Knowledge Distillation-based Domain-invariant Representation Learning for Domain Generalization

Ziwei Niu, Junkun Yuan, Xu Ma, Yingying Xu, Jing Liu, *Member, IEEE*
Yen-Wei Chen, *Member, IEEE*, Ruofeng Tong, Lanfen Lin✉, *Member, IEEE*

***Abstract*—Domain generalization (DG) aims to generalize the knowledge learned from multiple source domains to unseen target domains. Existing DG techniques can be subsumed under two broad categories, i.e., domain-invariant representation learning and domain manipulation. Nevertheless, it is extremely difficult to explicitly augment or generate the unseen target data. And when source domain variety increases, developing a domain-invariant model by simply aligning more domain-specific information becomes more challenging. In this paper, we propose a simple yet effective method for domain generalization, named Knowledge Distillation based Domain-invariant Representation Learning (KDDRL), that learns domain-invariant representation while encouraging the model to maintain domain-specific features, which recently turned out to be effective for domain generalization. To this end, our method incorporates multiple auxiliary student models and one student leader model to perform a two-stage distillation. In the first-stage distillation, each domain-specific auxiliary student treats the ensemble of other auxiliary students' predictions as a target, which helps to excavate the domain-invariant representation. Also, we present an error removal module to prevent the transfer of faulty information by eliminating incorrect predictions compared to the true labels. In the second-stage distillation, the student leader model with domain-specific features combines the domain-invariant representation learned from the group of auxiliary students to make the final prediction. Extensive experiments and in-depth analysis on popular DG benchmark datasets demonstrate that our KDDRL significantly outperforms the current state-of-the-art methods.**

***Index Terms*—Domain generalization, knowledge distillation, domain invariant representation.**

I. Introduction

DEEP learning methods have demonstrated exceptional progress in various fields over the past few years. However, the performance of these deep learning systems can be significantly degraded when encountering test data under different distributions from the training data. Such an issue is known as the domain shift problem. Recently, researchers have developed a number of Unsupervised Domain Adaptation (UDA) [1]–[5] and Domain Generalization (DG) [6]–[10] methods to address the domain shift problem. UDA aims to bridge the gap between source domain(s) and a specific target domain with the help of unlabeled target data. But for many real-world application scenarios, target data collection could be challenging and impossible. Additionally, the trained UDA model must be retrained or fine-tuned using the unlabeled target dataset before being deployed for entirely new scenario (i.e., a new target domain).

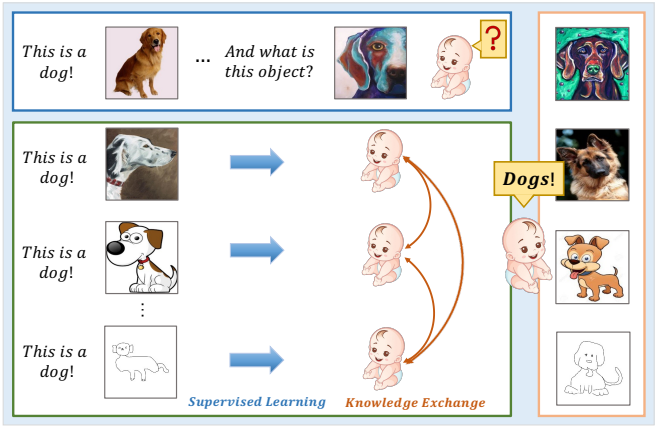


Fig. 1. An illustration of children learning for an unknown domain. Figure illustrates how it might be useful for children to interact with one another because information from several domains is shared and stored, making it easier to understand object variations and patterns. This strategy enhances the ability to generalize to unknown domains.

DG is considered as a more challenging but realistic task than UDA since it does not require target samples during the training phase. Regarding different strategies for transferring knowledge from source domains to the unseen target domain, existing DG technologies can be broadly classified into two categories, namely, domain-invariant representation learning [6], [11], [12] and domain manipulation [7], [9], [10]. Technically, domain-invariant representation learning methods aim to reach the consensus from multiple source domains and learn a domain-invariant latent representation for the target domain. However, when source domain variety increases, developing a domain invariant model becomes more challenging. This is because simply aligning more domain-specific information may be detrimental to model generalization. Whereas do-

This work was supported in part by the Major Technological Innovation Project of Hangzhou (No. 2022AIZD0147), the Zhejiang Provincial Natural Science Foundation of China (No. LZ22F020012), the China Postdoctoral Science Foundation (No.2020TQ0293), the Postdoctor Research from Zhejiang Province under Grant ZJ2021028,the Major Scientific Research Project of Zhejiang Lab (No. 2020ND8AD01), and the Japanese Ministry for Education, Science, Culture and Sports (No. 20KK0234, No. 21H03470 and No. 20K21821).

Ziwei Niu, Junkun Yuan, Xu Ma, Ruofeng Tong, and Lanfen Lin are with the College of Computer Science and Technology, Zhejiang University, Hangzhou 310027, China (e-mail: nzw@zju.edu.cn; yuanjk@zju.edu.cn; maxu@zju.edu.cn; llf@zju.edu.cn; trf@zju.edu.cn).

Yingying Xu, Jing Liu are with Research Center for Healthcare Data Science, Zhejiang Lab, Hangzhou 310027, China (e-mail: cs_ying@zju.edu.cn; liujinglj@zhejianglab.edu.cn).

Yen-Wei Chen is with the College of Information Science and Engineering, Ritsumeikan University, Kyoto 603-8577, Japan (e-mail: chen@is.ritsumei.ac.jp).

(*Corresponding author*: Lanfen Lin)

main manipulation techniques have an attention on augmenting the source samples or generating the pseudo-novel samples to assist in learning general representations. Although the data manipulation-based approaches perform well, it is extremely challenging to anticipate the distribution of the unseen test data to generate additional training samples due to large domain discrepancies and other complex factors.

To address the limitations of these current methods, we approach domain generalization from a novel perspective that is motivated by an intriguing scenario. We believe that each model can learn unique knowledge from the data alone. We hope to transfer knowledge between models, so as to further improve the performance of each model through knowledge sharing. An interesting example is shown in Figure 1. Children are trained to recognize different photos of dogs. Any child will soon be able to identify dogs from the given images after some experience. However, when a clild receive a collection of cartoon animal drawings from any cartoon book. The gap between real images and cartoon illustrations might make it challenging for the child to recognize and differentiate them. On the other hand, some children will certainly have no trouble classifying cartoon animals. As these children interact with one another, they will convey and incorporate the knowledge they have learnt, further they can properly classify animal representations in a variety of depictions such as line drawing and oil painting, etc. This process verifies our motivation that knowledge exchange between different individuals helps learning the regularity and generality of objects.

Based on this insight, we present Knowledge Distillation-based Domain-Invariant Representation Learning (KDDRL), a simple yet effective approach for domain generalization that is most relevant to the category of domain-invariant learning-based DG methods. Different from previous works that align feature distributions of multi-source domains to learn a generalized model, we design a multi-student network to learn domain-specific knowledge and perform two-stage distillation between the output distribution of all students to learn domain-invariant representation. The fine-grained output distributions following a high level abstraction (temperature) offer more interclass relations and semantic information than the complex feature distributions, making them more suitable for learning domain invariant representation.

The core idea of KDDRL is to communicate effectively with all domains to extract common information while discarding atypical information to generalize efficiently on unseen domains. To this end, our KDDRL establishes a multi-student network that involves multiple auxiliary student models and one student leader model. Each auxiliary student is assigned to a specific domain to learn different types of knowledge, while the student leader is used to acquire the combined information across all domains. However, training an ensemble of models instead of a single model means higher computational cost. To solve this problem, we design all auxiliary student models consisting of a shared convolutional neural network (CNN) feature extractor and multiple classifier heads. Each head is trained to classify images from a particular source domain. Therefore, different heads learn different patterns from the shared features for classification. To collaborate with all the domains and learn domain-invariant representation, a two-stage knowledge distillation is then performed. In the first-stage distillation, each auxiliary student derives its own target distribution from the ensemble of other auxiliary peers, which helps to eliminate the style information and emphasize the general semantic information. These steps increase model's resilience to semantic similarity information and make them insensitive to the changes in style attributes any longer. In addition, we have designed an error removal module to ensure the correctness and efficiency of knowledge transfer. Wrong predictions are removed during training by comparing auxiliary student model predictions with ground truth labels. The second-stage distillation is then performed to aggregate the knowledge in the ensemble of auxiliary students further to the student leader, i.e., the model for final prediction, which helps to ensure the stability of reasoning model. This strategy makes up for the performance degradation of the auxiliary student models in corresponding domains after the implementation of the first-stage distillation.

The main contributions of our work are highlighted as follows:

(1) We present a novel approach called KDDRL to address the issue of domain generalization, which establishes a multi-student network and performs a two-stage knowledge distillation procedure. The domain-specific student leader contains specific characteristics and when combined with domain-invariant information learned from auxiliary students can significantly promote performance on unseen domains.

(2) To solve the issue of wrong information transmission during distillation process, we devise an error removal module that compares the predicted outputs with the actual labels rather than simply distilling the predicted outputs directly.

(3) To demonstrate the merit of the proposed KDDRL framework, we extensively evaluate it on several state-of-the-art DG benchmark datasets, including PACS, VLCS, Office-Home, Digit-DG, Terra Incognita, and DomainNet. The results demonstrate that KDDRL significantly outperforms the current state-of-the-art methods.

II. Related Work

A. Domain shift

Most existing machine learning systems will suffer from performance degradation when encountering test (i.e., target) domain data, that is statistically different from the training (i.e., source) domain data, this discrepancy is known as domain shift [13]. In the field of computer vision, this domain difference is occurred due to the changes in background, style, and lighting conditions of an image as well as differences in the pose and position of objects in the image. To solve the issue of domain shifts, unsupervised domain adaptation was proposed, which aims to adapt the model trained on a labeled source domain to an unlabeled target domain. To date, most UDA methods are focused on feature distribution alignment. One popular line of approaches includes learning a domain invariant representation by minimizing the distributional shift between source and target feature distributions using MMD-based loss [14], [15] or adversarial loss [16], [17]. With the development of Generative Adversarial Networks (GANs) [18],

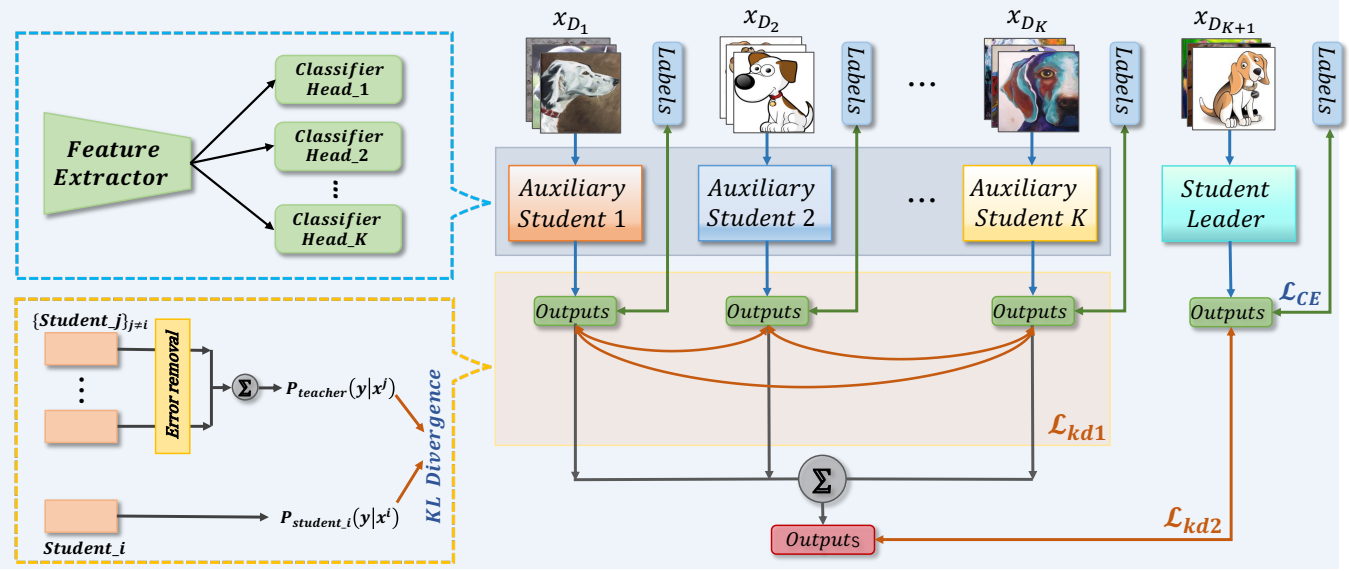


Fig. 2. The overall framework of the proposed method for DG: KDDRL adopts a multi-students network as shown in the blue dotted box, each student is trained to be specialized in a particular source domain. At the first-stage distillation, each auxiliary student derives its own target distribution from the ensemble of other auxiliary peers as shown in the yellow dotted box. The second-stage distillation is performed to transfer the knowledge in the ensemble of auxiliary students to the student leader, which is the model for final prediction.

additional papers [1], [16] were proposed to perform domain alignment in the feature space with adversarial learning. Then image translation-based methods aim to translate images from source domains to target domains to mitigate domain gaps [2], [3], [19]. More recently, few-shot DA [4], [5], [20] was proposed, where only a few labeled target samples together with source samples are available in the training phase.

B. Domain generalization

The most significant difference between UDA and DG is that DG will no longer require the utilization of target domain data or information, which makes it more challenging yet more practical. Existing DG methods can be broadly classified into two categories: domain-invariant representation learning [6], [15], [21]–[24] and domain manipulation [7]–[10]. The primary focus of domain-invariant representation learning is to reduce the divergence in representation between different source domains using domain-invariant feature learning. For example, Maundet et al. [22] first proposed a kernel-based method to obtain domain invariant features. Li et al. [15] reduce the domain gap across the domains by minimizing Maximum Mean Discrepancy (MMD) under adversarial autoencoder framework. Zhou et al. [24] explored domain invariant representation by performing explicit distribution alignment.

The other category is related to data manipulation. This line of work typically aims to simulate the invisible target domain data through data augmentation or data generation, which are then used to train the network along with the source samples to improve the generalization ability. For example, domain randomization [25] is an effective method of data augmentation by diversifying the training domain data using image space [26], feature space [27] and frequency space [28] to simulate deviations in color, texture, background, lighting conditions of any object. Data generation can be implemented utilizing techniques like the Variational Auto-encoder (VAE) [29], Wasserstein Auto-Encoder (WAE) [30] and Generative Adversative Network (GAN) [18]. Recently Mixup [31] has also emerged as a simple and efficient method of data generation in the DG field, generating new samples by performing mixing in the original space [32], feature space [10], or frequency space [33].

More recently, some learning strategy-based methods have also proved to be very effective. For instance, Qiao et al. [34] adopted meta-learning to divide the source domain into meta-train and meta-test at each iteration to stimulate domain shift. Bui et al. [35] introduced the use of the meta-training scheme to support domain-specific to adapt information from source domains to unseen domains. Li et al. [36] utilized ensemble learning to prove that a network is more robust to distribution shifts if its architecture aligns well with the invariant correlation. Li et al. [37] designed an ensemble network aware of model specialty is proposed to dynamically dispatch proper pre-trained models to predict each test sample. Lv et al. [38] utilized causality learning and introduced a general structural causal model to formalize the DG problem.

C. Knowledge distillation

Knowledge distillation is a common approach to transferring knowledge, which aims to train a model (student) by transferring knowledge extracted from another model (teacher) that is more powerful than the student. The idea of knowledge distillation is first introduced in [39]. They introduce soft targets associated with complex, but superior predictive accuracy teacher networks to induce the training of streamlined, low-complexity student networks that are more suitable for

reasoning and deployment. By introducing the KL divergence loss between the output of the teacher network and the streamlined network, the author redefined the global loss in [40]. It is proposed in [41] to train student models for knowledge refinement utilizing numerous teachers and multiple teacher labels. In the absence of a strong teacher model, the targets derived from a group of student models play a crucial role in knowledge transfer. Later, a novel two-level framework for effective online distillation is proposed in [42] with multiple distinct peers that is largely an inspiration to us.

III. Methodology

A. Problem definition.

In the typical setting of DG, giving K source domains $D_S = \{D_1, D_2, D_3, \cdots, D_K\}$, where each domain contains N_k labeled samples $\left\{\left(x_i^k, y_i^k\right)\right\}_{i=1}^{N_k}$. The goal of domain generalization is to train a model only with the source domains but generalize well on an arbitrary unseen target domain D_T. In this work, we have a specific domain D_{Mix} by mixing all the source domains, and for convenience, we represent it as the $(K+1)$-th domain, i.e., D_{K+1}.

B. Model design.

We design the KDDRL as a multi-student network with two-stage knowledge distillation as illustrated in Figure 2. In implementation, the multiple auxiliary student models share a CNN backbone for feature extraction, followed by domain-specific classification heads. The inputs of all students are sampled from various domains but share the same class.

For the first-stage distillation, each auxiliary student model for a specific domain derives its own target distribution from the average of other auxiliary peers' distributions. Meanwhile, to ensure the correctness and validity of the transmitted knowledge, we introduce an error removal module that compares the prediction results with the real labels. The wrong predictions will be removed to ensure the correctness of the delivered knowledge.

Finally, the second-stage distillation is performed to transfer the knowledge in the ensemble of auxiliary students further to the student leader which is used for inference.

C. Training all students with labeled data

Given images x^k from the k-th source domain, no matter the auxiliary students or the student leader, the primary goal is to learn a mapping by minimizing the cross entropy loss between the predicted class probabilities and the one-hot ground-truth label distribution. Let $CE(.,.)$ denote the cross entropy between two probability distributions, the loss function for all the students learning is

$$\mathcal{L}_{CE} = -\sum_{k=1}^{K+1} \mathbb{E}_{(x^k, y^k)\in\mathcal{D}_k}\left[CE\left(p_i^k, y\left(x^k\right)\right)\right], \tag{1}$$

where $y\left(x^k\right)$ is the one-hot ground-truth label distribution, $p_i^k = \sigma\left(z_i^k, \tau\right)$ is calculated with softmax of logits z_i^k, i.e., outputs of the last fully connected layer.

$$p_i^k = \frac{\exp\left(z_i^k/\tau\right)}{\sum_j \exp\left(z_j^k/\tau\right)}, \tag{2}$$

and parameter τ is usually set to 1.

D. First-stage distillation between auxiliary students

For the basic knowledge distillation, knowledge is transferred by aligning the student-predicted distribution to the teacher-predicted soft distribution (or target distribution) after a softmax with the same temperature τ. A higher τ means a softer distribution. Compare to hard ground-truth labels, fine-grained class information in soft predictions helps the student models to reach flatter local minima, which results in more robust performance and improves generalization ability [42]. In this paper, the value of τ is set to 2, which will be further discussed in the sensitivity analysis section.

During the first-stage knowledge distillation, with a population of K auxiliary students for k=1,2,...,K. Each auxiliary student model derives its own target distribution from the average of other auxiliary peers' distributions for further knowledge distillation. By doing so, it enables a model to learn domain-invariant features by enforcing prediction consistency between the data with the same label but from different domains. Specifically, we take turns picking one auxiliary student model's predicted distribution p^i as the student distribution, the average of all remaining auxiliary students' predicted distribution $\frac{1}{K-1}\sum_{j\neq i} p^j$ is the teacher (target) distribution. We align the student distribution to the ensemble distribution and employ the conventional Kullback-Leibler (KL) divergence as the distillation loss to transfer domain-specific information:

$$\mathcal{L}_{kd1} = \sum_{i=1}^{K} KL\left(\frac{1}{N-1}\sum_{j\neq i} p^j, p^i\right). \tag{3}$$

E. Error removal module

Each auxiliary student model has been considered as a "teacher" model and used to generate target distribution during the first step of distillation. Considering the fact that the "teacher" models lack a robust network in comparison to the actual sense of knowledge distillation. The wrong predictions may lead to the wrong ensemble consensus knowledge and further result in learning the wrong invariant representation, especially in the early stage of training. We design an error removal module to make sure the knowledge being delivered is accurate and effective. The predictions of the "teachers" are compared with the ground truth labels. Correct predicted distributions are kept and distilled, and incorrect ones are eliminated. The pseudo-code for error removal is shown in Algorithm. 1.

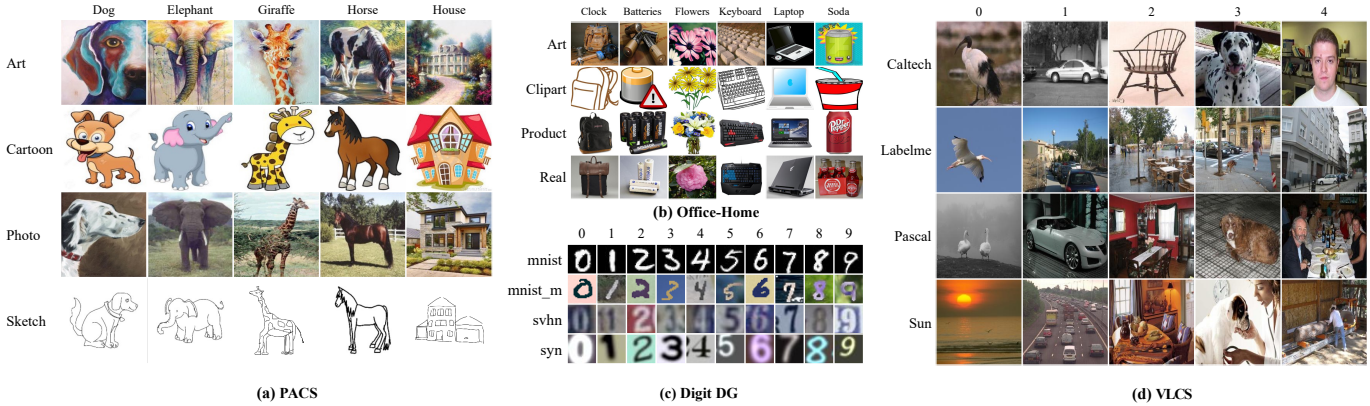


Fig. 3. Some example images of the adopted datasets for experiments, i.e., PACS (a), Office-Home (b), Digit-DG (c), and VLCS (d).

Algorithm 1 Pseudo-code for Error Removal.

Require: labeled source mini-batches $\{(X^i_{m\in[1,M]}, Y^i_{m\in[1,M]})\}^K_{i=1}$ with the same class Y, local models $\{M_i\}^K_{i=1}$.
Return: target distribution p_t.
$out\_list = []$ *// Initialize out_list for temporary storage of the local models' predictions*
// Choosing the k-th local model as student model
$p_k = M_k(X^k)$
for $j \neq k$ to K **do**
  $p_j = M_j(X^j)$ *// Calculating the predictions for the rest models*
  for $m = 1$ to M **do**
    if $argmax(p_j[m]) \neq argmax(Y_m)$ **then**
      $p_j[m] = 0$ *// Set incorrect predictions to 0 for removing error knowledge*
    end if
  end for
  $out\_list.append(p_j)$ *// Storing the correct predictions temporarily*
end for
$count = 0$ *// Initialize count for counting the number of correct predictions*
$count\_list = []$ *// Initialize count_list for storing the number of correct predictions for each model*
for $m = 1$ to M **do**
  for $i = 1$ to $len(out_list)$ **do**
    if $out\_list[i][m] \neq 0$ **then**
      $count+ = 1$ *// Counting the number of correct predictions*
    end if
  end for
  $count\_list.append(count)$ *// Storing the number of correct predictions for each model temporarily*
end for
$p_t = out\_list/count\_list$ *// Calculating the target distribution*

F. Second-stage distillation between auxiliary students and the student leader

The knowledge of the group (auxiliary student models) is further distilled to the student leader model to make the final prediction through the second-stage distillation. On the one hand, it is not reasonable to select a student arbitrarily or to use the student ensemble for prediction, on the other hand, previous work [35] indicates that domain invariant representation combined with domain-specific information can better generalize to the unseen target domain, while the domain-invariant representation learning during the first-stage distillation weakens their ability to distinguish domain-specific information. Thus, it is necessary to implement the second-stage distillation, i.e., the knowledge transfer from the auxiliary student's ensemble to the student leader. To be more precise, the knowledge of the auxiliary students' ensemble is just the mean of all their predictions. The second-stage distillation loss is

$$\mathcal{L}_{kd2} = KL\left(\frac{1}{K}\sum_{i=1}^{K} p^i, p^{K+1}\right), \tag{4}$$

where p_{K+1} is the predicted output of the student leader.

The full learning objective is a weighted sum of Eq. (1), (3) and (4),

$$\mathcal{L}_{Full} = (1-\alpha)\mathcal{L}_{CE} + \alpha\left(\tau^2\mathcal{L}_{kd1} + \tau^2\mathcal{L}_{kd2}\right). \tag{5}$$

α as a hyperparameter tunes the weighted average between two components of the loss. The first component of the total loss forces the student's predicted distribution closer towards the ground truth labels, while the second part forces it closer towards the soft target from temporary teacher models.

IV. Experiments

In this section, we demonstrate the superiority of our method on several DG benchmarks. We also conduct extensive experiments on the DomainBed which is a testbed for domain generalization.

A. Datasets and Settings

1) Datasets: To evaluate the efficacy of our method, we perform extensive experiments on four classic and frequently used domain generalization benchmarks: PACS [43], VLCS [44], Office-home [45], and Digit-DG [7]. Example images of these adopted datasets are given in Figure 4.

-PACS: consists of 9991 images from photo, art painting, cartoon, and sketch domains with 7 categories: dog, elephant, giraffe, guitar, horse, house, and person.

-VLCS: consists images with 5 categories over four domains, which are collected from the PASCAL VOC 2007, LabelMe, Caltech, and Sun datasets.

-Office-Home: contains four domains: Art, Clipart, Product, and Real-World. Each domain consists of 65 object categories with around 15500 images in total.

-Digit-DG: contains 4 different digit datasets including MNIST, MNIST_M, SVHN, and SYN, which are different in font style, stroke color, and background.

2) Evaluation Protocol: For evaluation, we follow the prior works [9], [46] to use the leave-one-domain-out protocol, i.e., one domain is chosen as 'unseen' target domain while the remaining domains are treated as source domains during model training. For the division of training and validation data, we use the same splits of the original datasets. Meanwhile, to ensure the fairness and confidence of results, we run each experiment 5 times and report the average accuracy as the final result.

B. Implementation Details

We follow the implementations of latest works [9], [46]. For PACS and Office-Home benchmark dataset, Resnet18 [47] is used as the CNN backbone and add Resnet50 [47] for Office-Home. We optimize the model by using SGD with batch size 40, momentum 0.9. The learning rate starts with $2e-3$ for Resnet18 and Resnet50, which is gradually decreased by using the cosine annealing schedule. For VLCS, we use Alexnet [48] (ImageNet [49] pretrained) on this benchmark. We train the model using the same optimization strategy and parameters as the above two benchmarks. As for Digit-DG, we adopt the same network architecture as [9], [26]. The SGD with a learning rate of 0.05 is used to optimize the model.

C. Comparisons with Other Methods

Comparisons on PACS. The results are shown in Table 1. Among all competitors, our method achieves the best average accuracy. However, we notice that naive DeepAll baseline performs well than KDDRL on the photo domain, this is because, on the one hand, the photo domain is similar to the pretrained dataset ImageNet. On the other hand, the precision of each student model will be reduced in the process of knowledge exchange. What's more, it is noticeable that KDDRL boosts the performance significantly on the Sketch domain with Resnet18 and Resnet50, which is the only colorless domain. The model may have to understand the semantics of objects to perform well on the sketch domain, which indicates that our proposed KDDRL method summarizes domain invariant information from multiple domains during knowledge distillation.

Comparisons on VLCS. We report the results in Table 2, it was found that our method did not perform very well, as it did on several other datasets. By observing the VLCS dataset, the images of each category will more or less contain objects of other categories, which will disturb the learning of domain invariant representation. Nevertheless, our KDDRL still achieves the comparable performance to all the recent DG methods except for EISNet [12].

Comparisons on Office-Home. From the results in Table 3. Among all the competitors, our method achieves the best performance on average accuracy as well as on the Clipart domain, and ranks second on the domain of Art and Product, which is slightly lower than L2A-OT [26]. Significantly, our method performs extremely well on the Clipart domain which consists of many colorless and simple line drawings. This is a further indication that KDDRL improves generalization by learning domain invariant representations.

TABLE 1
LEAVE-ONE-DOMAIN-OUT RESULTS ON PACS DATASET WITH RESNET18 AND RESNET50. THE BEST AND SECOND-BEST RESULTS ARE BOLDED AND UNDERLINED RESPECTIVELY

Methods	Art painting	Cartoon	Photo	Sketch	Avg.
Resnet18					
DeepAll	77.36	75.66	95.85	69.30	79.54
MedtaReg [50]	<u>83.70</u>	77.20	95.50	69.50	81.70
JiGen [6]	79.42	75.25	96.03	71.35	80.51
MASF [11]	80.29	77.17	94.99	71.69	81.03
Epi-FCR [51]	82.10	77.00	93.90	73.00	81.50
DDAIG [7]	84.20	78.10	95.30	74.70	83.10
EISNet [12]	81.89	76.44	95.93	74.33	82.15
L2A-OT [26]	83.30	<u>78.20</u>	**96.20**	73.60	82.80
DAEL [52]	84.60	74.40	95.60	<u>78.90</u>	83.40
SFA [9]	81.20	77.80	93.90	73.70	81.70
MixStyle [10]	**84.10**	78.80	<u>96.10</u>	75.90	<u>83.70</u>
KDDRL(ours)	82.26	**78.88**	95.61	**82.16**	**84.73**
Resnet50					
DeepAll	85.24	76.64	**97.64**	75.02	83.64
MetaReg [50]	**87.20**	79.20	<u>97.60</u>	70.30	83.60
MASF [11]	82.89	80.49	95.01	72.29	82.67
EISNet [12]	<u>86.64</u>	**81.53**	97.11	78.07	<u>85.84</u>
DAEL [52]	84.32	80.56	95.68	<u>82.79</u>	85.83
ERM [53]	84.87	80.80	97.20	79.30	85.50
KDDRL(ours)	85.55	<u>80.86</u>	96.04	**84.16**	**86.65**

TABLE 2
LEAVE-ONE-DOMAIN-OUT RESULTS ON VLCS DATASET WITH ALEXNET. THE BEST AND SECOND-BEST RESULTS ARE BOLDED AND UNDERLINED RESPECTIVELY

Methods	PASCAL	LabelMe	Caltech	Sun	Avg.
DeepAll	71.58	59.32	95.35	63.81	72.77
CCSA [54]	67.10	62.10	92.30	59.10	70.20
MMD-AAE [15]	67.70	62.60	94.40	64.40	72.30
JiGen [6]	70.62	60.90	96.93	64.30	73.19
S-MLDG [55]	68.70	<u>64.80</u>	96.40	64.00	73.50
MASF [11]	69.14	**64.90**	94.78	67.64	74.11
Epi-FCR [51]	67.10	64.30	94.10	65.90	72.90
EISNet [12]	69.83	63.49	<u>97.33</u>	<u>68.02</u>	**74.67**
DAEL [52]	68.73	60.12	96.98	66.47	73.08
SFA [9]	**70.40**	62.00	97.20	66.20	74.00
KDDRL(ours)	<u>69.85</u>	62.60	**97.37**	**68.08**	<u>74.48</u>

Comparisons on Digit-DG. The results are shown in Table 4. It is worth noting that our method achieves the best performance in the MNIST-M domain, SYN domain, and average accuracy. Especially on the SYN domain, one of the most difficult target domains, where involves clustered digits and low-quality images, KDDRL has a large margin of 5% compared with the second best method.

D. Evaluation on DomainBed

We also conduct extensive experiments on the DomainBed [53] which is a testbed for domain generalization to compare state-of-the-art methods across several benchmark datasets. The rationale behind the DomainBed is that the domain generalization performances are too much dependent

TABLE 3
LEAVE-ONE-DOMAIN-OUT RESULTS ON OFFICE-HOME DATASET WITH RESNET18. THE BEST AND SECOND-BEST RESULTS ARE BOLDED AND UNDERLINED RESPECTIVELY

Methods	Art	Clipart	Product	Real	Avg.
DeepAll	57.88	52.72	73.50	74.80	64.72
CCSA [54]	59.90	49.90	74.10	75.70	64.90
MMD-AAE [15]	56.50	47.30	72.10	74.80	62.70
CrossGrad [56]	58.40	49.40	73.90	75.80	64.40
JiGen [6]	53.04	47.51	71.47	72.79	61.20
DAEL [52]	59.40	<u>55.10</u>	74.00	75.70	<u>66.10</u>
L2A-OT [26]	**60.60**	50.10	<u>74.80</u>	**77.00**	65.60
DDAIG [7]	59.20	52.30	74.60	<u>76.00</u>	65.50
KDDRL(ours)	<u>59.42</u>	**55.12**	**74.83**	75.36	**66.18**

TABLE 4
LEAVE-ONE-DOMAIN-OUT RESULTS ON DIGIT-DG DATASET. THE BEST AND SECOND-BEST RESULTS ARE BOLDED AND UNDERLINED RESPECTIVELY

Methods	MNIST	MNIST-M	SVHN	SYN	Avg.
DeepAll	95.8	58.8	61.7	78.6	73.7
CCSA [54]	95.2	58.2	65.5	79.1	74.5
MMD-AAE [15]	96.5	58.4	65.0	78.4	74.6
CrossGrad [56]	<u>96.7</u>	61.1	65.3	80.2	75.8
JiGen [6]	96.5	61.4	63.7	74.0	73.9
L2A-OT [26]	**96.7**	63.9	68.6	<u>83.2</u>	78.1
DDAIG [7]	96.6	64.1	68.6	81.0	77.6
SFA [9]	96.5	<u>66.5</u>	<u>70.3</u>	80.5	<u>79.6</u>
KDDRL(ours)	96.6	**68.9**	**71.1**	**88.2**	**81.5**

on the hyperparameter tuning. For a fair comparison, we follow its standard protocols for training and evaluation. The results are shown in Table 5, our method generally shows competitive performances and ranks second out of 15 methods on average accuracy.

TABLE 5
DOMAIN GENERALIZATION ACCURACY (%) ON DOMAINBED. THE COLUMN "DNET" STANDS FOR DOMAINNET DATASET, THE COLUMN "TERRA" STANDS FOR TERRA INCOGNITA DATASET. NOTE THAT WE ADOPT LEAVE-ONE-DOMAIN OUT CROSS-VALIDATION AS A MODEL SELECTION CRITERIA. THE BEST AND SECOND-BEST RESULTS ARE BOLDED AND UNDERLINED RESPECTIVELY.

Methods	PACS	VLCS	OfficeHome	DNet	Terra	Avg.
ERM [57]	83.2	77.3	65.4	40.6	46.6	62.6
IRM [58]	81.9	76.6	64.7	34.8	46.2	60.8
GroupDRO [59]	83.5	77.2	65.5	33.4	44.3	60.8
Mixup [31]	83.2	<u>77.9</u>	67.8	38.8	47.9	63.1
MLDG [60]	83.0	77.3	66.3	<u>41.0</u>	46.7	62.9
CORAL [61]	82.4	**78.7**	**68.6**	**41.2**	47.2	**63.6**
MMD [15]	83.5	77.6	61.7	26.9	43.5	58.6
DANN [17]	81.1	76.4	65.0	38.5	40.0	60.2
CDANN [62]	79.5	77.6	64.8	37.8	38.6	59.6
MTL [63]	<u>83.6</u>	76.0	66.0	40.1	41.8	61.5
SagNet [64]	82.9	76.5	<u>67.8</u>	39.4	**48.3**	63.0
ARM [65]	82.7	76.3	64.4	35.6	43.3	60.5
VREx [66]	81.2	76.9	65.9	33.2	45.8	60.6
RSC [46]	83.6	77.5	65.8	38.5	46.0	61.7
KDDRL(ours)	**83.7**	77.3	66.9	39.6	<u>48.0</u>	<u>63.1</u>

TABLE 6
RESULTS OF THE ABLATION STUDY IN THE PACS DATASET WITH RESNET18

Method	$\mathcal{L}_{kd1}$	$\mathcal{L}_{kd2}$	Er	PACS	VLCS	Office-Home	Digit-DG
Baseline	-	-	-	79.54	72.77	64.72	73.70
Model A	✓	-	-	82.87	73.44	65.03	78.80
Model B	✓	-	✓	83.03	73.96	65.58	79.60
Model C	✓	✓	-	84.21	74.16	66.06	80.60
Model D	-	✓	-	80.58	72.92	64.80	76.40
KDDRL	✓	✓	✓	**84.73**	**74.48**	**66.18**	**81.50**

V. FURTHER ANALYSIS

Ablation Study. In this section, we have conducted extensive ablation experiments on four datasets to investigate the role of each component in our KDDRL model in Table 6. Starting from the baseline, model A is trained with the first-stage distillation and already works better than the baseline, which indicates that the first-stage distillation between auxiliary students is helpful to learn domain-invariant representation. Based on model A, we add an error removal module to obtain model B and add the second-stage distillation to obtain model C, which fully proves their effectiveness. We also create model D by only using the second-stage knowledge, which is a related baseline of traditional knowledge distillation that the performance is not significantly improved compared with the baseline. On the whole, our full KDDRL performs the best, which verifies the effectiveness of each component.

Parameter Sensitivity. To validate the significance of distillation temperature τ and the weight factor parameter α in the loss, we perform a sensitivity study on these hyperparameters, the results of which are displayed in Figure 4. In general, a temperature $\tau > 1$ can cause the predicted distribution to contain more information and lead to a favorable performance. Figure 4(a) illustrates that KDDRL is not sensitive to variations in τ and shows only slight fluctuations in average accuracy. When $\tau = 2 \sim 5$, the average performance gradually declines, primarily because the information in predictions is over-smoothed, confusing the model's decision-making process. As a result, we chose $\tau = 2$ as the temperature for all experiments. Figure 4(b) demonstrates that for the hyperparameter α, a weight of about 0.5 is optimal. KDDRL is not sensitive to the exact configuration of this hyperparameter as it outperforms the DeepAll baseline across all nine variables values of α.

Computational Cost. Our KDDRL designs all auxiliary student models consisting of a shared convolutional neural network (CNN) feature extractor and multiple classifier heads to reduce computational costs. However, as shown in Figure 5, compared with other types of methods, our KDDRL is not superior in terms of parameter quantity and running time, but they are within the acceptable range. At present, more works have proved the effectiveness of the ensemble-based approach, and they are also actively exploring how to reduce the computing requirements.

TABLE 7
FURTHER ANALYSIS OF MCKD ON OFFICE-HOME AND DIGIT-DG. OFFICE MEANS OFFICE-HOME DATASET, DIGIT MEANS DIGIT-DG DATASET.

	Office	Digit
Erm.	65.74	81.50
Erp.	64.86	80.30

(a) Error Removal vs. Error Replacement.

	Office	Digit
Col.	65.74	81.50
Ind.	63.27	80.04

(b) Collaborative vs. Individual distillation.

	Office	Digit
Ensemble.	65.74	81.50
Single.	64.86	80.90

(c) Single student vs. Students' ensemble.

	Office	Digit
Leader.	65.74	81.50
Ensemble.	65.29	81.20

(d) Student leader vs. Auxiliary students' ensemble.

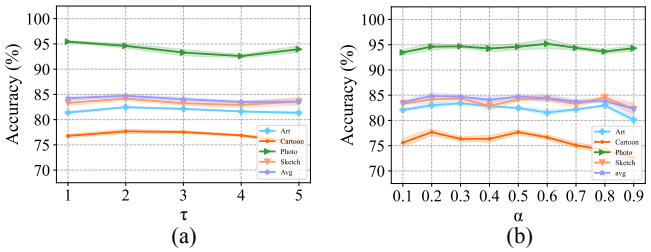


Fig. 4. Parameter sensitivity analysis of distillation temperature T and loss weight α.

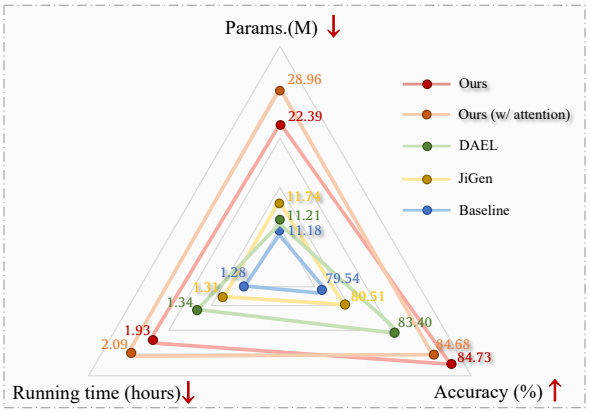


Fig. 5. Comprehensive evaluation of the parameter quantity, running time, and accuracy of KDDRL on the PACS dataset.

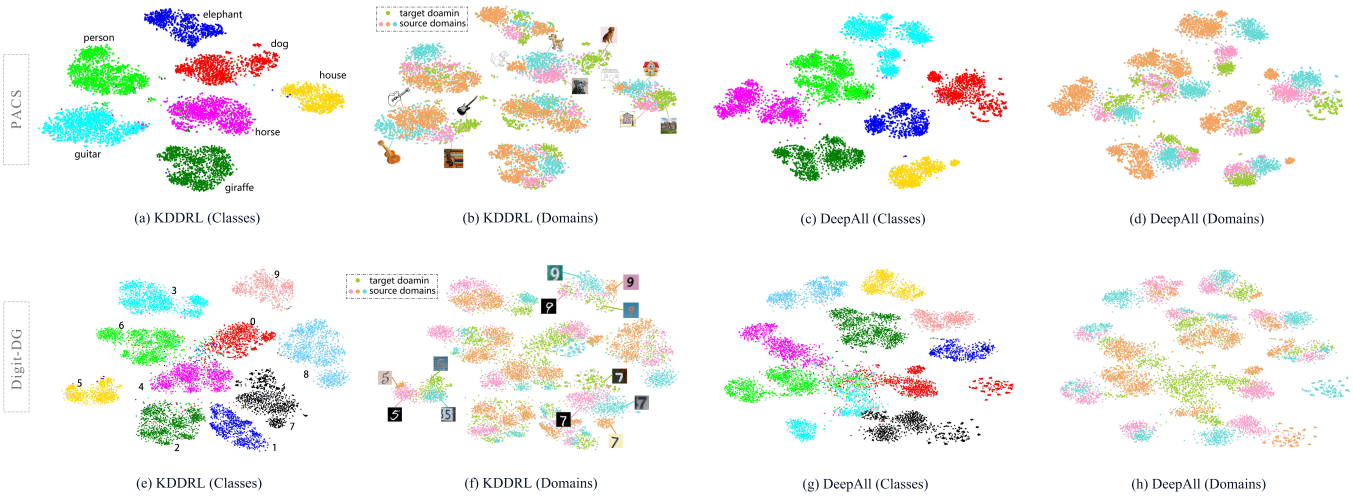
Error removal or error replacement? In addition to eliminating incorrect predictions immediately, replacing incorrect predicted distributions is also considered by a one-hot distribution of ground truth labels. We conduct experiments using the two datasets depicted in Table 7(a), using replacement is worse than using removal. This is because hard one-hot distributions will weaken the relations between classes reflected in the soft distribution after the averaging operation. In addition, we also added an experiment to replace the error elimination module with the attention module as shown in Figure 5, but it did not achieve the best performance and increased the amount of computation.

Collaborative distillation or individual distillation? Table 7(b) shows that collaborative distillation $\left(KL\left(p^t, \frac{1}{K}\sum_{k=1}^{K} p^k\right)\right)$ is better than individual distillation $\left(\frac{1}{K}\sum_{k=1}^{K} KL\left(p^t, p^k\right)\right)$. A similar approach in DAEL [52] explains that collaborative learning aggregates gradients from different models, which can better exploit the complementarity between different sources and further enhance generalization capability.

Who is more suitable for the temporary teacher? During the first-stage distillation, it also appears plausible that each auxiliary student model provides guidance for the rest of the auxiliary student models of the group to convey diverse knowledge. Table 7(c) discusses who is more reasonable as a temporary teacher. The results show that the ensemble of the remaining auxiliary students performs better as a teacher to provide guidance. This is mainly because one auxiliary student model is selected in turn, and the ensemble of all the remaining auxiliary student models is used to guide its progress. Till the end, each auxiliary student model is guided to achieve common learning progress.

Why second-stage distillation? Previous work [35] indicates that domain invariant representation combined with domain-specific information can better generalize to the unseen target domain. In our KDDRL, multiple domain-specific auxiliary student models perform distillation with each other to learn domain-invariant representation by aligning their soft predictions to the ensemble consensus. However, in the process of domain-invariant representation learning, their ability to distinguish domain-specific information is weakened. So we carry out the second-stage knowledge distillation to make up for this loss. No matter the results in Table 7(d), or the result in the ablation study, amply demonstrates that the student leader model following the second-stage distillation is more capable of accurate prediction than the group of auxiliary student models predictions.

Visualization of extracted features. We utilize t-SNE [67] to analyze the feature space learned with our proposed model KDDRL and DeepAll baseline method. The results shown in Figure 5. Figures (a)(c) and (e)(g) appear that our method yields a better class-wise separation than the baseline. Furthermore, it is obvious that our approach creates tighter feature clusters, while the features extracted by DeepAll within the same class have multiple sub-clusters. This indicates that KDDRL is able to learn more discriminative features among different object categories regardless of domains. Figures (b)(d) and (f)(h) show how KDDRL more effectively aligns the distribution of source and target domain data, we attribute it to the effectiveness of the two-stage distillation for learning domain-invariant representation and retaining domain-specific characteristics.



(a) KDDRL (Classes) (b) KDDRL (Domains) (c) DeepAll (Classes) (d) DeepAll (Domains)

(e) KDDRL (Classes) (f) KDDRL (Domains) (g) DeepAll (Classes) (h) DeepAll (Domains)

Fig. 6. The t-SNE visualization of feature representations extracted by the feature extractor of our KDDRL (a),(b),(e),(f) and DeepAll model (c),(d),(g),(h) on PACS and Digit-DG dataset. In (a),(c),(e), and (g), the different colors indicate different classes; correspondingly in (b),(d),(f), and (h), the different colors indicate different domains.

VI. Conclusion

This paper identifies the insufficiency of existing DG methods and presents a new view of DG. The main idea is to collaborate with all domains and using their complementary information to learn domain invariant representation. We then propose a framework KDDRL by performing a two-stage knowledge distillation with multiple student models to learn invariant representations that can generalize well on unseen domains. Comprehensive experiments demonstrate the effectiveness and superiority of KDDRL. Considering the mainstream related work is generally based on traditional DG methods, we hope our work can shed some lights into the community.

VII. Acknowledgements

Particular thanks to Rahul Kumar JAIN[1] (Ritsumeikan University, Kyoto), starting assistant professor who has provided language help and English proofing.

References

[1] M. Chen, S. Zhao, H. Liu, and D. Cai, "Adversarial-learned loss for domain adaptation," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, 2020, pp. 3521–3528.

[2] Y. Yang and S. Soatto, "Fda: Fourier domain adaptation for semantic segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 4085–4095.

[3] M. Klingner, J.-A. Termöhlen, J. Ritterbach, and T. Fingscheidt, "Unsupervised batchnorm adaptation (ubna): A domain adaptation method for semantic segmentation without using source domain representations," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2022, pp. 210–220.

[4] T. Wang, X. Zhang, L. Yuan, and J. Feng, "Few-shot adaptive faster r-cnn," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 7173–7182.

[1]rahulkumarjain16@gmail.com

[5] X. Yue, Z. Zheng, S. Zhang, Y. Gao, T. Darrell, K. Keutzer, and A. S. Vincentelli, "Prototypical cross-domain self-supervised learning for few-shot unsupervised domain adaptation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 13 834–13 844.

[6] F. M. Carlucci, A. D'Innocente, S. Bucci, B. Caputo, and T. Tommasi, "Domain generalization by solving jigsaw puzzles," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 2229–2238.

[7] K. Zhou, Y. Yang, T. Hospedales, and T. Xiang, "Deep domain-adversarial image generation for domain generalisation," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, 2020, pp. 13 025–13 032.

[8] X. Jin, C. Lan, W. Zeng, and Z. Chen, "Style normalization and restitution for domain generalization and adaptation," *IEEE Transactions on Multimedia*, 2021.

[9] P. Li, D. Li, W. Li, S. Gong, Y. Fu, and T. M. Hospedales, "A simple feature augmentation for domain generalization," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2021, pp. 8886–8895.

[10] K. Zhou, Y. Yang, Y. Qiao, and T. Xiang, "Domain generalization with mixstyle," in *International Conference on Learning Representations*, 2021.

[11] Q. Dou, D. Coelho de Castro, K. Kamnitsas, and B. Glocker, "Domain generalization via model-agnostic learning of semantic features," *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[12] S. Wang, L. Yu, C. Li, C.-W. Fu, and P.-A. Heng, "Learning from extrinsic and intrinsic supervisions for domain generalization," in *European Conference on Computer Vision*. Springer, 2020, pp. 159–176.

[13] S. Ben-David, J. Blitzer, K. Crammer, A. Kulesza, F. Pereira, and J. W. Vaughan, "A theory of learning from different domains," *Machine learning*, vol. 79, no. 1, pp. 151–175, 2010.

[14] M. Long, Y. Cao, J. Wang, and M. Jordan, "Learning transferable features with deep adaptation networks," in *International conference on machine learning*. PMLR, 2015, pp. 97–105.

[15] H. Li, S. J. Pan, S. Wang, and A. C. Kot, "Domain generalization with adversarial feature learning," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 5400–5409.

[16] E. Tzeng, J. Hoffman, K. Saenko, and T. Darrell, "Adversarial discriminative domain adaptation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[17] Y. Ganin, E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, M. Marchand, and V. Lempitsky, "Domain-adversarial training of neural networks," *The journal of machine learning research*, vol. 17, no. 1, pp. 2096–2030, 2016.

[18] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," *Advances in neural information processing systems*, vol. 27, 2014.

[19] W. Hong, Z. Wang, M. Yang, and J. Yuan, "Conditional generative adversarial network for structured domain adaptation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 1335–1344.

[20] S. Motiian, Q. Jones, S. Iranmanesh, and G. Doretto, "Few-shot adversarial domain adaptation," *Advances in neural information processing systems*, vol. 30, 2017.

[21] G. Blanchard, G. Lee, and C. Scott, "Generalizing from several related classification tasks to a new unlabeled sample," *Advances in neural information processing systems*, vol. 24, 2011.

[22] K. Muandet, D. Balduzzi, and B. Schölkopf, "Domain generalization via invariant feature representation," in *International Conference on Machine Learning*. PMLR, 2013, pp. 10–18.

[23] Y. Li, M. Gong, X. Tian, T. Liu, and D. Tao, "Domain generalization via conditional invariant representations," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 32, 2018.

[24] F. Zhou, Z. Jiang, C. Shui, B. Wang, and B. Chaib-draa, "Domain generalization with optimal transport and metric learning," *arXiv preprint arXiv:2007.10573*, 2020.

[25] J. Tobin, R. Fong, A. Ray, J. Schneider, W. Zaremba, and P. Abbeel, "Domain randomization for transferring deep neural networks from simulation to the real world," in *2017 IEEE/RSJ international conference on intelligent robots and systems (IROS)*. IEEE, 2017, pp. 23–30.

[26] K. Zhou, Y. Yang, T. Hospedales, and T. Xiang, "Learning to generate novel domains for domain generalization," in *European conference on computer vision*. Springer, 2020, pp. 561–578.

[27] Y. Wang, L. Qi, Y. Shi, and Y. Gao, "Feature-based style randomization for domain generalization," *arXiv preprint arXiv:2106.03171*, 2021.

[28] J. Huang, D. Guan, A. Xiao, and S. Lu, "Fsdr: Frequency space domain randomization for domain generalization," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 6891–6902.

[29] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," *arXiv preprint arXiv:1312.6114*, 2013.

[30] I. Tolstikhin, O. Bousquet, S. Gelly, and B. Schölkopf, "Wasserstein auto-encoders," in *6th International Conference on Learning Representations (ICLR 2018)*. OpenReview. net, 2018.

[31] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz, "mixup: Beyond empirical risk minimization," *arXiv preprint arXiv:1710.09412*, 2017.

[32] Y. Wang, H. Li, and A. C. Kot, "Heterogeneous domain generalization via domain mixup," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 3622–3626.

[33] Q. Liu, C. Chen, J. Qin, Q. Dou, and P.-A. Heng, "Feddg: Federated domain generalization on medical image segmentation via episodic learning in continuous frequency space," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 1013–1023.

[34] 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*, 2020, pp. 12 556–12 565.

[35] M.-H. Bui, T. Tran, A. Tran, and D. Phung, "Exploiting domain-specific features to enhance domain generalization," *Advances in Neural Information Processing Systems*, vol. 34, pp. 21 189–21 201, 2021.

[36] B. Li, J. Yang, J. Ren, Y. Wang, and Z. Liu, "Sparse fusion mixture-of-experts are domain generalizable learners," *arXiv e-prints*, pp. arXiv–2206, 2022.

[37] Z. Li, K. Ren, X. Jiang, B. Li, H. Zhang, and D. Li, "Domain generalization using pretrained models without fine-tuning," *arXiv preprint arXiv:2203.04600*, 2022.

[38] 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*, 2022, pp. 8046–8056.

[39] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," *Computer Science*, vol. 14, no. 7, pp. 38–39, 2015.

[40] H. Li, "Exploring knowledge distillation of deep neural nets for efficient hardware solutions," *CS2030 Report*, 2018.

[41] T. Fukuda, M. Suzuki, G. Kurata, S. Thomas, J. Cui, and B. Ramabhadran, "Efficient knowledge distillation from an ensemble of teachers." in *Interspeech*, 2017, pp. 3697–3701.

[42] D. Chen, J.-P. Mei, C. Wang, Y. Feng, and C. Chen, "Online knowledge distillation with diverse peers," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, 2020, pp. 3430–3437.

[43] 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)*, Oct 2017.

[44] 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)*, December 2013.

[45] 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)*, July 2017.

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

[47] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[48] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Communications of the ACM*, vol. 60, no. 6, pp. 84–90, 2017.

[49] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 2009, pp. 248–255.

[50] Y. Balaji, S. Sankaranarayanan, and R. Chellappa, "Metareg: Towards domain generalization using meta-regularization," *Advances in neural information processing systems*, vol. 31, 2018.

[51] D. Li, J. Zhang, Y. Yang, C. Liu, Y.-Z. Song, and T. M. Hospedales, "Episodic training for domain generalization," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019.

[52] K. Zhou, Z. Liu, Y. Qiao, T. Xiang, and C. C. Loy, "Domain generalization in vision: A survey," *arXiv preprint arXiv:2103.02503*, 2021.

[53] I. Gulrajani and D. Lopez-Paz, "In search of lost domain generalization," in *International Conference on Learning Representations*, 2020.

[54] S. Motiian, M. Piccirilli, D. A. Adjeroh, and G. Doretto, "Unified deep supervised domain adaptation and generalization," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.

[55] D. Li, Y. Yang, Y.-Z. Song, and T. Hospedales, "Sequential learning for domain generalization," in *European Conference on Computer Vision*. Springer, 2020, pp. 603–619.

[56] 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*, 2018.

[57] V. N. Vapnik, "An overview of statistical learning theory," *IEEE transactions on neural networks*, vol. 10, no. 5, pp. 988–999, 1999.

[58] M. Arjovsky, L. Bottou, I. Gulrajani, and D. Lopez-Paz, "Invariant risk minimization," *arXiv preprint arXiv:1907.02893*, 2019.

[59] S. Sagawa, P. W. Koh, T. B. Hashimoto, and P. Liang, "Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization," *International Conference on Learning Representations (ICLR 2021)*, 2021.

[60] D. Li, Y. Yang, Y.-Z. Song, and T. M. Hospedales, "Learning to generalize: Meta-learning for domain generalization," in *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[61] B. Sun and K. Saenko, "Deep coral: Correlation alignment for deep domain adaptation," in *Computer Vision–ECCV 2016 Workshops: Amsterdam, The Netherlands, October 8-10 and 15-16, 2016, Proceedings, Part III 14*. Springer, 2016, pp. 443–450.

[62] Y. Li, X. Tian, M. Gong, Y. Liu, T. Liu, K. Zhang, and D. Tao, "Deep domain generalization via conditional invariant adversarial networks," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 624–639.

[63] G. Blanchard, A. A. Deshmukh, Ü. Dogan, G. Lee, and C. Scott, "Domain generalization by marginal transfer learning," *The Journal of Machine Learning Research*, vol. 22, no. 1, pp. 46–100, 2021.

[64] 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*, 2021, pp. 8690–8699.

[65] M. M. Zhang, H. Marklund, N. Dhawan, A. Gupta, S. Levine, and C. Finn, "Adaptive risk minimization: A meta-learning approach for tackling group shift," 2021. [Online]. Available: https://openreview.net/forum?id=MA8eT-vUPvZ

[66] D. Krueger, E. Caballero, J.-H. Jacobsen, A. Zhang, J. Binas, D. Zhang, R. Le Priol, and A. Courville, "Out-of-distribution generalization via risk extrapolation (rex)," in *International Conference on Machine Learning*. PMLR, 2021, pp. 5815–5826.

[67] L. Van der Maaten and G. Hinton, "Visualizing data using t-sne." *Journal of machine learning research*, vol. 9, no. 11, 2008.

Ziwei Niu received his B.S. degree from the College of Computer Science and Technology at Anhui Agricultural University in 2021. He is currently pursuing the Ph.D. degree with the College of Computer Science and Technology at Zhejiang University since 2021. His research interests include domain adaptation, domain generalization and medical image analysis.

Junkun Yuan received his B.S. degree from the College of Information Engineering at Zhejiang University of Technology in 2019. He is currently pursuing the Ph.D. degree with the College of Computer Science and Technology at Zhejiang University since 2019. His research interests include domain adaptation and domain generalization.

Xu Ma received his B.S. degree from the College of Computer Science and Technology at Harbin Institute of Technology in 2020. He is currently pursuing Master degree at the College of Computer Science and Technology at Zhejiang University since 2020. His research interests include domain adaptation and domain generalization.

Yingying Xu recieved the Ph.D. degree in Computer Science and Technology from Zhejiang University, Hangzhou, China, in 2020. She held a post-doctoral position at the Research Center for Healthcare Data Science, Zhejiang Lab, Hangzhou, China, from December 2020 to March 2023. She is currently a research associate with Zhejiang Lab, Hangzhou, China. Her reserach interests include medical image processing, medical image segmentation, and medical image retrieval.

Jing Liu (Member, IEEE) received the B.S. degree in computer science and technology and the M.S. degree in electronic and communications engineering from Hainan University, Haikou, China, in 2010 and 2014, respectively, and the Ph.D. degree in information and communication engineering from Hainan University, Haikou, China, in 2020. She is currently a postdoctoral fellow and associate professor in Zhejiang laboratory, China. She has published more than 40 research articles in international journals and conferences. Her main research areas include artificial intelligence and machine vision, with special interests in image processing and medical image watermarking. Dr. Liu is a member of IEEE, Chinese Computer Federation (CCF) and Hainan PHD Association.

Yen-Wei Chen (Member, IEEE) received the B.E. degree from Kobe University, Kobe, Japan, in 1985, and the M.E. and D.E. degrees from Osaka University, Osaka, Japan, in 1987 and 1990, respectively. From 1991 to 1994, he was a Research Fellow with the Institute for Laser Technology, Osaka. From October 1994 to March 2004, he was an Associate Professor and a Professor with the Department of Electrical and Electronic Engineering, University of the Ryukyus, Okinawa, Japan. He is currently a Professor with the College of Information Science and Engineering, Ritsumeikan University, Kyoto, Japan. His research interests include computer vision, image processing, and deep learning. He has published more than 300 research articles in these fields. He is an Associate Editor of the International Journal of Image and Graphics (IJIG) and an Associate Editor of the International Journal of Knowledge-Based and Intelligent Engineering Systems.

Ruofeng Tong received his B.S. degree in mathematics from Fudan University and his Ph.D. degree in applied mathematics in 1996 from Zhejiang University. Currently, he is a professor in the College of Computer Science and Engineering, Zhejiang University, China. His research interests include CAD&CG, medical image reconstruction, and virtual reality.

Lanfen Lin (Member, IEEE) received Ph.D. degrees in Aircraft Manufacture Engineering from Northwestern Polytechnical University in 1995. She held a postdoctoral position with the College of Computer Science and Technology, Zhejiang University, China, from January 1996 to December 1997. She is currently a Full Professor and the Vice Director of the Artificial Intelligence Institute, Zhejiang University. Her research interests include medical image processing, big data analysis, and data mining.