Balancing Invariant and Specific Knowledge for Domain Generalization with Online Knowledge Distillation

Di Zhao[1], **Jingfeng Zhang**[1], **Hongsheng Hu**[2], **Philippe Fournier-Viger**[3], **Gillian Dobbie**[1] and **Yun Sing Koh**[1]
[1]University of Auckland
[2]University of Newcastle
[3]Shenzhen University
dzha866@aucklanduni.ac.nz, {g.dobbie, jingfeng.zhang, y.koh}@auckland.ac.nz,
Hongsheng.Hu@newcastle.edu.au, philfv@szu.edu.cn

Abstract

Recent research has demonstrated the effectiveness of knowledge distillation in Domain Generalization. However, existing approaches often overlook domain-specific knowledge and rely on an offline distillation strategy, limiting the effectiveness of knowledge transfer. To address these limitations, we propose Balanced Online knowLedge Distillation (BOLD). BOLD leverages a multi-domain expert teacher model, with each expert specializing in a specific source domain, enabling the student to distill both domain-invariant and domain-specific knowledge. We incorporate the Pareto optimization principle and uncertainty weighting to balance these two types of knowledge, ensuring simultaneous optimization without compromising either. Additionally, BOLD employs an online knowledge distillation strategy, allowing the teacher and student to learn concurrently. This dynamic interaction enables the teacher to adapt based on student feedback, facilitating more effective knowledge transfer. Extensive experiments demonstrate that BOLD outperforms state-of-the-art methods. Furthermore, we provide theoretical insights that highlight the importance of domain-specific knowledge and the advantages of uncertainty weighting.

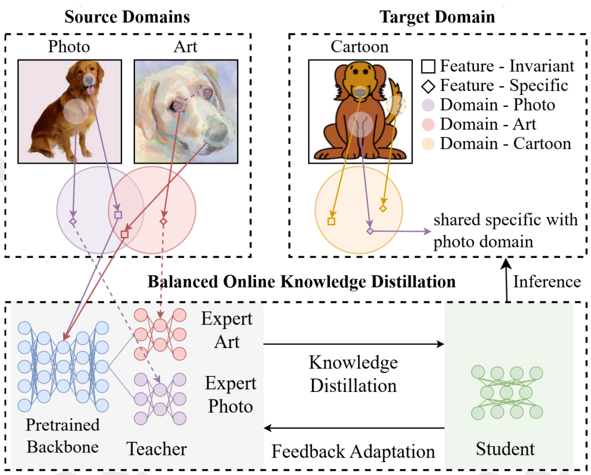


Figure 1: **Illustration of the significance of domain-specific knowledge in domain generalization.** Source domains contain domain-invariant features, common across all domains, and domain-specific features, unique to individual domains, *e.g.* edge features from the Art domain, and color features from the Photo domain. The target domain (Cartoon) shares domain-invariant features with all source domains and domain-specific features with some domains. Therefore, domain-specific features may enhance the model's generalization performance in addition to domain-invariant features.

1 Introduction

The success of deep neural networks largely depends on the assumption that training (source domain) and testing (target domain) data are independently and identically distributed (i.i.d.). However, this assumption is often violated in real-world scenarios due to discrepancies between training and testing data, known as the domain shift problem, leading to significant performance degradation [Wang *et al.*, 2022]. To address this problem, domain adaptation has been explored to transfer knowledge from source to target domains [Pan and Yang, 2009]. Unsupervised domain adaptation, in particular, leverages unlabelled data from target domains, thereby eliminating the need for target domain annotations [Xu *et al.*, 2019]. Despite their effectiveness, unsupervised domain adaptation methods necessitate data collection and model tuning for each target domain, making them impractical in many situations [Yue *et al.*, 2019]. Consequently, domain generalization (DG) has emerged as a prominent alternative. DG aims to learn a universal representation from multiple labelled source domains, enabling robust generalization to unseen domains [Wang *et al.*, 2022]. Existing approaches typically fall into three categories: data augmentation [Zhou *et al.*, 2020], domain-invariant representation [Wang *et al.*, 2022], and specialized training strategies [Zhao *et al.*, 2024].

Knowledge distillation is a training strategy that has demonstrated effectiveness in DG [Wang *et al.*, 2021; Huang *et al.*, 2023]. However, most DG methods based on knowledge distillation focus on extracting domain-invariant knowledge, assuming that domain-specific knowledge impedes generalization [Lee *et al.*, 2022]. This assumption does not always hold, as domain-invariant knowledge derived from a limited number of source domains may not generalize well to unseen domains [Zhang *et al.*, 2023b]. Conversely, domain-specific knowledge from source domains is able to enhance DG performance when target domains share characteristics with particular sources (Figure 1). Nonetheless, simultaneously distilling invariant and specific knowledge presents two key challenges. First, the teacher must ensure that learning specific knowledge does not compromise the invariant knowledge it has already obtained. Existing approaches that rely on ensembles [Zhou *et al.*, 2021] or batch normalization layers [Seo *et al.*, 2020; Zhang *et al.*, 2023b] without an independent extractor for invariant knowledge risk distorting the invariant knowledge while incorporating specific knowledge. Second, it is essential to balance distilled invariant and specific knowledge to prevent the distilled specific knowledge from undermining the distilled invariant knowledge, thereby ensuring the student's generalization performance when the source and target domains do not share characteristics. Additionally, current methods typically employ an offline distillation strategy, where the teacher remains fixed during the distillation process. This approach restricts the teacher from adapting to the student's evolving requirements during training, potentially resulting in suboptimal knowledge transfer.

To address these challenges, we propose Balanced Online knowLedge Distillation (BOLD), which distills both invariant and specific knowledge from a multi-domain expert teacher to a student, dynamically balancing their contributions using uncertainty weighting within an online distillation strategy. BOLD leverages adapter [Gao *et al.*, 2024] techniques to construct a multi-domain expert teacher. Specifically, it integrates multiple adapters into a pretrained backbone, with each adapter specializing in capturing knowledge for a specific domain. This design separates invariant knowledge within the backbone from specific knowledge in the adapters, allowing the student to distill invariant knowledge from the backbone and specific knowledge from the corresponding expert adapter. To tackle the second challenge, BOLD incorporates Pareto optimization principles [Lin *et al.*, 2019] and uncertainty weighting [Kendall *et al.*, 2018] to ensure that both types of knowledge are optimized simultaneously without compromising either. Furthermore, BOLD employs an online distillation strategy, enabling domain experts to train concurrently with the student. During this process, the domain experts minimize the discrepancy between their and student outputs. This online approach enables the domain experts to dynamically adapt based on the student feedback throughout training, facilitating more effective knowledge transfer.

Our **contributions** are as follows: (1) We demonstrate that leveraging an appropriate optimization strategy effectively enhances the model's generalizability by distilling both invariant and specific knowledge. (2) We illustrate that adapting the teacher in response to student feedback using an online distillation strategy improves knowledge transfer and strengthens the student's generalizability. (3) We provide the theoretical insights that underline the importance of domain-specific knowledge and establish the rationale for utilizing uncertainty weighting. Experiments against state-of-the-art baselines validate the effectiveness of the BOLD framework.

2 Related Work

Domain Shift refers to the degradation in performance caused by discrepancies between the source (training) and target (testing) domains [Pan and Yang, 2009]. Domain adaptation has been proposed to address this issue by aligning the marginal [Baktashmotlagh *et al.*, 2013] or conditional [Luo *et al.*, 2020] distributions of the source and target domains or by fine-tuning models trained on source domains to adapt to the target domain [Long *et al.*, 2015]. To reduce the cost associated with annotating target domain data, domain adaptation has been explored in semi-supervised [Saito *et al.*, 2019] and unsupervised [Long *et al.*, 2017] scenarios, utilizing partially labelled or unlabelled target domain data during training. However, these methods still rely on pre-collected target domain data, which presents a practical limitation, as obtaining such data is not always feasible [Yue *et al.*, 2019]. This limitation highlights the need for approaches that are able to generalize to unseen domains without requiring target domain data collection in advance [Wang *et al.*, 2022].

Domain Generalization (DG) was first introduced by [Blanchard *et al.*, 2011] and later formalized by [Muandet *et al.*, 2013]. Existing DG approaches primarily fall into three categories: data augmentation [Zhou *et al.*, 2020], domain-invariant representation learning [Wang *et al.*, 2022], and specialized learning strategies [Zhao *et al.*, 2024; Pang *et al.*, 2025]. Recently, knowledge distillation has attracted attention in the context of DG. [Wang *et al.*, 2021] first proposed a gradient regularization method to regularize the domain-invariant knowledge distilled from the teacher. [Lee *et al.*, 2022] introduced a self-distillation framework where a group of students collectively form a teacher, with each student distilling domain-invariant knowledge from the ensemble teacher. [Huang *et al.*, 2023] leverages the text encoder of a Vision-Language model to distil domain-invariant knowledge. [Zhang *et al.*, 2023b] suggested distilling domain-aware knowledge from a large pre-trained teacher model. Most existing methods focus exclusively on distilling domain-invariant knowledge, overlooking the significance of domain-specific knowledge [Seo *et al.*, 2020; Bui *et al.*, 2021]. Additionally, these methods typically employ an offline distillation strategy, where the teacher remains fixed after initial training. In contrast, our framework distills both invariant and specific knowledge using an online distillation strategy, allowing the teacher to adapt based on feedback from the student.

Knowledge Distillation was initially developed for model compression, with the goal of making the output of a smaller student model similar to that of a larger, existing teacher model [Hinton *et al.*, 2014]. [Luo *et al.*, 2016] demonstrated that training a student model using knowledge from a teacher via knowledge distillation can lead to better performance than direct training with one-hot ground truth labels. Knowledge distillation methods can be categorized into offline and online approaches, depending on whether the teacher is updated concurrently with the student [Gou *et al.*, 2021]. In offline distillation, knowledge is transferred from a pre-trained teacher to a student, typically following a two-stage training process [Zagoruyko and Komodakis, 2017;

Mirzadeh *et al.*, 2020; Li *et al.*, 2020]. Conversely, online distillation allows for the simultaneous updating of both teacher and student and supports an end-to-end trainable knowledge distillation framework [Anil *et al.*, 2018; Zhang *et al.*, 2018; Chen *et al.*, 2020; Wu and Gong, 2021]. While offline distillation has proven effective in DG [Wang *et al.*, 2021; Lee *et al.*, 2022; Huang *et al.*, 2023], the potential of online distillation remains unexplored. To our knowledge, this work is the first to explore how online distillation enhances DG.

3 Balanced Online Knowledge Distillation

This section begins with an overview of domain generalization and knowledge distillation. We then present Balanced Online Knowledge Distillation (BOLD) in three parts: 1) We describe how the teacher acquires specific knowledge and how the student distills invariant and specific knowledge from the teacher; 2) We explain how BOLD incorporates the Pareto optimization principle and uncertainty weighting to balance the distilled invariant and specific knowledge; 3) We discuss how the teacher dynamically adapts to student feedback, enhancing the knowledge transfer. Additionally, we provide theoretical insights that highlight the importance of domain-specific knowledge and justify the use of uncertainty weighting. Figure 2 illustrates the BOLD framework.

3.1 Preliminary

Notation. Let $\mathcal{X}$ denote an input feature space, with dimension d, and $\mathcal{Y}$ a target class label space. A domain, $\mathcal{D}$, is composed of data sampled from a joint distribution $\mathbb{P}(X,Y)$ on $\mathcal{X} \times \mathcal{Y}$, where $\mathcal{D} = (x_i, y_i)_{i=1}^{n} \sim \mathbb{P}(X,Y)$, $x \in \mathcal{X} \subset \mathbb{R}^d$, $y \in \mathcal{Y} \subset \mathbb{R}$ and n is the number of data in the domain. Here, X and Y denote the corresponding random variables [Zhou *et al.*, 2022; Wang *et al.*, 2022].

Domain Generalization. For the task of domain generalization, the input is N source domains (training set), $\mathcal{S} = \{\mathcal{D}^j \mid j = 1, \cdots, N\}$, where $\mathcal{D}^j = \{(x_i^j, y_i^j)\}_{i=1}^{n_j}$ denotes the j-th domain and n_j denotes the number of examples in j-th domain. The joint distributions between each pair of domains are different: $\mathbb{P}(X,Y)^{(j)} \neq \mathbb{P}(X,Y)^{(k)}, j \neq k$. The goal of domain generalization is to learn a robust and generalizable predictive function $f : \mathcal{X} \rightarrow \mathcal{Y}$ from the N source domains to achieve a minimum prediction error on an unseen target domain $\mathcal{T}$, where $\mathcal{T}$ cannot be accessed during training and $\mathbb{P}(X,Y)^{(\mathcal{T})} \neq \mathbb{P}(X,Y)^{(j)}$ for $j \in \{1, \cdots, N\}$.

Knowledge Distillation. Let $T(x)$ and $S(x)$ denote the outputs of the teacher and student models for a given input x. The knowledge distillation loss $\mathcal{L}_{KD}$ is typically defined as the Kullback-Leibler (KL) divergence between the outputs of the teacher and student models: $\mathcal{L}_{KD} = \mathrm{KL}\left(T(x) \parallel S(x)\right)$.

3.2 Distilling Invariant and Specific Knowledge

Teacher Model. We adopt Contrastive Language-Image Pretraining (CLIP) [Radford *et al.*, 2021] as the backbone for the teacher model, which includes both an image encoder and a text encoder. CLIP was selected for its strong generalization ability in associating images with their corresponding textual descriptions. For extracting invariant knowledge, the teacher leverages the pretrained image encoder without additional fine-tuning. To capture specific knowledge, we integrate adapters [Gao *et al.*, 2024], a parameter-efficient tuning method, where each adapter specializes in a specific domain. Figure 2 illustrates that multiple domain-specific adapters are appended to the image encoder, with the number of adapters corresponding to the number of source domains. This design segregates invariant knowledge within the backbone and specific knowledge within the adapters. By ensuring the teacher preserves invariant knowledge while acquiring specific knowledge, the student is able to effectively distill invariant knowledge from the backbone and specific knowledge from the corresponding expert adapter.

We employ cross-entropy loss, $\mathcal{L}_{\text{ce}}$, for each expert adapter E. Unlike approaches that rely solely on maximizing similarity, cross-entropy allows us to maximize the similarity between an image and its ground-truth prompt while minimizing the similarity with unmatched prompts, ensuring comprehensive optimization [Radford *et al.*, 2021]. For each class c, we generate $m \times N$ prompts in the format: "a picture of a $\{D^j\}\{c^k\}$.", where m is the number of classes, D^j represents the j-th domain and c^k represents the k-th class. The text encoder of the teacher model converts these prompts into text embeddings, yielding m text embeddings per domain, corresponding to the m classes. When processing an image from domain D^i, the corresponding expert adapter E^i calculates the cross-entropy loss $\mathcal{L}_{\text{E}}$ for each domain, as defined in Equation 1. Here, T_{img} denotes the image encoder of the teacher model, E^j and T^j represent the expert and text embeddings for the j-th domain, respectively, where $j \in \{1, \cdots, N\}$. The similarity measurement, sim$(\cdot,\cdot)$, evaluates the similarity of image-text pairs, and we adopt cosine similarity following prior works [Radford *et al.*, 2021].

$$\mathcal{L}_{\text{E}}^{j} = \mathcal{L}_{\text{ce}}(\text{sim}(E^j(T_{\text{img}}(x)), T^j), y) \tag{1}$$

After calculating $\mathcal{L}_{\text{E}}$ for each domain, BOLD computes the domain loss $\mathcal{L}_{\text{domain}}^{i}$ for expert adapter i by minimizing the loss for its corresponding domain while maximizing the loss for other domains, as outlined in Equation 2.

$$\mathcal{L}_{\text{domain}}^{i} = \mathcal{L}_{\text{E}}^{i} - \frac{1}{N-1} \sum_{j=1, j\neq i}^{N} \mathcal{L}_{\text{E}}^{j} \tag{2}$$

Student Model. To distill both invariant and specific knowledge from the teacher, we introduce two distillation losses: Invariant Distillation Loss ($\mathcal{L}_{\text{inv}}$) and Student-Specific Distillation Loss ($\mathcal{L}_{\text{sspc}}$), as defined in Equations 3 and 4. The loss $\mathcal{L}_{\text{inv}}$ minimizes the KL divergence between the outputs of the student and the image encoder of the teacher, while $\mathcal{L}_{\text{sspc}}^{i}$ minimizes the KL divergence between the student's outputs and the outputs of the relevant domain expert E^i corresponding to the domain of the input data. Since KL divergence is an asymmetric distance measure, the direction of distribution guidance is crucial. In our approach, the distribution of the teacher's output is used to guide the student's output when distilling knowledge from the teacher to the student.

$$\mathcal{L}_{\text{inv}} = \text{KL}(T_{\text{img}}(x) \parallel S(x)) \tag{3}$$

$$\mathcal{L}_{\text{sspc}}^{i} = \text{KL}(E^i(T_{\text{img}}(x)) \parallel S(x)) \tag{4}$$

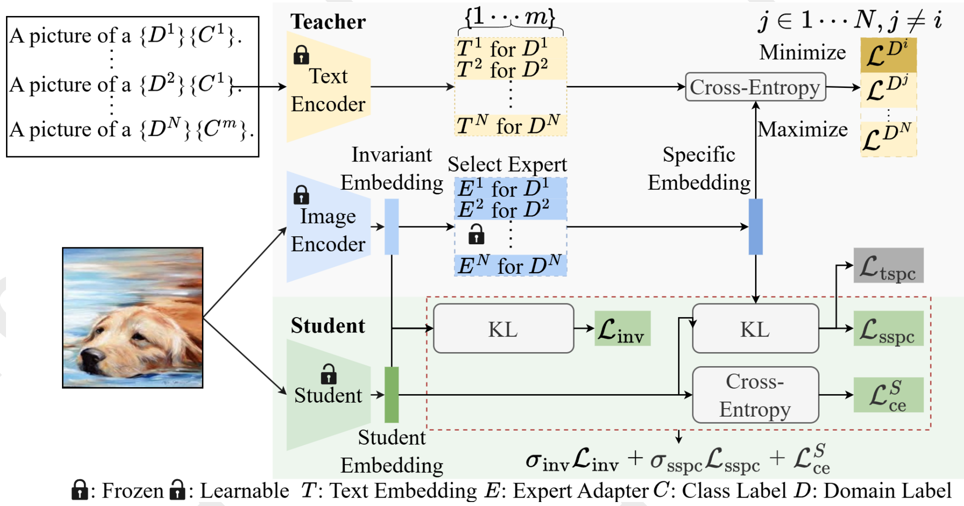


Figure 2: Overview of BOLD. BOLD employs a teacher-student architecture, where the teacher is based on Contrastive Language-Image Pretraining (CLIP) and consists of an image encoder and a text encoder. The image encoder is augmented with multiple domain expert adapters, each designed to retain domain-specific knowledge for a particular source domain. The student distills invariant knowledge by minimizing the KL divergence between its output and the invariant embeddings produced by the image encoder ($\mathcal{L}_{\text{inv}}$) and distills specific knowledge by minimizing the KL divergence between its output and the specific embeddings produced by the adapters ($\mathcal{L}_{\text{sspc}}$). These two losses are balanced using two learnable weights, σ_{inv} and σ_{sspc}. The domain expert adapters capture specific knowledge by minimizing the image-to-text loss for matched domains while maximizing it for unmatched domains. Additionally, they minimize the teacher-specific distillation loss ($\mathcal{L}_{\text{tspc}}$) to incorporate student feedback, further enhancing the effectiveness of knowledge transfer.

Additionally, the student model learns independently by minimizing the cross-entropy of the given input. Equation 5 outlines the complete loss function.

$$\mathcal{L}_S = \mathcal{L}_{\text{inv}} + \mathcal{L}_{\text{sspc}} + \mathcal{L}_{\text{ce}}(S(x), y) \tag{5}$$

The combination of $\mathcal{L}_{\text{inv}}$ and $\mathcal{L}_{\text{sspc}}$ enables the student to capture both shared and unique features across domains, which is essential for enhancing the model's ability to generalize to unseen domains when the target domain shares features with some source domains. Furthermore, minimizing the divergence between student and teacher outputs acts as regularization, mitigating the risk of overfitting to the source domain data. Since the teacher's output represents a full probability distribution over all classes, the student learns to not only predict the correct label but also to approximate this distribution, thereby accounting for uncertainty. Additionally, reducing the divergence between student and teacher outputs enables the student to capture implicit information embedded in the teacher's soft outputs regarding inter-class relationships. This includes subtle correlations and patterns that are not evident through hard labels [Wang *et al.*, 2021].

3.3 Balancing Invariant and Specific Knowledge

Simultaneously distilling invariant and specific knowledge into a single model presents a critical challenge: balancing the contributions of potentially conflicting losses. To address this, we leverage the principles of Pareto optimization [Lin *et al.*, 2019], which suggest that when conflicts arise between multiple optimization objectives, a well-designed weighting strategy can optimize one objective without compromising the others. Specifically, we adopt uncertainty weighting [Kendall *et al.*, 2018], introducing two learnable parameters, σ_{inv} and σ_{sspc}, to dynamically adjust the contributions of the invariant and specific distillation losses.

We begin by recalling the key definitions of Pareto optimality [Liang *et al.*, 2021]. Let $\theta \in \Theta$ represent the model parameters and consider n loss functions. Generally, it is infeasible to find a single θ that minimizes all losses simultaneously due to inherent conflicts among them. However, it is possible to identify a set of solutions known as Pareto optimal solutions, which balance these competing losses effectively.

Definition 1 (Pareto Dominance)**.** Let $\theta^A, \theta^B \in \Theta$ represent two parameter vectors. We say that θ^A Pareto dominates θ^B (denoted as $\theta^A \prec \theta^B$) if $l_i(\theta^A) \leq l_i(\theta^B), \forall i \in \{1, 2, \cdots, n\}$ and $l_i(\theta^A) < l_i(\theta^B), \exists i \in \{1, 2, \cdots, n\}$.

Definition 2 (Pareto Optimality)**.** A parameter vector θ^* is Pareto optimality if there is no other parameter vector θ that dominates it. Formally, θ^* is Pareto optimal if there does not exist θ such that $\theta \prec \theta^*$.

Definition 3 (Pareto Front)**.** The Pareto front is the set of all Pareto optimal parameter vectors in the loss space, where each point corresponds to a unique parameter vector.

We address the simultaneous learning of invariant and specific knowledge within a Pareto optimization framework. Traditional Pareto-based methods, such as weighted sums, the ϵ-constraint technique, Chebyshev distance minimization, and evolutionary algorithms, can approximate the Pareto front but only rely on static or manually tuned weight assignments [Miettinen, 1999], which are difficulty to calibrate. To overcome this limitation, we employ uncertainty weighting to dynamically adjust the contributions of each loss function. This approach eliminates the need for repeated hyperparameter tuning and ensures robustness against gradient fluctuations. The updated loss function for the student is defined in Equation 6. Compared to directly learning the weights using a simple linear combination of losses, uncertainty weighting offers a significant advantage by preventing the weights from converging to zero [Kendall *et al.*, 2018].

$$\mathcal{L}_S = \frac{1}{\sigma_{\text{inv}}^2}\mathcal{L}_{\text{inv}} + \frac{1}{\sigma_{\text{sspc}}^2}\mathcal{L}_{\text{sspc}} + \log(\sigma_{\text{inv}} \cdot \sigma_{\text{sspc}}) + \mathcal{L}_{\text{ce}}(S(x), y) \tag{6}$$

To justify our approach in relation to uncertainty weighting, we demonstrate that minimizing the KL divergence between student features and both invariant and specific features, corresponding to the invariant and specific distillation losses, is equivalent to maximizing a multi-task likelihood for these two objectives (Equation 7). This probabilistic interpretation aligns seamlessly with the uncertainty framework, providing a robust theoretical foundation. Here, z^{inv} denotes the invariant features output by the teacher's backbone, z^{spc} denotes the specific features output by the domain expert adapter, and z^{student} denote the features output by the student.

$$\begin{aligned}
\mathcal{L} &= \min\left(\text{KL}\left(z^{\text{inv}} \parallel z^{\text{student}}\right) + \text{KL}\left(z^{\text{spc}} \parallel z^{\text{student}}\right)\right) \\
&= \min\left(\mathbb{E}_{z^{\text{inv}}}\left[\log\left(\frac{z^{\text{inv}}}{z^{\text{student}}}\right)\right] + \mathbb{E}_{z^{\text{spc}}}\left[\log\left(\frac{z^{\text{spc}}}{z^{\text{student}}}\right)\right]\right) \\
&\propto \min\left(-\mathbb{E}_{z^{\text{inv}}}\left[\log z^{\text{student}}\right] - \mathbb{E}_{z^{\text{spc}}}\left[\log z^{\text{student}}\right]\right) \\
&\Leftrightarrow \max\left(\mathbb{E}_{z^{\text{inv}}, z^{\text{spc}}}\left[\log\left(z^{\text{student}}(z^{\text{inv}}) \cdot z^{\text{student}}(z^{\text{spc}})\right)\right]\right)
\end{aligned} \tag{7}$$

3.4 Online Knowledge Distillation

In contrast to existing DG methods based on knowledge distillation that utilize a fixed teacher model, we implement an online distillation strategy, enabling the teacher to adapt to student feedback. To accomplish this, we introduce the Teacher-Specific Distillation Loss, $\mathcal{L}_{\text{tspc}}$, defined in Equation 8 and integrate it into the teacher's learning objective, as illustrated in Equation 9. Unlike the Student-Specific Distillation Loss, the Teacher-Specific Distillation Loss leverages the student's output to guide the teacher's output. During training, only the domain expert corresponding to the input data's domain is updated, while the teacher's image encoder and domain experts for unrelated domains remain unaffected. Here, $\mathcal{L}_T^i$ denotes the loss for the domain expert associated with the i-th domain while $\mathcal{L}_{\text{domain}}^i$ and $\mathcal{L}_{\text{tspc}}^i$ are the domain and teacher-specific distillation losses for the i-th domain.

$$\mathcal{L}_{\text{tspc}}^i = \text{KL}(S(x) \parallel E^i(T_{\text{img}}(x))) \tag{8}$$

$$\mathcal{L}_T^i = \mathcal{L}_{\text{domain}}^i + \mathcal{L}_{\text{tspc}}^i \tag{9}$$

The online distillation strategy enables the teacher to adapt in real-time based on feedback from the student. Unlike fixed teacher models, which may become outdated as the student evolves, this dynamic adaptation ensures that the transferred knowledge remains relevant and continuously refined, resulting in more effective knowledge transfer. Moreover, the online distillation approach supports an end-to-end training process, eliminating the need for a separate training phase.

4 Experiments

We evaluate our approach using the DomainBed [Gulrajani and Lopez-Paz, 2021] benchmark across five datasets: PACS [Li *et al.*, 2017], OfficeHome [Venkateswara *et al.*, 2017], VLCS [Fang *et al.*, 2013], Terra Incognita [Beery *et al.*, 2018], and DomainNet [Peng *et al.*, 2019]. Additionally, we assess performance on Digits [Zhou *et al.*, 2020] and on NICO++ [Zhang *et al.*, 2023a].

4.1 Experimental Results

Table 1 reports the average accuracy for all baselines across all benchmarks. The best performance is highlighted in bold, while the second-best performance is underlined.

Overall Average Accuracy Across Benchmarks. Table 1 highlights three key findings: (1) BOLD consistently outperforms state-of-the-art approaches, achieving the highest accuracy on the PACS, OfficeHome, VLCS, DomainNet, and NICO++ datasets, demonstrating its effectiveness in improving model generalizability to unseen domains. Notably, BOLD's strong performance on large-scale datasets such as DomainNet and NICO++ underscores its scalability. (2) Knowledge distillation-based methods (NKD, RISE, and BOLD) show weaker performance on the Terra and Digits datasets. This underperformance is attributed to limitations of the teacher model, CLIP, which performs poorly on these datasets. Consequently, students trained to mimic the teacher's outputs inherit these limitations. (3). Despite overall lower performance on Terra and Digits, BOLD surpasses NKD and RISE by a clear margin, achieving an improvement of approximately 5% on the Terra dataset.

Effectiveness Across Different Backbones. Table 2 presents evaluation results of BOLD, NKD, and RISE using different backbones across all benchmarks. The table includes evaluations of knowledge distillation from ResNet-50 and ViT-B/32 to ResNet-18 and from ResNet-50 to ResNet-50. Results for distilling knowledge from ViT-B/32 to ResNet-50 are shown in Table 1. These results demonstrate that BOLD consistently outperforms NKD and RISE across all backbones, highlighting its effectiveness in domain generalization.

4.2 Ablation Study

Effectiveness of Domain-Specific Knowledge and Online Distillation Strategy. Table 3 presents the ablation study results, validating the effectiveness of distilling domain-specific knowledge and employing the online distillation strategy. Here, Invariant represents the setup where the student distills only domain-invariant knowledge. $\text{Spc}_{\text{offline}}$ refers to the setup where the student distills both types of knowledge but offline, meaning the teacher does not adapt to feedback from the student. $\text{Spc}_{\text{online}}$ refers to the setup where the student distills both types of knowledge online, allowing the teacher to adapt dynamically based on student feedback during training.

Based on the results in Table 3, we make three key observations: (1) When invariant knowledge is highly representative, the benefits of distilling specific knowledge are relatively minor. (2) When the target domain shares knowledge with the source domains, distilling specific knowledge results in substantial improvements, as observed in PACS and

	PACS	OfficeHome	VLCS	Terra	DomainNet	Digits	NICO++
ERM [Vapnik, 1999]	83.0 ± .4	68.2 ± .6	77.2 ± .5	41.7 ± .6	40.7 ± .4	79.4 ± .3	79.8 ± .3
CrossGrad [Shankar *et al.*, 2018]	81.7 ± .3	69.8 ± .3	76.1 ± .3	44.7 ± .3	38.5 ± .2	79.5 ± .4	80.6 ± .3
MLDG [Li *et al.*, 2018a]	82.8 ± .3	68.6 ± .4	77.2 ± .4	46.2 ± .5	41.0 ± .4	79.7 ± .4	79.7 ± .4
MMD [Li *et al.*, 2018b]	83.2 ± .7	67.7 ± .6	77.2 ± .4	46.6 ± .6	31.7 ± .5	79.9 ± .4	80.2 ± .4
IRM [Arjovsky *et al.*, 2019]	81.5 ± .3	66.9 ± .4	76.4 ± .4	43.1 ± .6	36.0 ± .4	79.2 ± .4	79.3 ± .4
DDAIG [Zhou *et al.*, 2020]	83.2 ± .3	69.9 ± .3	76.7 ± .3	45.2 ± .2	41.5 ± .3	80.2 ± .3	81.4 ± .2
RSC [Huang *et al.*, 2020]	82.7 ± .5	68.4 ± .6	77.5 ± .4	40.6 ± .3	39.0 ± .4	79.9 ± .4	82.1 ± .4
MixStyle [Zhou *et al.*, 2021]	82.3 ± .3	70.5 ± .3	77.5 ± .3	49.0 ± .3	42.8 ± .3	81.4 ± .3	82.3 ± .3
MTL [Blanchard *et al.*, 2021]	83.6 ± .5	68.1 ± .5	76.6 ± .4	46.2 ± .6	40.5 ± .3	80.3 ± .4	82.0 ± .4
DomainMix [Sun *et al.*, 2022]	82.2 ± .3	69.8 ± .3	76.1 ± .3	48.1 ± .3	42.3 ± .2	80.0 ± .4	82.7 ± .3
EFDMix [Zhang *et al.*, 2022]	84.6 ± .4	71.2 ± .2	78.3 ± .3	**49.9 ± .3**	44.2 ± .3	**82.1 ± .3**	82.6 ± .3
SSPL [Zhao *et al.*, 2024]	84.0 ± .3	71.2 ± .2	77.9 ± .4	48.5 ± .3	42.8 ± .3	81.1 ± .3	82.3 ± .3
NKD [Wang *et al.*, 2021]	84.7 ± .2	70.5 ± .2	80.3 ± .3	32.7 ± .3	44.5 ± .3	49.9 ± .3	81.7 ± .4
RISE [Huang *et al.*, 2023]	86.3 ± .4	71.1 ± .2	80.6 ± .3	34.4 ± .3	45.4 ± .2	51.6 ± .3	82.9 ± .4
BOLD (**Our method**)	**88.7** ± .3	**72.8 ± .3**	**81.7 ± .4**	39.6 ± .3	**48.1 ± .3**	53.7 ± .2	**84.7 ± .4**

Table 1: Comparison with the state-of-the-art methods on PACS, OfficeHome, VLCS, Terra, DomainNet, Digits, and NICO++.

	ResNet50 $\rightarrow$ ResNet18			ViT-B/32 $\rightarrow$ ResNet18			ResNet50 $\rightarrow$ ResNet50		
	NKD	RISE	BOLD	NKD	RISE	BOLD	NKD	RISE	BOLD
PACS	79.7 ± .3	80.9 ± .2	**82.0 ± .2**	81.2 ± .3	82.3 ± .2	**83.9 ± .2**	83.3 ± .4	85.0 ± .3	**85.7 ± .2**
OfficeHome	63.4 ± .2	64.1 ± .2	**66.2 ± .2**	64.0 ± .2	65.0 ± .2	**66.9 ± .2**	71.1 ± .3	71.5 ± .2	**72.6 ± .2**
VLCS	75.7 ± .4	76.2 ± .5	**76.9 ± .3**	76.0 ± .2	76.9 ± .3	**77.7 ± .3**	77.1 ± .3	77.6 ± .2	**78.7 ± .2**
Terra	22.1 ± .4	23.4 ± .3	**29.4 ± .3**	21.4 ± .3	22.4 ± .4	**28.6 ± .3**	37.2 ± .3	39.0 ± .3	**44.3 ± .3**
DomainNet	35.8 ± .2	38.4 ± .2	**39.5 ± .2**	39.4 ± .2	41.9 ± .3	**43.2 ± .2**	42.4 ± .2	45.2 ± .2	**46.9 ± .2**
Digits	47.9 ± .3	49.7 ± .4	**51.6 ± .2**	41.2 ± .3	43.7 ± .4	**46.3 ± .3**	49.9 ± .4	52.0 ± .3	**54.4 ± .2**
NICO++	76.3 ± .3	77.5 ± .3	**78.1 ± .2**	77.5 ± .3	78.8 ± .3	**79.4 ± .2**	80.8 ± .3	81.8 ± .3	**83.3 ± .3**

Table 2: Comparison with various knowledge distillation-based domain generalization approaches using different backbones.

DomainNet. (3) The online distillation strategy effectively enhances knowledge transfer, particularly when the original teacher demonstrates limited capability, as observed in Terra.

	Invariant	$\text{Spc}_{\text{offline}}$	$\text{Spc}_{\text{online}}$
PACS	84.7 ± .2	87.2 ± .2	**88.7 ± .3**
OfficeHome	70.5 ± .2	72.0 ± .2	**72.8 ± .3**
VLCS	80.3 ± .3	80.9 ± .2	**81.7 ± .4**
Terra	32.7 ± .3	33.9 ± .3	**39.6 ± .3**
DomainNet	44.5 ± .3	47.0 ± .4	**48.1 ± .3**
Digits	49.9 ± .3	51.7 ± .3	**53.7 ± .2**
NICO++	81.7 ± .4	83.9 ± .2	**84.7 ± .4**

Table 3: Ablation Study of Domain-Specific Knowledge and Online Distillation Strategy.

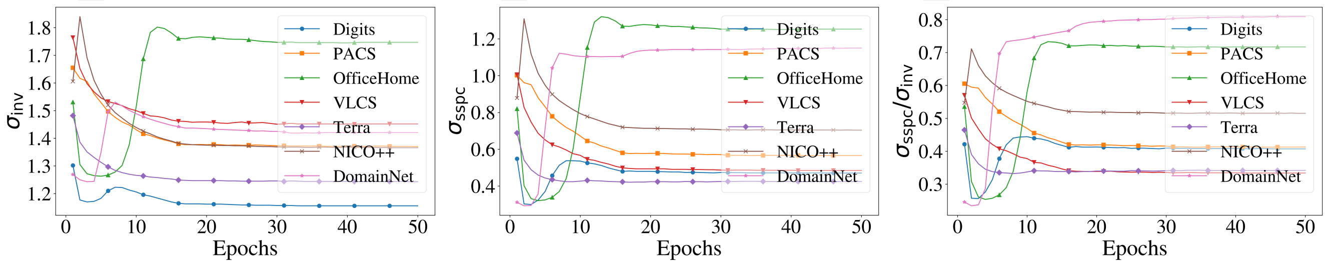
Effectiveness of Uncertainty Weighting. Table 4 presents an ablation study evaluating the uncertainty weighting for balancing invariant and specific distillation losses. We compare it with three alternatives: static weighting (fixed weight of 0.5 for each loss), GradNorm [Chen *et al.*, 2018], and ParetoMTL [Lin *et al.*, 2019]. The results show that while GradNorm and ParetoMTL outperform Uncertainty Weighting in specific domains, Uncertainty Weighting consistently achieves either the best or second-best performance across all domains, leading to the highest overall performance. These findings underscore its robustness and effectiveness.

	Static	GradNorm	ParetoMTL	Uncertainty
PACS	87.2 ± .3	87.4 ± .3	87.9 ± .3	**88.7 ± .3**
OfficeHome	71.3 ± .4	71.8 ± .3	71.9 ± .5	**72.8 ± .3**
VLCS	79.9 ± .3	80.9 ± .5	80.6 ± .4	**81.7 ± .4**
Terra	37.6 ± .3	38.5 ± .3	37.4 ± .4	**39.6 ± .3**
DomainNet	46.3 ± .3	47.1 ± .4	47.5 ± .3	**48.1 ± .3**
Digits	51.8 ± .4	52.6 ± .2	52.0 ± .5	**53.7 ± .2**
NICO++	82.5 ± .3	83.5 ± .3	83.0 ± .4	**84.7 ± .4**

Table 4: Ablation Study of Different Weighting Strategy.

4.3 Further Analysis

Varying of σ_{Inv} and σ_{sspc}. Figure 3 illustrates changes in σ_{inv}, σ_{sspc}, and their ratio $\sigma_{\text{sspc}}/\sigma_{\text{inv}}$. From these results, we make three key observations: (1) The optimal values of σ_{inv} and σ_{sspc} vary across datasets, making manual weight tuning impractical and highlighting the necessity for a dynamic weighting strategy. (2) A higher ratio of $\sigma_{\text{sspc}}/\sigma_{\text{inv}}$ in datasets such as DomainNet and OfficeHome suggests that specific knowledge is more important in these contexts than in datasets like Terra, Digits, and VLCS. This observation aligns with Figure 4, where Terra is the only dataset exhibiting gradient conflict between invariant and specific distillation losses, while Digits and VLCS exhibit low gradient similarity. (3) The values of σ_{inv} consistently exceed those of σ_{sspc}, indicating that the invariant distillation loss contributes more to learning than the specific distillation loss. This finding supports the intuition that specific knowledge complements rather than dominates



(a) σ_{inv} (b) σ_{sspc} (c) $\sigma_{\text{sspc}}/\sigma_{\text{inv}}$

Figure 3: Varying of σ_{inv}, σ_{sspc} and $\sigma_{\text{sspc}}/\sigma_{\text{inv}}$

invariant knowledge during training.

Knowledge Conflict. Figure 4 illustrates how cosine similarity between the gradients of invariant and specific distillation losses evolves over 50 epochs across all benchmarks. As training progresses, the gradient similarity converges. For the Terra dataset, the similarity converges to approximately -0.2, indicating a gradient conflict between invariant and specific distillation losses. In contrast, for the remaining benchmarks, the similarity converges to a positive value. These findings suggest that invariant and specific knowledge are not inherently conflicting; rather, their conflict is dataset-dependent, reinforcing the need for an appropriate balancing strategy.

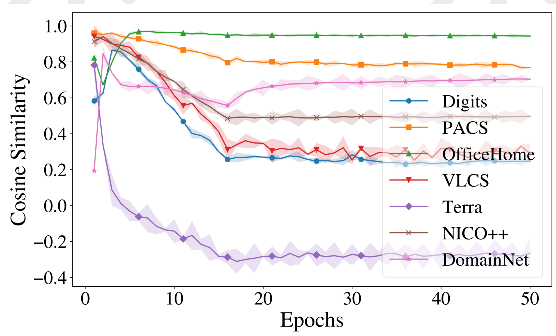
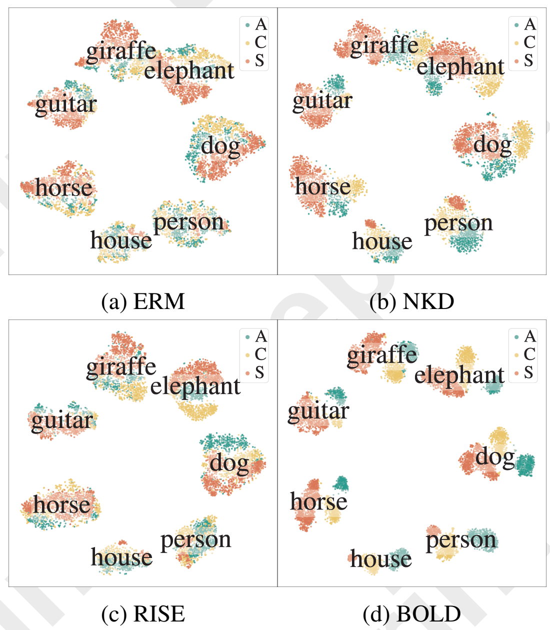


Figure 4: Knowledge Conflict Validation.

T-SNE Visualization. Figure 5 presents the T-SNE visualization for ERM, NKD, RISE, and BOLD on PACS. As shown, distilling knowledge from a large teacher allows NKD, RISE, and BOLD to produce a more separable embedding space than ERM, highlighting the effectiveness of knowledge distillation. Furthermore, by incorporating specific knowledge, BOLD achieves an even more distinct and well-separated embedding space than NKD and RISE, demonstrating the potential of domain-specific knowledge for effective DG.

Imbalanced Dataset & Scalability. Imbalanced dataset distribution poses a practical challenge in providing sufficient training for domain experts. However, our framework employs lightweight adapters as domain experts instead of large-scale neural networks. This design enables effective training even in domains with only a few hundred images. We also compare the parameter count across different backbones and expert adapters relative to the number of source domains.



(a) ERM (b) NKD

(c) RISE (d) BOLD

Figure 5: T-SNE visualization. **A**rt, **C**artoon, and **S**ketch.

5 Conclusion

Our Balanced Online Knowledge Distillation (BOLD) framework leverages both domain-invariant and domain-specific knowledge through an online distillation strategy to improve domain generalization. BOLD employs uncertainty weighting to dynamically balance loss contributions, eliminating the need for manual tuning and enhancing robustness across diverse datasets. The framework is supported by theoretical insights, providing a theoretical foundation for its design. Experiments and ablation studies validate the effectiveness of BOLD. Future work will explore developing more advanced distillation strategies to address limitations in the teacher model's capabilities and extend BOLD to more complex tasks, such as object detection and re-identification.

Acknowledgements

This research was partially supported by New Zealand MBIE Strategic Science Investment Fund (SSIF) Data Science platform - Time-Evolving Data Science / Artificial Intelligence for Advanced Open Environmental Science (UOWX1910).

We also gratefully acknowledge the Centre for Machine Learning for Social Good at the University of Auckland.

References

[Anil *et al.*, 2018] Rohan Anil, Gabriel Pereyra, Alexandre Passos, Robert Ormandi, George E Dahl, and Geoffrey E Hinton. Large scale distributed neural network training through online distillation. In *ICLR*, 2018.

[Arjovsky *et al.*, 2019] Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization. *ArXiv:1907.02893*, 2019.

[Baktashmotlagh *et al.*, 2013] Mahsa Baktashmotlagh, Mehrtash T Harandi, Brian C Lovell, and Mathieu Salzmann. Unsupervised domain adaptation by domain invariant projection. In *ICCV*, pages 769–776, 2013.

[Beery *et al.*, 2018] Sara Beery, Grant Van Horn, and Pietro Perona. Recognition in terra incognita. In *ECCV*, pages 456–473, 2018.

[Blanchard *et al.*, 2011] Gilles Blanchard, Gyemin Lee, and Clayton Scott. Generalizing from several related classification tasks to a new unlabeled sample. *NeurIPS*, 24, 2011.

[Blanchard *et al.*, 2021] Gilles Blanchard, Aniket Anand Deshmukh, Urun Dogan, Gyemin Lee, and Clayton Scott. Domain generalization by marginal transfer learning. *PMLR*, 22(2):1–55, 2021.

[Bui *et al.*, 2021] Manh-Ha Bui, Toan Tran, Anh Tran, and Dinh Phung. Exploiting domain-specific features to enhance domain generalization. *NeurIPS*, 34:21189–21201, 2021.

[Chen *et al.*, 2018] Zhao Chen, Vijay Badrinarayanan, Chen-Yu Lee, and Andrew Rabinovich. Gradnorm: Gradient normalization for adaptive loss balancing in deep multitask networks. In *ICML*, pages 794–803. PMLR, 2018.

[Chen *et al.*, 2020] Defang Chen, Jian-Ping Mei, Can Wang, Yan Feng, and Chun Chen. Online knowledge distillation with diverse peers. In *AAAI*, volume 34, pages 3430–3437, 2020.

[Fang *et al.*, 2013] Chen Fang, Ye Xu, and Daniel N Rockmore. Unbiased metric learning: On the utilization of multiple datasets and web images for softening bias. In *ICCV*, pages 1657–1664, 2013.

[Gao *et al.*, 2024] Peng Gao, Shijie Geng, Renrui Zhang, Teli Ma, Rongyao Fang, Yongfeng Zhang, Hongsheng Li, and Yu Qiao. Clip-adapter: Better vision-language models with feature adapters. *IJCV*, 132(2):581–595, 2024.

[Gou *et al.*, 2021] Jianping Gou, Baosheng Yu, Stephen J Maybank, and Dacheng Tao. Knowledge distillation: A survey. *IJCV*, 129(6):1789–1819, 2021.

[Gulrajani and Lopez-Paz, 2021] Ishaan Gulrajani and David Lopez-Paz. In search of lost domain generalization. In *ICLR*, 2021.

[Hinton *et al.*, 2014] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *NeurIPS*, 2014.

[Huang *et al.*, 2020] Zeyi Huang, Haohan Wang, Eric P Xing, and Dong Huang. Self-challenging improves cross-domain generalization. In *ECCV*, pages 124–140. Springer, 2020.

[Huang *et al.*, 2023] Zeyi Huang, Andy Zhou, Zijian Ling, Mu Cai, Haohan Wang, and Yong Jae Lee. A sentence speaks a thousand images: Domain generalization through distilling clip with language guidance. In *ICCV*, pages 11685–11695, 2023.

[Kendall *et al.*, 2018] Alex Kendall, Yarin Gal, and Roberto Cipolla. Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. In *CVPR*, pages 7482–7491, 2018.

[Lee *et al.*, 2022] Kyungmoon Lee, Sungyeon Kim, and Suha Kwak. Cross-domain ensemble distillation for domain generalization. In *ECCV*, pages 1–20. Springer, 2022.

[Li *et al.*, 2017] Da Li, Yongxin Yang, Yi-Zhe Song, and Timothy M Hospedales. Deeper, broader and artier domain generalization. In *ICCV*, pages 5542–5550, 2017.

[Li *et al.*, 2018a] Da Li, Yongxin Yang, Yi-Zhe Song, and Timothy Hospedales. Learning to generalize: Meta-learning for domain generalization. In *AAAI*, volume 32, 2018.

[Li *et al.*, 2018b] Haoliang Li, Sinno Jialin Pan, Shiqi Wang, and Alex C Kot. Domain generalization with adversarial feature learning. In *CVPR*, pages 5400–5409, 2018.

[Li *et al.*, 2020] Tianhong Li, Jianguo Li, Zhuang Liu, and Changshui Zhang. Few sample knowledge distillation for efficient network compression. In *CVPR*, pages 14639–14647, 2020.

[Liang *et al.*, 2021] Jian Liang, Kaixiong Gong, Shuang Li, Chi Harold Liu, Han Li, Di Liu, Guoren Wang, et al. Pareto domain adaptation. *NeurIPS*, 34:12917–12929, 2021.

[Lin *et al.*, 2019] Xi Lin, Hui-Ling Zhen, Zhenhua Li, Qing-Fu Zhang, and Sam Kwong. Pareto multi-task learning. *NeurIPS*, 32, 2019.

[Long *et al.*, 2015] Mingsheng Long, Yue Cao, Jianmin Wang, and Michael Jordan. Learning transferable features with deep adaptation networks. In *ICML*, pages 97–105, 2015.

[Long *et al.*, 2017] Mingsheng Long, Han Zhu, Jianmin Wang, and Michael I Jordan. Deep transfer learning with joint adaptation networks. In *ICML*, pages 2208–2217. PMLR, 2017.

[Luo *et al.*, 2016] Ping Luo, Zhenyao Zhu, Ziwei Liu, Xiaogang Wang, and Xiaoou Tang. Face model compression by distilling knowledge from neurons. In *AAAI*, volume 30, 2016.

[Luo *et al.*, 2020] Yadan Luo, Zijian Wang, Zi Huang, and Mahsa Baktashmotlagh. Progressive graph learning for open-set domain adaptation. In *ICML*, pages 6468–6478. PMLR, 2020.

[Miettinen, 1999] Kaisa Miettinen. *Nonlinear multiobjective optimization*, volume 12. Springer Science & Business Media, 1999.

[Mirzadeh *et al.*, 2020] Seyed Iman Mirzadeh, Mehrdad Farajtabar, Ang Li, Nir Levine, Akihiro Matsukawa, and Hassan Ghasemzadeh. Improved knowledge distillation via teacher assistant. In *AAAI*, volume 34, pages 5191–5198, 2020.

[Muandet *et al.*, 2013] Krikamol Muandet, David Balduzzi, and Bernhard Schölkopf. Domain generalization via invariant feature representation. In *ICML*, pages 10–18. PMLR, 2013.

[Pan and Yang, 2009] Sinno Jialin Pan and Qiang Yang. A survey on transfer learning. *TKDE*, pages 1345–1359, 2009.

[Pang *et al.*, 2025] Bo Pang, Tingrui Qiao, Caroline Walker, Chris Cunningham, and Yun Sing Koh. Libra: Measuring bias of large language model from a local context. In *ECIR*, pages 1–16. Springer, 2025.

[Peng *et al.*, 2019] Xingchao Peng, Qinxun Bai, Xide Xia, Zijun Huang, Kate Saenko, and Bo Wang. Moment matching for multi-source domain adaptation. In *ICCV*, pages 1406–1415, 2019.

[Radford *et al.*, 2021] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, pages 8748–8763. PMLR, 2021.

[Saito *et al.*, 2019] Kuniaki Saito, Donghyun Kim, Stan Sclaroff, Trevor Darrell, and Kate Saenko. Semi-supervised domain adaptation via minimax entropy. In *ICCV*, pages 8050–8058, 2019.

[Seo *et al.*, 2020] Seonguk Seo, Yumin Suh, Dongwan Kim, Geeho Kim, Jongwoo Han, and Bohyung Han. Learning to optimize domain specific normalization for domain generalization. In *ECCV*, pages 68–83. Springer, 2020.

[Shankar *et al.*, 2018] Shiv Shankar, Vihari Piratla, Soumen Chakrabarti, Siddhartha Chaudhuri, Preethi Jyothi, and Sunita Sarawagi. Generalizing across domains via cross-gradient training. *ICLR*, 2018.

[Sun *et al.*, 2022] Zhishu Sun, Zhifeng Shen, Luojun Lin, Yuanlong Yu, Zhifeng Yang, Shicai Yang, and Weijie Chen. Dynamic domain generalization. *IJCAI*, 2022.

[Vapnik, 1999] Vladimir N Vapnik. An overview of statistical learning theory. *TNN*, 10(5):988–999, 1999.

[Venkateswara *et al.*, 2017] Hemanth Venkateswara, Jose Eusebio, Shayok Chakraborty, and Sethuraman Panchanathan. Deep hashing network for unsupervised domain adaptation. In *CVPR*, pages 5018–5027, 2017.

[Wang *et al.*, 2021] Yufei Wang, Haoliang Li, Lap-pui Chau, and Alex C Kot. Embracing the dark knowledge: Domain generalization using regularized knowledge distillation. In *ACM MM*, pages 2595–2604, 2021.

[Wang *et al.*, 2022] Jindong Wang, Cuiling Lan, Chang Liu, Yidong Ouyang, Tao Qin, Wang Lu, Yiqiang Chen, Wenjun Zeng, and Philip Yu. Generalizing to unseen domains: A survey on domain generalization. *TKDE*, 2022.

[Wu and Gong, 2021] Guile Wu and Shaogang Gong. Peer collaborative learning for online knowledge distillation. In *AAAI*, volume 35, pages 10302–10310, 2021.

[Xu *et al.*, 2019] Ruijia Xu, Guanbin Li, Jihan Yang, and Liang Lin. Larger norm more transferable: An adaptive feature norm approach for unsupervised domain adaptation. In *ICCV*, pages 1426–1435, 2019.

[Yue *et al.*, 2019] Xiangyu Yue, Yang Zhang, Sicheng Zhao, Alberto Sangiovanni-Vincentelli, Kurt Keutzer, and Boqing Gong. Domain randomization and pyramid consistency: Simulation-to-real generalization without accessing target domain data. In *ICCV*, pages 2100–2110, 2019.

[Zagoruyko and Komodakis, 2017] Sergey Zagoruyko and Nikos Komodakis. Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer. In *ICLR*, 2017.

[Zhang *et al.*, 2018] Ying Zhang, Tao Xiang, Timothy M Hospedales, and Huchuan Lu. Deep mutual learning. In *CVPR*, pages 4320–4328, 2018.

[Zhang *et al.*, 2022] Yabin Zhang, Minghan Li, Ruihuang Li, Kui Jia, and Lei Zhang. Exact feature distribution matching for arbitrary style transfer and domain generalization. In *CVPR*, pages 8035–8045, 2022.

[Zhang *et al.*, 2023a] Xingxuan Zhang, Yue He, Renzhe Xu, Han Yu, Zheyan Shen, and Peng Cui. Nico++: Towards better benchmarking for domain generalization. In *CVPR*, pages 16036–16047, 2023.

[Zhang *et al.*, 2023b] Zhongqiang Zhang, Ge Liu, Fuhan Cai, Duo Liu, and Xiangzhong Fang. Boosting domain generalization by domain-aware knowledge distillation. *KBS*, 280:111021, 2023.

[Zhao *et al.*, 2024] Di Zhao, Yun Sing Koh, Gillian Dobbie, Hongsheng Hu, and Philippe Fournier-Viger. Symmetric self-paced learning for domain generalization. In *AAAI*, volume 38, pages 16961–16969, 2024.

[Zhou *et al.*, 2020] Kaiyang Zhou, Yongxin Yang, Timothy Hospedales, and Tao Xiang. Deep domain-adversarial image generation for domain generalisation. In *AAAI*, pages 13025–13032, 2020.

[Zhou *et al.*, 2021] Kaiyang Zhou, Yongxin Yang, Yu Qiao, and Tao Xiang. Domain generalization with mixstyle. *ICLR*, 2021.

[Zhou *et al.*, 2022] Kaiyang Zhou, Ziwei Liu, Yu Qiao, Tao Xiang, and Chen Change Loy. Domain generalization: A survey. *TPAMI*, 2022.