}$$

where, $\sqrt{d_Q}$ is the dimension of Query Q. Then, the attention features F_{att} are obtained by the cross-product operation of attention score A and Value V:

$$F_{att} = \mathcal{A} \cdot \mathcal{V} \tag{2}$$

In addition, the slots are updated with loop iteration. During each iteration $t = 1, \cdots, T$, we use the GRU function to update the features of each iteration. Based on Slot-Attention, each prompt embedding is gradually refined according to relevant visual features. This approach allows us to explicitly model and extract object-level feature representation. With the prompt embeddings and visual features, the Slot-Attention aggregates the multi-modal features by weighting them based on the importance and relevance of the objects. The fused features are converted into gate functions:

$$slots = \text{GRU}(state = slots, inputs = F_{att}) \tag{3}$$

Guided by the prompt embeddings, the Slot-Attention fusion module can obtain the features that are relevant to the objects or concepts specified in the prompt. The gating function takes the fused features as input and generates gating masks. The gating masks act as a gate or filter that controls the flow of information within the model. Since the gating function is a binary function that is not differentiable, during the training process, the Gumbel-Softmax technique is employed to transform the discrete binary function into a continuous variable.

3.3. Dynamic Selective Module

Based on the designed object-centric gating module, we embed the gating unit into the model to achieve dynamic activation of the model. Here, we take ResNet [17] as an example and selectively activate connections from both spatial and channel levels to improve the generalization of the model. For channel-wise selective modules, we insert the selective module between the two convolutions of the block and dynamically select the feature information that should be input to the next layer. The binary mask output by the gate module is multiplied by the activation results of the convolutional layer to filter out the unimportant features. The binary mask can be expressed as follows:

$$M(i) = \begin{cases} 1 & Slot(i) \geq threshold \\ 0 & \text{Otherwise} \end{cases}, \tag{4}$$

where, $Slot_c(i)$ is the feature of the i-th output by the Slot-Attention multi-modal fusion module.

For the dynamic selective module, in each block of ResNet [17], the binary masks are obtained with the visual features and the prompt embedding through the above gate module. For the feature pyramid and the problem of different feature scales, we use the upsampling method to generate new gated features to adapt to the feature size of each layer. The masks are multiplied by the normalized features after convolution, thus filtering irrelevant spatial area features. By dynamically activating features in the network at both spatial and channel levels, different levels of sparsity can be achieved in blocks. The dynamic object-centric perception approach prevents the model from overfitting and enhances the generalization ability on single-domain generalization tasks.

3.4. Overall Training Objective

To ensure stable training of the dynamic model, we adopt the approach proposed by Verelst et al. [40] and introduce a bound loss to guide the model optimization. This bound loss constrains the sparsity of features in both spatial and channel dimensions, limiting it within the range of $\left[p\sqrt{T_d}, 1-p\left(1-\sqrt{T_d}\right)\right]$. Here T_d denotes the target rate. The lower and upper bounds of the regularization term can be expressed as:

$$\begin{aligned} L_{b,\text{ low }} &= \sum_{l=1}^{L} \sum_{k\in\{s,c\}} \max\left(0, p\sqrt{T_d} - \left|M_k^l\right|_d\right)^2 \\ L_{b,\text{ up }} &= \sum_{l=1}^{L} \sum_{k\in\{s,c\}} \max\left(0, p\left(1-\sqrt{T_d}\right) - 1 + \left|M_k^l\right|_d\right)^2 \end{aligned} \tag{5}$$

where $|\cdot|_d$ is the density of the binary masks, and the exponential annealing function $p = \exp(-\alpha\cdot \text{ epoch })$ is used to gradually loose the bound. We set the α to be 0.05 in our experiments.

By combining the loss function of the task and the bound loss function, the joint training loss function for our proposed method can be expressed as:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{task} + \lambda_b(L_{b,\text{ low }} + L_{b,\text{ up }}), \tag{6}$$

where λ_b are the weight of the bound loss.

4. Experiments

To evaluate the effectiveness of our method, we conducted experiments on various visual task scenarios, such as image classification and object detection.

4.1. Datasets

PACS [24] is a generalization benchmark data set in the image classification domain, which contains four fields, namely art paintings, cartoons, photos, and sketches. Each domain contains 7 categories of images, a total of 9,991 images, and the image size is 224 × 224 pixels. This dataset has large stylistic differences between domains and is more challenging. For a fair comparison, we use the official split strategy to obtain the training set, validation set, and test set.

Diverse-Weather Dataset. We also evaluated our method on the urban-scene detection domain generalization benchmark diverse weather dataset built by [45]. It contains five domains with different weather conditions, namely Daytime Clear, Night Clear, Dusk Rainy, Night Rainy, and Daytime Foggy. Here we use Daytime Clear data as the source domain and other domains as the target domain. The Daytime Clear domain consists of 19,395 training images, and 8,313 images are used as the validation set for model selection. The four other domains are set as target domains, including 26,158 images in the Night Clear scene, 3,501 images in the Dusk Rainy scene, 2,494 images in the Night Rainy scene, and 3,775 images in the Daytime Foggy scene.

4.2. Image Classification

4.2.1 Implementation Details

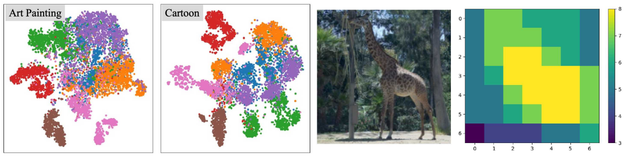
For the domain generalization task of image classification, we conducted evaluation experiments on single-source domain generalization and multi-domain generalization on the PACS dataset. For single-source domain generalization experiments, four sets of experiments were conducted with one domain as the source domain and the others as the target domain. For multi-domain generalization experiments, four sets of experiments were conducted with one of the four domains as the target domain and the other domains as the source domain. We have designed various prompts based on the designed template (such as "*an image taken in* {*scene name*}") for different scenarios. ResNet-18 [17] pretrained on ImageNet is used as the backbone network of the model and fine-tuned on the source domain. The four-layer block of ResNet-18 integrates a prompt-based dynamic selective module to connect the features in the block at the spatial level and channel level. During the training process, we train the model in 70 epochs, the batch size is set to 256. We also set the learning optimizer as SGD with a weight decay of 0.0001, and the initial learning rate is 0.01.

4.2.2 Experimental Results and Analysis

Single Domain Generalization. Table 1 shows the experimental results of our single-domain generalization method on the PACS dataset. We compared our method with state-of-the-art methods such as RSC [21], ASR [14], L2D [44], P-RC [9] and Meta-Casual [5]. Our method outperforms the state-of-the-art method with 1.2% on average classification accuracy. Specifically, our method can boost the performance by 2.5% than other methods in the cartoon domain

Table 1. Single domain generalization image classification results (%) on PACS with backbone of ResNet-18 [17].

Method	Year	Art	Cartoon	Sketch	Photo	Avg
RSC [21]	ECCV'20	73.40	75.90	56.20	41.60	61.80
RSC+ASR [14]	CVPR'21	76.70	79.30	61.60	54.60	68.10
L2D [44]	ICCV'21	76.91	77.88	53.66	52.29	65.18
P-RC [9]	CVPR'23	76.98	78.54	62.89	57.11	68.88
Meta-Casual [5]	CVPR'23	77.13	80.14	62.55	59.60	69.86
Ours	-	**78.77**	**82.69**	**62.94**	**60.09**	**71.12**



(a) t-SNE of Feature Embedding

(b) Visualization of Spatial Mask

Figure 4. (a) The t-SNE of feature embedding on the target domain of PACS, where the upper left domain name is the source domain. (b) The visualization results of the spatial mask accumulated from each block.

with relative margins. The results verify the advantages of our proposed prompt-based dynamic object-centric learning method on single-domain generalization tasks.

Multiple Domain Generalization. We also extended our method to multi-domain generalization and conducted evaluation experiments on the PACS dataset. The experimental results are shown in Table 2. In line with other multi-domain generalization approaches [5, 12], we adopted the leave-one-domain-out paradigm for our experiments. We compared our approach with existing state-of-the-art methods, such as MetaReg [1], EpiFCR [25], MASF [8], DMG [4], ME-ADA [49], MMLD [29] and Meta-Casual [5]. In addition, we also conducted comparisons with the DFRL [12], which is also based on dynamic networks. From Table 2 we can see that our method boosts the average classification accuracy with 1.0% compared to the baseline methods. This result demonstrates the effectiveness of our proposed method on multi-domain generalization.

Visualization Analysis. We conducted a visualization analysis on the learned representations of image classification in Figure 4. From the visualization results of feature embeddings in Figure 4 (a), it can be seen that our method can effectively distinguish samples from the target domain in classification tasks. Figure 4 (b) shows the visualization results of spatial masks accumulated from each block, which demonstrates the object-centric characteristics of our method.

4.3. Object Detection

4.3.1 Implementation Details

In order to further verify the effectiveness of our method, we also evaluate it on more complex object detection tasks.

Table 2. Multiple domain generalization image classification results (%) on PACS with backbone of ResNet-18 [17]. The domain name in the column is set as the target domain.

Method	Year	Art	Cartoon	Photo	Sketch	Avg
MetaReg [1]	NeurIPS'18	83.70	77.20	95.50	70.30	81.70
GUD [43]	NeurIPS'18	78.32	77.65	95.61	74.21	81.44
Epi-FCR [25]	ICCV'19	82.10	77.00	93.90	73.00	81.50
MASF [8]	NeurIPS'19	80.29	77.17	94.99	71.68	81.03
DMG [4]	ECCV'20	76.90	80.38	93.55	75.21	81.46
DDAIG [50]	AAAI'20	84.20	78.10	95.30	74.70	83.10
CSD [34]	ICML'20	78.90	75.80	94.10	76.70	81.40
RSC [21]	ECCV'20	83.43	80.31	95.99	80.85	85.15
ME-ADA [49]	NeurIPS'20	78.61	78.65	95.57	75.59	82.10
MMLD [29]	AAAI'20	81.28	77.16	96.09	72.29	81.83
L2D [44]	ICCV'21	81.44	79.56	95.51	80.58	84.27
FACT [47]	CVPR'21	85.37	78.38	95.15	79.15	84.51
MatchDG [28]	ICML'21	81.32	80.70	96.53	79.72	84.57
CIRL [27]	CVPR'22	86.08	80.59	95.93	82.67	86.32
DFRL [12]	INS'23	85.60	80.10	96.00	79.80	85.40
Meta-Casual [5]	CVPR'23	85.30	80.93	96.53	85.24	87.00
Ours	-	**86.94**	**82.50**	**97.30**	**85.55**	**88.07**

Compared with image classification tasks, object detection tasks not only require the correct classification of objects but also the accurate positioning of objects. Similar to other single-domain generalization methods for object detection, the Faster-RCNN [37] was used in the experiment with the backbone of ResNet-101 [17]. Here we conduct experiments on a dataset of urban scenes. Following other object detection domain generalization methods, here we use the data of the Daytime Clear domain as the training set, and other domains are set as four target domains in the experiments. We train the model in 100,000 iterators with a batch size of 4. The learning optimizer is set as SGD with a weight decay of 0.0005, and the learning rate is 0.001.

4.3.2 Experimental Results and Analysis

Comparison with SOTA Methods. We compared with the state-of-the-art single-domain generalization object detection method Single-DGOD [45] and CLIP-Gap [41] and the feature normalization domain generalization methods SW [33], IBN-Net [32], IterNorm [20], and ISW [7]. Faster-RCNN [37] is a simple baseline method that initializes the parameters of the model through ImageNet pre-trained weights. We set the Daytime Clear domain as the source domain and test the generalization performance on four unseen target domains (Daytime Foggy, Night Rainy, Dusk Rainy, and Night Clear) with more complex scenes. Table 3 shows the results of single-domain generalization for object detection. It can be seen that, due to the domain shift, the test performance of all the methods on the target domain drops sharply. This phenomenon reflects the importance

Table 3. Single domain generalization object detection results (%).

Method	Year	Day Clear	Night Clear	Dusk Rainy	Night Rainy	Daytime Foggy
Faster-RCNN [37]	NeurIPS'15	48.1	34.4	26.0	12.4	32.0
IBN-Net [32]	ECCV'18	49.7	32.1	26.1	14.3	29.6
IterNorm [20]	CVPR'19	43.9	29.6	22.8	12.6	28.4
SW [33]	ICCV'19	50.6	33.4	26.3	13.7	30.8
ISW [7]	CVPR'21	51.3	33.2	25.9	14.1	31.8
S-DGOD [45]	CVPR'22	**56.1**	36.6	28.2	16.6	33.5
CLIP-Gap [41]	CVPR'23	51.3	36.9	32.3	18.7	38.5
Ours	-	53.6	**38.5**	**33.7**	**19.2**	**39.1**

Table 4. Per-class results(%) on Daytime Clear to Night Clear.

Method	bus	bike	car	motor	person	rider	truck	mAP
Faster-RCNN [37]	34.7	32.0	56.6	13.6	37.4	27.6	38.6	34.4
IBN-Net [32]	37.8	27.3	49.6	15.1	29.2	27.1	38.9	32.1
IterNorm [20]	38.5	23.5	38.9	15.8	26.6	25.9	38.1	29.6
SW [33]	38.7	29.2	49.8	16.6	31.5	28.0	40.2	33.4
ISW [7]	38.5	28.5	49.6	15.4	31.9	27.5	41.3	33.2
S-DGOD [45]	40.6	35.1	50.7	19.7	34.7	32.1	43.4	36.6
CLIP-Gap [41]	37.7	34.3	58.0	19.2	37.6	28.5	42.9	36.9
Ours	**40.9**	**35.0**	**59.0**	**21.3**	**40.4**	**29.9**	**42.9**	**38.5**

Table 5. Per-class results(%) on Daytime Clear to Dusk Rainy.

Method	bus	bike	car	motor	person	rider	truck	mAP
Faster-RCNN [37]	28.5	20.3	58.2	6.5	23.4	11.3	33.9	26.0
IBN-Net [32]	37.0	14.8	50.3	11.4	17.3	13.3	38.4	26.1
IterNorm [20]	32.9	14.1	38.9	11.0	15.5	11.6	35.7	22.8
SW [33]	35.2	16.7	50.1	10.4	20.1	13.0	38.8	26.3
ISW [7]	34.7	16.0	50.0	11.1	17.8	12.6	38.8	25.9
S-DGOD [45]	37.1	19.6	50.9	13.4	19.7	16.3	40.7	28.2
CLIP-Gap [41]	37.8	22.8	60.7	16.8	26.8	**18.7**	42.4	32.3
Ours	**39.4**	**25.2**	**60.9**	**20.4**	**29.9**	16.5	**43.9**	**33.7**

Table 6. Per-class results(%) on Daytime Clear to Night Rainy.

Method	bus	bike	car	motor	person	rider	truck	mAP
Faster-RCNN [37]	16.8	6.9	26.3	0.6	11.6	9.4	15.4	12.4
IBN-Net [32]	24.6	10.0	28.4	0.9	8.3	9.8	18.1	14.3
IterNorm [20]	21.4	6.7	22.0	0.9	9.1	10.6	17.6	12.6
SW [33]	22.3	7.8	27.6	0.2	10.3	10.0	17.7	13.7
ISW [7]	22.5	11.4	26.9	0.4	9.9	9.8	17.5	14.1
S-DGOD [45]	24.4	11.6	29.5	9.8	10.5	11.4	19.2	16.6
CLIP-Gap [41]	**28.6**	12.1	**36.1**	9.2	12.3	9.6	22.9	18.7
Ours	25.6	**12.1**	35.8	**10.1**	**14.2**	**12.9**	**22.9**	**19.2**

of model generalization performance. Compared with the other methods, the performance of our method on the target domain is higher than that of the baseline method. Among them, there is a significant improvement on the Night Clear and Dusk Rainy domains, which are improved by 1.6% and 1.4% respectively. Our method improved by 0.6% in the Daytime Foggy scene, and by 0.5% in the challenging composite domain Night Rainy (Includes two stylistic transformations: nighttime and rainy conditions). The experimental results demonstrate the effectiveness of our object-centric learning method in single-domain generalization for object detection.

Daytime Clear to Night Clear. Table 4 shows the detection results on the Night Clear scene. Compared to the daytime scenes in the source domain, nighttime scenes pose challenges for object recognition and detection due to low visibility conditions. From the experimental results, it can be observed that our method outperforms other methods in various object categories. Specifically, the performance on bus, motor, and person categories has been improved by 3.2%, 2.1%, and 2.8% respectively. These results demonstrate the effective generalization ability of our dynamic network method to Daytime Clear to Night Clear scene.

Daytime Clear to Dusk Rainy. Table 5 shows the detection results on the Dusk Rainy scene. This scene is affected by low light conditions and rain and has a large domain shift from the source daytime image. Compared with other methods, our method has comparable performance on various categories of objects. Particularly, our method improves about 2.4%, 3.6%, and 3.1% on the bike, motor, and person categories, respectively. This shows that our dynamic network method can effectively improve the generalization performance of the model from Daytime Clear to Dusk Rainy.

Daytime Clear to Night Rainy. Table 6 shows the results on the Dusk Rainy scene. The nighttime rainy scene contains the effects of both low-light and rainy weather environments, and there is a large domain shift from the source daytime image. The influence of this composite domain shift brings huge challenges to object detection, which leads the model to suffer serious performance degradation. Compared with other methods, our method improves the average mAP by 0.5% and improves in the person and rider categories by 1.9% and 3.3%, respectively. The effectiveness of our method for challenging target domain scenarios is further verified.

Daytime Clear to Daytime Foggy. Table 7 shows the detection results on the Daytime Foggy scene. Objects in foggy scene images are blurred, which brings challenges to object detection. Our method shows comparable performance on various categories of objects in this scene. This shows that our dynamic network method can effectively improve the generalization performance of the model.

Visualization Analysis. We also conducted a visualization analysis on object detection as shown in Figure 5. The visualization results indicate that, compared to the CLIP-Gap [41] baseline methods, our approach achieves more accurate classification and localization of objects such as cars, person, buses, and trucks in Night Clear, Night Rainy, Daytime Foggy, and Dusk Rainy which are four complex target domain street scenes. This also validates the effectiveness

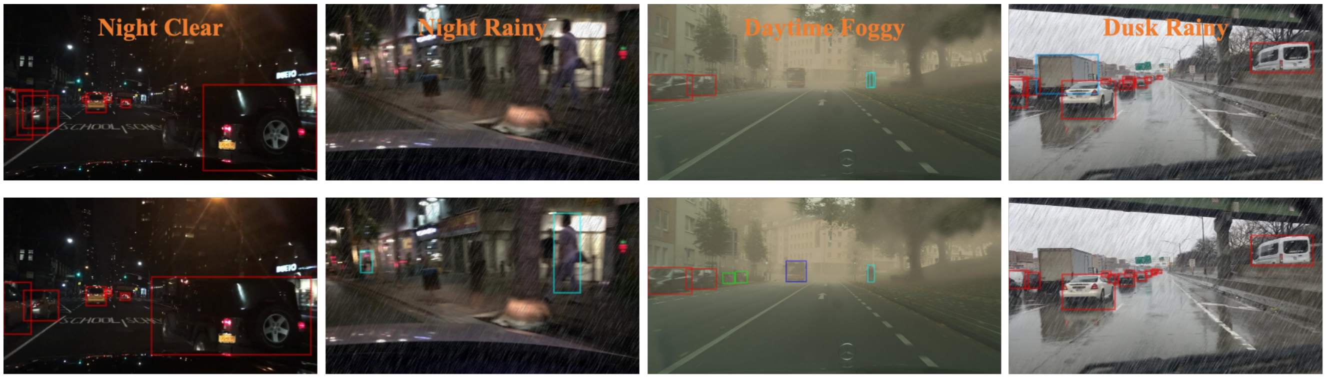


Figure 5. Detection results of the target domain on the urban scene Diverse-Weather Dataset, where the top row represents the detection results of CLIP-Gap [41], and the bottom row corresponds to our proposed method. In the "Night Clear" scene, our method achieves more accurate car detection compared to CLIP-Gap [41]. In the complex "Night Rainy" scene, CLIP-Gap [41] fails to detect the person, while our method successfully detects the person. In the "Daytime Foggy" scene, our method accurately detects small-sized buses. Furthermore, in the "Dusk Rainy" scene, our method exhibits improved accuracy in identifying and localizing trucks.

Table 7. Per-class results(%) on Daytime Clear to Daytime Foggy.

Method	bus	bike	car	motor	person	rider	truck	mAP
Faster-RCNN [37]	28.1	29.7	49.7	26.3	33.2	35.5	21.5	32.0
IBN-Net [32]	29.9	26.1	44.5	24.4	26.2	33.5	22.4	29.6
IterNorm [20]	29.7	21.8	42.4	24.4	26.0	33.3	21.6	28.4
SW [33]	30.6	26.2	44.6	25.1	30.7	34.6	23.6	30.8
ISW [7]	29.5	26.4	49.2	27.9	30.7	34.8	24.0	31.8
S-DGOD [45]	32.9	28.0	48.8	29.8	32.5	38.2	24.1	33.5
CLIP-Gap [41]	36.1	34.3	58.0	33.1	39.0	43.9	25.1	38.5
Ours	**36.1**	**34.5**	**58.4**	**33.3**	**40.5**	**44.2**	**26.2**	**39.1**

of the object-centric features of our method for object detection single-domain generalization tasks.

4.4. Ablation Study

Some ablation studies are conducted to analyze the impact of different components in our proposed method. First, we perform an ablation study to assess the contribution of the Slot-Attention mechanism by replacing it with a traditional attention method. Second, we also conduct an additional ablation analysis by removing the prompt-based adaptation mechanism from our dynamic network approach. This analysis aims to assess the significance of prompts in guiding the network dynamic adjustments.

Table 8 shows the results of the ablation experiment. It can be seen that when introducing dynamic networks for training, the average accuracy of the model reaches 64.27%, marking a significant improvement over the baseline method. The average accuracy of the model is 68.94% when introducing traditional attention methods. Finally, when introducing the prompts-driven object-centric learning module based on the Slot-Attention mechanism, the generalization performance of our method is further improved, with an average accuracy of 71.12%. We also reimplement our method with the MindSpore [22] framework to validate our method on various deep learning frameworks.

Table 8. Ablation study (%) on PACS dataset with backbone of ResNet-18 [17]. The domain name in the column is used as the source domain, and the other domains are used as the target domains. '♯' indicates the results that we reimplement with the MindSpore [22].

Method	Prompt	Dynamic	Attention	A	C	S	P	Avg
Base	×	×	×	71.26	67.64	43.97	36.99	54.97
Ours	✓	✓	×	74.29	78.54	56.54	47.74	64.27
Ours	✓	✓	Normal	75.78	81.94	59.94	58.09	68.94
Ours	×	✓	Slot	76.48	76.68	57.55	56.51	66.81
Ours♯	×	✓	Slot	76.50	75.82	58.97	57.32	67.15
Ours	✓	✓	Slot	**78.77**	**82.69**	**62.94**	**60.09**	**71.12**

5. Conclusion

Due to the domain shift, models trained on a single domain often suffer from significant performance degradation when tested on unseen target domains. Furthermore, different visual scenes in real-world scenarios require varying model complexities, while static networks are prone to overfitting. In this paper, we propose a dynamic object-centric learning approach via prompts to dynamically adjust the network to perceive object-centric features, thereby enhancing the generalization performance. First, we propose a multi-modal fusion module based on the Slot-Attention to extract object-centric features from objects. In addition, a prompt-based object-centric gating module is introduced to leverage the various scene prompts to guide the learning of the gating masks for various scenes. Finally, the object-centric gating masks are used to dynamically select the relevant object-centric feature within a model leading to more accurate and robust predictions. Extensive experiments conducted on image classification and object detection tasks have validated the effectiveness of our proposed method.

Acknowledgment. This work was supported in part by the NSFC (under Grant 62376186, 61932009, 62102293) and in part by the CAAI-Huawei MindSpore Open Fund.

References

[1] Yogesh Balaji, Swami Sankaranarayanan, and Rama Chellappa. Metareg: Towards domain generalization using meta-regularization. *Advances in neural information processing systems*, 31, 2018. 6

[2] 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. 1

[3] Fabio M Carlucci, Antonio D'Innocente, Silvia Bucci, Barbara Caputo, and Tatiana Tommasi. Domain generalization by solving jigsaw puzzles. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2229–2238, 2019. 1

[4] Prithvijit Chattopadhyay, Yogesh Balaji, and Judy Hoffman. Learning to balance specificity and invariance for in and out of domain generalization. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IX 16*, pages 301–318. Springer, 2020. 6

[5] Jin Chen, Zhi Gao, Xinxiao Wu, and Jiebo Luo. Meta-causal learning for single domain generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7683–7692, 2023. 5, 6

[6] Zhourong Chen, Yang Li, Samy Bengio, and Si Si. You look twice: Gaternet for dynamic filter selection in cnns. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9172–9180, 2019. 2

[7] 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, 8

[8] 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. 6

[9] Choi et al. Progressive random convolutions for single domain generalization. *CVPR*, 2023. 5, 6

[10] Lin et al. Parameter exchange for robust dynamic domain generalization. *ACM MM*, 2023. 2

[11] Sun et al. Dynamic domain generalization. *CVPR*, 2022. 2

[12] Wang et al. Enhanced dynamic feature representation learning framework by fourier transform for domain generalization. *Information Sciences*, 2023. 2, 6

[13] Zhou et al. Learning to prompt for vision-language models. *IJCV*, 2022. 3

[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. 5, 6

[15] Muhammad Ghifary, W Bastiaan Kleijn, Mengjie Zhang, and David Balduzzi. Domain generalization for object recognition with multi-task autoencoders. In *Proceedings of the IEEE international conference on computer vision*, pages 2551–2559, 2015. 1

[16] Yizeng Han, Gao Huang, Shiji Song, Le Yang, Honghui Wang, and Yulin Wang. Dynamic neural networks: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(11):7436–7456, 2021. 2

[17] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016. 1, 4, 5, 6, 8

[18] Weizhe Hua, Yuan Zhou, Christopher M De Sa, Zhiru Zhang, and G Edward Suh. Channel gating neural networks. *Advances in Neural Information Processing Systems*, 32, 2019. 2

[19] Gao Huang, Danlu Chen, Tianhong Li, Felix Wu, Laurens Van Der Maaten, and Kilian Q Weinberger. Multi-scale dense networks for resource efficient image classification. *arXiv preprint arXiv:1703.09844*, 2017. 2

[20] 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, 7, 8

[21] Zeyi Huang, Haohan Wang, Eric P Xing, and Dong Huang. Self-challenging improves cross-domain generalization. In *European conference on computer vision*, pages 124–140. Springer, 2020. 5, 6

[22] Huawei. Mindspore. https://www.mindspore.cn/, 2023. 8

[23] Muhammad Uzair Khattak, Hanoona Rasheed, Muhammad Maaz, Salman Khan, and Fahad Shahbaz Khan. Maple: Multi-modal prompt learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19113–19122, 2023. 3

[24] Da Li, Yongxin Yang, Yi-Zhe Song, and Timothy M Hospedales. Deeper, broader and artier domain generalization. In *Proceedings of the IEEE international conference on computer vision*, pages 5542–5550, 2017. 5

[25] 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. 6

[26] Francesco Locatello, Dirk Weissenborn, Thomas Unterthiner, Aravindh Mahendran, Georg Heigold, Jakob Uszkoreit, Alexey Dosovitskiy, and Thomas Kipf. Object-centric learning with slot attention. *Advances in neural information processing systems*, 33:11525–11538, 2020. 4

[27] Fangrui Lv, Jian Liang, Shuang Li, Bin Zang, Chi Harold Liu, Ziteng Wang, and Di Liu. Causality inspired representation learning for domain generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8046–8056, 2022. 6

[28] Divyat Mahajan, Shruti Tople, and Amit Sharma. Domain generalization using causal matching. In *International Conference on Machine Learning*, pages 7313–7324. PMLR, 2021. 6

[29] Toshihiko Matsuura and Tatsuya Harada. Domain generalization using a mixture of multiple latent domains. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 11749–11756, 2020. 6

[30] Krikamol Muandet, David Balduzzi, and Bernhard Schölkopf. Domain generalization via invariant feature representation. In *International conference on machine learning*, pages 10–18. PMLR, 2013. 1

[31] Zak Murez, Soheil Kolouri, David Kriegman, Ravi Ramamoorthi, and Kyungnam Kim. Image to image translation for domain adaptation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4500–4509, 2018. 1

[32] 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, 7, 8

[33] 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, 7, 8

[34] Vihari Piratla, Praneeth Netrapalli, and Sunita Sarawagi. Efficient domain generalization via common-specific low-rank decomposition. In *International Conference on Machine Learning*, pages 7728–7738. PMLR, 2020. 6

[35] 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, 3, 4

[36] Mengye Ren, Andrei Pokrovsky, Bin Yang, and Raquel Urtasun. Sbnet: Sparse blocks network for fast inference. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8711–8720, 2018. 2

[37] 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. 6, 7, 8

[38] Swami Sankaranarayanan, Yogesh Balaji, Carlos D Castillo, and Rama Chellappa. Generate to adapt: Aligning domains using generative adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8503–8512, 2018. 1

[39] Masashi Sugiyama and Amos J Storkey. Mixture regression for covariate shift. *Advances in neural information processing systems*, 19, 2006. 1

[40] Thomas Verelst and Tinne Tuytelaars. Dynamic convolutions: Exploiting spatial sparsity for faster inference. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition*, pages 2320–2329, 2020. 5

[41] 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. 2, 6, 7, 8

[42] Riccardo Volpi and Vittorio Murino. Addressing model vulnerability to distributional shifts over image transformation sets. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7980–7989, 2019. 2

[43] 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, 6

[44] 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. 2, 5, 6

[45] 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

[46] Aming Wu, Rui Liu, Yahong Han, Linchao Zhu, and Yi Yang. Vector-decomposed disentanglement for domain-invariant object detection. In *ICCV*, pages 9342–9351, 2021. 1

[47] Qinwei Xu, Ruipeng Zhang, Ya Zhang, Yanfeng Wang, and Qi Tian. A fourier-based framework for domain generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14383–14392, 2021. 6

[48] Yichi Zhang, Ritchie Zhao, Weizhe Hua, Nayun Xu, G Edward Suh, and Zhiru Zhang. Precision gating: Improving neural network efficiency with dynamic dual-precision activations. *arXiv preprint arXiv:2002.07136*, 2020. 2

[49] Long Zhao, Ting Liu, Xi Peng, and Dimitris Metaxas. Maximum-entropy adversarial data augmentation for improved generalization and robustness. *Advances in Neural Information Processing Systems*, 33:14435–14447, 2020. 2, 6

[50] Kaiyang Zhou, Yongxin Yang, Timothy Hospedales, and Tao Xiang. Deep domain-adversarial image generation for domain generalisation. In *Proceedings of the AAAI conference on artificial intelligence*, pages 13025–13032, 2020. 6

[51] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Conditional prompt learning for vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16816–16825, 2022. 3

[52] Zhengxia Zou, Keyan Chen, Zhenwei Shi, Yuhong Guo, and Jieping Ye. Object detection in 20 years: A survey. *Proceedings of the IEEE*, 2023. 1