Simulate, Refocus and Ensemble: An Attention-Refocusing Scheme for Domain Generalization

Ziyi Wang, Zhi Gao, Jin Chen, Qingjie Zhao, Xinxiao Wu*, *Member, IEEE,* Jiebo Luo, *Fellow, IEEE*

***Abstract*—Domain generalization (DG) aims to learn a model from source domains and apply it to unseen target domains with out-of-distribution data. Owing to CLIP's strong ability to encode semantic concepts, it has attracted increasing interest in domain generalization. However, CLIP often struggles to focus on task-relevant regions across domains, *i.e.,* domain-invariant regions, resulting in suboptimal performance on unseen target domains. To address this challenge, we propose an attention-refocusing scheme, called *Simulate, Refocus and Ensemble (SRE)*, which learns to reduce the domain shift by aligning the attention maps in CLIP via attention refocusing. SRE first simulates domain shifts by performing augmentation on the source data to generate simulated target domains. SRE then learns to reduce the domain shifts by refocusing the attention in CLIP between the source and simulated target domains. Finally, SRE utilizes ensemble learning to enhance the ability to capture domain-invariant attention maps between the source data and the simulated target data. Extensive experimental results on several datasets demonstrate that SRE generally achieves better results than state-of-the-art methods. The code is available at: https://github.com/bitPrincy/SRE-DG.**

***Index Terms*—Domain generalization; CLIP; Attention refocusing**

I. Introduction

DOMAIN Generalization (DG) aims to train models that can perform well on unseen target domains of different distributions by using labeled data from source domains [1]–[6]. Recently, CLIP [7] has emerged as a promising candidate for domain generalization [8]–[10], since it is pre-trained on larger scale data and encodes richer semantic concepts than traditional neural networks [7], [11]–[15].

However, CLIP has not yet achieved excellent performance in domain generalization. One important reason is that CLIP tends to understand images from a global perspective, as it is trained by aligning images and texts instead of learning discriminative features for classification. Hence, CLIP often struggles to focus on task-relevant regions, making it difficult to capture domain-invariant representations. As shown in the top part of Figure 1, the attention maps of input images generated by CLIP mainly focus on the background regions unrelated to classification.

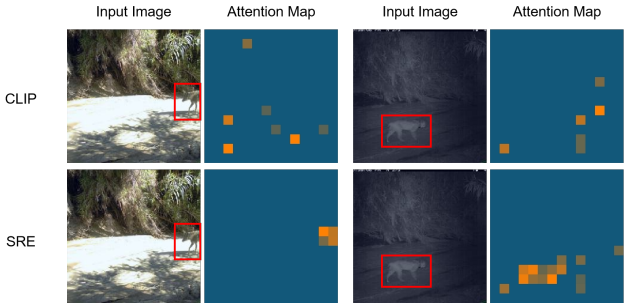

Fig. 1. Attention maps of input images generated by CLIP and the proposed SRE. CLIP often cannot focus on task-relevant regions (*i.e.*, regions of the animals), while SRE can focus on those regions, thus facilitating the subsequent classification.

To address the above challenge, we introduce attention refocusing into CLIP, thus reducing the domain shift by refocusing the attention across domains. Attention refocusing usually uses additional attention masks [16], fine-tuned attention weights [17], or a learned bias to modify the attention mechanism of inputs [18]. In this case, the model's attention is directed to task-relevant image regions, showing effectiveness in multiple visual tasks, such as image synthesis [17], [19] and image classification [18], [20].

Inspired by this, we introduce an attention-refocusing scheme for domain generalization, by which CLIP adaptively focuses more on task-relevant regions containing domain-invariant information, bridging the source and unseen target domains.

In this paper, we propose a three-stage DG scheme, named *Simulate, Refocus and Ensemble (SRE)*, where the domain shifts are simulated and bridged over the attention maps of data. The *Simulate* stage generates simulated target domains by augmenting source images with consistent spatial relationships. The *Refocus* stage introduces an attention-refocuser to capture the data distributions of the source images and simulated target images, and align the attention maps of the two kinds of images based on their data distributions. The *Ensemble* stage selects the parameters of the attention-refocuser with high-consistent attention between the source and the simulated target images, filtering out the parameters that are skewed on challenging images (e.g., the object and background are similar in color and texture). Figure 2 shows

Ziyi Wang, Zhi Gao, Jin Chen, Qingjie Zhao are with the Beijing Key Laboratory of Intelligent Information Technology, School of Computer Science and Technology, Beijing Institute of Technology, Beijing 100081, China.

Xinxiao Wu is with the Guangdong Laboratory of Machine Perception and Intelligent Computing, Shenzhen MSU-BIT University, Shenzhen 518172, and also with the Beijing Key Laboratory of Intelligent Information Technology, School of Computer Science and Technology, Beijing Institute of Technology, Beijing 100081, China.

Jiebo Luo is with the Department of Computer Science, University of Rochester, Rochester, NY 14627 USA.

*Corresponding author: Xinxiao Wu (e-mail: wuxinxiao@bit.edu.cn).

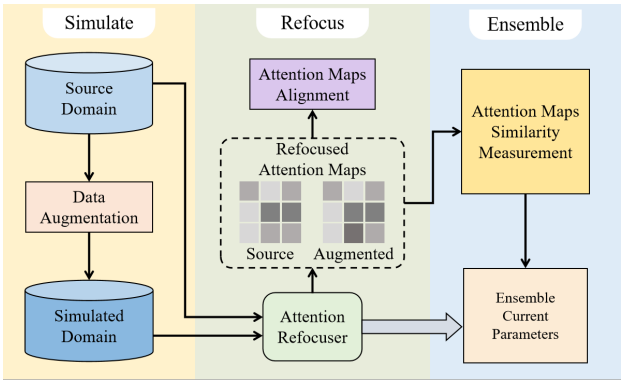

Fig. 2. An overview of the proposed DG attention-refocusing scheme SRE.

the overview of SRE.

Our scheme benefits from attention refocusing and allows CLIP to adaptively focus on domain-invariant regions within each sample, thereby learning domain-invariant features. Different from existing attention-refocusing methods, such as TOAST [18] that requires an additional training stage to refocus on task-relevant regions in downstream tasks, our scheme learns to perform refocusing on diverse simulated domains, through which it can effectively adapt to unseen target domains without any extra training. Experimental results demonstrate that our method achieves competitive performance compared to state-of-the-art methods, verifying that attention refocusing significantly improves generalization.

The remainder of this paper is organized as follows. In Section II, we summarize previous works related to our method. In Section III, we illustrate the proposed attention-refocusing scheme SRE for domain generalization. Section IV demonstrates experimental results on various benchmark datasets and the conclusion is presented in Section V.

II. Related Work

A. Domain Generalization

Domain generalization (DG) aims to develop models that can generalize well to unseen target domains given data from multiple source domains. This problem is critical in real-world applications where models face significant domain shifts, such as changes in environments, sensor configurations, or image styles. Unlike domain adaptation, which assumes access to unlabeled or labeled data in the target domain during training, DG methods operate under a more rigorous setting where no target domain information is available, making it a more challenging yet more practical research problem.

Traditional domain generalization methods primarily include data augmentation, representation learning, and learning strategy. The data augmentation methods [5], [6], [21]–[23] diversify the source domain by perturbing original data and expanding the data distribution within the source domain. The representation learning methods [1], [4], [24]–[26] extract domain-invariant features and learn to separate semantic information for classification. The learning strategy methods [27]–[30] modify the optimization process to improve generalization.

B. CLIP-based Domain Generalization

Large-scale pre-trained models [7], [31], [32] have emerged as powerful tools in modern machine learning, achieving impressive generalization across diverse tasks and domains. By training extensively on vast amounts of data, these models output expressive representations, making them particularly well-suited for tasks that require adaptation to new or unseen scenarios.

Recently, the large-scale pre-trained model, Contrastive Language-Image Pre-Training (CLIP) [7], has achieved impressive performance in various tasks such as image classification [33]–[36], image-text retrieval [37], [38], visual question answering [39], [40], and person re-identification [41], providing a promising option for domain generalization. Many methods design prompt tuning schemes for domain generalization with CLIP. Zhang *et al.* [42] propose a textual prompt learning framework, which constructs batch-wise text prompts based on visual features. Bose *et al.* [43] construct sample-level text prompts using feature statistics and introduce projectors to incorporate intermediate image features into text features. Niu *et al.* [10] introduce a source-free manner for CLIP, which contrastively pulls text prompts of the same class across domains closer and separates text prompts of different categories. In addition to those methods focusing on text prompts, Li *et al.* [9] perform visual prompt tuning (VPT) [44] at each layer of the vision transformer (ViT) [45] and extend VPT by adding sample-specific tokens mapped from the input context tokens.

C. Attention Refocusing in Transformer

Transformers [46] are widely used in both natural language processing and computer vision due to their ability to model complex relationships in data. The attention mechanism plays an important role in Transformers, enabling the model to focus on the most relevant parts of the input and making Transformers effective for a variety of tasks. Attention refocusing refers to adjusting the attention mechanism in Transformer-based networks based on input data. This process typically involves using additional attention masks, fine-tuning attention weights, or learning a bias for the attention mechanism to emphasize the visual or textual regions of interest. This concept has been used in computer vision and neural language processing tasks, such as image classification [18], [47], image synthesis [17], image segmentation [48], [49], and human-model interaction [16], [50]. Phung *et al.* [17] design a cross-attention loss and a self-attention loss to update the weights of the attention mechanism to adjust the focus on correct image regions. Shi *et al.* [18] propose a top-down attention fine-tuning framework to learn bias to adjust the input of value matrices in self-attention.

Although attention refocusing has achieved certain success in various tasks, its capability in domain generalization has not been fully exploited. Existing studies primarily focus on optimizing attention for known tasks or domains, leaving the question of how to adaptively refocus attention for unseen domains unanswered. This paper introduces attention refocusing in CLIP and proposes a novel framework to adaptively adjust attention in CLIP based on data distributions. Our

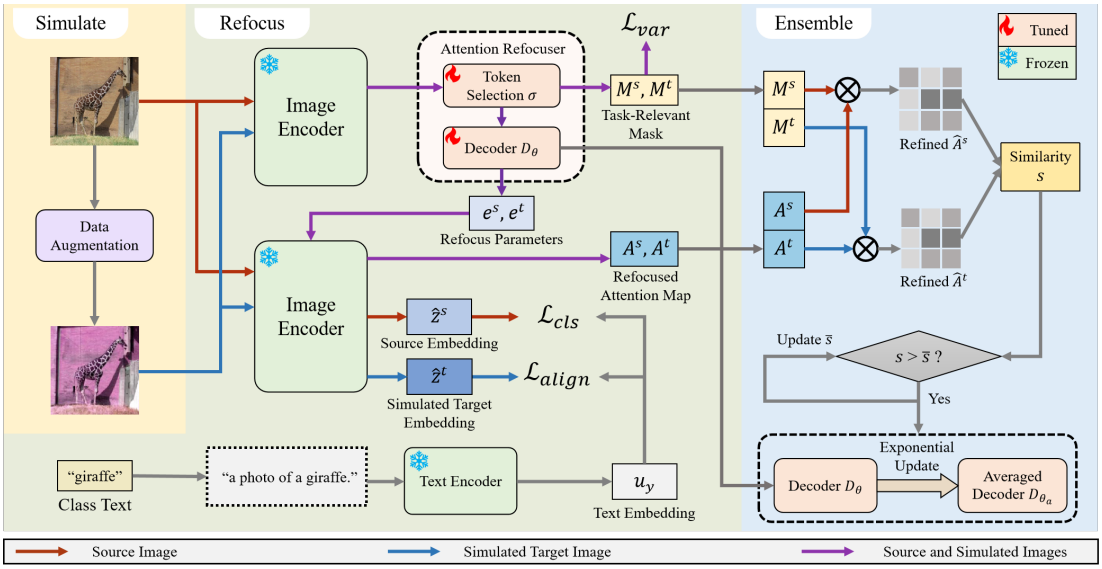

Fig. 3. Overview of the three stages of the proposed SRE. The *Simulate* stage performs data augmentation to simulate domain shift. The *Refocus* stage introduces an attention-refocuser and learns to align the attention maps of the source and the simulated target images. The *Ensemble* stage selects and ensembles parameters of the attention-refocuser that captures domain-invariant attention maps between the source and the simulated target images.

method not only extends the utility of attention refocusing but also demonstrates its effectiveness in enhancing model generalization across various unknown domains.

III. Our Method

A. Overview

In this paper, we focus on the problem of multi-source domain generalization, where multiple source domains $\mathcal{S}$ are available during training, while the target domains $\mathcal{T}$ have distributions differing from the source domains and are inaccessible during training. To tackle this problem, we propose *Simulate, Refocus and Ensemble (SRE)*, a three-stage scheme for multi-source domain generalization, which learns to reduce domain shift by aligning the attention maps in CLIP via attention refocusing. In the *Simulate* stage, we augment each source image through data transformations to generate simulated target domains, mimicking domain shifts between source and target domains. In the *Refocus* stage, we introduce an attention-refocuser that simultaneously adjusts the attention of source and simulated target images, mitigating the simulated domain shifts and helping CLIP focus on domain-invariant regions. In the *Ensemble* stage, we stabilize the training process by selecting and combining robust parameters of the attention-refocuser while removing parameters that are sensitive to the domain shifts, and thus ensure that the focused regions are similar between the source and simulated target images. Figure 3 illustrates the overview of our method.

B. Simulate via Data Transformation

In multi-source domain generalization, domain shifts between the source and target domains, such as differences in lighting, texture, and color distribution, pose a significant challenge to generalization. Since the target domain data is unavailable during training, our goal is not to precisely simulate the specific domain shifts between the source and target domains, which is inherently infeasible due to the unknown distribution of the target domain. Instead, we aim to create a training scenario that exposes the model to diverse domain shifts, thereby forcing the model to learn to address the domain shifts and enhancing its ability to generalize across unseen domains.

To achieve this, we simulate different degrees of domain shift to diversify the training data as much as possible. At the same time, to refocus CLIP's attention on domain-invariant regions, we create simulated target domains $\mathcal{S}^t$ with controlled variability by performing data transformations on source images without changing their spatial location information. This allows us to model domain shifts during training and ensures that the simulated target images retain the same task-relevant regions as the source images.

The augmentation process is formalized as a sequence of transformations applied to the source image x^s. One transformation is represented as a function $F_i(x^s, \phi_i)$, where ϕ_i is a parameter controlling the augmentation intensity, uniformly sampled from a predefined range. Given the input source image x^s, we apply three transformations to x^s and the corresponding simulated target image x^t is given by

$$x^t = F_3(F_2(F_1(x^s, \phi_1), \phi_2), \phi_3), \tag{1}$$

where F_1, F_2, and F_3 represent specific augmentation operations. In our implementation, the three transformations are **ColorJitter**, **GaussianBlur**, and **GrayScale**. **ColorJitter** adjusts brightness, contrast, saturation, and hue to simulate lighting changes; **GaussianBlur** introduces slight blurring to mimic variations in focus; **GrayScale** reduces color information to simulate grayscale conditions. By using the three transformations, the simulated target domain $\mathcal{S}^t$ is created to mimic the distribution discrepancy between source and target domains.

C. Refocus via Attention Map Alignment

A recent study [51] provides a theoretical analysis that fine-tuning the query or value matrices within the attention mechanism can improve generalization bounds and enhance memory efficiency. The goal of domain generalization is to develop models that can perform effectively on unseen target domains by leveraging knowledge from multiple source domains. Attention mechanisms contribute to this goal by emphasizing domain-invariant features while minimizing reliance on domain-specific attributes [52].

Given the source domains and the simulated target domains, we introduce an attention-refocuser to bridge the domain shifts between them. Specifically, the attention-refocuser learns to adaptively refine attention maps of source images and simulated target images based on their feature distributions, through which attention maps of simulated target images are aligned with source images. We display the detailed architecture of the attention-refocuser in Figure 4.

Denote f_v as the image encoder of CLIP and E as the original attention parameter in f_v, the attention-refocuser R_g learns to modify the attention parameter E in f_v to reduce the domain shifts between source domains and simulated target domains. $g = \{\sigma, \theta\}$ is the parameter of the attention-refocuser R_g, consisting of a learnable prompt σ for a token selection module and an attention decoder D_θ parameterized by θ. The attention parameter E in f_v for a source image x^s is refined based on its visual embedding $f_v(x^s, E)$, denoted by $\hat{E}^s = \{E, e^s\}$, where $e^s = R_g(f_v(x^s, E))$ is the newly added parameter for the attention mechanism. Similarly, the refined attention parameter for the simulated target image x^t is given by $\hat{E}^t = \{E, e^t\} = \{E, R_g(f_v(x^t, E))\}$. In this case, CLIP extracts the domain-invariant embeddings $\hat{z}^s$ and $\hat{z}^t$ of the source image x^s and the simulated target image x^t, respectively, formulated by

$$\hat{z}^s = f_v\left(x^s, \hat{E}^s\right), \ \hat{z}^t = f_v\left(x^t, \hat{E}^t\right). \tag{2}$$

Specifically, both the source image and its simulated target image are passed through two forward processes of f_v. In the first forward process, visual embeddings $z^s = f_v(x^s, E)$ and $z^t = f_v(x^t, E)$ are extracted from x^s and x^t and then used to generate refocus parameters e^s and e^t for each self-attention layer. In the second forward process, e^s and e^t are used in turn to adjust the self-attention mechanism in f_v, yielding the refocused visual embeddings $\hat{z}^s$ and $\hat{z}^t$ for the source image and the simulated target image, respectively.

During the first forward process, the image encoder f_v takes the source image x^s and the simulated target image x^t as input and produces corresponding visual embeddings z^s and z^t, consisting of L tokens each, represented as $z^s = \{z_l^s\}_{l=1}^L$ and $z^t = \{z_l^t\}_{l=1}^L$, where $z_l^s, z_l^t \in \mathbb{R}^d$, $z^s, z^t \in \mathbb{R}^{L\times d}$, with d representing the embedding dimension of the transformer. Then, we feed z^s, z^t into the attention-refocuser R_g which consists of a token selection module and a decoder. The token selection module extracts task-relevant tokens from z^s, z^t by calculating two task-relevant masks $M^s, M^t \in \mathbb{R}^L$ using a learnable task-relevant prompt $\sigma \in \mathbb{R}^d$:

$$M_l^s = \langle z_l^s, \sigma \rangle, \quad M_l^t = \langle z_l^t, \sigma \rangle, \tag{3}$$

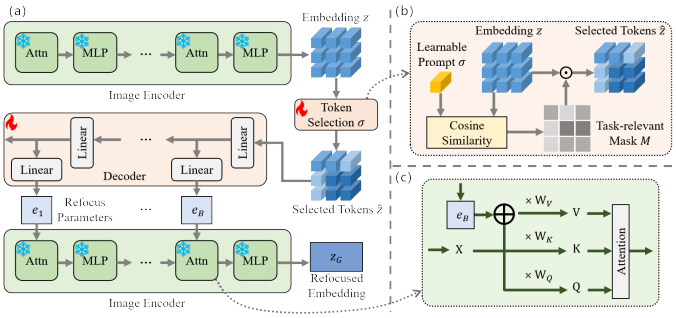

Fig. 4. The architecture of the attention-refocuser in the *Refocus* stage. (a) The attention-refocuser consists of a token selection module and a decoder. (b) The token selection module uses a learnable prompt to select task-relevant tokens. (c) The modified attention mechanism adds the refocus parameters to the input of the value matrix calculation.

where M_l^s and M_l^t denote the l-th element in M^s and M^t, respectively, measuring the task relevance of the l-th token in z^s and z^t. $\langle \cdot, \cdot \rangle$ is the cosine similarity. Selected tokens are denoted by $\tilde{z}^s = (M^s \odot z^s)$ and $\tilde{z}^t = (M^t \odot z^t)$, where $\odot$ is the Hadamard product. To increase the disparity of the task-relevant mask at different token positions, we further compute a variance loss on M^s and M^t:

$$\mathcal{L}_{var} = \mathbb{E}\left[(M^s - \mathbb{E}[M^s])^2\right] + \mathbb{E}\left[(M^t - \mathbb{E}[M^t])^2\right]. \tag{4}$$

The decoder D_θ comprises B layers, matching the number of layers in f_v, to convert the selected tokens into refocus parameters for attention. Given $\tilde{z}^s$ as the input for the first layer of D_θ, the output of the last b-th layer in the decoder D_θ is $e_b^s \in \mathbb{R}^{L\times d}$, which is used as the refocus parameter to adjust the attention mechanism in the b-th layer in f_v. All outputs of the decoder are represented as $e^s = \{e_b^s\}_{b=1}^B$. The refocus parameters e^t for $\tilde{z}^t$ can be obtained with the same procedure.

When it comes to the second forward process, the self-attention mechanism within each layer in the image encoder f_v integrates the corresponding refocus parameters into the computation of the value matrix. Conventionally, given the input x^s, the value matrix $\mathbf{V}_b$ in the b-th layer is derived as $\mathbf{V}_b = \mathbf{W}_{V_b}\mathbf{X}_b$, where $\mathbf{W}_{V_b}$ denotes the weight matrix corresponding to the value operation and $\mathbf{X}_b$ denotes the input visual embedding corresponding to x^s in the b-th layer. The refocus parameters e_b^s serve as a bias, modifying the calculation to $\hat{\mathbf{V}}_b = \mathbf{W}_{V_b}(\mathbf{X}_b + e_b^s)$, where $\hat{\mathbf{V}}_b$ represents the value matrix adjusted by the refocus parameters e_b^s. So we obtain the refined attention parameters $\hat{E}^s = \{E, e^s\}$ and the refocused visual embeddings $\hat{z}^s = f_v(x^s, \hat{E}^s)$. Similarly, we derive $\hat{z}^t = f_v(x^t, \hat{E}^t)$ for the simulated target images.

We use the text encoder f_t of CLIP to obtain text embeddings $u_c = f_t(t_c)$ for each category c, where t_c is the text prompt, constructed as $t_c =$ "a photo of a $[CLASS(c)]$", following the same manner in zero-shot CLIP [7]. With the incorporation of attention refocusing, the prediction probabilities of the source image and the simulated target image for

the label y are given by

$$p(y|x^s) = \frac{\exp(\langle u_y, \hat{z}^s \rangle / \tau)}{\sum_{c \in \mathcal{Y}} \exp(\langle u_c, \hat{z}^s \rangle / \tau)}, \\ p(y|x^t) = \frac{\exp(\langle u_y, \hat{z}^t \rangle / \tau)}{\sum_{c \in \mathcal{Y}} \exp(\langle u_c, \hat{z}^t \rangle / \tau)}, \tag{5}$$

where $\langle \cdot, \cdot \rangle$ denotes cosine similarity and τ is a temperature coefficient. We then employ the cross-entropy loss function on the source domains $\mathcal{S}$ to optimize the attention-refocuser, formulated by

$$\mathcal{L}_{cls} = \mathbb{E}_{(x^s,y)\sim\mathcal{S}}\big[-\log p(y|x^s)\big], \tag{6}$$

while forcing the attention-refocuser to simultaneously correctly refocus the attention map of the simulated target image, thereby aligning the source domain $\mathcal{S}$ and simulated target domain $\mathcal{S}^t$:

$$\mathcal{L}_{align} = \mathbb{E}_{(x^t,y)\sim\mathcal{S}^t}\big[-\log p(y|x^t)\big]. \tag{7}$$

Consequently, the parameters $R = \{\sigma, \theta\}$ of the attention-refocuser R_g are jointly optimized by the total loss:

$$\mathcal{L} = \mathcal{L}_{cls} + \mathcal{L}_{align} + \lambda \mathcal{L}_{var}, \tag{8}$$

where λ is a scaling hyper-parameter.

D. Ensemble via Parameter Selection

Although the *Refocus* stage is capable of aligning source domains and simulated target domains, it may also suffer from the problem of training instability. Instead of explicitly aligning the attention maps of source images and simulated target images, which may lead to over-fitting, we propose to ensemble parameters of the attention-refocuser to generate domain-invariant attention maps.

Considering that CLIP relies solely on class embedding in the last layer for classification, the attention map in the last layer contains much semantic information. Therefore, we use it to represent the focus of CLIP, denoted as $A = \mathrm{softmax}(\frac{\mathbf{Q}_B \mathbf{K}_B^T}{\sqrt{d_k}})$, where $\frac{1}{\sqrt{d_k}}$ serves as a scaling factor. We extract the attention maps A^s and A^t of to effectively represent the focus of CLIP on the source image x^s and the simulated target image x^t, respectively. Then, we use the task-relevant masks M^s and M^t calculated in Eq. (3) to refine the attention maps A^s and A^t, respectively, better representing the regions of interest in CLIP on the source image and the simulated target image. In this case, the refined attention maps are computed as

$$\hat{A}^s = A^s \odot M^s, \tag{9}$$

while the refined regions of interest $\hat{A}^t$ for the simulated target image x^t are calculated in the same way.

To stabilize the training procedure, we select the parameters θ of the decoder D_θ in the attention-refocuser that leads to high consistency between $\hat{A}^s$ and $\hat{A}^t$. The consistency is measured by the cosine similarity between the two attention maps, formalized as

$$s = \left\langle \hat{A}^s, \hat{A}^t \right\rangle. \tag{10}$$

We maintain an exponential updated score $\overline{s}$ of the similarity as a threshold to select the parameters θ. At the t-th step, when the average similarity s_t of all cosine similarities s in the current batch is larger than the threshold $\overline{s}$, we ensemble the current parameter θ_t of the decoder and simultaneously update the threshold using

$$\theta_a \leftarrow \begin{cases} \omega\theta_a + (1-\omega)\theta_t, & s_t > \overline{s} \\ \theta_a, & s_t \le \overline{s} \end{cases}, \tag{11}$$

$$\overline{s} \leftarrow \begin{cases} \omega\overline{s} + (1-\omega)s_t, & s_t > \overline{s} \\ \overline{s}, & s_t \le \overline{s} \end{cases}, \tag{12}$$

where $\overline{s}$ is initialized to zero, and ω is the update ratio for the ensembled parameter θ_a and the threshold $\overline{s}$. The ensembled parameter θ_a is initialized to all zeros. The ensembled decoder D_{θ_a} will be used in the inference stage.

E. Inference

During testing, each target sample x^t is fed into the image encoder f_v of CLIP to obtain its original visual embedding $z^{\mathcal{T}}$. Then the attention-refocuser adaptively adjusts attention through the task-relevant prompt σ and the ensembled decoder D_{θ_a} to produce the refocus parameters $e^{\mathcal{T}}$. The second forward process inputs the original test sample $x^{\mathcal{T}}$ and the refined parameters $E^{\mathcal{T}} = \left\{E, e^{\mathcal{T}}\right\}$, and outputs the refocused visual embedding $\hat{z}^{\mathcal{T}}$. Finally, $\hat{z}^{\mathcal{T}}$ is compared with text embeddings of all categories, and the category with the highest similarity is selected as the prediction result:

$$\hat{y} = \arg\max_{c \in \mathcal{Y}} p(c|x^{\mathcal{T}}), \tag{13}$$

where $p(c|x^{\mathcal{T}})$ is calculated according to Eq. (5).

IV. Experiments

A. Datasets

We conduct experiments on five commonly used datasets: VLCS [53], PACS [54], OfficeHome [55], TerraIncognita [56] and DomainNet [57]. VLCS consists of 10,729 images sourced from five distinct classes originating from four separate datasets: PASCAL VOC 2007 [58], LabelMe [59], Caltech [60], and Sun [61]. PACS contains 9,991 images classified into seven categories, exhibiting significant distribution shifts across four domains: art painting, cartoon, photo, and sketch. OfficeHome contains 15,500 images across 65 categories in both office and home environments. These categories in OfficeHome are distributed among four domains: Art, Clipart, Product, and Real-World, characterized by distinct viewpoints and image styles. TerraIncognita (Terra) contains 24,788 animal images captured in the wild across different locations. It encompasses a total of ten animal classes and each location is treated as a distinct domain, denoted as Location100, Location38, Location43, and Location46. DomainNet contains 586,575 images across 345 classes, categorized into six types of domains: clipart, infograph, painting, quickdraw, real, and sketch.

TABLE I
COMPARISON RESULTS BETWEEN SRE AND OTHER CLIP-BASED METHODS ON PACS, VLCS, OFFICEHOME, TERRA, AND DOMAINNET FOR MULTI-SOURCE DOMAIN GENERALIZATION USING DIFFERENT BACKBONES. THE RESULTS OF THESE COMPARED METHODS ARE COPIED BY DEFAULT FROM THEIR ORIGINAL PAPERS. "*" DENOTES OUR REPRODUCED RESULTS, "†" DENOTES RESULTS COPIED FROM SPG [63].

Method	PACS	VLCS	Off.H.	Terra	Dom.N.	Avg.
ViT-32/B						
ZS-CLIP [7]	94.9	80.9	78.9	22.6	53.9	66.2
VPT* [44]	**95.1**	81.9	80.3	40.3	50.8	69.7
MaPLe* [64]	94.6	81.6	81.3	41.5	55.8	70.9
SRE (Ours)	**95.1**	**82.8**	**82.5**	**49.0**	**57.3**	**73.3**
ViT-16/B						
ERM [65]	92.9	81.4	78.9	53.6	56.1	72.6
DANN [66]	92.2	80.1	78.0	47.9	57.5	71.1
CORAL [67]	92.6	79.6	78.5	51.7	56.4	71.8
MIRO [68]	95.2	81.1	82.5	52.9	56.6	73.7
ZS-CLIP [7]	96.1	82.5	82.1	33.9	57.5	70.4
DPL† [42]	96.4	80.9	83.0	46.6	59.5	73.6
VPT† [44]	96.9	82.0	83.2	46.7	58.5	73.6
CSVPT [9]	96.6	82.7	85.5	57.7	59.8	76.5
MaPLe† [64]	96.5	82.2	83.4	50.2	59.5	74.4
SPG [63]	97.0	82.4	83.6	50.2	60.1	74.7
CLIP-Adapter [69]	96.4	84.3	82.2	–	59.9	–
Gallop* [70]	96.6	83.8	86.0	51.7	61.8	76.0
CLIPCEIL [71]	**97.6**	**88.4**	85.4	53.0	**62.0**	77.3
SRE (Ours)	97.0	83.7	**86.1**	**60.8**	61.8	**77.9**
ViT-14/L						
ZS-CLIP [7]	98.3	81.9	86.6	43.7	63.0	74.7
VPT* [44]	98.0	83.2	89.4	60.4	63.9	79.0
CSVPT [9]	98.5	83.0	**90.9**	**65.3**	65.3	80.6
MaPLe* [64]	98.3	81.9	89.0	57.0	63.5	77.9
SRE (Ours)	**98.7**	**84.3**	**90.9**	64.8	**67.0**	**81.1**

B. Implementation Details

We use ViT-B/16 [45], ViT-B/32 [45] and ViT-L/14 [45] as the backbone for CLIP. We train the parameters g of the attention-refocuser via 2000 iterations for PACS, VLCS, and OfficeHome, while 5000 iterations for Terra and DomainNet. The AdamW optimizer is adopted. The learning rate is set to 0.0004 for the decoder parameter θ and 0.001 for the task-relevant prompt σ, and the weight decay is set to 0.005 across all datasets. θ is randomly initialized while σ is initialized to constant 1 for all positions. The hyperparameters ω and λ are set to 0.98 and 0.1, respectively, across all datasets. We opt for a mini-batch size of 16 and perform four gradient accumulation steps during training to achieve a larger batch size. It's noteworthy that we adhere to the training-domain validation strategy [62] that utilizes subsets from each training domain for validation to select the best-performing model. We split the data of each training domain into 80% for training and 20% for validation, subsequently combining the validation sets of all domains for model selection. All experiments are conducted on a single NVIDIA RTX 4090 GPU with 24GB of graphics memory.

C. Comparison with CLIP-based methods

We conduct a comparative analysis of our method (SRE) with CLIP-based methods, including the baseline zero-shot CLIP (ZS-CLIP) [7], parameter-efficient tuning methods (DPL [42], VPT [44], CSVPT [9], MaPLe [64], SPG [63], CLIP-Adapter [69], Gallop [70], CLIPCEIL [71]), representation learning methods (DANN [66], CORAL [67], MIRO [68]) with their baseline ERM [65]. The comparison results on the five datasets are shown in Table I. We have the following observations. Our method achieves the overall best performance on almost all backbones, which validates the effectiveness of the proposed SRE scheme for domain generalization. While CLIPCEIL shows strong results on VLCS, our method's significant lead on the challenging TerraIncognita benchmark (+7.8%) highlights its enhanced robustness against substantial domain shifts.

Table II-VI show the detailed comparison results of all domains on the PACS, VLCS, OfficeHome, Terra and DomainNet datasets, respectively, using the ViT-16/B backbone.

TABLE II
COMPARISON RESULTS BETWEEN SRE AND OTHER CLIP-BASED METHODS ON PACS UNDER THE LEAVE-ONE-DOMAIN-OUT SETTING. "*" DENOTES OUR REPRODUCED RESULTS, "†" DENOTES RESULTS COPIED FROM SPG [63]

Method	Art	Cartoon	Photo	Sketch	Avg.
ZS-CLIP [7]	97.1	**99.1**	**99.9**	88.1	96.1
DPL† [42]	97.8	98.5	**99.9**	89.5	96.4
VPT† [44]	97.9	98.9	**99.9**	91.0	96.9
MaPLe† [64]	97.9	98.7	99.7	89.8	96.5
SPG [63]	97.7	99.0	**99.9**	**91.3**	**97.0**
Gallop* [70]	**98.3**	98.9	**99.9**	89.3	96.6
SRE (Ours)	**98.3**	99.0	99.8	91.1	**97.0**

TABLE III
COMPARISON RESULTS BETWEEN SRE AND OTHER CLIP-BASED METHODS ON VLCS UNDER THE LEAVE-ONE-DOMAIN-OUT SETTING. "*" DENOTES OUR REPRODUCED RESULTS, "†" DENOTES RESULTS COPIED FROM SPG [63]

Method	Caltech	LabelMe	SUN	VOC	Avg.
ZS-CLIP [7]	99.8	**69.1**	75.5	85.7	82.5
DPL† [42]	99.8	61.5	77.8	84.6	80.9
VPT† [44]	**99.9**	64.8	78.2	85.2	82.0
MaPLe† [64]	98.3	64.8	80.6	85.1	82.2
SPG [63]	99.7	64.7	80.7	84.4	82.4
Gallop* [70]	99.8	68.8	79.3	**87.3**	**83.8**
SRE (Ours)	99.0	67.8	**81.5**	86.4	83.7

TABLE IV
COMPARISON RESULTS BETWEEN SRE AND OTHER CLIP-BASED METHODS ON OFFICEHOME UNDER THE LEAVE-ONE-DOMAIN-OUT SETTING. "*" DENOTES OUR REPRODUCED RESULTS, "†" DENOTES RESULTS COPIED FROM SPG [63]

Method	Art	Clipart	Product	Real	Avg.
ZS-CLIP [7]	82.1	67.9	89.1	89.2	82.1
DPL† [42]	81.0	71.2	90.0	89.6	83.0
VPT† [44]	80.9	72.5	89.0	90.4	83.2
MaPLe† [64]	81.6	72.6	90.2	89.0	83.4
SPG [63]	81.6	72.7	90.2	89.9	83.6
Gallop* [70]	85.9	**73.9**	**92.7**	91.5	86.0
SRE (Ours)	**86.8**	73.0	92.4	**92.1**	**86.1**

D. Single Domain Generalization

To investigate the effectiveness of our method on single domain generalization, we conduct comparative experiments on the OfficeHome [55] and TerraIncognita [56] datasets with other CLIP-based methods, including DPL [42], VPT [44] and MaPLe [64]. To measure the performance of a model on a certain domain, we train the model on this domain, test it on the left three domains, and average the results of the model on the three domains. As shown in Table VII and Table VIII,

TABLE V
COMPARISON RESULTS BETWEEN SRE AND OTHER CLIP-BASED METHODS ON TERRAINCOGNITA UNDER THE LEAVE-ONE-DOMAIN-OUT SETTING. "*" DENOTES OUR REPRODUCED RESULTS, "†" DENOTES RESULTS COPIED FROM SPG [63]

Method	L100	L38	L43	L46	Avg.
ZS-CLIP [7]	52.1	19.9	33.5	30.0	33.9
DPL† [42]	41.6	54.3	49.0	41.6	46.6
VPT† [44]	45.5	46.8	52.8	41.8	46.7
MaPLe† [64]	52.4	52.4	53.0	43.1	50.2
SPG [63]	49.8	51.0	49.2	**50.7**	50.2
Gallop* [70]	59.3	51.8	52.6	43.3	51.7
SRE (Ours)	**73.7**	**58.4**	**61.1**	49.8	**60.8**

TABLE VI
COMPARISON RESULTS BETWEEN SRE AND OTHER CLIP-BASED METHODS ON DOMAINNET UNDER THE LEAVE-ONE-DOMAIN-OUT SETTING. "*" DENOTES OUR REPRODUCED RESULTS, "†" DENOTES RESULTS COPIED FROM SPG [63]

Method	Clipart	Info	Paint	Quick	Real	Sketch	Avg.
ZS-CLIP [7]	71.2	48.1	66.3	14.0	83.4	63.2	57.7
DPL† [42]	72.5	50.4	68.3	15.8	83.9	66.0	59.5
VPT† [44]	71.0	48.5	66.2	16.3	83.6	65.2	58.5
MaPLe† [64]	73.1	49.9	67.8	16.6	83.5	65.9	59.5
SPG [63]	68.7	50.2	**73.2**	16.6	83.3	68.5	60.1
Gallop* [70]	76.1	**54.2**	70.9	17.3	**84.5**	68.0	**61.8**
SRE (Ours)	**76.2**	52.0	70.3	**18.9**	84.3	**68.9**	**61.8**

the proposed SRE achieves the highest average accuracy on both datasets.

E. Open-Set Domain Generalization

To further explore the potential of our method when the target domain exhibits semantic shifts with source domains, we conduct experiments on OfficeHome and TerraIncognita datasets under the Open-Set Domain Generalization (OSDG) setting. OSDG constructs a scenario where both data distribution shift and semantic shift exist between the target domain and the source domain, and the target domain contains unknown categories that do not belong to the source domain's category space. In this scenario, the model must not only accurately classify known class samples outside the source distribution, but also distinguish samples of unknown classes.

TABLE VII
COMPARISON RESULTS BETWEEN SRE AND OTHER CLIP-BASED METHODS ON OFFICEHOME UNDER SINGLE DOMAIN GENERALIZATION SETTING. "*" DENOTES OUR REPRODUCED RESULTS.

Method	Art	Clipart	Product	Real	Avg.
ZS-CLIP [7]	82.1	86.8	79.7	79.7	82.1
DPL* [42]	**83.7**	87.4	80.6	81.6	83.3
VPT* [44]	78.1	82.1	75.1	78.9	78.5
MaPLe* [64]	83.0	87.3	80.6	81.4	83.1
SRE (Ours)	83.1	**87.8**	**82.1**	**83.5**	**84.1**

TABLE VIII
COMPARISON RESULTS BETWEEN SRE AND OTHER CLIP-BASED METHODS ON TERRAINCOGNITA UNDER SINGLE DOMAIN GENERALIZATION SETTING. "*" DENOTES OUR REPRODUCED RESULTS.

Method	L100	L38	L43	L46	Avg.
ZS-CLIP [7]	27.8	38.5	34.0	35.2	33.9
DPL* [42]	42.4	29.8	40.9	50.3	40.9
VPT* [44]	34.7	23.0	42.8	46.5	36.5
MaPLe* [64]	41.2	33.5	44.8	46.6	41.5
SRE (Ours)	**45.1**	**36.0**	**55.1**	**52.9**	**47.3**

To make a fair comparison with other CLIP-based methods, we combine all methods with a common open-set recognition strategy that distinguishes unknown classes by setting a specific threshold for prediction probabilities. Following previous works, we choose three evaluation metrics to verify the performance: the close-set accuracy (Acc), the harmonic mean of known class accuracy and unknown class accuracy (H-score), and the open-set classification rate (OSCR), which plots the true positive rate against the false positive rate by a varying threshold.

Table IX and Table X show the results on both OfficeHome and TerraIncognita datasets, which validates the effectiveness of SRE in Open-Set Domain Generalization. SRE not only achieves the highest close-set accuracy (Acc), but also excels in H-score and OSCR, demonstrating its ability to accurately classify known classes while effectively identifying unknown classes. Compared to existing CLIP-based methods, SRE consistently outperforms baselines in handling domain shifts and semantic shifts, making it a robust solution in real-world applications where unknown categories are prevalent.

F. Ablation Studies

a) Results of different stages: To evaluate the effectiveness of each stage of our method, We compare our method with the following variants: (1) ZS-CLIP, where the original CLIP is used to perform zero-shot classification; (2) AR-CLIP, where only the attention-refocuser is incorporated into ZS-CLIP for fine-tuning; (3) SR, where the *Ensemble* stage is removed from the SRE scheme; (4) SR + EMA, where an exponential moving average strategy is added to SR to average all weights instead of selecting and combining parameters as done in the *Ensemble* stage.

Table XI shows the results of different modules on the five datasets. It is interesting to observe that: (1) AR-CLIP outperforms ZS-CLIP, which suggests that attention refocusing improves the generalization capability of CLIP; (2) SR outperforms AR-CLIP, which indicates that the simulated target domains and attention maps aligning boost the generalization of attention refocusing; (3) SR + EMA performs worse than SRE and similar to SR, which indicates that the performance improvement of the *Ensemble* stage is not benefited from simple weight average, but from adaptive selection to the robust parameters.

In addition, we explicitly compare our method with TOAST [18] on five datasets and report the results in Table XII. We observe that directly applying TOAST to fine-tune the CLIP image encoder can only achieve marginal performance gains. Our combination of the attention-refocuser and zero-shot classification, denoted as AR-CLIP, alleviates over-fitting to some extent. Most importantly, the proposed SRE greatly improves DG performance.

b) Results of different losses: We conduct ablation studies on different losses in Eq. 8. As shown in Table XIII, the combination of $\mathcal{L}_{cls}$ and $\mathcal{L}_{align}$ achieves significantly better performance than either loss alone, demonstrating that the

TABLE IX
COMPARISON RESULTS BETWEEN SRE AND OTHER CLIP-BASED METHODS ON OFFICEHOME UNDER THE OPEN-SET DOMAIN GENERALIZATION SETTING. THE RATIO OF KNOWN TO UNKNOWN CLASSES IS 50:15. "*" DENOTES OUR REPRODUCED RESULTS.

Method	Art			Clipart			Product			Real World			Avg		
	Acc	H-score	OSCR	Acc	H-score	OSCR	Acc	H-score	OSCR	Acc	H-score	OSCR	Acc	H-score	OSCR
ZS-CLIP [7]	84.1	75.9	76.6	70.3	65.9	61.8	90.0	80.2	83.5	90.6	82.2	85.4	83.7	76.1	76.8
DPL* [42]	85.0	74.5	75.5	73.7	66.7	64.0	**92.9**	**81.9**	**86.2**	91.8	82.6	85.9	85.9	76.4	77.9
VPT* [44]	83.8	75.3	75.8	74.1	67.1	64.8	91.4	81.1	85.0	91.3	81.7	85.6	85.1	76.3	77.8
MaPLe* [64]	84.7	75.1	76.4	73.3	66.1	63.3	92.1	81.7	85.9	92.3	**83.4**	86.9	85.6	76.6	78.1
SRE (Ours)	**87.5**	**77.5**	**79.7**	**75.7**	**67.2**	**65.5**	**92.9**	81.8	**86.2**	**93.0**	83.1	**87.9**	**87.3**	**77.4**	**79.8**

TABLE X
COMPARISON RESULTS BETWEEN SRE AND OTHER CLIP-BASED METHODS ON TERRAINCOGNITA UNDER THE OPEN-SET DOMAIN GENERALIZATION SETTING. THE RATIO OF KNOWN TO UNKNOWN CLASSES IS 8:2. "*" DENOTES OUR REPRODUCED RESULTS.

Method	Location 100			Location 38			Location 43			Location 46			Avg		
	Acc	H-score	OSCR	Acc	H-score	OSCR	Acc	H-score	OSCR	Acc	H-score	OSCR	Acc	H-score	OSCR
ZS-CLIP [7]	44.2	44.8	36.3	26.1	29.3	19.4	38.8	40.5	31.2	32.6	28.8	19.5	34.3	34.3	25.1
DPL* [42]	64.4	54.1	46.7	61.0	55.5	49.3	56.9	49.3	42.4	47.6	43.6	34.5	57.5	50.6	43.3
VPT* [44]	61.1	48.3	40.6	55.3	53.3	45.4	56.7	49.4	41.1	50.9	46.2	37.8	56.0	49.3	41.2
MaPLe* [64]	68.4	55.7	51.3	61.3	55.9	49.6	58.6	51.8	44.6	48.1	43.0	34.3	59.1	51.6	45.0
SRE (Ours)	**75.8**	**68.3**	**65.2**	**64.5**	**58.0**	**53.1**	**61.3**	**59.6**	**53.7**	**54.9**	**53.0**	**45.2**	**64.1**	**59.7**	**54.3**

TABLE XI
RESULTS OF DIFFERENT STAGES ON PACS, VLCS, OFFICEHOME, TERRA, AND DOMAINNET WITH VIT-B/16 AS THE BACKBONE. "SR + EMA" DENOTES REPLACING THE ENSEMBLE STAGE WITH THE EXPONENTIAL MOVING AVERAGE (EMA), WHICH SERVES AS A BASELINE FOR THE PARAMETER ENSEMBLING.

Method	Refocus	Simulate	Ensemble	PACS	VLCS	OfficeHome	Terra	DomainNet	Avg.
ZS-CLIP [7]				96.1	82.5	82.1	33.9	57.5	70.4
AR-CLIP	✓			95.9	82.8	83.6	54.3	59.7	75.3
SR	✓	✓		96.3	83.3	84.3	59.8	61.2	77.0
SR + EMA	✓	✓		96.1	83.5	84.6	60.0	61.6	77.1
SRE (Ours)	✓	✓	✓	**97.0**	**83.7**	**86.1**	**60.8**	**61.8**	**77.9**

TABLE XII
COMPARISON RESULTS BETWEEN TOAST [18] AND OUR METHOD ON PACS, VLCS, OFFICEHOME, TERRA, AND DOMAINNET WITH VIT-B/16 AS THE BACKBONE.

Method	PACS	VLCS	Off.H.	Terra	Dom.N.	Avg.
TOAST [18]	95.9	82.1	83.8	44.0	58.6	72.9
AR-CLIP	95.9	82.8	83.6	54.3	59.7	75.3
SRE (Ours)	**97.0**	**83.7**	**86.1**	**60.8**	**61.8**	**77.9**

TABLE XIII
RESULTS OF DIFFERENT LOSS ON VLCS, OFFICEHOME, TERRA WITH VIT-B/16 AS THE BACKBONE.

Method	$\mathcal{L}_{cls}$	$\mathcal{L}_{align}$	$\mathcal{L}_{var}$	VLCS	Off.H.	Terra	Avg.
ZS-CLIP [7]				82.5	82.1	33.9	70.4
+ $\mathcal{L}_{cls}$	✓			82.8	83.6	54.3	73.6
+ $\mathcal{L}_{cls}, \mathcal{L}_{var}$	✓		✓	83.5	83.2	55.3	74.0
+ $\mathcal{L}_{align}$		✓		83.0	82.1	54.1	73.1
+ $\mathcal{L}_{align}, \mathcal{L}_{var}$		✓	✓	83.4	82.2	54.9	73.5
+ $\mathcal{L}_{cls}, \mathcal{L}_{align}$	✓	✓		**83.7**	85.9	60.1	76.5
SR (Ours)	✓	✓	✓	**83.7**	**86.1**	**60.8**	**76.9**

design in the *Simulate* stage can better align the source domain and simulation target domain. The introduction of $\mathcal{L}_{var}$ further improves the robustness, especially in challenging domains like Terra (+2.0% with $\mathcal{L}_{cls}+\mathcal{L}_{var}$). Our full model integrating all three losses achieves the best performance (75.8% average, 59.8% on Terra), verifying that $\mathcal{L}_{var}$ effectively stabilizes the joint optimization process while preserving alignment benefits.

c) Analysis on the Simulate *stage:* We analyze the impact of using different data augmentation strategies in the *Simulate* stage. We use the combination of used augmentation (ColorJitter, GaussianBlur, and GrayScale) as a baseline method. We explore two types of variants: (1) Incremental variants (adding Polarize, Equalize, Solarize, or their combination to the baseline); (2) Alternative variants (replacing the entire baseline with Fourier Transform [72] or Random Convolution [73]). Table XII presents the results on the VLCS, OfficeHome, and TerraIncognita datasets, where the baseline method achieves stable cross-domain performance. Incremental variants slightly improve the accuracy on VLCS but degrade the performance on TerraIncognita, suggesting that excessive enhancement may hurt feature extraction. Alternative variants exhibit task-dependent effectiveness: Fourier transform performs well on OfficeHome but poorly on TerraIncognita, while random convolution reduces performance across all datasets. These results show that adding more augmentation strategies does not bring a larger overall performance improvement. Our choice reflects a systematic trade-off between simplicity and performance.

To further investigate the generalization ability of our model on unseen simulated target domains using the three augmentation strategies, we apply other augmentations to the target domains to emulate a broader spectrum of distribution shifts. As shown in Table XIII, we add augmentation strategies,

TABLE XIII
RESULTS OF OUR METHOD ON VLCS, OFFICEHOME, AND TERRAINCOGNITA WHEN ADDING DIFFERENT DATA AUGMENTATION STRATEGIES ON TARGET DOMAIN TO SIMULATE VARIOUS DOMAIN SHIFTS. "ALL" DENOTES COMBINING POLARIZE, EQUALIZE, AND SOLARIZE TOGETHER.

Method	Equalize			Polarize			Solarize			ALL			Fourier			RandConv		
	VLCS	Off.H.	Terra	VLCS	Off.H.	Terra	VLCS	Off.H.	Terra	VLCS	Off.H.	Terra	VLCS	Off.H.	Terra	VLCS	Off.H.	Terra
ZS-CLIP [7]	82.2	80.2	34.5	82.5	81.0	27.5	81.7	79.2	33.8	81.4	76.5	27.5	82.6	75.9	31.4	80.1	72.0	21.8
VPT* [44]	82.3	81.9	44.8	82.2	82.6	38.2	81.7	80.9	44.6	81.7	78.5	35.7	82.7	78.1	40.2	80.0	75.7	26.3
MaPLe* [64]	82.2	82.3	46.3	82.0	83.2	39.9	82.2	81.3	45.9	82.0	78.8	37.5	82.8	78.7	42.9	81.5	75.7	29.2
SRE (Ours)	**83.6**	**84.6**	**59.7**	**83.7**	**85.2**	**48.6**	**83.4**	**84.0**	**59.1**	**83.3**	**81.8**	**46.8**	**83.3**	**81.7**	**56.7**	**81.8**	**78.9**	**42.4**

including Equalize, Polarize, Solarize, and their combination ("ALL"), as well as Fourier Transform and Random Convolution, to the test set of VLCS, OfficeHome, and TerraIncognita. The results demonstrate that our method (SRE) consistently outperforms all compared methods under all simulated shifts. These findings underscore that SRE exhibits better generalization across diverse domain shifts, including those far beyond the original augmentation (e.g., Fourier transform and random convolution), suggesting its potential applicability to real-world domain shifts.

TABLE XII
RESULTS OF OUR METHOD ON VLCS, OFFICEHOME, AND TERRAINCOGNITA WHEN USING DIFFERENT DATA AUGMENTATION STRATEGIES. "CJ", "GB", AND "GS" DENOTE COLORJITTER, GAUSSIANBLUR, AND GRAYSCALE, RESPECTIVELY. "ALL" DENOTES COMBINING POLARIZE, EQUALIZE, AND SOLARIZE TOGETHER. THE ROW IN GRAY DENOTES THE AUGMENTATION STRATEGIES USED IN SRE.

Category	Augmentation Variant	VLCS	Off.H.	Terra
ZS-CLIP [7]	None	82.5	82.1	33.9
Incremental	Base (CJ+GB+GS)	83.7	**86.1**	**60.8**
	Base + Polarize	**83.9**	86.0	60.0
	Base + Equalize	**83.9**	86.0	59.5
	Base + Solarize	83.6	**86.1**	59.7
	Base + All	83.8	86.0	59.4
Alternative	Fourier Transform	83.7	85.7	58.3
	Random Convolution	83.7	**86.1**	58.6

d) Ablation on the Ensemble *stage:* We conducted additional ablation studies on the four domains of the OfficeHome dataset to evaluate the effectiveness of the *Ensemble* stage. Figure 5 shows the test accuracy of SRE and SR (without the *Ensemble* stage) at intervals of 100 training steps. The results show that SRE demonstrates better generalization capability compared to SR and significantly stabilizes the training process. In the early training stage, SRE performs worse than SR because the ensembled parameters are initialized to zero. This design ensures that the ensembled parameters select the weights of models once the training process is more stable.

e) Results of different data scales: To evaluate our method in the scenario of insufficient training data, we conduct experiments on the OfficeHome dataset with different training data scales and record the average accuracy across four domains. We gradually decrease the percentage of available training data from 80% to 10%. We compare our method with ZS-CLIP [7], VPT [44] and MaPLe [64], and report the results in Figure 6. From the results, we notice that our method outperforms VPT, MaPLE, and ZS-CLIP in all percentages of available training data. Notably, when only 10% of the data is available for training, our method still achieves

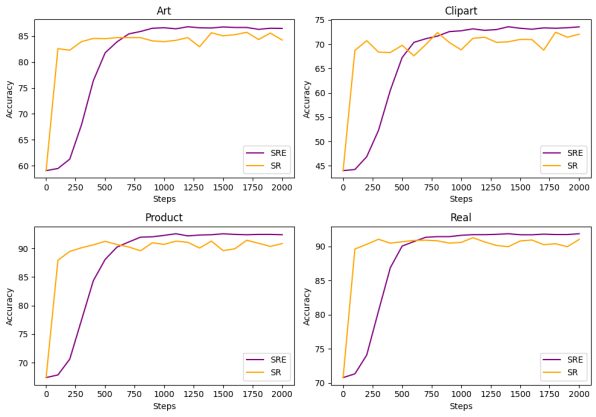

Fig. 5. Comparison of the test accuracy between SRE and its variant SR on the four target domains of the OfficeHome dataset.

performance gains compared to ZS-CLIP, while VPT suffers from overfitting and performs worse generalization ability than ZS-CLIP. The results verify the capability of our method on insufficient training data.

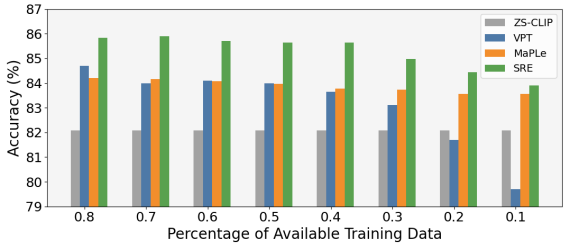

Fig. 6. Results of percentages of available training data on OfficeHome.

f) Sensitivity to hyperparameters: We investigate how sensitive our method is to the weight λ of the variance loss and the update ratio ω by changing their values and reporting the corresponding results. Figure 7 shows the average accuracy over three runs on OfficeHome. The results indicate that our method is insensitive to both λ and ω.

G. Visualization Study

To further demonstrate the capability of refocusing attention of our method, we visualize CLIP's attention using Grad-CAM [74] in the form of heatmaps. As shown in Figure 8, our method focuses better on task-relevant regions while paying less attention to task-irrelevant regions, especially on images with noisy backgrounds, which verifies the effectiveness of our method to refocus the attention in CLIP.

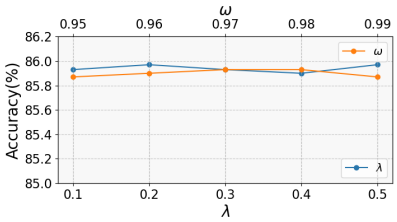

Fig. 7. Results of different weights λ of the variance loss and different update ratios ω for ensembling parameter θ_a on OfficeHome.

To further demonstrate the effectiveness of the proposed *Simulate, Refocus and Ensemble* scheme, we present visualization results of attention maps of our method and several variants in Figure 9. We observe that our method performs better than SR in focusing on task-relevant regions and capturing visual details of both easy and difficult images, which indicates the effectiveness of the *Ensemble* stage. Moreover, SR performs better than ZS-CLIP, which suggests that the *Simulate* and the *Refocus* stages improve the focus of CLIP on task-relevant regions.

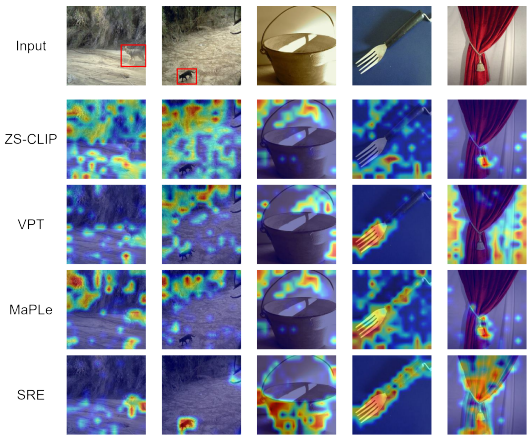

Fig. 8. Attention visualization of SRE and several existing methods.

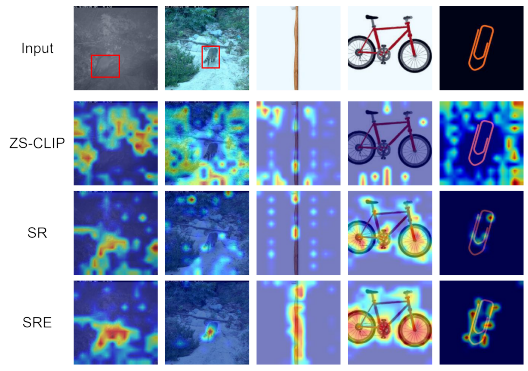

Fig. 9. Attention visualization of SRE and several variants.

H. Results on the NICO Challenge

We also conduct experiments on the NICO++ [75] dataset, as shown in Table XIV. Unlike traditional DG datasets, NICO++ focuses more on significant distribution shifts in real-world scenarios and recommends evaluating the model's generalization performance across multiple target domains. Environmental changes primarily cause the distribution shifts across different domains in NICO++. The NICO++ dataset consists of six domains: Autumn, Rock, Dim, Grass, Outdoor, and Water. Following the guidelines in [75], we categorize these six domains into three groups based on environmental similarity: Autumn-Rock, Dim-Grass, and Outdoor-Water. We train the model on two groups while leaving the third group out for testing, which is analogous to the leave-one-domain-out setting. For comparison, we reproduce three methods: DPL [42], VPT [44] and MaPLe [64]. Additionally, ZS-CLIP [7] represents the zero-shot CLIP approach, which uses the template "a photo of a [class]" for text prompts.

TABLE XIV
COMPARISON RESULTS BETWEEN SRE AND OTHER CLIP-BASED METHODS ON NICO++ FOR MULTI-SOURCE DOMAIN GENERALIZATION. "*" DENOTES OUR REPRODUCED RESULTS.

Method	Autumn	Rock	Dim	Grass	Outdoor	Water	Avg.
ZS-CLIP [7]	89.0	89.8	87.9	90.4	85.8	82.5	87.6
DPL* [42]	91.1	92.0	**90.6**	92.6	90.5	85.2	90.3
VPT* [44]	89.6	91.9	89.0	92.6	90.1	85.6	89.8
MaPLe* [64]	90.7	92.2	89.8	92.5	90.4	85.8	90.2
SRE (Ours)	**91.4**	**92.3**	90.4	**93.2**	**90.8**	**86.4**	**90.8**

I. Complexity Analysis

We compare the complexity of SRE with other CLIP-based methods on PACS (7 classes) and OfficeHome (65 classes) under the same mini-batch size of 16 for training. As shown in Table XV, while SRE requires higher memory usage and slightly longer latency during training compared to methods like VPT [44] and MaPLe [64], its resource requirements are still lower than DPL [42]. These additional resources are essential for enabling SRE to focus on domain-invariant regions, leading to superior generalization performance. Furthermore, compared to methods like DPL [42] and SPG [63], where the computational cost increases with the number of categories during training, our method achieves a better trade-off between generalization and computational cost. During testing, SRE demonstrates a balanced profile by requiring less memory to methods such as DPL [42], while maintaining a competitive latency of 3.1 ms/img.

We also compare the total parameters and trained parameters of SRE with other CLIP-based methods. As shown in Table XVI, SRE has 164.4M parameters, slightly higher than methods like DPL (159.6M) and MaPLe (154.4M). This is due to the introduced attention-refocuser, which refocuses the attention of CLIP on domain-invariant features. SRE trains 14.7M parameters, more than prompt-based methods (e.g., VPT: 0.1M, MaPLe: 4.8M). This reflects our design philosophy: instead of limiting updates to prompts alone, we introduce an extra network to adaptively adjust the attention of CLIP. While this increases trainable parameters, it ensures richer generalization ability, as evidenced by the performance gains in Tables I-VI. Furthermore, SRE remains far more parameter-efficient than full fine-tuning approaches (e.g., 149.6M trained parameters for full tuning).

TABLE XV
MODEL COMPLEXITY COMPARISON BETWEEN DIFFERENT CLIP-BASED METHODS USING VIT-16/B.

Method	Train		Test	
	Memory	Latency	Memory	Latency
PACS (7 classes)				
ZS-CLIP [7]	0 G	0 ms/img	2.1 G	1.5 ms/img
DPL [42]	13.4 G	3.6 ms/img	4.1 G	3.3 ms/img
VPT [44]	5.2 G	1.6 ms/img	2.7 G	1.5 ms/img
MaPLe [64]	5.5 G	1.8 ms/img	2.9 G	1.5 ms/img
SPG [63]	5.1 G	16.7 ms/img	2.1 G	4.4 ms/img
SRE (Ours)	11.7 G	5.2 ms/img	3.5 G	3.1 ms/img
OfficeHome (65 classes)				
ZS-CLIP [7]	0 G	0 ms/img	2.1 G	1.5 ms/img
DPL [42]	80.6 G	26.4 ms/img	15.9 G	17.5 ms/img
VPT [44]	5.3 G	1.6 ms/img	2.7 G	1.5 ms/img
MaPLe [64]	6.4 G	1.9 ms/img	2.9 G	1.7 ms/img
SPG [63]	5.1 G	87.5 ms/img	2.1 G	5.1 ms/img
SRE (Ours)	11.7 G	5.2 ms/img	3.5 G	3.1 ms/img

TABLE XVI
MODEL PARAMETERS COMPARISON BETWEEN DIFFERENT CLIP-BASED METHODS USING VIT-16/B.

Method	Total Parameters	Trained Parameters
ZS-CLIP [7]	149.6 M	0.0 M
DPL [42]	159.6 M	10.0 M
VPT [44]	149.7 M	0.1 M
MaPLe [64]	154.4 M	4.8 M
SPG [63]	154.3 M	4.7 M
SRE (Ours)	164.4 M	14.7 M

V. CONCLUSION AND LIMITATION

We present an attention-refocusing scheme, named *Simulate, Refocus and Ensemble (SRE)*, for domain generalization, which can adaptively refocus CLIP's attention on task-relevant regions for unseen domains and effectively reduce training instability. The *Simulate* stage can well mimic diverse domain shifts, the *Refocus* stage learns to reduce domain shifts by aligning the attention maps of the source data and simulated target data in CLIP, and the *Ensemble* stage can effectively enhance the efficiency of attention refocusing by selecting insensitive parameters. The above designs have been validated via extensive experiments across multiple domain generalization datasets.

Although SRE has achieved promising results, it has larger model complexity and higher computational cost compared to conventional prompt-tuning methods. This may hinder its deployment in resource-constrained environments. To mitigate this issue, future work will explore optimization strategies such as memory optimization, model pruning, and lightweight architecture design to improve efficiency without significantly compromising performance. In addition, we will investigate techniques like knowledge distillation or quantization to further reduce computational overhead.

ACKNOWLEDGMENT

This work was supported by the Shenzhen Science and Technology Program under Grant No. JCYJ20241202130548062, the Natural Science Foundation of Shenzhen under Grant No. JCYJ20230807142703006, and the Key Research Platforms and Projects of the Guangdong Provincial Department of Education under Grant No.2023ZDZX1034.

REFERENCES

[1] J. Chen, Z. Gao, X. Wu, and J. Luo, "Meta-causal learning for single domain generalization," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023, pp. 7683–7692.

[2] F. Lv, J. Liang, S. Li, B. Zang, C. H. Liu, Z. Wang, and D. Liu, "Causality inspired representation learning for domain generalization," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 8046–8056.

[3] S. Shankar, V. Piratla, S. Chakrabarti, S. Chaudhuri, P. Jyothi, and S. Sarawagi, "Generalizing across domains via cross-gradient training," in *International Conference on Learning Representations (ICLR)*, 2018.

[4] D. Mahajan, S. Tople, and A. Sharma, "Domain generalization using causal matching," in *International Conference on Machine Learning (ICML)*, 2021, pp. 7313–7324.

[5] K. Zhou, Y. Yang, T. Hospedales, and T. Xiang, "Learning to generate novel domains for domain generalization," in *Proceedings of the European Conference on Computer Vision (ECCV)*. Springer, 2020, pp. 561–578.

[6] H. Nam, H. Lee, J. Park, W. Yoon, and D. Yoo, "Reducing domain gap by reducing style bias," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 8690–8699.

[7] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, "Learning transferable visual models from natural language supervision," in *International Conference on Machine Learning (ICML)*. PMLR, 2021, pp. 8748–8763.

[8] J. Cho, G. Nam, S. Kim, H. Yang, and S. Kwak, "Promptstyler: Prompt-driven style generation for source-free domain generalization," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023, pp. 15 656–15 666.

[9] A. Li, L. Zhuang, S. Fan, and S. Wang, "Learning common and specific visual prompts for domain generalization," in *Proceedings of the Asian Conference on Computer Vision (ACCV)*, 2022, pp. 4260–4275.

[10] H. Niu, H. Li, F. Zhao, and B. Li, "Domain-unified prompt representations for source-free domain generalization," *arXiv preprint arXiv:2209.14926*, 2022.

[11] W. Tu, W. Deng, and T. Gedeon, "A closer look at the robustness of contrastive language-image pre-training (clip)," in *Proceedings of the 37th International Conference on Neural Information Processing Systems (NeurIPS)*, 2024.

[12] S. Zhao, X. Wang, L. Zhu, and Y. Yang, "Test-time adaptation with CLIP reward for zero-shot generalization in vision-language models," in *International Conference on Learning Representations (ICLR)*, 2024.

[13] D. Cheng, Y. Hu, N. Wang, D. Zhang, and X. Gao, "Achieving plasticity-stability trade-off in continual learning through adaptive orthogonal projection," *IEEE Transactions on Circuits and Systems for Video Technology (TCSVT)*, 2025.

[14] D. Cheng, Y. Li, D. Zhang, N. Wang, J. Sun, and X. Gao, "Progressive negative enhancing contrastive learning for image dehazing and beyond," *IEEE Transactions on Multimedia (TMM)*, 2024.

[15] D. Cheng, Y. Ji, D. Gong, Y. Li, N. Wang, J. Han, and D. Zhang, "Continual all-in-one adverse weather removal with knowledge replay on a unified network structure," *IEEE Transactions on Multimedia (TMM)*, 2024.

[16] Z. Sun, Y. Fang, T. Wu, P. Zhang, Y. Zang, S. Kong, Y. Xiong, D. Lin, and J. Wang, "Alpha-clip: A clip model focusing on wherever you want," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024, pp. 13 019–13 029.

[17] Q. Phung, S. Ge, and J.-B. Huang, "Grounded text-to-image synthesis with attention refocusing," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024, pp. 7932–7942.

[18] B. Shi, S. Gai, T. Darrell, and X. Wang, "Toast: Transfer learning via attention steering," *arXiv preprint arXiv:2305.15542*, vol. 5, no. 7, p. 13, 2023.

[19] Y. Wang, W. Zhang, J. Zheng, and C. Jin, "Primecomposer: Faster progressively combined diffusion for image composition with attention steering," in *Proceedings of the 32nd ACM International Conference on Multimedia (ACM MM)*, 2024, pp. 10 824–10 832.

[20] J. Yoo, T. J. Jun, and Y.-H. Kim, "xecgnet: Fine-tuning attention map within convolutional neural network to improve detection and explainability of concurrent cardiac arrhythmias," *Computer Methods and Programs in Biomedicine (CMPB)*, vol. 208, p. 106281, 2021.

[21] Z. Wang, Y. Luo, R. Qiu, Z. Huang, and M. Baktashmotlagh, "Learning to diversify for single domain generalization," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021, pp. 834–843.

[22] F. Qiao, L. Zhao, and X. Peng, "Learning to learn single domain generalization," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 12 556–12 565.

[23] W. Huang, Y. Shi, Z. Xiong, and X. X. Zhu, "Representation enhancement-stabilization: Reducing bias-variance of domain generalization," in *Proceedings of the European Conference on Computer Vision (ECCV)*. Springer, 2024, pp. 108–125.

[24] S.-Y. Jo and S. W. Yoon, "Poem: polarization of embeddings for domain-invariant representations," in *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, vol. 37, no. 7, 2023, pp. 8150–8158.

[25] K. Akuzawa, Y. Iwasawa, and Y. Matsuo, "Adversarial invariant feature learning with accuracy constraint for domain generalization," in *European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML PKDD)*. Springer, 2020, pp. 315–331.

[26] Y. Luo, J. Feinglass, T. Gokhale, K.-C. Lee, C. Baral, and Y. Yang, "Grounding stylistic domain generalization with quantitative domain shift measures and synthetic scene images," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024, pp. 7303–7313.

[27] Z. Huang, H. Wang, E. P. Xing, and D. Huang, "Self-challenging improves cross-domain generalization," in *Proceedings of the European Conference on Computer Vision (ECCV)*. Springer, 2020, pp. 124–140.

[28] Y. Shi, J. Seely, P. Torr, S. N, A. Hannun, N. Usunier, and G. Synnaeve, "Gradient matching for domain generalization," in *International Conference on Learning Representations (ICLR)*, 2022.

[29] X. Chu, Y. Jin, W. Zhu, Y. Wang, X. Wang, S. Zhang, and H. Mei, "Dna: Domain generalization with diversified neural averaging," in *International Conference on Machine Learning (ICML)*. PMLR, 2022, pp. 4010–4034.

[30] P. Wang, Z. Zhang, Z. Lei, and L. Zhang, "Sharpness-aware gradient matching for domain generalization," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023, pp. 3769–3778.

[31] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," in *North American Chapter of the Association for Computational Linguistics (ACL)*, 2019.

[32] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 10 684–10 695.

[33] Z. Wang, J. Liang, L. Sheng, R. He, Z. Wang, and T. Tan, "A hard-to-beat baseline for training-free clip-based adaptation," *arXiv preprint arXiv:2402.04087*, 2024.

[34] B. Tang, J. Zhang, L. Yan, Q. Yu, L. Sheng, and D. Xu, "Data-free generalized zero-shot learning," in *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, vol. 38, no. 6, 2024, pp. 5108–5117.

[35] D. Cheng, Z. Xu, X. Jiang, N. Wang, D. Li, and X. Gao, "Disentangled prompt representation for domain generalization," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024, pp. 23 595–23 604.

[36] Y. Xing, Q. Wu, D. Cheng, S. Zhang, G. Liang, P. Wang, and Y. Zhang, "Dual modality prompt tuning for vision-language pre-trained model," *IEEE Transactions on Multimedia (TMM)*, vol. 26, pp. 2056–2068, 2023.

[37] J. Tang, G. McGoldrick, M. Al-Ghossein, and C.-W. Chen, "Captions are worth a thousand words: Enhancing product retrieval with pretrained image-to-text models," *arXiv preprint arXiv:2402.08532*, 2024.

[38] B. Yang, Y. Dai, X. Cheng, Y. Li, A. Raza, and Y. Zou, "Embracing language inclusivity and diversity in clip through continual language learning," in *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, vol. 38, no. 6, 2024, pp. 6458–6466.

[39] M. Tschannen, B. Mustafa, and N. Houlsby, "Clippo: Image-and-language understanding from pixels only," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023, pp. 11 006–11 017.

[40] J. Merullo, L. Castricato, C. Eickhoff, and E. Pavlick, "Linearly mapping from image to text space," in *International Conference on Learning Representations (ICLR)*, 2023.

[41] L. He, D. Cheng, N. Wang, and X. Gao, "Exploring homogeneous and heterogeneous consistent label associations for unsupervised visible-infrared person reid," *International Journal of Computer Vision (IJCV)*, pp. 1–20, 2024.

[42] X. Zhang, Y. Iwasawa, Y. Matsuo, and S. S. Gu, "Amortized prompt: Lightweight finetuning for clip in domain generalization," *arXiv preprint arXiv:2111.12853*, vol. 2, no. 3, p. 5, 2021.

[43] S. Bose, A. Jha, E. Fini, M. Singha, E. Ricci, and B. Banerjee, "Stylip: Multi-scale style-conditioned prompt learning for clip-based domain generalization," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2024, pp. 5542–5552.

[44] M. Jia, L. Tang, B.-C. Chen, C. Cardie, S. Belongie, B. Hariharan, and S.-N. Lim, "Visual prompt tuning," in *Proceedings of the European Conference on Computer Vision (ECCV)*. Springer, 2022, pp. 709–727.

[45] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," in *International Conference on Learning Representations (ICLR)*, 2021.

[46] A. Vaswani, N. M. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proceedings of the 31st International Conference on Neural Information Processing Systems (NeurIPS)*, 2017.

[47] Y. Zhong, J. Yang, P. Zhang, C. Li, N. Codella, L. H. Li, L. Zhou, X. Dai, L. Yuan, Y. Li *et al.*, "Regionclip: Region-based language-image pretraining," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 16 793–16 803.

[48] X. Xu, T. Xiong, Z. Ding, and Z. Tu, "Masqclip for open-vocabulary universal image segmentation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023, pp. 887–898.

[49] J. Xu, S. Liu, A. Vahdat, W. Byeon, X. Wang, and S. De Mello, "Open-vocabulary panoptic segmentation with text-to-image diffusion models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023, pp. 2955–2966.

[50] Q. Zhang, C. Singh, L. Liu, X. Liu, B. Yu, J. Gao, and T. Zhao, "Tell your model where to attend: Post-hoc attention steering for llms," *arXiv preprint arXiv:2311.02262*, 2023.

[51] X. Yao, H. Qian, X. Hu, G. Xu, and Y. Liu, "Theoretical insights into fine-tuning attention mechanism: Generalization and optimization," *arXiv preprint arXiv:2410.02247*, 2024.

[52] R. Meng, X. Li, W. Chen, S. Yang, J. Song, X. Wang, L. Zhang, M. Song, D. Xie, and S. Pu, "Attention diversification for domain generalization," in *Proceedings of the European Conference on Computer Vision (ECCV)*. Springer, 2022, pp. 322–340.

[53] C. Fang, Y. Xu, and D. N. Rockmore, "Unbiased metric learning: On the utilization of multiple datasets and web images for softening bias," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2013, pp. 1657–1664.

[54] D. Li, Y. Yang, Y.-Z. Song, and T. M. Hospedales, "Deeper, broader and artier domain generalization," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 5542–5550.

[55] H. Venkateswara, J. Eusebio, S. Chakraborty, and S. Panchanathan, "Deep hashing network for unsupervised domain adaptation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 5018–5027.

[56] S. Beery, G. Van Horn, and P. Perona, "Recognition in terra incognita," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 456–473.

[57] X. Peng, Q. Bai, X. Xia, Z. Huang, K. Saenko, and B. Wang, "Moment matching for multi-source domain adaptation," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2019, pp. 1406–1415.

[58] M. Everingham, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes (voc) challenge," *International Journal of Computer Vision (IJCV)*, vol. 88, pp. 303–338, 2010.

[59] B. C. Russell, A. Torralba, K. P. Murphy, and W. T. Freeman, "Labelme: a database and web-based tool for image annotation," *International Journal of Computer Vision (IJCV)*, vol. 77, pp. 157–173, 2008.

[60] L. Fei-Fei, R. Fergus, and P. Perona, "Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshop*. IEEE, 2004, pp. 178–178.

[61] J. Xiao, J. Hays, K. A. Ehinger, A. Oliva, and A. Torralba, "Sun database: Large-scale scene recognition from abbey to zoo," in *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2010, pp. 3485–3492.

[62] I. Gulrajani and D. Lopez-Paz, "In search of lost domain generalization," *arXiv preprint arXiv:2007.01434*, 2020.

[63] S. Bai, Y. Zhang, W. Zhou, Z. Luan, and B. Chen, "Soft prompt generation for domain generalization," in *Proceedings of the European Conference on Computer Vision (ECCV)*. Springer, 2024, pp. 434–450.

[64] M. U. Khattak, H. Rasheed, M. Maaz, S. Khan, and F. S. Khan, "Maple: Multi-modal prompt learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023, pp. 19 113–19 122.

[65] V. N. Vapnik, V. Vapnik *et al.*, "Statistical learning theory," 1998.

[66] Y. Ganin, E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, M. March, and V. Lempitsky, "Domain-adversarial training of neural networks," *Journal of Machine Learning Research (JMLR)*, vol. 17, no. 59, pp. 1–35, 2016.

[67] B. Sun and K. Saenko, "Deep coral: Correlation alignment for deep domain adaptation," in *Proceedings of the European Conference on Computer Vision (ECCV)*. Springer, 2016, pp. 443–450.

[68] J. Cha, K. Lee, S. Park, and S. Chun, "Domain generalization by mutual-information regularization with pre-trained models," in *Proceedings of the European Conference on Computer Vision (ECCV)*. Springer, 2022, pp. 440–457.

[69] P. Gao, S. Geng, R. Zhang, T. Ma, R. Fang, Y. Zhang, H. Li, and Y. Qiao, "Clip-adapter: Better vision-language models with feature adapters," *International Journal of Computer Vision (IJCV)*, vol. 132, no. 2, pp. 581–595, 2024.

[70] M. Lafon, E. Ramzi, C. Rambour, N. Audebert, and N. Thome, "Gallop: Learning global and local prompts for vision-language models," in *Proceedings of the European Conference on Computer Vision (ECCV)*. Springer, 2024, pp. 264–282.

[71] X. Yu, S. Yoo, and Y. Lin, "Clipceil: Domain generalization through clip via channel refinement and image-text alignment," *Proceedings of the 37th International Conference on Neural Information Processing Systems (NeurIPS)*, vol. 37, pp. 4267–4294, 2024.

[72] Q. Xu, R. Zhang, Y. Zhang, Y. Wang, and Q. Tian, "A fourier-based framework for domain generalization," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 14 383–14 392.

[73] Z. Xu, D. Liu, J. Yang, C. Raffel, and M. Niethammer, "Robust and generalizable visual representation learning via random convolutions," in *International Conference on Learning Representations (ICLR)*, 2021.

[74] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 618–626.

[75] X. Zhang, Y. He, R. Xu, H. Yu, Z. Shen, and P. Cui, "Nico++: Towards better benchmarking for domain generalization," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023, pp. 16 036–16 047.

Ziyi Wang received the B.E. degree from Beijing Institute of Technology, Beijing, China, in 2023. He is currently pursuing an M.S. degree with the School of Computer Science and Technology, Beijing Institute of Technology, Beijing, China. His research interests include domain adaptation, domain generalization, and video understanding.

Zhi Gao is an associate professor (tenure-track) at the School of Computer Science and Technology, Beijing Institute of Technology (BIT). He received the B.S. and Ph.D. degrees in computer science from the Beijing Institute of Technology (BIT), Beijing, China, in 2017 and 2023, respectively. After that, he was a postdoctoral research fellow at the School of Intelligence Science and Technology, Peking University. His current research interests include computer vision, machine learning, multimodal learning, and Riemannian geometry.

Jin Chen received the B.S. and Ph.D. degrees in computer science from the Beijing Institute of Technology (BIT), Beijing, China, in 2017 and 2023, respectively. Her current research interests include computer vision, machine learning, and transfer learning.

Qingjie Zhao received the Ph.D. degree in Computer Science and Technology of Tsinghua University, in 2003. She has been working in the school of Computer Science and Technology, Beijing Institute of Technology, China. She was a visiting Fellow in the school of Computer Science and Electronic Engineering, University of Essex, U.K., from 2008 to 2009, and a senior visiting scholar in the Department of Information, University of Hamburg, from 2017 to 2018, Germany. She is a member of China Computer Federation, Chinese Association of Automation, China Association of Artificial Intelligence. Her current research interests include Image and Video Processing, machine learning and intelligent system.

Xinxiao Wu (Member, IEEE) received the B.S. degree in computer science from the Nanjing University of Information Science and Technology, Nanjing, China, in 2005, and the Ph.D. degree in computer science from the Beijing Institute of Technology (BIT), Beijing, China, in 2010. From 2010 to 2011, she was a Postdoctoral Research Fellow with Nanyang Technological University, Singapore. She is currently a Full Professor with the School of Computer Science, BIT. Her research interests include vision and language, machine learning, and video understanding. She serves on the Editorial Boards of the IEEE Transactions on Multimedia.

Jiebo Luo (Fellow, IEEE) is currently the Albert Arendt Hopeman Professor of Engineering and Professor of Computer Science with the University of Rochester, Rochester, NY, USA, which he joined in 2011 after a prolific career of 15 years with the Kodak Research Laboratories. He has authored over 600 technical papers and holds more than 90 U.S. patents. His research interests include computer vision, natural language processing (NLP), machine learning, data mining, computational social science, and digital health. Dr. Luo is a fellow of NAI, ACM, AAAI, SPIE, and IAPR. He has been involved in numerous technical conferences, including serving as Program Co-Chair for ACM Multimedia 2010, IEEE CVPR 2012, ACM ICMR 2016, and IEEE ICIP 2017, and General Co-Chair for ACM Multimedia 2018 and IEEE ICME 2024. He has served on the editorial boards of IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE (TPAMI), IEEE TRANSACTIONS ON MULTIMEDIA (TMM), IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY (TCSVT), IEEE TRANSACTIONS ON BIG DATA (TBD), ACM Transactions on Intelligent Systems and Technology (TIST), Pattern Recognition, Knowledge and Information Systems (KAIS), Machine Vision and Applications, and Intelligent Medicine. He was an Editor-in-Chief of IEEE TRANSACTIONS ON MULTIMEDIA from 2020 to 2022.