EPVT: Environment-Aware Prompt Vision Transformer for Domain Generalization in Skin Lesion Recognition

Siyuan Yan[1,2,3,6], Chi Liu[2,3,6], Zhen Yu[2,3,6], Lie Ju[1,2,3,6], Dwarikanath Mahapatra[5,6], Victoria Mar[3,6], Monika Janda[4,6], Peter Soyer[4,6], and Zongyuan Ge[2,6](✉)

[1] Faculty of Engineering, Monash University, Melbourne, Australia
[2] AIM for Health Lab, Monash University, Victoria, Australia
`zongyuan.ge@monash.edu`
[3] Monash Medical AI, Monash University, Victoria, Australia
[4] Victorian Melanoma Service, Alfred Health, Victoria, Australia
[5] The University of Queensland Diamantina Institute, Dermatology Research Centre, The University of Queensland, Brisbane, Australia
[6] Inception Institute of AI, Abu Dhabi, UAE

Abstract. Skin lesion recognition using deep learning has made remarkable progress, and there is an increasing need for deploying these systems in real-world scenarios. However, recent research has revealed that deep neural networks for skin lesion recognition may overly depend on disease-irrelevant image artifacts (*i.e.* dark corners, dense hairs), leading to poor generalization in unseen environments. To address this issue, we propose a novel domain generalization method called EPVT, which involves embedding prompts into the vision transformer to collaboratively learn knowledge from diverse domains. Concretely, EPVT leverages a set of domain prompts, each of which plays as a domain expert, to capture domain-specific knowledge; and a shared prompt for general knowledge over the entire dataset. To facilitate knowledge sharing and the interaction of different prompts, we introduce a domain prompt generator that enables low-rank multiplicative updates between domain prompts and the shared prompt. A domain mixup strategy is additionally devised to reduce the co-occurring artifacts in each domain, which allows for more flexible decision margins and mitigates the issue of incorrectly assigned domain labels. Experiments on four out-of-distribution datasets and six different biased ISIC datasets demonstrate the superior generalization ability of EPVT in skin lesion recognition across various environments. Code is available at https://github.com/SiyuanYan1/EPVT.

Keywords: Skin lesions · Prompt · Domain generalization · Debiasing

1 Introduction

Skin cancer is a serious and widespread form of cancer that requires early detection for successful treatment. Computer-aided diagnosis systems (CAD) using

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/978-3-031-43990-2_24.

H. Greenspan et al. (Eds.): MICCAI 2023, LNCS 14226, pp. 249–259, 2023.
https://doi.org/10.1007/978-3-031-43990-2_24

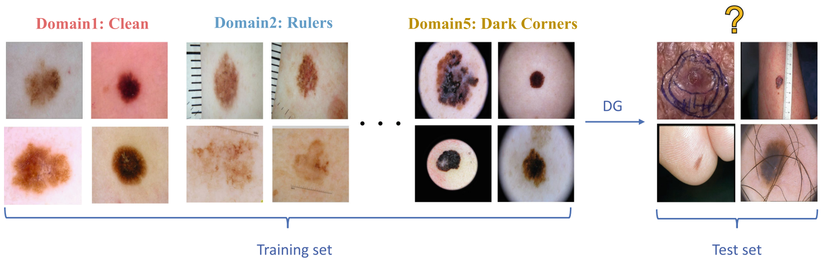

Fig. 1. The training data is split into five domains: clean, rulers, hairs, air pockets, and dark corners. Domain generalization aims to train the model to learn from these domains to generalize well in unseen domains.

deep learning models have shown promise in accurate and efficient skin lesion diagnosis. However, recent research has revealed that the success of these models may be a result of overly relying on "spurious cues" in dermoscopic images, such as rulers, gel bubbles, dark corners, and hairs [3–5,29], which leads to unreliable diagnoses. When a deep learning model overfits specific artifacts instead of learning the correct dermoscopic patterns, it may fail to identify skin lesions in real-world environments where the artifacts are absent or inconsistent.

To alleviate the artifact bias and enhance the model's generalization ability, we rethink the problem from the domain generalization (DG) perspective, where a model trained within multiple different but related domains are expected to perform well in unseen test domains. As illustrated in Fig. 1, we define the domain labels based on the types of artifacts present in the training images, which can provide environment-aware prior knowledge reflecting a range of noisy contexts. By doing this, we can develop a DG algorithm to learn the generalized and robust features from diverse domains.

Previous DG algorithms learning domain-invariant features from source domains have succeeded in natural image tasks [2,17,19], but cannot directly apply to medical images, in particular skin images, due to the vast cross-domain diversity of skin lesions in terms of shapes, colors, textures, etc. As each domain contains ad hoc intrinsic knowledge, learning domain-invariant features is highly challenging. One promising way is, as suggested in some recent works [7,24,32], exploiting multiple learnable domain experts (e.g., batch norm statistic, auxiliary classifiers, etc.) to capture domain-specific knowledge from different source domains individually. Still, two significant challenges remain. First, previous work only exploits some weak experts, like the batch norm, to capture knowledge, which naturally hampers the capability of capturing essential domain-specific knowledge. Second, previous methods such as [30] focused on learning domain knowledge independently while overlooking the rich cross-domain information that all domain experts can contribute collectively for the target domain prediction.

To overcome the above problems, we propose an environment-aware prompt vision transformer (EPVT) for domain generalization of skin lesion recognition.

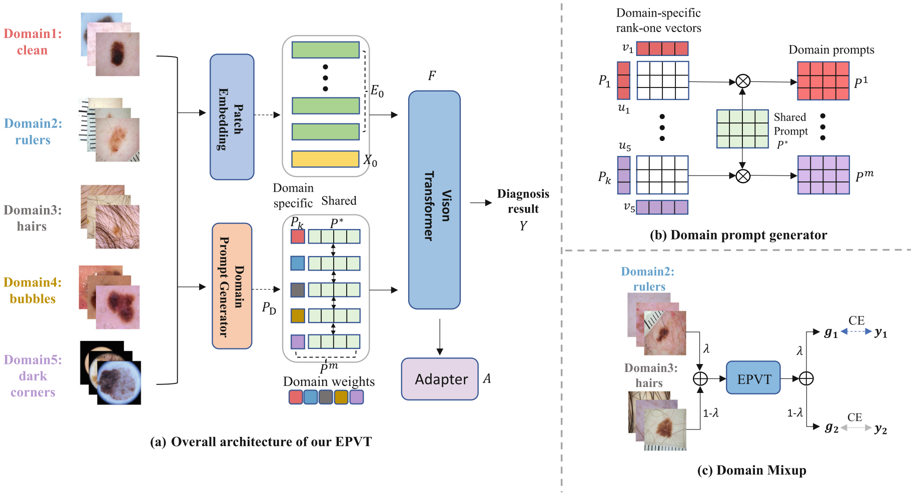

Fig. 2. The overview of our environment-aware prompt vision transformer (EPVT).

On the one hand, inspired by the emerging prompt learning techniques that embed prompts into a model for adaptation to diverse downstream tasks [12,26,31], we construct different prompt vectors to strengthen the learning of domain-specific knowledge for adaptation to diverse domains. Then, the self-attention mechanism of the vision transformer (ViT) [8] is adopted to fully model the relationship between image tokens and prompt vectors. On the other hand, to encourage cross-domain information sharing while preserving the domain-specific knowledge of each domain prompt, we propose a domain prompt generator based on low-rank weights updating. The prompt generator enables multiple domain prompts to work collaboratively and benefit from each other for generalization to unknown domains. Additionally, we devise a domain mixup strategy to resolve the problem of co-occurring artifacts in dermoscopic images and mitigate the resulting noisy domain label assignments.

Our contributions can be summarized as: (1) We resolve an artifacts-derived biasing problem in skin cancer diagnosis using a novel environment-aware prompt learning-based DG algorithm, EPVT; (2) EPVT takes advantage of a ViT-based domain-aware prompt learning and a novel domain prompt generator to improve domain-specific and cross-domain knowledge learning simultaneously; (3) A domain mixup strategy is devised to reduce the co-artifacts specific to dermoscopic images; (4) Extensive experiments on four out-of-distribution skin datasets and six biased ISIC datasets demonstrate the outperforming generalization ability and robustness of EPVT under heterogeneous distribution shifts.

2 Method

In domain generalization (DG), the training dataset D_{train} consists of M source domains, denoted as $D_{train} = \{D_k | k = 1, ..., M\}$. Here, each source domain D_k is represented by n labeled instances $\{(x_j^k, y_j^k)\}_{j=1}^n$. The goal of DG is to learn a model $G : X \rightarrow Y$ from the M source domains so that it can generalize well in unseen target domains D_{test}. The overall architecture of our proposed model, EPVT, is shown in Fig. 2a. We will illustrate its details in the following sections.

2.1 Domain-Specific Prompt Learning with Vision Transformer

To enable the pre-trained vision transformer (ViT) to capture knowledge from different domains, as shown in Fig. 2a, we define a set of M learnable domain prompts produced by a domain prompt generator (introduced in Sect. 2.2), denoted as $P_D = \{P^m \in \mathbb{R}^d\}_{m=1}^M$, where d is the same size as the feature embedding of the ViT and each prompt P^m corresponds to one domain (*i.e.* dark corners). To incorporate these prompts into the model, we follow the conventional practice of visual prompt tuning [12], which prepends the prompts P_D into the first layer of the transformer. Particularly, for each prompt P^m in P_D, we extract the domain-specific features as:

$$F_m(x) = F([\ X_0, P^m, E_0\ ]) \tag{1}$$

where F is the feature encoder of the ViT, X_0 denotes the class token, E_0 is the image patch embedding, F_m is the feature extracted by ViT with the m-th prompt, and 0 is the index of the first layer. Domain prompts P_D are a set of learnable tokens, with each prompt P^m being fed into the vision transformer along with the image and corresponding class tokens from a specific domain. Through optimizing, each prompt becomes a domain expert only responsible for the images from its own domain. By the self-attention mechanism of ViT, the model can effectively capture domain-specific knowledge from the domain prompt tokens.

2.2 Cross-Domain Knowledge Learning

To facilitate effective knowledge sharing across different domains while maintaining its own parameters of each domain prompt, we propose a domain prompt generator, as depicted in Fig. 2b. Our approach is inspired by model adaptation and multi-task learning techniques used in natural language processing [13,26]. Aghajanyan *et al.* [1] have shown that when adapting a model to a specific task, the updates to weights possess a low intrinsic rank. Similarly, each domain prompt P^m should also have a unique low intrinsic rank when learning knowledge from its own domain. To this end, we decompose each P^m into a Hadamard product between a randomly initialized shared prompt P^* and a rank-one matrix P_k obtained from two randomly initialized learnable vectors u_k and v_k, which is:

$$P^m = P^* \odot P_k \quad \text{where} \quad P_k = u_k \cdot v_k^T \tag{2}$$

where P^m represents the domain-specific prompt, computed by Hadamard product of P^* and P_k. Here, $P^* \in \mathbb{R}^{s \times d}$ is utilized to learn general knowledge, with s and d representing the dimensions of the prompt vector and feature embedding respectively. On the other hand, P_k is computed using domain-specific trainable vectors: $u_k \in \mathbb{R}^s$ and $v_k \in \mathbb{R}^d$. These vectors capture domain-specific information in a low-rank space. The decomposition of domain prompts into rank-one subspaces ensures that the model effectively encodes domain-specific information. By using the Hadamard product, the model can efficiently leverage cross-domain knowledge for target domain prediction.

2.3 Mitigating the Co-artifacts Issue

The artifacts-based domain labels can provide domain information for dermoscopic images. However, a non-trivial issue arises due to the possible co-occurrence of different artifacts from other domains within each domain. To address this issue, we employ a domain mixup strategy [27,28]. Instead of assigning a hard prediction label ("0" or "1") to each image, in each batch, we mix every image using two randomly selected images from two different domains. This allows us to learn a flexible margin relative to both domains. We then apply the cross-entropy loss to the corresponding labels of bot images, as shown in Fig. 2c and can be represented by the following equation:

$$\mathcal{L}_{mixup} = \lambda \mathcal{L}_{CE}(G(x_{mix}), y_i) + (1 - \lambda)\mathcal{L}_{CE}(G(x_{mix}), y_j) \tag{3}$$

where $x_{mix} = \lambda x_i^k + (1-\lambda)x_j^q$; x_i^k and x_j^q are samples from two different domains k and q, and y_i^k and y_j^q are the corresponding labels. This strategy can overcome the challenge of ambiguous domain labels in dermoscopic images and improve the performance of our model.

2.4 Optimization

So far, we have introduced $\mathcal{L}_{mixup}$ in Eq. 3 for optimizing our model. However, since our goal is to generalize the model to unseen environments, we also need to take advantage of each domain prompt. Instead of assigning equal weights to each domain prompt, we employ an adapter [30] that learns the linear correlation between the domain prompts and the target image prediction. To obtain the adapted prompt for inference in the target domain, we define it as a linear combination of the source domain prompts:

$$P_{adapted} = A(F(x)) = \sum_{m=1}^{M} w_m \cdot P^m, \quad \text{s.t.} \quad \sum_{m=1}^{M} w_m = 1 \tag{4}$$

where A represents an adapter containing a two-layer MLP with a softmax layer, and w_m denotes the learned weights.

To train the adapter A, we simulate the inference process for each image in the source domain by treating it as an image from the pseudo-target domain.

Specifically, we first extract features from the ViT: $\hat{F_m}(x) = F([X_0, E_0])$. Then we calculated the adapted prompt $P_{adapted}$ for the pseudo-target environment image x using the adapter A: $P_{adapted} = A(\hat{F_m}(x))$. Next, we extract features from ViT using the adapted prompt: $\hat{F_m}(x) = F([\hat{F_m}(x), P_{adapted}, E_0])$. Finally, the classification head H is applied to predict the label y: $y = H(\hat{F}_m(x))$. Additionally, the inferece process is the same as the simulated inference process and our final prediction will be conditioned on the adapted prompt $P_{adapted}$.

To ensure that the adapter learns the correct linear correlation between the domain prompts and the target image, we use the domain label from source domains to directly supervise the weights w_m. We also use the cross-entropy loss to maintain the model performance with the adapted prompt:

$$\mathcal{L}_{adapted} = \mathcal{L}_{CE}(H(\hat{F}_m(x)), y) + \lambda(\frac{1}{M}\sum_{m=1}^{M}\frac{1}{M}(\mathcal{L}_{CE}(w_m^m, 1) + \sum_{t\neq m}\mathcal{L}_{CE}(w_t^m, 0)) \tag{5}$$

where $\hat{F}_m(x)$ is the obtained feature map conditioned on the adapted prompt $P_{adapted}$, and H is the classification head. The total loss is then defined as $\mathcal{L}_{total} = \mathcal{L}_{mixup} + \mathcal{L}_{adapted}$.

3 Experiments

Experimental Setup: We consider two challenging melanoma-benign classification settings that can effectively evaluate the generalization ability of our model in different environments and closely mimic real-world scenarios. (1) Out-of-distribution evaluation: The task is to evaluate the model on test sets that contain different artifacts or attributes compared to the training set. We train and validate all algorithms on *ISIC2019* [6] dataset, following the split of [3]. We use the artifacts annotations from [3] and divide the training set of *ISIC2019* into five groups: *dark corner*, *hair*, *gel bubble*, *ruler*, and *clean*, with 2351, 4884, 1640, 672, and 2796 images, respectively. We evaluate models on four out-of-distribution (OOD) datasets, including *Derm7pt-Dermoscopic* [14], *Derm7pt-Clinical* [14], *PH2* [18], and *PAD-UFES-20* [21]. It's worth noting that *ISIC2019*, *Derm7pt-Dermoscopic*, and *PH2* are dermoscopic images, while *Derm7pt-Clinical* and *PAD* are clinical images. (2) Trap set debiasing: We train and test our EPVT with its baseline on six trap sets [3] with increasing bias levels, ranging from 0 (randomly split training and testing sets from the ISIC2019 dataset) to 1 (the highest bias level where the correlation between artifacts and class label is in the opposite direction in the dataset splits). More details about these datasets and splits are provided in the complementary material.

Implementation Details: For a fair comparison, we train all models using ViT-Base/16 [8] backbone pre-trained on Imagenet and report the ROC-AUC with five random seeds. Hyperparameter and model selection methods are crucial for domain generalization algorithms. We conduct a grid search over learning rate (from $3e^{-4}$ to $5e^{-6}$), weight decay (from $1e^{-2}$ to $1e^{-5}$), and the length of the

Table 1. The comparison on out-of-distribution datasets

Method	derm7pt_d	derm7pt_c	pad	ph2	Average
ERM	81.24 ± 1.6	71.61 ± 1.9	82.62 ± 1.6	83.06 ± 1.9	79.63 ± 1.5
DRO [23]	82.46 ± 1.7	72.88 ± 1.9	81.52 ± 1.2	84.64 ± 1.8	81.27 ± 1.6
CORAL [25]	81.42 ± 1.9	71.45 ± 1.3	**88.13** ± 1.2	85.2 ± 2.2	81.55 ± 1.5
MMD [17]	82.08 ± 1.7	71.8 ± 1.5	85.89 ± 1.9	87.17 ± 1.4	81.73 ± 1.5
DANN [10]	81.79 ± 1.1	73.12 ± 1.6	84.12 ± 1.6	85.18 ± 1.9	81.87 ± 1.7
IRM [2]	79.07 ± 1.7	71.3 ± 1.8	77.82 ± 3.4	79.37 ± 1.2	76.64 ± 1.7
SagNet [20]	82.28 ± 1.8	73.19 ± 1.6	78.89 ± 4.5	88.79 ± 1.9	81.79 ± 1.8
MLDG [16]	81.06 ± 1.6	71.79 ± 1.6	83.41 ± 1.0	84.22 ± 1.8	79.87 ± 1.2
CAD [22]	82.72 ± 1.5	69.57 ± 1.6	81.36 ± 1.9	88.4 ± 1.5	81.51 ± 1.6
DoPrompt [30]	82.38 ± 1.0	71.61 ± 1.7	83.81 ± 1.4	91.33 ± 1.8	82.06 ± 1.6
SelfReg [15]	81.83 ± 1.9	73.29 ± 1.4	85.27 ± 1.3	85.16 ± 3.3	81.12 ± 1.0
EPVT (Ours)	**83.69** ± 1.4	**73.96** ± 1.6	86.67 ± 1.5	**91.91** ± 1.5	**84.11** ± 1.4

prompt (from 4 to 16, when available) and report the best performance of all models. We employ the training-domain validation set method [11] for model selection. After the grid search, we use the AdamW optimizer with a learning rate of $5e^{-6}$ and a weight decay of $1e^{-2}$. The batch size is 130, and the length of the prompt is 10. We resize the input image to a size of 224×224 and adopt the standard data augmentation like random flip, crop, rotation, and color jitter. An early stopping with the patience of 22 is set and with a total of 60 epochs for OOD evaluation and 100 epochs for trap set debiasing. All experiments are conducted on a single NVIDIA RTX 3090 GPU.

Out-of-Distribution Evaluation: Table 1 presents a comprehensive comparison of our EPVT algorithm with existing domain generalization methods. The results clearly demonstrate the superiority of our approach, with the best performance on three out of four OOD datasets and remarkable improvements over the ERM algorithm, especially achieving 4.1% and 8.9% improvement on the *PAD* and *PH2* datasets, respectively. Although some algorithms may perform similarly to our model on one of the four datasets, none can consistently match the performance of our method across all four datasets. Particularly, our approach showcases the highest average performance, with a 2.05% improvement over the second-best algorithm across all four datasets. These findings highlight the effectiveness of our algorithm in learning robust features and its strong generalization abilities across diverse environments.

Ablation Study: We perform ablation studies to analyze each component of our model, as shown in Table 2. We set our baseline as the Empirical Risk Minimization (ERM) algorithm, and we gradually add P (prompt [12]), A (Adapter), M (Mixup), and G (domain prompt generator) into the model. Firstly, we observe that the baseline model with prompt only improves the average performance by

Table 2. Ablation study on out-of-distribution datasets

Method	derm7pt_d	derm7pt_c	pad	ph2	Average
Baseline	81.24 ± 1.6	71.61 ± 1.9	82.62 ± 1.6	83.06 ± 1.9	79.63 ± 1.5
+P	82.13 ± 1.1	71.41 ± 1.3	82.15 ± 1.6	84.21 ± 1.4	79.73 ± 1.3
+P+A	82.55 ± 1.6	72.86 ± 1.1	81.02 ± 1.5	84.97 ± 1.8	81.10 ± 1.6
+P+A+M	81.43 ± 1.4	73.18 ± 1.5	85.78 ± 1.9	89.28 ± 1.3	82.42 ± 1.7
+P+A+M+G	**83.69** ± 1.4	**73.96** ± 1.6	**86.67** ± 1.5	**91.91** ± 1.5	**84.11** ± 1.4

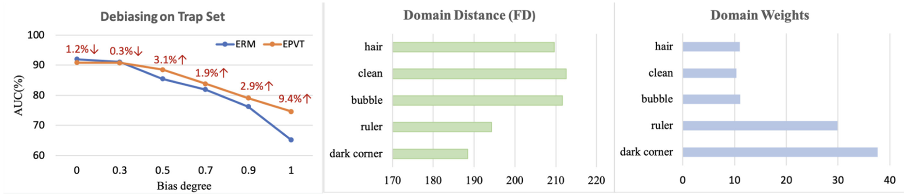

Fig. 3. (a) Deibiasing evaluation (b) domain distance (c) domain weights

0.1%, showing that simply combining prompt does not very helpful for domain generalization. When we combine the adapter, the model's average performance improves by 1.37%, but it performs worse than ERM on *PAD* dataset. Subsequently, we added domain mixup and domain prompt generator to the model, resulting in significant further improvements in the model's average performance by 1.32% and 1.69%, respectively. The consistently better performance than the baseline on all four datasets also highlights the importance of addressing co-artifacts and cross-domain learning for DG in skin lesion recognition.

Trap Set Debiasing: In Fig. 3a, we present the performance of the ERM baseline and our EPVT on six biased ISIC2019 datasets. Each point on the graph represents an algorithm that is trained and tested on a specific bias degree split. The graph shows that the ERM baseline performs better than our EPVT when the bias is low (0 and 0.3). However, this is because ERM relies heavily on spurious correlations between artifacts and class labels, leading to overfitting on the training set. As the bias degree increases, the correlation between artifacts and class labels decreases, and overfitting the train set causes the performance of ERM to drop dramatically on the test set with a significant distribution difference. In contrast, our EPVT exhibits greater robustness to different bias levels. Notably, our EPVT outperforms the ERM baseline by 9.4% on the bias 1 dataset.

Prompt Weights Analysis: To verify whether our model has learned the correct domain prompts for target domain prediction, we analyze and plot the results in Fig. 3b and 3c. Firstly, we extract the features of each domain from our training set and extract the feature from one target dataset, *Derm7pt-Clin*. We then calculate the Frechet distance [9] between each domain and the target dataset using the extracted feature, representing the domain distance between them. The results are recorded in Fig. 3b. Next, we record the learned weights of each domain prompt in Fig. 3c; it shows that our model assigns the highest weight to the "dark corner" group, as the domain distance between "dark corner" and *Derm7pt-Clin* is the closest, as shown in Fig. 3b. This suggests that they share the most similar domain information. Further, the "clean" group is assigned the smallest weight as the domain distance between them is the largest, indicating that their domains are significantly different and contain less useful information for target domain prediction. In summary, we observe a negative correlation between domain distance and the prompt's weights, indicating that our model can learn the correct knowledge from different domains precisely.

4 Conclusion

In this paper, we propose a novel DG algorithm called EPVT for robust skin lesion recognition. Our approach addresses the co-artifacts problem using a domain mixup strategy and cross-domain learning problems using a domain prompt generator. Compared to other competitive domain generalization algorithms, our method achieves outstanding results on three out of four OOD datasets and the second-best on the remaining one. Additionally, we conducted a debiasing experiment that highlights the shortcomings of conventional training using empirical risk minimization, which leads to overfitting in dermoscopic images due to artifacts. In contrast, our EPVT model effectively reduces overfitting and consistently performs better in different biased environments.

References

1. Aghajanyan, A., Zettlemoyer, L., Gupta, S.: Intrinsic dimensionality explains the effectiveness of language model fine-tuning. In: Annual Meeting of the Association for Computational Linguistics (2020)
2. Arjovsky, M., Bottou, L., Gulrajani, I., Lopez-Paz, D.: Invariant risk minimization. arXiv abs/1907.02893 (2019)
3. Bissoto, A., Barata, C., Valle, E., Avila, S.: Artifact-based domain generalization of skin lesion models. In: ECCV Workshops (2022)
4. Bissoto, A., Fornaciali, M., Valle, E., Avila, S.: (de) constructing bias on skin lesion datasets. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pp. 2766–2774 (2019)
5. Bissoto, A., Valle, E., Avila, S.: Debiasing skin lesion datasets and models? Not so fast. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pp. 3192–3201 (2020)
6. Combalia, M., et al.: Validation of artificial intelligence prediction models for skin cancer diagnosis using dermoscopy images: the 2019 international skin imaging collaboration grand challenge. Lancet Digit. Health **4**(5), e330–e339 (2022)
7. Dai, Y., Li, X., Liu, J., Tong, Z., Duan, L.Y.: Generalizable person re-identification with relevance-aware mixture of experts. In: 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 16140–16149 (2021)
8. Dosovitskiy, A., et al.: An image is worth 16x16 words: transformers for image recognition at scale. In: ICLR (2021)

9. Dowson, D., Landau, B.: The fréchet distance between multivariate normal distributions. J. Multivar. Anal. **12**(3), 450–455 (1982)
10. Ganin, Y., et al.: Domain-adversarial training of neural networks. J. Mach. Learn. Res. **17**(1), 2096-2030 (2016)
11. Gulrajani, I., Lopez-Paz, D.: In search of lost domain generalization. In: International Conference on Learning Representations (2021). https://openreview.net/forum?id=lQdXeXDoWtI
12. Jia, M., et al.: Visual prompt tuning. In: Avidan, S., Brostow, G., Cissé, M., Farinella, G.M., Hassner, T. (eds.) ECCV 2022. LNCS, vol. 13693, pp. 709–727. Springer, Cham (2022). https://doi.org/10.1007/978-3-031-19827-4_41
13. Karimi Mahabadi, R., Henderson, J., Ruder, S.: Compacter: efficient low-rank hypercomplex adapter layers. Adv. Neural. Inf. Process. Syst. **34**, 1022–1035 (2021)
14. Kawahara, J., Daneshvar, S., Argenziano, G., Hamarneh, G.: Seven-point checklist and skin lesion classification using multitask multimodal neural nets. IEEE J. Biomed. Health Inform. **23**(2), 538–546 (2018)
15. Kim, D., Yoo, Y., Park, S., Kim, J., Lee, J.: Selfreg: self-supervised contrastive regularization for domain generalization. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 9619–9628 (2021)
16. Li, D., Yang, Y., Song, Y.Z., Hospedales, T.: Learning to generalize: meta-learning for domain generalization. In: AAAI Conference on Artificial Intelligence (2018)
17. Li, H., Pan, S.J., Wang, S., Kot, A.C.: Domain generalization with adversarial feature learning. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 5400–5409 (2018)
18. Mendonça, T., Celebi, M., Mendonca, T., Marques, J.: PH2: a public database for the analysis of dermoscopic images. Dermoscopy Image Anal. (2015)
19. Motiian, S., Piccirilli, M., Adjeroh, D.A., Doretto, G.: Unified deep supervised domain adaptation and generalization. In: 2017 IEEE International Conference on Computer Vision (ICCV), pp. 5716–5726 (2017)
20. Nam, H., Lee, H., Park, J., Yoon, W., Yoo, D.: Reducing domain gap by reducing style bias. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 8690–8699 (2021)
21. Pacheco, A.G., et al.: PAD-UFES-20: a skin lesion dataset composed of patient data and clinical images collected from smartphones. Data Brief **32**, 106221 (2020)
22. Ruan, Y., Dubois, Y., Maddison, C.J.: Optimal representations for covariate shift. In: International Conference on Learning Representations (2022). https://openreview.net/forum?id=Rf58LPCwJj0
23. Sagawa, S., Koh, P.W., Hashimoto, T.B., Liang, P.: Distributionally robust neural networks. In: International Conference on Learning Representations (2020). https://openreview.net/forum?id=ryxGuJrFvS
24. Seo, S., Suh, Y., Kim, D., Kim, G., Han, J., Han, B.: Learning to optimize domain specific normalization for domain generalization. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J.-M. (eds.) ECCV 2020. LNCS, vol. 12367, pp. 68–83. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-58542-6_5
25. Sun, B., Saenko, K.: Deep CORAL: correlation alignment for deep domain adaptation. In: Hua, G., Jégou, H. (eds.) ECCV 2016. LNCS, vol. 9915, pp. 443–450. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-49409-8_35
26. Wang, Z., Panda, R., Karlinsky, L., Feris, R., Sun, H., Kim, Y.: Multitask prompt tuning enables parameter-efficient transfer learning. In: International Conference on Learning Representations (2023). https://openreview.net/forum?id=Nk2pDtuhTq

27. Xu, M., Zhang, J., Ni, B., Li, T., Wang, C., Tian, Q., Zhang, W.: Adversarial domain adaptation with domain mixup. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 34, pp. 6502–6509 (2020)
28. Yan, S., Song, H., Li, N., Zou, L., Ren, L.: Improve unsupervised domain adaptation with mixup training. arXiv preprint arXiv:2001.00677 (2020)
29. Yan, S., et al.: Towards trustable skin cancer diagnosis via rewriting model's decision. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 11568–11577 (2023)
30. Zheng, Z., Yue, X., Wang, K., You, Y.: Prompt vision transformer for domain generalization. arXiv abs/2208.08914 (2022)
31. Zhou, K., Yang, J., Loy, C.C., Liu, Z.: Learning to prompt for vision-language models. Int. J. Comput. Vision **130**, 2337–2348 (2021)
32. Zhou, K., Yang, Y., Qiao, Y., Xiang, T.: Domain adaptive ensemble learning. IEEE Trans. Image Process. **30**, 8008–8018 (2020)