Generalizing to Unseen Domains in Diabetic Retinopathy Classification

Chamuditha Jayanga Galappaththige Gayal Kuruppu Muhammad Haris Khan
Mohamed bin Zayed University of Artificial Intelligence, UAE.
{chamuditha.jayanga,gayal.kuruppu,muhammad.haris}@mbzuai.ac.ae

Abstract

Diabetic retinopathy (DR) is caused by long-standing diabetes and is among the fifth leading cause for visual impairment. The prospects of early diagnosis and treatment could be helpful in curing the disease, however, the detection procedure is rather challenging and mostly tedious. Therefore, automated diabetic retinopathy classification using deep learning techniques has gained interest in the medical imaging community. Akin to several other real-world applications of deep learning, the typical assumption of i.i.d data is also violated in DR classification that relies on deep learning. Therefore, developing DR classification methods robust to unseen distributions is of great value. In this paper, we study the problem of generalizing a model to unseen distributions or domains (a.k.a domain generalization) in DR classification. To this end, we propose a simple and effective domain generalization (DG) approach that achieves self-distillation in vision transformers (ViT) via a novel prediction softening mechanism. This prediction softening is an adaptive convex combination of one-hot labels with the model's own knowledge. We perform extensive experiments on challenging open-source DR classification datasets under both multi-source and more challenging single-source DG settings with three different ViT backbones to establish the efficacy and applicability of our approach against competing methods. For the first time, we report the performance of several state-of-the-art domain generalization (DG) methods on open-source DR classification datasets after conducting thorough experiments. Finally, our method is also capable of delivering improved calibration performance than other methods, showing its suitability for safety-critical applications, including healthcare. We hope that our contributions would instigate more DG research across the medical imaging community. Code is available at github.com/Chumsy0725/SPSD-ViT.

1. Introduction

Diabetic retinopathy (DR) is a global medical problem over the last few decades; it arises from a complication in Diabetes Mellitus. In DR, the development of glucose in blood vessels can block them, thereby causing possible swelling and/or leaking of blood or fluid which could eventually lead to visual impairment [45, 46]. Approximately 33% of 285 million people with diabetes mellitus across the globe have signs of DR [30]. The prevailing practice requires doctors to manually examine the fundus images of the eye to understand the severity of DR. However, this is a time-consuming process and also there is a scarcity of medical professionals compared to the number of patients. Therefore, without much surprise, the development of AI-powered tools capable of accurately detecting DR has gained importance in the recent past. Several studies in this pursuit utilize fundus images, which visually depict the current ophthalmic appearance of a person's retina [54]. The existence of DR symptoms in these fundus images can be used to classify them using several steps such as retinal blood vessel segmentation, lesion segmentation, and DR detection [51]. We can detect DR and its current stage by examining the presence/absence of several lesions. The lesions that are important for diagnosis are microaneurysms (MAs), superficial retinal hemorrhages (SRHs), exudates (Exs) both soft exudates (SEs) and hard exudates (HEs) intraretinal hemorrhages (IHEs), and cotton wool spots (CWSs) [54]. According to recent studies, DR can be classified into five different categories: namely no DR, mild DR, moderate DR, severe DR, and proliferative DR [25].

The typical assumption of i.i.d data which belong to training and testing sets is often violated in many real-world applications of deep learning, e.g., methods for DR classification [3]. Unsupervised domain adaptation is a line of research for handling domain shift [16, 23, 39, 58, 69], but it requires the availability of unlabelled data and an adaptation phase, which typically consists of model re-training. Such requirements are often difficult to fulfill in most medical diagnosis applications. A viable direction is domain generalization (DG), which does not require the availability of target domain data and any adaptation phase and therefore it is more rewarding but also challenging [6, 26, 31, 33, 42, 73]. Unfortunately, very little attention is paid to the develop-

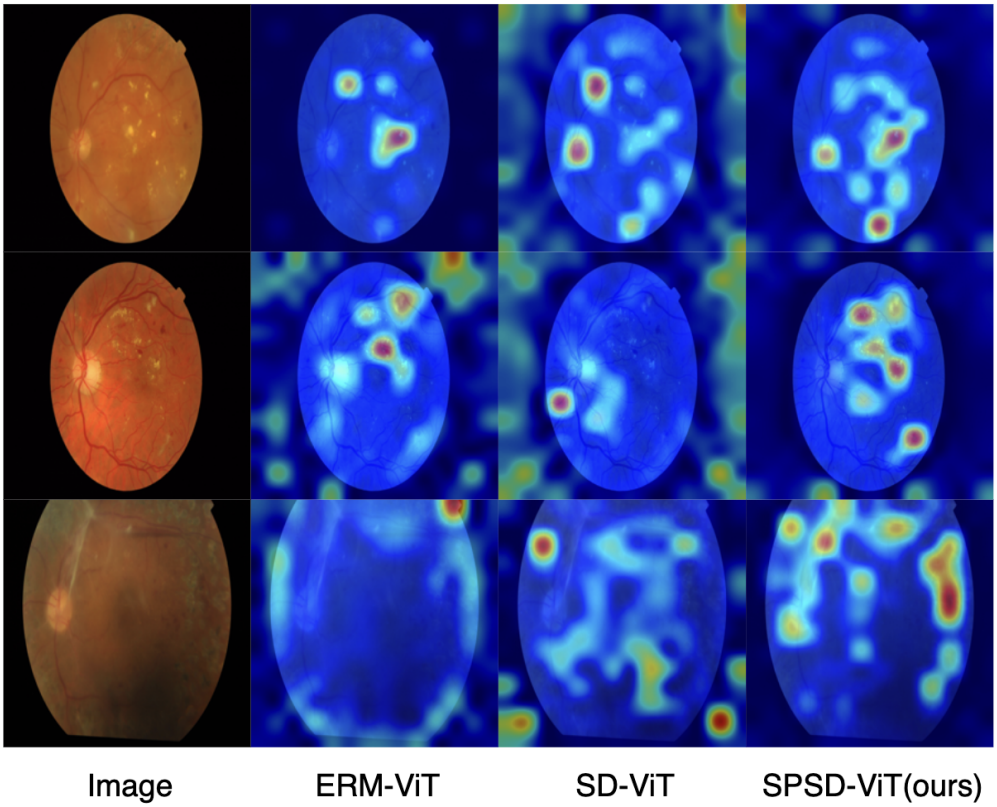

Figure 1. Attention heatmaps obtained from the final ViT block. We can observe that the baselines: ERM-ViT and SD-ViT are suspectible to relying on non-generalizable domain-specific features such as the background. On the other hand, our method (SPSD-ViT) is capable of exploiting cross-domain generalizable features which mostly correspond to haemorrhages, microaneurysms, exudates and cotton wool spots.

ment of DG methods for DR classification [3], which will play a pivotal role in realizing robust DR classification systems. To this end, we explore the problem of generalizing to unseen domains (DG) for the DR classification tasks. We summarise our key contributions as follows:

Contributions: **(1)** Owing to the increasing popularity of vision transformers (ViTs) [13, 59, 61, 67], to our knowledge, we propose a first ViT-based DG approach for DR classification that self-distills the knowledge of the full network to its intermediate blocks via a new prediction softening mechanism. **(2)** We construct softened predictions by formulating an adaptive convex combination of one-hot labels with the model's own knowledge. It strengths intermediate representations and induces model regularization to alleviate the overfitting to source domains, thereby encouraging the learning of more robust (typically domain-invariant) features and reduces reliance on non-robust (typically domain-specific) features (Fig. 1). **(3)** We conduct experiments on challenging DR datasets following Domainbed [17] protocol under both multi-source and single-source DG settings with three different ViT backbones. Results show the efficacy and applicability of our approach against baselines and established methods. **(4)** For the first time, we report the performance of several state-of-the-art DG methods on open-source DR classification datasets.

2. Related Work

Domain Generalization: Domain generalization [6, 26, 31, 33, 41, 64] utilizes data from multiple source domains for training to generalize to a new (unseen) domain. Some DG methods explicitly aimed at reducing domain gap in the feature space [12, 15, 36, 43, 71]. Another class of methods attempted the learning of generalizable model parameters through variants of meta-learning [4, 8, 14, 32, 33, 50]. Through leveraging different auxiliary tasks, some methods were proposed to robustify the model against domain shifts [6, 65]. Furthermore, various DG methods resorted to devising data augmentation techniques for improving cross-domain generalization [24, 26, 55, 73] while some employ test time adaptation methods [9, 21]. Recently, Gulrajani et al. [17] proposed a new benchmark for DG, named "Do-

mainbed", which includes a rigorous evaluation protocol that ensures a fair comparison between different DG algorithms. It showed that even a simple Empirical Risk Minimization (ERM) method can be competitive with many of the current state-of-the-art DG approaches. As such, the Domainbed protocol has quickly gained popularity and is now considered a standard for evaluating DG algorithms. In this study, we have also adopted the Domainbed protocol to report results. By using this protocol, we can provide a reliable and fair comparison of our DG approach with other state-of-the-art methods. Our results demonstrate the effectiveness of our approach, even under the strict evaluation criteria of the Domainbed benchmark.

A few methods have proposed variants of teacher-student formulation [19] to tackle the DG problem. [66] proposed a CNN-based teacher-student distillation scheme along with a gradient filter as an efficient regularization term. Recently [57] developed a self-distillation strategy to improve the DG capabilities of ViTs. It scales the logits of both the full network and a randomly sampled block with a fixed temperature parameter prior to distillation. We adapt this self-distillation for ViTs in DG for DR classification and propose a new prediction softening mechanism, featuring an adaptive convex combination of zero-entropy labels with the model's own knowledge. We empirically show that it is more effective in improving the model's generalizability to unseen domains in DR classification.

DG in medical image analysis: The distribution of data originating from different hospitals or even different sensors could be sparingly different and hence it is important that a model should generalize to a different data distribution than it is trained on. Despite carrying significant importance, DG for medical imaging analysis remains largely unexplored. Among a few DG methods, [35] developed a meta-leaning approach based on episodic training with task augmentation for medical image classification, and [34] leveraged variational encoding to realize a characteristic feature space through linear-dependency regularization. DG in medical imaging has also been explored in the context of Federated Learning (FL). To allow privacy-protected distribution of information among clients, [38] presented episodic learning in Continuous Frequency Space (ELCFS) approach. We note that there is very little work on studying domain generalization for DR classification. Recently, [3] proposed the very first approach for robustifying the model under data from unseen domains in DR classification. It achieves flatness during the training of convolutional neural network (CNN) and also employs domain-level gradient variance regularization. We also propose a new DG approach for DR categorization which transfers the model's (ViTs) full knowledge to its intermediate feature routes by a new prediction softening scheme.

3. Proposed Method

3.1. Preliminaries

DG problem settings: In the typical domain generalization (DG) setting, as outlined in [17], we assume access to data from a set of training (source) domains, denoted as $\mathcal{D} = \{\mathcal{D}\}_{n=1}^{N}$. Each domain $\mathcal{D}_n$ represents a distribution over the input space $\mathcal{X}$, and there are a total of N training domains. From each domain $\mathcal{D}_n$, we sample K training images consisting of pairs of inputs $\mathbf{x}_n^k \in \mathcal{X}$ and labels $y_n^k \in \mathcal{Y}$, where k ranges from 1 to K. Moreover, we assume the existence of a set of target domains, denoted as $\{\mathcal{T}\}_{t=1}^{T}$, where T is the total number of target domains and is typically 1. The core objective in DG is to learn a function $\mathcal{F}_\theta : \mathcal{X} \rightarrow \mathcal{Y}$, parameterized by θ which is capable of predicting accurate labels for input data from an unseen target domain $\mathcal{T}_t$.

ViT-based ERM for DG: We first briefly revisit empirical risk minimization (ERM) in the context of DG and then describe the ViT-based ERM in DG for DR classification task. We assume the availability of a loss function $\mathcal{L}$ that can measure the discrepancy between the predicted label and the desired label. The ERM for DG accumulates data from all training (source) domains and trains a classifier that finds a predictor by minimizing [62]: $\frac{1}{M}\sum_{i=1}^{M}\mathcal{L}(\mathcal{F}_\theta(\mathbf{x}^i, y^i))$. Where $M = N \times K$ denotes the total number of images from all training (source) domains. The work of [17] established that this simple ERM-based DG baseline reveals competitive performance against many preceding state-of-the-art DG methods under a fair evaluation protocol.

Now we assume that the model $\mathcal{F}_\theta$ in ERM-DG (for DR classification) contains J intermediate blocks/layers and a final classifier h, which can be written as: $\mathcal{F} = (f_1 \circ f_2 \circ f_3 \circ \ldots f_J) \circ h$, where f_j denotes an intermediate block/layer. If this model is a ViT (e.g., DeIT-Small [59]), then f_j is a self-attention transformer block. An important characteristic of this network design is that any intermediate transformer block generates features of the same dimensions: $\mathbb{R}^{m \times d}$, where m denotes the number of input features or tokens and each lies in d dimensions.

Self-Distilled (SD) ViT for DG: Owing to the monolithic design of ViT it is possible to create several intermediate classifiers. For instance, the output of each transformer block can be provided to the final classifier h to obtain an intermediate classifier: $\mathcal{F}_j = f_j \circ h$. Whereby each intermediate classifier manifests a feature route through the network.

Through exploiting the ability to seamlessly create intermediate classifiers, [57] developed a technique that randomly samples an intermediate classifier from all the possible ones at each training iteration. The output of the final classifier is then distilled to this randomly sampled intermediate classifier. Specifically, the discrepancy between the

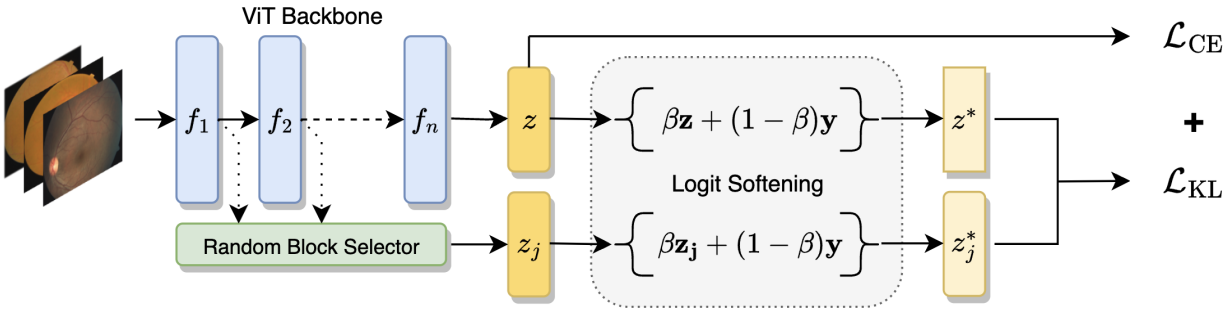

Figure 2. Overall architecture of our proposed self-distilled ViT with the prediction softening mechanism.

final and randomly sampled intermediate classification classifier outputs is computed by comparing the KL divergence between their logit distributions:

$$\mathcal{L}_{\text{KL}}(\mathbf{z}\|\mathbf{z}_j) = \sum_{c=1}^{C} \sigma\left(\mathbf{z}/\tau\right)_c \log \frac{\sigma\left(\mathbf{z}/\tau\right)_c}{\sigma\left(\mathbf{z}_j/\tau\right)_c}, \tag{1}$$

where $\mathbf{z}, \mathbf{z}_j \in \mathbb{R}^C$ are the logit vectors produced by $\mathcal{F}$ and $\mathcal{F}_j$, respectively, and C is number of classes. σ denotes the softmax operation and τ represent temperature used to rescale the logit vector [19]. The model is optimized by jointly minimizing the Eq.(1) and $\mathcal{L}_{\text{CE}}$:

$$\mathcal{L} = \mathcal{L}_{\text{CE}} + \lambda \mathcal{L}_{\text{KL}}, \tag{2}$$

where λ balances the contribution of $\mathcal{L}_{\text{KL}}$ towards the overall loss $\mathcal{L}$.

3.2. Softening Predictions for Self-distillation (SPSD)

We notice that, in Eq.(1), the τ is a fixed hyperparameter during training, which is used to rescale the logits from both the full classifier and the sampled intermediate classifier. This can likely act as a bottleneck towards fully harnessing the potential of self-distillation. For instance, it ignores the fact that during self-distillation: (1) the full classifier is not static and it is learning and, (2) during early training, the predictions from both the full and intermediate classifiers are unreliable. It can lead to model overfitting to source domains causing more reliance on the brittle, non-generalizable domain-specific features (Fig. 1).

To this end, inspired by [27], we propose to replace this fixed rescaling with an adaptive convex combination of logits vector (from the full classifier or sampled intermediate classifier) with one-hot ground truth vector. As training evolves, this combination gradually increases the influence of the model's own knowledge in the self-distillation process. Fig. 2 displays the overall architecture of our method. We believe that this would better allow enhancing the intermediate feature routes via the process of self-distillation (Fig. 3), thereby facilitating the learning of cross-domain generalizable features.

Adaptive convex combination: Let $\mathbf{z}$ be the logit vector produced by the full/intermediate classifier and let $\mathbf{y}$ be the one-hot vector representation corresponding to ground-truth label y. We propose an adaptive convex combination of both $\mathbf{z}$ and $\mathbf{y}$ to generate the soft prediction from the full/intermediate classifier as: $\beta\mathbf{z} + (1-\beta)\mathbf{y}$, where β is a mixing coefficient that determines how much the model should trust its own prediction from the full network or intermediate classifier.

We now discuss how to set the value of β. The full classifier along with the intermediate classifier does not stay fixed during training and they are constantly evolving. Any fixed value of β would become sub-optimal at some iteration during training. As discussed earlier, a model is mostly unreliable during the early phases of the training since it has not gone through the data enough times to be able to generate reliable predictions. Therefore, we make β as a function of training evolution. The β is gradually increased as the training progresses to reflect the increasing reliability of the model. Specifically, the β at t^{th} training iteration is computed as: $\beta_t = \beta_T \times \frac{t}{T}$ where T is the total number of training iterations and β_T is the β_t at final iteration. Note that, β_T is the only hyperparameter to be sought using the validation set. With β_t, our proposed convex combination becomes adaptive and is now expressed as: $\beta_t\mathbf{z} + (1-\beta_t)\mathbf{y}$. After including our proposed soft prediction, the KL divergence between the outputs of the full classifier and the randomly sampled intermediate classifier can be computed as:

$$L_{\text{KL}}(\mathbf{z}\|\mathbf{z}_j) = \sum_{c=1}^{C} \sigma\left(\beta_t\mathbf{z} + (1-\beta_t)\mathbf{y}\right)_c \log \frac{\sigma(\beta_t\mathbf{z}+(1-\beta_t)\mathbf{y})_c}{\sigma(\beta_t\mathbf{z}_j+(1-\beta_t)\mathbf{y})_c}. \tag{3}$$

4. Experiments

Datasets: Following the DG method for DR classification [3], we evaluate the effectiveness of our proposed method

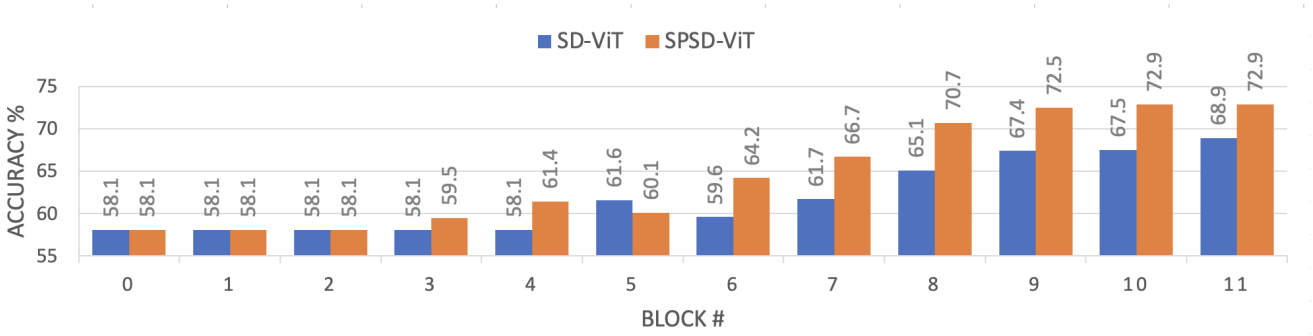

Figure 3. Block-wise accuracy (top-1 %) for SDViT and ours (SPSD-ViT) on Messidor 2 as target domain.

Figure 4. Sample of Images taken from the aptos dataset. Images contain many artefacts such as being out of focus, underexposed, overexposed to light, etc.

on four existing datasets namely, APTOS [1], EYEPACS [22], MESSIDOR and MESSIDOR2 [10]. It should be noted that each dataset has a high class imbalance (e.g. 'No DR' class itself takes up to 74% of the EYEPACS dataset). Each offers 3662, 88702, 1200, and 1744 fundus images, respectively. Also, each dataset has a set of retina images taken under a variety of imaging conditions belonging to 5 classes: no DR, mild DR, moderate DR, severe DR, and proliferative DR. Images contain many artefacts such as being out of focus, underexposed, overexposed to light, etc. Note that, the images in these datasets are collected under multiple clinical settings using various cameras over an extended period of time which introduces more noise and variations (Fig. 4). We consider each dataset as a separate domain to conduct our experiments. We also report results on WildCamelyon [5] dataset which contains histopathological images of breast cancer metastases in lymph node sections taken over 5 hospitals. Each hospital is considered a domain. There are significant variations in each domain that arise from sources such as differences in patient populations, slide staining, and equipment of image acquisition [28,63]. Each domain has two classes, namely tumorous and non-tumorous.

Implementation & training/testing details: We follow the rigorous training and evaluation protocol of DomainBed [17] to allow a fair comparison among methods. We use the same data augmentations proposed in DomainBed [17]: crops of random size and aspect ratio, resizing to 224 × 224 pixels, random horizontal flips, random color jitter, grayscaling the image with 10% probability, and normalization. We consider each dataset as a domain and use the standard training/validation split of 80%/20%. We use AdamW [40] optimizer with its default hyperparameters as in DomainBed [17] for ERM, with a learning rate of 5e-05 and a batch size of 32. We conduct hyperparameter search only for our model-specific hyperparameters λ and β in the range of $\{0.1\ 0.3\ 0.5\ 0.7\ 0.9\}$ and $\{0.2\ 0.4\ 0.6\ 0.8\}$, respectively. We report classification top-1 accuracy (%) on each target domain averaged over three different trials with 3 different seeds. Following [17], we use the training domain validation settings (IID) as our model selection criteria. Initially, each training domain is split into a subset of training and validation and then we pool all the validation subsets of each training domain to create an overall validation set. The model maximizing the accuracy on the overall validation set is selected as the best model. We use PyTorch [49] for implementation and train on 2 V100 GPUs. Table 2 shows that our method adds very little training overhead while signifi-

Method	Backbone(#Param)	Aptos	Eyepacs	Messidor	Messidor 2	Avg.
ERM [62]	ResNet50(23.5M)	47.6±1.7	71.3±0.3	63.0±0.4	69.0±1.5	62.7
IRM [2]	ResNet50	52.1±1.7	73.2±0.3	51.3±3.8	57.2±1.7	58.4
ARM [72]	ResNet50	45.6±1.5	71.7±0.5	62.4±1.0	60.0±3.4	59.9
Fish [56]	ResNet50	44.6±2.2	72.7±0.7	62.1±0.7	66.4±1.7	61.4
Fishr [52]	ResNet50	47.0±1.8	71.9±0.6	63.3±0.5	66.4±0.2	62.2
GroupDRO [53]	ResNet50	44.9±3.8	72.0±0.4	63.1±0.9	67.8±1.9	62.0
MLDG [32]	ResNet50	44.1±1.6	71.9±0.5	62.7±0.6	64.4±0.4	60.8
Mixup [69]	ResNet50	47.3±1.7	72.0±0.3	59.8±2.8	65.8±1.4	61.2
Coral [58]	ResNet50	44.8±2.2	71.7±0.9	58.6±2.8	68.2±0.6	63.2
MMD [36]	ResNet50	49.3±0.1	69.3±1.1	64.6±1.4	69.6±0.6	60.8
DANN [15]	ResNet50	**54.4±0.8**	72.9±0.4	57.0±1.1	58.6±1.7	60.7
CDANN [37]	ResNet50	48.1±0.7	73.1±0.3	55.8±1.8	61.2±1.3	59.5
VREX [29]	ResNet50	49.6±2.3	73.2±0.3	58.5±0.6	65.4±1.6	61.7
SagNet [47]	ResNet50	41.4±3.5	70.9±0.9	60.8±0.3	66.1±0.7	59.8
RSC [20]	ResNet50	46.7±0.6	71.7±0.9	62.5±0.3	66.4±1.5	61.8
SWAD [7]	ResNet50	43.8±2.2	71.6±1.3	58.9±1.7	67.7±2.0	60.5
DRGen [3]	ResNet50	51.2±2.1	72.6±0.8	59.1±1.8	65.2±0.6	62.1
ERM-ViT [62]	DeitSmall(22M)	48.5±0.9	70.7±0.7	62.7±1.6	69.5±2.5	62.9
ERM-ViT [62]	T2T-14(21.5M)	54.0±3.0	73.2±0.4	60.8±1.7	72.0±0.2	62.5
ERM-ViT [62]	CvT-13(20M)	49.3±3.8	69.3±0.1	65.2±0.5	70.6±1.8	63.6
SD-ViT [57]	DeitSmall(22M)	48.2±2.5	69.6±1.5	61.5±0.2	69.4±0.2	62.2
SD-ViT [57]	T2T-14(21.5M)	46.5±0.8	71.7±0.7	63.9±0.9	71.4±0.2	63.4
SD-ViT [57]	CvT-13(20M)	47.8±2.3	70.9±0.8	63.9±1.4	72.4±0.6	63.1
SPSD-ViT(ours)	DeitSmall(22M)	51.6±1.1	73.3±0.3	64.0±0.4	72.9±0.1	65.5
SPSD-ViT(ours)	T2T-14(21.5M)	50.0±2.8	**73.6±0.3**	**65.2±0.3**	**73.3±0.2**	65.5
SPSD-ViT(ours)	CvT-13(20M)	51.7±1.2	73.3±0.2	64.8±0.5	72.4±0.6	**65.6**

Table 1. Multi-Source Domain Generalization Results.

cantly improving the DG capabilities.

Method	Average step time
ERM-ViT [62]	0.353
SD-ViT [57]	0.361
SPSD-ViT(ours)	0.368

Table 2. Training overhead in terms of average step time (secs.),

Evaluation with different ViT backbones: We conduct experiments with three different ViT-based backbones, namely DeiT [60], CvT [68], and T2T-ViT [70] to establish the applicability and generalizability of our method. DeiT is a data-efficient image transformer trained on ImageNet [11] using a student-teacher strategy. Note that, we do not utilize the distillation token and the student-teacher formulation settings in our experiments. We use the DeiT-small model having 22M parameters as our default ViT backbone, unless otherwise specified. The DeiT-small model can be regarded as the ViT counterpart of the ResNet-50 [18] which has 23.5M parameters. CvT [68] improves vision transformer performance and efficiency by introducing convolutions to ViTs to get the best out of both designs. We consider CvT-13 trained on ImageNet [11] for our experiments which has 20M parameters in size. T2T-ViT [70] propose a method to encode the local structure of the surrounding token and to reduce the length of tokens iteratively by relying on progressive tokenization. The T2T-14 model trained on ImageNet [11] containing 21.5M parameters is adopted as the backbone in our experiments. Table 1 shows that our proposed method achieves superior results over all other existing methods irrespective of the backbone architecture used.

Multi-source DG results: We report an extensive comparison with the existing SOTA methods in DG literature on DR datasets as shown in Table 1. We believe that our experiments will offer insights into how the existing SOTA DG methods on natural datasets behave on DR datasets. Moreover, we compare our proposed method with existing SOTA methods [3] in the DR context. We report ERM results with both CNN and ViT backbones as a baseline as it shows competitive performance against many existing DG methods [17]. We achieve a notable **+2.1%** increase in (overall)

Method	Hosp. 0	Hosp. 1	Hosp. 2	Hosp. 3	Hosp. 4	Average
ERM-ViT [62]	97.4±0.5	93.5±0.8	94.5±0.4	96.2±0.3	92.7±0.9	94.9
SD-ViT [57]	97.3±0.1	93.6±0.5	**96.2±0.6**	96.3±0.3	91.6±1.3	95.0
SPSD-ViT(ours)	**97.4±0.1**	**95.3±0.2**	95.8±0.4	**96.4±0.2**	**93.5±0.9**	**95.7**

Table 3. Multi-Source Domain Generalization Results (IID) on WildCamelyon [5] dataset.

Method	Aptos	Eyepacs	Messidor 2	Average
DRGen [3]	41.7±4.3	43.1±7.9	44.8±0.9	43.2
ERM-ViT [62]	45.3±1.3	52.4±3.2	58.2±3.2	51.9
SD-ViT [57]	44.3±0.9	53.2±1.6	57.8±2.4	51.7
SPSD-ViT(ours)	**48.3±1.1**	**57.4±2.1**	**62.2±1.6**	**55.9**

Table 4. Single-source domain generalization results on Messidor dataset.

Method	Aptos	Eyepacs	Messidor	Average
DRGen [3]	40.9±3.9	69.3±1.0	61.3±0.8	57.7
ERM-ViT [62]	47.9±2.1	67.4±0.9	59.6±3.9	58.3
SD-ViT [57]	51.8±0.9	68.7±0.6	**62.0±1.7**	60.8
SPSD-ViT(ours)	**52.8±2.0**	**72.5±0.3**	61.0±0.8	**62.1**

Table 5. Single-source domain generalization results on Messidor 2 dataset.

average accuracy over the second-best contestant.

Method	Eyepacs	Messidor	Messidor 2	Average
DRGen [3]	67.5±1.8	**46.7±0.1**	**61.0±0.1**	58.4
ERM-ViT [62]	67.8±1.4	45.5±0.2	58.8±0.4	57.3
SD-ViT [57]	**72.0±0.8**	45.4±0.1	58.5±0.2	**58.6**
SPSD-ViT(ours)	71.4±0.8	45.6±0.1	58.8±0.2	**58.6**

Table 6. Single-source domain generalization results on Aptos dataset.

Method	Aptos	Messidor	Messidor 2	Average
DRGen [3]	61.3±1.9	**54.6±1.5**	**65.4±0.1**	60.4
ERM-ViT [62]	69.1±1.4	50.4±0.3	62.8±0.2	60.8
SD-ViT [57]	69.3±0.3	50.0±0.5	62.9±0.2	60.7
SPSD-ViT(ours)	**75.1±0.5**	50.5±0.8	62.2±0.4	**62.5**

Table 7. Single-source domain generalization results on Eyepacs. dataset.

Single-source DG results: In the single-source DG setting, we train our model only on one dataset and evaluate on the other 3 datasets. The results in Table 4,5,6,7 show that our method achieves superior performance in all the datasets even under heavy class-imbalance. Single-source DG is even more challenging in messidor dataset as the model does not have any data for class id 4 (proliferative DR). Table 4 shows that our method can generalize to new unseen domains even under a heavy class-imbalanced scenario with a significant improvement in the performance (+4.2%).

DG capability in other medical images: We also show results for multi-source domain generalization results on the WildCamelyon dataset [5]. This dataset constitutes an extensive collection of histopathological images representing breast cancer metastases. The detailed results can be found in Table 3. Our method SPSD-ViT achieves superior results with 95.7% accuracy over both ERM-ViT and SD-ViT. These results underscore the versatility of our proposed SPSD-ViT method. While the core scope of our study includes only the Diabetic Retinopathy (DR) data, we found that the method's application is not limited to this area.

Calibration performance: We also evaluate the calibration performances of our proposed method under multi-source settings, utilizing two widely accepted evaluation metrics to measure the miscalibration of a model - Expected Calibration Error (ECE) [48], and Static Calibration Error (SCE) [44]. The calibration performance of a predictive model is crucial in decision-making processes, which are an important part of healthcare applications. It establishes the credibility and reliability of the model's predictions. Table 8 illustrates the superior performance of our method compared to established baselines. Notably, our method not only exceeds the baselines in domain generalization (DG) performance, but also shows superior calibration performance. This dual achievement underscores the method's robustness, reinforcing the reliability of its predictions while maintaining high performance levels. This makes our method a promising tool for effective, reliable decision-making in various applications e.g., healthcare.

Ablations on prediction softening and hyperparameter analysis: We show results with possible ablations of our prediction softening in Table 9. In the final classifier only case, logits are softened at the final block only using our adaptive convex combination, and in the intermediate classifier only case, logits are softened at the randomly selected intermediate block. Results show that our prediction softening on both full and intermediate classifier, as proposed, achieves superior results compared to applying only on either the full or the intermediate classifier. In Table 10, we

Dataset	Aptos		Eyepacs		Messidor		Messidor 2	
	ECE	SCE	ECE	SCE	ECE	SCE	ECE	SCE
ERM-ViT [62]	33.80 ± 2.60	16.21 ± 1.40	22.66 ± 1.18	9.88 ± 0.34	25.40 ± 3.69	10.94 ± 1.11	16.88 ± 2.74	8.19 ± 1.02
SD-ViT [57]	27.13 ± 1.03	15.04 ± 2.20	$\mathbf{16.97 \pm 1.77}$	8.31 ± 0.14	22.22 ± 1.73	10.12 ± 0.83	12.67 ± 1.92	7.08 ± 0.40
SPSD-ViT(ours)	$\mathbf{23.20 \pm 4.00}$	$\mathbf{14.11 \pm 1.41}$	17.06 ± 2.12	$\mathbf{7.88 \pm 0.99}$	$\mathbf{20.86 \pm 1.84}$	$\mathbf{9.41 \pm 0.47}$	$\mathbf{11.36 \pm 2.77}$	$\mathbf{6.41 \pm 0.47}$

Table 8. Calibration performance in SCE [44] and ECE [48] (in scale of 10^{-2}).

Method	Aptos	Eyepacs	Messidor	Messidor 2	Average
None	48.2±2.5	69.6±1.5	61.5±0.2	69.4±0.2	62.2
Final classifier only	51.5±0.4	**73.5±0.1**	60.7±0.7	69.7±1.7	63.9
Interm. classifier only	49.1±0.4	71.7±0.5	63.0±2.4	68.4±3.4	63.0
Final & interm. classifier (SPSD-ViT)	**51.6±1.1**	73.2±0.3	**64.0±0.4**	**72.9±0.1**	**65.5**

Table 9. Ablation studies on proposed prediction softening.

Hyper-parameters	Aptos	Eyepacs	Messidor	Messidor 2	Average
$\lambda = 0.1$	47.7±0.7	70.9±1.2	63.3±0.2	71.9±0.7	63.5
$\lambda = 0.3$	50.6±1.2	73.2±0.3	60.2±2.0	71.0±0.7	63.7
$\lambda = 0.5$	**52.7±0.8**	**73.5±0.2**	61.2±2.8	71.9±0.3	64.8
$\lambda = 0.7$(Default)	51.6±1.1	73.3±0.7	64.0±0.4	**72.9±0.1**	**65.5**
$\lambda = 0.9$	50.5±0.7	73.4±0.4	**64.5±0.3**	72.2±0.5	65.2
Fixed $\beta = 0.5$	44.2±1.9	71.6±1.0	62.0±1.4	71.3±0.3	62.2
$\beta = 0.2$	48.2±0.4	69.6±0.9	63.9±0.2	72.6±0.6	63.5
$\beta = 0.4$	47.3±1.6	68.1±0.7	62.9±1.2	71.7±0.6	62.5
$\beta = 0.6$	**53.1±0.6**	72.5±0.3	63.4±0.9	71.9±0.8	65.2
$\beta = 0.8$(Default)	51.6±1.1	**73.3±0.7**	**64.0±0.4**	**72.9±0.1**	**65.5**
$\beta = 0.1$	49.7±0.8	73.2±0.2	63.5±0.6	72.3±0.6	64.7

Table 10. Detailed results on the sensitivity of SPSD-ViT(ours) to λ and β.

present a detailed analysis focusing on the sensitivity of two key parameters in our method: λ derived from Equation (2), and β which features in our adaptive convex combination. We note that, our method is relatively resilient to minor perturbations in the β and λ parameters, continuing to deliver comparable performance even when these variables deviate slightly from the best-found values. This highlights the stability of our method and suggests that it is not overly reliant on hyperparameter fine-tuning. However, it is notable that constraining the value of β to a fixed level tends to lead to sub-optimal performance, emphasizing the importance of this adaptive parameter. Note that, all these experiments are conducted using the DeiT-small backbone.

5. Conclusion

We present a new DG approach for DR classification based on distilling the model's own knowledge to its intermediate blocks by constructing a new prediction softening scheme, which is an adaptive convex combination of one-hot labels and the model's own knowledge. We reported comprehensive results derived from multiple Diabetic Retinopathy (DR) datasets, with both multi-source and single-source domain generalization (DG) settings, in conjunction with various Vision Transformer (ViT) backbones. These wide-ranging experiments corroborate the effectiveness and versatility of our method against previously established techniques. Beyond delivering superior DG performance (top-1 accuracy), our method also shows improved out-domain calibration performance (ECE and SCE). Notably, the importance of these improved calibration performances cannot be overstated in a number of safety-critical applications, with healthcare being a prime example. In such critical fields, the reliability of a model's prediction is of paramount importance.

References

[1] APTOS. APTOS 2019 Blindness Detection. https://www.kaggle.com/competitions/aptos2019-blindness-detection/data, 2019. 5

[2] Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*, 2019. 6

[3] Mohammad Atwany and Mohammad Yaqub. Drgen: Domain generalization in diabetic retinopathy classification. In *MICCAI*, pages 635–644. Springer, 2022. 1, 2, 3, 4, 6, 7

[4] Yogesh Balaji, Swami Sankaranarayanan, and Rama Chellappa. Metareg: Towards domain generalization using meta-regularization. In *NeurIPS*, pages 998–1008, 2018. 2

[5] Bandi et al. From detection of individual metastases to classification of lymph node status at the patient level: the camelyon17 challenge. *IEEE Transactions on Medical Imaging*, 2018. 5, 7

[6] Fabio M Carlucci, Antonio D'Innocente, Silvia Bucci, Barbara Caputo, and Tatiana Tommasi. Domain generalization by solving jigsaw puzzles. In *CVPR*, pages 2224–2233, 2019. 1, 2

[7] Junbum Cha, Sanghyuk Chun, Kyungjae Lee, Han-Cheol Cho, Seunghyun Park, Yunsung Lee, and Sungrae Park. Swad: Domain generalization by seeking flat minima. *NeurIPS*, pages 22405–22418, 2021. 6

[8] Chaoqi Chen, Jiongcheng Li, Xiaoguang Han, Xiaoqing Liu, and Yizhou Yu. Compound domain generalization via meta-knowledge encoding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7119–7129, June 2022. 2

[9] Liang Chen, Yong Zhang, Yibing Song, Ying Shan, and Lingqiao Liu. Improved test-time adaptation for domain generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 24172–24182, June 2023. 2

[10] Decencière and et al. Feedback on a publicly distributed image database: The messidor database. *Image Analysis and Stereology*, 2014. 5

[11] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009. 6

[12] Yu Ding, Lei Wang, Bin Liang, Shuming Liang, Yang Wang, and Fang Chen. Domain generalization by learning and removing domain-specific features. In *Advances in Neural Information Processing Systems*, volume 35, pages 24226–24239, 2022. 2

[13] Dosovitskiy and et al. An image is worth 16x16 words:transformers for image recognition at scale. In *ICLR*, volume abs/2010.11929, 2021. 2

[14] Qi Dou, Daniel Coelho de Castro, Konstantinos Kamnitsas, and Ben Glocker. Domain generalization via model-agnostic learning of semantic features. In *NeurIPS*, pages 6447–6458, 2019. 2

[15] Ganin et al. Domain-adversarial training of neural networks. *The Journal of Machine Learning Research*, 17(1):2096–2030, 2016. 2, 6

[16] Yaroslav Ganin and Victor Lempitsky. Unsupervised domain adaptation by backpropagation. In *Proceedings of the 32nd International Conference on International Conference on Machine Learning-Volume 37*, pages 1180–1189. JMLR. org, 2015. 1

[17] Ishaan Gulrajani and David Lopez-Paz. In search of lost domain generalization. *ArXiv*, abs/2007.01434, 2021. 2, 3, 5, 6

[18] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *CoRR*, abs/1512.03385, 2015. 6

[19] Hinton et al. Distilling the knowledge in a neural network (2015). *arXiv*, 2, 2015. 3, 4

[20] Zeyi Huang, Haohan Wang, Eric P. Xing, and Dong Huang. Self-challenging improves cross-domain generalization. In *ECCV*, 2020. 6

[21] Yuxuan Jiang, Yanfeng Wang, Ruipeng Zhang, Qinwei Xu, Ya Zhang, Xin Chen, and Qi Tian. Domain-conditioned normalization for test-time domain generalization. In Leonid Karlinsky, Tomer Michaeli, and Ko Nishino, editors, *Computer Vision – ECCV 2022 Workshops*, pages 291–307, Cham, 2023. Springer Nature Switzerland. 2

[22] Kaggle. Diabetic Retinopathy Detection. https://www.kaggle.com/competitions/diabetic-retinopathy-detection/data, 2015. 5

[23] Guoliang Kang, Lu Jiang, Yi Yang, and Alexander G Hauptmann. Contrastive adaptation network for unsupervised domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4893–4902, 2019. 1

[24] Juwon Kang, Sohyun Lee, Namyup Kim, and Suha Kwak. Style neophile: Constantly seeking novel styles for domain generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7130–7140, June 2022. 2

[25] Kempen et al. The prevalence of diabetic retinopathy among adults in the united states. *Archives of ophthalmology*, page 552—563, 2004. 1

[26] Khan et al. Mode-guided feature augmentation for domain generalization. In *BMVC*, 2021. 1, 2

[27] Kyungyul Kim, ByeongMoon Ji, Doyoung Yoon, and Sangheum Hwang. Self-knowledge distillation with progressive refinement of targets. In *CVPR*, 2021. 4

[28] Daisuke Komura and Shumpei Ishikawa. Machine learning methods for histopathological image analysis. *Computational and Structural Biotechnology Journal*, 16:34–42, 2018. 5

[29] Krueger and et al. Out-of-distribution generalization via risk extrapolation (rex). In *ICML*, 2021. 6

[30] PN Sharath Kumar, RU Deepak, Anuja Sathar, V Sahasranamam, and R Rajesh Kumar. Automated detection system for diabetic retinopathy using two field fundus photography. *Procedia computer science*, 93:486–494, 2016. 1

[31] Li et al. Deeper, broader and artier domain generalization. In *ICCV*, 2017. 1, 2

[32] Li et al. Learning to generalize: Meta-learning for domain generalization. In *AAAI Conference on Artificial Intelligence*, 2018. 2, 6

[33] Li et al. Episodic training for domain generalization. *In ICCV*, 2019. 1, 2

[34] Li and et al. Domain generalization for medical imaging classification with linear-dependency regularization. In *NeurIPS*, 2020. 3

[35] Chenxin Li, Xin Lin, Yijin Mao, Wei Lin, Qi Qi, Xinghao Ding, Yue Huang, Dong Liang, and Yizhou Yu. Domain generalization on medical imaging classification using episodic training with task augmentation. *Computers in biology and medicine*, 141:105144, 2022. 3

[36] Haoliang Li, Sinno Jialin Pan, Shiqi Wang, and Alex C Kot. Domain generalization with adversarial feature learning. In *CVPR*, 2018. 2, 6

[37] Ya Li, Xinmei Tian, Mingming Gong, Yajing Liu, Tongliang Liu, Kun Zhang, and Dacheng Tao. Deep domain generalization via conditional invariant adversarial networks. In *ECCV*, 2018. 6

[38] Liu et al. Feddg: Federated domain generalization on medical image segmentation via episodic learning in continuous frequency space. In *CVPR*, 2021. 3

[39] Mingsheng Long, Han Zhu, Jianmin Wang, and Michael I Jordan. Unsupervised domain adaptation with residual transfer networks. In *Advances in Neural Information Processing Systems*, pages 136–144, 2016. 1

[40] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *ICLR*, 2019. 5

[41] Fangrui Lv, Jian Liang, Shuang Li, Bin Zang, Chi Harold Liu, Ziteng Wang, and Di Liu. Causality inspired representation learning for domain generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8046–8056, June 2022. 2

[42] Saeid Motiian, Marco Piccirilli, Donald A Adjeroh, and Gianfranco Doretto. Unified deep supervised domain adaptation and generalization. In *ICCV*, 2017. 1

[43] Krikamol Muandet, David Balduzzi, and Bernhard Schölkopf. Domain generalization via invariant feature representation. In *ICML*, 2013. 2

[44] Mahdi Pakdaman Naeini, Gregory F. Cooper, and Milos Hauskrecht. Obtaining well calibrated probabilities using bayesian binning. *Proceedings of the ... AAAI Conference on Artificial Intelligence. AAAI Conference on Artificial Intelligence*, 2015:2901–2907, 2015. 7, 8

[45] Nair and et al. Categorization of diabetic retinopathy severity levels of transformed images using clustering approach. *Int J Comput Sci Eng*, 2019. 1

[46] Nair and et al. Classification of diabetic retinopathy severity levels of transformed images using k-means and thresholding method. *Int J Eng Adv Technol*, 2019. 1

[47] Nam et al. Reducing domain gap by reducing style bias. In *CVPR*, pages 8690–8699. 6

[48] Jeremy Nixon, Michael W. Dusenberry, Linchuan Zhang, Ghassen Jerfel, and Dustin Tran. Measuring calibration in deep learning. In *CVPR Workshops*, pages 38–41. Computer Vision Foundation / IEEE, 2019. 7, 8

[49] Paszke et al. Pytorch: An imperative style, high-performance deep learning library. *arXiv*, 2019. 5

[50] Xiaorong Qin, Xinhang Song, and Shuqiang Jiang. Bi-level meta-learning for few-shot domain generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15900–15910, June 2023. 2

[51] C Raja and L Balaji. An automatic detection of blood vessel in retinal images using convolution neural network for diabetic retinopathy detection. *PRIA*, 2019. 1

[52] Alexandre Rame, Corentin Dancette, and Matthieu Cord. Fishr: Invariant gradient variances for out-of-distribution generalization. In *ICML*, 2022. 6

[53] Shiori Sagawa, Pang Wei Koh, Tatsunori B Hashimoto, and Percy Liang. Distributionally robust neural networks for group shifts. In *ICLR*, 2020. 6

[54] Anila Sebastian, Omar Elharrouss, Somaya Al-Maadeed, and Noor Almaadeed. A survey on deep-learning-based diabetic retinopathy classification. *Diagnostics*, 13(3):345, 2023. 1

[55] Shiv Shankar, Vihari Piratla, Soumen Chakrabarti, Siddhartha Chaudhuri, Preethi Jyothi, and Sunita Sarawagi. Generalizing across domains via cross-gradient training. *arXiv*, 2018. 2

[56] Yuge Shi, Jeffrey Seely, Philip H. S. Torr, N. Siddharth, Awni Hannun, Nicolas Usunier, and Gabriel Synnaeve. Gradient matching for domain generalization. *arXiv*, 2021. 6

[57] Maryam Sultana, Muzammal Naseer, Muhammad Haris Khan, Salman Khan, and Fahad Shahbaz Khan. Self-distilled vision transformer for domain generalization. In *ACCV*, 2022. 3, 6, 7, 8

[58] Baochen Sun and Kate Saenko. Deep coral: Correlation alignment for deep domain adaptation. In *European conference on computer vision*, pages 443–450. Springer, 2016. 1, 6

[59] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In *ICML*, 2021. 2, 3

[60] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Herve Jegou. Training data-efficient image transformers and amp; distillation through attention. In *ICML*, 2021. 6

[61] Shikhar Tuli, Ishita Dasgupta, Erin Grant, and Thomas L Griffiths. Are convolutional neural networks or transformers more like human vision? *arXiv*, 2021. 2

[62] Vladimir Vapnik. *The nature of statistical learning theory*. Springer science & business media, 1999. 3, 6, 7, 8

[63] Mitko Veta, Paul J Van Diest, and Mehdi Jiwa. Shaimaa al-janabi, and josien pw pluim. mitosis counting in breast cancer: Object-level interobserver agreement and comparison to an automatic method. *PloS one*, 11(8):e0161286, 2016. 5

[64] Pengfei Wang, Zhaoxiang Zhang, Zhen Lei, and Lei Zhang. Sharpness-aware gradient matching for domain generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3769–3778, June 2023. 2

[65] Shujun Wang, Lequan Yu, Caizi Li, Chi-Wing Fu, and Pheng-Ann Heng. Learning from extrinsic and intrinsic supervisions for domain generalization. In *ECCV*, 2020. 2

[66] Yufei Wang, Haoliang Li, Lap-pui Chau, and Alex C Kot. Embracing the dark knowledge: Domain generalization using regularized knowledge distillation. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 2595–2604, 2021. 3

[67] Wu et al. Cvt: Introducing convolutions to vision transformers. In *ICCV*, 2021. 2

[68] Haiping Wu, Bin Xiao, Noel Codella, Mengchen Liu, Xiyang Dai, Lu Yuan, and Lei Zhang. Cvt: Introducing convolutions to vision transformers, 2021. 6

[69] Shen Yan, Huan Song, Nanxiang Li, Lincan Zou, and Liu Ren. Improve unsupervised domain adaptation with mixup training. *arXiv preprint arXiv:2001.00677*, 2020. 1, 6

[70] Li Yuan, Yunpeng Chen, Tao Wang, Weihao Yu, Yujun Shi, Zi-Hang Jiang, Francis EH Tay, Jiashi Feng, and Shuicheng Yan. Tokens-to-token vit: Training vision transformers from scratch on imagenet. In *ICCV*, 2021. 6

[71] Hanlin Zhang, Yi-Fan Zhang, Weiyang Liu, Adrian Weller, Bernhard Schölkopf, and Eric P. Xing. Towards principled disentanglement for domain generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8024–8034, June 2022. 2

[72] Marvin Zhang, Henrik Marklund, Nikita Dhawan, Abhishek Gupta, Sergey Levine, and Chelsea Finn. Adaptive risk minimization: Learning to adapt to domain shift. In *NeurIPS*, 2021. 6

[73] Zhou et al. Learning to generate novel domains for domain generalization. In *ECCV*, 2020. 1, 2