Pivot-Guided Embedding for Domain Generalization

HYUN SEOK SEONG\(^1\), JAEHYUN CHOI\(^2\), (Member, IEEE), WOOJIN JEONG\(^1\), AND JAE-PIL HEO\(^1,3\), (Member, IEEE)

\(^1\)Department of Artificial Intelligence, Sungkyunkwan University, Suwon 16419, South Korea
\(^2\)Department of Electrical Engineering, Korea Advanced Institute of Science and Technology (KAIST), Daejeon 34141, South Korea
\(^3\)Department of Computer Science and Engineering, Sungkyunkwan University, Suwon 16419, South Korea
Corresponding author: Jae-Pil Heo (jaepilheo@skku.edu)

This work was supported in part by the Ministry of Culture, Sports and Tourism and the Korea Creative Content Agency under Project R2020070002; in part by the Institute of Information & Communications Technology Planning & Evaluation funded by the Korean Government through the Ministry of Science and ICT (MSIT) under Project 2020-0-00973 and Project 2019-0-00421; and in part by the Korea Institute of Police Technology funded by the Korean National Police Agency & MSIT under PoliceLab 2.0 Project 210121M06.

ABSTRACT Neural networks have suffered from a distribution gap between training and test data, known as domain shift. Domain generalization (DG) methods aim to learn domain invariant representations only with limited source domain data to cope with unseen target domains. The main assumption is that the model trained to extract semantically consistent features without any domain specific information is highly adaptable to the unseen target domain. Metric learning allows embedding representations to be class-separated and domain-mixed, which is an optimal condition for DG but has been downplayed in recent works. Even the most popular triplet embedding has limitations in forming an optimal embedding space for DG due to instability. In this paper, we present a novel deep metric learning method for domain invariant representations. Specifically, we propose Pivot-Guided Embedding (PGE), which explicitly forms the entire feature distribution of the embedding space with a novel pivot-guided attraction-repulsion mechanism, to address the instability problem that triplet embedding has. In particular, we leverage pivot features representing a coarse distribution of the entire space as reference points to guide other features toward domain invariant feature distribution. To this end, a pivot selection algorithm is presented to reliably reflect the entire feature distribution. Furthermore, we define Guide-Field, a subspace spanned by a subset of pivots chosen for individual samples, to guide each sample to domain invariant feature space. In a nutshell, the attraction-repulsion mechanism based on pivots, the reliable set of features representing the entire feature distribution, enables the model to extract domain invariant feature representations and also settles the instability problem of triplet loss. Experimental results on three different benchmarks validate the performance advantages of the proposed method over the state-of-the-art DG techniques.

INDEX TERMS Deep learning, domain generalization, image classification, metric learning.

I. INTRODUCTION

Deep neural networks have been a great success in various visual tasks. However, most of the studies assume that training and test data are in the same distribution, therefore, the performance drops when the distribution gap exists between training and test data, known as the domain shift problem. To alleviate this problem, there have been many studies on domain generalization (DG) that build models applicable to new “unseen” target domains by utilizing training data from multiple source domains [1], [2], [3], [4], [5], [6], [7], [8].

Recent DG methods focus on learning domain invariant representation by aligning the feature distributions of source domains [9], [10], [11], [12], expecting to capture semantically consistent features that may generalize better to unseen domains [6]. Accordingly, under the constraint of inaccessibility to the target domain, an optimal feature distribution of source domains for DG can be accomplished by
In this paper, we address the DG problem by introducing a novel Pivot-Guided Embedding (PGE) which explicitly forms the entire feature distribution to learn domain invariant representations. Specifically, PGE employs representative features of each domain and class as pivot features, which define the coarse distribution of the embedding space. To overcome the aforementioned problem of triplet embedding and realize stable embedding for DG as illustrated in Figure 1, we first introduce the pivot selection algorithm to discover a reliable set of features reflecting the entire feature distribution. Moreover, to utilize the pivots for guiding the embedding samples toward domain invariant feature distribution, we define Guide-Field (GF) spanned by a subset of pivots to interact with samples in either attractive or repulsive ways. The attraction mechanism gathers the samples toward the positive GF defined by pivot features of the same class regardless of the domain, while the repulsion mechanism separates the samples from the negative GF constructed by pivots of the same domain but different classes. Note that, GF is temporarily generated for each sample through interpolation among sample-wise chosen pivots since the most influential hard positive and hard negative pivots are different for each sample. This pivot-based attraction-repulsion mechanism allows domain-mixing but class-separated embedding in a stable manner. In short, our metric learning method enables the model to learn domain invariant features with high stability by exploiting the pivots that reflect the entire feature distributions well.

Our main contributions are summarized as follows:

- To address the DG problem, we propose a novel domain invariant metric learning, Pivot-Guided Embedding, which involves the entire feature distribution during training based on pivot features.
- We introduce a novel pivot selection algorithm and Guide-Field construction scheme to elect reliable pivot features and to guide the samples to an appropriate position for domain invariant feature distribution.
- We achieve the state-of-the-art DG performances on three benchmark datasets; PACS, Office-Home, and Digits-DG. Further experimental analysis verifies the stability of our metric learning method compared to the standard triplet embedding.

II. RELATED WORK

A. DOMAIN GENERALIZATION (DG)

DG aims to generalize well on the unseen target domains when multiple source domains are given. To achieve this goal, various DG methods are proposed to align feature distributions among multiple source domains, to learn domain invariant representations [2], [3], [4], [6], [7], [9], [10], [11], [12], [13], [15], [16], [17]. A common intuition of these methods is that the model will be capable of extracting domain-invariant features as learning to mix source domains [2], [11], [12]. For instance, MMD-AAE [2] matches feature distributions of multiple source domains using an adversarial method and MMLD [12] learns domain invariant representations by deceiving the domain discriminator with latent domains defined by local features latent in each layer. In addition, meta-learning methods, in which models are trained by meta-training and meta-test simulations to expose it to domain shift conditions, have been widely used for domain generalization recently [3], [4], [7], [17]. Especially, Epi-FCR [7] trains the model with an episodic strategy...
of iteratively exposing the network to novel domains and mDSDF [17], a recent method using meta-training scheme, efficiently utilizes both domain specific and domain invariant information under the theoretical analysis of the efficiency of domain specific information. Furthermore, MASF [6] and EISNet [13] conduct metric learning objectives that explicitly regularize the semantic structure of feature space. Although a simple but direct approach for domain invariant representation is metric learning which can explicitly handle the embedding features, it has been downplayed in recent approaches. In this respect, we tackle the DG problem by proposing a novel metric learning scheme, PGE to reduce the domain gap between source domain data. In addition, training instability problems arise in recent metric learning works [6], [13]. We argue that PGE addresses such training instability and shows much higher performance. Detailed differences between existing metric learning methods and PGE are described in the following subsection.

Another mainstream of DG is based on data augmentation, a way to generate new samples by applying slight variation to training data. The purpose of data augmentation in DG is to enlarge the source domain distribution into a wider span [5], [8], [18], [19], [20], [21], [22], [23], [24], [25], [26], [27], which can improve the robustness of the models to novel domains. For instance, CrossGrad [19] trains a domain classifier with adversarial training to perturb input data, and Jigen [5] augments images by mixing squared patches within images to train a single model with both supervision and self-supervision. Very recently, the generation-based approaches to define novel domains by synthesizing new images in the embedding space have been proposed [8], [23], [25], [27]. In particular, MixStyle [25] generates novel domains by mixing implicit style information and demonstrates outstanding performances and EFDMix [27] provides more diverse feature augmentations by measuring accurate feature distributions. The above studies have an intuition that a simple way to improve generalization capability is to increase the diversity of source domains by generating additional data. To demonstrate that PGE works complementarily with the data augmentation, we additionally conduct the experiments combining the two methods and show performance improvements.

**B. METRIC LEARNING**

Metric learning, which directly handles the embedding representations by distance metric has been widely studied for various tasks [28], [29], [30], [31]. One of the most popular metric learning methods is triplet embedding, which brings positive samples closer and separates negative samples [6], [13], [32]. The triplet embedding is actively utilized due to the strength of local classwise retrieval [14].

In the DG problem, metric learning mostly involving the feature locations in the embedding space would be an appropriate choice to accomplish domain invariant feature distribution as it deals with learning a feature embedding consistent with semantic similarity. Recently, the triplet loss is exploited for DG by considering a vast number of inter-domain sample pairs [6], [10], [13]. However, triplet loss with hard negative moves the samples in a heterogeneous direction because we cannot consider a great number of samples or even entire training data for a single anchor point, as shown in Figure 1 (Left). Furthermore, while the performance of the triplet embedding improves as the number of hard negatives increases, the extremely large number of hard negatives disturbs the model training [13]. Since it is obviously beneficial to employ the entire feature distribution for metric learning, we propose to utilize a small number of pivots representing the entire distribution rather than using the triplet embedding.

In the end, our proposed PGE explicitly handles the domain gap and also performs domain invariant embedding in a more deterministic and stable manner than the triplet embedding, since our PGE restructures the embedding space only with the interactions between samples and the corresponding GF, spanned by the fixed pivots, in either attractive or repulsive ways.

**C. CONTRASTIVE LEARNING**

Recently, several approaches have been proposed that minimize the relative differences between the two differently augmented data (i.e., contrastive learning) for representation learning [33], [34], [35]. These approaches are similar to our pivot-guided attraction-repulsion mechanism since they interact with both positive and negative reference points. However, in PGE, the samples are assigned based on the reference space, GF, not the reference point. It facilitates the construction of domain invariant feature distribution since each sample interacts with prudently selected pivots reflecting the entire feature distribution without the influence of outliers considered as noise. Moreover, in the attraction mechanism, the feature of input data moves toward the feature of differently transformed data as well as a positive GF, enforcing the model to regularize instance consistency while losing domain-specific features. We argue that, for DG, this instance consistency regularizer facilitates enlarging microscale generalization capability, leading the model to be robust against minor differences in the images.

**III. PROPOSED METHOD**

**A. OVERVIEW**

In DG, the model utilizes $K$ source domains $\{D_1, D_2, \ldots, D_K\}$, where the $k$-th domain $D_k$ contains $N_k$ labeled instances $\{(x_{ki}^l, y_{ki}^l)\}_{l=1}^{N_k}$. Note that, $x_{ki}^l$ indicates the input images and $y_{ki}^l \in \{1, \ldots, L\}$ is the corresponding class label. During training, a feature extractor $f_{\theta}(\cdot)$ extracts the feature $f_{\theta}(x_{ki}^l)$ from an image $x_{ki}^l$, where $\theta$ denotes the parameters of the feature extractor. We further define a pivot feature $z_{j}^{l}$, the $j$-th pivot feature in a pivot set $Z^{k,l}$, a set of representative features of the $l$-th class in $D_k$. The pivot set stores $N_{p}$ pivots selected for each class in a domain in the past few iterations.

In this paper, we propose PGE to help the model to be aware of the entire feature distribution of the embedding...
space. PGE employs two objective functions to attract and repulse the embedding samples from the GF constructed by the pivots according to their class and domain labels. Specifically, the embedding feature of input data $x_i^k, f_o(x_i^k)$, is attracted by the positive GF, built upon the pivots of the same class, $z_j^k$, and the features of differently transformed input data, $f_o(x_i^k)$, while it is repulsed from the negative GF, built upon the pivots of the same domain but different classes $z^l_j$. Note that, we apply PGE to both $x_i^k$ and $\tilde{x}_i^k$, but only discuss the cases of $x_i^k$ in the following subsections.

The overview of the proposed method is illustrated in Figure 4. In the following, we introduce the pivot selection algorithm, Guide-Field construction method, and the objective functions of the proposed metric learning methods.

**B. PIVOT SELECTION PROCESS**

As the pivots are reference points of the embedding space, it is crucial to select highly reliable pivots. To do so, we maintain a candidate set throughout the training iterations. The candidate set is used to elect representatives to be stored in the pivot set. Specifically, in each iteration, we elect $K \times L$ pivots each from different class and domain. These new pivots replace the oldest pivots in the pivot set similar to circular queue. To identify a candidate that best reflects the distribution of its corresponding class and domain, a pivot selection is performed to discover pivots within locally dense regions. The entire process of the pivot selection process is illustrated in Figure 2.

1) CANDIDATE SET

While the size of the mini-batch is insufficient to measure the representability of each domain and class, we gather the candidates and maintain them with the first-in-first-out mechanism. Thus, such a rich set of candidates ensures the high reliability of pivots. Accordingly, the number of features stored in the candidate set remains constant during training. We denote the set of candidates of the class $i$ from the domain $k$ as $C_{i,k,l} = \{c_{i,q}^{k,l}\}_{q=1}^{N_{Q}}$, where $N_{Q}$ is the number of candidates for $l$-th class of $k$-th domain in the bank. We represent the sum of $N_{Q}^{k,l}$ over all $k$ and $l$ as $N_{Q}$.

2) PIVOT SELECTION ALGORITHM FOR PIVOT SET

To produce reference points for guiding the embedding features, we configure the pivot set with a few pivots selected from the candidate set. To do so, we present a pivot selection algorithm to elect a new pivot for a particular class of a domain. Intuitively, our criterion to select a pivot is to identify the candidate having many neighboring samples of the same class and domain. We formulate the criterion to select a new pivot ($N_{P}$-th pivot of the class $i$ and the domain $k$) based on the average distances to the other candidates of the same class and domain as follows:

$$z_{N_{P}}^{k,l} = \arg \min_{c \in C_{i,k,l}} \frac{1}{N_{Q}^{k,l}} \sum_{q=1}^{N_{Q}} d(c, z_{q}^{k,l}),$$

where $d(\cdot, \cdot)$ is Euclidean distance between two vectors.

One may ask why we do not utilize a mean vector of elements in $C_{i,k,l}$ as a new pivot since the result computed by Eq. 1 would be close to the mean vector in most cases. However, since the mean vectors are highly sensitive to outliers, it can disturb to achieve our goal. That’s why we utilize a slightly more complex but robust algorithm instead of the simple alternative.

**C. GUIDE-FIELD CONSTRUCTION**

The purpose of our metric learning is to move the individual sample in an optimal direction to accomplish domain invariant embedding, considering the entire feature distribution. To this end, we define the corresponding two reference fields for each sample. One is a field having the information of the same domain but different classes, and the other is a field with
mixed-domain information within the same class. We refer to them as negative and positive GF, respectively, and both GFs are defined as subspace spanned by the subset of pivots. The main purpose of constructing GF is to form the most influential reference spaces from the perspective of individual samples while being aware of the entire feature distribution. Another purpose of utilizing the GF rather than only pivots is to reduce undesired risks: losing similarity when the pivots selected by two similar samples are highly different, and even more serious when the target direction of each sample is fixed while being aware of the entire feature distribution.

To define the interaction between a GF and a feature vector, we mass-produce the Guide-Points by linearly interpolating the pivots defining the GF, motivated by the attractive-repulsive metric learning schemes. Specifically, for each sample \( x^k_i \), a positive GF is defined by the hard positive pivot per each domain. In other words, the positive GF is a subspace spanned by \( K \) hard positive pivots. To have an interaction between the sample \( x^k_i \) and such positive GF, we approximate the positive GF by sampling the Guide-Points \( g^p(x^k_i) \). The sample \( x^k_i \) then moves toward every Guide-Point.

In detail, for attraction, the hard positive pivots \( p^h(x^k_i) \) to build positive GF are defined by the farthest one from the feature \( f_0(x^k_i) \) within each set \( Z^k_i \) for \( n = 1, \ldots, K \) (i.e., \( K \) pivots are selected), as follows:

\[
p^h_n(x^k_i) = \text{argmax}_{z \in Z^k_i} d(z, f_0(x^k_i)).
\]

Then these pivots form a positive GF. To approximate the positive GF, we draw \( N_{GF} \) Guide-Points. Each Guide-Point \( g^p_n(x^k_i) \) for \( n = 1, \ldots, N_{GF} \) is described as follows:

\[
g^p_n(x^k_i) = \sum_{s=1}^{T} w_s p^h_s(x^k_i),
\]

where \( (t_1, t_2, \ldots, t_T) \) are randomly chosen from \( \{1, \ldots, K\} \) without duplicate and the mixing coefficient \( w \) is obtained from Dirichlet distribution. Note that, we found that \( T = 3 \) provides a good balance between the accuracy and learning cost, and \( N_{GF} = 64 \) is sufficient to approximate GF.
Given the above, the objective function for attraction is formulated as follows:

\[
L_{\text{att}}(x_i^k) = \frac{1}{K + N_{\text{GF}}} \left( \sum_{n=1}^{K} d(f(x_i^k), p_n^\theta(x_i^k)) + \sum_{u=1}^{N_{\text{GF}}} d(f(x_i^k), g_u^\omega(x_i^k)) \right),
\]

where the left term in parenthesis indicates that the hard positives pivots are also utilized for attraction.

As a result, the objective function by incorporating the instance consistency regularization is formulated as follows:

\[
L_{\text{ATT}}(x_i^k) = L_{\text{att}}(x_i^k) + \beta d(f_0(x_i^k), f_0(\tilde{x}_i^m)),
\]

where \(f_0(\tilde{x}_i^m)\) is the feature of differently transformed input data \(\tilde{x}_i^m\). Intuitively, a sample moves toward the class-specific domain with the sample are ignored. Through the same process with the attraction, we build a negative GF upon the domain-wise selected \(K\) pivots with the longest distance from the sample. In addition, the right term of Eq. 5 with balancing hyperparameter \(\beta\) is a consistency regularizer term which enforces the features of the two samples \(f(x_i^k)\) and \(f(\tilde{x}_i^m)\) to be closely located in the embedding space to learn instance-level discriminative features.

On the other hand, to locate the samples of different classes of the same domain far from each other, we explicitly repulse those samples from negative GF defined by the hard pivots of different classes of the same domain. In contrast to the attraction, the hard negatives for the repulsion are close to the sample. In detail, for each class, the influential pivot for the repulsion is the nearest one in each set \(Z^{k,m}\) for \(m = 1, \ldots, L(m \neq y_i^k)\), to the sample across different domains thus the negative GF is constructed by these pivots. Therefore, for repulsion, the hard negative pivots \(p_m^N(x_i^k)\) are defined as follows:

\[
p_m^N(x_i^k) = \mathbb{I}_{[m \neq y_i^k]} \arg\min_{z \in Z^{k,m}} d(z, f_0(x_i^k)),
\]

where \(\mathbb{I}_{[m \neq y_i^k]}\) ensures that pivots of the same class and domain with the sample are ignored. Through the same process with the attraction, we build a negative GF upon the hard negative pivots \(p_m^N(x_i^k)\). In other words, each Guide-Point \(g_v^N(x_i^k)\) for \(v = 1, \ldots, N_{\text{GF}}\) is computed by the same manner as positive GF as follows:

\[
g_v^N(x_i^k) = \sum_{s=1}^{T} w_s p_s^N(x_i^k).
\]

As a result, the objective function for the repulsive interaction for a sample \(x_i^k\) is described as follows:

\[
L_{\text{REP}}(x_i^k) = -\frac{1}{L + N_{\text{GF}} - 1} \left( \sum_{u=1}^{L-1} d(f(x_i^k), p_u^\theta(x_i^k)) + \sum_{v=1}^{N_{\text{GF}}} d(f(x_i^k), g_v^\omega(x_i^k)) \right).
\]

Therefore, the objective function of PGE for \(x_i^k\) is formulated as follows:

\[
L_{\text{PGE}}(x_i^k) = \frac{1}{2N} \sum_{i=1}^{N} (\gamma_1 L_{\text{ATT}}(x_i^k) + \gamma_2 L_{\text{REP}}(x_i^k)),
\]

where \(N\) denotes the number of images in a mini-batch, and hyperparameters \(\gamma_1\) and \(\gamma_2\) control the contributions of the attraction and repulsion terms, respectively. The overall objective function when PGE is applied to both \(x_i^k\) and \(\tilde{x}_i^m\) is described as follows:

\[
L_{\text{PGE}} = \frac{1}{2} (L_{\text{PGE}}(x_i^k) + L_{\text{PGE}}(\tilde{x}_i^m)).
\]

Finally, by putting the cross-entropy loss \(L_{\text{CE}}\) for the classification task on the source domain, the overall objective function is defined as follows:

\[
L_{\text{ALL}} = L_{\text{CE}} + L_{\text{PGE}}.
\]

IV. EXPERIMENTS
A. DATASETS
To demonstrate the effectiveness of our model, we evaluate our approach on three domain generalization benchmark datasets:

- **PACS** [1] dataset consists of 9,991 images in four domains, (Art Painting, Cartoon, Sketch, and Photo), with seven shared object categories (i.e., Dog, Elephant, Giraffe, Guitar, Horse, House and Person).
- **Office-Home** [38] dataset has 15,500 images containing 65 object categories of objects in office and home environment. It has 4 domains, which are (Artistic, Clipart, Product, and Real World).
- **Digits-DG** includes four different digit datasets MNIST [39], MNIST-M [40], SVHN [41], SYN [40] with different font, background, and stroke colors. Each dataset consists of the handwritten digit images, the variant of MNIST, street view house number images, and synthetic digit dataset with varying fonts, backgrounds and stroke color, respectively. Each class in a domain consists of 10 classes with total of 600 images and we split the dataset into 80% for training dataset and 20% for testing dataset.

We follow the leave-one-out evaluation protocol for a fair comparison with the prior domain generalization methods. In other words, we set one domain as a target domain to test a model while the remained domains are used as source domains.

B. IMPLEMENTATION DETAILS
1) TRAINING
For PACS and Office-Home datasets, we use the ImageNet [42] pre-trained ResNet-18 as the backbone and also use the ImageNet pre-trained ResNet-50 for the PACS dataset as did in the baselines. The network is trained with an SGD solver for 50 epochs. The initial learning rate, weight decay, and batch size are 0.001, 5e-4, and 32 (16 images each.
2) HYPERPARAMETERS

($\beta, \gamma_1, \gamma_2$) are set to (1, 0.01, 0.005) for Office-Home and PACS with ResNet-18, and (10, 0.001, 0.001) for ResNet-50 architecture. We maintain $N_Q = 4096$, $8192$ candidates for PACS and Office-Home, respectively, and $N_P = 64$ pivots in the pivot set. For Digits-DG dataset, ($\gamma_1, \gamma_2$) are set to (0.002, 0.002). Note that, there is no $\beta$ because data augmentation is not applied for Digits-DG dataset. In addition, we maintain ($N_Q$, $N_P$) as (1024, 16).

C. RESULTS

1) PACS DATASET

Table 1 reports top-1 classification accuracy on PACS dataset with comparison to previous works as well as DeepAll which trains a model on the aggregated source domain data without any technique. Compared to ER [43] and MMLD [12], which train models to extract domain invariant representations by adversarial training schemes, our metric learning method (PGE) that explicitly handle the embedding feature representations improves DG performance. Specifically, we state that PGE does not suffer from training instability which can be caused by adversarial training in MMLD. Besides, we can observe that PGE shows superior performance against the other metric-based methods including MASF [6] and EISNet [13] on both ResNet-18 and ResNet-50 architectures. In comparison to EISNet, it shows that PGE achieves a large improvement against the triplet embedding, despite the absence of additional self-supervision existing in EISNet. The superiority of our method over the other triplet-based methods comes from alleviating training instability that occurs when samples are shifted in a heterogeneous direction because the selected positive and negative samples for each instance are different (Figure 1). Reducing the domain gap from a macroscopic perspective improves performance. A detailed comparison of PGE against triplet loss utilized in EISNet is shown in Section V-A. A recent data augmentation method, MixStyle [25], pAdaIN [24] and EFDMix [27], achieve significant improvement over the other baselines. Although they diversify the source domain distribution by discovering novel domains implicitly synthesized between multiple source domains, such trained models are insufficient to produce domain invariant features because they do not explicitly reduce the domain gap to diminish the domain specific features. Further, although L2D [26] are to extract style invariant and class specific features by increasing the mutual information between multiple generated features and they improve the performance above the aforementioned data augmentation methods, PGE performs better than L2D because they do not explicitly reduce the domain gap between multiple source domains. Notably, for ResNet-50 architecture, PGE achieves remarkable performances surpassing all baselines in all domains except Photo domain.

Furthermore, we conducted additional experiments that combine PGE with the state-of-the-art data augmentation method, MixStyle [25], as shown in Table 2, to show the for two different transformations), respectively. The learning rate is decayed by 0.1 after 40 epochs. We use standard data augmentation protocol as Jigen [5], which includes resizing, horizontal flipping, random cropping, and color jittering for a fair comparison. Note that, the instance consistency regularizer takes effect by the variation arose by two different randomly chosen transformations. For Digits-DG dataset, the model is constructed with four $3 \times 3$ convolutional layers and a classifier. ReLU and $2 \times 2$ max-pooling layer follow after each convolutional layer. In addition, we train the model only with $L_C$ without $L_{PGE}$ until sufficient pivot candidates and pivots are gathered (i.e., $N_Q$ and $N_P$, respectively). After sufficient pivot candidates and pivots are collected, the model is trained using $L_{PGE}$ together with $L_C$ while continuously updating the candidate set and pivot set. We train the model for 50 epochs, with the initial learning rate set as 0.05, decayed by 0.1 after every 20 epochs. All of the experiment results are averaged over five independent runs.

### TABLE 1. DG performance on PACS dataset. \( \dagger \) results of the open-source implementations published by the authors. Bold and underline represent the best and second-best accuracy, respectively.

| Method       | Art | Cartoon | Sketch | Photo | Avg. |
|--------------|-----|---------|--------|-------|------|
| DeepAll      | 77.85 | 75.49 | 69.28 | 96.66 | 79.82 |
| Jigen [5]    | 79.42 | 75.25 | 71.35 | 96.03 | 80.51 |
| MASF [6]     | 80.29 | 77.17 | 71.69 | 94.99 | 81.04 |
| Epi-FCR [7]  | 82.10 | 77.00 | 73.00 | 93.90 | 81.50 |
| MMLD [12]    | 81.28 | 77.16 | 72.29 | 96.09 | 81.83 |
| ER [43]      | 80.70 | 76.40 | 71.77 | 96.65 | 81.38 |
| EISNet [13]  | 81.89 | 76.44 | 74.33 | 95.93 | 82.15 |
| RSC* [44]    | 80.11 | 78.17 | 77.50 | 93.90 | 82.42 |
| DDAI[8]      | 84.20 | 78.10 | 74.70 | 95.30 | 83.10 |
| pAdaIN [24]  | 81.74 | 76.91 | 75.13 | 96.29 | 82.51 |
| MixStyle [25]| 84.10 | 78.80 | 75.90 | 96.10 | 83.70 |
| L2D [26]     | 81.44 | 79.56 | 80.50 | 95.51 | 84.22 |
| EFDMix [27]  | 83.90 | 79.40 | 75.00 | 96.80 | 83.90 |
| PGE (Ours)   | 83.66 | 80.36 | 79.28 | 95.65 | 84.74 |

| Method       | Art | Cartoon | Sketch | Photo | Avg. |
|--------------|-----|---------|--------|-------|------|
| DeepAll      | 85.47 | 75.48 | 73.44 | 97.78 | 83.04 |
| MASF [6]     | 82.89 | 80.49 | 72.29 | 95.01 | 82.67 |
| ER [43]      | 87.51 | 79.31 | 76.30 | 98.25 | 85.34 |
| EISNet [13]  | 86.64 | 81.53 | 78.07 | 97.11 | 85.84 |
| RSC* [44]    | 85.47 | 82.12 | 82.32 | 95.41 | 86.33 |
| pAdaIN [24]  | 85.82 | 81.06 | 77.37 | 97.17 | 85.36 |
| MixStyle [25]| 87.86 | 82.26 | 81.16 | 97.45 | 87.18 |
| EFDMix [27]  | 90.60 | 82.50 | 76.40 | 98.10 | 86.90 |
| PGE (Ours)   | 88.76 | 83.85 | 83.40 | 97.00 | 88.25 |

### TABLE 2. DG performance on PACS dataset to combine PGE with MixStyle, a recent state-of-the-art data augmentation method. Bold represents the best accuracy.

| Method       | Art | Cartoon | Sketch | Photo | Avg. |
|--------------|-----|---------|--------|-------|------|
| DeepAll      | 85.47 | 75.48 | 73.44 | 97.78 | 83.04 |
| MixStyle [25]| 87.86 | 82.26 | 81.16 | 97.45 | 87.18 |
| + PGE (Ours) | **88.40** | **84.03** | **83.56** | **96.65** | **88.61** |
complementary benefits between PGE and data augmentation. We can observe that PGE boosts the performances of MixStyle in all domains except Photo domain. Note that, the purpose of data augmentation techniques, MixStyle [25], pAdaIN [24] and EFDMix [27], are to diversify the source domain as aforementioned. Therefore, when combined with our PGE, the performance improves because the diversified statistics of features through data augmentation can boosts the objective of PGE, reducing the gap between the features with given domain statistics. The promising results for MixStyle+PGE indicate the high adaptivity of PGE to the data augmentation technique.

2) OFFICE-HOME DATASET

From Table 3, we can observe that PGE outperforms the baselines including the regularization-based approach RSC [44] and JiGen [5] which trained with auxiliary self-supervision. In particular, CCSA [10] utilized metric learning for domain invariant embedding thus sharing the same intuition with PGE. However, our stabilized training method achieves a large improvement of about 1.82%p from CCSA. In addition, the results demonstrate that the advanced data augmentation methods, DDAIG [8] and MixStyle [25], show significant improvement in specific domains, such as Product and Real domain, over the baselines. Interestingly, PGE achieves competitive results in Art and Clipart domains while comparable performances in the remaining domain without advanced data augmentation technique, it eventually provides the best performance on average with 1.22%p improvement over the prior works, DDAIG and MixStyle.

3) DIGITS-DG DATASET

From the summarized results in Table 4, we can observe that PGE achieves significant improvement over the recent approaches, including the metric learning-based approach CCSA [10] and the state-of-the-art data augmentation method, DDAIG [8]. Compared with the metric learning method, PGE yields significantly enhanced performances over CCSA for all domains. Meanwhile, the data augmentation technique benefits in performance improvement especially for synthetic domains, such as MNIST-M and SYN, due to various background colors for monotonous objects in Digits-DG dataset. Nevertheless, we achieve competitive results without utilizing any data augmentation technique and improve average performance over the recent approaches.

V. FURTHER ANALYSIS

A. STABILITY ANALYSIS

To verify the stability of PGE compared to standard triplet loss, we compare the local neighborhood preservation ratio during the training. Specifically, the comparison target is the triplet loss using large hard negative samples utilizing a memory module. In other words, we refer to open-source implementations published by the authors of EISNet [13] to experiment with triplet loss. For stability analysis, we retrieve K-nearest neighbors (KNNs) of every sample within features of the same domain and class, then we measure how many neighbors are preserved between two consecutive training epochs. The experiments are conducted on the PACS dataset with ResNet-18 architecture. High preservation ratio indicates that the features are moving in a structurally stable manner.

As shown in Figure 5, our PGE better preserves the local neighborhood compared to the standard triplet-based method in every K and training epoch, indicating that PGE which forms a global feature distribution via fixed pivots is
more deterministic and stable, thus facilitating maintaining local similarity between instances. Specifically, the triplet in EISNet is composed of scattered local neighborhoods and does not even consider the global structure for DG. In other words, they indirectly construct the embedding space through the constraints of the local structure only for classification boundary, not for domain-mixing and class-separated embedding. However, in PGE, the displacements of samples are explicitly induced by the pivots containing the core class-diding. In addition, for both optimal class boundary and domain characteristics which most often violates domain invariant embedding, samples of the same class regardless of different, but the closest pivot has the most analogous character-

### TABLE 5. Ablation studies varying the size of pivot set N_p. The experiments are conducted on PACS dataset with ResNet-18. Bold represents the best accuracy.

| N_p  | Art  | Cartoon | Sketch | Photo | Avg. |
|------|------|---------|--------|-------|------|
| 1    | 81.52| 79.18   | 78.35  | 95.81 | 83.71|
| 4    | 83.00| 79.83   | 79.17  | 95.74 | 84.43|
| 16   | 82.21| 80.14   | 80.18  | 95.96 | 84.62|
| 64   | 83.66| 80.36   | 79.28  | 95.65 | 84.74|
| 256  | 82.18| 80.17   | 79.61  | 95.69 | 84.41|
| 512  | 81.94| 79.95   | 79.86  | 95.68 | 84.36|

### TABLE 6. Ablation study to show the contribution of Guide-Field. Bold represents the best accuracy.

| Method    | Art | Cartoon | Sketch | Photo | Avg. |
|-----------|-----|---------|--------|-------|------|
| DeepAll   | 77.85| 75.49   | 69.28  | 96.66 | 79.82|
| PGE w/o GF| 82.45| 80.70   | 79.03  | 95.37 | 84.44|
| PGE w/ GF | 83.66| 80.36   | 79.28  | 95.65 | 84.74|

### TABLE 7. Performance comparison on pivot selection methods for pivot set. Mean: A mean vector of the candidates is used as a pivot. Ours: A pivot is selected by our selection algorithm. Bold represents the best accuracy.

| Selection | Art  | Cartoon | Sketch | Photo | Avg. |
|-----------|------|---------|--------|-------|------|
| Mean      | 82.36| 79.23   | 78.73  | 95.31 | 83.96|
| Ours      | 83.66| 80.36   | 79.28  | 95.65 | 84.74|

### TABLE 8. Effect of hard sample selection strategy for attraction-repulsion mechanism. Random: A randomly selected pivot among N_p pivots. Mean: A mean vector of N_p pivots is used as a positive (or negative) pivot. Bold represents the best accuracy.

| Selection | Art  | Cartoon | Sketch | Photo | Avg. |
|-----------|------|---------|--------|-------|------|
| Random    | 81.61| 79.86   | 78.42  | 95.45 | 83.83|
| Mean      | 81.90| 80.35   | 78.95  | 95.53 | 84.18|
| Hard (Ours)| 83.66| 80.36   | 79.28  | 95.65 | 84.74|

more deterministic and stable, thus facilitating maintaining local similarity between instances. Specifically, the triplet in EISNet is composed of scattered local neighborhoods and does not even consider the global structure for DG. In other words, they indirectly construct the embedding space through the constraints of the local structure only for classification boundary, not for domain-mixing and class-separated embedding. However, in PGE, the displacements of samples are explicitly induced by the pivots containing the core class-diding. In addition, for both optimal class boundary and domain characteristics which most often violates domain invariant embedding, samples of the same class regardless of different, but the closest pivot has the most analogous character-

### B. ABLATION STUDY

#### 1) VARYING THE SIZE OF PIVOT SET

We explore the necessity of organizing the pivot set. Table 5 shows the experimental results, varying the size of the pivot set N_p from 1 to 512. Note that the corresponding number of pivot set indicates each domain and class. It shows that utilizing too few pivots under-constrains the feature embedding, while metric learning with a large pivot set is more likely to use the old feature as a pivot which can constrain features to shift in the wrong direction. Although the results indicate that it is fine as long as the size of pivot set N_p is not too meager, it demonstrates that maintaining an adequate pivot pool consisting of a few reliable pivots is beneficial to cross-domain generalization.

#### 2) CONTRIBUTION OF GUIDE-FIELD

We provide an ablation study to verify the performance contribution of Guide-Field. As shown in Table 6, the performance improves as GF is added. Specifically, the experiment PGE w/o GF is conducted only with the interaction between features f(x_i^k) and vertices p(x_i^k) for both L_ATT(x_i^k) and L_REP(x_i^k). This demonstrates that the GF constructed from the pivots by Dirichlet Manifold Mixup represents a more sophisticated characteristic of the reference space than the pivots themselves in terms of optimization. Therefore, it is better to interact with the reference field than simply interact with the pivots.

#### 3) INFLUENCE OF PIVOT SELECTION ALGORITHM

We design a pivot selection algorithm slightly more complex but robust to outliers instead of utilizing simple mean vectors. The performance comparison between our algorithm and utilizing the mean vectors is reported in Table 7. Our algorithm consistently outperforms the mean vectors in all domains. We argue that the pivots should be selected from the center of the crowd to avoid the impact of outliers rather than simply selecting the mean vectors. The results in Table 7 demonstrate that our pivot selection algorithm elects more reliable pivots and helps the model form a global feature distribution through the pivots.

#### 4) INFLUENCE OF HARD PIVOT SELECTION STRATEGY

We perform the attraction-repulsion mechanism with the Guide-Fields constructed upon the hard pivots, the closest and farthest pivots for negative and positive, respectively, from the pivot set. There are reasons for selecting hard pivots. For DG, the information between a sample and the pivots from the same domain but different classes must be completely different, but the closest pivot has the most analogous characteristics which most often violates domain invariant embedding. In addition, for both optimal class boundary and domain invariant embedding, samples of the same class regardless of the domain should be mixed and closely located, but the pivot that most influentially violates this is the farthest.

To show the performance improvement of the hard pivot selection strategy, we experimented with two more strategies, ‘mean vector’ and ‘random vector’. As shown in Table 8, we observed that the ‘mean vector’ performs a little less (84.18%) because the hard pivot (i.e., the closest and
FIGURE 6. Ablation study varying the hyperparameters for PGE. The experiments are conducted on PACS dataset with ResNet-18 architecture. (a) Ablation study varying $\beta$ while the other hyperparameters $\gamma_1$ and $\gamma_2$ are set to 0.01, 0.005, respectively. It shows the influence of the consistency regularizer term of attraction mechanism. $\beta = 0$ shows the experiment results training without a consistency regularizer. (b) Ablation study varying $\gamma_1$ while $\beta$ and $\gamma_2$ are set to 1 and 0.005, respectively. (c) Ablation study varying $\gamma_2$ while $\beta$ and $\gamma_1$ are set to 1 and 0.01, respectively. The two hyperparameters $\gamma_1$ and $\gamma_2$ control the contributions of the attraction and repulsion term, respectively.

5) INFLUENCE OF CONSISTENCY REGULARIZER

The attraction mechanism in PGE involves consistency regularizer to learn instance-level discriminative features. Specifically, since PGE without consistency regularizer term is not involved in forming local structure while enabling the model to form a global structure of the embedding space, we utilize consistency regularizer to organize instance-level discriminative embedding within local neighborhoods. In other words, a consistency regularizer enables the model to learn micro-scale differences in images which cannot be captured when learning only with PGE that learns the domain gap in macro-scale. As shown in Figure 6 (a), we conduct an ablation study varying $\beta$ to explore the influence of consistency regularizer. We observed that the consistency regularizer improves the performance up to 0.41% when $\beta$ is increased from 0 to 1. Specifically, PGE performs 84.23% and 84.74% when $\beta$ is set to 0 and 1, respectively. However, the performance drops for too large $\beta$ due to overfitting. It validates that the consistency regularizer with adequate $\beta$ improves the generalization capability in DG. Furthermore, PGE shows competitive performance even without a consistency regularizer.

6) CONTRIBUTIONS OF ATTRACTION AND REPULSION LOSSES

To show the effect when varying the scales of attraction and repulsion losses, we conduct the ablation analysis on varying the hyperparameters $\gamma_1$ and $\gamma_2$ which control the contributions of the attraction and repulsion terms. As shown in Figure 6 (b), PGE without attraction ($\gamma_1 = 0$) performs 82.68% for the fixed hyperparameters $\beta$ and $\gamma_2$. We also observed that the performance improves as $\gamma_1$ increases and peaks when $\gamma_1 = 0.01$. Likewise, the accuracy for PGE without repulsion ($\gamma_2 = 0$) is 83.58% and also improves as $\gamma_2$ increases until $\gamma_2 = 0.005$ as shown in Figure 6 (c). These ablation studies indicate that PGE has low sensitivity to the choice of the hyperparameter $\gamma_2$. In other words, attraction loss has more influence on performance than repulsion loss when without using each other. Although the repulsion loss helps to eliminate the possibility of domain clusters that occurs by domain specific information, the class specific information is also bound to be lost when repulsion loss is used alone. We argue that the attraction loss can reduce damage to these shortcomings of repulsion loss due to its ability to cluster the same class regardless of domain. In specific, the attraction loss not only induces robust classification boundaries but also forms relatively dense and domain-mixed clusters within classes.

VI. CONCLUSION

In this paper, we have presented a metric learning method for domain generalization that imposes PGE to learn domain-invariant representations. PGE incorporates a novel pivot-based attraction-repulsion metric learning based on a reliable pivot selection algorithm and GF construction scheme. It can efficiently consider the entire feature distribution thus stably maintaining the local similarity. In addition, since the GF is temporarily generated in the perspective of individual samples to define the most influential reference space, the induced displacements are stably performed to accomplish domain invariant feature space. Through various experiments, we achieve state-of-the-art performances on three DG benchmark datasets.

the farthest pivot from negative pivots and positive pivots, respectively) to individual sample has the greatest influence, and even worse for the "random vector" (83.83%) that shifts each sample in a more random direction. This proves that hard selection is a more appropriate way for DG.
TABLE 9. DG performances on DomainBed [55]. PGE shows competitive results among the recent DG methods. Note that, we choose the test-domain validation set as a model selection method. The experimental results of 14 baseline methods are sorted according to the order of publication.

| Algorithm   | ColoredMNIST | RotatedMNIST | VLCS   | PACS   | OfficeHome | Avg   |
|-------------|--------------|--------------|--------|--------|------------|-------|
| ERM [45]    | 57.8 ± 0.2   | 97.8 ± 0.1   | 77.6 ± 0.3 | 86.7 ± 0.3 | 66.4 ± 0.5 | 77.3  |
| CORAL [46]  | 58.6 ± 0.5   | 98.0 ± 0.0   | 77.7 ± 0.2 | 87.1 ± 0.5 | 68.4 ± 0.2 | 78.0  |
| DANN [47]   | 57.0 ± 1.0   | 97.9 ± 0.1   | 79.7 ± 0.5 | 85.2 ± 0.2 | 65.3 ± 0.8 | 77.0  |
| MTL [48]    | 57.6 ± 0.3   | 97.9 ± 0.1   | 77.7 ± 0.5 | 86.7 ± 0.2 | 66.5 ± 0.4 | 77.3  |
| MMD [2]     | 63.3 ± 1.3   | 98.0 ± 0.1   | 77.9 ± 0.1 | 87.2 ± 0.1 | 66.2 ± 0.3 | 78.5  |
| CDANN [11]  | 59.5 ± 2.0   | 97.9 ± 0.0   | 79.9 ± 0.2 | 85.8 ± 0.8 | 65.3 ± 0.5 | 77.7  |
| MLDG [3]    | 58.2 ± 0.4   | 97.8 ± 0.1   | 77.5 ± 0.1 | 86.8 ± 0.4 | 66.6 ± 0.3 | 77.4  |
| IRM [49]    | 67.7 ± 1.2   | 97.5 ± 0.2   | 76.9 ± 0.6 | 84.5 ± 1.1 | 63.0 ± 2.7 | 77.9  |
| GroupDRO [50]| 61.1 ± 0.9   | 97.9 ± 0.1   | 77.4 ± 0.5 | 87.1 ± 0.1 | 66.2 ± 0.6 | 77.9  |
| ARM [51]    | 63.2 ± 0.7   | 98.1 ± 0.1   | 77.8 ± 0.3 | 85.8 ± 0.2 | 64.8 ± 0.4 | 77.9  |
| Mixup [52]  | 58.4 ± 0.2   | 98.0 ± 0.1   | 78.1 ± 0.3 | 86.8 ± 0.3 | 68.0 ± 0.2 | 77.9  |
| RSC [44]    | 58.5 ± 0.5   | 97.6 ± 0.1   | 77.8 ± 0.6 | 86.2 ± 0.5 | 66.5 ± 0.6 | 77.3  |
| VREx [53]   | 67.0 ± 1.3   | 97.9 ± 0.1   | 78.1 ± 0.2 | 87.2 ± 0.6 | 65.7 ± 0.3 | 79.2  |
| SagNet [54] | 58.2 ± 0.3   | 97.9 ± 0.0   | 77.6 ± 0.1 | 86.4 ± 0.4 | 67.5 ± 0.2 | 77.5  |
| PGE (Ours)  | 61.0 ± 2.1   | 97.7 ± 0.1   | 78.3 ± 0.2 | 86.7 ± 0.5 | 68.2 ± 0.1 | 78.4  |

APPENDIX A
ADDITIONAL EXPERIMENT RESULTS

We additionally conduct experiments on five datasets (i.e., Colored MNIST [49], Rotated MNIST [15], VLCS [56], PACS [1], Office-Home [38]) which are included in DomainBed [55], a unified testbed for domain generalization. In addition, DomainBed provides experiment results of 14 baseline methods (ERM [45], IRM [49], GroupDRO [50], Mixup [52], MLG [3], CORAL [46], MMD [2], DANN [47], CDANN [11], MTL [48], SagNet [54], ARM [51], VREx [53], and RSC [44]) for a fair comparison.

Note that, the experiments of PGE (Ours) are conducted based on PGE w/o GF.

As shown in Table 9, we compare PGE with the recent DG methods on this standard benchmark. We observed that our PGE shows competitive results among the recent DG methods, especially for VLCS and OfficeHome datasets. These results imply that our PGE is robust to various domain gaps. Since the domain gap in VLCS dataset is large, while the domain gap in the OfficeHome dataset is relatively small. Specifically, we achieved the second-highest performance on OfficeHome dataset after CORAL [46]. Note that, CORAL trains the model leveraging some target domain data, while our PGE utilizes only training data.

APPENDIX B
ROLE OF REPULSION MECHANISM

The purpose of repulsion is to deplete domain specific information to adapt well to unseen target domains. We noticed that shortcuts of DeepAll are not removing domain information despite high classification accuracy on training data, as shown in Fig 7 (a, b-left). A key way to achieve optimal domain-mixing and class-separated embedding as Fig 7 (b-right) is to eliminate the possibility of domain cluster (red circle in Fig 7 (b-left)). In optimal embedding, domain specific features should be diminished. Hence, we focus on separating the samples from the same domain but different classes in the repulsion mechanism, which is an explicit way for domain-agnostic feature learning.

REFERENCES

[1] D. Li, Y. Yang, Y.-Z. Song, and T. M. Hospedales, “Deeper, broader and artier domain generalization,” in Proc. IEEE Int. Conf. Comput. Vis. (ICCV), Oct. 2017, pp. 5542–5550.
[2] H. Li, S. J. Pan, S. Wang, and A. C. Kot, “Domain generalization with adversarial feature learning,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., Jun. 2018, pp. 5400–5409.
[3] D. Li, Y. Yang, Y.-Z. Song, and T. Hospedales, “Learning to generalize: Meta-learning for domain generalization,” in Proc. AAAI Conf. Artif. Intell., 2018, pp. 1–8.
[4] Y. Balaji, S. Sankaranarayanan, and R. Chellappa, “MetaReg: Towards domain generalization using meta-regularization,” in Proc. Adv. Neural Inf. Process. Syst., 2018, pp. 998–1008.
[5] F. M. Carlucci, A. D’Innocente, S. Bucci, B. Caputo, and T. Tommasi, “Domain generalization by solving jigsaw puzzles,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2019, pp. 2229–2238.
[6] Q. Dou, D. C. de Castro, K. Kamnitsas, and B. Glocker, “Domain generalization via model-agnostic learning of semantic features,” in Proc. Adv. Neural Inf. Process. Syst., 2019, pp. 6450–6461.
[7] D. Li, J. Zhang, Y. Yang, C. Liu, Y.-Z. Song, and T. Hospedales, “Episodic training for domain generalization,” in Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV), Oct. 2019, pp. 1446–1455.
[38] H. Venkateswara, J. Eusebio, S. Chakraborty, and S. Panchanathan, “Deep hashing network for unsupervised domain adaptation,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2015, pp. 815–823.

[39] Y. LeCun. (1998).

[40] Y. Ganin and V. Lempitsky, “Unsupervised domain adaptation by back-propagation,” in Proc. Int. Conf. Mach. Learn., 2015, pp. 1180–1189.

[41] Y. Netzer, T. Wang, and D. Anguelov, “Reading digits in natural images with unsupervised feature learning,” in Proc. NIPS Workshop Deep Learn. Unsupervised Feature Learn., 2011.

[42] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, “ImageNet: A large-scale hierarchical image database,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., Jun. 2009, pp. 248–255.

[43] S. Zhang, M. Gong, T. Liu, H. Fu, and D. Tao, “Domain generalization via entropy regularization,” in Proc. Adv. Neural Inf. Process. Syst., vol. 33, 2020, pp. 16096–16107.

[44] Z. Huang, H. Wang, E. Xing, and D. Huang, “Self-challenging improves cross-domain generalization,” in Proc. Eur. Conf. Comput. Vis., 2020, pp. 124–140.

[45] V. N. Vapnik, “An overview of statistical learning theory,” IEEE Trans. Neural Netw., vol. 10, no. 5, pp. 988–999, Sep. 1999.

[46] S. Yan, H. Song, N. Li, L. Zou, and L. Ren, “Adaptive risk minimization: A meta-learning approach for tackling group shift,” in Proc. Adv. Neural Inf. Process. Syst., vol. 23, 2010, pp. 1335–1344.

[47] S. Arjovsky, L. Bottou, I. Gulrajani, and D. Lopez-Paz, “Invariant risk minimization,” 2019, arXiv:1907.02893.

[48] Y. Ganin, E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, M. Marchand, and V. Lempitsky, “Domain-adversarial training of neural networks,” J. Mach. Learn. Res., vol. 17, no. 1, pp. 2030–2096, 2016.

[49] G. Blanchard, A. A. Deshmukh, U. Dogan, G. Lee, and C. Scott, “Domain generalization by marginal transfer learning,” 2017, arXiv:1711.07910.

[50] M. Arjovsky, L. Bottou, I. Gulrajani, and D. Lopez-Paz, “Invariant risk minimization,” 2019, arXiv:1907.02893.

[51] S. Sagawa, P. W. Koh, T. H. Hashimoto, and P. Liang, “Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization,” 2019, arXiv:1911.08731.

[52] M. M. Zhang, H. Marklund, N. Dhallan, A. Gupta, S. Levine, and C. Finn, “Adaptive risk minimization: A meta-learning approach for tackling group shifts,” 2020, arXiv:2007.02931.

[53] S. Yan, H. Song, N. Li, L. Zou, and L. Ren, “Improving unsupervised domain adaptation with mixup training,” 2020, arXiv:2001.00677.

[54] D. Krueger, E. Caballero, J.-H. Jacobsen, A. Zhang, J. Binns, D. Zhang, R. Le Priol, and A. Courville, “Out-of-distribution generalization via risk extrapolation (REx),” in Proc. Int. Conf. Mach. Learn., 2021, pp. 5815–5826.

[55] H. Nam, H. Lee, J. Park, W. Yoon, and D. Yoo, “Reducing domain gap by reducing style bias,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2021, pp. 8690–8699.
HYUN SEOK SEONG received the B.S. degree in electronic and electrical engineering from Sungkyunkwan University (SKKU), South Korea, in 2019, where he is currently pursuing the combined M.S. and Ph.D. degrees in artificial intelligence. His research interests include metric learning for image categorization, machine learning, and deep learning.

JAEHYUN CHOI (Member, IEEE) received the B.S. degree in electrical engineering and computer science from The State University at New York at Buffalo (SUNY at Buffalo), USA, in 2019, and the M.S. degree in artificial intelligence from Sungkyunkwan University (SKKU), South Korea, in 2021. He is currently pursuing the Ph.D. degree in electrical engineering with the Korea Advanced Institute of Science and Technology (KAIST), South Korea. His research interests include transfer learning, domain generalization, computer vision, and machine learning.

WOOJIN JEONG received the B.S. degree in computer science and engineering and the M.S. degree in electrical and computer engineering from Sungkyunkwan University (SKKU), South Korea, in 2019 and 2021, respectively, where he is currently pursuing the Ph.D. degree in artificial intelligence. His research interests include content-based image retrieval, person re-identification, domain generalization, and deep learning.

JAE-PIL HEO (Member, IEEE) received the B.S., M.S., and Ph.D. degrees in computer science from the Korea Advanced Institute of Science and Technology (KAIST), in 2008, 2010, and 2015, respectively. He is currently an Assistant Professor with Sungkyunkwan University (SKKU), South Korea. Before joining SKKU, he was a Researcher with the Electronics and Telecommunications Research Institute (ETRI). His research interests include scalable algorithms for high dimensional data, computer vision, machine learning, and deep learning.

**References**

[55] I. Gulrajani and D. Lopez-Paz, “In search of lost domain generalization,” 2020, arXiv:2007.01434.

[56] C. Fang, Y. Xu, and D. N. Rockmore, “Unbiased metric learning: On the utilization of multiple datasets and web images for softening bias,” in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2013, pp. 1657–1664.